MiCoRe Vignette

Kevin McGregor

6/4/2020

Microbiome **Co**variance **Re**gression (**MiCoRe**) allows the estimation of how OTU co-occurrence networks vary with respect to a covariate profile using principles of covariance regression. This work was developed in the Greenwood Lab at McGill University.

Installation

MiCoRe can be installed easily from Github. Note that the name of the R package is all lowercase: `micore`.

```r
if (!require(devtools)) {
  install.packages("devtools")
  library(devtools)
}
install_github("kevinmcgregor/micore", dependencies=TRUE)
```

The model

The goal of **MiCoRe** is to estimate how covariance matrices vary with repsect to a covariate profile in the context of microbiome data.

Assume that the matrix $\mathbf{Y}_{n\times(p+1)}$ contains the counts of $p+1$ taxa over n samples. Taxon $p+1$ will be used as a reference taxon, and will not be included in the estimated covariance matrices. The matrix $\mathbf{X}_{n\times q}$ contains the $q-1$ covariates over which the covariance (or precision) matrix is assumed to vary, along with an intercept column. The vector $\mathbf{x}_i = (1, x_{i1}, \dots, x_{i(q-1)})^\top$ contains the covariates for individual i.

We assume a multinomial logistic regression framework for the taxon counts. We denote the total count for individual i as $M_i = \sum_{j=1}^{p+1} \mathbf{Y}_{ij}$. We also assume that the true proportions of all the taxa in individual i's microbiome is $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{i(p+1)})$, with $0 < \pi_{ij} < 1$ for all $j \in \{1, \dots, p+1\}$ and $\sum_{j=1}^{p+1} \pi_{ij} = 1$. Then we assume the observed counts for individual i, denoted by $\mathbf{Y}_{i\cdot}$, follow a multinomial distribution. The full model is written as.

$$
\begin{aligned}
\mathbf{Y}_{i\cdot} | \boldsymbol{\eta}_{i\cdot}, \mathbf{A}, \mathbf{B}, \gamma_i, \boldsymbol{\Psi}, \boldsymbol{\Gamma} &\sim \text{Multinomial}(M_i, \boldsymbol{\pi}_i) \\
\boldsymbol{\eta}_{i\cdot} | \mathbf{A}, \mathbf{B}, \gamma_i, \boldsymbol{\Psi}, \boldsymbol{\Gamma} &\sim \text{Normal}\left([\mathbf{A} + \gamma_i \mathbf{B}]\mathbf{x}_i, \boldsymbol{\Psi}\right) \\
\mathbf{C} = (\mathbf{A}, \mathbf{B}) | \boldsymbol{\Psi}, \boldsymbol{\Gamma} &\sim \text{Matrix-Normal}\left(\mathbf{C}_0, \boldsymbol{\Psi}, \boldsymbol{\Gamma}\right) \\
\boldsymbol{\Psi} &\sim \text{inv-Wishart}\left(\nu_\Psi, \boldsymbol{\Psi}_0\right) \\
\boldsymbol{\Gamma} &\sim \text{inv-Wishart}\left(\nu_\Gamma, \boldsymbol{\Gamma}_0\right) \\
\gamma_i &\sim \text{Normal}(0, 1). \qquad (1)
\end{aligned}
$$

where the proportions $\boldsymbol{\pi}_i$ are parameterized using a matrix of latent parameters, $\boldsymbol{\eta}_{n\times p}$, whose elements are denoted by η_{ij}:

$$
\boldsymbol{\pi}_i = \left(\frac{\exp(\eta_{i1})}{1 + \sum_{j=1}^p \exp(\eta_{ij})}, \dots, \frac{\exp(\eta_{ip})}{1 + \sum_{j=1}^p \exp(\eta_{ij})}, \frac{1}{1 + \sum_{j=1}^p \exp(\eta_{ij})} \right).
$$

The elements of $\boldsymbol{\eta}$ can be thought of as the additive log-ratio transformed proportions with respect to the reference taxon $p+1$:

$$\boldsymbol{\eta}_{i\cdot} = \left[\log\left(\frac{\pi_{i1}}{\pi_{i(p+1)}}\right), \ldots, \log\left(\frac{\pi_{ip}}{\pi_{i(p+1)}}\right)\right],$$

where $\boldsymbol{\eta}_{i\cdot}$ represents row i of $\boldsymbol{\eta}$.

