Dynamic Network Regression Using R Package dnr

Abhirup Mallik

July 26, 2018

R package 'dnr' enables the user to fit dynamic network regression models for time variate network data available mostly in social sciences or social network analysis using the methodology described in [1]. In this document, we demonstrate the process of building a model to fit a dynamic network data set and using that model for prediction.

1 Analysis of Beach data

First, we consider the beach data for our demo.

```r
suppressMessages(library(dnr))
data(beach)

## get the number of time points
length(beach)

## [1] 31

## network size (that allows NA)
network.size.1 <- function(x){
  if(!network::is.network(x)){
    if(is.na(x)) return(0)
  } else return(network::network.size(x))
}

## get the size of networks at each time point
sapply(beach, network.size.1)
```

The beach data is a rapidly changing data set with possible periodic effects. We visualize the adjacency matrix from four time points of the data.
For vertex model, we define our own term dictionary. We use the similar approach as the edge model for specifying the time dependence of the terms using a matrix of lag terms.
Table 1: Index of the terms for specifying the vertex model

<table>
<thead>
<tr>
<th>Term</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>degree (Freeman)</td>
<td>1</td>
</tr>
<tr>
<td>in degree</td>
<td>2</td>
</tr>
<tr>
<td>out degree</td>
<td>3</td>
</tr>
<tr>
<td>Eigen centrality</td>
<td>4</td>
</tr>
<tr>
<td>Between centrality</td>
<td>5</td>
</tr>
<tr>
<td>Info centrality</td>
<td>6</td>
</tr>
<tr>
<td>Closeness centrality</td>
<td>7</td>
</tr>
<tr>
<td>Log K cycle</td>
<td>8</td>
</tr>
<tr>
<td>Log size</td>
<td>9</td>
</tr>
</tbody>
</table>

1.1 Model Fitting

We first try to build the model for vertex regression. We consider a maximum lag of 3. We need to specify the lag structure using a binary vector of size 3. We also need to specify the dependence on the vertex parameters up to 3 lags. There are 9 vertex parameters available in the current version of the library. We use a binary matrix of size $3 \times 9$ for specifying the lag dependence of the parameters.

```r
define the model fitting
nvertexstats <- 9
maxLag = 3
VertexLag = rep(1, maxLag)
VertexLagMatrix1 <- matrix(1, maxLag, nvertexstats)
VertexLagMatrix1

As for this data set there is expected seasonal effect, for example weekends would have different effect than weekdays, we would like to model that using a time variate intercept parameter. We write a function to extract the day information from the data.

```r
getWeekend <- function(z){
  weekends <- c("Saturday", "Sunday")
  if(!network::is.network(z)){
    if(is.na(z)) return(NA)
  } else {
    zDay <- get.network.attribute(z, attrname = "day")
    out <- ifelse(zDay %in% weekends, 1, 0)
    return(out)
  }
}
```
} 

```r
## for(i in 1:31) print(getWeekend(beach[[i]]))
## generate a vector of network level exogenous variable
dayClass <- numeric(length(beach))
for(i in seq_along(dayClass)) {
  dayClass[i] <- getWeekend(beach[[i]])
}

We then use the function paramVertexOnly() to fit the model specified above to the beach data. Most of the options are kept at their default value. For a detail description of the model specification, please refer to the help pages of the function. We use the default 'bayesGLM' option for the logistic regression. We print the model object, which is an object from arm package, with its own summary method.

```r
out <- paramVertexOnly(InputNetwork = beach, 
  maxLag = 3, 
  VertexStatsvec = rep(1, nvertexstats), 
  VertexLag = rep(1, maxLag), 
  VertexLagMatrix = VertexLagMatrix1, 
  dayClass = dayClass)

summary(out$VertexFit$fit)
```

```r
## Call:
## arm::bayesglm(formula = y ~ . - 1, family = binomial(link = "logit"), 
## data = XYdata)
##
## Deviance Residuals:
##    Min       1Q   Median       3Q      Max
## -2.0506  -1.1774  -1.0569  -0.6421   2.5063
##
## Coefficients:
##             Estimate Std. Error  z value  Pr(>|z|)
## lag1       1.028333   0.769727   1.336  0.18156
## lag2       1.572503   0.892916   1.761  0.07822 .
## lag3       0.843973   0.764983   1.103  0.26992
## Day       -0.852764   0.091841  -9.285  < 2e-16 ***
## DegreeLag1. 0.021577   0.235425   0.092  0.92698
## InDegreeLag1.043153   0.470850   0.092  0.92698
## OutDegreeLag1.0.043153   0.470850   0.092  0.92698
## EigenCentralityLag1.0.493574   0.524196   0.942  0.34641
## BetweenCentralityLag1. -0.006343  0.005935  -1.069  0.28520
## InfoCentralityLag1. 0.204742   0.141982   1.442  0.14929
## CloseCentralityLag1. 2.724521   0.842090   3.235  0.00121 **
```

