Journal of Chromatography A, 656 (1993) 577–590
Elsevier Science Publishers B.V., Amsterdam

CHROM. 25 195

Review

Modelling retention of ionogenic solutes in liquid chromatography as a function of pH for optimization purposes

Peter J. Schoenmakers* and Robert Tijssen

Koninklijke/Shell-Laboratorium, Amsterdam, (Shell Research BV), P.O. Box 3003, 1003 AA Amsterdam (Netherlands)

ABSTRACT

In liquid chromatography, the retention of ionogenic solutes is a strong function of the pH of the mobile phase, with different solutes showing different behaviour, both qualitatively (*e.g.*, acids *vs.* bases) and quantitatively (values of dissociation constants). Thus, pH will also affect selectivity and it can be used as a parameter for optimizing separations. In many instances, such optimization studies will require an accurate description of retention as a function of pH. In this paper, attention is focused on basic models describing retention as a function of pH and their use in practice. The theoretical, sigmoidal curve is discussed and a number of possible causes of deviations are considered. The inaccuracies introduced by linearly or quadratically interpolating part of a sigmoidal curve are addressed, in addition to the sensitivity of sigmoidal interpolation to experimental errors.

CONTENTS

* Corresponding author.

0021-9673/93/$24.00

1. INTRODUCTION

In developing and optimizing methods for high-performance liquid chromatography (HPLC), the manipulation of retention and selectivity through variation of suitable parameters plays a key role [1]. The main reason for this is the lack of alternative strategies. Unlike the situation in, for example, capillary gas chromatography and capillary zone electrophoresis, theoretical plates in HPLC are difficult to achieve. The typical approach to optimizing HPLC separations, therefore, involves the following steps: (1) simplifying the chromatogram, *i.e.*, minimizing the number of peaks, using selective sample preparation methods, multi-column techniques and selective detection; (2) optimizing the selectivity, *i.e.*, moving from an initial, more or less random distribution of peaks towards an optimum distribution of the relevant peaks over the chromatogram; and (3) optimizing the system (column dimensions, particle size, flow-rate, etc.), *i.e.*, minimizing the costs of the analysis in terms of analysis time, pressure drop, eluent consumption, etc., while achieving adequate resolution and sensitivity.

Optimizing the solvent pH could be a factor in all three stages. By establishing appropriate pH values, potentially interfering compounds can be selectively removed from the sample (step 1). For instance, if at a certain pH the relevant components in the sample are neutral, positively and negatively charged compounds can be removed using cation- and anion-exchange materials, respectively. Essentially, the pH used for the actual separation (step 2) is independent of the sample preparation process. The two steps can be independently optimized, with great advantages in terms of simplicity and requirements (time, number of experiments). An effect of pH on both the selectivity and the sensitivity of detection is feasible (see, *e.g.*, ref. 2), yet in most instances such effects will not be dramatic and the advantages of individually addressing the above three steps easily outweigh the disadvantages. In cases in which optimizing the detection pH leads to important gains, one possible solution is the postcolumn addition of a high-capacity buffer of suitable pH. In this paper we shall focus exclusively on the role of pH for optimizing the separation, *i.e.*, the effects of pH within the analytical column.

2. THEORETICAL MODELS

For a weak acid (denoted by HA), the effect of pH on retention can easily be expressed in an algebraic equation. Two assumptions are typically made:

(i) the dissociation equilibrium of the acid can be expressed as

$$K_{\mathrm{a,HA}} = \frac{a_{\mathrm{H^+}} a_{\mathrm{A^-}}}{a_{\mathrm{HA}}} \approx \frac{c_{\mathrm{H^+}} c_{\mathrm{A^-}}}{c_{\mathrm{HA}}} \tag{1}$$

where $K_{\mathrm{a,HA}}$ is the acid dissociation constant of the compound, a_i is the thermodynamic activity of the indicated species in solution and c_i is its concentration; and

(ii) the observed capacity factor of species A is a weighted average of those of the two forms (HA and $\mathrm{A^-}$):

$$k_{\mathrm{A}} = \frac{c_{\mathrm{HA,m}} k_{\mathrm{HA}} + c_{\mathrm{A^-,m}} k_{\mathrm{A^-}}}{c_{\mathrm{HA,m}} + c_{\mathrm{A^-,m}}} \tag{2}$$

where $c_{i,\mathrm{m}}$ is the concentration of species i in the mobile phase. For reasons of simplicity, the subscript m will be omitted below, as concentrations in the stationary phase are not relevant in the context. This second assumption is equivalent to Horváth *et al.*'s assumption of an independent distribution of each solute species over the mobile and stationary phases [3].

Combining eqns. 1 and 2 leads to a general equation relating the observed capacity factor to the acid dissociation constant K_{a} or its negative logarithm pK_{a} and the acidity of the mobile phase ($c_{\mathrm{H^+}}$ or pH):

$$k_{\mathrm{A}} = \frac{c_{\mathrm{H^+}} k_{\mathrm{HA}} + K_{\mathrm{a,A}} k_{\mathrm{A^-}}}{c_{\mathrm{H^+}} + K_{\mathrm{a,A}}} = \frac{10^{-\mathrm{pH}} k_{\mathrm{HA}} + 10^{-\mathrm{p}K_{\mathrm{a,A}}} k_{\mathrm{A^-}}}{10^{-\mathrm{pH}} + 10^{-\mathrm{p}K_{\mathrm{a,A}}}} \tag{3}$$

Eqn. 3 contains one variable (pH) and three coefficients ($K_{\mathrm{a,A}}$, k_{HA} and $k_{\mathrm{A^-}}$). For a given set of values for the three solute-specific coefficients, eqn. 3 represents the well known sigmoidal curve

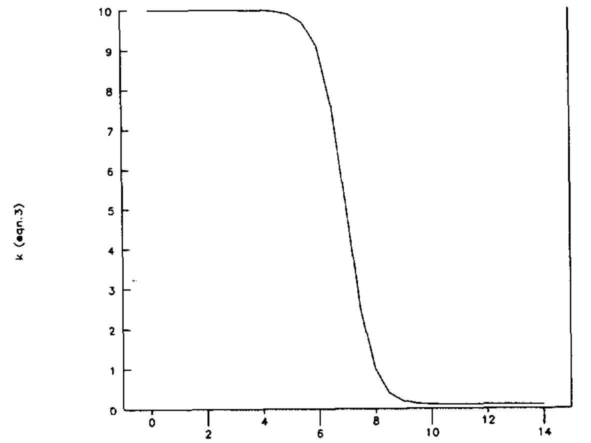

Fig. 1. Typical sigmoidal curve, depicting retention as a function of pH according to eqn. 3. Coefficients: $k_{HA} = 10$, $k_{A^-} = 0.1$ and $pK_a = 7$.

illustrated in Fig. 1. In reversed-phase liquid chromatography (RPLC) the retention of ionized species (such as A^-) is typically low, whereas the retention of neutral species (such as HA) is much higher. For a weak acid, this results in the curve depicted in Fig. 1. At low pH ionization is suppressed and retention is high. At high pH the solute is completely ionized and retention is low. For a basic compound (*e.g.*, an amine) the situation is reversed. At high pH the solute will be neutral and retained more or less strongly. At low pH it will be charged (*i.e.*, protonated, RNH_3^+) and eluted much earlier. Indeed, we can use the typical behaviour of acidic and basic solutes outlined above for a pragmatic definition of acidic and basic solutes in RPLC, as follows [4]:

(i) A weakly acidic solute is retained more strongly at the lower end of the practical pH range than it is at the higher end.