Interpretations of parameters

Parameter interpretations come from marginalizing out the individual-specific term γ_i. The expected value for $\boldsymbol{\eta}_{i\cdot}$ (i.e. the additive log-ratio transformed proportions for individual i) is written as:

$$\mathbb{E}(\boldsymbol{\eta}_{i\cdot}|\mathbf{A}, \mathbf{B}, \boldsymbol{\Psi}, \boldsymbol{\Gamma}) = \mathbf{A}\mathbf{x}_i.$$

Hence, $\mathbf{A}$ characterizes how the covariates in $\mathbf{x}_i$ affect the expected value of the additive log-ratio transformed proportions for individual i, and ultimately the relative abundances of the taxa for individual i. Likewise, the covariance matrix for $\boldsymbol{\eta}_{i\cdot}$ is calculated as:

$$\begin{aligned} \mathrm{var}(\boldsymbol{\eta}_{i\cdot}|\mathbf{A}, \mathbf{B}, \boldsymbol{\Psi}, \boldsymbol{\Gamma}) &= \boldsymbol{\Psi} + \mathbf{B}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{B}^\top \\ &= \boldsymbol{\Sigma}_{\mathbf{x}_i}. \end{aligned} \tag{2}$$

The matrix $\boldsymbol{\Sigma}_{\mathbf{x}_i}$, or perhaps its corresponding correlation matrix, can then be used to define a taxon co-occurrence network for individual i based on the covariates. In this expression, $\boldsymbol{\Psi}$ can be thought of as a baseline covariance matrix and $\mathbf{B}$ describes how the covariates in $\mathbf{x}_i$ affect $\boldsymbol{\Sigma}_{\mathbf{x}_i}$.

Running MiCoRe

After installing the `micore` package, running the method is simple. Let's simulate some data and run the function. Note that you need to supply the model matrix $\mathbf{X}$, and you specifically need to give it an intercept column. Also note that, in this example, we run only 500 burn-in and 500 MCMC samples, but in practice, you should likely run for longer. For example, the default is 4000 burn-in and 4000 MCMC samples.

```r
n <- 100
p <- 5
q <- 2

# Simulating data
x <- rnorm(n)
# Model matrix with intercept column
X <- cbind(1, x)
counts <- matrix(0, n, p+1)
for (i in 1:n) {
  counts[i,] <- rmultinom(1, size=100, prob=rep(1,p+1))
}

# Number of burn-in samples and number of MCMC samples to save
n.burn <- 500
n.samp <- 500

# Running micore
library(micore)
mc.fit <- micore(counts, X, n.burn = n.burn, n.samp = n.samp,
                 n.chain=4, n.cores=4, verbose=TRUE)
```

Note that the `micore` object contains one list element for each MCMC chain run. In this example, we ran 4 chains, so each chain's data can be accessed like so:

```
# Chain 1
tmp <- mc.fit[[1]]
attributes(tmp)
```

```
## $names
##  [1] "eta"          "Psi"          "A"            "B"           
##  [5] "gamma"        "eta.accepted" "sigma.zero"   "Gamma"       
##  [9] "acc.probs"    "counts"       "X"
```

```
# Chain 2
tmp <- mc.fit[[2]]
attributes(tmp)
```

```
## $names
##  [1] "eta"          "Psi"          "A"            "B"           
##  [5] "gamma"        "eta.accepted" "sigma.zero"   "Gamma"       
##  [9] "acc.probs"    "counts"       "X"
```

```
# etc...
```

MCMC samples from any of the chains can be extracted from any of the parameters directly from this object. Each parameter is an array where the first dimension represents the . For example, we can extract the **B** parameter from chain 3:

```
# Extracting the B parameter from chain 3
B.3 <- mc.fit[[3]]$B
dim(B.3)
```

```
## [1] 500   5   2
```

```
# Get 101th sample of B in chain 3
B.3[101,,]
```

```
##              [,1]        [,2]
## [1,]  0.005188425  0.01789568
## [2,] -0.018644328  0.04194454
## [3,]  0.045020203  0.00136607
## [4,] -0.011090151 -0.01213980
## [5,]  0.004421874 -0.01100698
```

The names of the parameters available to extract are:

- `eta`: $\boldsymbol{\eta}$, the additive log-ratio transformed proportions
- `Psi`: $\boldsymbol{\Psi}$, the baseline covariance matrix
- `A`: $\mathbf{A}$, the "fixed effect" parameter
- `B`: $\mathbf{B}$, the "random effect" parameter
- `gamma`: γ_i, $i \in 1, \ldots, n$, the individual-specific parameter (not to be confused with `Gamma` with a capital G)
- `Gamma`: $\boldsymbol{\Gamma}$ the column covariance matrix in the Matrix-Normal prior (not to be confused with `Gamma` with a lowercase g)

The MCMC samples from all chains can be merged together for a particular parameter in order to run summary statistics on all MCMC samples from the parameter:

```r
# Merging all 4 chains into single array
B.merge <- mergeChains(mc.fit, par="B")
# Mean of B over all chains
apply(B.merge, 2:3, mean)
```

```
##               [,1]         [,2]
## [1,]  0.0238358743  0.022455626
## [2,] -0.0130634742  0.031215807
## [3,] -0.0004464394 -0.001412758
## [4,] -0.0041749373  0.011745255
## [5,] -0.0098944464  0.019360617
```

