



# VARIANCE REDUCTION IN ANALYTICAL CHEMISTRY

New Numerical Methods in Chemometrics and  
Molecular Simulation

Magnus Åberg

Doctoral Thesis  
Stockholm University  
Department of Analytical Chemistry

2004

Akademisk avhandling framlägges för avläggande av filosofie doktorsexamen vid Stockholms Universitet och offentligen försvaras i Magnélisalen, Kemiska övningslaboratoriet, Svante Arrhenius väg 12, fredagen den 3 december 2004, kl. 10:00. Avhandlingen kommer att försvaras på engelska.

ISBN 91-7265-968-8

© Magnus Åberg, 2004

Intellecta DocuSys AB, Sollentuna 2004

# Abstract

This thesis is based on five papers addressing variance reduction in different ways. The papers have in common that they all present new numerical methods.

Paper I investigates quantitative structure-retention relationships from an image processing perspective, using an artificial neural network to preprocess three-dimensional structural descriptions of the studied steroid molecules.

Paper II presents a new method for computing free energies. Free energy is the quantity that determines chemical equilibria and partition coefficients. The proposed method may be used for estimating, *e.g.*, chromatographic retention without performing experiments.

Two papers (III and IV) deal with correcting deviations from bilinearity by so-called peak alignment. Bilinearity is a theoretical assumption about the distribution of instrumental data that is often violated by measured data. Deviations from bilinearity lead to increased variance, both in the data and in inferences from the data, unless invariance to the deviations is built into the model, *e.g.*, by the use of the method proposed in III and extended in IV.

Paper V addresses a generic problem in classification; namely, how to measure the goodness of different data representations, so that the best classifier may be constructed.

Variance reduction is one of the pillars on which analytical chemistry rests. This thesis considers two aspects on variance reduction: before and after experiments are performed. Before experimenting, theoretical predictions of experimental outcomes may be used to direct which experiments to perform, and how to perform them (papers I and II). After experiments are performed, the variance of inferences from the measured data are affected by the method of data analysis (papers III–V).

**Key words:** chemometrics, pulse-coupled neural networks, peak alignment, class separability, molecular dynamics, Monte Carlo, expanded ensembles, free energy.

ISBN 91-7265-968-8



# Preface

This thesis is based on work carried out as a PhD student between February 2000 and October 2004 at the Department of Analytical Chemistry, Stockholm University, Sweden. The thesis is based on the following publications and manuscripts:

- I. Pre-processing of three-way data by pulse-coupled neural networks—an imaging approach  
K. M. Åberg and S. P. Jacobsson  
*Chemom. Intell. Lab. Syst.* **57**, 25–36 (2001).
- II. Determination of solvation free energies by adaptive expanded ensemble molecular dynamics  
K. M. Åberg, A. P. Lyubartsev, S. P. Jacobsson, and A. Laaksonen  
*J. Chem. Phys.* **120**, 3770–3776 (2004).
- III. Peak alignment using reduced set mapping  
R. J. O. Torgrip, M. Åberg, B. Karlberg, and S. P. Jacobsson  
*J. Chemometrics* **17**, 573–582 (2003).
- IV. Extensions to peak alignment using reduced set mapping and classification of LC-UV data from peptide mapping  
K. M. Åberg, R. J. O. Torgrip and S. P. Jacobsson  
Submitted to: *J. Chemometrics* (2004).
- V. A measure of class separation  
K. M. Åberg and S. P. Jacobsson  
Submitted to: *J. Chemometrics* (2004).

In Paper I the author was responsible for developing the ideas, implementing the code, modelling and analyzing the data, and writing the paper. In Paper II the author was responsible for the idea, developing the algorithm, doing the MD simulations, and writing most of the paper. In Paper III the author was responsible for everything involving the breadth first search, including the idea, its implementation, and writing part of the paper. In Paper IV the author was responsible for the ideas, implementing the code, modelling and analyzing the data, and writing the paper. In Paper V the author was responsible for the idea, developing the algorithm, and writing the paper.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Chemometrics</b>	<b>5</b>
2.1	Bilinear data . . . . .	6
2.2	Rank . . . . .	6
2.3	Principal components analysis . . . . .	7
2.4	Shrinkage Regression . . . . .	9
2.4.1	Ridge regression and the variance–bias dilemma . . . . .	9
2.4.2	Principal component regression . . . . .	10
2.4.3	PLS . . . . .	10
2.4.4	Continuum regression . . . . .	11
2.5	Classification . . . . .	17
2.5.1	Nearest neighbour classification . . . . .	18
2.5.2	Linear discriminant analysis . . . . .	18
2.5.3	SIMCA . . . . .	18
2.5.4	PLS-DA . . . . .	19
2.5.5	Nonlinear classifiers . . . . .	19
2.5.6	Class separability measures . . . . .	19
2.6	Model validation . . . . .	21
2.7	Pulse-coupled neural networks . . . . .	23
2.8	Peak alignment . . . . .	25
<b>3</b>	<b>Computation of solvation free energies</b>	<b>29</b>
3.1	Statistical mechanics . . . . .	31
3.2	Forcefields: describing the interaction between atoms . . . . .	32
3.2.1	Bonds . . . . .	33
3.2.2	Angles . . . . .	33
3.2.3	Torsional angles . . . . .	34
3.2.4	Non-bonded interactions . . . . .	34
3.3	Monte Carlo . . . . .	36
3.4	Molecular dynamics . . . . .	37
3.5	Expanded ensemble molecular dynamics . . . . .	38

<b>4</b>	<b>Discussion</b>	<b>41</b>
4.1	Paper I: The PCNN as a preprocessor for QSRR . . . . .	41
4.2	Paper II: Adaptive expanded ensembles . . . . .	43
4.3	Papers III and IV: Solutions to the peak alignment problem . . .	45
4.3.1	Complexity of PARS and DTW . . . . .	45
4.3.2	The target question . . . . .	46
4.4	Paper V: A measure of class separation . . . . .	47
<b>5</b>	<b>Conclusions</b>	<b>49</b>

# Chapter 1

## Introduction

The aim of applied analytical chemistry is to answer a specific question of some kind. The nature of the question differs—it can be to qualitatively determine the contents of a sample of biological origin or quantify the amount of a certain substance in indoor air.

A central concept in analytical chemistry is the so-called analytical chain. The first link in this chain is the sampling procedure, including a strategy for sampling and the actual taking of a sample. The subsequent links are defined as the steps and procedures applied to the sample in order to make the desired assessment and provide an answer to the question at hand. The answer should be more than a single figure describing, for instance, the concentration of sugar in natural juices. It must be delivered with some measure of certainty, or rather uncertainty, commonly expressed using standard deviations and confidence intervals. The chain is no stronger than its weakest link; the same holds for an analytical method. If you do not have control over all the steps from sampling to data evaluation, you cannot say anything about the uncertainty of the answer. And if you are uncertain about the uncertainty, be assured, it is huge!

Researchers in analytical chemistry do not generally approach the above questions other than for demonstration purposes, but are more devoted to the task of developing and refining methods and instruments that the practitioner can use to get more reliable answers, preferably in shorter a time and with smaller amounts of sample. Researchers deal with different links of the analytical chain—sometimes a single link, at other times the complete chain. There exists a multitude of instrumental techniques and methods, most of them developed with one aim in mind—variance reduction. Björklund *et al.*<sup>1</sup> provide an excellent example of method improvement for variance reduction in the analysis of the fully brominated di-phenyl ether, *deca*-BDE, a flame retardant and environmental pollutant.

The four most common measures of goodness of an analytical method are the limit of detection (LOD), the repeatability, the reproducibility, and the limit of quantification (LOQ). The limit of detection (LOD) used to have a simple

definition—three times the standard deviation of the noise, but now has a more detailed and intricate definition. ISO<sup>2,3</sup> and IUPAC<sup>4</sup> provide LOD definitions and their similarities and differences are discussed by Currie.<sup>5</sup> The IUPAC orange book<sup>4</sup> defines repeatability as the standard deviation for “independent results obtained with the same method on identical test material, under the same conditions (same operator, same apparatus, same laboratory and after short intervals of time).” Reproducibility has a similar definition with the difference that the results are to be obtained under *different* conditions and the differences should be specified. LOQ is commonly defined as the level of the desired quantity where it can be determined with a relative standard deviation of ten percent. It is all about variance.

Selectivity is another key word in analytical chemistry. Increasing the selectivity of a method or detector means reducing the overlap between the signal of interest and other signals, which might vary uncontrollably. This, of course, decreases the uncertainty (variance) in the final answer. But do not pursue minimal variance heedlessly, pursue low enough variance: the ultimate goal should be to answer efficiently with tolerable uncertainty, not as accurately as possible. To quote a professor at the department: “It’s not about doing things right, it’s about doing the right things.” Even so, variance reduction retains its importance. Variance criteria must be met in high throughput analysis as well as in miniaturization.

Chemometrics is where (analytical) chemistry and statistics meet. The focus in chemometrics is on the use of more sophisticated mathematical and statistical models, not just the usual univariate regression and statistics based on the assumption of normally distributed errors. The outspoken aim of chemometrics is variance reduction. The chemometrician usually tries to minimize the expected squared prediction error; that is, find the model and the parameters thereof, which minimize  $\text{Var}(\hat{y} - y)$ , where  $y$  is the quantity of interest and  $\hat{y}$  is the model estimate. This minimization is about using all the information present in the measured data.

Another form of variance reduction is to perform as few experiments as possible. The total variance is roughly proportional to the number of experiments performed, “ $\text{Var}(\text{Study}) \propto \text{tr}(\mathbf{X}\mathbf{X}^T)$ ,”  $\mathbf{X}$  being the measured data. *Design of experiments* is concerned with obtaining maximum information from as few samples as possible. At the next level of abstraction the experiments can be performed *in silico*. In other words, by performing relevant computer calculations and drawing conclusions from them, we may reduce even further the number of actual experiments to include only those we believe important. When doing “relevant computer calculations” we are treading on the very outskirts of analytical chemistry where it intersects with physical chemistry. The objective, however, still lies within analytical chemistry. The transition between the two disciplines is smooth, for instance via quantitative structure-activity relationships (QSAR), considered to be a part of chemometrics. An example of QSAR modelling is the use of molecular descriptors to predict the binding affinity of ligands to an enzyme, or the activity of an enzyme in the presence of potential ligands. The

molecular descriptors are often based on quantum chemical calculations, quantum chemistry being a core part of physical chemistry. The disciplines also meet where we seek detailed physical and/or mechanistic explanations for the response of a detector, or for the retention mechanisms of a chromatographic media. The latter problem may be studied by means of statistical mechanics. In statistical mechanics, variance reduction is a key word. When high-dimensional integrals are to be evaluated, variance reduction techniques are needed to make the calculations feasible. The two dominating techniques are Monte Carlo and molecular dynamics.

