Journal of Pharmaceutical and Biomedical Analysis 47 (2008) 659–669

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

Review

The Basis Function Regression in pharmaceutical analysis Theory and example of application

Łukasz Komsta*, Robert Skibiński, Marta Paryło, Karolina Dudek

Department of Medicinal Chemistry, Medical University of Lublin, Jaczewskiego 4, 20-090 Lublin, Poland

ARTICLE INFO

Article history:
Received 29 November 2007
Received in revised form 13 February 2008
Accepted 18 March 2008
Available online 26 March 2008

Keywords:
Basis Function Regression
B-splines
Dimension reduction
UV spectrophotometry
Chemometrics

ABSTRACT

The BFR (Basis Function Regression) is an interesting alternative to common techniques (such as PCR or PLS) in chemometrics. It is based on projecting the spectral information onto some number of equally spaced spline bases, than obtaining integrals of resulted curves. Existing references show that in certain cases it can reduce almost twice the RMSEP values. As this technique is not so popular in chemometrics nor applied in pharmaceutical analysis, it is desirable to present its theoretical considerations and implementation (with example MATLAB/Octave code). As an illustrative example we present the chemometric model for content recognition of a tablet (12 possible compounds in binary or ternary combinations) from the UV spectrum of its methanolic extract. The BFR technique gave lowest prediction error and the estimators obtained have more meritorical meaning than in case of PCR, PLS and other techniques used for comparison. In our opinion this technique should be considered in any chemometric approach as it often shows better performance.

Contents

1. Introduction

The last decade of the 20th century resulted in a significant development in computational technology. An average personal computer has now enough performance to handle large datasets numerically and to apply the complicated algorithms to them. Parallel development of measurement instrument technology (and possibility of generating large datasets in computer-transferable form) resulted in the expansion of chemometric techniques.

The main advantage of chemometric approaches to spectroscopic data is the use of complete spectra (saved as vectors of absorbance values) as the explaining variable. There is no need to perform analytical wavelength selection, because the chemometric algorithms perform all necessary extraction of interesting information from the data (and filter it from the remaining information, or noise). The information extracted (predicted) from the

* Corresponding author. Tel.: +48 81 742 3692, fax: +48 81 742 3691.
E-mail address: lukasz.komsta@am.lublin.pl (Ł. Komsta).

0731-7085/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2008.03.016

spectrum can be very complex—for example, the famous datasets from Kalivas [1] are used to predict the octane number of gasoline, or water and protein content in wheat, from their NIR spectra. The chemometric approaches in UV region can extensively increase the performance of the methods, when the spectra of several compounds are overlapped and classical methods are very problematic.

1.1. Spectroscopic calibration

In the spectroscopic calibration dataset, consisting of n samples, each recorded on p wavelengths, the spectra can be arranged into the matrix $\mathbf{X}(n \times p)$, where the columns correspond to the wavelengths, and the rows are the samples. Each sample has then arranged an y value which is the explained (searched) variable. They form the column vector $\mathbf{y}(n \times 1)$.

The task for chemometric method is to find the row vector β (called estimator), which conforms to the following equation:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{1}$$

where ϵ is a vector of normally distributed measurement errors. Such model does not contain the intercept term. Incorporating the intercept would significantly increase the complication of the computations. The better way is to center around the mean (and sometimes scale around the variance) each column of $\mathbf{X}$ and remember these values. But in practice, the centering of spectroscopic calibration matrix is in many cases unnecessary [2]. The scaling is almost always unnecessary, because the columns contain the same units.

1.2. Ordinary least squares disadvantages

The simpliest method for obtaining the β vector would be the use of Ordinary Least Squares (OLS):

$$\beta = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{2}$$

Unfortunately, two main problems appear here. First, the absorbance values are collinear, so the crossproduct $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is almost singular. Computing its inverse is often impossible, because its determinant lies besides the computational accuracy. Although this problem can be partially omitted by using the Moore-Penrose pseudoinverse of $\mathbf{X}$, the second problem appears here. The wavelength number (p) is in most cases larger than sample number (n). This leads to infinite number of correct (ideal) solutions, which fit Eq. (1) without any error. The solutions are instable, with high internal variance, and additionally overfitted, because their predictive ability [3] is unnacceptable.

1.3. The popular chemometric regression techniques

A number of techniques dealing with the collinearity and ambiguity have been proposed. They produce biased estimates (not ideally fitted), but with acceptable prediction ability. The optimization of such method is done by cross-validation to minimize RMSEP (Root Mean Square Error of Prediction) value. Historically, the first technique was the Ridge Regression (RR) [4]. The idea is based on adding the small constant to diagonal elements of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ crossproduct, decreasing its singularity.

The most popular techniques used today in chemometrics are based on Principal Component Analysis [5]. The Principal Component Regression (PCR) [6] uses the k first principal components of the calibration matrix as the regressors. The Partial Least Squares (PLS) [7] takes into account also the response and maximises the covariance between them. The Cyclic Subspace Regression (CSR) [8] and Continuum Power Regression (CPR) [9] are the generalized techniques. With certain parameter values they can be equivalent to PCR, PLS or OLS. The intermediate cases often show better predictive ability. The variable selection methods, such as Forward Stepwise Regression (FSR) [10] or Least Angle Regression (LARS) [11] are often used as the comparative methods.

1.4. The Basis Function Regression

The idea of Basis Function Regression, presented in this paper, was first introduced by Hastie and Mallows [12]. They discussed the use of smoothing splines for the estimation of coefficient function over the wavelength. Goutis [13] applied the similar method to the second derivative of the spectrum. Marx and Eilers [14] proposed for the first time the projection onto some number of the equally spaced B-spline bases. The use of B-splines as the tool for selection of spectral variables was published recently by Rossi et al. [15].

Rasmussen [16] extended the Eilers approach to any number of basis functions (even more than the number of samples) by use of the integral over the wavelengths (and such variant is presented in our paper). He applied the method to the Kalivas datasets [1] and obtained about twice lower prediction error compared to classical methods such as PCR and PLS, which is an obviously significant improvement.

Although the BFR seems to be very interesting alternative, there is general lack of its use. Searching bibliographic databases (Scopus, etc.) we did not find no other papers describing the use of BFR in chemical analysis. Therefore we have decided to present this technique with sample implementation and application.

2. Theory

2.1. The basis function transformation

Denote $\mathbf{X}$ as a matrix of calibration spectra, containing n rows (samples) and p columns (wavelengths) and X as one of its spectra (p-length vector). Next, denote $\mathbf{B}$ as a matrix of the k B-spline basis functions $B_1 \ldots B_k$, containing p rows (wavelengths) and k columns. The basis functions are obtained numerically and the details about their generation are given in literature [17]. From practical point of view, $\mathbf{B}$ is the matrix containing peak-shaped curves. If the functions are linear, the peaks are triangular, otherwise they are smooth.

