Maximum A Posteriori Bayesian Estimation of Chromatographic Parameters by Limited Number of Experiments

ARTICLE *in* ANALYTICAL CHEMISTRY · JUNE 2015

Impact Factor: 5.64 · DOI: 10.1021/acs.analchem.5b01195 · Source: PubMed

3 AUTHORS, INCLUDING:

Paweł Wiczling
Medical University of Gdansk
52 PUBLICATIONS **754** CITATIONS
SEE PROFILE

Roman Kaliszan
Medical University of Gdansk
141 PUBLICATIONS **3,431** CITATIONS
SEE PROFILE

analytical chemistry

Article

pubs.acs.org/ac

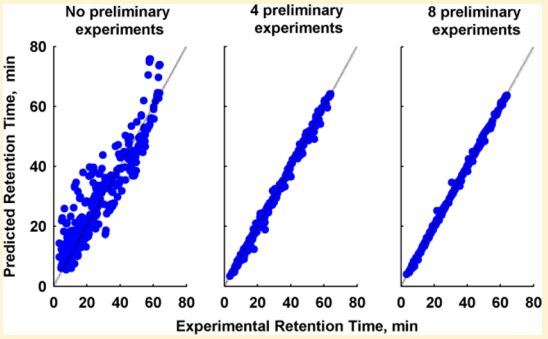
Maximum *A Posteriori* Bayesian Estimation of Chromatographic Parameters by Limited Number of Experiments

Paweł Wiczling,* Łukasz Kubik, and Roman Kaliszan

Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gen. J. Hallera 107, 80-416 Gdańsk, Poland

Supporting Information

ABSTRACT: The aim of this work was to develop a nonlinear mixed-effect chromatographic model able to describe the retention times of weak acids and bases in all possible combinations of organic modifier content and mobile-phase pH. Further, we aimed to identify the influence of basic covariates, like lipophilicity (log *P*), dissociation constant (pK_a), and polar surface area (PSA), on the intercompound variability of chromatographic parameters. Lastly, we aimed to propose the optimal limited experimental design to the estimation process of parameters through a maximum *a posteriori* (MAP) Bayesian method to facilitate the method development process. The data set comprised retention times for two series of organic modifier content collected at different pH for a large series of acids and bases. The obtained typical parameters and their distribution were subsequently used as priors to improve the estimation process from reduced design with a variable number of preliminary experiments. The MAP Bayesian estimator was validated using two external-validation data sets. The common literature model was used to relate analyte retention time with mobile-phase pH and organic modifier content. A set of QSRR-based covariate relationships was established. It turned out that four preliminary experiments and prior information that includes analyte pK_a, log *P*, acid/base type, and PSA are sufficient to accurately predict analyte retention in virtually all combined changes of pH and organic modifier content. The MAP Bayesian estimator of all important chromatographic parameters controlling retention in pH/organic modifier gradient was developed. It can be used to improve parameter estimation using limited experimental design.



The application of mathematical models of reversed-phase high-performance liquid chromatography (RP HPLC) for prediction of analyte retention times can considerably facilitate the process of finding an optimal separation. Nowadays, there is a great number of such models in the literature. They allow researchers to predict analytes' retention time for various organic modifier contents, pHs, temperatures, and flow rates in either isocratic and gradient conditions.[1] Nevertheless, all those models contain a certain number of adjustable parameters that need to be estimated or guessed prior to any practical application. There are two major ways of finding all analyte and column-specific parameters: (1) by using Quantitative Structure Retention Relationships (QSRR), that is, linking analyte properties, like lipophilicity, with its retention factor;[2] or (2) by using a set of preliminary experiments. The latter is a well-established technique in chromatography, already implemented in commercial software.[3] In routine practice, a trial-and-error approach that combines both those scenarios is sometimes utilized. It is based on the general expertise of the chromatographer, literature, and some preliminary chromatographic experiments to guide the further steps in a search for the best solution. Literature very often provides substantial information on an analyte (e.g., in the form of its pK_a value and lipophilicity, log *P*). If those measures are unavailable from direct measurements, one can easily obtain them from analyte structure by using free or commercial software.[4] Generally, numerous research has been done to relate analyte retention time to its structure, with the hope of eventually being able to predict retention and separation in the absence of any prior experiments. In general, it has not proved possible to predict retention times in HPLC with a sufficient accuracy to support method development, however.[5] Thus, the approach that combines both analyte structure and preliminary experiments might be of help to overcome the limitations of techniques that are based solely on QSRR equation or preliminary experiments.

In this work, a nonlinear mixed-effect modeling is proposed to simultaneously analyze a set of analytes and to determine one single mathematical model describing retention of a "population" of analytes. This method is widely used in biometrical studies, especially in pharmacokinetic research,[6] to account for the various sources of variability, and to characterize a large set of data from different sources and experimental designs. It also allows researchers to seek for specific relationships of covariates to relate various known physicochemical properties of analyte to the chromatographically specific parameters through the QSRR-based equations. The determination of the parameter distribution

Received: March 30, 2015
Accepted: June 22, 2015
Published: June 22, 2015

DOI: 10.1021/acs.analchem.5b01195
Anal. Chem. 2015, 87, 7241−7249

characterizing the whole "population" of analytes provides a possibility to use a maximum *a posteriori* (MAP) Bayesian of parameter estimation from the limited set of chromatographic experiments to obtain the most likely estimates of parameters for the specific analyte (and uncertainty around these estimates). In this technique, the estimation of parameters is based on the weighted influence of experimental data and prior information. The more individual data is provided (more experiments are conducted), the less is the reliance on the "population" typical data.

The application of the MAP Bayesian estimation will be illustrated in the example of predicting analyte HPLC retention time for various pH and organic modifier content changes. It is a challenging problem due to the complex relationship linking retention time and mobile-phase properties, especially for gradient mode. Usually, a large number of experiments is required to precisely determine all important parameters controlling analyte retention, like hydrophobicity of ionized and nonionized form of analyte, dissociation constant, pK_a, and the slope determining the changes of retention factor and pK_a with content of organic modifier in the eluent.[7] Also, a simplified method of prediction of the chromatographic retention of acid−base compounds was recently proposed.[8] Thus, the experimental designs that would help to reduce the time and costs of analysis are of great importance. The MAP Bayesian estimation of chromatographic parameters from a series of preliminary experiments and prior knowledge seems to be particularly appealing for that purpose.