Getting estimated OTU abundances

Though the `micore` object contains all samples from all parameters from all chains, these data structures can be difficult to work with directly. This package contains a functions to get estimates (and credible intervals) of the OTU abundances for a given covariate profile $\mathbf{x}_i$. These estimates can be done on either the additive log-ratio scale or on the proportions scale:

```r
# Want to estimate OTU abundances for individual with x=1.3.
# Create the covariate profile with intercept
x.try <- matrix(c(1, 1.3), nrow=1)
# Get predicted abundances on additive log-ratio scale
p1 <- predict(mc.fit, newdata=x.try)
p1$fit
```

```
##            [,1]        [,2]       [,3]       [,4]       [,5]
## [1,] 0.01893807 -0.05250767 0.01406322 0.03901543 0.02684453
```

```r
# Credible intervals
p1$quant
```

```
## $`2.5%`
##             [,1]       [,2]        [,3]        [,4]        [,5]
## [1,] -0.05963265 -0.1270207 -0.06767106 -0.03504107 -0.05271376
## 
## $`97.5%`
##           [,1]       [,2]      [,3]      [,4]      [,5]
## [1,] 0.1157733 0.02037507 0.1033794 0.1428498 0.1140935
```

```r
# Get predicted abundances on proportions scale
p2 <- predict(mc.fit, newdata=x.try, type="prop")
p2$fit
```

```
##           [,1]      [,2]     [,3]      [,4]      [,5]      [,6]
## [1,] 0.1684856 0.1568517 0.167661 0.1718857 0.1698036 0.1653124
```

```r
# Credible intervals
p2$quant
```

```
## $`2.5%`
##           [,1]      [,2]     [,3]      [,4]      [,5]      [,6]
## [1,] 0.1589318 0.1473431 0.158928 0.1633428 0.1613305 0.1559439
## 
## $`97.5%`
##          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,] 0.178784 0.1642617 0.1788653 0.1819505 0.1787398 0.1741472
```

Getting estimated covariance matrix (or precision, correlation, partial correaltion matrix)

It's also possible to directly extract the estimated covariance matrix based on a covariate profile $\mathbf{x}_i$. The function `getPredCov()` allows the user to provide a covariate profile and will return the corresponding estimated covariance, precision, correlation, or partial correlation matrix along with interval estimates. Note that the returned object saves the estimate value in and array in the `fit` slot, and the first dimension corresponds to the individuals that covariances matrices are being calculated for.

```
# Get estimated covariance matrix using covariate profile defined earlier...
cov1 <- getPredCov(mc.fit, newdata=x.try)
# Extract the covariance matrix for the first individual in x.try
cov1$fit[1,,]
```

```
##             [,1]        [,2]        [,3]        [,4]        [,5]
## [1,] 0.013980108 0.005095244 0.004253667 0.005711540 0.004858979
## [2,] 0.005095244 0.012658911 0.005968791 0.008232957 0.006593006
## [3,] 0.004253667 0.005968791 0.015556365 0.008374043 0.007307498
## [4,] 0.005711540 0.008232957 0.008374043 0.016627968 0.008041835
## [5,] 0.004858979 0.006593006 0.007307498 0.008041835 0.012586562
```

```
# Credible intervals
cov1$quant$`2.5%`[1,,]
```

```
##              [,1]          [,2]          [,3]          [,4]          [,5]
## [1,]  0.005718736 -0.0024561262 -0.0046997147 -0.0057084431 -0.0028223058
## [2,] -0.002456126  0.0053827421 -0.0017447453  0.0006495886  0.0004294459
## [3,] -0.004699715 -0.0017447453  0.0072396795 -0.0006317702 -0.0002814817
## [4,] -0.005708443  0.0006495886 -0.0006317702  0.0060101249  0.0002010204
## [5,] -0.002822306  0.0004294459 -0.0002814817  0.0002010204  0.0048094354
```

```
cov1$quant$`97.5%`[1,,]
```

```
##            [,1]       [,2]       [,3]       [,4]       [,5]
## [1,] 0.03093163 0.01914557 0.01707407 0.02148487 0.01764162
## [2,] 0.01914557 0.03064196 0.01891168 0.02393119 0.02146732
## [3,] 0.01707407 0.01891168 0.03302960 0.02647445 0.02330200
## [4,] 0.02148487 0.02393119 0.02647445 0.04398590 0.02540641
## [5,] 0.01764162 0.02146732 0.02330200 0.02540641 0.03484826
```

```
# Partial correlation instead
pc1 <- getPredCov(mc.fit, newdata=x.try, type="pcor")
# Extract the partial correlation matrix for the first individual in x.try
pc1$fit[1,,]
```

```
##            [,1]       [,2]       [,3]      [,4]      [,5]
## [1,] 1.00000000 0.13619988 0.02108654 0.1432840 0.1505555
## [2,] 0.13619988 1.00000000 0.07156671 0.3225661 0.2057916
## [3,] 0.02108654 0.07156671 1.00000000 0.2659102 0.2604549
## [4,] 0.14328402 0.32256614 0.26591022 1.0000000 0.2338781
## [5,] 0.15055548 0.20579157 0.26045493 0.2338781 1.0000000
```

```
# Credible intervals
pc1$quant$`2.5%`[1,,]
```

```
##            [,1]       [,2]       [,3]       [,4]       [,5]
## [1,]  1.0000000 -0.3422461 -0.4275397 -0.4514547 -0.3286014
## [2,] -0.3422461  1.0000000 -0.4176462 -0.1673022 -0.2498067
## [3,] -0.4275397 -0.4176462  1.0000000 -0.2060636 -0.2055425
## [4,] -0.4514547 -0.1673022 -0.2060636  1.0000000 -0.2021749
## [5,] -0.3286014 -0.2498067 -0.2055425 -0.2021749  1.0000000
```

