## Data

**Abstract**

**Real data:** At the University of Texas M.D. Anderson Cancer Center, a study was conducted using matrix assisted laser desorption and ionization, time-of-flight mass spectrometry (MALDI-TOF) to discover potential proteomic markers of pancreatic cancer. In this study, researchers collected the blood serum samples from 139 pancreatic cancer patients and 117 normal controls and ran them on a MALDI-TOF mass spectrometer to produce a mass spectrum for each sample. These data were preprocessed using the Cromwell pipeline (Coombes, et al. 2005 Proteomics), including baseline correction, normalization, denoising and transformation, as detailed in Koomen et al. (2005, Clinical Cancer Research). In our analysis here, we focused on the spectral region [5000D, 8000D] of the original dataset, which includes 1659 observations per spectrum. The 256 samples in this dataset were measured in four different blocks over a span of several months, so we estimated and subtracted the block-specific mean from the preprocessed mass spectra to adjust for the block effects before performing functional quantile regression. The raw dataset, preprocessed dataset, and the block effect adjusted dataset are all provided.

**Simulated data:** For two simulation settings described in the paper, i.e., symmetric heavy tailed setting (i) and right skewed setting (ii), 100 replicate datasets were generated. Each replicate dataset includes simulated functions consisting of 501 observations for each of 400 subjects coming from 2 groups. For space considerations, only 1 replicate simulated dataset is provided for each setting. However, all 100 replicates are available upon request, and can be generated using the provided MATLAB script. They will also be deposited on the senior author’s github repository upon acceptance of paper.

**Availability**

There are no restrictions on the availability of data used in this paper.

**Description**

* These data have been used in other publications (Morris, et al. 2008 Biometrics; Morris 2012 Statistics and Its Interface), and have been publicly available on the website (<https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software_Id=70>) for the WFMM software download website for numerous years.
* The data we use are provided in part (B) of the supplementary. They will also be deposited on the senior author’s github repository upon acceptance of paper.

File format

* **Real data:** The raw and preprocessed mass spectrometry data are stored as matrices in “.mat” file; the block effect adjusted dataset ready for model fitting is stored as a structure array in “.mat” file along with design matrix, spectral locations (in Daltons) and wavelet transform specifications.
* **Simulated data:** The simulation datasets are stored both as structure arrays in “.mat” files and matrices in “.txt” files.

**Optional Information**

Unique identifier / DOI

## Code

**Abstract**

We provided the code to reproduce all the figures in the manuscript and part (A) of the supplementary. In addition, we provided code to adjust for block effects from the preprocessed mass spectrometry data and to generate the simulation datasets, as well as a set of MATLAB scripts to implement our proposed Bayesian Functional Quantile Regression model and the naïve Bayesian Quantile Regression. We also provided R scripts to implement the bootstrap-based approaches which were compared to our proposed model in the paper.

**Description**

How delivered (R package, Shiny app, etc.)

* MATLAB and R scripts

Licensing information

* The code is open source on a MIT license

Link to code/repository (e.g., github.com, bitbucket.org)

* Github page: <https://github.com/MorrisStatLab/FunctionalQuantileRegression>

Version information (e.g., for a Git repository, the number or branch+commit)

* Github: branch 1 and commit 5.

**Optional Information**

Hardware requirements (e.g., operating system with version number, access to cluster, GPUs, etc.)

* Can be run on any windows, mac, or Unix system; we used a 64-bit operating system with 2 processors and an RAM of 256GB

Supporting software requirements (e.g., libraries and dependencies, including version numbers)

* MATLAB version 2016b or later (wavelet toolbox needed), and R version 3.2.2 or later (the packages “quantreg”, “coda” and “FDboost” are needed).

## Instructions for Use

**Reproducibility**

What is to be reproduced (e.g., "All tables and figure from paper", "Tables 1-4”, etc.)