THEORETICAL

Let i denote the i^{th} compound (i = 1,..., N) and j the j^{th} chromatographic retention time for a compound (j = 1,..., n_i), where n_i is the number of observations for analyte i. Let $t_{R,ij}$ = $\{t_{Ri1},..., t_{R,ini}\}$ be the n_i-vector of measurement performed for compound i. Let the function f denotes the nonlinear structural model relating retention $t_{R,ij}$ with analyte properties and certain experimental design, D_{ij}. The statistical model for the observation $t_{R,ij}$ in compound i under the design D_{ij}, is given by

$$t_{R,ij} = f(D_{ij}, R_i) + \varepsilon_{ij} \quad (1)$$

where R_i is the p-vector of the individual parameters (it includes log k_w, pK_a, S, and others), D_{ij} is a vector of all adjustable system parameters influencing analyte retention (i.e., organic modifier content, pH, flow rate, etc.), and ε_{ij} is the residual error, which is assumed to be normal, with zero mean. The variance of ε_{ij} may depend on the predicted retention times $f(D_{ij}, R_i)$ through a (known) variance model. In this work, the following error model is assumed:

$$\mathrm{var}(\varepsilon_{ij}) = (\sigma_{\mathrm{add}} + \sigma_{\mathrm{prop}} f(D_{ij}, R_i))^2 \quad (2)$$

where σ_{add} and σ_{prop} are an additive and a proportional component of unexplained residual variability, respectively.

Another usual assumption in nonlinear mixed-effect model is that the distribution of the individual parameters R_i follows a normal distribution:

$$R_i = h(\theta, X_i) + \eta_{R,i} \quad (3)$$

where $\boldsymbol{\theta}$ is the population vector of parameters, $\boldsymbol{X}_i$ a vector of covariates, h is a function giving the expected value of the parameters depending on the covariates, and $\eta_{R,i}$ represents the vector of random effects for compound i. $\eta_{R,i}$ is assumed to follow a normal distribution N (0,Ω), where Ω is the variance-covariance matrix of the random effects. An example of such a relationship might be given by the following QSRR equation:

$$\log k_{w,i} = \theta_{\log k_w} + \theta_{\log k_w - \log P} \log P_i + \eta_{\log k_w,i} \quad (4)$$

which assumes a linear QSRR relationship between analyte lipophilicity and retention factor extrapolated to neat water as an eluent.[9] $\theta_{\log k_w}$ and $\theta_{\log k_w - \log P}$ are the intercept and the slope of that relationship. $\eta_{\log k_w,i}$ is a deviation of the individual estimate of log k_w for i^{th} analyte from the population mean. $\eta_{\log k_w,i}$ can be also understood as a residual error for that particular QSRR relationship.

The structural model (f) used in this work assumes the three-parameter relationship between retention factor and organic modifier content for both ionized and nonionized form of an analyte,[10] a linear dependence of pK_a with organic modifier content, and a sigmoidal dependence of retention factor with mobile-phase pH. It allows us to model the retention time as a function of pH and organic modifier content for isocratic and gradient conditions:[11,12]

$$\text{Acids: } \int_0^{t_R - t_0} \frac{1}{t_0} \frac{1 + 10^{{}^s_s pH(t) - pK_a(\varphi(t))}}{10^{\log k_{w,N} - \frac{S_{1,N}\varphi(t)}{1+S_2\varphi(t)}} + 10^{\log k_{w,I} - f_A(pH(t)) - \frac{S_{1,I}\varphi(t)}{1+S_2\varphi(t)}} 10^{{}^s_s pH(t) - pK_a(\varphi(t))}} dt = 1 \quad (5)$$

$$\text{Bases: } \int_0^{t_R - t_0} \frac{1}{t_0} \frac{1 + 10^{pK_a(\varphi(t)) - {}^s_s pH(t)}}{10^{\log k_{w,N} - \frac{S_{1,N}\varphi(t)}{1+S_2\varphi(t)}} + 10^{\log k_{w,I} - \frac{S_{1,I}\varphi(t)}{1+S_2\varphi(t)}} 10^{pK_a(\varphi(t)) - {}^s_s pH(t)}} dt = 1 \quad (6)$$

where ${}^s_s$pH and $\varphi(t)$ are functions describing changes of mobile-phase pH and its composition at column inlet during a chromatographic run; log $k_{w,N}$ and log $k_{w,I}$ represent retention factors for neat water as the mobile phase of the individual ionized (N) and un-ionized (I) form of the analyte; $S_{1,N}$, $S_{1,I}$, S_2 are parameters showing how rapidly retention factor is changing with changes in organic modifier content also for neutral and ionized form of analyte; t_0 denotes column hold-up time, p$K_a(\varphi(t))$ denotes the pK_a value of an analyte for different organic modifier contents. In this work, a linear relationship between pK_a and φ is assumed. The presence of silanols is a known factor affecting analytes retention.[13] In this work it was evident for acids, thus the following empirical function was used to accounts for this effect:

$$f_A(\mathrm{pH}(t)) = 1 + a(\mathrm{pH}(t) - 7) \quad (7)$$

where a denotes a fractional increase in log $k_{w,I}$ for acids per unit increase in pH.

MATERIALS AND HPLC EQUIPMENT

All experiments were done using a Merck-Hitachi LaChrome (Darmstadt, Germany-San Jose, CA, U.S.A.) apparatus equipped with a diode array detector, autosampler and thermostat. Chromatographic data were collected using a D-7000 HPLC System Manager, version 3.1 (Merck-Hitachi). An Xterra MS C-18 column, 150 × 4.6 mm I.D., particle size 5 μm (Waters Corporation, Milford, MA, U.S.A.) was used. 1% urea was injected to determine the column dead volume, V_0, which was 1.44 mL. The system dwell volume V_d equaled 1.74 mL. The extra column volume equaled 0.59 mL. The extra column time was subtracted from all the measured retention times data prior to any calculation. The chromatographic measurements were done at 25 °C with flow rate of 1 mL/min.

