Quantile Regression-Based Drift Correction

Applied to Metabolomics Data from Human Serum Samples

Sophie Narath[1,2], *Michael G. Schimek*[2], *Gunnar Libiseller*[1], *Edgar Gander*[1], *Harald Sourij*[3], *Frank M. Sinner*[1,3], *Thomas R. Pieber*[1,3], *Christoph Magnes*[1]

1
JOANNEUM RESEARCH
Forschungsgesellschaft mbH

HEALTH
Institute for Biomedicine
and Health Sciences

Sophie Narath

Elisabethstrasse 5
8010 Graz, Austria

Phone +43 316 876-40 00
Fax +43 316 8769-40 00

health@joanneum.at
www.joanneum.at/health

2
Medical University of Graz

Institute for Medical Informatics,
Statistics and Documentation

Graz, Austria

3
Medical University of Graz

Clinic of Internal Medicine
Division of Endocrinology and
Metabolism

Graz, Austria

Objective

We aim to compose a workflow to identify and compensate analytical data for bias from various sources (sample collection and preparation, HILIC-FTMS analysis). The workflow comprises data filtering and drift correction. The statistical methods to be applied for signal drift correction depend primarily on data structure, study size, the number of QC samples, and technical specificities of the analytical method. QC samples are generally used to correct metabolomics data[1]. We present here our quantile regression approach.

Methods

R[2] was used for peak detection, peak grouping and the complete workflow for drift correction.

Filtering

Filtering steps are based on excluding technical artefacts, redundant information, batch differentiation, and highly spread features.

Quantile Regression

The main advantage of quantile regression over other regression techniques is its flexibility for modelling data with heterogeneous conditional distributions. Since the variability of features was high, smoothing by a locally adaptive regression technique was required to retrieve the maximum valuable information.

In this case, the 90 % quantile of the QCs(y's) was estimated via a nonparametric quantile regression using regression splines depending on the sample number (x's). This procedure fits a piecewise cubic polynomial with 5 knots (df) (breakpoints in the third derivative) arranged at the 90 % quantile of the x's: rq(y ~ bs(x, df = 5), tau = 0.9).

Through a multiplicative correction factor based on the median of the original QC-values, a further quantile regression model is estimated to correct all samples (Figure 1).

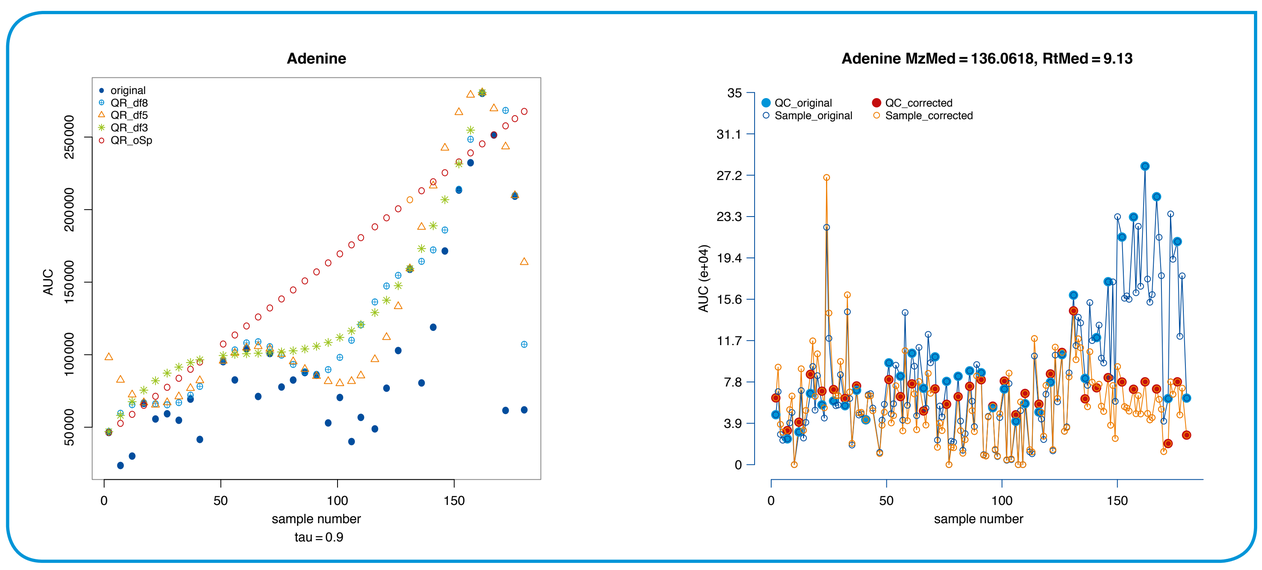

Figure 1: *Drift correction using a quantile regression approach. Left: model fits for the QCs. Right: final correction using df = 5 and tau = 0.9*