* All the figures in the manuscript can be reproduced by running the provided code. To reproduce Table 2 that summarizes the simulation performance for various methods, the posterior (or bootstrap) samples obtained by each approach at each considered quantile level based on each of 100 replicate datasets in each simulation setting are needed. To reproduce Figure S1-4, the posterior (or bootstrap) samples of the regression coefficients obtained by each approach at each quantile level based on the proteomics dataset are needed. For space considerations, we do not provide these posterior (or bootstrap) samples for now. However, they can be generated using the scripts we provided, and are available upon request. They will also be deposited on the senior author’s github repository upon acceptance of paper. We also provide the script to implement the Bayesian FQR model and various alternatives on a given functional dataset and generate posterior (or bootstrap) samples, and the code to do estimation and inference on functional coefficients and generate Figure S1-4 using these posterior (or bootstrap) samples.

How to reproduce analyses (e.g., workflow information, makefile, wrapper scripts)

* The MATLAB scripts “figures\_and\_tables.m” and “supplementary\_figures.m” and the R script “supplementary\_figures\_1\_2\_3.R” in the main folder contain the code to reproduce all of the figures in the manuscript and part (A) of the supplementary.
* In the subfolder “realdata/”, the file “ProteomicsData.mat” stores the raw and preprocessed pancreatic cancer mass spectrometry data, the cancer/normal status of each of 256 subjects and the spectral locations of observations (in Daltons); the MATLAB script “realdata\_preparation.m” records how to adjust for the block effect of the preprocessed data and generate the file “model.mat” which stores the adjusted mass spectra, design matrix, spectral locations and wavelet transform specifications as a structure array; the MATLAB script “realdata\_modelfit.m” and R script “realdata\_modelfit\_bootstrap.R” record how to implement our proposed Bayesian FQR and various alternative approaches on the adjusted mass spectrometry dataset and obtain posterior (or bootstrap) samples for the quantile regression coefficient functions. We set the seed when implementing each approach, so users should obtain exactly the same posterior (and bootstrap) samples as ours by running the provided scripts.
* In the subfolder “simulations1/”, the MATLAB script “data\_generation.m” records how to generate the 100 replicate simulation datasets for the symmetric heavy tailed setting (i), and obtain the ground truth for the quantile regression coefficient functions characterizing the group effect for multiple quantile levels; the MATLAB script “data\_modelfit.m” and R script “data\_modelfit\_bootstrap.R” record how to implement our proposed Bayesian FQR and alternative approaches on the simulated datasets and obtain posterior (or bootstrap) samples; the MATLAB script “data\_estimation\_inference\_example.m” illustrates how to summarize the estimation and inference performance of each approach using posterior (or bootstrap) samples based on a simulated dataset given the ground truth.
* The subfolder “simulations2/” contains the same contents for the right skewed setting (ii), and the R script “FDboost.R” which records how to implement the functional linear array model by Brockhaus et al. (2015, Statistical Modelling) on a simulated dataset in this setting to perform FQR for .
* The subfolder “BayesianFQR/” contains a set of MATLAB scripts to implement the Bayesian FQR model with discrete wavelet transform (DWT) on regression coefficient functions and a horseshoe prior on wavelet coefficients. These scripts are called when performing FQR on the real or simulation datasets.
* The subfolder “BayesianQR/” contains a set of MATLAB scripts to implement the naïve Bayesian Quantile Regression. These scripts are called when running this naïve approach on the real or simulation datasets.
* Other scripts or files not described above are either auxiliary functions used for plot or ROC analysis, or various output from running the scripts above.

How to use software in other settings (or links to such information, e.g., R package vignettes, demos or other examples)

* The provided scripts can be straightforwardly applied to other data sets given the matrix of response functions **Y**, design matrix **X** for quantile regression, and the basis transform specifications provided by users. Some scripts in the subfolder “BayesianFQR/” (i.e., FQR\_HS.m, UpdateLambda.m, UpdateTau.m, and UpdateSa2.m) need to be modified if other shrinkage priors are placed on the basis coefficients.

References

[1] Coombes K R, Tsavachidis S, Morris J S, et al (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface - enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics, 5*(16): 4107-4117.

[2] Koomen J M, Shih L N, Coombes K R, et al (2005). Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. *Clinical Cancer Research, 11*(3): 1110-1118.

[3] Brockhaus S, Scheipl F, Hothorn T, et al (2015). The functional linear array model. *Statistical Modelling*, *15*(3): 279-300.