4
As we can see the model is hardly parsimonious. So, we decided to remove the terms that were not significant. We report the result of the refitted model.

```r
VertexLagMatrix <- matrix(0, maxLag, nvertexstats)
VertexLagMatrix[, c(4, 7)] <- 1
VertexLagMatrix[c(2, 3), 7] <- 0
VertexLagMatrix
```

As we can see the model is hardly parsimonious. So, we decided to remove the terms that were not significant. We report the result of the refitted model.
maxLag = 3,
VertexStatsvec = rep(1, nvertexstats),
VertexLag = rep(1, maxLag),
VertexLagMatrix = VertexLagMatrix,
dayClass = dayClass)

summary(out$VertexFit$fit)

##
## Call:
## arm::bayesglm(formula = y ~ . - 1, family = binomial(link = "logit"),
## data = XYdata)
##
## Deviance Residuals:
##     Min 1Q Median 3Q Max
## -1.9023 -1.1774 -1.0664 -0.7149 2.6037
##
## Coefficients:
##                  Estimate Std. Error   z value  Pr(>|z|)
## lag1           -0.49443  0.17426   -2.837 0.004549 **
## lag2           -1.23852  0.19443  -6.370 1.89e-10 ***
## lag3           -1.23395  0.19007  -6.492 8.46e-11 ***
## Day           -0.83117  0.09079  -9.155 < 2e-16 ***
## EigenCentralityLag1  1.20172  0.28613    4.200 2.67e-05 ***
## CloseCentralityLag1  2.53228  0.73493    3.446 0.000570 ***
## EigenCentralityLag2  1.30669  0.31145    4.196 2.72e-05 ***
## EigenCentralityLag3  1.18579  0.31646    3.747 0.000179 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 3555.8  on 2565  degrees of freedom
## Residual deviance: 3268.0  on 2557  degrees of freedom
## AIC: 3284
##
## Number of Fisher Scoring iterations: 4

Now, we have a model with mostly significant parameters, so we select this model for vertex generation.

As the edge model and vertex model are separable, we can expect this model to work for the joint model as well. We use the function paramVertex() for fitting the joint vertex-edge model to the beach data.
out <- paramVertex(InputNetwork = beach,
  maxLag = 3,
  VertexStatsvec = rep(1, nvertexstats),
  VertexModelGroup = "regular",
  VertexLag = rep(1, maxLag),
  VertexLagMatrix = VertexLagMatrix,
  dayClass = dayClass,
  EdgeModelTerms = NA,
  EdgeModelFormula = NA,
  EdgeGroup = NA,
  EdgeIntercept = c("edges"),
  EdgeNetparam = c("logSize"),
  EdgeExvar = NA,
  EdgeLag = c(1, 1, 0),
  paramout = TRUE)

summary(out$VertexFit$fit)

