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.

May 25, 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 Subtypes is 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 data frame 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 'cgdsr', 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.

The '`TCGA`' is a list object with four elements, including the '`geneexpr`' data frame of z scores for the mRNA expression, the '`t`' vector of the survival time, the '`d`' vector of the survival status indicator, and the '`pathList`' list of the pathway information. The '`pathList`' has four elements, each of which contains names of genes belonging to each pathway. Z scores for the mRNA expression data of 389 genes are provided for 50 cancer patients, along with survival times and survival statuses.

This dataset can be loaded as follows:

```
R> data(TCGA)
R> TCGA$geneexpr[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> TCGA$t[1:5]

[1] 43.89 40.97 49.12  2.00 46.59

R> TCGA$d[1:5]

[1] 1 1 0 1 0

R> TCGA$pathList[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

To eliminate the most unlikely predictors, we first conduct 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 are included in the subsequent analysis. By default, p = 0.5 is used as cut-off point.

```
R> prefilter.results <- prefilter(data=TCGA$geneexpr, time=TCGA$t, status=TCGA$d, plist=TCGA$pa
R> prefilter.results

Summary: Pre-filtering results (class: Prefiltered)
-------------------------------------------------
Number of genes before prefiltering: 389
Number of genes after prefiltering: 213
-------------------------------------------------
```

4 Gene Selection

This section is about gene-level analysis using a SPLS Cox regression[3]. 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.

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 can search 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 set fold as 5, '`K`' as 5, and search for '`eta`' between 0.1 and 0.9 with the following command:

```
R> gene.results <- selectGene(prefilter.results)

R> gene.results

Summary: Gene-level analysis results (class: FitGene)
--------------------------------------------------
Number of prefiltered genes: 213
Number of selected genes: 132
--------------------------------------------------

R> head(coef(gene.results)[[1]])

  gene       coef
1 ACLY  0.0000000
2   CS  0.0000000
3 DLAT  0.0000000
4  DLD  0.0000000
5 MDH1 -0.3560516
6 PCK1 -0.1988449
```

5 Pathway Selection

Next, in order to identify a parsimonious set of pathways associated with patient survivals, we fit a LASSO-penalized Cox regression[4] on latent components derived from all the pathways. Specifically, a pathway was selected if at least one of its latent components had non-zero LASSO coefficient estimate.

Hazard ratio plot associated with each latent component in selected pathways can be shown using the function `plot()` with the argument `type= "HR"`.

```
R> plot(path.results, type="HR")
```

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 ($HR \geq 1.15$) are `KEGG_CITRATE_CYCLE_TCA_CYCLE` and `KEGG_TGF_BETA_SIGNALING_PATHWAY` pathways.

This approach has the following two strengths: 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)
R> path.results

Summary: Pathway-level analysis results (class: FitPath)
-------------------------------------------------
Number of all pathways: 4
Number of selected pathways: 4

List of selected pathways:
        KEGG_CITRATE_CYCLE_TCA_CYCLE:
        KEGG_MAPK_SIGNALING_PATHWAY:
        KEGG_TGF_BETA_SIGNALING_PATHWAY:
        KEGG_THYROID_CANCER:
-------------------------------------------------

R> head(coef(path.results))

                          pathway       coef
1    KEGG_CITRATE_CYCLE_TCA_CYCLE 0.44388849
2    KEGG_CITRATE_CYCLE_TCA_CYCLE 0.00000000
3    KEGG_CITRATE_CYCLE_TCA_CYCLE 0.00000000
4    KEGG_CITRATE_CYCLE_TCA_CYCLE 0.00000000
5     KEGG_MAPK_SIGNALING_PATHWAY 0.08378862
6 KEGG_TGF_BETA_SIGNALING_PATHWAY 0.37094205
```

6 Risk Group Prediction

Risk group predictions can be made using the function `predict()`

```
R> predicted <- predict(path.results)
R> predicted[1:3]

$risk.index
 [1] 3 4 0 4 2 4 4 0 0 1 4 0 2 3 3 3 0 1 0 2 1 0 1 1 2 2 2 0 0 0 2 0 1 0 0 0 1 0
[39] 4 4 2 4 4 4 4 4 3 4 3 3

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

$cuts
[1] 0.00 3.75
```