The following properties of the basis function matrix are interesting in our case. The sum of all values for the same wavelength (sum of one row) is equal to 1:

$$\sum_{i=1}^{k} B_i(\lambda) = 1 \tag{3}$$

When multiplying any spectrum by series of the basis functions, the resulting sum of the values at any wavelength is equal to the absorbance at that wavelength at original spectrum:

$$\sum_{i=1}^{k} B_i(\lambda) X_\lambda = X_\lambda \tag{4}$$

Therefore, by computing sums of $B_i(X)$ over the wavelength one can obtain k integrals, and their sum is equal to the integral of the original spectrum (sum of absorbances):

$$\sum_{i=1}^{p}\sum_{j=1}^{k} B_j(\lambda_i) X_{\lambda_i} = \sum_{i=1}^{p} X_{\lambda_i} \tag{5}$$

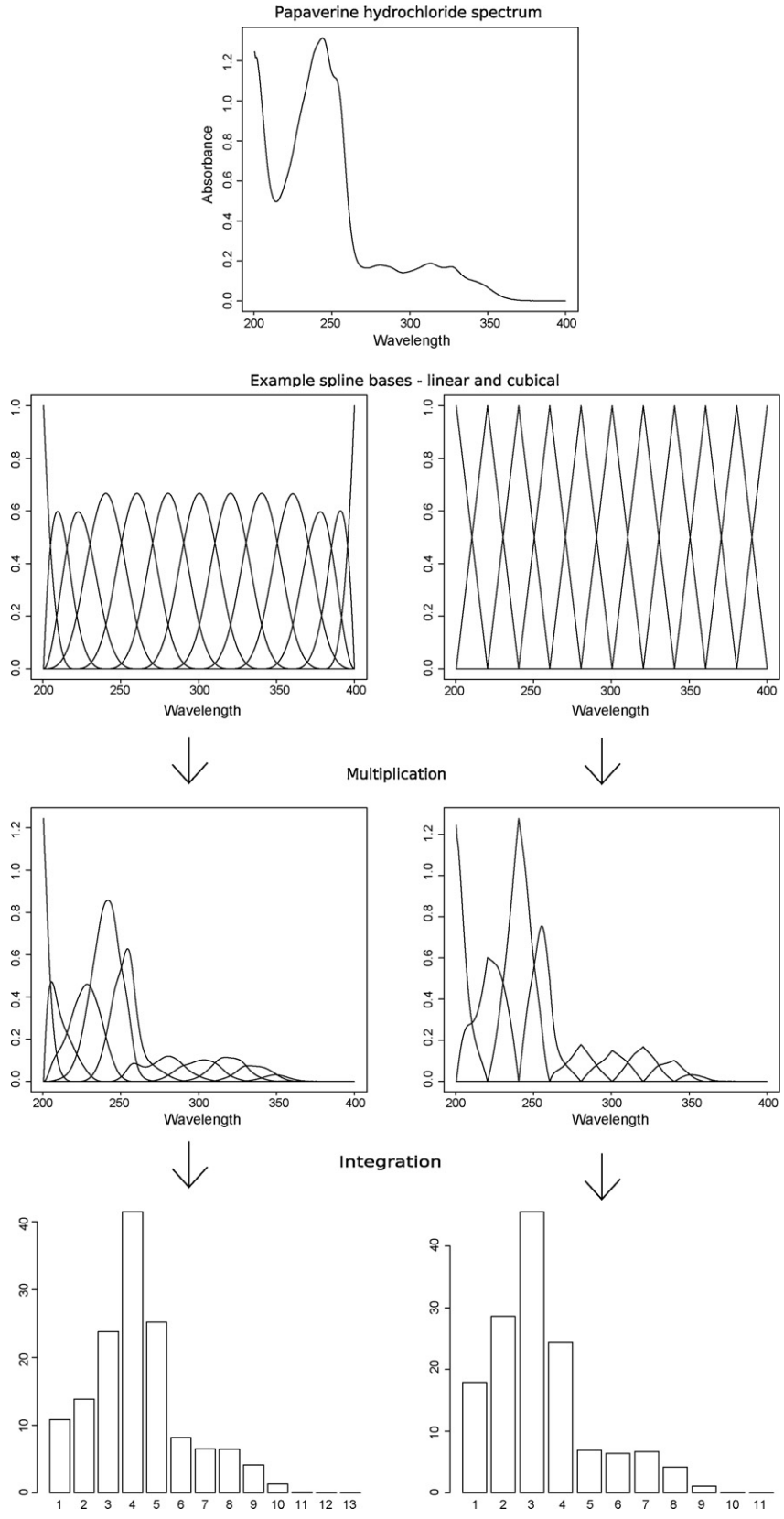

Fig. 1. The example of BFR transformation using 13 linear and cubical splines.

This property guarantees that no information from spectrum is lost. A spectrum of p individual absorbances is transformed into k variables, which can be treated as "subintegrals", emphasizing the subsequent regions of the spectrum. The graphical illustration of an example spectrum transformation is presented in Fig. 1.

The obtained subintegrals can be then used (instead of p absorbances) in classical least squares regression, preventing the

ambiguity and overfitting problem. The collinearity problem cannot be totally avoided here, but at some optimal number of k the obtained variables are not collinear.

2.2. The algorithm and its implementation

As above, we have a matrix $\mathbf{X}(n \times p)$. We generate a matrix $\mathbf{B}(p \times k)$ containing k basis functions of degree r. By simple matrix multiplication, we obtain reduced subintegrals:

$$\mathbf{Z} = \mathbf{XB} \tag{6}$$

The matrix $\mathbf{Z}$ contains then n rows (samples) and k columns (subintegrals). It is used in classical Ordinary Least Squares manner to predict $\mathbf{y}$:

$$\mathbf{U} = [(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y}]^{\mathrm{T}} \tag{7}$$

Resulting estimators $\mathbf{U}$ lie still in reduced dimensions ($k \times 1$). To convert it to the estimator $\beta(p \times 1)$ we use the final multiplication:

$$\beta = \mathbf{UB}^{\mathrm{T}} \tag{8}$$

The implementation of the above algorithm in any language which supports matrix multiplication is fairly simple. The only difficulty is the method used to generate the basis function matrix. In the GNU R computational environment (www.r-project.org [18]), there is a built-in function `bs()`, which do the work for us. In MATLAB (www.mathworks.com) and GNU Octave (www.octave.org), which are the unofficial standard for the chemometric calculations, there is no easy way to generate spline bases. We have found no ready-to-use routines either in Internet, or in the literature, and finally obtained a sample code from Dr. Graeme A. Chandler (see Acknowledgement). The obtained code snippet was cleaned up, reorganized and tuned especially for the specified task.

