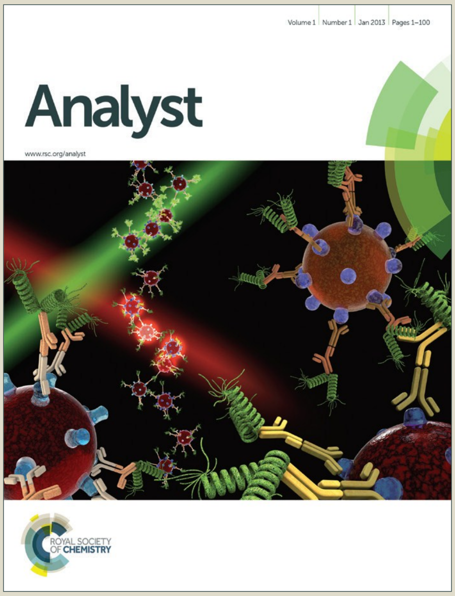
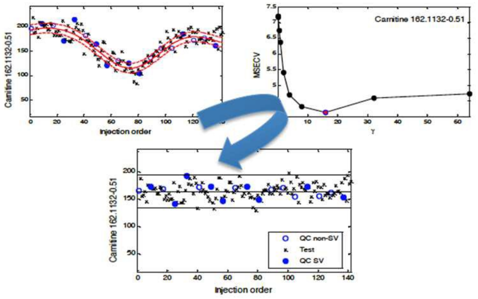

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Analyst
www.rsc.org/analyst
ROYAL SOCIETY OF CHEMISTRY



Intra-batch effects in liquid chromatography-mass spectrometry are corrected using quality control samples and support vector regression.

Intra-batch effect correction in liquid chromatography-mass spectrometry using quality control samples and Support Vector Regression (QC-SVRC)

Julia Kuligowski[a], Ángel Sánchez-Illana[a], Daniel Sanjuán-Herráez[b], Máximo Vento[a,c,d], Guillermo Quintás[b,e]*

[a]Neonatal Research Unit, Health Research Institute La Fe, Valencia, Spain

[b]Safety and sustainability Division, Leitat Technological Center, Valencia, Spain

[c]Division of Neonatology, University & Polytechnic Hospital La Fe, Valencia, Spain

[d]National Coordinator Spanish Maternal and Child Health and Development Network SAMID; Instituto Carlos III (Spanish Ministry of Economy and Competitiveness)

[e]Analytical Unit, Health Research Institute La Fe, Valencia, Spain

*Corresponding author: guira@uv.es

ABSTRACT

Instrumental developments in sensitivity and selectivity boost the application of liquid chromatography – mass spectrometry (LC-MS) in metabolomics. Gradual changes in the LC-MS instrumental response (i.e. intra-batch effect) are often unavoidable and reduce the repeatability and reproducibility of the analysis; decrease the power to detect biological responses and hinder the interpretation of the information provided. Because of that, there is an interest in the development of chemometric techniques for the post-acquisition correction of batch effects. In this work, the use of quality control (QC) samples and Support Vector Regression and a radial basis function kernel (QC-SVRC) is proposed to correct intra-batch effects. The repeated analysis of a single sample is used for the assessment of both, the correction accuracy and the effect of the distribution of QC samples throughout the batch. The QC-SVRC method is evaluated and compared to a recently developed method based on QC samples and robust cubic smoothing splines (QC-RSC). Results show that QC-SVRC slightly outperformed QC-RSC and allows a straightforward fitting of the SVRC parameters to the instrument performance by using the ε-insensitive loss parameter.

Keywords

Batch effect, Support Vector Regression (SVR), liquid chromatography – mass spectrometry, Metabolomics.

1. INTRODUCTION

Metabolomics has emerged as a promising technology that is gaining broader recognition and it is increasingly being used in biomedical research. Metabolomics aims at the comprehensive quantitative analysis of all metabolites in a cellular system in a given state at a given point in time[1]. Metabolites reflect the interaction of the genome, proteome and transcriptome with the environment and so, the identification of perturbed metabolic pathways is a suitable approach to e.g. identify phenotypes associated to pathological conditions, getting insight into biological processes or for the monitoring of responses to therapy.

High resolution liquid chromatography - mass spectrometry (LC-MS) is becoming the method of choice in metabolomics because of its increased sensitivity, high throughput and metabolite coverage as compared to other techniques such as gas chromatography - mass spectrometry (GC-MS) or nuclear magnetic resonance (NMR). Metabolomic profiles typically show a high variability among subjects. Besides biological variation, LC-MS data includes instrumental variation due to e.g. minor changes in the injection volume, ideally as a normal white noise process with mean zero and constant variance, and also from e.g. gradual inlet interface contamination, drifts in ionization efficiency or column performance[2,3] that introduce a gradual change in the instrumental response during the measurement of a batch of samples (i.e. intra-batch effect). The intra-batch effect decreases the power to detect biological responses and hinders data interpretation[4] and joint analysis of data from several batches. There are two approaches to limit this potential pitfall. The most straightforward solution is to avoid it happening in the first place by using improved sample clean up, robust LC columns, or more stable ionization sources and detectors. However, intra-batch effects are often unavoidable and so, complementary strategies enabling the signal correction after the measurement has been taken are desirable.