My contribution to variance reduction in analytical chemistry lies in data preprocessing within the field of chemometrics. The thesis includes two fundamentally different fields of study: Paper I examines the use of pulse-coupled neural networks for quantitative structure-retention relationships (QSRR), Paper III and Paper IV study the benefits of peak alignment prior to multivariate modelling with chemometric standard methods. Paper V is related to Paper IV and deals with the objective function for variance reduction in a classification context.

Paper II may seem like an outlier next to the other four papers. We started a QSRR project where chromatographic retention was to be predicted from first principles rather than from a regression model, which is the usual procedure. So the step from Paper I was not all that far. However, the project turned out to be less straightforward and more complicated than we initially thought. Paper II addresses the first major issue that we came across in our molecular simulations.

The thesis is organized in two theoretical chapters discussing chemometrics (Chapter 2) and the theory of solvation-free energies (Chapter 3). These chapters are followed by a discussion of the individual articles (Chapter 4) and some general conclusions and considerations (Chapter 5).



## Chapter 2

# Chemometrics

As in all experimental sciences and in analytical chemistry in particular, researchers are dependent on statistical measures for reporting their results. The field of chemometrics is founded on the border between chemistry and (mathematical) statistics. Chemometrics can be said to encompass all the statistics used in chemistry. Multivariate statistics is the core of chemometrics, especially experimental design, multivariate optimization, and multivariate regression. Among the new statistical methods that have seen the light of day within the chemometrics community, partial least squares is the most prominent example. The statistical methods are often not directly applicable to the measured data. Therefore, a lot of effort is put into research on data pretreatments, transformations of the data that will enable the chemometrician to use a standard multivariate regression method.

An introductory book on the subject is *Chemometrics: Data Analysis for the Laboratory and the Chemical Plant* by Brereton.<sup>6</sup> An excellent book of reference character covering a wide range of subjects is the *Handbook of Chemometrics and Qualimetrics* in two volumes A<sup>7</sup> and B.<sup>8</sup> More specialized books include *Multivariate Calibration* by Martens and Næs<sup>9</sup> and *A User's Guide to Principal Components* by Jackson.<sup>10</sup>

The fundamental assumption in modelling is that the measured data consist of information and noise:

$$\text{data} = \text{information} + \text{noise},$$

the mathematical formulation being:

$$\mathbf{x} = f(y) + \varepsilon. \quad (2.1)$$

On occasions, when the situation is more complicated, the relationship is written more generally as  $\mathbf{x} = f(y, \varepsilon)$ . These equations refer to the measurement process. As data analysts we are interested in the inverse relationship: how to determine  $y$  (information) given  $\mathbf{x}$  (data). The rest of this chapter will be devoted to the issues of multivariate modelling, assuming that Eq. (2.1) holds.

## 2.1 Bilinear data

In chemometrics a matrix consisting of first-order data, *e.g.*, spectra, from several samples is almost exclusively represented by variables in columns and samples as rows. What is perhaps the most common situation nowadays is that the number of variables exceeds the number of samples; the problem is underdetermined. When this is the case, multiple linear regression (MLR)\* cannot be used directly. A common and old-fashioned approach is to select a small number of variables (wavelengths if the measured data are spectrometric) and apply MLR to find the regression of  $y$  on  $\mathbf{x}$ . By doing this, we dispose of most of our data and the information contained therein. The chemometric approach is to try to utilize all of the observed data and to extract the information they contain.

Consider the issue of determining the concentration of  $k$  compounds in water using UV/Vis spectroscopy. Assume further that there is no selective wavelength for any of the compounds. A natural approach would be to design a set of  $m$  calibration samples with varying amounts of the analytes and measure the absorbance spectrum,  $\mathbf{x}$  (at  $n$  wavelengths), for each sample. If the Beer-Lambert law is obeyed, each sample may be described as a sum of the pure component spectra,  $\mathbf{s}$ , times their concentration,  $c$ :

$$\mathbf{x}_i = c_a \mathbf{s}_a + c_b \mathbf{s}_b + \dots + c_k \mathbf{s}_k. \quad (2.2)$$

This equation may be written for all samples simultaneously in matrix form:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^T, \quad (2.3)$$

where  $\mathbf{S}$  is a  $(n \times k)$  matrix of the pure spectra and  $\mathbf{C}$  is  $(m \times k)$  containing the concentrations of all components in all samples. Matrices like  $\mathbf{X}$  are called bilinear. In chemometrics it is customary to structure the data matrix,  $\mathbf{X}$ , with samples as rows and variables as columns. In other disciplines  $\mathbf{X}$  may be structured the other way around. Exploiting the bilinearity of  $\mathbf{X}$  is a major concern for the chemometrician, the two main applications being regression and classification.

## 2.2 Rank

The number of independent rows or columns in a matrix,  $\mathbf{X}$ , is called the rank of the matrix. Any experimentally measured data matrix has full mathematical rank. This fact is due to instrumental noise. The rank is then equal to the number of rows or columns, whichever is the smaller. If data free of noise could be obtained, then the rank would be equal to the number of chemical phenomena that vary between samples (or possibly the number of phenomena minus one, if

---

\*Multiple linear regression is also known by the names ordinary least squares (OLS) and inverse least squares (ILS). These are not to be confused with classical least squares (CLS) which solves the problem  $\mathbf{X} = \mathbf{K}\mathbf{Y} + \mathbf{E}$ ,  $\mathbf{K}$  being, *e.g.*, pure spectra

the measured system exhibits closure). An example of closure is when the sum of all the variables is constant from sample to sample; all the constituents of a mixture add up to 100%. The number of phenomena is called the chemical rank and this is important in multivariate regression. For real data, the chemical rank is determined as the number of large eigenvalues of the  $\mathbf{X}^T\mathbf{X}$ . The eigenvalues tend to fall off rapidly at first and smoothly thereafter. It is a simple task to say that there are at least  $l$  phenomena present. The true chemical rank can be more difficult to specify; small effects may be difficult to discern from noise. A major concern of chemometrics is to estimate the chemical rank.

Instrumental artefacts are common sources of misspecification of the chemical rank. The most trivial example of this is where there is a time shift that differs between samples in chromatographic data. The position of a peak varies between samples and this destroys the bilinear structure of the data. The structure no longer satisfies Eq. (2.3). A single peak, present in all samples, which could be described by a single latent variable, were it not for the varying shift, may need several latent variables to be described equally well. Imagine what this can do to a full chromatogram with tens, hundreds, or even thousands of peaks. A deviation from bilinearity increases the apparent rank, and this is degenerative to multivariate modelling.

## 2.3 Principal components analysis

The history of the method dates back to the early 19th century. Cauchy is considered to have been the first to derive principal components analysis (PCA), which he did as early as 1829.<sup>6</sup> The next time PCA makes an appearance is in two papers by Adcock from 1877 and 1878. In the 1877 paper<sup>11</sup> he derives the one- and two-dimensional subspaces for overdetermined systems of point measurements in three dimensions. In the following paper<sup>12</sup> he uses the term principal axis when computing the regression line  $y = ax + b$  with errors in both  $x$  and  $y$ . The solution is optimal in a least squares sense, where the residuals are measured orthogonal to the regression line. Pearson<sup>13</sup> commented on the strange habit of assuming that only  $y$ , the dependent variable, is prone to error, while the independent variable,  $x$ , is free of error, while we know that this is not the case. Next, he goes on to give a description which is more easily recognized as PCA than the earlier ones. The method was further developed by Hotelling in 1933<sup>14</sup> and used in a context similar to the way it is often used today. Examples of the use of PCA in chemometrics can be found in any textbook, see, *e.g.*, Refs.<sup>6-8</sup> The book by Jackson<sup>10</sup> is perhaps the most comprehensive source of information about PCA.

Consider the matrix  $\mathbf{X}$ , a data table with  $m$  samples and  $n$  variables. PCA relates to the second statistical moment of  $\mathbf{X}$ , which is proportional to  $\mathbf{X}^T\mathbf{X}$ . PCA partitions  $\mathbf{X}$  into two matrices  $\mathbf{T}$  and  $\mathbf{P}$ , which are called scores and loadings respectively, such that:

$$\mathbf{X} = \mathbf{TP}^T \quad (2.4)$$

The loadings matrix contains the eigenvectors of  $\mathbf{X}^T\mathbf{X}$  ordered by their eigenvalues with the largest first and in descending order. If  $\mathbf{P}$  has the same rank as  $\mathbf{X}$ , *i.e.*,  $\mathbf{P}$  contains the eigenvectors to all non-zero eigenvalues, then  $\mathbf{T} = \mathbf{X}\mathbf{P}$  is a rotation of  $\mathbf{X}$ . The first loading vector,  $\mathbf{p}_1$ , points in the direction that minimizes the orthogonal distances from the samples to their projection onto this vector. This means that the first column of  $\mathbf{T}$  captures the largest possible sum of squares as compared to any other direction in  $\mathbf{R}^n$ .

In statistics it is customary to center the variables in the matrix  $\mathbf{X}$  prior to using PCA. This makes  $\mathbf{X}^T\mathbf{X}$  proportional to the variance-covariance matrix. The first principal axis is then the direction in which the data have the largest spread. This property of PCA opens up a possibility for data compression and noise suppression. When only the  $k$  first loading vectors are used  $\mathbf{T}_k = \mathbf{X}\mathbf{P}_k$  is a projection onto the subspace of  $\mathbf{R}^n$  with the smallest residual in least squares sense. The data can be reconstructed as  $\hat{\mathbf{X}} = \mathbf{T}_k\mathbf{P}_k^T$ . Noise suppression is achieved with little loss of information if  $k$  equals the chemical rank. Since the same phenomena are measured  $m$  times,  $m - k$  samples contribute to the noise smoothing.

There is a considerable similarity between Eq. (2.3) and Eq. (2.4). With the same number of components and without mean centering of the variables, the PCA loadings will span the same space as  $\mathbf{S}$ . If the  $\mathbf{X}$  of Eq. (2.3) is decomposed using PCA, the pure spectra  $\{\mathbf{s}_i\}_{i=1}^k$  can be found as linear combinations of the loadings,  $\{\mathbf{p}_i\}_{i=1}^k$ . This can be written as  $\mathbf{s}_i = \mathbf{P}\mathbf{a}$ , where the coefficients  $\mathbf{a}$  may be found from the PCA score space  $\mathbf{T}$ . Imagine the scores inscribed in a pyramid with  $k$  edges with the top of this pyramid at  $\mathbf{t} = \mathbf{c} = \mathbf{0}$ . If there are samples in which only one component has nonzero concentration, then the coefficients  $\mathbf{a}$  are determined by the equations of the edges of the pyramid. The samples may be consecutive spectra from a liquid chromatographic system coupled to a diode array detector (LC-DAD). In this context PCA can be used to mathematically resolve overlapping chromatographic peaks. Once the pure spectra are known, the concentration profiles of the individual components can easily be computed. This subject is known as multivariate curve resolution (see for instance Ref.<sup>15</sup>). PCA plays an important part in many other areas or subjects, examples of which include exploratory data analysis,<sup>16</sup> classification, variable decorrelation prior to the use of neural networks,<sup>17</sup> analysis of sensory data,<sup>8</sup> data compression,<sup>18</sup> and noise reduction,<sup>19</sup> to mention a few.