(ii) A weakly basic solute is one that is retained less at lower pH than at higher pH.

(iii) A strongly acidic solute is one that is dissociated (and hence negatively charged) throughout the pH range. Retention is not affected by pH. It tends to be low, but can be decreased with the aid of a cationic pairing ion (*e.g.*, a tetraalkylammonium compound).

(iv) A strongly basic solute is one that is protonated (and hence positively charged) throughout the pH range. Retention is not affected by pH. It tends to be low, but can be increased with the aid of an anionic pairing ion (*e.g.*, an alkylsulphonate).

(v) A neutral solute is uncharged throughout the pH range. Retention is not affected by pH or by the addition of pairing ions.

Using the above pragmatic definitions, a compound may be classified differently when the practical pH range changes. For example, silica-based columns are typically restricted to a range of $2 < pH < 7$, whereas polystyrene- or graphitic carbon-based columns may allow much wider ranges (*e.g.*, $1 < pH < 13$). As a consequence, a compound may be classified as a strong base on one column, whereas it may be a weak base on another column. Using the above conventions, therefore, does not result in an absolute classification of solutes, but in a practical classification in relation to the type of chromatography being performed.

The effects of pH on retention and peak shape and of the addition of negatively and positively charged pairing ions on retention for the five identified classes of solutes is summarized in Table 1.

TABLE 1

INDICATION OF THE EFFECTS OF pH AND OF NEGATIVELY (PI^-) AND POSITIVELY CHARGED (PI^+) ION-PAIRING AGENTS ON THE CHROMATOGRAPHIC BEHAVIOUR (CAPACITY FACTOR, *k*, AND PEAK SHAPE) OF FIVE CLASSES OF SOLUTES

Solute classes: SA = strong acid; WA = weak acid; N = neutral; WB = weak base; SB = strong base. Magnitude of effects: L = large; S = small; V = variable; N = negligible.

Solute class	Effect of pH on *k*	Effect of pH on peak shape	Effect of PI^- on *k*	Effect of PI^+ on *k*
SA	S	V	S	L
WA	L	L	S	V
N	N	N	N	N
WB	L	L	V	S
SB	S	V	L	S

3. COMPLICATIONS

Having come to a very simple description of retention as a function of pH, we should add that there are several factors causing deviations in real life. These are summarized below.

(i) Eqn. 1 strictly applies in terms of *activities* rather than *concentrations*. If there is no strict proportionality between the two, eqn. 3 will not strictly be valid. Practical consequences are as follows:

(a) The dissociation constant will be affected by the total ionic strength, as the latter causes different ratios between activities and concentrations (*i.e.*, activity coefficients) for ionic and non-ionic species.

Chromatographers often vary the ionic strength of their mobile phase unintentionally, but significantly. For example, Fig. 2 shows the (calculated) ionic strength for buffers prepared by titrating a 0.287 *M* solution of triethylamine with a 0.1764 *M* solution of phosphoric acid. Optimistically (*i.e.*, assuming the widest column stability range of $1 < \text{pH} < 8$), the resulting buffers may be applied on silica-based LC columns in the ranges $1 < \text{pH} < 3$ and $6 < \text{pH} < 8$. In the important "central" region ($3 < \text{pH} < 6$) phosphate (and triethylamine) solutions have no buffer capacity. Variations in the ionic strength of up to a factor three are seen to occur in Fig. 2. In order to maintain a constant ionic strength, measured amounts of a dissociating salt must be added to the mobile phase.

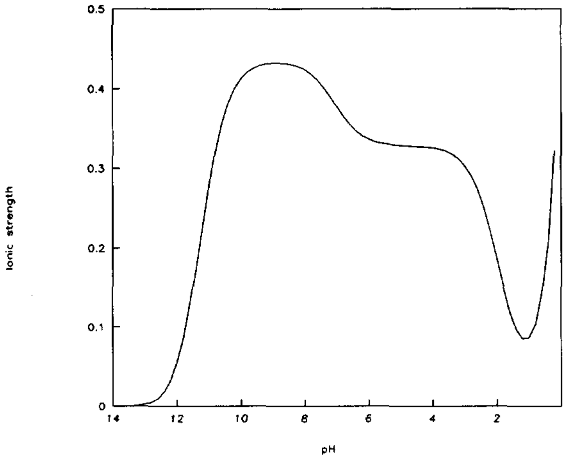

Fig. 2. Calculated total ionic strength as a function of (calculated) pH for buffers prepared by titrating a 0.287 *M* solution of triethylamine with a 0.1764 *M* solution of phosphoric acid.

(b) Activity coefficients, may be affected by the solvent, so that K_a is a function of the solvent composition.

(ii) The true pH of the mobile phase is the negative logarithm of the activity of the H^+ species. Again, the activity coefficient will be affected by the environment. Thus, if an aqueous buffer of pH 5 is mixed with an organic solvent, the true pH as defined above will be different from that of the original buffer [5]. While corrections can be made [6,7], this is an undesirable complicating factor for LC optimization studies. Rather than repeatedly converting between the pH of the aqueous fraction and true pH in a mixed solvent, we consistently refer to the former as the pH of the eluent. Using such an "operational pH" is a perfectly legitimate and very practical convention in the context. However, other results from the optimization process than the optimum mobile phase composition and pH should be interpreted with care. For example, pK_a values obtained from fitting an equation such as eqn. 3 to a series of data are not absolute ("true") values, but refer to the above convention for defining the pH. For a more extensive discussion, see ref. 8.

(iii) The stationary phase may affect the validity of eqn. 3. By postulating eqn. 2 it is assumed that there is a given, pH-independent capacity factor for each individual species (HA and A^-). This assumption is no longer reasonable if the stationary phase is affected considerably by variations in pH. An obvious example is the dissociation equilibrium of silanol groups on silica-based stationary phases. This is a gradual process, *i.e.*, the fraction of ionized silanol groups increases gradually with increasing pH over a broad range centered around 6.5 [9]. Ionic interactions (attraction or repulsion) between electrically charged solute molecules and dissociated silanols may significantly affect the shape of retention *vs.* pH curves in RPLC [10].

(iv) Solutes may show multiple ionization equilibria. In the case of diprotic acids or bases, Fig. 1 may show two steps rather than one. However, if the pH range studied is limited, as is typically the case on silica-based columns (2 <

pH<7), it is uncommon that both steps are observed within the practical "window". However, deviations from eqn. 3 have been observed, for example, for phthalic acid [11]. The situation is more complicated for zwitterions, *i.e.*, compounds which contain both acidic and basic functions. In principle, characteristic bell-shaped curves may be obtained for retention *vs.* pH, which cannot be described by eqn. 3. Horváth *et al.* [3] have published the equations equivalent to eqn. 3 for diprotic acids and zwitterions. A general formalism to deal with multi-protic acids and based can be found in ref. 8.