Spiked internal standards (ISs) with similar physico-chemical properties as the analytes of interest can be used for data normalization. However, while the number of spiked standards is limited, the number of detected analytes in untarget LC-MS experiments is *a priori* unknown and so, this approach may bias the results due to ion suppression affecting the IS[5]. On the other hand, if the drift in the instrumental response for each detected variable over the batch is fit to a function, its accurate estimation would allow an effective correction of the intra-batch effect and the shrinkage of the instrumental error. The use of the analytical variation in the response observed in pooled quality control (QC) samples dispersed evenly throughout the batch has been proposed to model the instrumental drift. The algorithm, Quality Control – Robust Spline Correction (QC-RSC)[6], has been shown to be useful for the evaluation and correction of both, intra-and inter-batch instrumental stability, and for an accurate identification of unreliable variables. Briefly, in the QC-RSC method an adaptive cubic spline function f is used to describe the change in the intensities of each variable in QC samples (y_{QC}) as a function of the injection order (x_{QC}), with the value of f at the data point $x_{QC}(i)$ approximating the value $y_{QC}(i)$ for $i = 1 \dots n$, where n=$length(x_{QC})$.

The smoothing spline function f is calculated to minimize the penalyzed sum of squares:

$$p\sum_{i=1}^{n}\left(y_{QC}(i)-f\left(x_{QC}(i)\right)\right)^2+(1-p)\int\left(\frac{d^2f}{dx_{QC}^2}\right)^2 dy$$

The parameter p is the smoothing parameter $(0 < p < 1)$ with small values corresponding to smoother estimates. The first term of the equation denotes the residual sum of squares and penalizes the lack of fit of the function. The second term is weighted by $(1 - p)$ and penalizes the curvature of the function[7]. The selection of the smoothing parameter is based on leave one out - cross validation (LOO-CV) to avoid overfitting as a compromise between the smoothness of the function and the lack of fit. After selecting p, the smoothing spline function calculated using data extracted from the analysis of QCs, is used to normalize each variable at each point within the analysis order vector. However, the fit of the functions has several difficulties such as: data is typically noisy; the intra-batch effect can be seen as a stationary or non-stationary process depending on the position within the batch (e.g. it can be negligible at the beginning and very significant at the end of the batch); and the intra-batch effect varies across metabolites.

Support Vector Regression (SVR) is a non-parametric and distribution free model derived from the Statistical Learning Theory developed by Vapnik[8], that provides high generalization capabilities at a low estimation cost. SVR maps the input data to a higher dimensional kernel variable space by means of a non-linear mapping function and then solves a linear model there[9]. The estimation of the SVR turns to be a quadratic optimization problem with linear restrictions and a global solution. The SVR parameter selection requires the specification of a loss function and a kernel function. A popular loss function is the ε-insensitive loss function that ignores absolute error lower than ε, growing linearly for errors higher than ε. In the present study, SVR using a radial basis function (RBF) kernel was tested to correct the instrumental drift within a batch using data acquired from QC samples. The proposed QC-SVRC approach was evaluated and compared to the reference QC-RSC approach using the repeated analysis of a single plasma sample as a model example to facilitate the evaluation of the correction accuracy and provide a straightforward assessment of the effect of the distribution of QC samples throughout the batch. Based on the distribution of RSD% values in 'non-QC' samples after intra-batch effect correction, the correction could be slightly improved using QC-SVRC approach. Besides, the use of the ε-insensitive loss parameter allowed the fit of the SVRC parameters to the instrument performance.

2. EXPERIMENTAL SECTION

Reagents and materials

All solvents were of LC-MS grade and were purchased from Scharlau (Barcelona, Spain). Ultra-pure water was generated with a Milli-Q water purification system from Merck Millipore (Darmstadt, Germany). The internal standards used for LC-TOF-MS were: Phenylalanine-D_5, Methionine-D_3, Caffeine-D_9 (Cambridge Isotopes Laboratory Inc., Andover, MA, USA), 98% purity and Reserpine. Formic acid (≥95%) was obtained from Sigma-Aldrich Química SA (Madrid, Spain).

Sample preparation

A single 3 mL blood sample was collected from a volunteer in a heparin vacutainer tube. The blood sample was centrifuged at 1800 x g during 10 min at 4°C. Then, the plasma layer was collected, centrifuged at 2000 x g during 5 min at 4°C and the supernatant transferred to an Eppendorf vial. Then, 600 µL of the heparinized plasma were withdrawn and 2 mL of cold methanol was added for protein precipitation. The sample was homogeneized (vortex, 10 s) and centrifuged at 10000 x g (10 min, 4°C). Then, 2 mL of the supernatant was collected and evaporated to dryness under vacuum and at 25°C. The residue was redissolved in 800 µL of a 2 µM internal standard solution containing Phenylalanine-D_5 and Methionine-D_3 in H_2O:CH_3OH (1:1, 0.1% HCOOH).