The scores and loadings can be computed using many different algorithms, of which the power method, eigenvalue decomposition, and singular value decomposition (based on a generalization of eigenvalues to non-square matrices) are the most common (see Ref.<sup>20</sup> for details). The number of components can be chosen by examining the eigenvalues or, for instance, considering the residual error from cross-validation.

In essence, PCA is nothing but a rotation of the coordinate system, though a very useful one. It is perhaps the most frequently used method in chemometrics. Every analyst of multivariate data uses it and comes up with new applications all the time. SciFinder<sup>21</sup> returns about 17 000 hits for the search words “principal

components analysis.” This figure is to be compared to the number of answers to “linear calibration” (21 000), “linear regression” (31 000), and “UV-VIS spectroscopy” (74 000).

## 2.4 Shrinkage Regression

Consider the model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}. \quad (2.5)$$

The formal least squares solution using multiple linear regression (MLR) is  $\hat{\mathbf{b}}_{MLR} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ . If the problem is underdetermined,  $\mathbf{X}^T\mathbf{X}$  is singular and there is no *unique* solution,  $\hat{\mathbf{b}}_{MLR}$ , but a solution space of  $\hat{\mathbf{b}}_{MLR}$ -vectors in which  $\mathbf{e} = \mathbf{0}$ , *i.e.*, the problem is solved exactly. Since underdetermined regression problems can always be solved exactly using a linear model,<sup>†</sup> yielding non-sensical results, there is a need for methods that can handle this situation. One such class of methods is called shrinkage regression methods. The word shrinkage refers to the fact that  $\|\hat{\mathbf{b}}_{SR}\| < \|\hat{\mathbf{b}}_{MLR}\|$  and that  $\text{Var}(\hat{\mathbf{b}}_{SR}) < \text{Var}(\hat{\mathbf{b}}_{MLR})$ , where the subscript *SR* indicates a shrinkage method. A direct effect of shrinkage methods is that  $\text{Var}(\hat{y}) < \text{Var}(y)$ ; hence the predictions are shrunk towards the mean,  $E(y)$ . These methods usually have parameters that need to be determined; hence an element of model selection enters into the regression modelling (see Section 2.6 for a discussion of this topic).

### 2.4.1 Ridge regression and the variance–bias dilemma

The simplest way to stabilize the matrix inverse is to add a constant to the diagonal. This is the basis of ridge regression. The formal solution for the regression coefficients is

$$\hat{\mathbf{b}}_{RR} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \quad (2.6)$$

where  $\lambda$  is the *ridge coefficient*. The regression coefficients are biased as a result of the use of the ridge coefficient. As their variance decrease, however, the model becomes more stable with respect to the prediction error.

The variance–bias dilemma is a general problem in multivariate regression. The model bias can only be reduced at the expense of increased model variance and vice versa. The expected prediction error is the sum of two components:  $E((\hat{y} - y)^2) = (\text{model bias})^2 + (\text{model variance})$ . The better the assumptions about the data are, the better a model can be expected to perform. A suitable transform  $\mathbf{Z} = f(\mathbf{X})$  prior to the regression may simultaneously reduce the model bias and the model variance. The transform can be either linear or nonlinear. The regression model becomes  $\mathbf{y} = \mathbf{Z}\mathbf{b}$ .

---

<sup>†</sup>An important special case when an underdetermined problem cannot be solved exactly is when two samples have the exact same location in predictor space, *i.e.*,  $\mathbf{x}_j = \mathbf{x}_k$ , while their responses differ,  $y_j \neq y_k$ . This will, of course, never occur when  $\mathbf{x}$  is composed of measured data.

### 2.4.2 Principal component regression

As its name implies, this method is closely related to principal components analysis. In essence, the method is just multiple linear regression of PCA scores on  $\mathbf{y}$  using a suitable number of principal components. The formal solution may be written as:

$$\hat{\mathbf{b}}_{PCR} = \mathbf{P}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y}. \quad (2.7)$$

New samples are predicted as  $\hat{\mathbf{y}} = \mathbf{x}^T\hat{\mathbf{b}}_{PCR}$ . In principal components regression (PCR) the matrix inverse is stabilized in an altogether different way than in ridge regression. The scores vectors (columns in  $\mathbf{T}$ ) of different components are orthogonal. The product between two score vectors can be written as  $\mathbf{t}_i^T\mathbf{t}_j = \delta_{ij}\lambda_i$ , where  $\delta_{ij}$  is the Kronecker delta and  $\lambda_i$  the  $i$ :th largest eigenvalue of  $\mathbf{X}^T\mathbf{X}$ . PCR uses a truncated inverse where only the scores corresponding to large eigenvalues are included. The main drawback of PCR is that the largest variation in  $\mathbf{X}$  might not correlate with  $\mathbf{y}$  and therefore the method may require the use of more latent variables than PLS (described below). More latent variables means a more complex model, and according to Occam's razor, or the parsimonious principle, one should choose the least complex of otherwise comparable models.

### 2.4.3 PLS

The acronym PLS stands for partial least squares or, as others prefer to put it, projection onto latent structures. The interpretation of the acronym is not unique and varies with the author. The method was first described in the early 1970s by Herman Wold (see, *e.g.*, Ref.<sup>22</sup>). Since its appearance, PLS has become enormously popular; it is used, not only in chemistry, but also in, *e.g.*, econometrics,<sup>23</sup> psychometrics,<sup>24</sup> and biometrics.<sup>25</sup> In practice, PCR is almost never used due to the excellent performance of PLS. PLS is a latent variable method, as the second interpretation of the acronym implies. It differs from PCR in the way the latent variables are chosen. In PCR the latent variables maximize the explained variance of  $\mathbf{X}$ , or equivalently:  $\max_{\mathbf{p}} \text{Var}(\mathbf{t})$ , with  $\|\mathbf{p}\| = 1$ , while PLS maximizes the covariance between the  $\mathbf{X}$  and  $\mathbf{y}$  data,  $\max_{\mathbf{w}} \text{Cov}(\mathbf{y}, \mathbf{t})$ , subjected to  $\|\mathbf{w}\| = 1$ . There are a number of algorithms that perform PLS regression. The NIPALS algorithm, as it was introduced in chemometrics by S. Wold and H. Martens in the early 1980s, is still used today. Its advantages are its speed and simplicity. However, if there is more than one response variable, it does not exactly maximize the covariance between  $\mathbf{X}$  and  $\mathbf{Y}$ . This was pointed out in 1993 by de Jong,<sup>26</sup> who in the same paper proposed an alternative PLS algorithm that fulfils the covariance criterion. His algorithm is called SIMPLS or Statistically Inspired Modification of PLS. The differences between NIPALS and SIMPLS are small from a practical point of view. Furthermore, NIPALS is not as efficient when the response is multivariate, and this has led to the development of more efficient PLS algorithms (often SVD based), among which the so-called kernel-PLS algorithms are worth mentioning. Lindgren and Rännar<sup>27</sup> reviewed the various algorithms in 1998. The most efficient algorithm depends on the

problem: the number of variables compared to the number of samples, whether  $\mathbf{Y}$  is multivariate or not, and how the validation is done.

Stepping through the NIPALS algorithm for a univariate response, *e.g.*, the protein content in samples of minced meat, and a multivariate predictor, in this case consisting of NIR spectra. The first task is to find the direction,  $\mathbf{w}$ , in the predictor space, along which  $\mathbf{X}$  has maximum covariance with  $\mathbf{y}$ . This is almost trivial; the weight vector  $\mathbf{w}$  is computed as:

$$\mathbf{w} = \mathbf{X}^T \mathbf{y} / \|\mathbf{y}^T \mathbf{X} \mathbf{X}^T \mathbf{y}\| \quad (2.8)$$

The scores are the values obtained when  $\mathbf{X}$  is projected onto  $\mathbf{w}$ :

$$\mathbf{t} = \mathbf{X} \mathbf{w} \quad (2.9)$$

The next step is just as easy: determine the regression coefficient of the model  $\mathbf{y} = \mathbf{t}b$  as

$$b = \mathbf{y}^T \mathbf{t} / (\mathbf{t}^T \mathbf{t}) \quad (2.10)$$

The PLS loadings are defined as:

$$\mathbf{p} = \mathbf{X}^T \mathbf{t} / (\mathbf{t}^T \mathbf{t}) \quad (2.11)$$

Eq. (2.11) looks very similar to Eq. (2.10) and, indeed,  $\mathbf{p}$  gives us the least squares reconstruction of  $\mathbf{X}$  given  $\mathbf{t}$ . So far everything is intuitive and not difficult at all. The first PLS component has been found. Next, the data are *deflated* before the next component can be computed:

$$\mathbf{y}_1 = \mathbf{y} - \mathbf{t}b \quad (2.12)$$

$$\mathbf{X}_1 = \mathbf{X} - \mathbf{t}\mathbf{p}^T \quad (2.13)$$

The step that makes PLS hard to understand is the deflation of  $\mathbf{X}$ . One would expect  $\mathbf{X}$  to be deflated as  $\mathbf{X}_1 = \mathbf{X} - \mathbf{t}\mathbf{w}^T$ . It seems natural to subtract what has already been used from  $\mathbf{X}$ . The problem with this alternative deflation is that the scores on the weight vector of the next PLS component do not become orthogonal to the previous scores. As a result, the second regression coefficient  $b_2$  cannot be determined independently of  $b_1$  and vice versa; hence the algorithm does not work. Strictly speaking, it is not necessary to update both  $\mathbf{X}$  and  $\mathbf{y}$  for the algorithm to give orthogonal PLS vectors.<sup>27</sup> The excellence of PLS regression has been established by its success in a wide variety of problems. The relationship between MLR, PLS, and PCR is analyzed and discussed in the following section.