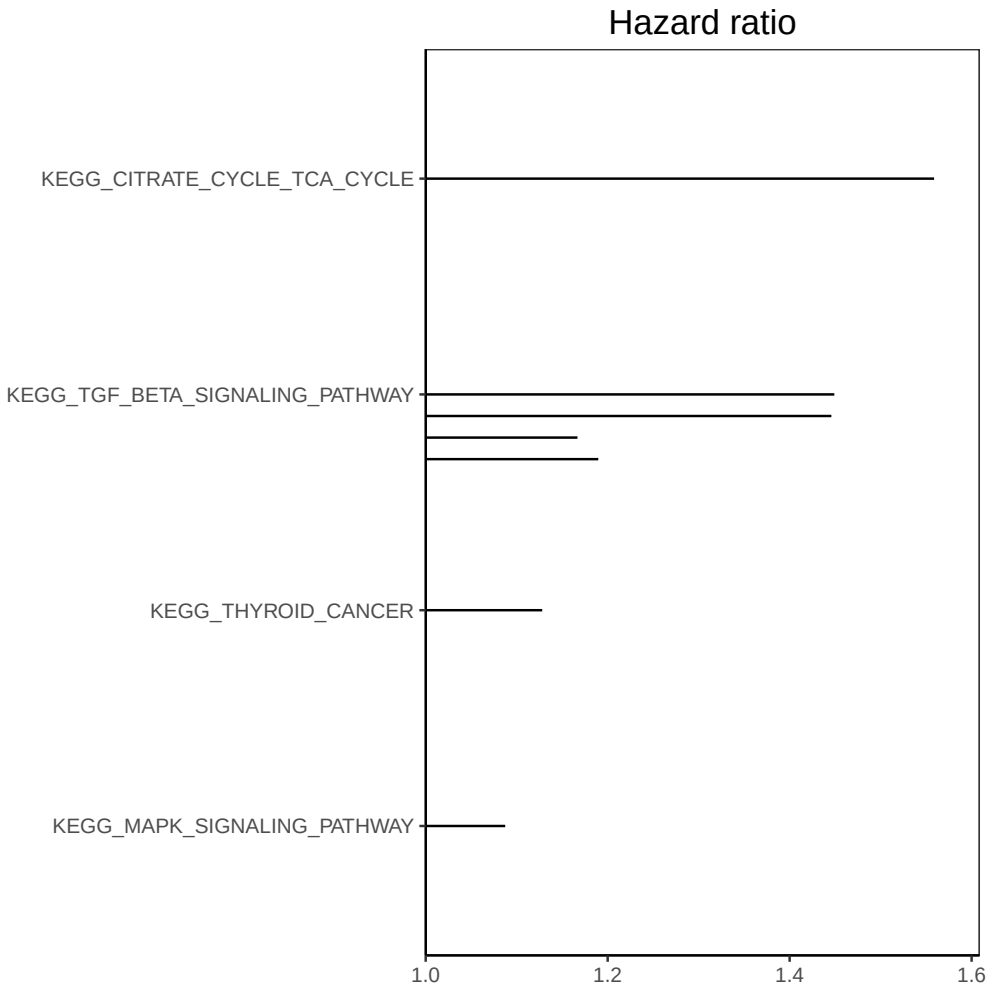

Figure 1: Hazard ratio (HR) associated with each latent component in selected pathways

7 Survival Curve

The predictive performance of `PICS` method can be presented by Kaplan-Meier curves. Kaplan-Meier curves of predicted patient subgroups based on the `PICS` approach can be shown with `plot()` function with argument `type= "KM"`.

```
R> plot(path.results,type="KM")
```

Figure 2 shows the Kaplan-Meier curves of predicted patient subgroups based on the `PICS` approach. The `PICS` approach distinguish the high, medium and low risk group from each other.

8 Survival ROC

The predictive performance of `PICS` method can be further evaluated based on area under the time dependent receiver operating curve (ROC). ROC plot can be shown using `plot()` function with argument `type= "ROC"`.