The resulting MATLAB/Octave code is presented below and consists of two functions. The first function `bfr()` is the main BFR routine. It accepts the spectral matrix $\mathbf{X}$, the column vector $\mathbf{y}$, the degree of the basis functions r and their number k. The subroutine `bspline()` is called from this function and generates the basis function matrix with $k+r+1$ columns ($k+r-1$ internal knots and 2 boundary knots). The result of the `bfr()` function is a matrix with k columns, each of them containing the β for $1 \ldots k$ used knots.

```
function [beta] = bfr(X,y,r,k);

        [n,p] = size(X);
        [nk] = length(k);

        beta = zeros(p,nk);

        for i=1:nk
                B = bspline(r,k(i),p);
                Z = X*B;
                U = (inv(Z'*Z)*Z'*y)';
                beta(:,i) = (U*B')';
        end
end

function [sp] = bspline(r,k,p);

        xx = (1:p)';
        t = round((1:k)./(k+1).*p);
        t = [repmat(1,1,r+1) t repmat(p,1,r+1)]';
        n = length(t);
        a = eye(length(t)-r-1);
```

```
        k = round(r(1)+1);
        na = n-r-1;
        tt = [t(1)*ones(k-1,1) ; t ; t(n)*ones(k-1,1)];
        aa = [zeros(k-1,na) ; a ; zeros(k-1,na)];
        nx0 = sum(xx<t(1));
        nx1 = sum(xx>t(n));
        xx = xx(t(1)<=xx & xx <= t(n));
        nx = length(xx);
        ox = ones(nx,1);
        zx = zeros(nx,1);
        ix = ones(nx,1);

        for j=(2:n-1)
                ix = ix + (xx>=t(j)).*(t(j) < t(n));
        end

        ix = ix+k-1;
        b = ones(nx,1);
        ik = ix;

        for kk = 2:k
                ok = ones(1,kk);
                ik = ix*ok-kk+ox*(1:kk);
                xk = xx*ok;
                d1 = (tt(ik+kk)-tt(ik+1));
                d1 = d1+(d1==0);
                d1 = (tt(ik+kk)-xk)./d1;
                d0 = (tt(ik+kk-1)-tt(ik));
                d0 = d0+(d0==0);
                d0 = (xk - tt(ik))./d0;
                b = [zx b].*d0+[b zx].*d1;
        end

        b = reshape(b,nx,k);

        sp = (b(:,1)*ones(1,na)).*(aa(ik(:,1),:));

        for kk = 2:k
                sp = sp + (b(:,kk)*ones(1,na)) ...
                        .*(aa(ik(:,kk),:));
        end

        sp = [zeros(nx0,na) ; sp ; zeros(nx1,na)];

end
```

3. Experimental

The aim of the example application elaborated in our laboratory was to obtain the chemometric model useful for OTC antipyretic and analgesic tablet content recognition. The model was trained against 12 possible ingredients: acetaminophen (ACE), aspirin (ASP), caffeine (CAF), codeine phosphate (COD), dextrometorphan hydrobromide (DEX), dipyrone (DIP), ethoxybenzamide (ETO), ibuprofen (IBU), phenylephrine hydrochloride (PHE), propyphenazone (PRO), pseudoephedrine hydrochloride (PSE) and vitamin C (VIT).

All the substances used were of appropriate purity (Sigma–Aldrich, USA). The spectra were recorded in spectroscopic grade methanol (POCH, Gliwice, Poland) using Hitachi UV-2001 double-beam spectrophotometer in 1 cm quartz cells, in range 200–300 nm with 0.5 nm resolution (200 absorbances).

The formulations used in the investigation were bought in local a drugstore:

- Antidol 15—tablets (500 mg acetaminophen, 15 mg codeine phosphate), produced by Lek, series No. 7420111H.
- Apap C plus—effervescent tablets (500 mg acetaminophen, 300 mg ascorbic acid), produced by US Pharmacia, series No. 607275.
- Aspirin C—effervescent tablets (400 mg aspirin, 240 mg ascorbic acid), produced by Bayer, series No. BTA 5902.
- Cefalgin—tablets (250 mg acetaminophen, 150 mg prophyphenazone, 50 mg caffeine), produced by Polfa Pabianice, series No. 60506.
- Coffepirine—tablets (450 mg aspirin, 50 mg caffeine), produced by Marcmed Lublin, series No. 060407.
- Codipar plus—tablets (500 mg acetaminophen, 65 mg caffeine), produced by GlaxoSmithKline, series No. K06005.
- Coldrex HotRem—sachets (750 mg acetaminophen, 10 mg phenylephrine hydrochloride, 60 mg ascorbic acid), produced by GlaxoSmithKline, series No. 6033.
- Coldrex MaxGrip—sachets (1000 mg acetaminophen, 10 mg phenylephrine hydrochloride, 40 mg ascorbic acid), produced by GlaxoSmithKline, series No. Z217.
- Dafalgan Codeine—tablets (500 mg acetaminophen, 30 mg codeine phosphate), produced by UPSA, series No. J9752.
- Efferalgan—effervescent tablets (330 mg acetaminophen, 200 mg ascorbic acid), produced by UPSA, series No. J8581.
- Etopiryna—tablets (300 mg aspirin, 100 mg ethoxybenzamide, 50 mg caffeine), produced by Polpharma, series No. 71106.
- Gardan P—tablets (200 mg prophyphenazone, 300 mg dipyrone), produced by Polfa Pabianice, series No. 10107.
- Grypostop—tablets (325 mg acetaminophen, 30 mg pseudoephedrine hydrochloride, 15 mg dextrometorphan hydrobromide), produced by PERRIGO, series No. 5K1591.
- Ibuprom—tablets (200 mg ibuprofen), produced by US Pharmacia, series No. 1687707.
- Modafen—tablets (200 mg ibuprofen, 30 mg pseudoephedrine hydrochloride), produced by ZENTIVA, series No. 3821204.
- Neopyrin ASA—tablets (300 mg aspirin, 100 mg acetaminophen, 50 mg caffeine), produced by BIOFARM, series No. 020706.
- Nurofen Plus—tablets (200 mg ibuprofen, 12.8 mg codeine phosphate), produced by Boots Healthcare, series No. 87J.
- Panadol Extra—tablets (500 mg acetaminophen, 65 mg caffeine), produced by GlaxoSmithKline, series No. 060774.
- Saridon—tablets (250 mg acetaminophen, 150 mg prophyphenazone, 50 mg caffeine), produced by Roche, series No. L2F368.

