Journal of Data Science xx (xx), 1–3
July 2022
DOI: 0000

Supplementary Material for geeVerse Ultra-high Dimensional Heterogeneous Data Analysis with Generalized Estimating Equations

Tianhai Zu[1], Brittany Green[2], and Yan Yu[3,*]

[1] University of Texas at San Antonio, United States of America
[2] University of Louisville, United States of America
[3] University of Cincinnati, United States of America

1 Computational Algorithms

This supplement contains: 1) an in-depth explanation of the computational algorithms used for the models in the *geeVerse* package; and 2) detailed instructions on generating the genetic data simulations presented in the main paper.

The *geeVerse* package aims to provide simultaneous variable selection and estimation across diverse data scenarios, including quantile and mean regression for heterogeneous data, analysis of both longitudinal and cross-sectional data, and applications to moderate, high, or even ultra-high dimensional datasets. Given this, several challenges must be addressed to estimate coefficients and select variables. The data is sparse and potentially ultra-high dimensional, making variable selection challenging, since only a few covariates are important among such a large set. If the SCAD penalty is adopted, the non-convex penalty function creates an additional challenge. When the conditional quantile is of interest, there is added complexity due to the discontinuous quantile check function. To deal with these particular challenges, we utilize two computational approaches as described in Zu et al. (2023). We provide an iterative computational algorithm for quantile regression and a separate iterative computational algorithm for mean regression.

When conditional quantiles are of interest, we implement a version of induced smoothing to handle the discontinuous quantile check function (Zhao et al., 2017; Brown and Wang, 2005), while also implementing a minorization–maximization algorithm (Hunter and Li, 2005; Wang et al., 2012) to provide a local linear approximation of the nonconvex SCAD penalty function. Combining these approaches of smoothing and local linear approximation, the quantile penalized estimating equations can be approximated as

$$\sum_i \mathbf{X}_i^{\mathrm{T}}\mathbf{W}_i\mathbf{R}_i^{-1}\left(\Phi\left(\frac{\mathbf{Y}_i-\mathbf{X}_i\widehat{\boldsymbol{\beta}}}{h}\right)-(1-\tau)\right)-n\mathbf{q}_\lambda(|\widehat{\boldsymbol{\beta}}|)\mathbf{sgn}(\widehat{\boldsymbol{\beta}})\frac{|\widehat{\boldsymbol{\beta}}|}{c+|\widehat{\boldsymbol{\beta}}|}=\mathbf{0}. \tag{1}$$

This version of induced smoothing uses the standard normal cumulative distribution function, Φ. Here we set $h=n^{-\frac{1}{2}}$ as recommended in Brown and Wang (2005) and $c=10e^{-6}$. We omit subscript τ throughout for notational simplicity.

Next, we utilize an iterative algorithm similar to the Newton-Rhapson algorithm. Specifically, the coefficient estimates are updated every iteration using

$$\widehat{\boldsymbol{\beta}}^{(k)}=\widehat{\boldsymbol{\beta}}^{(k-1)}+\left[\mathbf{H}\left(\widehat{\boldsymbol{\beta}}^{(k-1)}\right)+n\mathbf{E}\left(\widehat{\boldsymbol{\beta}}^{(k-1)}\right)\right]^{-1}\times\left[\mathbf{S}\left(\widehat{\boldsymbol{\beta}}^{(k-1)}\right)-n\mathbf{E}\left(\widehat{\boldsymbol{\beta}}^{(k-1)}\right)\widehat{\boldsymbol{\beta}}^{(k-1)}\right],$$

*Corresponding author Email: tianhai.zu@utsa.edu.

with

$$\mathbf{H}\left(\widehat{\boldsymbol{\beta}}^{(k-1)}\right) = \frac{1}{h}\sum_{i=1}^{n}\mathbf{X}_i^{\mathrm{T}}\mathbf{W}_i\mathbf{R}_i^{-1}\,\mathrm{diag}\left\{\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{-\frac{1}{2}}\phi\left(\frac{\mathbf{Y}_i-\mathbf{X}_i\widehat{\boldsymbol{\beta}}}{h\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\right)^{\frac{1}{2}}}\right)\right\}\mathbf{X}_i,$$

$$\mathbf{E}\left(\widehat{\boldsymbol{\beta}}^{(k-1)}\right) = \mathrm{diag}\left\{\frac{q_{\lambda_n}\left(\left|\widehat{\beta}_1\right|_+\right)}{c+\left|\widehat{\beta}_j\right|},\ldots,\frac{q_{\lambda_n}\left(\left|\widehat{\beta}_{p_n}\right|_+\right)}{c+\left|\widehat{\beta}_j\right|}\right\},$$

and $\phi(\cdot)$ is the standard normal probability density function.