#### 2.4.4 Continuum regression

Why we get a good predictor from maximizing the covariance between  $\mathbf{X}$  and  $\mathbf{y}$  is perhaps not obvious. By expanding the PLS criterion:

$$\text{Cov}(t, y) = \text{Var}(t) \text{Var}(y) \rho_{ty} \quad (2.14)$$

where  $\rho_{ty}$  is the correlation coefficient between  $t$  and  $y$ , we can identify the components thereof. Recall that the first latent variable of PCR is the direction

in which  $\mathbf{X}$  varies the most. PCR chooses the latent variable based on  $\mathbf{X}$  only and this is the same as maximizing  $\text{Var}(t)$ . Multiple linear regression, on the other hand, is only concerned with explaining the variance in  $\mathbf{y}$  and this is identical to maximizing  $\text{Var}(y)\rho_{ty}$ . Since MLR is not a latent variable method, only one score vector can be computed:  $\mathbf{t}_{MLR} = \mathbf{X}\hat{\mathbf{b}}_{MLR}/(\hat{\mathbf{b}}_{MLR}^T\hat{\mathbf{b}}_{MLR})^{1/2}$ . Stone and Brooks<sup>28</sup> parametrized the maximum covariance criterion of PLS [Eq. (2.14)] to produce a new method—continuum regression, which includes MLR, PLS, and PCR as special cases.

$$C = \text{Var}(t)^{\alpha/(1-\alpha)}\text{Var}(y)\rho_{ty} \quad (2.15)$$

At  $\alpha = 0$  we have MLR, while  $\alpha = \frac{1}{2}$  yields PLS, and PCR is found in the limit  $\alpha \rightarrow 1$ . The MLR part of Eq. (2.15) is the explained  $y$ -variance, *i.e.*,  $\text{Var}(y)\rho_{ty} \approx \hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}/(N-1)$ , assuming that the columns of  $\mathbf{X}$  are centered. Note that  $\hat{\mathbf{b}}$  can be expressed as  $\mathbf{w}\hat{b}$  where the latent variable is defined by the unit length vector  $\mathbf{w}$ . Following the same argument, we identify  $\text{Var}(t) \approx \mathbf{t}^T\mathbf{t}/(N-1)$ .

Two numerical examples will be used to illustrate the differences between the regression methods. Both examples use simulated data.

### Example 1

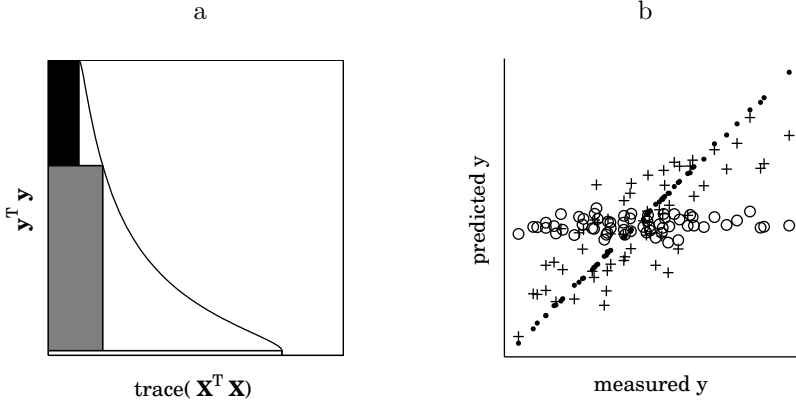
The first example serves to illustrate the differences between MLR, PLS, and PCR for an overdetermined problem. The predictor space,  $\mathbf{x}$ , is three-dimensional and the samples are independent and multinormal with  $\Sigma_{\mathbf{x}} = \text{diag}([1 \ 1 \ 9])$ . This means that the three elements of  $\mathbf{x}$  are independent as well. The number of samples is 60. The true relationship between  $\mathbf{x}$  and  $y$  is:

$$y = 3x_1 \quad (2.16)$$

The “measured” data are, of course, noisy in both  $\mathbf{x}$  and  $y$ , so after the true  $\mathbf{x}$ ’s are drawn and the true  $y$  are computed, white noise is added. Independent and normal random numbers with standard deviation 0.01 were added to  $\mathbf{X}$ . The random vector added to  $\mathbf{y}$  had a standard deviation of 0.1.

A geometrical interpretation of continuum regression is given in Figure 2.1 (a). The  $x$ -axis shows explained variance along the dimension  $\mathbf{w}$ . The  $y$ -axis shows explained  $y$  variance. The MLR solution is depicted as the black rectangle. It explains practically all the variance in  $y$ , while the explained  $\mathbf{x}$  variance is low. The grey rectangle is the one-component PLS solution and the white one near the bottom of the figure is the one-component PCR solution, which explains most of the  $\mathbf{x}$  variance but explains almost nothing of  $y$ . The continuum regression solutions lie on the solid curve.

Figure 2.1 (b) shows predicted vs. measured values of  $y$  for the three regression methods. Since the number of samples are much greater than the number of variables, we dispense with validation in this example. As to be expected, MLR is best at predicting  $y$ . In order for PLS and PCR to explain more of the variance in  $y$ , we need at least one extra latent variable. The effect of using two



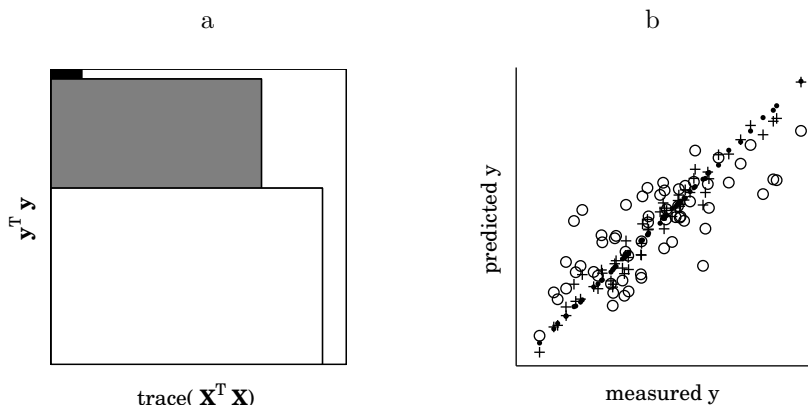
**Figure 2.1.** The first principal component of example 1. (a) Solutions that maximize  $C$  [Eq. (2.15)] lie on the solid curve. Three special cases are: MLR (black rectangle), PLS (grey rectangle), and PCR (white rectangle). (b) Predicted vs. measured values of  $y$  from MLR ( $\bullet$ ), PLS ( $+$ ), and PCR ( $\circ$ ).

latent variables is shown in Figure 2.2. The depicted solutions are the ones closest to the average solutions from 1000 simulations. The MLR solution cannot be improved and stays exactly the same. PLS explains nearly 70% of the  $\mathbf{X}$  data and is almost as good at explaining  $y$  as MLR. The second component is crucial for the PCR predictions. It explains about 50% of the variance in  $y$ . The degree of explanation for PCR does, however, vary considerably between simulations. The first principal component almost exclusively describes  $x_3$ , while the second component can take any direction in the  $x_1 x_2$  plane. The observed direction is determined by small correlations between the random numbers.

## Example 2

Our next example shows the relationship between MLR, PLS, and PCR for an underdetermined problem. The underdetermined situation is likely to occur in calibration against spectrophotometric measurements. The data are once again randomly distributed. Each sample contains a mixture of four pure “spectra.” The concentration of the species in a sample is denoted by  $\mathbf{x}$  and its spectrum is denoted by  $\mathbf{s}$ . Each spectrum has a single Gaussian peak (see Figure 2.3, solid curves). The distribution of each component is uniform within the ranges  $x_1 \in [0, 0.7]$ ,  $x_2 \in [0, 0.5]$ ,  $x_3 \in [0, 0.6]$ , and  $x_4 \in [0, 1.2]$ .

We want to quantify the third compound with the dashed spectrum. The relationship between  $\mathbf{x}$  and  $y$  is once again described by Eq. (2.16). As before, white noise is added to  $\mathbf{S}$  and  $\mathbf{y}$ . The standard deviation of the noise in  $\mathbf{s}$  is 0.01, and in  $y$  it is 0.1. There is no selective region in which only the desired signal is present.

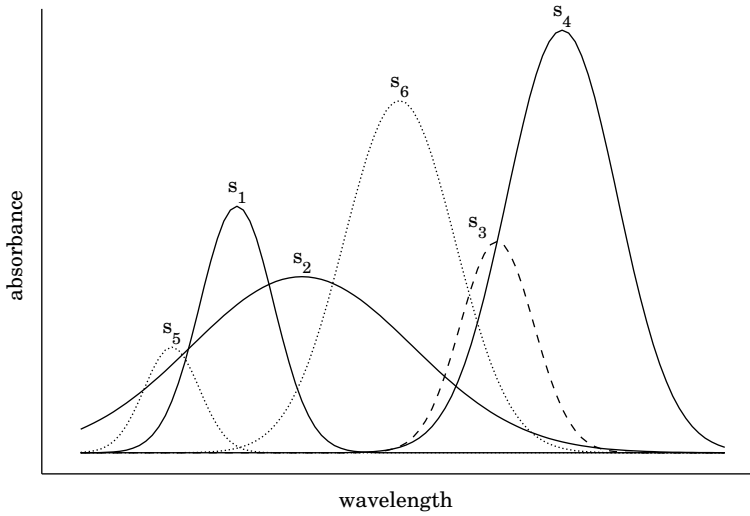


**Figure 2.2.** The second principal component, otherwise the same as in Figure 2.1. (a) Explained variance. (b) Predicted vs. measured  $y$ .

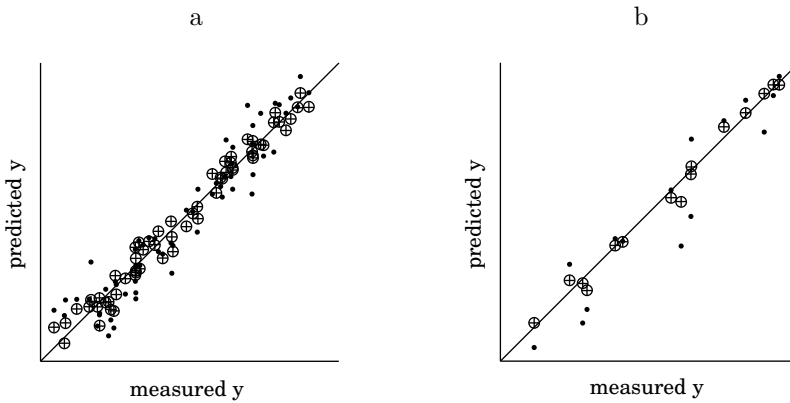
Three test sets are created with 15 objects each. The first test set is created precisely as the calibration data. To the second set of test samples the peak  $\mathbf{s}_5$  (dotted) of Figure 2.3 is added. Its intensity is drawn from the uniform distribution in the range of  $[0, 1]$ . The third and last test set contains a peak  $\mathbf{s}_6$  (dotted) that overlaps with our peak of interest. The intensity of the interfering peak has a uniform distribution in the range of  $[0, 0.3]$ . How will the methods respond to the extra sources of variance not present in the calibration data?