(v) In addition to dissociation, there may be other pH-dependent processes involving the solute species. One example concerns ion-pair interactions between solute ion and buffer components. For example, a solute cation may associate with phosphate ions. If the degree of association and/or the retention of the ion pair is different for the different phosphate ions (phosphate, hydrogenphosphate, dihydrogenphosphate), this causes a pH-dependent contribution to the observed retention of the solute.

Despite the possible deviations due to all the sources listed above, eqn. 3 is a very good starting point [3,12]. If necessary, the equation can be modified to deal with deviations. Finally, it is worth mentioning that there are many other factors that may complicate the retention behaviour of solutes in RPLC, but do not affect the validity of eqn. 3. Notably, all factors that affect the values of k_{HA} and/or k_{A^-} are irrelevant, as long as the assumptions to treat these two parameters as constants (independent of pH) and to calculate k_{A} as a weighted average remain unaffected. The mechanism of separation is thus not an issue here. The retention of either species may be controlled by "solvophobic", "silanophilic" or electrostatic interactions, or by any combination thereof. In fact, it is possible, both theoretically (eqn. 3) and practically [13], to deal with negative values of k_{A^-}, which may result from electrostatic exclusion of ions.

4. EFFECT OF SOLVENT COMPOSITION

RPLC is usually performed with mobile phases consisting of a mixture of water (or an aqueous buffer) and an organic solvent. When optimizing solvent selectivity, *i.e.*, the type(s) and concentrations(s) of organic solvent(s), the concept of isoeluotropic solvents [14] is frequently employed. In this case, the selectivity (solvent composition) can be varied in such a way that the overall retention is kept roughly constant. If pH is used to vary the selectivity, this (or a similar) concept cannot be applied. Varying the pH can greatly affect the retention of ionogenic solutes (see Fig. 1) and the pH corresponding to optimum selectivity between two or more solutes may lead to very high or very low retention at a specific mobile phase composition. The logical response is to adapt the mobile phase composition so as to derive the maximum benefit from pH-induced selectivity effects.

The variation of retention (expressed as the natural logarithm of the capacity factor) with binary composition (volume fraction of organic solvent in water, φ) can be expressed by a quadratic function [15]:

$$\ln k = A\varphi^2 + B\varphi + C \qquad (4)$$

where A, B and C are coefficients which depend on the type of organic solvent (modifier), the solute and the stationary phase. Over a limited range of capacity factors (typically $1 < k < 10$) a straight-line approximation often suffices

$$\ln k = \ln k_0 - S\varphi \qquad (5)$$

The coefficients $\ln k_0$ and S, which are susceptible to the same influences as the coefficients in eqn. 4, can be interpreted as the extrapolated retention in pure water and the rate of variation of retention with varying modifier content. Eqn. 4 or the simplified eqn. 5 can be thought to represent the variation of the retention of each individual species (*e.g.*, HA or A^-) with composition. As mentioned before, the third coefficient in eqn. 3, K_{a}, should also be assumed to vary with pH. The variation of $\ln K_{\mathrm{a}}$ with φ may be described by a quadratic equation similar to eqn. 4 or by a cubic equation [8].

By starting with eqn. 3 and assuming the coefficients in that equation to be (logarithmically) linear (eqn. 5), quadratic (eqn. 4) or cubic functions of composition, one can obtain equations that express the capacity factor as a function of both pH and composition. Alternatively, eqn. 4 or 5 can be taken as the starting point and eqn. 3 can be assumed to describe the variation of the coefficients with pH. Lopes Marques and Schoenmakers [8] have evaluated a number of such equations. They obtained the best results by starting with eqn. 3 and assuming ln K_{HA}, ln k_{A^-} and ln K_a all to vary quadratically with φ according to

$$\ln k_{HA} = \ln k^0_{HA} + s_{HA}\varphi + T_{HA}\varphi^2 \quad (6)$$

$$\ln k_{A^-} = \ln k^0_{A^-} + S_{A^-}\varphi + T_{A^-}\varphi^2 \quad (7)$$

$$\ln K_a = \ln K^0_a + Q_1\varphi + Q_2\varphi^2 \quad (8)$$

The following equation was derived:

$$k = \frac{k^0_{HA}\exp(S_{HA}\varphi + T_{HA}\varphi^2)c_{H^+} + k^0_{A^-}K^0_a\exp[(S_{A^-}+Q_1)\varphi + (T_{A^-}+Q_2)\varphi^2]}{c_{H^+} + K^0_a\exp(Q_1\varphi + Q_2\varphi^2)} \quad (9)$$

Eqn. 9 contains nine coefficients (three each from eqns. 6–8) and the two variables pH and φ. While there is some theoretical foundation for eqns. 3, 6 and 7, eqn. 8 is essentially empirical. Eqn. 9 is only one of a number of partly theoretical, partly empirical equations that can be derived. It was found to behave best in a practical evaluation [8] of possible model equations. For describing the variation of retention with pH and composition by a single equation it is the recommended form.

5. EMPIRICAL AND MODEL-FREE APPROACHES

A common way of modelling curves or surfaces is by interpolation. Characteristics of interpolation are that (i) experimental data are available in all directions from the point at which a value is needed (*e.g.*, obtaining a value at pH 5 from data at pH 4 and pH 6) and (ii) a model is used that fits exactly through all data points (*e.g.*, a straight line through two points or a quadratic curve through three points).

Interpolations can be performed with theoretical equations, such as eqn. 3 in the pH domain or eqn. 4 in the composition domain (both requiring three experimental data points), or even with eqn. 9 in a two-dimensional pH–composition space (requiring nine experimental data points). As an alternative to theoretical models, linear, quadratic or higher polynomial functions can be used. Quadratic [16,17] and piecewise-quadratic [18] interpolants are frequently used for describing the influence of pH in RPLC. Lopes Marques *et al.* [19] have suggested the use of so-called smooth-surface interpolants in situations in which adequate model equations are lacking.

Linear and quadratic interpolations can easily be performed. An analytical expression for interpolating data according to a sigmoidal function (eqn. 3) can also be derived. Assume that the pH is varied (with other parameters, including composition constant) and that three data points (numbered 1, 2 and 3) have been recorded. With χ denoting the equivalent hydrogen ion concentration ($\chi = 10^{-pH}$) the three data points can be described as

$$k_1 = \frac{\chi_1 k_{HA} + K_a k_A}{\chi_1 + K_a} \quad (10)$$

$$k_2 = \frac{\chi_2 k_{HA} + K_a k_A}{\chi_2 + K_a} \quad (11)$$

$$k_3 = \frac{\chi_3 k_{HA} + K_a k_A}{\chi_3 + K_a} \quad (12)$$

From these three equations, the unknown coefficients (K_a, k_{HA} and k_A) can be eliminated, after which the retention at any pH can be immediately predicted. The resulting expression for the capacity factor at χ_* is

$$k_* = -\frac{k_1k_2 + \theta k_2k_3 - (\theta+1)k_1k_3}{\theta k_1 - (\theta+1)k_2 + k_3} \quad (13)$$

with

$$\theta = \left(\frac{\chi_* - \chi_1}{\chi_* - \chi_3}\right)\left(\frac{\chi_2 - \chi_3}{\chi_1 - \chi_2}\right) \quad (14)$$

When interpolation takes place at regular pH intervals, θ becomes a simple constant. For example, when $\text{pH}_1 = \text{pH}_* - 0.5$, $\text{pH}_2 = \text{pH}_* + 0.5$ and $\text{pH}_3 = \text{pH}_* + 1.5$, we find $\theta = -0.2233$.