To select the penalty parameter for the conditional quantile model, we recommend utilizing the high-dimensional Bayesian information criterion (HBIC), especially for high-dimensional data, due to its computational efficiency citep{Wangetal2009,chenchen08,Zuetal2023}) To implement this, the optimal λ for the quantile model is identified from a grid of candidate values that minimizes

$$HBIC\left(\lambda\right) = \log\left(\sum_{i=1}^{n}\rho_\tau\left(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\right)\right) + p_{1,\lambda}\frac{\log(n)}{2n}C_n, \tag{2}$$

as in Zu et al. (2023). Here $p_{1,\lambda}$ refers to the count of parameters that are not zero. C_n is $\log\left(\log(p_n)\right)$, which is suggested from the previous literature.

For the conditional mean model, estimating coefficients and selecting variables also requires an iterative algorithm. Coefficients of the mean penalized generalized estimating equations are estimated using a combination of the minorization–maximization algorithm (Hunter and Li, 2005) which provides a local linear approximation of the nonconvex SCAD penalty function and the Newton-Raphson algorithm for the GEE. See Wang et al. (2012) for the full computational algorithm.

To determine the penalty parameter for the mean approach, cross-validation is used as in Inan and Wang (2017). Cross-validation is also an option for the quantile approach. When dealing with longitudinal data, cross-validation is performed by sampling subjects to preserve the working correlation structure, while with non-longitudinal data, cross-validation is conducted on the observational level. For the mean regression approach, we utilize mean squared errors to evaluate model performance, and the implementation was inherited from `PGEE`, whereas, for quantile regression models, we employ the mean check loss function to evaluate model performance. Similar to the `pgee()` function in the package `PGEE`, we did not implement HBIC for the mean regression function in our package. However, HBIC can be easily attained for the mean model by substituting the check loss function portion in equation (2) with the mean squared loss (Wang et al., 2009; Chen and Chen, 2008).

Note that although cross-validation offers a robust evaluation for λ choices, this robustness often comes at the cost of increased computational resources. The need to repeatedly fit the model to different subsets of data significantly escalates the computational burden, particularly for complex models or large datasets. While we provide both HBIC and cross-validation as tools for λ selection, the choice between them should be informed by the specific dataset and scenarios at hand, empowering users to balance robustness with practical computation resources.

2 Genetic Data Simulation Instructions

We employ a commonly used tool for genetic data, HapGen2 (Su et al., 2011), to resample existing genotype data in Section 2.2 in the main paper. This resampling-based approach amalgamates existing genotype data to construct the genotypes of simulated samples and provides a robust methodology for generating simulated genotype data. This ensures the preservation of allele frequency and linkage disequilibrium (LD) patterns, thereby providing a realistic simulation environment. The genotype data we resample from is the publicly available 1000 Genomes Project data (1000 Genomes Project Consortium, 2010) from the Impute2 webpage. To simulate the response variable and the other control variables, users can use `generate_data` and provide the SNPs data in the optional argument `SNPs`.

3 Replication Script for the Main Paper

The replication scripts for the main paper are hosted at a public repository on Github.

References

1000 Genomes Project Consortium (2010). A map of human genome variation from population scale sequencing. *Nature*, 467(7319): 1061.

Brown BM, Wang YG (2005). Standard errors and covariance matrices for smoothed rank estimators. *Biometrika*, 92(1): 149–158.

Chen J, Chen Z (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3): 759–771.

Hunter DR, Li R (2005). Variable selection using MM algorithms. *The Annals of Statistics*, 33(4): 1617–1642.

Inan G, Wang L (2017). `PGEE`: an r package for analysis of longitudinal data with high-dimensional covariates. *R Journal*, 9(1): 393.

Su Z, Marchini J, Donnelly P (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics (Oxford, England)*, 27(16): 2304–2305.

Wang H, Li B, Leng C (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3): 671–683.

Wang L, Zhou J, Qu A (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2): 353–360.

Zhao W, Lian H, Liang H (2017). GEE analysis for longitudinal single-index quantile regression. *Journal of Statistical Planning and Inference*, 187: 78–102.

Zu T, Lian H, Green B, Yu Y (2023). Ultra-high dimensional quantile regression for longitudinal data: an application to blood pressure analysis. *Journal of the American Statistical Association*, 118(541): 97–108.