UPLC-TOF-MS analysis

Chromatographic analysis of plasma samples was performed on an Acquity UPLC chromatograph using an Acquity UPLC HSS T3 (100 x 2.1 mm, 1.8 µm) analytical column (Waters, Wexford, Ireland). Autosampler and column temperatures were set to 4°C and 40°C, respectively and the injection volume was 5 µL. A gradient elution with a total run time of 14 min was performed at a flow rate of 500 µL min^{-1} as follows: initial conditions of 100% of mobile phase A (H_2O (0.1% v/v HCOOH)) were kept for 1 min, followed by three linear gradients from 0% to 15% of mobile phase B (CH_3CN (0.1% v/v HCOOH)) in 2 min; from 15% to 50% B in 3 min; and from 50% to 95% B in 3 min; 95% B was held for 3 min and finally, a 0.5 min gradient was used to return to the initial conditions, which were held for 1.5 min. Full scan MS data from 50 to 1000 mass to charge ratio (m/z) with a scan time of 0.1 s was collected on a quadrupole time of flight (QTOF) SYNAPT HDMS spectrometer (Waters, Manchester, UK) in the TOF MS mode. The following electrospray ionization parameters were selected: capillary and cone voltages were set at 3.2 kV and 20 V; source and desolvation temperatures were set at 120°C and 380°C, respectively; flow rates of cone and nebulization gases were set at 50 and 800 L/h, respectively. The plasma sample was injected 150 times in a single batch. Blank samples were analyzed at the beginning and end of the batch to remove background contamination. The first 8 injections were used for column conditioning and were not included in the study.

Data analysis

Centroid raw LC-TOF-MS data (.raw files) was converted into .netCDF format using DataBridge software (Waters) before generating peak tables using XCMS software (http://metlin.scripps.edu/xcms/). The centWave method was used for peak detection with the following parameters: ppm=20, peakwidth=(3, 20), snthresh=25. A minimum difference in m/z of 7.5 mDa was selected for peaks with overlapping RTs. Intensity weighted m/z values of each feature were calculated using the wMean function. Peak limits used for integration were found through descent on the Mexican hat filtered data. Peak grouping was carried out using the 'nearest' method, using the following parameters, mzVsRT=1 and RT and m/z tolerances of 6 s and 20 mDa, respectively. After peak grouping, the fillPeaks method with the default parameters was applied to fill missing peak data. The obtained peak tables were imported into MATLAB (Mathworks Inc., Natick, MA, USA) for data analysis. A total of 3320 features were initially detected after peak detection, chromatographic de-convolution and peak alignment of the entire batch. Blank samples were used to identify features arising from source contaminants and other sample components originating from e.g. tubes, solvent impurities and anticoagulant that were removed, leaving a total of 552 variables.

Support Vector Regression was carried out in MATLAB using the LIBSVM library[10] (software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm). MATLAB scripts and the raw peak table used in this work are available from the authors or the Neonatal Research Unit website (www.perinox.es). Batch effect correction using robust splines (QC-RSC) was performed using MATLAB scripts kindly provided by Dr. D. Broadhurst.

3. METHODS

Correction of the intra-batch effect using QC-SVRC

Detailed descriptions of Support Vector Regression theory can be found in [11,12,13]. Consider a training dataset of QCs $\{(x_{QC(i)}, y_{QC(i)}) | x_{QC(i)} \in R,\ y_{QC(i)} \in R, i = 1, \dots, n\}$, where n = number of QCs included in the training set. A regression function is defined as $\widehat{y_i} = f(\boldsymbol{x_i}, \boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{x_i} \rangle + b$, where $\boldsymbol{w}$ is the weight vector, b is the bias term and $\widehat{y_i}$ is the estimated value of y_i ($y_i = \widehat{y_i} + e_i,\ e_i \in R$ are the residuals of the model). In ε-SVR the objective is to find a function as flat as possible that has $e_i < \varepsilon$ for the whole training set. If there is no function $f(\boldsymbol{x})$ that approximates all pairs $(\boldsymbol{x_i}, y_i)$ with ε accuracy, slack variables (ξ_i, ξ_i^*) can be introduced leading to the following formulation of the SVR optimization problem[13]:

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\|\boldsymbol{w}\|^2 + C\textstyle\sum_{i=1}(\xi_i + \xi_i^*) \\ &\text{subject to} && y_i - \langle \boldsymbol{w}, \boldsymbol{x_i} \rangle - b \le \varepsilon + \xi_i \\ & && \langle \boldsymbol{w}, \boldsymbol{x_i} \rangle + b - y_i \ \le \varepsilon + \xi_i^* \\ & && \xi_i, \xi_i^* \ge 0 \end{aligned}$$

The SVR initially maps the sample data into a second high dimensional feature space by means of a function $\varphi(\boldsymbol{x})$. Optimization is solved as a quadratic programming problem through the use of Lagrange multipliers and a kernel function $K(\boldsymbol{x_i} \cdot \boldsymbol{x_j}) = \langle \varphi(\boldsymbol{x_i}), \varphi(\boldsymbol{x_j}) \rangle$ to replace the inner product operation of the high dimensional feature space. After training and solving the quadratic problem the solution is expressed as:

$$\hat{y} = \sum_{x_i \in SV} (\alpha_i - \alpha_i^*) K(\boldsymbol{x_i}, \boldsymbol{x}) + b$$

where α_i, α_i^* are Lagrange multipliers, and the samples with non zero weights α_i, α_i^* are called the Support Vectors (SVs). Non SVs are redundant and are not used for the construction of the function $f(\boldsymbol{x_i}, \boldsymbol{w})$[13].

In this work, the kernel function used was the Radial Basis Function (RBF).

$$K(\boldsymbol{x_i}, \boldsymbol{x}) = \exp\left(-\gamma \|\boldsymbol{x} - \boldsymbol{x_i}\|^2\right)$$

where γ is the width of the RBF kernel.