A universal buffer was used to control the pH of the mobile phase. The base buffer solution was formed using three compounds: citric acid (CIT), tris(hydroxymethyl)-amino-methane (TRIS), glycine (GLY), each at a concentration of

DOI: 10.1021/acs.analchem.5b01195
Anal. Chem. 2015, 87, 7241−7249

0.008 M. The necessary volume of 3 M KOH and 1 M HCl was added to the base solution to get the buffer of a high (solvent D) and a low (solvent C) pH. The mobile phases contained buffers D and C in different proportions and methanol as the organic modifier (solvent B). Prior to any experiments, ${}^{s}_{w}$pH, value has been determined for a multiple combinations of B, C, and D. The linear interpolation served to find the pH value for a desired combination of the mobile-phase components. The ${}^{s}_{w}$pH was converted to the ${}^{s}_{s}$pH before use in calculations.[14] The ${}^{s}_{w}$pH measurements were done at 25 °C with an HI 9017 pH meter (Hanna Instruments, Bedfordshire, U.K.). The left side symbols denote the pH scale as described elsewhere.[15]

EXPERIMENTAL DATA

In this work, we reanalyzed previously collected data.[11,16] The first data set comprised 18 RP HPLC runs for 93 monoprotic acids and bases. The experiments differed in gradient duration, which equaled 20 min for series I and 60 min for series II. Each series differed in pH of the mobile phase spanning the range from 2.5 to 10.5, as indicated in Table 1. The second data set comprised a series of 38 isocratic, 30 methanol gradients, and 25 pH-gradient retention time measurements for a weak acid (ketoprofen) and a weak base (papaverine). The isocratic measurements were conducted for six different pHs of the mobile phase and methanol contents ranging from 20 to 80% (v/v). Organic modifier gradients were conducted as a wide methanol gradients ranging from 5 to 80%, developed for varying gradient durations of 10, 20, 40, 60, and 90 min and for six different pHs of the mobile phase. pH gradients comprised a series of gradients with increasing pH for acids (from 2.5 to 10.5) and decreasing pH for bases (from 10.5 to 2.5), developed at different gradient durations (from 4 to 90 min) and for three different methanol contents of 30, 40, and 50%.

Table 1. Parameters of Chromatographic Run for the First Data Set

run number	average pH	gradient duration, min	organic modifier content
1	3.45	20	5%-80%
2	4.21	20	5%-80%
3	5.04	20	5%-80%
4	5.92	20	5%-80%
5	6.80	20	5%-80%
6	7.89	20	5%-80%
7	8.77	20	5%-80%
8	9.66	20	5%-80%
9	10.53	20	5%-80%
10	3.45	60	5%-80%
11	4.21	60	5%-80%
12	5.04	60	5%-80%
13	5.92	60	5%-80%
14	6.80	60	5%-80%
15	7.89	60	5%-80%
16	8.77	60	5%-80%
17	9.66	60	5%-80%
18	10.53	60	5%-80%

NONLINEAR MIXED-EFFECTS MODELING

Nonlinear mixed-effects (NLME) modeling was done in MONOLIX software. It implements a stochastic approximation of the standard expectation maximization (SAEM) algorithm without approximations.[17] All data processing and MONOLIX operations were done with Matlab Software version 2014b. The integration of eqs 5 and 6 was done numerically by means of the matlab *trapz* function. The minimum value of the objective function (OFV), typical goodness-of-fit diagnostic plots, the evaluation of the precision of parameters and variability estimates, and various internal evaluations (like Visual Predictive Check, Numerical Prediction Distribution Errors) were used to discriminate between models during the model-building process.[18]

The log P, pK_a, acid/base type, polar surface area (PSA), number of hydrogen donor and acceptor groups, and molecular mass were investigated as potential covariates on the chromatographic parameters. Literature values of log P and aqueous pK_a were used. PSA, acid/base type, and number of hydrogen bonds donors and acceptors, and molecular mass were obtained from the Advanced Chemistry Development, Inc. (ACD/Laboratories) software.

The systematic covariate analysis process and likelihood ratio test was used to test the effect of each variable. The selection of variables was determined using a forward and backward selection process. During forward selection, a covariate was selected only if a significant ($P < 0.05$, χ^2 distribution with one degree of freedom) decrease (reduction >3.84) in the objective function value (OFV) from the basic model was obtained. Then all the variables found to be significant were added simultaneously into a "full" model. The importance of each variable was re-evaluated by backward selection. Each variable was independently removed from the full model to confirm its relevance. An increase in the OFV of more than 6.635 ($P < 0.01$, χ^2 distribution) was required for confirmation. The resulting model which included all significant variables was called the "final".

MAP BAYESIAN ESTIMATOR

The nonlinear data-fitting problem for a single analyte was solved by maximum *a posteriori* approach. The idea behind the estimation method is to determine the individual parameters for each compound (vector $\boldsymbol{\mu}$) that maximize the penalized version of the maximum likelihood of the sample data.[19] The objective function (negative log likelihood) is given by

$$\Phi(\boldsymbol{\mu}) = \sum_{j=1}^{n}\left[\frac{1}{2}\frac{(t_{R,j} - f(D_j, \boldsymbol{\mu}))^2}{\mathrm{var}_j} + \log(\mathrm{var}_j) + (\boldsymbol{\mu} - \boldsymbol{\theta})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu} - \boldsymbol{\theta})\right] \quad (8)$$

Where the symbols have the following meanings: Φ − objective function corresponding to the Bayesian posterior distribution; n − number of experimental points; $t_{R,j}$ − retention time for experimental design D_j ; $f(D_j, \boldsymbol{\mu})$ − predicted retention time for the design D_j ; var_j − variance of measured retention times; $\boldsymbol{\mu}$ − vector of estimated model parameters; $\boldsymbol{\theta}$ − mean parameter values of the prior distribution; $\boldsymbol{\Omega}$ − variance-covariance matrix of parameters of the prior distribution; the symbol T denotes matrix transposition. The var_j denotes the variance of the residual error. It was modeled using the additive and a proportional error model as given by eq 2. The $\boldsymbol{\theta}$, $\boldsymbol{\Omega}$ and var_j were obtained from the final nonlinear mixed-effect model. The minimum of eq 8 was found using the Nelder−Mead Simplex Method implemented in the *fminsearch* function using Matlab Software version 2014b.

DOI: 10.1021/acs.analchem.5b01195
Anal. Chem. 2015, 87, 7241−7249

DATA ANALYSIS

The general overview of the used methodology is presented in the Supporting Information. The first data set (93 analytes) was randomly divided into two subsets: a training set (66 analytes) which was used to develop the nonlinear mixed-effect model and a test set (27 analytes) which provided a means to evaluate the induced model. The second data set (ketoprofen and papaverine) was used to evaluate the final model. The model evaluation was based on MAP Bayesian estimation of parameters and comparison of the estimated retention times with actual experimental data. The visual comparison and calculated median prediction error and median absolute prediction error were used. The prediction error (PE) was calculated for each measurement as PE = 100(measured − predicted)/predicted and was summarized as median for each individual compound. The median prediction error (MDPE) and median absolute prediction error (MDAPE) were calculated according to the formulas:

$$\mathrm{MDPE} = \mathrm{median}\ (\mathrm{PE}_1, \mathrm{PE}_2, \ldots \mathrm{PE}_N)$$

$$\mathrm{MDAPE} = \mathrm{median}\ (|\mathrm{PE}_1|, |\mathrm{PE}_2|, \ldots |\mathrm{PE}_N|) \quad (9)$$

where N denotes number of compounds. MDPE reflects the bias of the model, whereas MDAPE reflects the inaccuracy of the prediction.