Table 1
The concentrations (mg/L) of compounds in binary calibration series

Solution	Compound 1	Compound 2
1	1.0	2.0
2	7.0	7.0
3	4.0	1.0
4	2.0	3.0
5	1.0	5.0
6	3.0	4.0
7	6.0	2.0
8	8.0	1.0
9	2.0	8.0
10	5.0	6.0
11	4.5	1.5
12	4.5	5.0
13	3.5	3.5
14	1.5	6.5
15	4.0	5.0

- Solpadeine—tablets (500 mg acetaminophen, 30 mg caffeine, 8 mg codeine phosphate), produced by SmithKline Beecham, series No. 060925.
- Solpadeine—effervescent tablets (500 mg acetaminophen, 30 mg caffeine, 8 mg codeine phosphate), produced by GlaxoSmithKline, series No. 065168.
- Solpadeine—capsules (500 mg acetaminophen, 30 mg caffeine, 8 mg codeine phosphate), produced by SmithKline Beecham.
- Talvosilen forte—capsules (500 mg acetaminophen, 30 mg codeine phosphate) produced by Bene-Arzneimittel GmbH, series No. 652105.

3.1. Calibration dataset

The calibration set consisted of binary and ternary methanolic mixtures in concentration not exceeding 14 mg/L of the sum of the ingredients. Due to extremely large number of possible combinations only the sets present together in one pharmaceutical formulation were used (see Section 4). We used 15 solutions for each of

Table 2
The concentrations (mg/L) of compounds in ternary calibration series

Solution	Compound 1	Compound 2	Compound 3
1	2.0	3.0	4.0
2	1.0	4.0	2.0
3	4.0	1.0	3.0
4	3.0	2.0	1.0
5	1.5	3.5	3.0
6	3.0	2.0	3.5
7	2.5	2.5	2.5
8	2.0	3.0	1.5
9	3.5	1.5	2.0

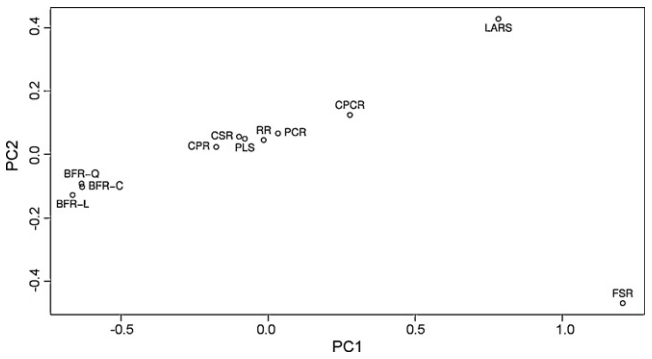

Fig. 2. The similarities between RMSEP values obtained with different techniques presented by Principal Component Analysis.

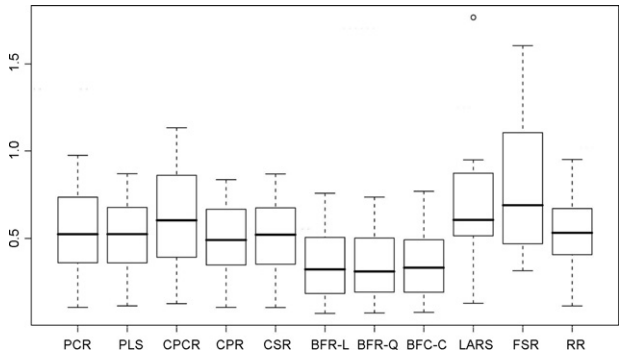

Fig. 3. The boxplot of RMSEP values (mg/L) obtained within different techniques.

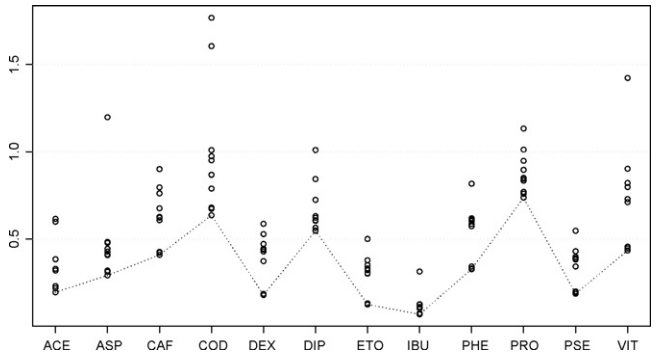

Fig. 4. The stripchart of RMSEP values (mg/L) for 12 substances. The lowest values (marked by line) are obtained using different variants of BFR.

the following combinations (concentrations given in Table 1):

(1) aspirin and vitamin C;
(2) ibuprofen and codeine phosphate;
(3) ibuprofen and pseudoephedrine hydrochloride;
(4) aspirin and caffeine;
(5) acetaminophen and codeine phosphate;
(6) acetaminophen and codeine;
(7) propyphenazone and dipyrone;