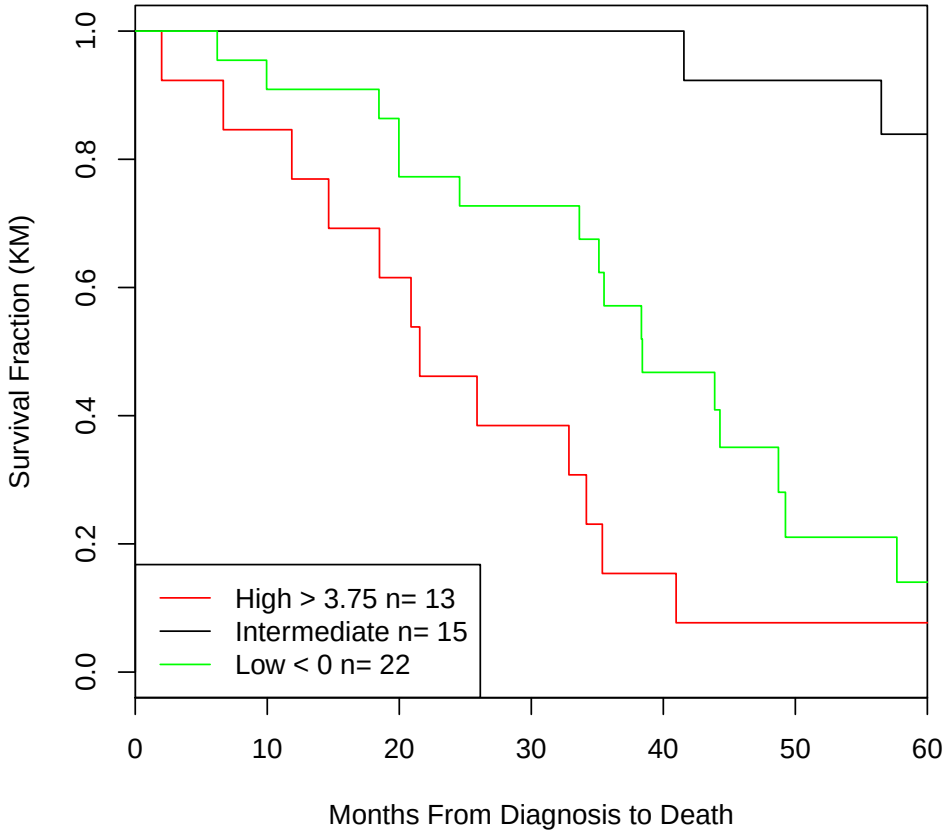

Figure 2: The observed survival curves for patient subgroups identified by the `PICS`

```
R> plot(path.results,type="ROC")
```

Figure 3 shows the ROC curves for survival and demonstrates an advantage of the `PICS` approach. For the TCGA data, the area under curve (AUC) associated with the `PICS` approach was 0.886.

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] Bastien, P., Bertrand, F., Meyer, N., Maumy-Bertrand, M. (2014), " Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data". Bioinformatics, 31(3), 397-404.

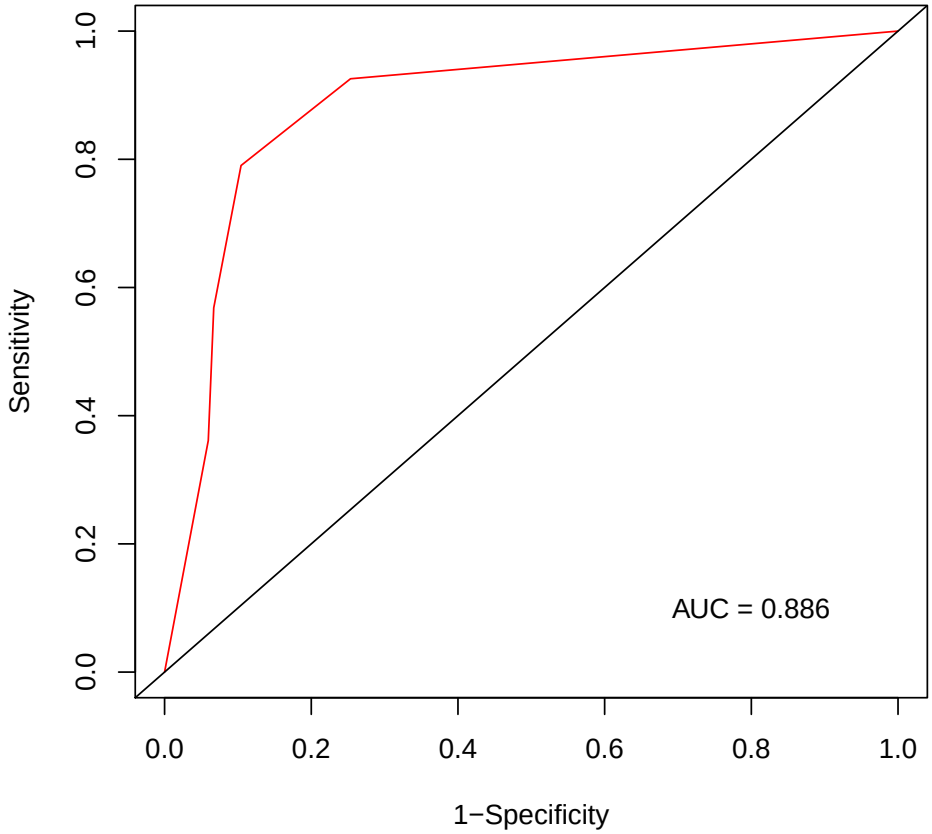

Figure 3: Time dependent Receiver Operating Curves

[4] Tibshirani, R. (1997), "The Lasso Method For Variable Selection In The Cox Model". Statistics in Medicine, 16(4), 385-395.