RESULTS AND DISCUSSION

The reversed-phase high-performance liquid chromatography (RP HPLC), unlike the majority of chemical/biological tests, can provide a large number of reproducible compound property data (retention time) for a large set of analytes in a short period of time. Especially, if gradient technique is combined with mass spectrometry detection.[20] In this work, we reanalyzed such type of data using nonlinear mixed-effect modeling approach to obtain a universal model describing the behavior of monoprotic weak acids and bases for various pH and/or organic modifier content changes. The proposed model combines the structural, covariate, interindividual, and residual intracompound variability model. Figure 1 and Figure 2 show the typical goodness-of-fit plots of the final population model.

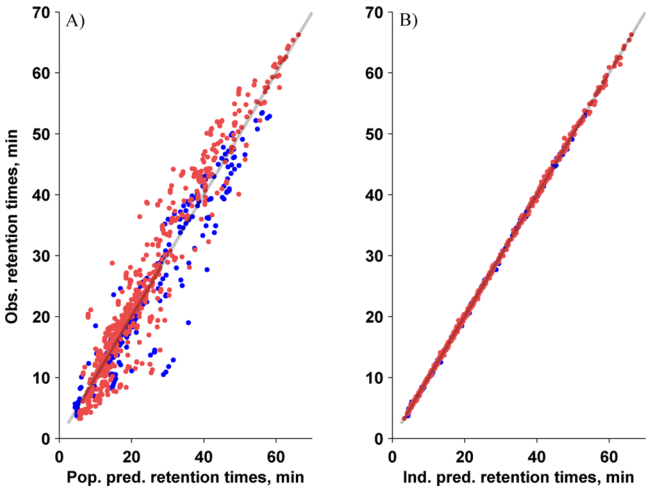


Figure 1. Goodness of fit plots of the final nonlinear mixed-effect. (A) The observed versus the population predicted retention times and (B) the observed versus the individual population predicted retention times. The red symbols denote bases and blue acids.

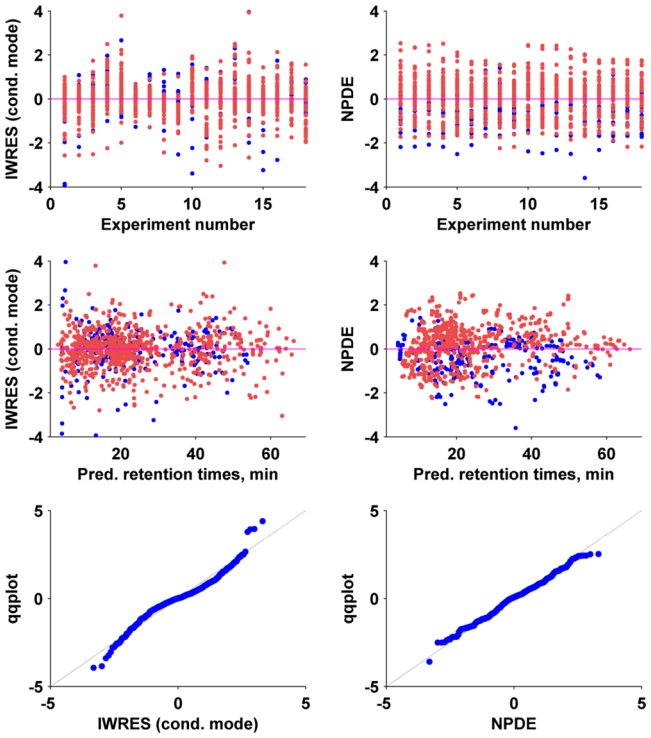


Figure 2. Individual weighted residuals (IWRES) and normalized prediction distribution errors (NPDE) of the final population model in relation to time, population predicted concentrations, and as qq plots. The red symbols denote bases and blue acids.

The individual and population predictions are very close to experimental data, indicating good performance of the model. It is also confirmed by the individual weighted residual (IWRES) and normalized prediction distribution errors (NPDE) graphs. The data points of IWRES and NPDE versus experiment number and model predicted retention times are evenly scattered around the horizontal 0 line. There is also a close proximity of IWRES and NPDE to the expected normal distribution (qq plot). Also, the plots of experimental data and model predictions for selected analytes, as shown in Figure 3, confirm good model performance in predicting analytes retention times.

The model parameter estimates are listed in Table 2. All parameters, intersubject and residual error variances were estimated precisely with low (lower than 50%) coefficients of variation (CV). The following QSRR bases relationships were identified during the model building process:

$$\log k_{\mathrm{w},N,i} = \theta_{\log k_\mathrm{w}} + \theta_{\log k_\mathrm{w}-\log P}\log P_i + \theta_{\log k_\mathrm{w}-\mathrm{PSA}}\mathrm{PSA}_i + \eta_{\log k_\mathrm{w},N,i} \quad (10)$$

$$S_{1,N,i} = \theta_{SN} + \theta_{SN-\log P}\log P_i + \theta_{\log SN-\mathrm{PSA}}\mathrm{PSA}_i + \eta_{SN,i} \quad (11)$$

$$\log k_{\mathrm{w},I,i} = \log k_{\mathrm{w},N} + \theta_{\Delta\log k_\mathrm{w}} + \eta_{\log k_\mathrm{w},I,i} \quad (12)$$

$$S_{1,I,i} = S_{1,N} + \theta_{\Delta S} + \theta_{\mathrm{AB}-\alpha}\mathrm{AB}_i + \eta_{SI,i} \quad (13)$$

$$\mathrm{p}K_\mathrm{a}(\varphi(t))_i = {}^{\mathrm{w}}\mathrm{p}K_{\mathrm{a},i} + (\theta_\alpha + \theta_{\mathrm{AB}-\alpha}\mathrm{AB}_i)\varphi(t) + \eta_{\mathrm{pKa},i} \quad (14)$$

where AB equals 0 for acids and 1 for bases. It has been found that log k_w is linearly related to drug lipophilicity (log P) and polar surface area (PSA). The same type of relationships was evident for $S_{1,N}$. The retention factor of the ionized form of an analyte, was lower by 1.06 compared to the neutral form. This