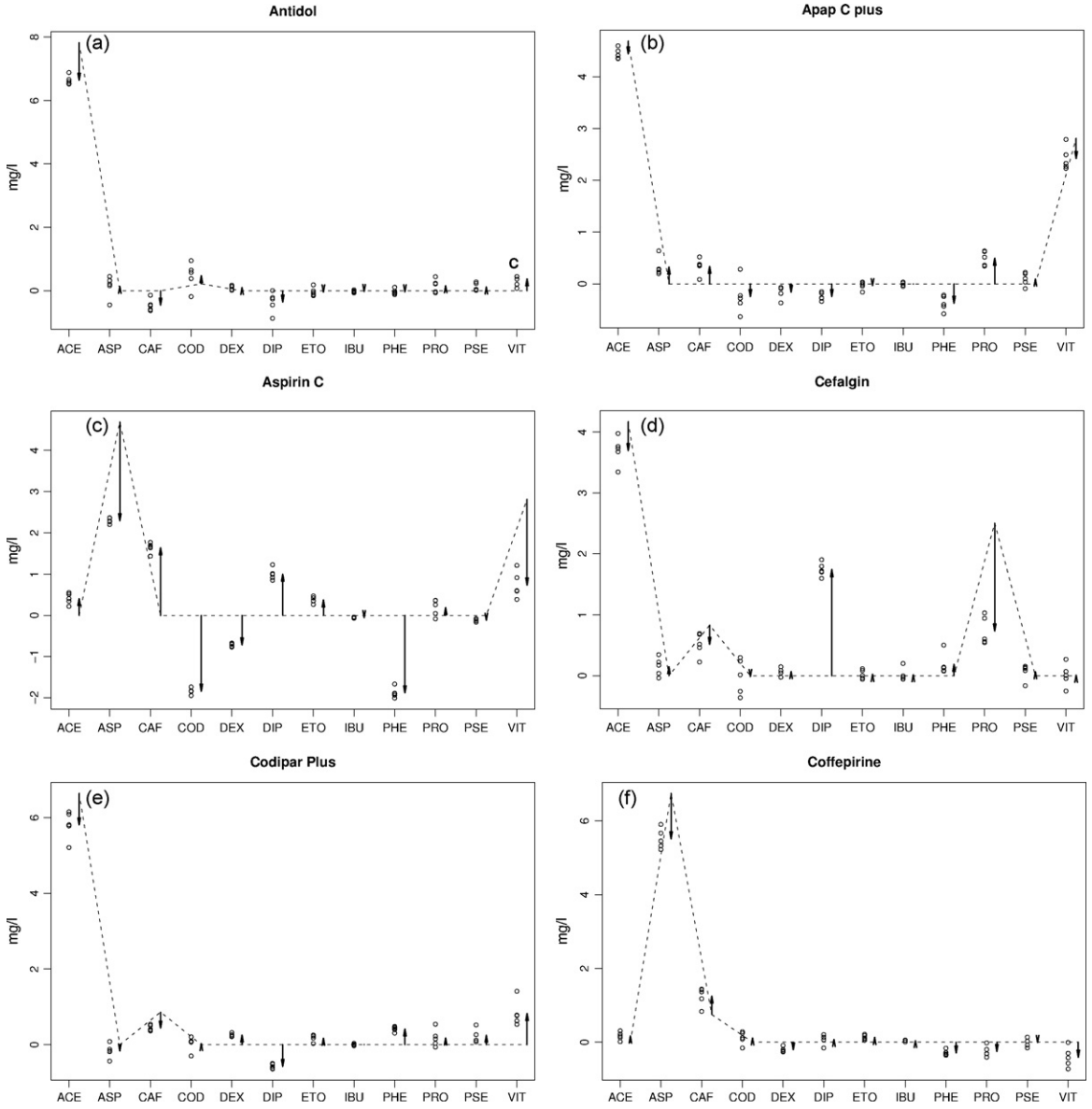

Fig. 5. Concentrations of pharmaceutical formulations ingredients (names of formulations above plots), calculated by BFR method. Arrows indicate error of the result against expected value (expected values are connected by dashed line). Part 1

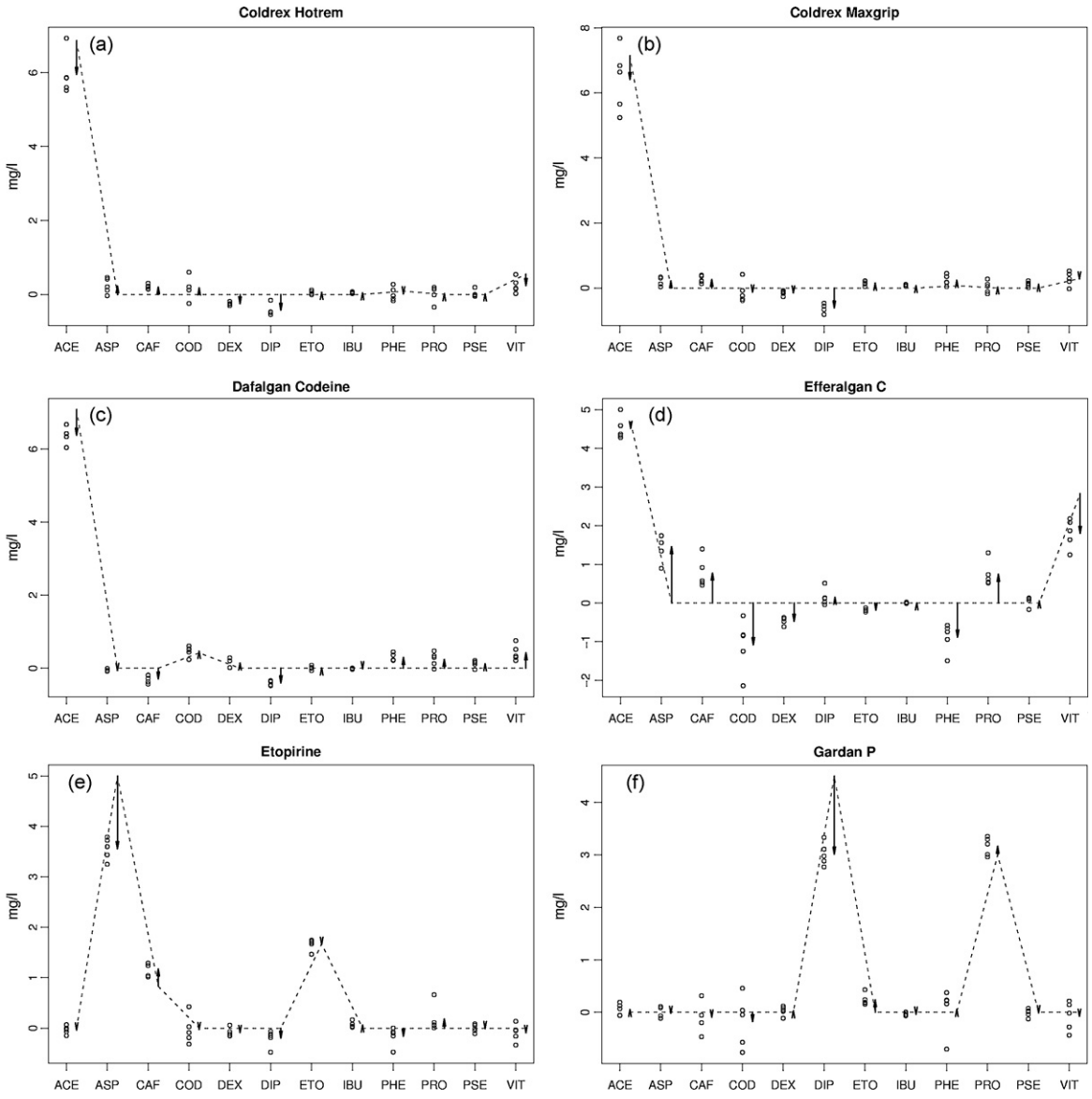

Fig. 6. Concentrations of pharmaceutical formulations ingredients (names of formulations above plots), calculated by BFR method. Arrows indicate error of the result against expected value (expected values are connected by dashed line). Part 2

and 9 solutions of each ternary mixture (concentrations given in Table 2):

(1) aspirin, ethoxybenzamide and caffeine;
(2) aspirin, acetaminophen and caffeine;
(3) acetaminophen, phenylephrine hydrochloride and vitamin C;
(4) acetaminophen, codeine and caffeine;
(5) acetaminophen, propyphenazone and caffeine;
(6) acetaminophen, pseudoephedrine and dextrometorphan hydrobromide;

which resulted in $7 \times 15 + 6 \times 9 = 159$ total calibration spectra.