The performance of the RBF-SVR is determined by the optimal selection of the ε-insensitive loss parameter, the error penalty parameter C and the kernel parameter γ. The ε-insensitive loss function does not penalize training errors $e_i < \varepsilon$, whereas further deviations are linearly penalized and so, its value affects the number of SVs. Larger ε threshold values result in fewer SVs selected and less complex regression estimates which are less prone to model overfitting. The user-defined constant C is the cost associated with the training error. While selecting large C values may lead to many SVs and overfit the model, small C values may lead to underfitting. The kernel parameter γ controls the width of the Gaussian kernel. Large γ values reduce the radius of the area of influence of the SVs and lead to model overfitting. Cross-validation (CV) is typically used to trade off training error against model complexity using different strategies (e.g. grid search) for the selection of the optimal set of C , γ and ε values. Once obtained the QC-SVR curve, the complete sample set is corrected by subtracting the predicted $\hat{y}$ value at each point $\boldsymbol{x_i}$ ($i = number\ of\ samples\ in\ the\ batch$).

4. RESULTS AND DISCUSSION

4.1 Data overview

A common way to evaluate intra-batch effects is by visualization means. In the absence of an intra-batch effect, the intensity of every metabolite in QCs as a function of the injection order should only show random variation as a normal white noise process. On the contrary, strong batch influence will make the intensity profiles deviate from this behavior. An illustration of the intra-batch effect observed is depicted in Figure 1, where different drifts in the intensities of endogenous metabolites and ISs were found as a function of the injection order.

Unsupervised principal component analysis (PCA) can also be used for a visual analysis of the intra-batch effect. PCA scores plots provide an unbiased overview of the data structure and can be used to identify trends in the data as a function of the injection order. Visual analysis of PCA scores plots showed that the first two PCs accounting for 67% of the total variance of the original data revealed a significant association with the injection order (see Figure 2).

Visualization tools are useful for a fast inspection of the data but for a more rigorous assessment quantitative measures should also be used. The distribution of values of %RSD in QCs (%RSD$_{QC}$) provides an overview of the precision of the analysis throughout the batch. The difference in the %RSD$_{QC}$ distributions before and after batch effect elimination can be used to evaluate the accuracy of the correction. The main shortcoming of this approach is that the distribution of %RSD$_{QC}$ after batch effect correction is overoptimistic because the same data is used for the estimation of the RS or SVR functions. On the other hand, the analysis of the %RSD distributions in the samples (%RSD$_{samples}$) fails to detect whether the correction modified the unknown biological distributions. To avoid these potential pitfalls, in this study we used the repeated analysis of a single sample that allowed a straightforward quantification of both, intra-batch effects and the accuracy of their correction. It also facilitated the interpretation of changes in distribution of %RSD$_{samples}$ after batch effect correction and the evaluation of the correction performance. However, an inherent disadvantage of this design is that the inter individual variation per peak cannot be estimated and so, the identification of variables containing biological information after batch effect correction is not as efficient as by including the analysis of sample replicates from several different individuals. Figure 3 (left) shows the cumulative distribution function of %RSD$_{samples}$ before batch effect correction using one injection out of 8 as QC. Data depicted showed a clear intra-batch effect that reduced the percentage of variables with %RSD$_{samples}$<15 down to 4% (22/552).

4.2 Intra-batch effect correction

The performance of the QC-SVRC is determined by an optimal selection of the ε-insensitive loss parameter, the error penalty C and the kernel parameter γ. The selection of the optimal set of C, γ and ε values can be made through a grid search, using the root mean square error of leave-one-out cross validation (LOOCV) to trade off training error against model complexity. In a basic (*n x n x n)* uniform grid search the function performance is evaluated at each point represented by a $\{\varepsilon, C, \gamma\}$ combination. The optimum values of C, γ and ε are then selected as those for which the RMSECV shows the minimum value. This approach achieves high accuracy and reproducibility, but costs too much computing time and the intervals have to be carefully selected. To reduce computation time and facilitate QC-SVRC, the parameter C was calculated for each variable as the difference between the 10th and 90th quartile of the output values in the QC samples. Mattera et al.[14] proposed the parameter C to be equal to the range of output values. However, the difference between the 10th and 90th quartile was selected in this work to reduce the impact of potential outliers in the C value. Furthermore, Smola et al.[15] proposed a ε-insensitive loss parameter proportional to the noise variance and Cherkassky et al.[16] proposed the selection of ε based on the estimated noise. Accordingly, the ε-insensitive loss parameter for each variable was selected as 7.5% of the output value in the first QC, based on our previous experience with the precision of the UPLC-MS system (ca 15%). Finally, the kernel parameter γ was selected by LOOCV in the [2^{-3}, 2^{-2},... , 2^{6}] range reducing the search down to 10 iterations. Besides, outlying or missing QC values arising from e.g. incorrect peak alignment or low injection volume were left out from the SVR function calculation. If the number of retained QCs was <5, the variable was not corrected. Likewise, outlying 'zero samples' were not corrected. In this example, for each variable, an injection was classified as outlier using 20% of the median intensity in QCs as lower threshold value. QC-RSC follows a similar protocol when the number of QCs <5 also replacing zeros with missing values. In QC-RSC, a simple linear regression is used if the number of QCs is <5. A peak is assumed to be too unstable for correction if de number of QCs is <3; in this case, the peak is deleted.