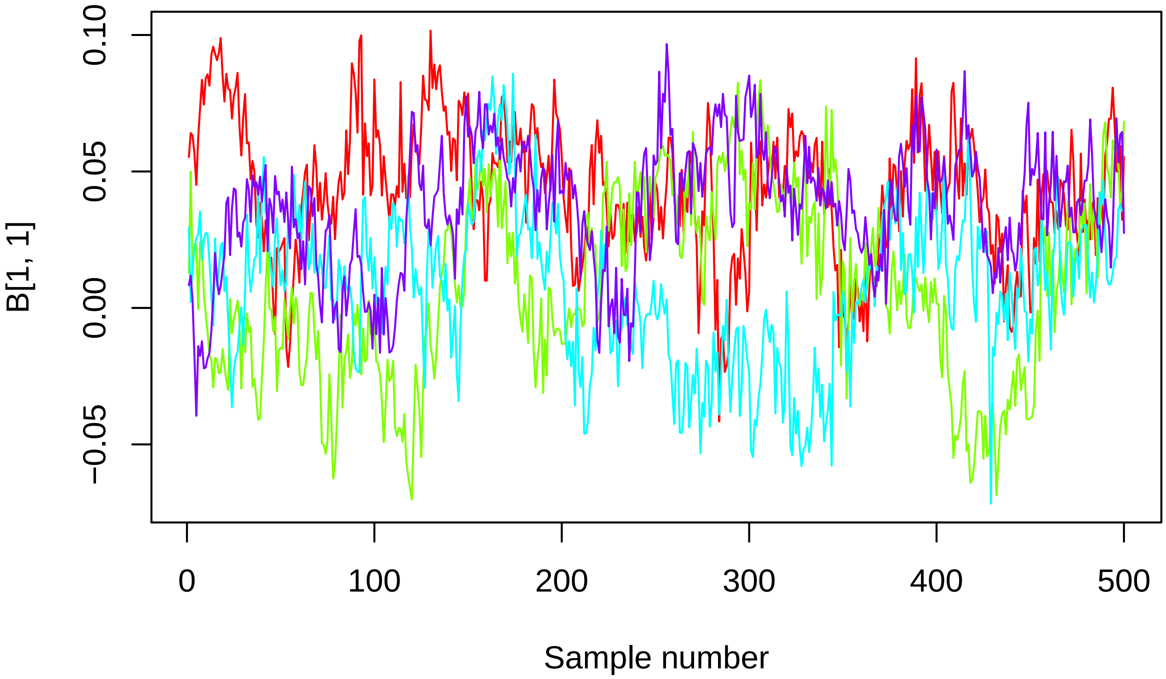
```
pc1$quant$`97.5%`[1,,]
```

```
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.5535032 0.5377571 0.6076847 0.6104813
## [2,] 0.5535032 1.0000000 0.4750256 0.6679584 0.6279180
## [3,] 0.5377571 0.4750256 1.0000000 0.6659402 0.6511696
## [4,] 0.6076847 0.6679584 0.6659402 1.0000000 0.5981701
## [5,] 0.6104813 0.6279180 0.6511696 0.5981701 1.0000000
```

Model diagnostics

When running the model, you should always run multiple chains (the default is 4), to check convergence of the parameters. This can be investigated using the `trplot()` function. Let's check the traceplot for the $\mathbf{B}_{11}$ matrix element:

```
trplot(mc.fit, par="B", ind=c(1,1))
```



In this example, the chains have not yet converged, meaning that it is possible that not enough MCMC samples have been run. In this case it would be wise to increase the parameters `n.burn` and `n.samp` in the `micore()` function.