Since the problem is underdetermined, MLR cannot be used. Instead, PCR with as many components as samples has been used. This enables  $\mathbf{X}^T \mathbf{X}$  to be inverted while retaining most of the MLR properties of the solution (except the infinite variance of the regression coefficients). We will still refer to the method as MLR. For PLS and PCR, the number of latent variables was determined using fivefold cross-validation and selecting the model with the lowest root mean squared error of prediction.<sup>‡</sup> Four latent variables were selected for both methods, as would be expected from knowing how the data set was constructed. The predicted values in Figure 2.4 (a) are those from the validation set in the cross-validation. In this figure all three methods seem to give reasonable results. Surprisingly enough, even MLR performs well, although not as good as PCR or PLS. The latter two models are about as good as they can be. The test set predictions are the real test of the methods. The predictions of the first test set are shown in Figure 2.4 (b), and these are of about the same quality as the cross-validated predictions. Exposing the methods to the non-overlapping extra source of variance in test set two is degenerative to the MLR predictions, while

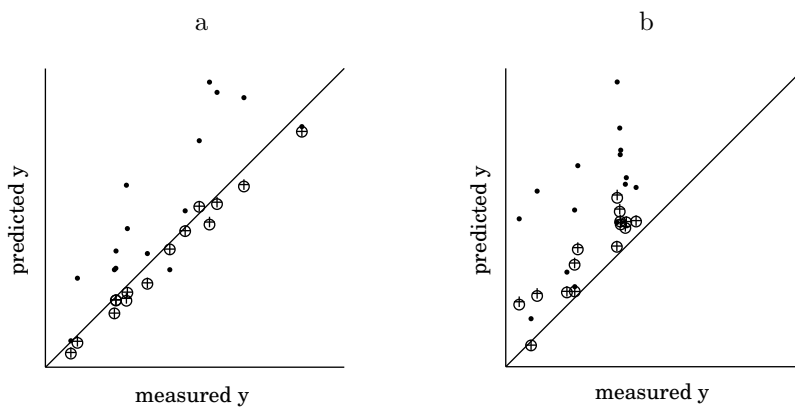
<sup>‡</sup>From doing this, the PLS and PCR predictions suffer from some selection bias. However, being the inventors of the data, we could have chosen the correct number of latent variables *a priori* and exactly the same numerical values would have been without bias. Something to think long and hard about?



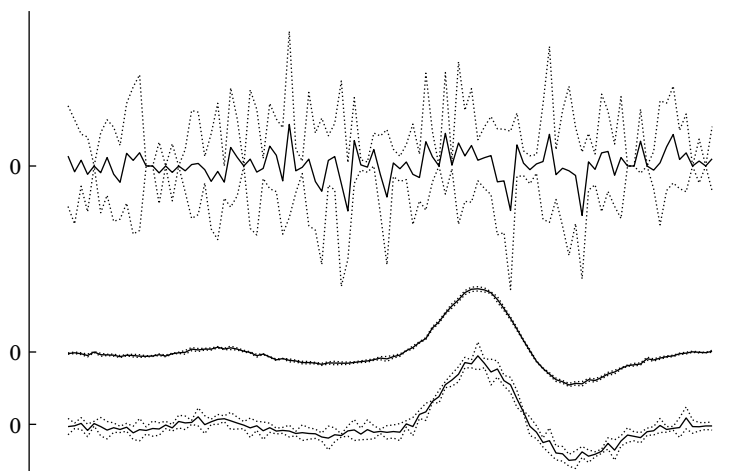
**Figure 2.3.** Simulated pure spectra used in example 2. Spectra 1–4 are included in the calibration models for  $s_3$ . The spectra  $s_5$  and  $s_6$  are added to test sets 1 and 2 respectively. The depicted spectral intensities correspond to the maximum concentrations/intensities in the samples.



**Figure 2.4.** Predicted vs. measured values for example 2 using MLR ( $\bullet$ ), PLS ( $+$ ), and PCR ( $\circ$ ). (a) Predictions from cross-validation. (b) Predictions from test set 1.



**Figure 2.5.** Predicted vs. measured values for example 2 using MLR (●), PLS (+), and PCR (○). (a) Predictions from test set 2. (b) Predictions from test set 3.



**Figure 2.6.** Regression coefficients of example 2 (solid curves) with 95% confidence intervals (dotted). MLR at the top, PLS in the middle, and PCR at the bottom.

PLS and PCR stand out by being almost unaffected by this extra peak. The third test set with peak overlap causes all three methods to overestimate the content of  $x_3$ . MLR again displays the worst performance. An explanation of this behavior can be found by looking at the regression coefficients. These are depicted in Figure 2.6 as solid curves with the 95 percent confidence interval indicated by the dotted curves. The  $\mathbf{b}$ -vector of MLR mostly resembles noise and the uncertainty in the elements is huge; we cannot even tell the sign of any of them. PLS has the most stable regression vector, so stable in fact that the confidence interval can hardly be seen. The regression vector of PCR resembles that of PLS while being noisier. The predictions of PLS and PCR do not exhibit any discernible differences, although PLS can be preferred to PCR based on the stability of its  $\mathbf{b}$  vector. To sum up, PLS and PCR can handle interferences that are present in the calibration set. They are affected by other interfering signals, though only a little if this signal does not overlap with the signal of interest. MLR, on the other hand, is unstable and is seriously affected by new sources of variation. If the interfering signal overlaps the signal from our analyte, the predictions from all three methods will be biased; however, that is only to be expected.

## 2.5 Classification

In everyday life we use our senses to classify things all the time. We perceive a chair as a chair and seldom confuse it with a table, and we can tell classical music from techno, rock, or pop. All these decisions are based on certain features or characteristics of the objects or phenomena. Chairs and tables share many characteristics: they usually have four legs and a horizontal surface. The leg of one chair may be very different from the leg of another. There is a continuous range of manifestations of characteristics, and these may be represented as integers or real numbers. The everyday division of objects into groups can be viewed as a mathematical problem. In chemometrics classification can serve to tell whether a batch of chemicals is pure enough or whether the impurity levels in a drug are low enough so it can be sold. The mathematical task is to find the boundaries between the groups. Discrimination between objects based on sensory input can be highly nonlinear and very tricky to do with computers. A well-studied subject, where the success is moderate, is the recognition of handwritten letters and numbers from photos or some other digital representation, such as the touch pad of a hand-held computer. Difficult pattern recognition problems within chemometrics include, for instance, the classification of rats that have been fed a drug and those that have not, based on NMR spectra of their urine,<sup>29</sup> or discriminating between apple varieties in apple juices, also based on NMR spectra.<sup>30</sup> In this thesis, only the case of continuous valued representations will be considered. There, the class boundary is a surface in  $\mathbf{R}^n$ . An object is, or is not, a member of a class, depending on which side of the class boundary it lies. The surface is sometimes found easily; it may be an upper limit to the amount of impurities

in a drug, according to some document—a yes or no situation, is the product acceptable or not? At other times it is a nontrivial task when we know that the measured multivariate data belong to one of two or more classes.

### 2.5.1 Nearest neighbour classification

One intuitive and very efficient classifier is the  $k$  nearest neighbour method ( $k$ -NN). The distances between an unknown sample and other known objects are computed. The class in the majority among the  $k$  closest objects determines the class of the sample. This decision rule may lead to a complex surface of separation. Among its pros is its simplicity—the method has only one parameter, the odd integer  $k$ , which is usually larger than or equal to 5, yet smaller than 20. An implicit parameter is the distance measure that is used. The Euclidean distance is a common choice.

### 2.5.2 Linear discriminant analysis

Linear discriminant analysis<sup>31</sup> finds an optimal hyperplane that separates two multivariate normal classes  $A$  and  $B$ . The classes are assumed to have an equal variance-covariance structure and a pooled covariance matrix,  $\Sigma^{-1}$ , is used. The normal to this hyperplane,  $\mathbf{n}$ , is found by maximizing the Fisher criterion:

$$F = \frac{\mathbf{n}^T(\mathbf{m}_A - \mathbf{m}_B)(\mathbf{m}_A - \mathbf{m}_B)^T \mathbf{n}}{\mathbf{n}^T \Sigma^{-1} \mathbf{n}}, \quad (2.17)$$

where  $\mathbf{m}_A$  is the center of class  $A$ . The use of a pooled covariance matrix requires the classes are not to differ too much in shape. If the classes differ in direction or spread, this classifier will display problems with nonoptimal performance.<sup>8</sup>

### 2.5.3 SIMCA

Linear discriminant analysis is limited to situations where a sample belongs to exactly one of  $m$  classes. Sometimes the problem is such that a sample may belong to more than one class at the same time, or not belong to any class. If this is the case, one remedy can be to use the SIMCA method. SIMCA stands for soft independent modelling of class analogy and was developed by Wold<sup>32,33</sup> in 1976. In this method each class is modelled by a multivariate normal in the score space from PCA. Two measures are used to determine whether a sample belongs to a specific class or not. One is the leverage—the Mahalanobis distance<sup>34</sup> to the center of the class, the class boundary being computable as an ellipse using a multivariate Student's  $t$ -distribution at a suitable level of significance. The other is the norm of the residual, which must be lower than a critical value from an  $F$ -test, *i.e.*, not significantly larger than the residuals of the calibration samples.

### 2.5.4 PLS-DA

When classes cannot be assumed to follow a normal distribution, one may try to find a hyperplane that is optimal with respect to some other measure, for instance, the separation between two adjacent classes based on the sample distances to the hyperplane. Partial least squares-discriminant analysis<sup>35,36</sup> (PLS-DA) is of this kind. For each class, a response vector is formed using ones and zeros to indicate class membership of each object. A PLS regression model is fitted to the data, possibly using cross-validation to determine a suitable number of latent variables. An object is considered to belong to the class with the highest predicted response. A binary classification requires only one response vector and the decision plane is defined by  $\mathbf{P}^T \mathbf{X} \mathbf{b} = 0.5$ .

### 2.5.5 Nonlinear classifiers

As long as the data has more variables than samples, all classification problems are linearly separable. However, the classifier will become unstable if it uses more dimensions than the chemical rank of the data. This situation can arise if the problem is inherently nonlinear, *i.e.*, the class boundary is curved in a space with the dimensionality equal to the chemical rank. The LDA and PLS-DA cannot handle this. If objects truly belong to one and only one class, SIMCA might not work. One remedy is to transform the data so that they become linearly separable. Finding a transform that does the job could be prohibitively difficult if it is to be done by hand.

The simplest nonlinear classifier is the quadratic discriminant.<sup>37</sup> It is based on the assumption that the classes are normal distributed. An object is assigned to the class which has the highest value of the probability density at the position of the object. The boundary between two classes is defined as the surface where the probability densities are equal. This is a quadratic surface and, hence the name.

Neural networks belong to a flexible class of methods that can be used for classification and are well suited to nonlinear problems. Just as with PLS, the data are projected onto a subspace with as many dimensions as the number of neurons in the hidden layer. An easy introduction to the subject is given by Zupan and Gasteiger,<sup>38</sup> while Haykin<sup>39</sup> gives a more comprehensive and detailed treatment. See also Section 2.7.