Using eqns. 13 and 14, sigmoidal interpolation is computationally not more difficult than quadratic interpolation. However, Lewis *et al.* [11] have found that sigmoidal interpolation may in some instances give rise to anomalous results in terms of "imaginary" values of k_{A^-} (or k_{HB^+}) or K_a. Especially when none of the three data points is close to the inflection point (say $pK_a - 1 \leq pH \leq pK_a + 1$), unrealistic values for the dissociation coefficient may be obtained. However, as long as k_* and not K_a is the desired outcome of the calculations, anomalies in the latter parameter are not necessarily indicative of poor interpolations. Likewise, negative values for the capacity factor of the ionized species may occur, but such values may even be physically meaningful, *i.e.*, they may arise from electrostatic exclusion of the ions. In the Results section the usefulness of sigmoidal interpolation for retention as a function of pH will be discussed in more detail.

In addition to k_*, the dissociation coefficient K_a can easily be calculated from three experimental data points. The appropriate equation is

$$K_a = -\frac{\beta k_1 \chi_1 + (1-\beta)k_2\chi_2 - k_3\chi_3}{\beta k_1 + (1-\beta)k_2 - k_3} \quad (15)$$

with

$$\beta = \frac{\chi_2 - \chi_3}{\chi_2 - \chi_1} \quad (16)$$

6. PRACTICAL CONSIDERATIONS

The different models described above have different requirements for their practical implementation. The complexity of using the model roughly increases with increasing complexity of the model itself. Linear interpolation between individual data points is obviously easiest in practice. Even in two dimensions (pH and φ) it is relatively straightforward to identify the three experimental locations surrounding each particular point in the parameter space and to perform the interpolation. Linear interpolation is used extensively in commercial software.

Non-linear interpolation is more difficult. The algorithms required may be complex, but they have typically been developed for other ("general") purposes and standard software procedures can usually be called upon. Eqn. 13 facilitates rapid sigmoidal interpolation. All interpolation routines tend to be quick and undemanding as far as processing time and computer memory are concerned.

Fitting non-linear models to data sets is a more involved approach in practice. Routines for non-linear regression are contained in many software packages, but the iterative process requires a set of initial estimates for the coefficients to be determined. These initial estimates have to be provided by the chromatographer and the regression process may put very high demands on their accuracy. Lopes Marques and Schoenmakers [20] have described a method for automatically obtaining accurate initial estimates for a particular (3 × 4) experimental design. They also described a parameter transformation, which made the accuracy of the initial estimates less critical.

Non-linear least-squares procedures also suffer from a general lack of robustness. Convergence of the process to the optimum set of coefficients may be hampered by the (un)availability of data points (leading to singular matrices) and by the occurrence of very large or very small numbers at intermediate stages (leading to overflow or underflow situations).

7. GENETIC ALGORITHMS

Genetic algorithms (GA) offer an alternative approach for establishing the coefficients in a complex non-linear equation such as eqn. 9 [20]. The idea behind GA has been borrowed from evolution theory. In essence, a set of trial solutions is formed, from which only the best are retained. These serve as the basis for creating a new set of solutions through "evolutionary" processes (cross-over and mutation).

The application of a GA for the present purpose can be described by the following three-step process (see also ref. 20). First, a possible solution to the problem is represented by a bit string. For example, when trying to fit eqn. 3 every solution is a set of values for the three parameters k_{HA}, k_{A^-} and K_a. These can be expressed as binary numbers over a given range. The precision depends on the number of bits dedicated to each parameter. If 8-bit resolution is chosen and the ranges are $0 \leqslant k_{\mathrm{HA}} \leqslant 50$, $-1 \leqslant k_{\mathrm{A}} \leqslant 2$ and $0 \leqslant \mathrm{p}K_{\mathrm{a}} \leqslant 14$, then the string 10101000-00101111-01001000 (where the hyphens are added for clarity), corresponding to the decimal numbers 168, 47 and 72, represents the solution $k_{\mathrm{HA}} = 33$, $k_{\mathrm{A}} = -0.45$ and $\mathrm{p}K_{\mathrm{a}} = 3.95$. Every other bit string corresponds to a set of three values within the indicated ranges and, conversely, every set of three parameter values within these ranges can be approximated by a bit string.

Second, one must be able to assign an objective quality to every possible bit string. In our case, this can be done by comparing the model that corresponds to the bit string with experimental data. For example, using the above values for k_{HA}, $k_{\mathrm{A^-}}$ and K_{a} we can calculate capacity factors of 29.6, 2.3 and −0.4 at pH 3, 5 and 7, respectively. If the experimental values at these three pH values are known to be 22, 7 and 1, respectively, then we can calculate the sum of squared deviations (SSQ) (calculated − experimental) as $7.6^2 + (-4.7)^2 + (-1.4)^2 = 81.7$. Every possible solution (bit string) can be measured against the same three experimental data points in terms of an SSQ. The best solution is that with the lowest SSQ value.

Third, the best solution can be found by starting with an arbitrary set of randomly generated bit strings and manipulating these in an evolutionary manner. If the initial set contains 100 solutions, we can select the best ten of these based on the SSQ values. We may then generate a new set of 100 by randomly combining parts of one bit string with part of another. For example, the above string 10101000-00101111-01001000 maybe crossed with the string 01001011-01110100-00111010 to yield 101010000-00*x*110100-00111010 or 010010*x*00-00101111-01001000. Finally, a few of the bit strings in the new set may be altered at one position (mutation), with the idea of covering all possible areas of the parameter space. Thus, the above string 010010*x*00-00101111-01001000 may be transformed into the string 010010*x*00-00101$\underline{0}$11-01001000.

GA offer a robust means to find the global optimum, *i.e.*, the best values for the coefficients. However, the process is relatively slow and of an approximate nature (more accurate solutions require longer bit strings and, thus, more computation time). A much more complete description of genetic algorithms and their applications in analytical chemistry can be found in a tutorial article by Lucasius *et al.* [21].

8. EXPERIMENTAL

All calculations reported in this paper were performed in Lotus 1-2-3 Spreadsheets (Release 2.4; Lotus, Cambridge, MA, USA) on a Compaq (Houston, TX, USA) 386s/20 personal computer.