DOI: 10.1021/acs.analchem.5b01195
Anal. Chem. 2015, 87, 7241−7249

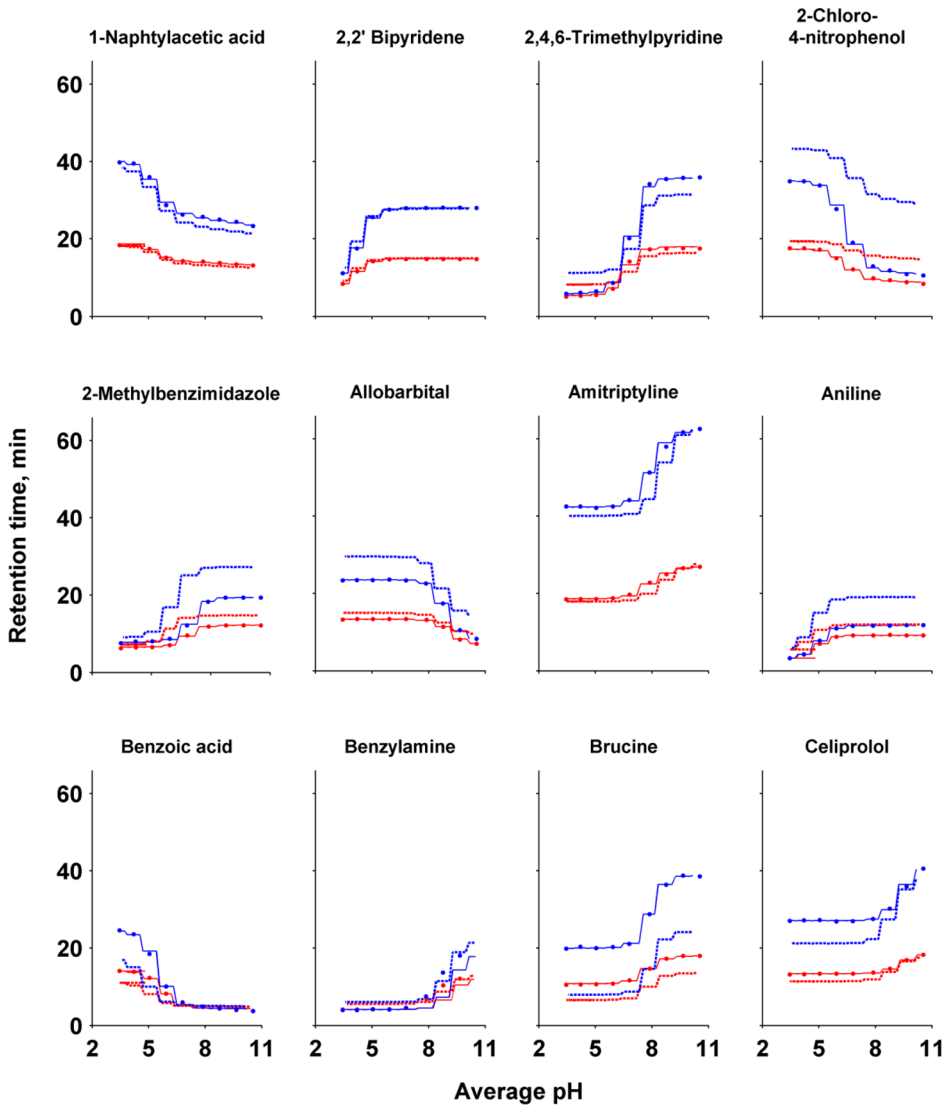


Figure 3. Plots of observed (point), population predicted (broken line), and individual predicted (solid line) retention times versus average pH during the experiment. Red color corresponds to 20 min gradient durations, whereas blue corresponds to 60 min gradient durations.

difference was found to be the same for acids and bases. The typical first slope parameter ($S_{1,I}$) of the ionized form of an analyte was higher for bases (1.01) and lower for acids (−0.831) compared to the nonionized form of an analyte. It indicates that the retention factor of ionized form increases more rapidly for acids than for bases with an increase in organic modifier content. The second slope parameter (S_2) equaled 0.183. The log k of ionized form of acids slightly decreased with increasing pH with the slope of −0.0172. $pK_a(\varphi(t))$ was proportional to the aqueous pK_a and dependent on organic modifier content with a slope α equaled 1.61 for acids and −0.365 for bases. It is consistent with the literature values, where generally high and positive slope is expected for acids, and small negative slope is expected for bases.[21] The ETAs in eq 10−14 denote a difference between an individual parameter estimate and the typical (mean) value of a parameter. In this manuscript, all ETAs were assumed to come from a multivariate normal distribution with variance-covariance matrix with diagonal elements (an independence of parameters from each other). All other elements of this matrix were zeros, except the ETA for log $k_{w,N}$ and S_N, for which a strong correlation (0.806) was evident.

The population (nonlinear mixed-effect) model is an excellent tool that can be used to obtain the typical (most likely) values of parameters along with uncertainty associate with those values. It also provides individual (analyte-specific) parameters that are estimated jointly for all compounds. It is a different method from the usual two-stage approach, in which individual parameters are estimated separately for each analyte before their further use in prediction of analyte retention or in search for QSRR relationships. The main advantage of the population approach is that one can obtain (1) a universal model for a whole "population" of analytes, (2) the model is obtained jointly for the whole data, so it can be used for a very sparse, imbalanced, and fragmentary data, (3) more detailed models can be used (there is no need to make some crude assumptions, like equal S for ionized and nonionized form of analyte, etc.) and (4) unbiased estimates are obtained. The last is very important if one wants to determine priors. For a very rich design, a nonlinear mixed-effect model and a traditional model fitting (one analyte per time) would yield similar results.

The accuracy of population and individual predictions is shown in Figure 1. The population predictions correspond to the retention times that are obtained from chromatographic

DOI: 10.1021/acs.analchem.5b01195
Anal. Chem. 2015, 87, 7241−7249

Table 2. Parameter Estimates Obtained for the Final Population Model Based on the 66 Analytes from the Calibration Data Set[a]