### 2.5.6 Class separability measures

The objective when constructing a classifier is to have the classes as well separated as possible. The separation is related to the within-class and the between-class variance. If their ratio (with the between-class variance in the numerator) is high, the classes are better separated than when the ratio is low. The Fisher criterion [Eq. (2.17)] is precisely this ratio. The criterion is therefore often used

to choose between transforms or to optimize the parameters of a transform. Another measure of separation is the Bhattacharyya distance.<sup>37,40</sup> It is directly related to the Bayes error of the quadratic classifier. The Bhattacharyya distance is uncommon within the field of chemometrics, but is more frequently used in other disciplines that are also concerned with classification problems. It is defined as:

$$B = \frac{1}{8}(\mathbf{m}_A - \mathbf{m}_B)^T \left( \frac{\Sigma_A + \Sigma_B}{2} \right)^{-1} (\mathbf{m}_A - \mathbf{m}_B) + \frac{1}{2} \ln \frac{|\frac{\Sigma_A + \Sigma_B}{2}|}{\sqrt{|\Sigma_A||\Sigma_B|}} \quad (2.18)$$

The Fisher criterion is limited by the assumption of equal variance-covariance structure of the classes. If this assumption does not hold, its use is questionable. The measure is much used anyway because of its robustness. The criterion relates to a linear classifier, and these tend to be more robust than nonlinear ones. Even if the covariance matrices are different, LDA usually performs well. A more complex classifier might not be as forgiving if its postulates about the data were violated. The quadratic classifier, for instance, is considered to be unstable if the covariance matrices are estimated from a small number of samples. This means that the Bhattacharyya distance is not a very good measure to optimize against if the number of samples is limited. One of its properties is that two classes can show a considerable amount of separation even if the class means coincide. Hence using the distance in combination with a linear classifier would be foolish.

SIMCA has a measure of class separation called discriminatory power. This is computed as the sums of squares of the residuals when samples from class A are projected onto the model of class B and vice versa. This sum is divided by the sum of the squared residuals when the samples are projected onto the model of the class to which they belong:

$$D = \sqrt{\frac{\text{tr}((\mathbf{X}_A - \mathbf{X}_A \mathbf{P}_B \mathbf{P}_B^T)^T (\mathbf{X}_A - \mathbf{X}_A \mathbf{P}_B \mathbf{P}_B^T)) + \text{tr}(\dots)}{\text{tr}((\mathbf{X}_A - \mathbf{X}_A \mathbf{P}_A \mathbf{P}_A^T)^T (\mathbf{X}_A - \mathbf{X}_A \mathbf{P}_A \mathbf{P}_A^T)) + \text{tr}(\dots)}} - 1 \quad (2.19)$$

This measure is good if the dimensionality of the measurement space is considerably higher than the space spanned by the class models. If, for instance, the classes are two-dimensional in a two-dimensional space—a not completely unrealistic situation, then the discriminatory power becomes zero. If you are unlucky, the measure will give an unrealistically small class distance also when the measurement space have one more dimension than the class models. As the difference in dimensionality increases, the likeliness of obtaining a misleading value decreases rapidly.

We have found neither of the described measures completely satisfactory and in Paper V we therefore present a new measure of class separation. It is closely related to the mentioned measures as it is based on the assumption of the classes being normally distributed.

## 2.6 Model validation

The problem in regression with more variables than samples is to know when to stop including more latent variables. The more we include, the better the fit of the model to the data. When the latent variables are as many as the number of samples, all regression problems can be solved exactly with a linear model. In other words, there is a unique solution to the equation  $\mathbf{y} = \mathbf{B}\mathbf{T}$ , where  $\mathbf{B}$  is the matrix of regression coefficients. This might seem tractable at a glance, but the solution is unstable with respect to measurement noise. Small changes in the data will induce large changes in the regression coefficients. The changes may be so large that the regression coefficients will appear purely random. The prediction error of a new/unknown sample becomes uncontrollably large.

The problem is called overfitting and the solution is to search for a model that generalizes well rather than fits the data as closely as possible. The most important ways of estimating the generalization error of a model, *i.e.*, validating the model, is to use a test set and cross-validation.<sup>41</sup> The bootstrap<sup>42,43</sup> method can be used to study the stability of the model parameters.

The purpose of validation is twofold. It is used to optimize the hyper-parameters of the model or to choose between competing models. Hyper-parameters are parameters of a model which are not changed by the model-fitting procedure, but which still affect the model. The number of latent variables in PLS regression is an example of a hyper-parameter. The second and most important purpose is to assess the expected performance of the selected model. To achieve the two goals, it is recommended and customary to divide the samples into a calibration set and a test set.

The test set is kept aside and is only used to put the final model to a realistic test of its generalization error. One recommendation is to use one-fifth of the samples as the test set. This fraction is not an absolute figure but will depend on the total number of samples, how many samples that are needed to fit the model, etc.

The calibration set is used for both fitting model parameters and choosing between models. Bootstrap methods and cross-validation are common ways to utilize the data as efficiently as possible.

In cross-validation the data are divided into, say,  $k$ , subsets. The model parameters are fitted using  $k - 1$  subsets and the model performance is measured on the subset that is left out. The procedure is repeated  $k$  times until every subset has been left out once. The performance of the model family with a certain set of hyper-parameters is simply the average performance over the  $k$  realizations. Additionally, it may be of interest to study how much the model parameters vary during the cross-validation. The model-selection criterion may be the generalization performance alone or be based on a combination of the parameter stability and the generalization performance. Leave-one-out cross-validation uses one sample per subset, although for computational or other reasons the subsets may contain almost any number of samples. If the predictors of the data

set are designed and not random variables, some care must be taken when partitioning the data into subsets. Replicates should preferably belong to the same subset, otherwise the prediction error may become unrealistically low. The data should also span the domain. If all the withheld samples are in one region of the predictor space and those used for model fitting in another region, then the prediction error can be expected to be overestimated.

Bootstrap validation is mostly concerned with computing statistics for model parameters. The idea is to use the available observations as representing the whole population of data and draw  $k$  sample subsets using sampling with replacement. These are used to fit the model parameters and the parameter statistics are computed from the  $k$  realizations. Used in this way, bootstrap methods do not provide estimates of the generalization error, although this may correlate with the magnitude of the variance of the model parameters.

In chemometrics cross-validation is the predominant method of validation. In statistics bootstrapping is also much used. Another related validation method is the so-called jackknife.<sup>43</sup>

Even if one goes about all of this in order to get as representative and accurate results as possible, the selected model will still suffer from selection bias. This is due to the fact that the model is fitted on the same data that are used to select it. Ideally, the model should be refitted on new samples after the selection and evaluated with a test set never used before. Chatfield<sup>44</sup> and Miller<sup>45</sup> provides excellent discussions of model selection and the implications thereof.

The model selection bias stems from the fact that choosing the model with the lowest cross-validation error might not correspond to the model with the lowest true generalization error. That is, the generalization error of the selected model has probably been underestimated and the parameters of the model are biased due to the model selection step. The greater the number of alternative models that are considered, the likelier it is that the selected model will suffer from selection bias. Of course, we cannot solve the problem by choosing the second or third best model; these are just as likely to suffer from selection bias and will represent a worse choice. By refitting the model using new data, the unbiased estimates of the model parameters can be obtained, together with a better estimate of the model's true generalization error.

A real problem in data analysis leading to even more biased results is the difficulty of only using the test set once. It seems really simple but requires a lot of self-discipline. If the test set results are worse than expected, it may be very tempting to make a few adjustments to the model and test it again. But in that case the test set results will no longer be unbiased. Information from the test set has seeped into the model. By using the test set more than once, its independence is compromised. The bias in the result will grow larger every time the test set is used to make corrections to the model. The test set simply becomes a part of the calibration data and a new test set needs to be obtained to assess the model performance.

## 2.7 Pulse-coupled neural networks

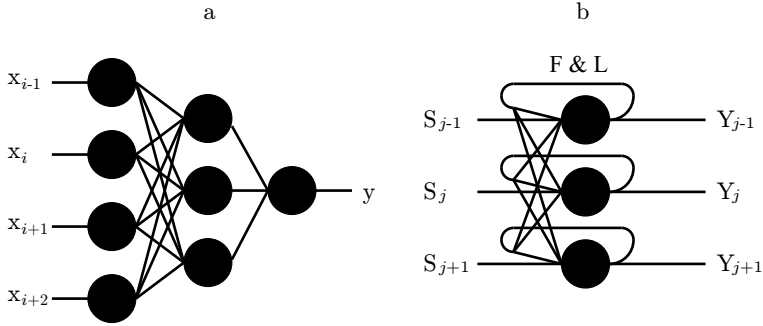
In Paper I, pulse-coupled neural networks (PCNNs) have been used to preprocess three-dimensional image data into time series. The time series have been used to model quantitative structure retention relationships. To put PCNNs into context, a brief description of nerve cells and the development of the neural networks field of research is appropriate.

In 1906 Santiago Ramón y Cajal was awarded the Nobel Prize in Physiology or Medicine for the discovery of neurons in brain tissue. He shared the prize with the Italian physician Camillo Golgi, who more than thirty years earlier, in 1873, had invented a silver staining technique still used today. Silver staining made it possible to see individual neurons using a microscope. Golgi had also seen individual brain cells, but the major breakthrough in knowledge of the organization of the nervous system came about in the late 1880s and the 1890s with the work of Cajal.<sup>46, 47</sup>

A neuron consists of three main parts: the cell body or soma, the dendrites, which are connected to the soma and can be likened to the root system of a tree—heavily branched, and the axon, a long outgrowth from the soma. The dendrites respond to stimuli from other neurons by synaptic connections. The stimuli is forwarded as an electrical signal along the cell membrane to the soma. The axon provides stimuli to other neurons. If the potential difference across the cell membrane exceeds than a threshold value at the base of the axon (near the soma), the axon fires. This means that an electrical pulse is sent along the axon to its synapses, which are either electrical or chemical. The electrical synapses propagate the signal to other neurons by barrel-shaped proteins called connexons. In chemical synapses a neurotransmitter is released, *e.g.*, acetylcholine. The transmitter diffuses across the synaptic cleft to the postsynaptic membrane of a dendrite, where it binds to a receptor and, thus, the signal has reached a second neuron, which in turn responds to it.<sup>48</sup>

Ever since the discovery of neurons, researchers have been intrigued by the network structure of the brain and have tried to model its functionality. One early milestone was when Hebb in 1949 formulated a learning rule for synaptic connections between neurons.<sup>39</sup> The brain is massively parallel and research into modelling neural networks, more often than not, requires huge numbers of numerical operations to be performed. Thus, progress in the field has followed the development of computers. Nowadays, two mainstreams in the research can be seen: detailed simulations of neurons mimicking biological systems (see, *e.g.*, Ref.<sup>49</sup>) and function approximation by the use of artificial neural networks.<sup>39</sup>

The biologically realistic simulations are generally slow and not usable for practical applications. The information flow between neurons is coded in the frequency with which an axon fires. The neural networks used for function approximations have neurons (called perceptrons) that are extremely simplified. As a result of stimuli, these give an immediate real-valued response rather than an analog pulsed one. The inputs from the dendrites,  $\mathbf{x}$ , are also real-valued and modified by a multiplication with a weights,  $\mathbf{w}$ . At the soma the inputs



**Figure 2.7.** (a) A feed forward neural network with one hidden layer. (b) A one-dimensional PCNN with feedback loops (F & L).

are summed and used as the argument of a transfer function. The perceptron is described by the function  $y = \varphi(\mathbf{x}^T \mathbf{w} + b_0)$ , where  $y$  is the axonal output, and  $\mathbf{w}$  and  $b_0$  (a constant) its parameters. The transfer function  $\varphi(\cdot)$  is often sigmoid, *e.g.*, arctan. It has been shown that perceptrons can be used to approximate any function, if connected in one so-called hidden layer and one output layer, see Figure 2.7 (a).