Evaluation of the drift correction

Criteria to evaluate the success of the workflow were overall lower variation in QC samples, graphical representations of drift features and multivariate modelling based on batches as class variables to determine the degree of batch overlap. Batch separation before and after data preparation is compared through multivariate modelling approaches (PCA & Random Forest). We have chosen Random Forest because of its adaptability for nonlinear properties of the data (Figure 2). In the present case 500 trees were used as a default parameter.

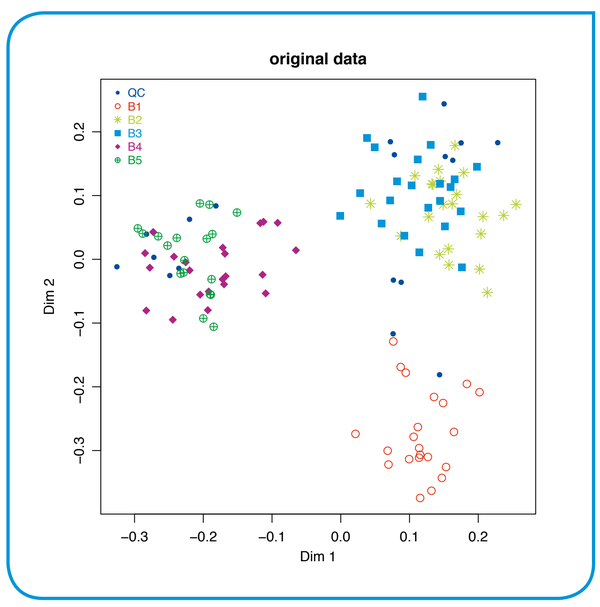

Figure 2: *Unsupervised Random Forest calculation from original data, plotting batches and QCs as class variable*

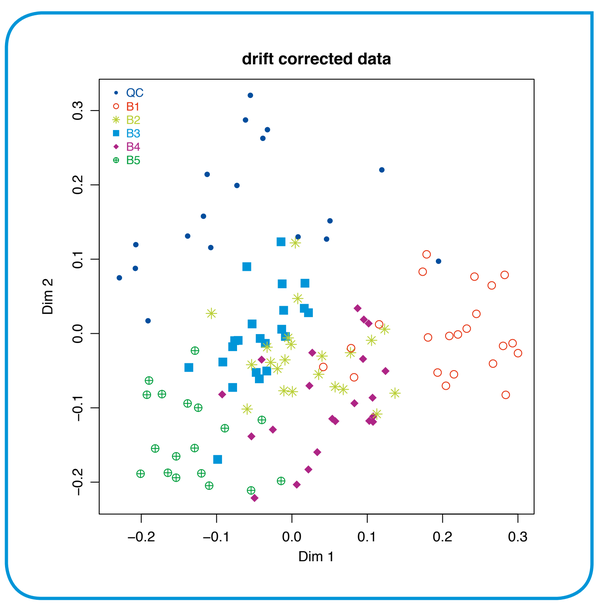

Figure 3: *Unsupervised Random Forest calculation from drift corrected data, plotting batches and QCs as class variable*

Results

In the present project, metabolite fingerprints of deproteinized serum samples from 106 type 1 diabetes patients were generated using HILIC-FTMS (LTQ Orbitrap XL). Study samples were divided into 5 different analysis batches.

The workflow application resulted in a feature reduction of more than 50 % (from 12,000 initially detected features), lower variation over the QC pool samples (from a RSD of 0.35 to 0.26), and less visible batch clustering (Figure 3).

Discussion

To consider various feasible drift patterns, a systematic analysis of the behaviour of the quantile regression approach for such data is currently under investigation. The workflow will be optimized with data from upcoming clinical studies.

References

[1]
Dunn, W.B., Broadhurst, D., et al. (2011). Nature protocols, 6 (7).

Kamleh, M.A., Ebbels, T.M.D., et al. (2012). Analytical chemistry, 84 (6).

Kirwan, J.A, Broadhurst, D.I., et al. (2013). Analytical and bioanalytical chemistry.

[2]
R Packages: "xcms", C.A. Smith et al. (2006), "quantreg", R. Koenker (2013), "randomForrest", A. Liaw and M. Wiener (2002)

Acknowledgement

This work was supported financially by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit), project Met2Net.