## Call:
## arm::bayesglm(formula = y ~ . - 1, family = binomial(link = "logit"),
##     data = XYdata)

## Deviance Residuals:
##    Min 1Q Median 3Q Max
## -1.9397 -1.1774 -1.0793 -0.6249 3.6880

## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## lag1          -1.69546   0.33745  -5.024  5.05e-07 ***
## lag2          -3.20981   0.59823  -5.365  8.07e-08 ***
## lag3          -2.79034   0.47850  -5.831  5.49e-09 ***
## Day           -0.79949   0.09136  -8.751   < 2e-16 ***
## attribLag1     1.44810   0.34010   4.258   2.06e-05 ***
## attribLag2     2.23119   0.60271   3.702   0.000214 ***
## attribLag3     1.79800   0.47927   3.752   0.000176 ***
## EigenCentralrtyLag1   1.07029   0.29102   3.678   0.000235 ***
## CloseCentralrtyLag1   2.37304   0.74049   3.205   0.001352 **
## EigenCentralrtyLag2   1.04992   0.31895   3.292   0.000996 ***
## EigenCentralrtyLag3   0.97343   0.32232   3.020   0.002527 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## (Dispersion parameter for binomial family taken to be 1)

## Null deviance: 3555.8 on 2565 degrees of freedom
The edge model parameters are specified using 'EdgeIntercept' term, as we are using an intercept only model. For this example, we have tried using time variate parameters, but finally decided on the intercept only model. The term 'EdgeNetParam' indicates the network level attribute. Currently the only attribute supported here is 'logSize', which is log of the network size at the present time point. The binary vector 'EdgeLag' indicates the lag dependence of the edges. The terms 'EdgeModelTerms' and 'EdgeModelFormula' has not been used for this example.
1.2 Prediction for Beach Data

As we have finalized on a model for the beach data, we can use this model to predict the future networks up to any arbitrary number of time points. As long as we do not run into the problems of degeneracy, the simulation method should be able to generate networks with this model.

```r
suppressWarnings(simResult <- engineVertex(InputNetwork = beach,
numSim = 3,
maxLag = 3,
VertexStatsvec = rep(1, nvertexstats),
VertexModelGroup = "regular",
VertexAttLag = rep(1, maxLag),
VertexLag = rep(1, maxLag),
VertexLagMatrix = VertexLagMatrix,
dayClassObserved = dayClass,
dayClassFuture = c(1, 0, 0, 0, 0),
EdgeModelTerms = NA,
EdgeModelFormula = NA,
EdgeGroup = NA,
EdgeIntercept = c("edges"),
EdgeNetparam = c("logSize"),
EdgeExvar = NA,
EdgeLag = c(0, 1, 0),
paramout = TRUE ))

## [1] 1

## Evaluating log-likelihood at the estimate.

## [1] 2

## Evaluating log-likelihood at the estimate.

## [1] 3

## Evaluating log-likelihood at the estimate.

par(mfrow = c(2,2))
binaryPlot(beach[[31]][, ], title = "Time point 31")
binaryPlot(simResult$SimNetwork[[1]][, ], title = "Time point 32 (simulated)"
binaryPlot(simResult$SimNetwork[[2]][, ], title = "Time point 33 (simulated)"
binaryPlot(simResult$SimNetwork[[3]][, ], title = "Time point 34 (simulated)"
```
2 Model for Fixed Vertex Case

Even though fixed vertex case can be considered as a special case of dynamic vertex-edge case, it is preferred that the fixed vertex case is handled in a simpler way. We have provided separate functions for this case, that we will demonstrate using the blog data set.

data(rdNets)
length(rdNets)

## [1] 484
rdNets[[1]]

## Network attributes:
## vertices = 47
## directed = TRUE
## hyper = FALSE
## loops = FALSE
## multiple = FALSE
## bipartite = FALSE
## total edges= 182
## missing edges= 0
## non-missing edges= 182