The pulse-coupled neural network is a third variety that falls between the two previously described. It is a simplified model of the cat's visual cortex,<sup>50</sup> with local connections to other neurons. The neurons are stimulated continuously and respond with a binary output which takes the value one if the internal state of the neuron goes above a threshold. This firing of the neuron results in inhibitory feedback, which forces the neuron to a period of inactivity before it can fire again. The frequency with which a neuron fires is roughly proportional to the intensity of its stimulus. The network is constructed as a single layer with one neuron per pixel in the image with which it is stimulated. The output,  $Y$ , from one neuron is fed to its neighbors via a topological link matrix,  $\mathbf{K}$ . The PCNN architecture is schematically described in Figure 2.7 (b). There are two types of input to each neuron: the feeding input  $F$  and the linking input  $L$ . These are so-called leaky integrators, meaning that input is summed over time (integrated) and current integral value is multiplied by a constant in the range  $(0, 1)$  at each time step (it leaks away). The feeding input of neuron  $ij$  receives stimuli,  $S$ , from the pixel with corresponding position and neighboring neurons, while the linking input only receives stimuli from the neighboring neurons. The two inputs are combined nonlinearly into the internal state of the neuron,  $U$  [Eq. (2.22)], which is compared to a threshold,  $T$  [Eq. (2.23)]. If the value of the internal state exceeds the threshold, the output  $Y$  of that neuron is set to 1 (the neuron fires), otherwise it is 0 (inactive). The equations of a two-dimensional PCNN

neuron are:

$$F_{ij}(t) = a_F F_{ij}(t-1) + S_{ij} + b_F \sum_{kl} Y_{kl}(t-1) K_{kl} \quad (2.20)$$

$$L_{ij}(t) = a_L L_{ij}(t-1) + b_L \sum_{kl} Y_{kl}(t-1) K_{kl} \quad (2.21)$$

$$U_{ij}(t) = F_{ij}(t)(1 + cL_{ij}(t)) \quad (2.22)$$

$$Y_{ijk}(t) = \begin{cases} 1, & \text{if } U_{ij} \geq T_{ij} \\ 0, & \text{otherwise} \end{cases} \quad (2.23)$$

$$T_{ij}(t) = a_T T_{ij}(t-1) + b_T Y_{ij}(t) \quad (2.24)$$

$$g(t) = \sum_{ijk} Y_{ijk}(t) \quad (2.25)$$

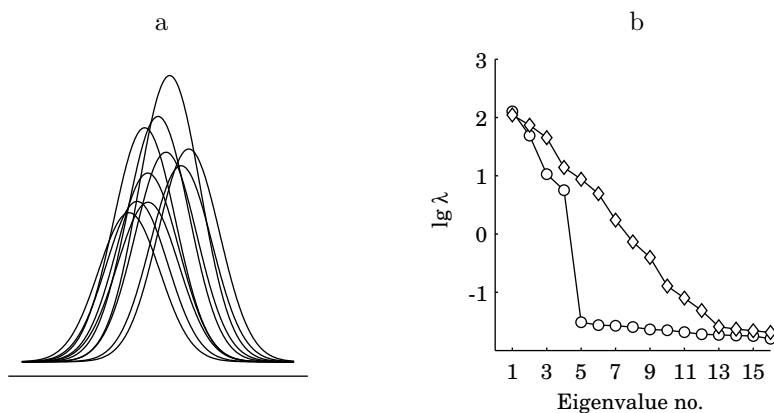
where the parameters  $a_F$ ,  $a_L$ , and  $a_T$  regulate how quickly the neuron forgets and the parameters  $b_F$ ,  $b_L$ , and  $b_T$  determine the importance of the stimuli from the neighbors, and  $c$  regulates the strength of the linking input effect. The summations in Eqns. (2.20) and (2.21) run over all elements of the matrix  $K$  centered on neuron  $ij$ . The generalization into three dimensions is straightforward, an extra index being added to all the equations.

A property of this neural network is that the time series  $g(t)$  is more or less invariant to the orientation of objects in an image.<sup>51</sup> Another noteworthy property of the times series is that it is something of a fingerprint of an image. It has also been reported that the PCNN is good at image segmentation and edge detection.<sup>52</sup> The network model has generated an interest as an image preprocessor for automated target recognition in military applications.<sup>53</sup> Kinser *et al.*<sup>54</sup> studied the unified cortical model (a simplified PCNN model) with chemical data in the form of  $17\beta$ -estradiol represented as a three-dimensional image. They concluded that there might be a potential use for the network in quantitative structure-property relationships (QSPR), although this was not investigated in that paper.

## 2.8 Peak alignment

Peak alignment is a form of data pretreatment. It can be applied to data which is bilinear in theory but shows deviations from bilinearity in measured data. The data should also have pronounced peaks, and typical instrumental techniques like, for instance, gas and liquid chromatography, NMR, and capillary electrophoresis generate the type of data that we are considering.

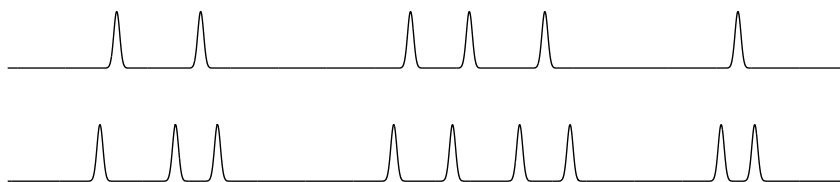
Let us revisit example 2 in Section 2.4.4. This time the peak positions are randomized. Figure 2.8 (a) shows typical peak shifts, while in (b) the degenerative effect of peak shifts on a PCA model of the data is shown. The eigenvalue structure in the data without shifts indicates a chemical rank of four. With added peak shifts the eigenvalues fall off smoothly down to the noise level, which



**Figure 2.8.** Example 2 revisited: (a) typical peak shifts, (b) effect of peak shifts on the eigenvalues of  $\mathbf{X}^T \mathbf{X}$ , original data ( $\circ$ ), data with peak shifts ( $\diamond$ ).

makes it much more difficult to estimate the chemical rank. Thus, there is a real need for peak alignment.

A method of peak alignment (called PARS, presented in Paper III and IV) can be designed as follows: first, we note that the original data representation is not needed. For the subsequent analysis of the data, after alignment, we will be interested in the information carried by the peaks. The information in a single peak can be characterized by the statistics position, height or area, and possibly one or more shape parameters. A logical conclusion is to compute the statistics for every peak in a chromatogram, resulting in a more compact data representation. The representation may use a matrix for each sample with peaks as rows and statistics as columns. Using this representation, there is a distinct possibility that the matrices for different samples will have different number of rows, as the number of peaks may vary between samples. The alignment problem may now be formulated as a graph problem. Figure 2.9 shows two chromatograms that are to be aligned. In the lower chromatogram the peaks are shifted between two and three FWHM (full width at half maximum) to shorter retention times. Three new peaks are added to complicate matters, by creating ambiguous solutions. For each peak in one of the chromatograms, we look for possible matching peaks in the other. In this example, we consider that peaks for which the difference in position is at most 10 FMWH to be possible matches. In Figure 2.10, all possible matches are plotted as circles with the positions in one chromatogram on the  $x$ -axis and the positions in the other on the  $y$ -axis. A graph is created by connecting these circles by lines. Any path from the lower left corner to the upper right corner represents a possible solution to the alignment problem. The lines between possible matches are called edges. Each edge is associated with a weight that governs the quality of the solution. The value of the weight is a complicated issue, but it contains a term that is proportional



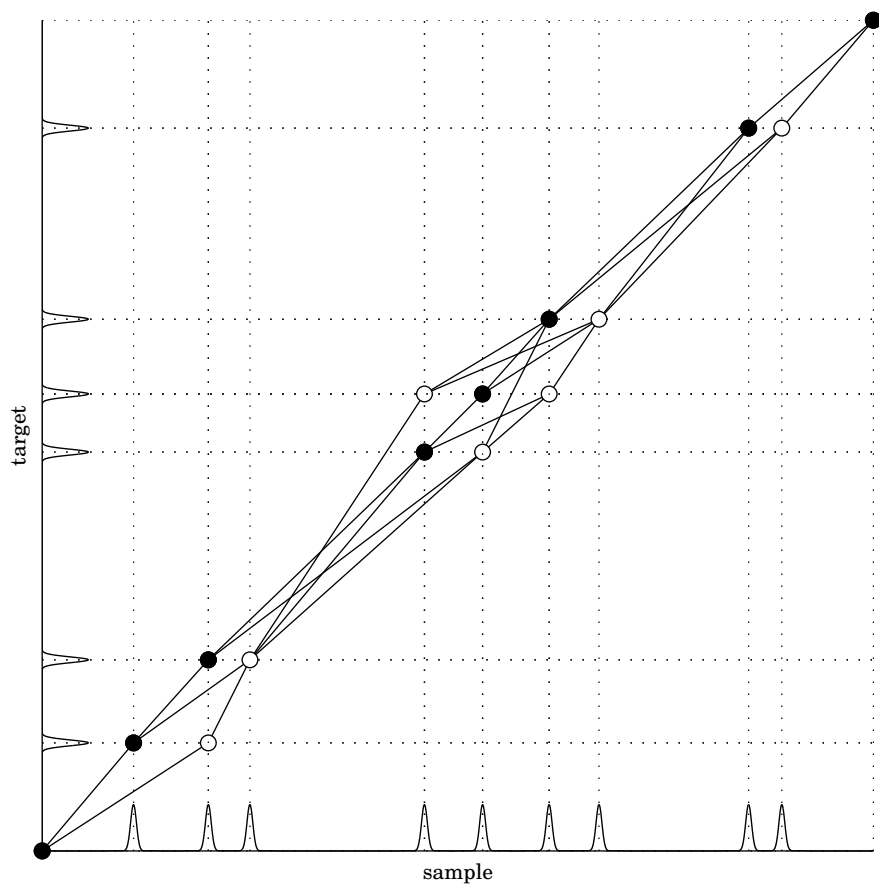
**Figure 2.9.** Simulated chromatograms: the uppermost is used as the target by the alignment procedure, the bottom one is the sample to be aligned. The sample peaks are shifted to retention times 2–3 FWHM lower than in the target. Three new peaks are also included.

to the required peak displacement. A more thorough description of the edge weights can be found in Paper IV. The solution to the peak alignment problem has been reduced to finding an optimal path through the graph. The optimal path has a minimal sum of edge weights.