As an example of the method performance, Figure 4 shows the accuracy as RMSECV using different γ values for the correction of the signals of L-carnitine, phenylalanine-D_5, caffeine and LysoPC(14:0), and 18 equally spaced QCs distributed throughout the batch (i.e. one injection out of 8 injections). Figure 5 shows the effect of different ε-insensitive values on the estimated SVR and the number of SVs. Over- or underestimation of the expected instrumental variation in the absence of batch effects would lead to inaccurate corrections and, hence it must be carefully selected. The resulting models depicted in Figure 1 showed the capacity of the SVR to model the intra-batch effect while controlling overfit with no support vectors within the ε-insensitive regions. Corrected intensities depicted in Figure 6 showed that the intra-batch effect was accurately corrected with most of the samples lying in the ε-insensitive regions. Variation around the initial value in samples after correction could be attributed to a poor injection volume precision and the automatic integration of the chromatographic peaks. Figure 7 shows results from the projection of the data before and after QC-RSC and QC-SVRC onto the PCA model built using QC samples before correction. The comparison of the PC1-PC2 scores showed that both; QC-RSC and QC-SVRC reduced the total amount of variance and the size of the batch-effect. Finally, Figure 3 shows the cumulative distribution functions of $\%RSD_{samples}$ after intra-batch effect correction. The correction clearly improved the precision and increased the percentage of variables with $\%RSD_{samples}<15$ up to 57% (313/552).

As abovementioned, the outcome of the correction depends on the distribution of QCs throughout the batch. To evaluate the performance of the correction using different QC distributions, the correction procedure was carried out with the number of samples between QCs ranging from 4 to 16. As shown in Figure 8, the use of a high number of QCs improved the accuracy of the correction, but it also reduced the applicability of the approach and so, the number of QCs should be as few as possible under the condition that the fitting accuracy is good enough. Using the number of variables with $\%RSD_{samples}<15$ as quality parameter of the correction accuracy, QC-SVRC slightly outperformed QC-RSC (see Figures 3 and 6). As an example, 275 (49.8%) and 313 (57%) variables showed %$RSD_{samples}$<15 after QC-RSC or QC-SVRC, respectively when one injection out of 8 injections was used as QC (see Figure 3, right).

5. Conclusions

Intra-batch effects affecting the LC-MS instrumental response reduce the repeatability and reproducibility of the analysis, and may decrease the power to detect biological responses. Results show that the use of QC samples and SVR with a Radial Basis Function kernel is a useful non-parametric approach to model changes in the instrument response. Besides, the ε-insensitive loss parameter allows adjusting the SVR parameters to the instrument performance. QC-SVRC was compared to a reference algorithm based on robust splines (QC-RSC) using data obtained from the repeated analysis of a plasma sample. This design enabled a fast evaluation of the effect of the distribution of QCs along the batch but did not allowed a comparison of the instrumental and biological variation for a subsequent identification of unreliable peaks after correction. The experimental design of the QC-RSC strategy includes the analysis of replicates from several individuals, thus allowing the assessment of the correction performance by measuring within and between individual variance per peak[6]. Moreover, the fact that QC-SVRC requires three parameters to be optimized compared to a single parameter for QC-RSC may increase the computational cost of the correction. Further work is required to assess whether the assumption presented in this work to reduce the number of iterations during the training phase is generalizable across new data sets and experimental conditions (e.g. multiple batches). Nonetheless, QC-SVRC slightly outperformed QC-RSC for the studied data set. Although it is difficult to estimate the statistical significance of the difference, QC-SVRC can be considered as a valuable tool complementary to QC-RSC. Further research will be carried out to compare the robustness of the QC-SVRC by analyzing the effect of noise in QC samples on both, the estimated functions and the correction accuracy, as well as the use of alternative kernels (e.g. polynomial, splines), optimization procedures (e.g. simplex) or type of SVR (e.g. least squares SVR, LSSVR) and experimental designs (e.g. including additional QCs or sample replicates).

Acknowledgements

JK acknowledge the 'Sara Borrell' grant (CD12/00667) from the *Instituto Carlos III* (Ministry of Economy and Competitiveness, Spain). ASI acknowledges the support of the Spanish Maternal and Child Health and Development Network SAMID RD12/0022/12 (*Instituto Carlos III*, Ministry of Economy and Competitiveness, Spain) and FEDER funds (European Community). MV acknowledges the FISPI14/0433 grant from the *Instituto Carlos III* (Ministry of Economy and Competitiveness, Spain). GQ acknowledges financial support from the Spanish Ministry of Economy and Competitiveness (SAF2012-39948).