## Vertex attribute names:
## dnc rnc vertex.names

## Edge attribute names:
## frequency

plot(rdNets[[1]])
We use the function `paramest()` to fit the edge model to the blog data. The function accepts ERGM style model formulas, however we require that the terms are expanded by the user. This is required to construct the lag dependence binary matrix. The interface for this function is similar to the dynamic vertex case.

```r
input_network=rdNets[1:6]
model.terms=c("triadcensus.003", "triadcensus.012", "triadcensus.102", "triadcensus.021D"
model.formula = net~triadcensus(0:3)+gwesp(decay=0, fixed=FALSE, cutoff=30)-1;
graph_mode='digraph';
group='dnc';
alpha.glmnet=1
directed=TRUE;
```
method <- 'bayesglm'
maxlag <- 3
lambda=NA
intercept = c("edges")
cdim <- length(model.terms)
lagmat <- matrix(sample(c(0,1),(maxlag+1)*cdim,replace = TRUE),ncol = cdim)
ylag <- rep(1,maxlag)
exvar <- NA
out <- suppressWarnings(paramEdge(input_network,
               model.terms,
               model.formula,
               graph_mode='digraph',
               group,intercept = c("edges"),exvar=NA,
               maxlag = 3,
               lagmat = matrix(sample(c(0,1),(maxlag+1)*cdim,
                                      replace = TRUE),ncol = cdim),
               ylag = rep(1,maxlag),
               lambda = NA, method='bayesglm',
               alpha.glmnet=1))
out$coef
## $coef
##  ## edges  edgecov.dnc11  edgecov.dnc01  edgecov.dnc10
##  -5.93868451 -1.21829474  0.01675488  0.44024470
##  edgecov.dnc00  triadcensus.003  triadcensus.003.1  triadcensus.012.1
##  -0.50213014  0.01525423  0.02936452  0.01556787
##  triadcensus.102.1  gwesp.1  triadcensus.102.2  triadcensus.021D.2
##   0.03517797 -0.21774123 -0.02354351 -0.02185304
##  gwesp.2  triadcensus.021D.3  lag1  lag2
##   0.13165204 -0.04354879  1.24929328  3.74340698
##  lag3
##   9.24430893
##  ## $se
##  ## edges  edgecov.dnc11  edgecov.dnc01  edgecov.dnc10
##  2.00643745  1.23889461  1.20426697  1.20689944
##  edgecov.dnc00  triadcensus.003  triadcensus.003.1  triadcensus.012.1
##  1.52610818  0.03384629  0.06830034  0.04326980
##  triadcensus.102.1  gwesp.1  triadcensus.102.2  triadcensus.021D.2
##  0.05672938  0.61251630  0.04632015  0.14329671
##  gwesp.2  triadcensus.021D.3  lag1  lag2
##  0.61267475  0.13047488  1.68397353  1.79150846
##  lag3
##  1.18981013
##  13
Here the model formula is an ERGM formula. However, we have also provided the expansion of all the terms in the formula. For example the term 'triadcensus(0:3)' has been expanded out to respective triadcensus terms. The 'group' parameter is a categorical attribute for the vertices. This was present in the dynamic vertex case also. The specification of the intercept term is similar as well. The lag terms and lag dependency of the parameters are represented with a binary vector or a binary matrix respectively.

We can use the model chosen to simulate the networks in future time points. Here we are simulating 10 future networks. We have kept the option of specifying the model and the initial network separate unlike the dynamic vertex case. However, using different model than the model fitted on the input network is not recommended as it is easily possible to create examples where these two inputs differ significantly, hurting the performance of the simulation.
method <- 'bayesglm'
maxlag <- 3
lambda=NA
intercept = c("edges")
cdim <- length(model.terms)
lagmat <- matrix(sample(c(0,1),(maxlag+1)*cdim,replace = TRUE),ncol = cdim)
ylag <- rep(1,maxlag)
lagmat[1,] <- rep(0,ncol(lagmat))
out <- suppressWarnings(paramEdge(input_network,model.terms, model.formula,
                               graph_mode="digraph",group,intercept = c("edges"),exvar=NA,
                               maxlag = 3,
                               lagmat = lagmat,
                               ylag = rep(1,maxlag),
                               lambda = NA, method='bayesglm',
                               alpha.glmnet=1))

start_network <- input_network
inputcoeff <- out$coef$coef
nvertex <- 47
ns <- 10
exvar <- NA
input_network <- rdNets[1:6]
maxlag <- 3
start_network <- input_network
inputcoeff <- out$coef$coef
nvertex <- 47
ns <- 10
exvar <- NA
tmp <- suppressWarnings(engineEdge(start_network=start_network,inputcoeff=inputcoeff,ns=
                                   model.terms=model.terms, model.formula=model.formula,
                                   graph_mode=graph_mode,group=group,intercept=intercept,
                                   exvar=exvar,
                                   maxlag=maxlag,
                                   lagmat=lagmat,
                                   ylag=ylag,
                                   lambda = NA, method='bayesglm',
                                   alpha.glmnet=alpha.glmnet))

