Estimation and inference for high dimensional, doubly-structured regression models

Parker Knight, Yue Wang, Timothy W. Randolph
Fred Hutchinson Cancer Research Center, Seattle, WA
University of Florida, Gainesville, FL

Background

Multiple regression models of the form

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

lack a unique solution when the number of variables exceeds the number of samples. This issue has been addressed by popular regularized regression models, such as the LASSO and ridge regression. Kernel Penalized Regression (KPR) [1] is an extension of ridge regression that incorporates external variable and subject structure, allowing researchers to include additional scientific information in high dimensional models.

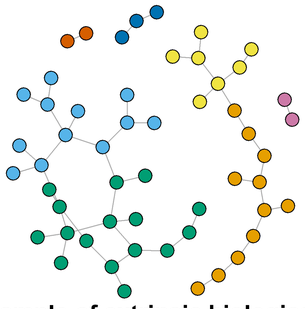

Figure 1: **Example of extrinsic biological structure among variables**, e.g., a metabolic pathway.

KPR models have the form

$$\mathbf{Y} = \mathbf{Z}\beta + \mathbf{E}\eta + \varepsilon$$

where **Z** is a matrix of penalized variables and **E** is a matrix of unpenalized covariates.

We have developed an R software package for fitting KPR models by solving the following minimization problem:

$$\hat{\beta}, \hat{\eta} = \underset{\beta,\eta}{\arg\min} \left\{ \|\mathbf{Y} - \mathbf{Z}\beta - \mathbf{E}\eta\|_H^2 + \lambda\|\beta\|_{Q^{-1}}^2 \right\}$$

Q represents a matrix of variable structure and **H** is a matrix of sample structure. The penalized coefficients are tested for significance with the GMD Inference test [2].

Microbiome Data Analysis

We evaluated the KPR models by analyzing microbiome bacterial abundance data from [3], with subject age as the outcome.

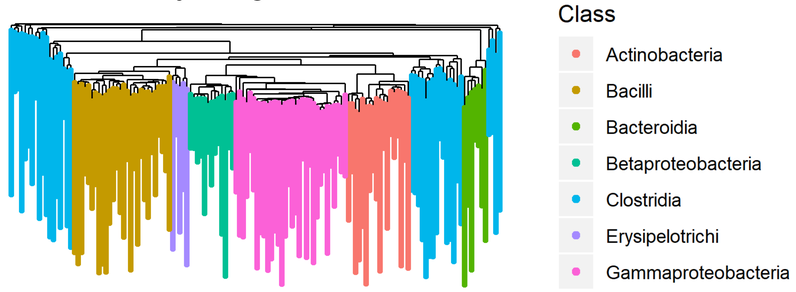

Figure 2: **A phylogenetic tree** (extrinsic structure) constructed from 149 genera included in the analysis, colored by taxonomic class.

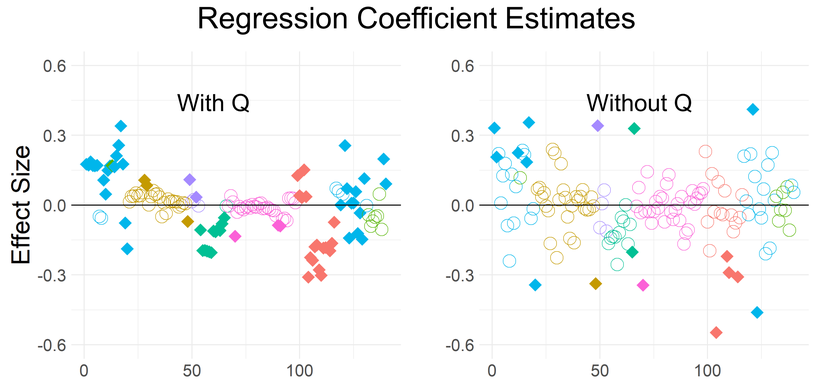

Figure 3: **KPR coefficient estimates by index**, colored by taxonomic class. Dark points indicate significance ($p < 0.0001$) with respect to age. Coefficients from a KPR model with a **Q** matrix of phylogenetic info (left). Coefficients from a model fit without a **Q** matrix (right).

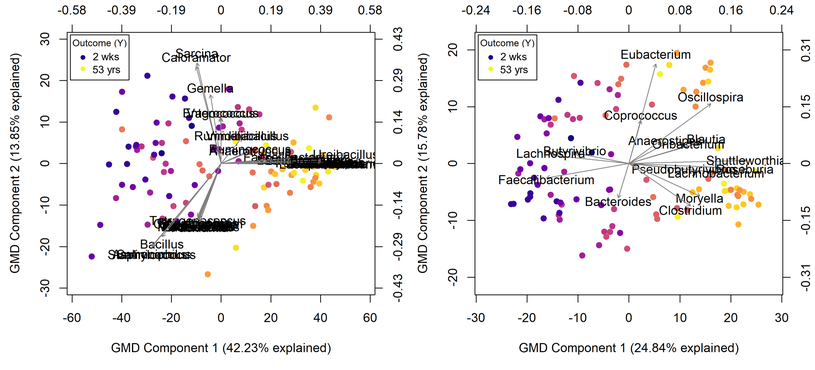

Figure 4: **Generalized matrix decomposition biplots** [4] of KPR models, colored by outcome. Biplot of significant taxa from model fit with **Q** (left). Biplot without **Q** (right).

Package Usage

The `KPR` package includes functions for processing structured data before it's included in the regression model. For example, the `generateSimilarityKernel` function converts distance matrices to similarity kernels.

```
library(KPR)
Q <- generateSimilarityKernel(patristic_dist)
H <- generateSimilarityKernel(unifrac_dist)
```

Users can fit models by passing their data matrices to the `KPR` function.

```
kpr.fit <- KPR(designMatrix = Z, covariates = E,
               Y = Y, Q = Q, H = H)
```

Conclusions

The `KPR` package provides an easy interface for fitting and interpreting Kernel Penalized Regression models. By using the package to analyze microbiome bacterial abundance data, we have shown that incorporating structural information into penalized regression models can improve interpretability and statistical power.

We plan on adding generalized KPR models (e.g., logistic regression), adaptive kernel penalties, and additional inference methods.

Acknowledgements

This work is funded by Machine Learning Tools for Discovery and Analysis of Active Metabolic Pathways, R01GM114029; and Statistical models for multi-modal brain imaging studies, R01 MH108467. The Summer Undergraduate Research Program is supported in parts by the Cancer Center Support Grant (CCSG) CURE Supplement: NCI 3P30CA015704, the Fred Hutch Internship Program, and individual labs/research groups.

References

[1] Randolph, T. et al (2018). Kernel-penalized regression for analysis of microbiome data. Ann. Appl. Stat. 12, no. 1, 540-566.
[2] Wang, Y. et al. Generalized Matrix Decomposition Regression. Technical report.
[3] Yatsunenko, T. et al (2012). Human gut microbiome viewed across age and geography. Nature, 486, 222-227.
[4] Wang, Y. et al. The GMD-biplot and its application to microbiome data. Technical report.