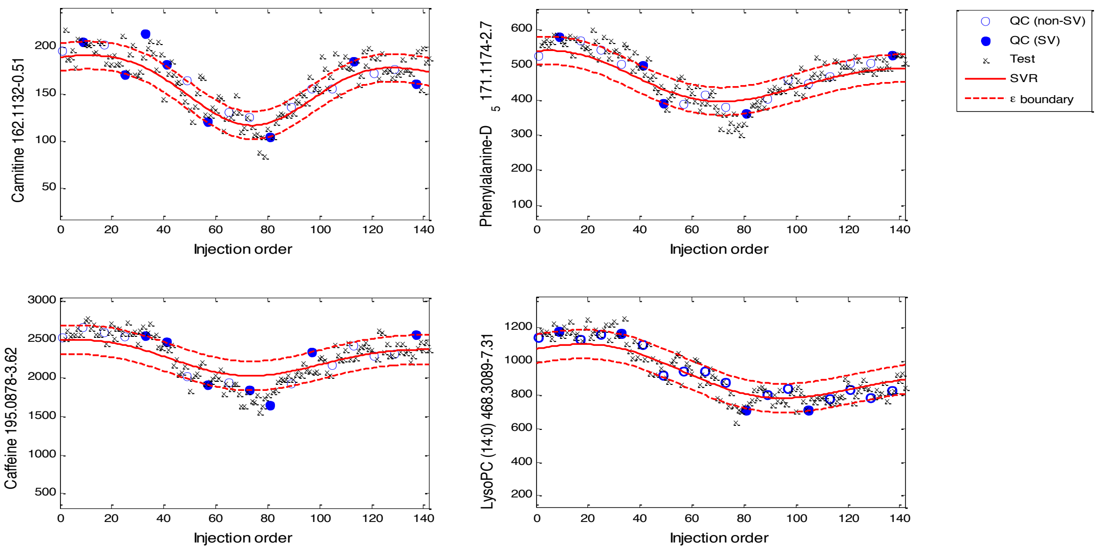


Figure 1. Intensity of L-carnitine, phenylalanine-D$_5$, caffeine and LysoPC(14:0) as a function of the injection order. The red line depicts the SVR calculated using one out of every 8 chromatograms as QC for QC-SVRC. The ε boundaries are given with the red dotted line.

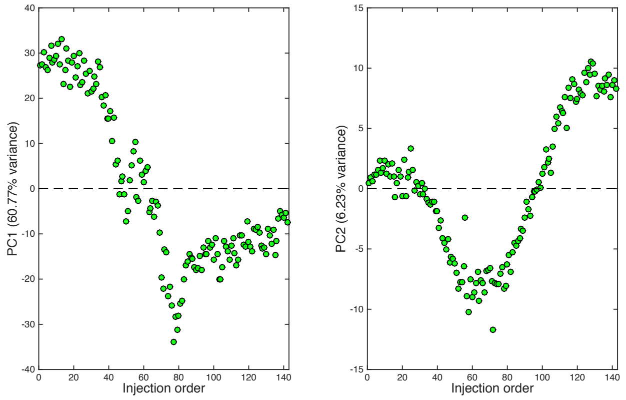


Figure 2. PCA of the data set before correction using autoscaling as data pretreatment.

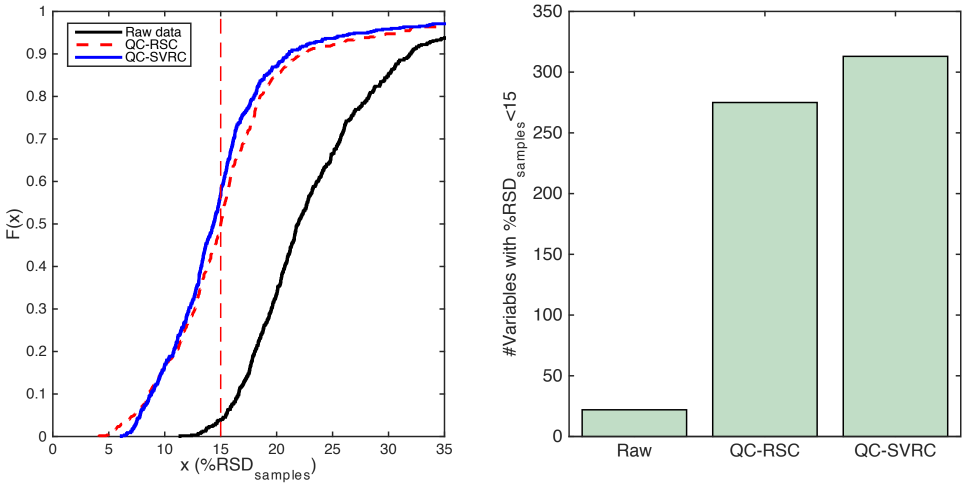


Figure 3. Left) Cumulative distribution functions of the $\%RSD_{samples}$ in the original data set and after QC-RSC and QC-SVRC using one of every 8 injections as QC (total number of variables = 552); Right) Number of variables showing $\%RSD_{samples}<15$. Note: Search range for B-spline smoothing parameter in QC-RSC: [0:0.5:7]. Search range for the kernel parameter γ in QC-SVRC: [2^{-3}, 2^{-2},… , 2^{6}].

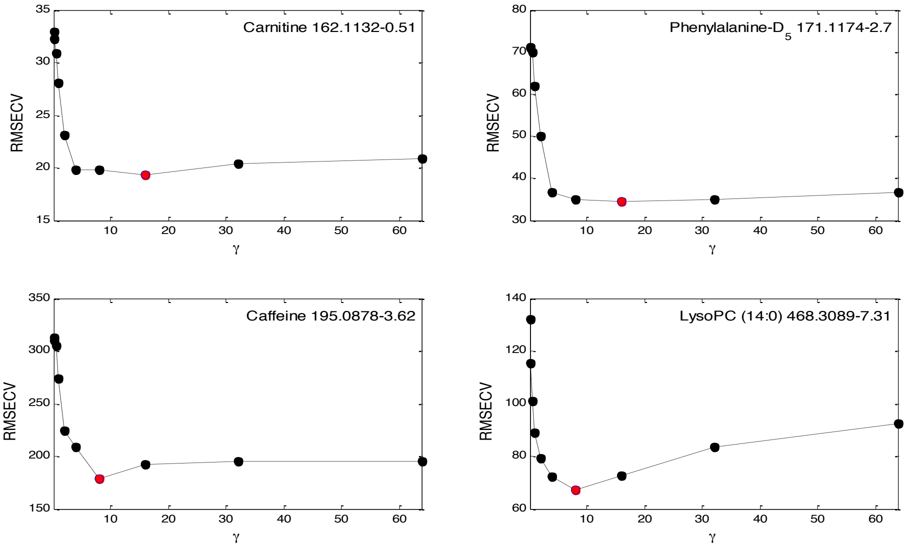


Figure 4. RMSECV obtained as a function of the γ value used for the correction of the signal of L-carnitine, phenylalanine-D$_5$, caffeine and LysoPC(14:0) using QC-SVRC and one of every 8 chromatograms as QC. Red dots indicate the selected value.

