gcimputeR Vignette on low rank Gaussian copula with quantified uncertainty

Yuxuan Zhao

9/1/2020

Generate data

We generate the high rank continuous data matrix $\mathbf{X} \in \mathbb{R}^{500\times 200}$ from a low rank Gaussian copula as in the experiments of https://arxiv.org/pdf/2006.10829.pdf and randomly mask 40% observation as test set.

```
p = 200
n = 500
rank = 10
sigma = 0.1
set.seed(410)
W = matrix(rnorm(p*rank), nrow = p)
W = t(apply(W, 1, function(x){x/sqrt(sum(x^2))  * sqrt((1-sigma))}))
Z = matrix(rnorm(n*rank), ncol = rank) %*% t(W) + matrix(rnorm(n*p, sd = sqrt(sigma)), ncol = p)
# cubic transformation
X = Z^3
# mask 40% of the original observation
X_obs = X
loc = sample(1:prod(n*p), size = floor(prod(n*p)*0.4))
X_obs[loc] = NA
```

Fitting low rank Gaussian copula

Simply specify the rank to the function call.

```
library(gcimputeR)
est = impute_mixedgc_ppca(X_obs, rank = 10) # around 8 secs
print("Normalized root mean squared error (NRMSE) is: ")
```

```
## [1] "Normalized root mean squared error (NRMSE) is: "
```

```
print(sqrt(mean((est$Ximp[loc] - X[loc])^2))/sqrt(mean(X[loc]^2)))
```

```
## [1] 0.5057409
```

Construct confidence interval

```
ct = ct_impute(X_obs, est, 0.95)
print("The empirical coverage is: ")
```

```
## [1] "The empirical coverage is: "
```

```r
print(mean(X[loc] >= ct$lower[loc] & X[loc] <= ct$upper[loc]))
```

```
## [1] 0.9252
```

```r
print("The mean confidence interval length is: ")
```

```
## [1] "The mean confidence interval length is: "
```

```r
mean(ct$upper[loc] - ct$lower[loc])
```

```
## [1] 3.6503
```