## [1] 1

## Evaluating log-likelihood at the estimate.
## [1] 2

## Evaluating log-likelihood at the estimate.
## Evaluating log-likelihood at the estimate.

```r
par(mfrow = c(2,2))
binaryPlot(input_network[[1]][, ], title = "Time point 6", axlabs = FALSE)
binaryPlot(tmp$out_network[[1]][, ], title = "Time point 7 (simulated)", axlabs = FALSE)
binaryPlot(tmp$out_network[[2]][, ], title = "Time point 8 (simulated)", axlabs = FALSE)
binaryPlot(tmp$out_network[[3]][, ], title = "Time point 9 (simulated)", axlabs = FALSE)
```
2.1 Time series of parameter estimates

As all the coefficients are calculated as a part of the simulation, they are also provided along with the simulated networks. We can plot the time series of the network parameters to see the quality of the simulations.

```r
plot.ts(tmp$coefmat[, 1:10], xy.labels=FALSE,
        main = "Estimated parameters from simulated networks", cex = 0.8)
```
2.2 Performance metrics

We also provide some functions for assessing the quality of the simulated networks and make comparisons with holdout set or some other benchmarked networks. Specifically, there are functions for number of triangles, cluster coefficient and expectation of degree distribution has been implemented. We report the performance metrics for the input networks as well as the simulated networks for our example on blog data.

```r
perfMetrics <-
  cbind(c(sapply(tmp$out_network, function(net) ntriangles(net[,])),
        sapply(input_network, function(net) ntriangles(net[,]))),
```
Table 2: Performance metrics for input and simulated networks.

<table>
<thead>
<tr>
<th>Triangles</th>
<th>ClustCoefs</th>
<th>ExpDeg</th>
</tr>
</thead>
<tbody>
<tr>
<td>234</td>
<td>0.438</td>
<td>9.255</td>
</tr>
<tr>
<td>265</td>
<td>0.463</td>
<td>9.468</td>
</tr>
<tr>
<td>248</td>
<td>0.470</td>
<td>9.000</td>
</tr>
<tr>
<td>244</td>
<td>0.471</td>
<td>8.915</td>
</tr>
<tr>
<td>248</td>
<td>0.474</td>
<td>8.957</td>
</tr>
<tr>
<td>252</td>
<td>0.473</td>
<td>9.043</td>
</tr>
<tr>
<td>261</td>
<td>0.477</td>
<td>9.170</td>
</tr>
<tr>
<td>255</td>
<td>0.457</td>
<td>9.383</td>
</tr>
<tr>
<td>248</td>
<td>0.474</td>
<td>8.957</td>
</tr>
<tr>
<td>248</td>
<td>0.474</td>
<td>8.872</td>
</tr>
<tr>
<td>248</td>
<td>0.476</td>
<td>8.745</td>
</tr>
<tr>
<td>257</td>
<td>0.482</td>
<td>9.085</td>
</tr>
<tr>
<td>247</td>
<td>0.475</td>
<td>8.915</td>
</tr>
<tr>
<td>246</td>
<td>0.473</td>
<td>8.915</td>
</tr>
<tr>
<td>245</td>
<td>0.472</td>
<td>8.915</td>
</tr>
<tr>
<td>248</td>
<td>0.474</td>
<td>9.000</td>
</tr>
</tbody>
</table>

```r
c(sapply(tmp$out_network, function(net) clustCoef(net[,])), sapply(input_network, function(net) clustCoef(net[,]))), c(sapply(tmp$out_network, function(net) expdeg(net[,])), sapply(input_network, function(net) expdeg(net[,])))
```

```r
colnames(perfMetrics) <- c("Triangles", "ClustCoefs", "ExpDeg")
perfMetrics <- data.frame(perfMetrics, row.names = NULL)
knitr::kable(perfMetrics, digits = 3,
             caption = "Performance metrics for input and simulated networks.")
```

References