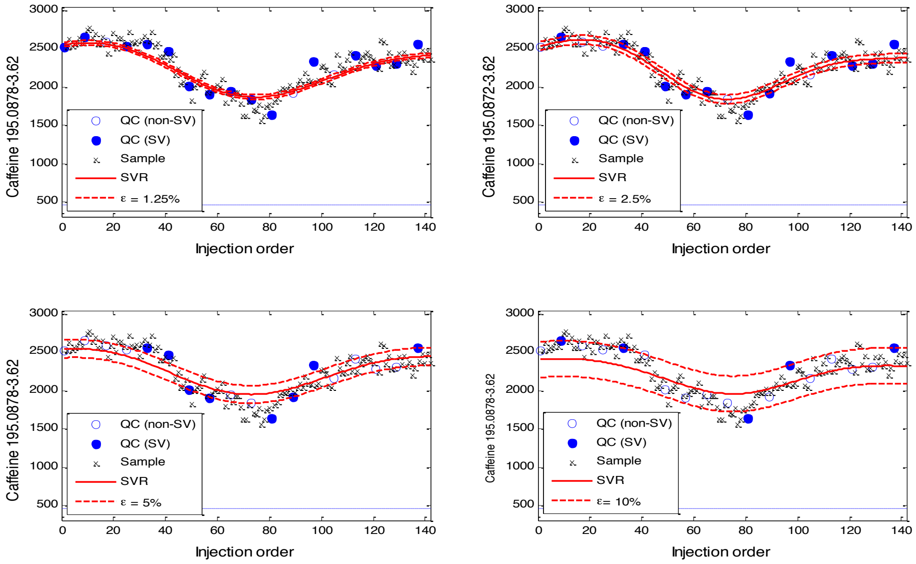


Figure 5. Effect of ε-insensitive loss parameter in the SVR estimated for the correction of the intra-batch effect for L-carnitine. The red line depicts the SVR calculated using one out of every 8 chromatograms as QC for QC-SVRC. The ε boundaries (1.25, 2.5, 5, and 10%) are indicated by the red dotted line.

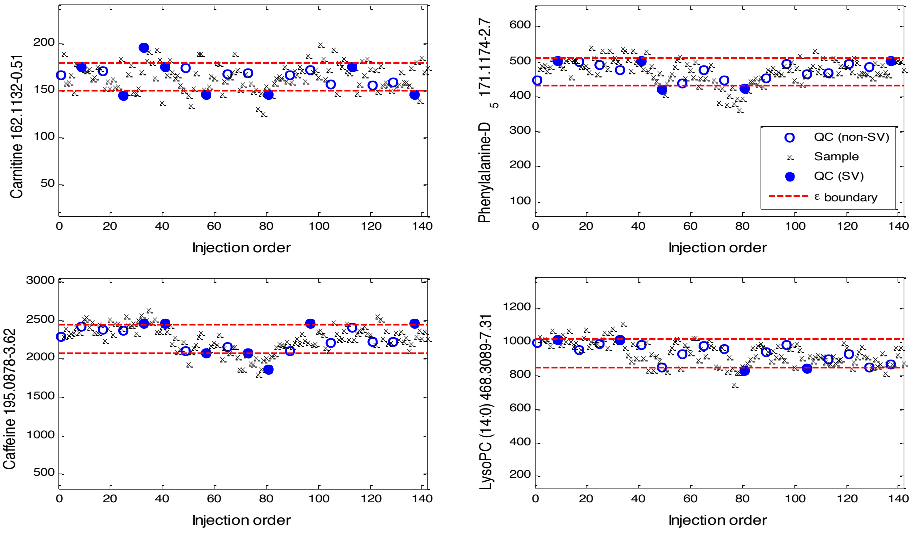


Figure 6. Intensity of L-carnitine, phenylalanine-D_5, caffeine and LysoPC(14:0) after QC-SVRC calculated using one of every 8 injections as QC.