parameters	description	fixed-effects estimate, $\boldsymbol{\theta}$ (%CV)	random-effects estimate, $\boldsymbol{\Omega}^b$ (%CV)
log $k_{w,N}$	retention factor of nonionized form of an analyte extrapolated to neat water as an eluent		0.217 (10)
$\theta_{\log k_w}$	intercept	0.433 (44)	
$\theta_{\log k_w - \log P}$	slope for log P	0.915 (6)	
$\theta_{\log k_w - \mathrm{PSA}}$	slope for PSA	0.0144 (15)	
log $k_{w,I}$	retention factor of ionized form of an analyte extrapolated to neat water as an eluent	−1.06 (5)	0.124 (11)
$\theta_{\Delta \log k_w}$	the difference of log k between the nonionized and ionized form of an analyte		
$S_{1,N}$	the first slope coefficient for nonionized form of an analyte		0.437 (11)
θ_{SN}	intercept	2.39 (12)	
$\theta_{SN - \log P}$	slope for log P	0.756 (11)	
$\theta_{SN - \mathrm{PSA}}$	slope for PSA	0.0281 (12)	
$S_{1,I}$	the first slope coefficient for ionized form of an analyte		0.503 (15)
$\theta_{\Delta S\ (\mathrm{acids})}$	the difference between S_1 of ionized and nonionized form of acid	−0.831 (29)	
$\theta_{\Delta S\ (\mathrm{bases})}$	the difference between S_1 of ionized and nonionized form of acid	1.01 (15)	
S_2	the second slope coefficient	0.183 (17)	
p$K_a(\varphi(t))$	the pK_a value		0.193 (9)
acids: θ_α	the slope of pK_a vs organic modifier content for acids	1.61 (10)	
bases: $\theta_\alpha+\theta_{\mathrm{AB}-\alpha}$	the slope of pK_a vs organic modifier content for bases	−0.365 (19)	
a	the empirical parameter accounting for the influence of pH on retention of anions due to nonhydrophobic interactions	−0.0172 (5)	
cov($\eta_{\log k_{w,N}}\eta_{S,N}$)	covariance between log $k_{w,N}$ and S_N		0.248 (6)
intracompound variability:			
σ_{add}	additive error model component	0.137 (6)	
σ_{prop}	proportional error model component	0.00628 (10)	

[a]%CV denotes a coefficient of variation of parameter estimates. [b]$\boldsymbol{\Omega}$ − denotes variance−covariance matrix of the random effects. The diagonal elements are given in the table. All other element of that matrix are zero, except for the covariance between $\eta_{\log k_{w,N}}$ and $\eta_{S,N}$.

parameters calculated on the basis of eqs 10−14 with all ETAs equal to zero. They can be viewed as predictions that are based solely on the parameter derived from the covariates, thus, for a compound for which no observations have been made or the observations have been ignored. Inclusion of eta (random effects) improves model prediction considerable (Figure 1B) and leads to the far more accurate analyte-specific parameters. The individual parameters are obtained from the population mean estimate of parameters (prior) and each individual data (likelihood). Here the difference between the nonlinear mixed-effect model and a traditional two-stage analysis is clearly evident. In the case of a two-stage analysis, the information about other analytes in not utilized, as it is in the case of the population modeling.

When a population model is established (i.e., the experimental data of a previously studied group of analytes are adequately described by a set of mathematical equations with identified typical parameter values and their uncertainties), it can be used to predict the performance of an unstudied set of compounds by means of Bayesian reasoning. Figure 4 presents an application of the MAP Bayesian estimator for our first validation data set, with varying number of preliminary experiments ranging from 0 to 18. Furthermore, as can be expected, the more experiments that are performed, the higher overall accuracy of model predictions. From that graph, and calculated bias (MDPE) and precision (MDAPE), it seems that 4−6 experiments are sufficient to get parameter estimates that lead to the unbiased prediction of retention times for a variety of analytes. The four best experiments corresponds to the short gradients conducted at lowest, middle, and highest pH values and one additional gradient with 3-fold longer organic modifier gradients conducted at middle pH value.

The second validation data set was used to illustrate the usefulness of our approach for other modes than organic modifier gradient. We used the MAP Bayesian estimation to predict the retention time of ketoprofen and papaverine based on four preliminary experiments for a set of experiments covering a large body of isocratic, organic modifier gradient and pH gradient data. The predictions and preliminary experiments are presented in Figure 5. This example also shows that four preliminary experiments are sufficient to completely characterize analyte retention, if the prior knowledge about an analyte is utilized.

A major benefit of MAP Bayesian forecasting is that any number of preliminary retention time data can be used to predict retention. Also, an adaptive design can be proposed in which one starts without any preliminary experiments to determine the required (desired) retention times. If the prediction is unsatisfactory, one can treat that experiment as a preliminary one and obtain another predictions. Such a trial-and-error approach, supported by MAP Bayesian estimation, seems to be more scientifically based than the usual educated or experience-based guessing, and can quickly lead to the satisfactory separation.

The prior information is crucial for valid predictions. Because the literature does not always offer experimental pK_a and log P values, we were also interested in assessing the accuracy of predictions obtained from the computationally calculated values