3.2. Data processing

The calibration matrix was used to obtain concentration prediction model for each of the 12 substances. The models were cross-validated by 10-fold validation against the RMSEP (mg/L) value. The BFR technique was validated using linear (BFR-L), quadratic (BFR-Q) and cubical (BFR-C) splines with 1–50 knots. The following techniques were also calculated and validated for comparison:

- Partial Least Squares (PLS), Principal Component Regression (PCR, CPCR)—1–50 components.
- Cyclic Subspace Regression (CSR)—variable l in range 1–50, for its each value variable j was in range $1 - l$ (1275 combinations).
- Continuum Power Regression (CPR) 1–50 factors, for each factor number continuum parameter was within range 0.1–1 in step 0.05 (19 values, 950 total combinations).
- Forward Stepwise Regression (FSR), Least Angle Regression (LARS)—1–50 chosen variables.
- Ridge Regression—the small constant k in range 0.01–0.3 with step 0.01 (29 values).

The model building and validation was performed under the GNU Octave numerical environment. For the BFR technique, the

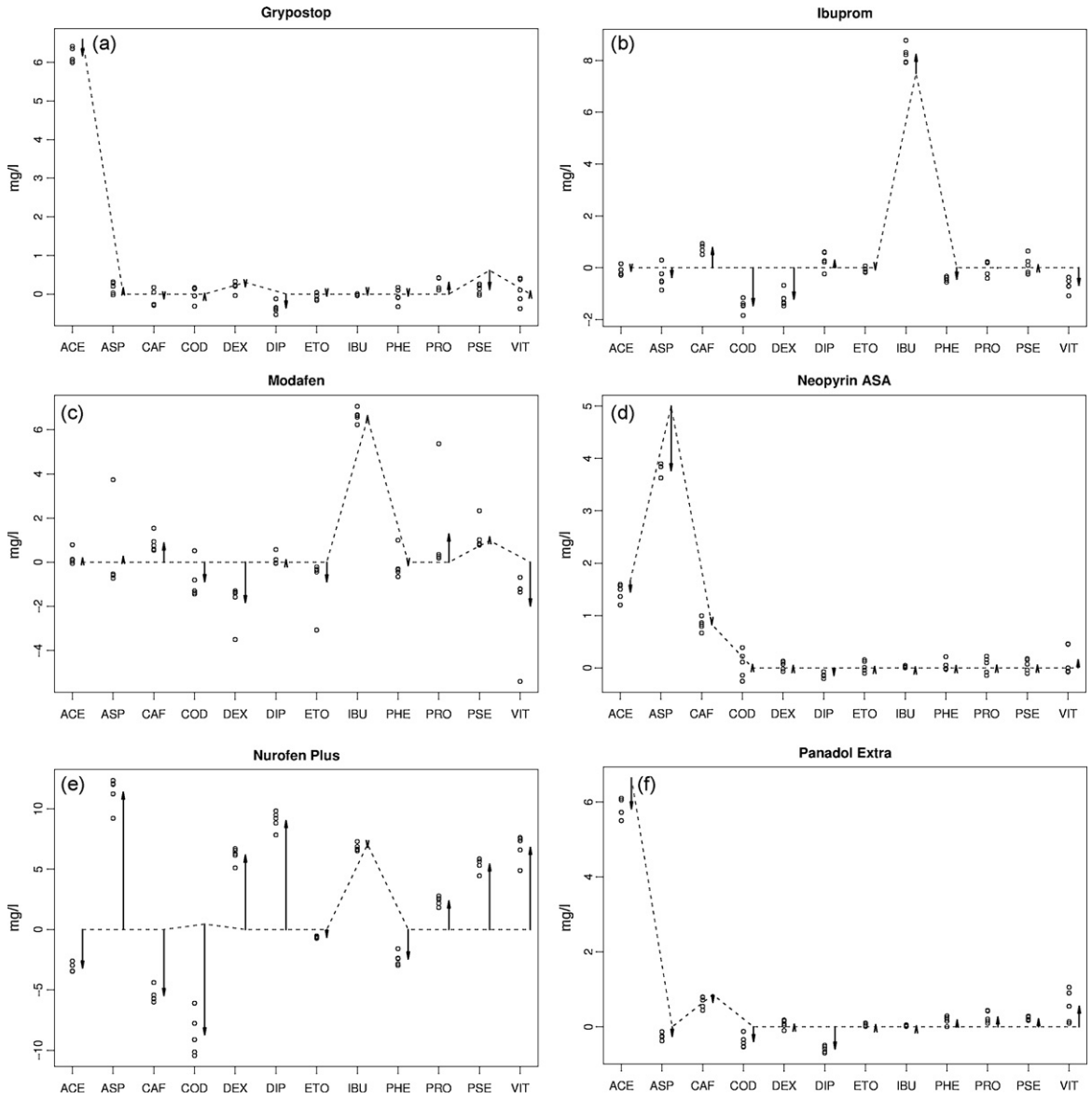

Fig. 7. Concentrations of pharmaceutical formulations ingredients (names of formulations above plots), calculated by BFR method. Arrows indicate error of the result against expected value (expected values are connected by dashed line). Part 3

code presented in this paper was used. For the other techniques we used the functions given by Rasmussen [16], Sjöstrand [19] and Daszykowski et al. (TOMCAT Toolbox for MATLAB [20]). The crossvalidation of all techniques lasted about 100 h of continuous running on Intel Dual Core Pentium 2 GHz with 2 GB RAM.

4. Results and discussion

The recognizing of single, binary or ternary combination of 12 possible compounds is quite a difficult task for a linear chemometric model. The spectra of several drugs are very similar (when investigating similarity by clustering technique, strong similarity is observed for propyphenazone and dipyrone or caffeine and ethoxybenzamide). There is a risk that the compound will be recognized as other similar compound, when spectra are mixed and additionally any small possible matrix effect is present. Although the performance of such large model cannot be high enough to use it in pharmaceutical control, it can be used as a tool for content recognition and estimation of the concentration, for example, in forensic toxicology.

The classical mixture design could not be used directly in our case due to following reasons: the task for chemometric model here is not to estimate proportions of ingredients of a mixture, but to predict absolute concentrations of possible ingredients. The tablet sample consists of active compounds (absorbing UV significantly) and other excipients (such as lactose, talc, etc.) which absorb UV light poorly and do not interfere in practice with active ingredients. Moreover, the sum of active ingredients in calibration samples should not exceed 14 mg/mL, because absorption bands would then exceed the scale of spectrophotometer. The calibration set must then contain compounds with different proportions and also different absolute concentrations. We have obtained the concentrations meeting these criteria by generating random sets until these criteria are met.