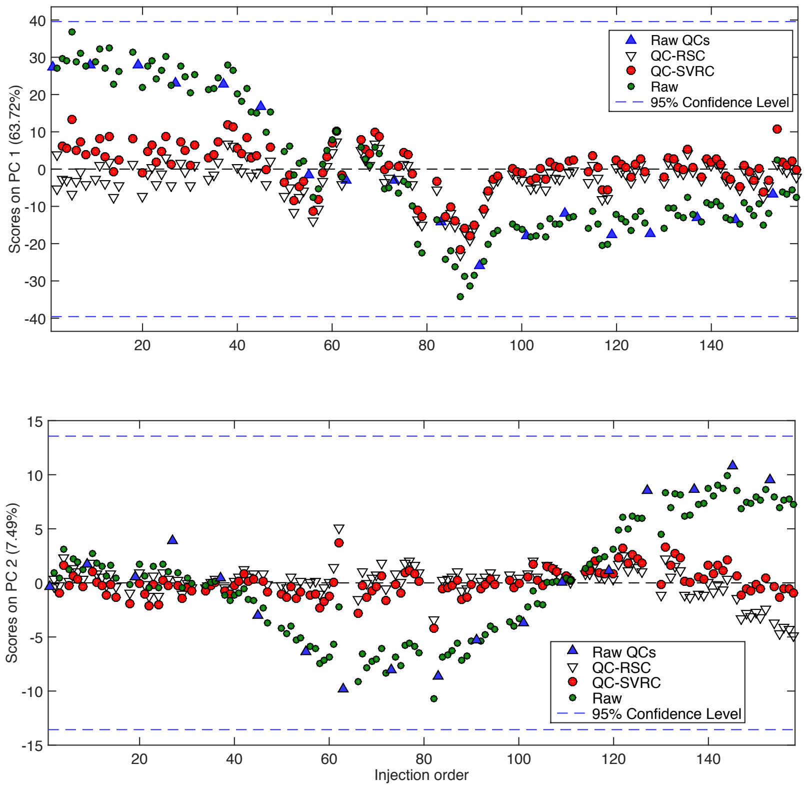


Figure 7. PCA of the data set before and after QC-RSC or QC-SVR correction using one out of every 8 chromatograms as QC. A PCA model was calculated using QC samples before intra-batch effect correction (i.e. raw data). PC scores of samples (i.e. non QCs) before and after QC-RSC or QC-SVRC were obtained by projection onto this model.

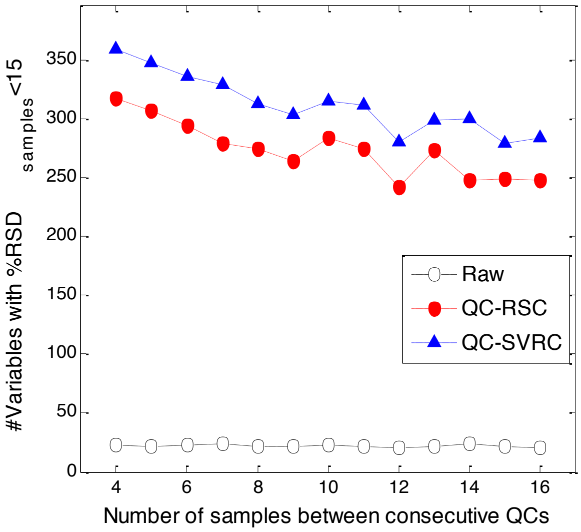


Figure 8. Number of variables showing $\%RSD_{samples}<15$ after QC-RSC and QC-SVRC correction as a function of the number of QCs dispersed evenly throughout the batch (total number of variables = 552). Note: Search range for B-spline smoothing parameter in QC-RSC: $1/(1+epsilon*10^{p})$, where p=[0:0.5:7]. Search range for the kernel parameter γ in QC-SVRC: [2^{-3}, 2^{-2},... , 2^{6}].

References

1 R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan and D. B. Kell, *Trends Biotechnol.*, 2004, **22**, 245–252.

2 M. Sysi-Aho, M. Katajamaa, L. Yetukuri and M. Orešič, *BMC Bioinformatics*, 2007, **8**, 93.

3 B. Zhou, J. F. Xiao, L. Tuli and H. W. Ressom, *Mol. Biosyst.*, 2012, **8**, 470–481.

4 J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly and R. A. Irizarry, *Nat. Rev. Genet.*, 2010, **11**, 733–739.

5 S.-Y. Wang, C.-H. Kuo and Y. J. Tseng, *Anal. Chem.*, 2013, **85**, 1037–1046.

6 J. A. Kirwan, D. I. Broadhurst, R. L. Davidson and M. R. Viant, *Anal. Bioanal. Chem.*, 2013, **405**, 5147–5157.

7 D. Aydin, 2007, WASET, 2007, **1**, 558-592.

8 V. N. Vapnik, *IEEE Trans. Neural Netw. Publ. IEEE Neural Netw. Counc.*, 1999, **10**, 988–999.

9 G. Camps-Valls, L. Bruzzone, J. L. Rojo-Alvarez and F. Melgani, *IEEE Geosci. Remote Sens. Lett.*, 2006, **3**, 339–343.

10 C.-C. Chang and C.-J. Lin, *ACM Trans. Intell. Syst. Technol. TIST*, 2011, **2**, 27.

11 C. Cortes and V. Vapnik, *Mach. Learn.*, 1995, **20**, 273–297.

12 V. Vapnik, S. E. Golowich and A. Smola, in *Advances in Neural Information Processing Systems 9*, MIT Press, 1996, pp. 281–287.

13 A. J. Smola and B. Schölkopf, *Stat. Comput.*, 2004, **14**, 199–222.

14 D. Mattera and S. Haykin, eds. B. Schölkopf, C. J. C. Burges and A. J. Smola, MIT Press, Cambridge, MA, USA, 1999, pp. 211–241.

15 A. J. Smola, N. Murata, B. Schölkopf and K.-R. Müller, in *ICANN 98*, eds. L. Niklasson, M. Bodén and T. Z. MSc, Springer London, 1998, pp. 105–110.

16 V. Cherkassky and Y. Ma, *Neural Netw.*, 2004, **17**, 113–126.