9. RESULTS AND EVALUATION

9.1. *Accuracy of linear and quadratic interpolation*

To discuss the accuracy of linear and quadratic interpolation of retention *vs.* pH data, we shall assume the simple model of eqn. 3 to be valid. The curve depicted in Fig. 1 will form the basis of the discussion. Linear interpolation can take place between any two points on the curve. It is obvious that on either side of the inflection point the interpolation error will increase with increasing distance between the points. Fig. 3a illustrates the error obtained by interpolating over 1-unit intervals in pH for all points between pH 0.5 and 13.5 at 0.5 unit intervals. The errors shown in Figs. 2 and 3 are obtained by subtracting the value obtained from the simple sigmoidal curve from that obtained through interpolation. For example, the error at pH 3 can be found from $e(3) = \frac{1}{2}f(2.5) + \frac{1}{2}f(3.5) - f(3)$, where $e(\mathrm{pH})$ denotes the error and $f(\mathrm{pH})$ the sigmoidal function (eqn. 3). Errors are largest in the vicinity of $\mathrm{p}K_{\mathrm{a}}$, except when the two data points happen to be on either side of the inflection point. The deviations between the sigmoidal curve and the linear interpolation are found to be considerable. Fig. 3b illustrates the same errors in terms of percentage points. The resulting curve is non-symmetrical, with the largest errors occurring at the lowest k values (high pH).

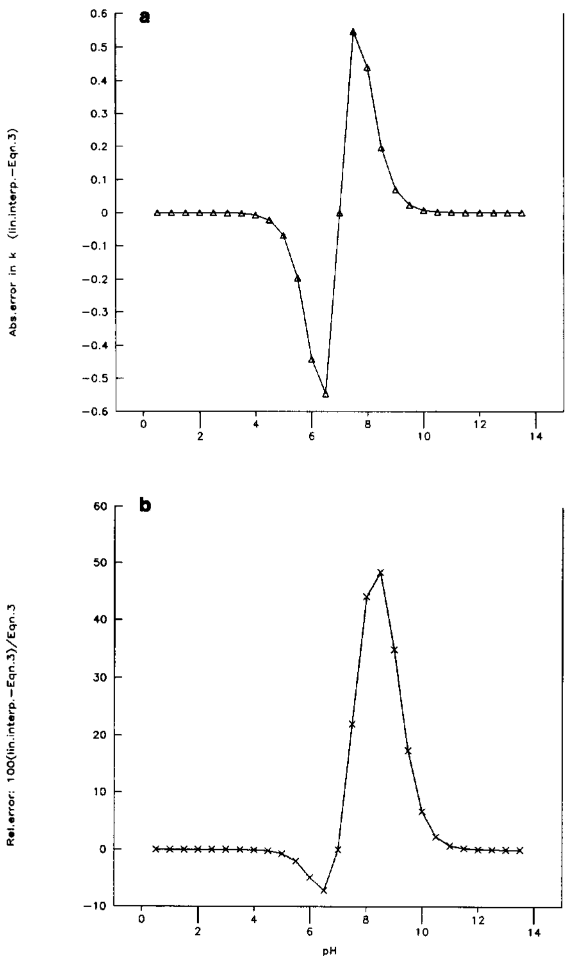

Fig. 3. Error induced by linearly interpolating the sigmoidal curve in Fig. 1 over intervals of 1 pH unit. (a) Absolute errors in k.; (b) relative errors (%).

Inevitably, prediction errors in capacity factors, as depicted in Fig. 3, will also lead to errors in predicted resolution values. When using linear interpolation, the errors in k for each solute will be characterized by a curve such as Fig. 3a. Large errors will obviously occur when the curves are "out-of-phase", as will be the case for an acidic and a basic compound with the same pK_a values. Also, large errors will result for two acidic or two basic compounds, the pK_a values of which are one or two units apart (two curves as Fig. 3a shifted by one or two units). However, even a much smaller shift will lead to large errors in resolution, as is illustrated in Fig. 4. Fig. 4a

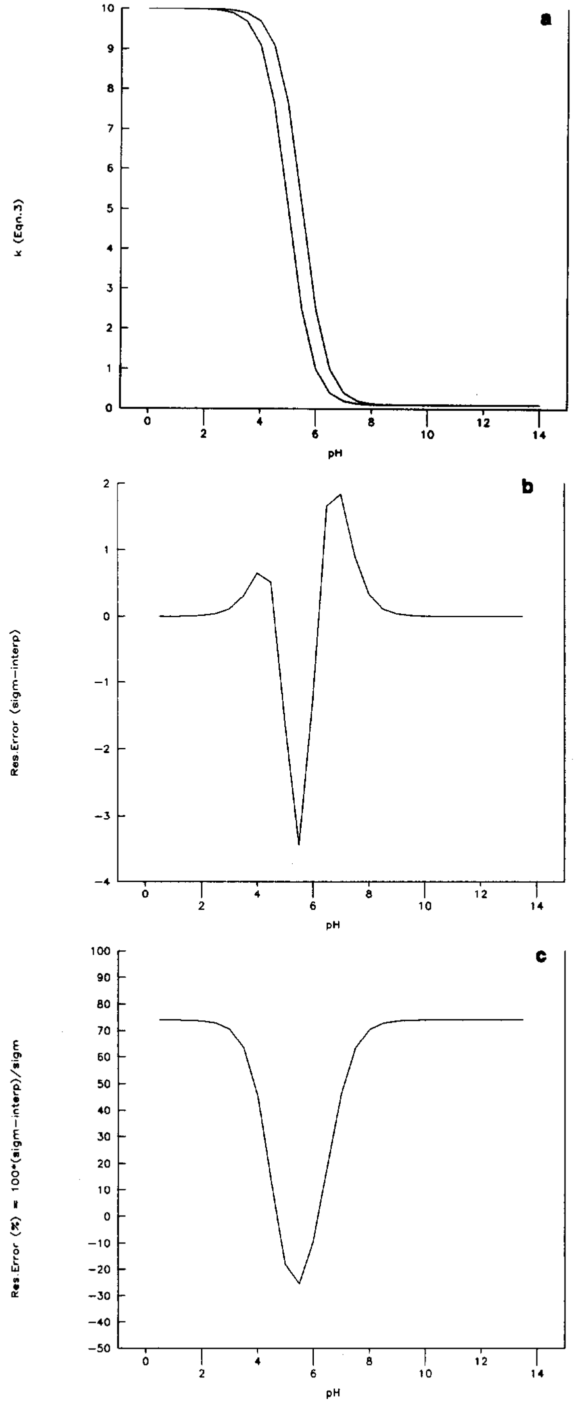

Fig. 4. Error obtained by predicting resolution values through linear interpolation of sigmoidal retention curves over 1 pH unit. (a) Retention curves ($k_{HA,1} = k_{HA,2} = 10$, $k_{A^-,1} = k_{A^-,2} = 0.1$, p$K_{a,1} = 5$, p$K_{a,2} = 5.5$); (b) absolute error in predicted resolution; (c) relative error in predicted resolution.

shows the retention curves for two acidic compounds, with identical values for k_{HA} (10) and k_{A^-} (0.1), but with different pK_a values (5 and 5.5, respectively). Fig. 4b illustrates the difference between the R_s values obtained by linear interpolation over one pH unit and the true value obtained from the sigmoidal retention curves. Large errors of up to several resolution units are found. Fig. 4c shows the relative errors in resolution, which are seen to be highest when resolution is negligibly low (at high and low pH values). However, over most of the region where the two compounds can be resolved the error in the resolution predicted by linear interpolation is much larger than the few percent thought tolerable for the purpose of pH optimization [11].

