PenPC: A Two-step Approach to Estimate the Skeletons of High Dimensional Directed Acyclic Graphs

Min Jin Ha

December 30, 2014

1 Overview

```
> library(PenPC)
```

This vignette describes how to use `R/PenPC` to estimate the skeleton of a high-dimensional directed acyclic graph (DAG) by a two-step approach. We first estimate the non-zero entries of a concentration matrix using penalized regression implemented in `R/PEN` package, and then fix the difference between the concentration matrix and the skeleton by evaluating a set of conditional independence tests.

2 An example

We illustrate the usage of `PenPC` package using simulation data. We simulate a random DAG following Barabási,A. and Albert, R. (1999). Specifically, the initial graph had one vertex and no edge. In the (t+1)-th step, e edges were proposed. For each edge, the new vertex was connected to the i-th ($1 \leq i \leq t$) existing vertex with probability $\nu_i^{(t)} / \sum_j \nu_j^{(t)}$, where $\nu_i^{(t)}$ is the degree of i at the t-th step. After constructing the DAG, observed data are simulated by structure equations under multivariate Gaussian assumption. For example, denote the parents of vertex j by $\mathtt{pa}_j$, then the $n \times 1$ vector of n observations for j, denoted by $\mathbf{x}_j$, is generated from $\mathbf{x}_j = \sum_{k \in \mathtt{pa}_j} b_{jk} \mathbf{x}_k + \epsilon_j$, where $\epsilon_j \sim N(0, I_{n \times n})$ and $b_{jk} \sim \text{Unif}(0.1, 1)$. The following is the example of generating the simulation data with the number of vertices, $p = 100$, the sample size, $n = 30$, and $e = 1$.

```
> p = 100
> n = 30
> e = 1
> simul=simul.BA(p,e,n)
```

The $p \times p$ adjacency matrix for the underlying DAG, $n \times p$ data matrix, and the underlying graph are displayed.

```
> dim(simul$A)
```

```
[1] 100 100
```

```
> dim(simul$X)
```

```
[1]  30 100
```

```
> simul$G
```

```
A graphNEL graph with directed edges
Number of Nodes = 100
Number of Edges = 99
```

In the first step of the PenPC, we estimate the non-zero entries of a concentration matrix by neighborhood selection. We select the neighborhood of vertex i by a penalized regression with i as a response and all other vertices as covariates. For the penalized regression, we employ the log penalty with two tuning parameters λ and τ, $p(|b|;\lambda,\tau) = \lambda \log(|b| + \tau)$, which was solved by a coordinate decent algorithm (Sun, Wei and Ibrahim, Joseph G and Zou, Fei , 2010). The two tuning parameters λ and τ are selected by two-grid search to minimize extended BIC (Chen, J. and Chen, Z. , 2008). In the following example code, we perform the neighborhood selection for all p vertices from 100 and 10 candidate λ and τ values. By setting `order=TRUE`, we perform the coordinate decent algorithm after sorting the covariates in the decreasing order of absolute correlation with the response.

```
> dat = simul$X
> coeff = ne.PEN(dat=dat,nlambda=100,ntau=10,V=1:p,order=TRUE)
> sum(coeff!=0)
```

```
[1] 119
```

The neighborhood selection for a selected set of vertices can be performed by setting `V` option.

```
> coeff.sel= ne.PEN(dat=dat,nlambda=100,ntau=10,V=c(1,2,3),order=TRUE)
```

After the p penalized regressions for each of the p vertices, we construct the structure of zeros in the concentration matrix (the moral graph) of the p vertices by adding an edge between vertices i and j if $\hat{b}_{ij} \neq 0$ or $\hat{b}_{ji} \neq 0$ where $\hat{b}_{ij}$ is the estimate of the regression coefficient for j in the penalized regression with i as the response.

```
> edgeWeights = matrix(0,p,p)
> edgeWeights[coeff!=0|t(coeff)!=0] =1
```

In the second step of the PenPC algorithm, we estimate the skeleton of the DAG, starting from the moral graph implied in the structure of zeros in the concentration matrix. To exclude co-parent edges, we perform a series of conditional independence tests using the p-value cutoff, 0.01.

```
> alpha = 0.01
> indepTest = gaussCItest
> suffStat  = list(C = cor(dat), n = n)
> fit.penpc  = skeletonPENstable(suffStat, indepTest, as.integer(p), alpha,
+ edgeWeights=edgeWeights, verbose=F)
```

```
Tue Dec 30 12:55:48 2014 : order= 0 , # of edges remaining =  70
Tue Dec 30 12:55:48 2014 : order= 1 , # of edges remaining =  41
Tue Dec 30 12:55:48 2014 : order= 2 , # of edges remaining =  41
Tue Dec 30 12:55:48 2014 : order= 3 , # of edges remaining =  41
Tue Dec 30 12:55:48 2014 : order= 4 , # of edges remaining =  41
Tue Dec 30 12:55:48 2014 : order= 5 , # of edges remaining =  41
```

```
> fit.penpc@graph
```

```
A graphNEL graph with undirected edges
Number of Nodes = 100
Number of Edges = 41
```

References

Barabási, A.L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286,** 509-512.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95,** 759-771.

Sun, Wei and Ibrahim, Joseph G and Zou, Fei (2010). Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression . *Genetics*, **185,** 349-359.