There is a very large number of possible combinations of active ingredients. We have decided to use only binary and ternary mixtures of compounds present together in pharmaceutical formu-

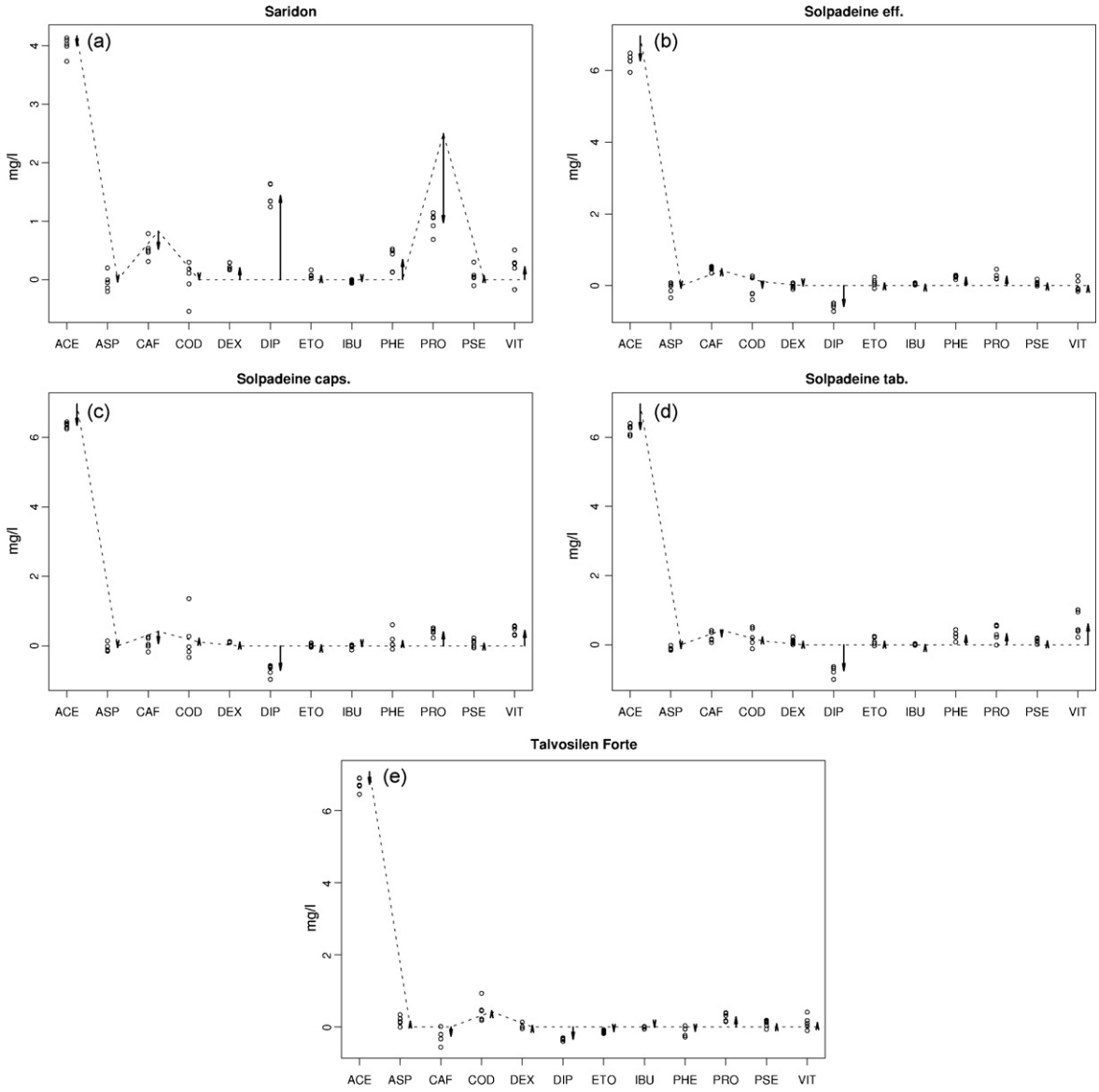

Fig. 8. Concentrations of pharmaceutical formulations ingredients (names of formulations above plots), calculated by BFR method. Arrows indicate error of the result against expected value (expected values are connected by dashed line). Part 4

lations. Such approach do not exhaust all possible combinations in context of DoE, but was found to be sufficient in our case. Moreover, the individual binary and ternary sets can be further used as more specialized calibration set for quantitative determination of particular drug, which is very important and costs little in terms of time and the quantity of reagents consumed.

From each technique, the optimal parameter set (number of factors, continuum, constant, etc.) was chosen to minimize RMSEP value. The minimal RMSEP values obtained with each technique were collected for each of the 12 compounds.

Fig. 2 show the multivariate similarity between RMSEP values (two first principal components). The BFR techniques are clustered together and clearly separated from the others. Fig. 3 shows the boxplot of RMSEP values (12 values, 1 for each compound) grouped by techniques. It is clearly seen, that BFR variants (BFR-L, BFR-Q, BFR-C) present visibly lower prediction error. For each drug (Fig. 4, errors grouped by substance) lowest RMSEP values (connected by line) were always obtained with one of BFR variants, where the optimal number of knots was in range 14–39 depending on a variant and a compound.

We have decided to use BFR for building final models and tested the models by content estimation of 23 multicomponent pharmaceuticals available on the Polish market. The grounded tablet, capsule content of powder amount equivalent to final concentration approximately 7.5 mg/L of sum of ingredients were placed in volumetric flask, dissolved in methanol, ultrasonicated (15 min) and then filtered. The spectra was measured in the same way and used for prediction.

Most of the formulations (16 of total 23, see Figs. 5–8) were recognized very well and small errors are fully acceptable for semi-quantitative content recognition. The following formulations were recognized with not acceptable error:

(1) Aspirin C and Efferalgan C—probably due to the same interfering effervescent excipients in tablets.
(2) Cefalgin and Saridon—due to strong similarity of the dipyrone and propyphenazone spectra (propyphenazone was recognized as dipyrone).
(3) Ibuprom, Modafen—due to relatively low absorbance of iburofen at 7.5 mg/L concentration.