Fig. 5a illustrates the error found between the sigmoidal curve and a quadratic interpolation using data points $-\frac{1}{2}$, $+\frac{1}{2}$ and $+1\frac{1}{2}$ pH units removed from the point for which an interpolated value is sought (the *target*, indicated by pH* and the interpolated capacity factor, k*). In this instance the errors are smaller around the upper and lower bends in the sigmoid, but larger around the inflection point. Similar, but opposite, errors are found when the data point at $+1\frac{1}{2}$ pH units is replaced by one at $-1\frac{1}{2}$ pH units (Fig. 5b). The largest errors still exceed 0.5 units in k, similar to the maximum error found with linear interpolation over 1 pH unit (Fig. 3a). The maximum error can be reduced by piecewise quadratic interpolation (Fig. 5c). This implies taking the average of two possible quadratic interpolations. To do so, four instead of three data points are required. For comparison purposes, the two quadratic interpolations (lines) and the average values (symbols) are all shown in Fig. 5c.

Table 2 summarizes the errors obtained by the various interpolation methods on the curve of Fig. 1. Errors are listed over all possible points across a wide range (*i.e.*, those points for which the encompassing data points exist for all interpolation procedures), as well as for a limited pH range around the pK_a value of 7. Because relative errors depend greatly on the value of k, absolute errors are thought to provide a better indication of model accuracy.

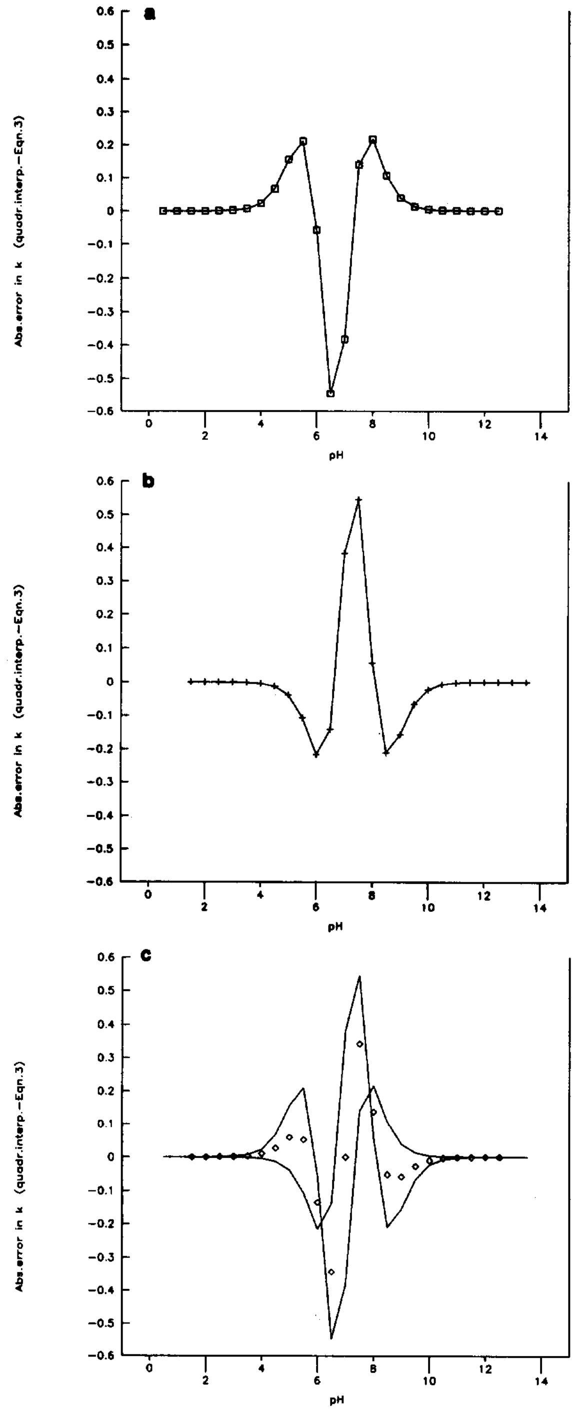

Fig. 5. Absolute errors induced by quadratic interpolation of the sigmoidal function in Fig. 1. (a) Using data points at $-\frac{1}{2}$, $+\frac{1}{2}$ and $+1\frac{1}{2}$ pH units from the target; (b) using data points at $-\frac{1}{2}$, $+\frac{1}{2}$ and $-1\frac{1}{2}$ pH units; (c) piecewise quadratic interpolation (symbols), using the average of (a) and (b) (lines).

TABLE 2

AVERAGE DEVIATION OF VARIOUS INTERPOLATIONS FROM THE CURVE SHOWN IN FIG. 1

Interpolation	Data points (relative)	pH range	N	Average error	
				Absolute	%
Linear	$-\frac{1}{2}, +\frac{1}{2}$	2–12	21	0.12	9.1
		6–8	5	0.39	15.6
Quadratic	$-\frac{1}{2}, +\frac{1}{2}, +1\frac{1}{2}$	2–12	21	0.09	5.2
		6–8	5	0.27	8.6
Quadratic	$-\frac{1}{2}, +\frac{1}{2}, -1\frac{1}{2}$	2–12	21	0.09	12.1
		6–8	5	0.27	7.9

9.2. *Precision of interpolation methods*

When interpolation procedures are applied to experimental data, deviations between predicted and experimental points can arise not only from model inaccuracies, but also from imprecisions in the data (*i.e.*, experimental error). To test the performance of interpolation methods in practice, we took a set of data from the literature [11]. We considered retention data for a series of ten aromatic amines recorded at ten different pH values at 0.5 unit intervals in the range $2 < \text{pH} < 6.5$. These data were selected for demonstration purposes. We do not wish to suggest that these data are more or less accurate or precise than other published data sets. Interpolation was performed using retention times rather than capacity factors. Interpolation algorithms such as eqn. 13 are equally valid in this case.

Table 3 summarizes the results obtained by various interpolation methods. The figures indicate the deviations between experimental and interpolated values. The first thing to notice is that the errors in Table 3 are small compared with those in Table 2. This is due to the fact that most of the data points are on the upper plateau of the sigmoidal curve. Also, relative errors in retention times are smaller than those in the corresponding capacity factors, especially for low values of k.

A second observation from Table 3 is that the inclusion of a data point $1\frac{1}{2}$ pH units below the target leads to better results than the inclusion of a point $1\frac{1}{2}$ pH units above the target. Generally (see below) a greater precision can be expected when the points needed for the interpolation algorithm are selected in the direction of the inflection point (*i.e.*, towards $\text{p}K_a$). This is because the greatest variation in retention occurs around the inflection point. The precision is highest when the magnitude of the variations in retention are much larger than the experimental error.

TABLE 3

DEVIATIONS BETWEEN EXPERIMENTAL DATA POINTS AND VALUES OBTAINED BY VARIOUS INTERPOLATION METHODS

Retention data on ten aromatic amines (mobile phase 65% buffer–35% water; 25 × 4.6 mm I.D. StableBond CN column; 25 mM buffers of sodium citrate ($\text{pH} \geq 4.0$) or potassium phosphate ($\text{pH} < 4$); temperature (35°C) at 0.5 unit intervals over the range $2 < \text{pH} < 6.5$ [12].

