Latentcor: An R Package for Latent Correlation Estimation

18 May 2021

Summary

The R package *latentcor* provides estimation for latent correlation with mixed data types (continuous, binary, truncated and ternary). Comparing to *MixedCCA*, which estimates latent correlation for canonical correlation analysis, our new package provides a standalone version for latent correlation estimation. Also we add new functionality for latent correlation between ternary/continous, ternary/binary, ternary/truncated and ternary/ternary cases.

Compare to MixedCCA, memory footprint.

Statement of need

Currently there is no standalone package dealing with latent correlation for mixed data type like we did in *latentcor*. The R package *stats* (Team and others 2013) have some functionality to calculate different type of correlations (Pearson, Kendall and Spearman). The R package *polycor* (Fox 2019) computes polycoric and polyserial correlations for ordinal data. The R package *pcaPP* (Croux, Filzmoser, and Fritz 2013) provides a fast calculation for Kendall's τ. The R package *MixedCCA* (Yoon, Carroll, and Gaynanova 2020) have functionality for latent correlation estimation as an intermediate step for canonical correlation analysis on mixed data.

Usage

Type	continuous	binary	truncated
continuous	Liu, Lafferty, and Wasserman (2009)	Fan et al. (2017)	Yoon, Carroll, and Gaynanova
binary	Fan et al. (2017)	Fan et al. (2017)	Yoon, Carroll, and Gaynanova
truncated	Yoon, Carroll, and Gaynanova (2020)	Yoon, Carroll, and Gaynanova (2020)	Yoon, Carroll, and Gaynanova
ternary	Quan, Booth, and Wells (2018)	Quan, Booth, and Wells (2018)	This paper

Definition 1 Fan et al. (2017) considered the problem of estimating Σ for the latent Gaussian copula model based on Kendall's τ. Given the observed data $(X_{1j}, X_{1k}), ..., (X_{nj}, X_{nk})$ for variables X_j and X_k, Kendall's τ is defined as

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} sign(X_{ij} - X_{i'j}) sign(X_{ik} - X_{i'k})$$

Theorem 1 Let $W_1 \in \mathcal{R}^{\surd^{\infty}}$, $W_2 \in \mathcal{R}^{\surd^{\in}}$, $W_3 \in \mathcal{R}^{\surd^{\ni}}$, $W_4 \in \mathcal{R}^{\surd^{\triangle}}$ be such that $W = (W_1, W_2, W_3, W_4) \sim NPN(0, \Sigma, f)$ with $p = p_1 + p_2 + p_3 + p_4$. Let $X = (X_1, X_2, X_3, X_4) \in \mathcal{R}^{\surd}$ satisfy $X_j = W_j$ for $j = 1, ..., p_1$, $X_j = I(W_j > c_j)$ for $j = p_1 + 1, ..., p_1 + p_2$, $X_j = I(W_j > c_j)W_j$ for $j = p_1 + p_2 + 1, ..., p$ and $X_j = I(W_j > c_j^1) + I(W_j > c_j^2)$ with $\Delta_j = f(c_j)$, $\Delta_j^1 = f(c_j^1)$ and $\Delta_j^2 = f(c_j^2)$. The rank-based estimator of Σ based on the observed n realizations of X is the matrix $\hat{R}$ with $\hat{r}_{jj} = 1$, $\hat{r}_{jk} = \hat{r}_{kj} = F^{-1}(\hat{\tau}_{jk})$ with block structure The original method is taking estimated Kendall's $\hat{\tau}$ and other parameters to calculate latent correlation $\hat{r}$. Whereas the approximated method is using multilinear interpolation to approximate latent correlation $\hat{r}$ via pre-calculated grid values (Yoon, Müller, and Gaynanova 2021).

refer to table for reference of formula.

Table to show memory improvement compare to mixedCCA.

```
library(latentcor)
### Data setting
n <- 1000; p1 <- 1; p2 <- 1 # sample size and dimensions for two datasets.

### Correlation structure within each data set
set.seed(0)
perm1 <- sample(1:(p1 + p2), size = p1);
Sigma <- autocor(p1 + p2, 0.7)[perm1, perm1]
mu <- rbinom(p1+p2, 1, 0.5)

# Data generation
simdata <- GenData(n=n, type1 = "binary", type2 = "continuous", p1 = p1, p2 = p2, copula1 = "exp",
copula2 = "cube",  muZ = mu, Sigma = Sigma, c1 = rep(1, p1), c2 =  NULL)
```

```
## Warning in GenerateData(n = n, trueidx1 = trueidx1, trueidx2 = trueidx2, : Same
## threshold is applied to the all variables in the first set.
```

```
X1 <- simdata$X1; X2 <- simdata$X2
# Estimate latent correlation matrix with original method
R_nc_org <- estR(X1 = X1, type1 = "ternary", X2 = X2, type2 = "continuous",
                              method = "original")$R
# Estimate latent correlation matrix with aprroximation method
R_nc_approx <- estR(X1 = X1, type1 = "ternary", X2 = X2, type2 = "continuous",
                              method = "approx")$R
```

Rendered R Figures

```
PlotPair(datapair = cbind(c(Sigma), c(R_nc_org)), namepair = c("Sigma", "R_nc_org"),
                  title = "Latent correlation (True vs. Estimated)")
PlotPair(datapair = cbind(c(Sigma), c(R_nc_approx)), namepair = c("Sigma", "R_nc_approx"),
                  title = "Latent correlation (True vs. Estimated)")
```

References

Croux, Christophe, Peter Filzmoser, and Heinrich Fritz. 2013. "Robust Sparse Principal Component Analysis." *Technometrics* 55 (2): 202–14.

Fan, Jianqing, Han Liu, Yang Ning, and Hui Zou. 2017. "High Dimensional Semiparametric Latent Graphical Model for Mixed Data." *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 79 (2): 405–21.

Fox, John. 2019. *Polycor: Polychoric and Polyserial Correlations*. https://CRAN.R-project.org/package=polycor.

Liu, Han, John Lafferty, and Larry Wasserman. 2009. "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs." *Journal of Machine Learning Research* 10 (10).

Quan, Xiaoyun, James G Booth, and Martin T Wells. 2018. "Rank-Based Approach for Estimating Correlations in Mixed Ordinal Data." *arXiv Preprint arXiv:1809.06255*.

Team, R Core, and others. 2013. "R: A Language and Environment for Statistical Computing."

Yoon, Grace, Raymond J Carroll, and Irina Gaynanova. 2020. "Sparse Semiparametric Canonical Correlation Analysis for Data of Mixed Types." *Biometrika* 107 (3): 609–25.

Yoon, Grace, Christian L Müller, and Irina Gaynanova. 2021. "Fast Computation of Latent Correlations." *Journal of Computational and Graphical Statistics*, 1–8.