DOI: 10.1021/acs.analchem.5b01195
Anal. Chem. 2015, 87, 7241−7249

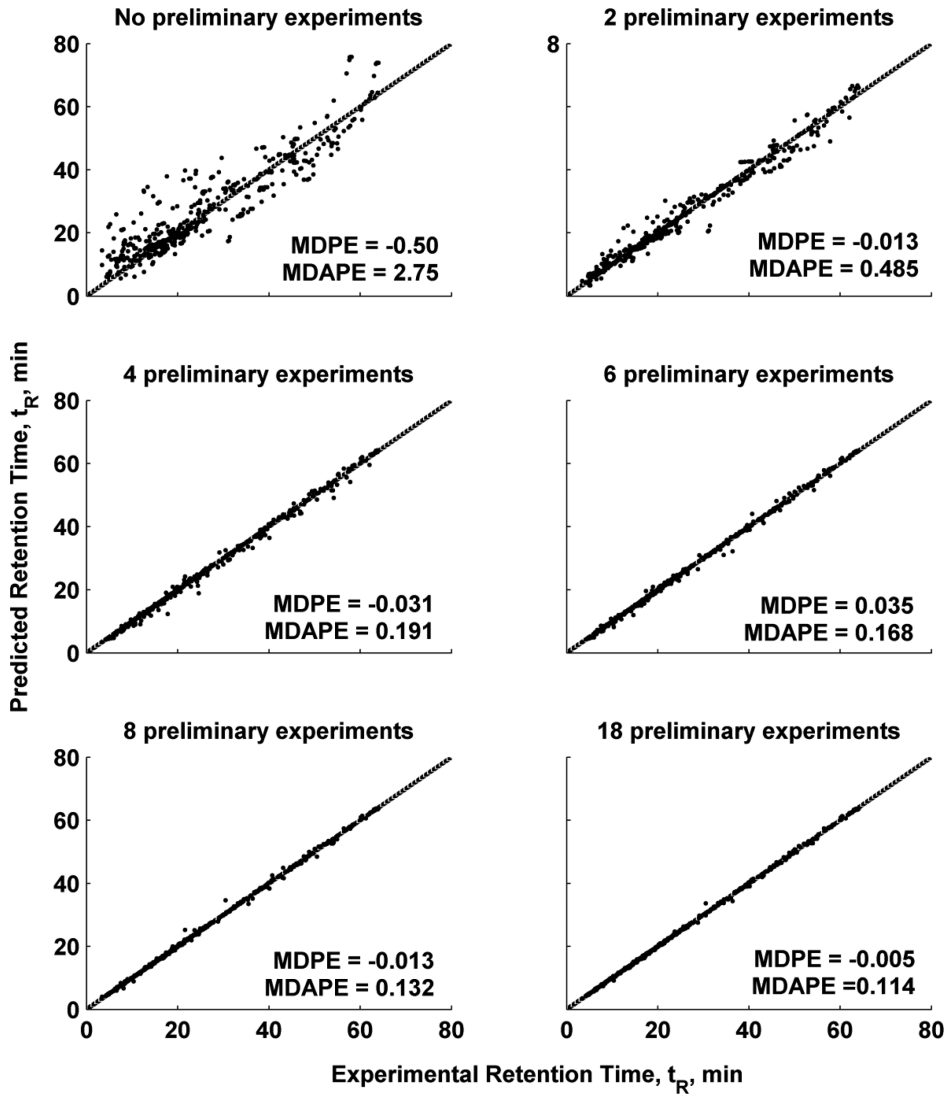


Figure 4. Experimental and model predicted retention times for external validation data set by means of MAP Bayesian Estimator with a varying set of preliminary experiments. The following experiments (see Table 1) were used as preliminary for each graph: 2 − {5, 14}, 4 − {1, 5, 9, 14}, 6 − {1, 5, 9, 10, 14, 18}, and 8 − {2, 4, 6, 8, 11, 13, 15, 17}.

(ACD/Laboratories). Nearly the same predictions as presented for experimentally measured values were obtained (data not shown). It is a particularly useful result for practical applications, as one can utilize the information from different sources to get the required parameters. We purposely limited ourselves to the very simple QSRR equations, with log P and pK_a, as those values are readily available from the chemical structures. More complex equations can be proposed, if they are derived from 3D molecular structure.[2] However, more studies are needed to address this issue.

Recently, a simplified method of prediction of the chromatographic retention of acid−base compounds in pH buffered methanol−water mobile phases in gradient mode was proposed.[8] It solves a similar problem as posted in this work, although using different methodology. The authors used several preliminary experiments to determine selected parameters of the model, like ratio between the retention factors of the pure ionized and pure neutral forms of the compound, and two parameters reflecting changes of retention factor with methanol content. The pK_a value was predicted from analyte structure. A simplified model (with the same slope of changes of retention factor with the increase of organic modifier content for ionized and neutral form of analyte) was used to predict analyte retention time. There are several limitations of such an approach as (1) the initially determined parameters are treated as fixed effects without any uncertainty and as such might lead to poorer predictions, especially if they are for some reason biased (i.e., the ratio between the retention factors of the pure ionized and pure neutral forms of the compound with very high or very low pK_a values is usually difficult to determine); (2) the simplest model needs to be used, with some crude assumptions, like the same behavior of ionized and neutral form of analyte due to the organic modifier changes. In our opinion, the MAP Bayesian method offers a more natural platform for combining prior knowledge about an analyte with a limited set of preliminary experiments and does not have the above-mentioned limitations.

Disadvantage of the proposed methodology includes reliance on the existence of an appropriate population models, prior knowledge of important covariates, and rather complex calculation requiring specialized software. Nevertheless, we believe that the proposed approach provides a means for fast

DOI: 10.1021/acs.analchem.5b01195
Anal. Chem. 2015, 87, 7241−7249

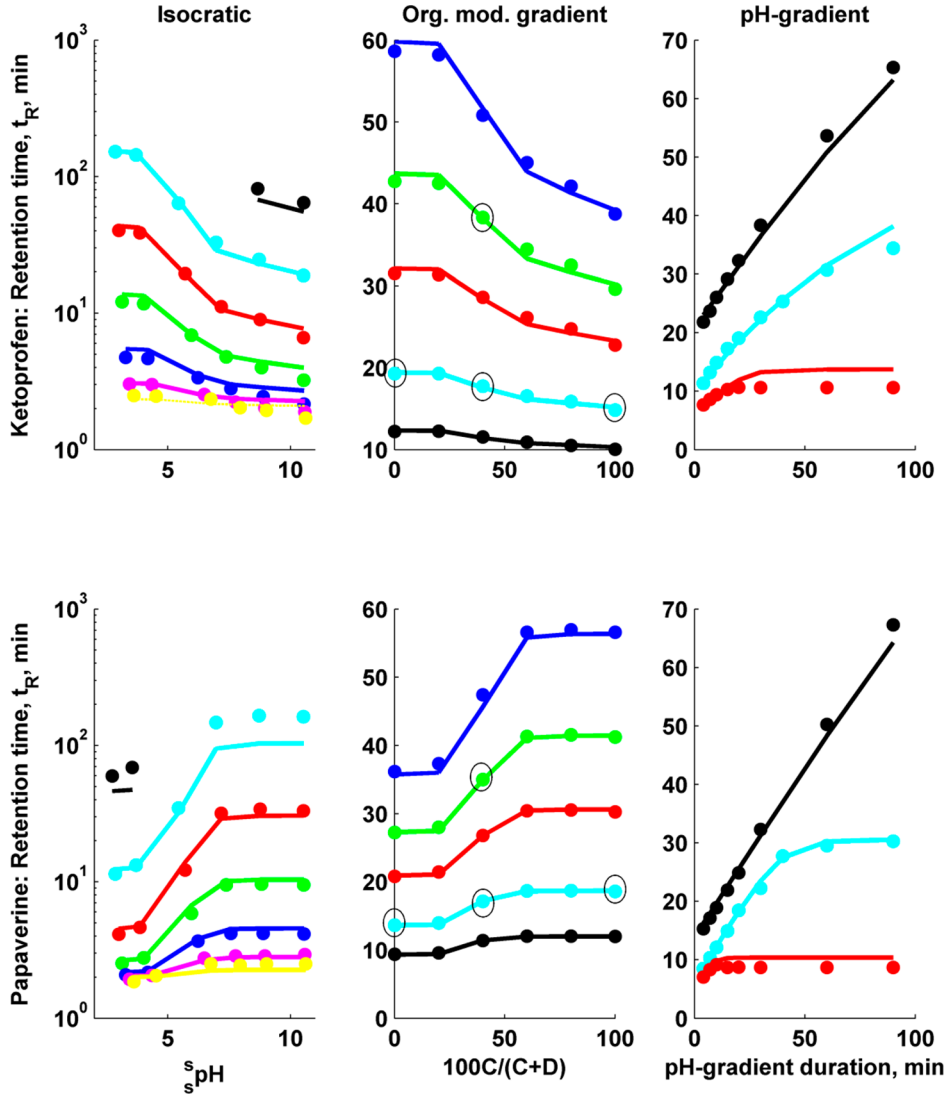


Figure 5. Experimental and model predicted retention times for ketoprofen (acid) and papaverine (base) by means of MAP Bayesian Estimator with 4 preliminary experiments (shown as circles). The RP HPLC experiments were carried out in a series of isocratic (at different pH and methanol contents of 20%, 30%, 40%, 50%,60%, 70%, 80%), organic modifier gradient (at approximately constant pH determined by the fraction of buffers C and D, and at different gradient durations of 10 min, 20 min, 40 min, 60 min, 90 min), and linear pH gradients (at different gradient durations and for different methanol contents of 30%, 40%, 50%) .

development of analytical methods. Especially when a population model for large samples of analytes will be publicly available.