Interpolation	Data points (relative)	N	Average error	
			Absolute	%
Linear	$-\frac{1}{2}, +\frac{1}{2}$	80	0.23	2.4
Quadratic	$-\frac{1}{2}, +\frac{1}{2}, +1\frac{1}{2}$	60	0.25	2.6
Quadratic	$-\frac{1}{2}, +\frac{1}{2}, -1\frac{1}{2}$	60	0.16	1.6
Average quadratic	$-\frac{1}{2}, +\frac{1}{2}, +1\frac{1}{2}, -1\frac{1}{2}$	40	0.19	1.9
Sigmoidal	$-\frac{1}{2}, +\frac{1}{2}, +1\frac{1}{2}$	60	0.48	4.3
Sigmoidal	$-\frac{1}{2}, +\frac{1}{2}, -1\frac{1}{2}$	60	0.20	1.8

Unlike the situation in Table 2, the four-point piecewise quadratic procedure does not lead to consistently better results than the (best of the) three-point quadratic interpolations.

The results of the sigmoidal interpolation are not as good as might have been expected. Especially when a point $1\frac{1}{2}$ pH units above the target is included, the results are worse than with any of the other interpolation methods. Sigmoidal interpolation towards $\text{p}K_a$ does yield better results, but these are still inferior to those obtained by quadratic interpolation using the same data points.

An explanation for these observations can be found by inspecting graphical representations of the results of sigmoidal interpolation. Fig. 6 shows two representative examples. In each of

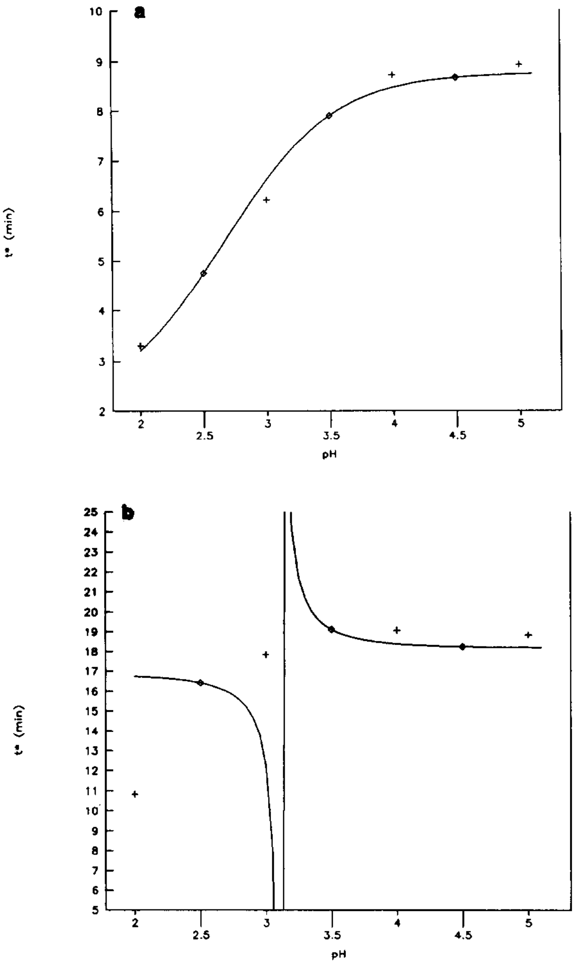

Fig. 6. Examples of the results obtained by sigmoidal interpolation through three experimental data points. ◇ = Points used for the interpolation; + = other available data points. (a) 4-Chloroaniline; (b) 3,4-dichloroaniline. Data taken from ref. 11. For experimental details, see Table 3.

the two graphs three data points (pH 2.5, 3.5 and 4.5, indicated by diamonds) have been used to establish values for retention times (t_*) for all pH values in the range between 2 and 5. For most of this range interpolation occurs. At either extreme extrapolation takes place.

The results of the sigmoidal interpolations (drawn line) in Fig. 6a are seen to provide a good representation of the experimental data. The data points not used for the interpolation (indicated by +) fall close to the line. Fig. 6b provides a different perspective. The resulting curve is seen to be hyperbolic (with a vertical asymptote around pH 3.1) rather than sigmoidal and the additional experimental data turn out to be nowhere near the calculated curve. Many examples of such behaviour were encountered when applying sigmoidal interpolation to the data set in question.

Qualitatively, the behaviour illustrated in Fig. 6b can be understood easily. The interpolation procedure calculates a curve through three data points, which is not a fitted line but an exact solution. In the case of Fig. 6b the central data point used for the interpolation (t_2) is the highest of the three values. This implies that a sigmoidal curve as illustrated in Fig. 1 cannot be fitted through the data. The "line" that does obey eqn. 3 and passes through the three data points is indeed hyperbolic. The value found for K_a is such that the denominator of eqn. 3 approaches zero around pH 3.1. Clearly, there are situations in which the use of sigmoidal interpolation should be avoided. Two such situations have been identified above, *viz.* (i) when t_2 is the highest (or lowest) of the three experimental values and (ii) when eqn. 15 yields a value for K_a that is negative and the magnitude of which is within the range of χ ($10^{-\mathrm{pH}}$) values studied ($\chi_{\mathrm{low}} \leq -K_a \leq \chi_{\mathrm{high}}$).

A more general indication of the risks involved in applying sigmoidal interpolation can be obtained by considering the error in k_* (or t_*) caused by experimental errors in the data points. By derivatizing eqn. 13 towards k_1, k_2 and k_3 the following equations can be obtained:

$$\Delta k_* = \frac{(1+\theta)(k_2-k_3)^2}{[\theta k_1-(1+\theta)k_2+k_3]^2}\cdot \Delta k_1 \tag{17}$$

$$\Delta k_* = \frac{-\theta(k_1-k_3)^2}{[\theta k_1-(1+\theta)k_2+k_3]^2}\cdot \Delta k_2 \tag{18}$$

and

$$\Delta k_* = \frac{\theta(1+\theta)(k_1-k_2)^2}{[\theta k_1-(1+\theta)k_2+k_3]^2}\cdot \Delta k_3 \tag{19}$$

Eqns. 17–19 predict the effects of small errors in the experimental data on interpolated values.

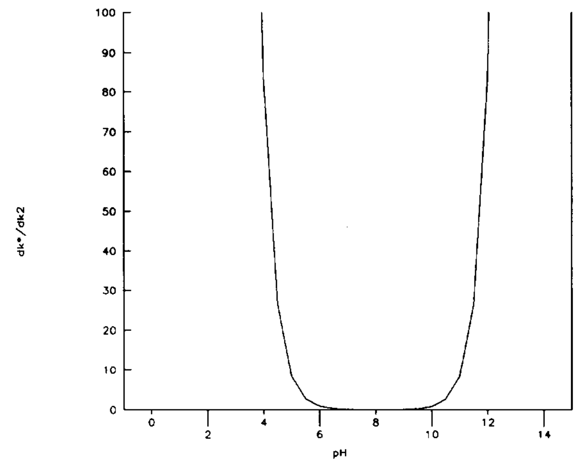

Fig. 7. Derivative function by which errors in the central interpolation point (k_2) are multiplied to obtain errors in the result (k_*) according to eqn. 18.