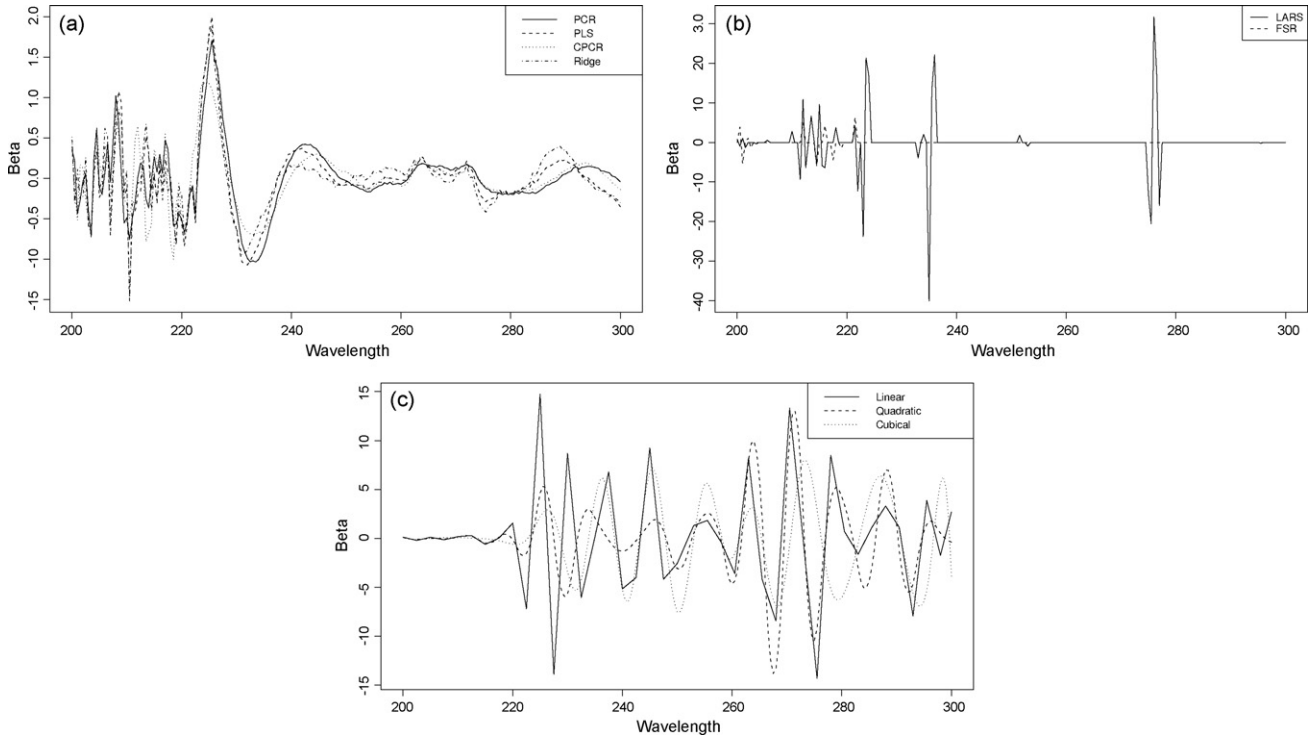

Fig. 9. The estimator for ibuprofen in function of wavelength (nm) for (a) PCR, PLS, CPCR and Ridge; (b) FSR and LARS; (c) BFR.

(4) Nurofen plus—due to relatively low absorbance of iburofen at 7.5 mg/L concentration and additional interfering excipients.

Comparing the estimator as function of wavelengths, we present an example of ibuprofen in Fig. 9, but for all the other substances the conclusions are similar. We see that PCR, PLS, CPCR and Ridge Regression produce very similar estimators. They find good and collinear dependence in 220–300 nm (where ibuprofen differs from other substances), but also "take care" on 200–220 nm non-specific region, producing the chaotic estimator values here. The FSR technique locates selected variables almost only at inspecific region (and therefore has worst predictive ability), LARS founds more important wavelengths at appropriate region, still focusing at 200–220 nm. The BFR techniques almost completely reject the inspecific region (the chaotic values here ale "smoothed" and averaged to zero), and all found dependences are located in the appropriate analytical region. This seems to be a reason that BFR works better in our case. It is clearly seen from this Figure that the BFR technique can be treated as spline-smoothing method of the estimator function, where the knots are located by the algorithm, and between the knots the function is smoothly interpolated.

5. Conclusion

The Basis Function Regression is an interesting alternative to PLS and PCR. It can (and even should) be always considered during predicting any property of the sample from the spectral data as comparative method, both in semiquantitative and quantitative analysis. It can show better performance than PCR and PLS in certain cases, especially when the considered ingredients of the sample have strong similarity of their individual spectra. The presented MATLAB routines allow any chemometrician to use it in everyday practice.

Acknowledgement

We gratefully acknowledge obtaining the MATLAB code snippet for spline basis function generation from Dr. Graeme A. Chandler, Department of Mathematics, The University of Queensland, Australia.

References

[1] J. Kalivas, Chemom. Intell. Lab. Syst. 37 (1997) 255–259.
[2] R. Sundberg, Scand. J. Stat. 26 (1999) 161–191.
[3] M. Stone, J. R. Stat. Soc. B 36 (1974) 111–147.
[4] A.E. Hoerl, R.W. Kennard, R.W. Hoerl, J. R. Stat. Soc. C 34 (1985) 114–120.
[5] J. Edwards, P. Oman, R. News, 3 (2003) 2–7, www.r-project.org.
[6] Y.L. Xie, J. Kalivas, Anal. Chim. Acta 348 (1997) 19–27.
[7] J. Hinkle, W. Rayens, Chemom. Intell. Lab. Syst. 30 (1995) 159–172.
[8] J. Kalivas, Chemom. Intell. Lab. Syst. 45 (1999) 215–224.
[9] S. De Jong, R. Farebrother, Chemom. Intell. Lab. Syst. 25 (1994) 179–181.
[10] L. Breiman, Ann. Stat. 24 (1996) 2350–2383.
[11] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, H. Ishwaran, K. Knight, J.M. Loubes, P. Massart, D. Madigan, G. Ridgeway, S. Rosset, J. Zhu, Ann. Stat. 32 (2004) 407–499.
[12] T.J. Hastie, C. Mallows, Technometrics 35 (1993) 140–143.
[13] C. Goutis, J. R. Stat. Soc. B 60 (1998) 103–114.
[14] B.D. Marx, P.H.C. Eilers, Technometrics 41 (1999) 1–13.
[15] F. Rossi, D. Francois, V. Wertz, M. Meurens, M. Verleysen, Chemom. Intell. Lab. Syst. 86 (2007) 208–218.
[16] M. Rasmussen, Generalized methods for calibration, MSc Thesis, http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=729.
[17] C. de Boor, A Practical Guide to Splines, Springer Verlag, 1978.
[18] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2005, ISBN 3-900051-07-0.
[19] K. Sjöstrand, Matlab implementation of LASSO, LARS, the elastic net and SPCA, 2005, http://www2.imm.dtu.dk/pubdb/p.php?3897.
[20] M. Daszykowski, S. Serneels, K. Kaczmarek, P. Van Espen, C. Croux, B. Walczak, Chemom. Intell. Lab. Syst. 85 (2007) 269–277.