CONCLUSIONS

The fitting of retention data simultaneously for a large group of analytes within the NMLE framework allows the elucidation of the QSRR relationship that could otherwise be difficult to obtain. If the proper sample of analytes is selected from the whole "population" of interest, valid relationships for that population can be obtained and used as priors for the prediction of retention times.

A very general strategy has been proposed for a robust and effective determination of chromatographic parameters, like log k_w and p$K_{a,chrom}$, from a limited number of experiments and *a priori* information easily accessible from the chemical structure of an analyte using MAP Bayesian estimation. Four preliminary experiments were proved to be sufficient to obtain an accurate estimate of chromatographically important parameters, leading to unbiased retention predictions when pH and organic modifier content are varied. These results can easily be extended to other optimization problems, including temperature and different type of organic modifiers.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b01195.

AUTHOR INFORMATION

Corresponding Author

*E-mail: wiczling@gumed.edu.pl. Tel.:++48 58 349 1493. Fax: ++48 58 349 3262.

Notes

The authors declare no competing financial interest.

DOI: 10.1021/acs.analchem.5b01195
Anal. Chem. 2015, 87, 7241−7249

■ ACKNOWLEDGMENTS

This Project was supported by the Ministry of Science and Higher Education of the Republic of Poland, from the quality-promoting subsidy, under the Leading National Research Centre (KNOW) programme for the years 2012−2017 and partially by the Polish National Science Centre grant 2014/13/N/NZ7/04218. We thank Academic Computer Center in Gdańsk for access to the computer cluster.

■ REFERENCES

(1) Nikitas, P.; Pappa-Louisi, A.; Zisi, C. *Anal. Chem.* **2012**, *84*, 6611−6618. Rosés, M.; Subirats, X.; Bosch, E. *J. Chromatogr A* **2009**, *1216*, 1756−75. Téllez, A.; Rosés, M.; Bosch, E. *Anal. Chem.* **2009**, *81*, 9135−45. Andrés, A.; Téllez, A.; Rosés, M.; Bosch, E. *J. Chromatogr A* **2012**, *1247*, 71−80.

(2) Kaliszan, R. *Chem. Rev.* **2007**, *107*, 3212−46.

(3) Molnar, I. *J. Chromatogr A* **2002**, *965*, 175−94.

(4) Tetko, I.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V.; Radchenko, E.; Zefirov, N.; Makarenko, A.; Tanchuk, V.; Prokopenko, V. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453−463.

(5) Snyder, L. R.; Kirkland, J. J.; Dolan, J. W. *Introduction to modern liquid chromatography*, 3rd ed.; Wiley-Blackwell: Oxford, 2010.

(6) Mould, D. R.; Upton, R. N. *CPT: Pharmacometrics Syst. Pharmacol.* **2013**, *2*, e38.

(7) Wiczling, P.; Markuszewski, M. J.; Kaliszan, R. *Anal. Chem.* **2004**, *76*, 3069−77. Wiczling, P.; Kaliszan, R. *Anal. Chem.* **2008**, *80*, 7855−61. Nasal, A.; Siluk, D.; Kaliszan, R. *Curr. Med. Chem.* **2003**, *10*, 381−426. Valko, K.; Bevan, C.; Reynolds, D. *Anal. Chem.* **1997**, *69*, 2022−2029. Kaliszan, R.; Haber, P.; Baczek, T.; Siluk, D.; Valko, K. *J. Chromatogr A* **2002**, *965*, 117−127.

(8) Andrés, A.; Rosés, M.; Bosch, E. *J. Chromatogr A* **2015**, *1385*, 42−8.

(9) Kaliszan, R.; van Straten, M.; Markuszewski, M.; Cramers, C.; Claessens, H. *J. Chromatogr A* **1999**, *855*, 455−486.

(10) Neue, U.; Phoebe, C.; Tran, K.; Cheng, Y.; Lu, Z. *J. Chromatogr A* **2001**, *925*, 49−67. Pappa-Louisi, A.; Nikitas, P.; Balkatzopoulou, P.; Malliakas, C. *J. Chromatogr A* **2004**, *1033*, 29−41.

(11) Wiczling, P.; Kawczak, P.; Nasal, A.; Kaliszan, R. *Anal. Chem.* **2006**, *78*, 239−49.

(12) Neue, U. *Chromatographia* **2006**, *63*, S45−S53.

(13) Neue, U.; Tran, K.; Mendez, A.; Carr, P. *J. Chromatogr A* **2005**, *1063*, 35−45. Mendez, A.; Bosch, E.; Roses, M.; Neue, U. *J. Chromatogr A* **2003**, *986*, 33−44.

(14) Canals, I.; Oumada, F.; Roses, M.; Bosch, E. *J. Chromatogr A* **2001**, *911*, 191−202.

(15) Rosés, M. *J. Chromatogr A* **2004**, *1037*, 283−98.

(16) Wiczling, P.; Kaliszan, R. *J. Chromatogr A* **2010**, *1217*, 3375−81.

(17) Kuhn, E.; Lavielle, M. *Comput. Stat Data Anal* **2005**, *49*, 1020−1038.

(18) Comets, E.; Brendel, K.; Mentre, F. *Comput. Meth Prog. Bio* **2008**, *90*, 154−166. Kiang, T.; Sherwin, C.; Spigarelli, M.; Ensom, M. *Clin. Pharmacokinet.* **2012**, *51*, 515−525.

(19) Davidian, M.; Giltinan, D. M.; *Nonlinear Models for Repeated Measurement Data*; Chapman & Hall: New York, 1995.

(20) Wiczling, P.; Struck-Lewicka, W.; Kubik, L.; Siluk, D.; Markuszewski, M. J.; Kaliszan, R. *J. Pharm. Biomed. Anal.* **2014**, *94*, 180−7.

(21) Canals, I.; Valko, K.; Bosch, E.; Hill, A.; Roses, M. *Anal. Chem.* **2001**, *73*, 4937−4945.

DOI: 10.1021/acs.analchem.5b01195
Anal. Chem. 2015, 87, 7241−7249