The effects of errors in the central point (eqn. 18) turn out to most severe. Fig. 7 provides an illustration for the case of sigmoidal interpolation using data points at $\mathrm{pH}_* + 0.5$, $\mathrm{pH}_* - 0.5$, and $\mathrm{pH}_* - 1.5$ on the curve shown in Fig. 1. The vertical axis refers to the derivative $\mathrm{d}k_*/\mathrm{d}k_2$, *i.e.*, the factor relating the two errors in eqn. 18. For $\mathrm{pH}_* = 8$ the smallest errors are obtained. When substituting the data point at $\mathrm{pH}_* + 1.5$ for that at $\mathrm{pH}_* + 1.5$ the minimum shifts to pH_* 6. In either case, the most favourable range for interpolation is $\chi_2 < K_a < \chi_3$, *i.e.*, the inflection point is included in the interpolation range. Away from the inflection point, the error increases, taking on dramatic values on either side of the curve. The values of over 100 on either side of the curve in Fig. 7 imply that a small experimental error in k_2 of 0.01 unit result in an error in the predicted (interpolated) values of k_* of more than a full unit. Clearly, sigmoidal interpolation based on three data points can be a "frustrating experience" [11] if it is not applied with care. The process should be restricted to situations in which pH has a large effect on retention. Moreover, an automated system should contain built-in checks, "warning flags" [11] and, ideally, "self-correction" procedures in the form of fall-back options (*e.g.*, changing from sigmoidal to quadratic interpolation) once problems have been diagnosed.

9.3. Regression analysis

One possible way of reducing the effects of experimental errors on predicted capacity factors is to increase the number of data points and to broaden the pH range covered. Clearly, the hyperbolic function shown in Fig. 6b will not result from a process in which all seven data points shown in this figure are used to fit a sigmoidal curve. The simultaneous use of more data points is not by definition inefficient. For example, when data points at different mobile phase compositions (*e.g.*, different methanol-to-water ratios) are available, the simultaneous use of all these points will help avoid the occurrence of artifactual asymptotes. When twelve data points (four pH values at three compositions) are used, the effect is not observed [8,13,20]. However, when interpolating first in one direction (pH or composition) and then in the other, the danger persists [22].

The use of genetic algorithms may also be advantageous, because in this case a sensible search area has to be defined in terms of minimum and maximum values for each coefficient. The occurrence of asymptotes in the pH range studied can thus easily be avoided.

10. CONCLUSIONS

The variation of retention with pH for monoprotic acids and bases is typically described by a sigmoidal function. However, a number of causes of deviation from this simple model can be identified. Models for multiprotic acids and bases and for zwitterionic solutes still need to be investigated.

When linear or quadratic interpolation is used to approximate a sigmoidal function, significant errors may be introduced. For quadratic (and sigmoidal) interpolations it turns out to be advantageous to select data points stretching towards (and ideally beyond) the inflection point ($\mathrm{p}K_a$), rather than away from it.

Sigmoidal interpolation using three experimental data points can easily be performed using simple equations. However, for real data it is not a robust process. The effect of experimental errors can be dramatic, especially in regions where pH has little effect on retention, *i.e.*, for neutral solutes or on the first parts of a sigmoidal curve. In that case quadratic interpolation using data points shifted towards pK_a is the preferred process.

The availability and simultaneous use of more data points than the required minimum of three greatly increases the reliability of retention modelling. When considering simultaneous variations of pH and eluent composition, all data points must ideally be used for building a single reliable model describing retention as a function of both parameters.

11. ACKNOWLEDGEMENT

We acknowledge the contribution of Sinéad Murray to the construction of Fig. 2.

REFERENCES

1 P.J. Schoenmakers, *Optimization of Chromatographic Selectivity; a Guide to Method Development*, Elsevier, Amsterdam, 1986.

2 M. Morvai, V. Fábián and I. Molnár-Perl, *J. Chromatogr.*, 600 (1992) 87.

3 Cs. Horváth, W. Melander and I. Molnár, *Anal. Chem.*, 49 (1977) 142.

4 G.K.C. Low, Á. Bartha, H.A.H. Billiet and L. de Galan, *J. Chromatogr.*, 478 (1989) 21.

5 R.G. Bates, *Determination of pH, Theory and Practice*, Wiley, New York, 1973.

6 M. Paabo, R.A. Robinson and R.G. Bates, *J. Am. Chem. Soc.*, 87 (1965) 415.

7 D.B. Rorabacher, W.J. McKellar, F.R. Shu and S.M. Bonavita, *Anal. Chem.*, 43 (1971) 561.

8 R.M. Lopes Marques and P.J. Schoenmakers, *J. Chromatogr.*, 592 (1992) 157.

9 R.K. Iler, *The Chemistry of Silica. Solubility, Polymerization, Colloid and Surface Properties, and Biochemistry*, Wiley, New York, 1979, Ch. 6.

10 P.J. Schoenmakers, S. van Molle, C.M.G. Hayes and L.G.M. Uunk, *Anal. Chim. Acta*, 250 (1991) 1.

11 J.A. Lewis, D.C. Lommen, W.D. Raddatz, J.W. Dolan, L.R. Snyder and I. Molnar, *J. Chromatogr.*, 592 (1992) 183.

12 C. Herrenknecht, D. Ivanovic, E. Guernet-Nivaud and M. Guernet, *J. Pharm. Biomed. Anal.*, 8 (1990) 1071.

13 P.J. Schoenmakers, N. Mackie and R.M. Lopes Marques, *Chromatographia*, 35 (1993) 18.

14 P.J. Schoenmakers, H.A.H. Billiet and L. de Galan, *J. Chromatogr.*, 205 (1981) 13.

15 P.J. Schoenmakers, H.A.H. Billiet, R. Tijssen and L. de Galan, *J. Chromatogr.*, 149 (1978) 519.

16 T. Hamoir, M. de Smet, H. Pirijns, P. Conti, N. Vandendriesche, D.L. Massart, F. Maris, H. Hindriks and P.J. Schoenmakers, *J. Chromatogr.*, 589 (1992) 31.

17 J.P. Westlake, B.W. King and P. Meyers, presented at the *19th International Symposium on Chromatography, Aix-en-Provence, September 23–28, 1992.*

18 R.J. Lynch and C. Measures, presented at the *Pittsburgh Conference, New Orleans, Marach 9–13, 1992*, Paper 463P.

19 R.M. Lopes Marques, P.J. Schoenmakers, C.B. Lucasius and G. Kateman, presented at the *19th International Symposium on Chromatography, Aix-en-Provence, September 23–28, 1992.*

20 R.M. Lopes Marques, P.J. Schoenmakers, C.B. Lucasius and L. Buydens, *Chromatographia*, 36 (1993) 83.

21 C.B. Lucasius and G. Kateman, *J. Chemometr. Intell. Lab. Syst.*, 19 (1993) 1.

22 J.A. Lewis, J.W. Dolan, L.R. Snyder and I. Molnar, *J. Chromatogr.*, 592 (1992) 197.