An analogy to this problem is finding the quickest route between address A and address B in a city, when travelling by car. There are a lot of alternative routes to choose from: some routes are longer but the extra distance may be compensated for by higher speed limits; other routes are short but have many traffic lights or may suffer from traffic jams. An edge weight corresponds to the time it takes to drive between two intersections, at which alternative routes diverge or converge. Such an intersection corresponds to a possible peak match in our problem.

The optimal solution is sought using breadth first search (BFS). BFS is one of two fundamental graph search algorithms, the other one being depth first search. BFS is good at solving problems of finding optimal paths through graphs, while depth first search answers questions about existence: is it possible to get to point B starting from point A? An introduction to graph theory can, for instance, be found in Ref.<sup>55</sup>

The solution using this method of peak alignment is indicated by the filled circles in Figure 2.10.



**Figure 2.10.** The search graph used by PARS with the optimal path marked by filled circles.

## Chapter 3

# Computation of solvation free energies

Free energy is a thermodynamic quantity that governs the direction of chemical reactions, physical processes, and chemical equilibria. At constant temperature and pressure it is called Gibbs free energy, while at constant temperature and volume it is called Helmholtz free energy. In most practical situations, constant pressure is simpler to achieve and maintain than a constant volume. Therefore, our interest has been directed towards Gibbs free energy.

The solvation free energy is the energy required to transfer a molecule from an ideal gas phase into a bulk solvent (*e.g.*, water or octanol) at constant temperature and pressure. This free energy is known as the Gibbs free energy of solvation. The free energy can be used to predict the partition of different compounds between octanol and water, *i.e.*, the ratio of the concentration in the octanol phase and in the aqueous phase at equilibrium. The partition coefficient,  $\log P_{o/w}$ , can be computed from the solvation free energies in octanol,  $G_{\text{octanol}}$ , and water,  $G_{\text{water}}$ , as:<sup>56</sup>

$$\log P_{o/w} = \frac{1}{RT \ln 10} (G_{\text{octanol}} - G_{\text{water}}) \quad (3.1)$$

The partition coefficient may be seen as a very crude model of a chromatographic system, one phase being polar and the other nonpolar. Once we master how to predict the partition accurately, or at least consistently, we may transfer the computational scheme to make purely theoretical predictions of phase equilibria for chromatographic separations.

There are many ways of estimating solvation free energies (or the desired equilibrium constant directly). The procedure in quantitative structure-activity relationship (QSAR) modelling is to use multivariate linear regression to correlate a set of molecular descriptors with an experimentally determined response (free energy). The molecular descriptors are variables that indicate, for instance,

the presence of a functional group (OH, NO<sub>2</sub>, NH<sub>2</sub>, ...) at a specific position in a molecule, or the number of OH groups attached. Free–Wilson analysis is based on this type of descriptors. The solvent accessible surface area<sup>57,58</sup> is another well-studied molecular descriptor used in Hansch analysis. Combining Free–Wilson and Hansch analysis seems to be the more powerful than being confined to a particular type of descriptors.<sup>59</sup> A drawback of this class of methods is their limited generalization ability. The calibration model is often only valid for a class of compounds with a common substructure, onto which different functional groups are attached. Also, the need for experimental data is a significant drawback. A lot of effort can be wasted creating the calibration model; it might be more efficient to experiment directly on the substances of interest—if available, that is.

The limitations of QSAR models can be avoided by computing the solvation free energies from “first principles,” or as closely as is possible. Ideally, we should use some kind of quantum dynamics to determine the free energy, but the sheer complexity of such computations is prohibitive with present day computers. Therefore, a compromise between accuracy and computational feasibility is necessary. One approximation is to neglect the quantum properties of matter and treat it as if it followed Newton’s laws of motion—this was our approach. The free energy is computed using a special class of molecular simulation methods. (Computing free energy is considered to be a difficult problem in statistical mechanics, the reason being that it depends on the volume of the phase space.) Another approximation is to look upon the solvent as a continuum, defined by its dielectricity constant, and treat the solute according to quantum mechanics. Within this approximation, the solvation free energy is determined by the difference in energy when the solute is in vacuum and when the solute is present in a cavity in the continuum solvent, plus the work required to create the cavity. The solute is accurately described, although entropic contributions that the solute induces in the solvent are neglected.

We have been studying the computation of absolute free energies of solvation by direct insertion of a solute molecule into a molecular solvent. A more common approach is to compute the free energies relative to a known (by computation or experiment) free energy of solvation for a single compound and then use a coupling parameter to gradually transform that molecule into the desired solute molecule. The absolute free energy of solvation is computed as the sum of the known solvation free energy and the computationally observed difference in free energy for the transformation. This approach is most efficient when studying a class of molecules whose structures share a common backbone. Absolute free energy calculations, however, allow maximum flexibility for structural diversity among the studied molecules.

The rest of this chapter will be devoted to an introduction to statistical mechanics, models for interactions between atoms, the two dominant simulation techniques: Monte Carlo and molecular dynamics (MD), and the method of expanded ensemble molecular dynamics. Expanded ensemble MD is a hybrid Monte Carlo-MD method that we have used to compute solvation free energies.

### 3.1 Statistical mechanics

When matter is regarded from a microscopic point of view while macroscopic thermodynamic quantities or properties are of interest, statistical mechanics bridges the gap between the microscopic and macroscopic scales. Statistical mechanics connects the two extremes via what is known as the partition function. This is a weighted sum of all possible states of the system. The sum transforms to an integral, if the system is considered to be classical instead of quantized. In the classical treatment, the interior states of a system are given by the positions and momenta of all its particles. Every thermodynamical quantity can be derived from this function. The expression for the partition function differs somewhat, depending on which ensemble it represents. An ensemble is the allowed combinations of positions and momenta of particles in the system under study. The name of an ensemble derives from the state variables that are kept constant: we can, for instance, study 256 water molecules at 298 K at a pressure of 1 atm. At least one of the three state variables must be extrinsic, *i.e.*, determine the size of the system. The canonical ensemble ( $NVT$ ) has a constant number of particles, constant volume, and constant temperature, while the isobaric-isothermal ensemble ( $NPT$ ) has constant pressure instead of volume. The  $NTP$  ensemble is the only ensemble used in Paper II and its quasi-classical partition function looks like:

$$Q_{NPT} = \iiint \exp(-(\mathcal{H}(\mathbf{r}^N, \mathbf{p}^N) + PV)/k_B T) d\mathbf{p}^N d\mathbf{r}^N dV, \quad (3.2)$$

where  $\mathcal{H}$  is the Hamiltonian of the system with potential energy and kinetic energy terms for all particles (see Section 3.2). Positions,  $\mathbf{r}$ , and momenta,  $\mathbf{p}$ , together with the volume,  $V$ , constitute what is called the phase space. The value of a property  $A$  is determined as

$$\langle A \rangle_{NPT} = \iiint A(\mathbf{r}^N, \mathbf{p}^N, V) \exp(-(\mathcal{H}(\mathbf{r}^N, \mathbf{p}^N) + PV)/k_B T) d\mathbf{p}^N d\mathbf{r}^N dV. \quad (3.3)$$

The excess free energy of solvation,  $\mu_{\text{ex}}$ , in the  $NPT$  ensemble can be calculated approximately as the energy required to transfer the solute from a noninteracting (ideal gas) state to a dissolved state, while the bulk solution is kept at constant temperature and pressure. For Gibbs free energy,  $A(\cdot)$  takes the form:  $\exp((\mathcal{H}(\mathbf{r}^N, \mathbf{p}^N) + PV)/k_B T)$ .

The integral in eq. (3.3) can be factorized into a product of integrals over positions and momenta separately. The integral over momenta can be solved exactly and the result is the ideal gas. The integral over positions is commonly referred to as the configurational partition function. Approximate analytical solutions to the configurational partition function can be found for simple pair potentials and at low density. These analytical solutions represent equations of state for gases that are more complex and realistic than the ideal gas. Examples include the van der Waals equation and other similar generalizations.<sup>60</sup> For

condensed liquid phases, it is not possible to find explicit analytical equations of state. The configurational partition function must be evaluated using numerical methods—molecular simulation techniques.

The two dominating numerical methods for evaluating the partition function are Monte Carlo<sup>61</sup> and molecular dynamics.<sup>62,63</sup> It is not possible to use simple interval-based integration schemes due to the high dimensionality of the integral: six times the number of particles,  $6N$ . Ideally,  $N$  should be a truly macroscopic number, in the order of Avogadro’s number ( $\sim 10^{23}$ ), but since such integrals cannot be evaluated, the usual number of particles is usually somewhere between 100 and  $10^6$ . To mimic bulk behaviour, the system is allowed to interact with periodic images of itself. This self-interaction is called periodic boundary conditions, or sometimes toroid boundary conditions because of their geometric interpretation. If an interval-based integration scheme were to be used with ten points along each space coordinate, the total number of function evaluations would be  $10^{3N}$ , *i.e.*, at least  $10^{300}$ . Even with this immense computation the accuracy of the thermodynamic property would be ridiculously low.\* Thus, there is a need for more efficient integration schemes. These are termed importance sampling techniques. The idea is to concentrate the function evaluations to the volumes of phase space that contribute the most to the integral. The important volumes often constitute a very small fraction of the total phase space and high accuracy can be achieved using as few function evaluations as  $10^5$  to  $10^8$ . The efficient computational methods perform variance reduction; a lower standard error in the results is obtained even though fewer “samples” are needed.

## 3.2 Forcefields: describing the interaction between atoms

The Hamiltonian,  $\mathcal{H}$ , is an energy operator for the total energy of a molecular system. It is made up of two parts, kinetic and potential energy:

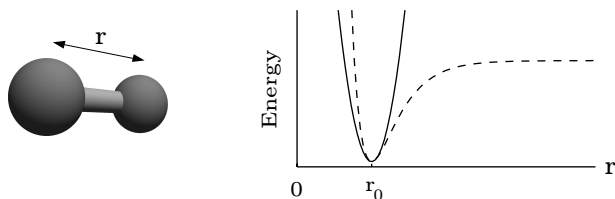
$$\mathcal{H} = T(\mathbf{p}) + U(\mathbf{r}), \quad (3.4)$$

where the kinetic energy is  $T = \sum_{j=1}^N \frac{\mathbf{p}_j^2}{2m_j}$ . The potential energy,  $U(\mathbf{r})$  describes the interactions between atoms and is called the forcefield.

Existing forcefields have been parameterized for somewhat different purposes: the MM2<sup>64</sup> and MM3<sup>65</sup> forcefields are mainly used for “small” molecules, OPLS is an acronym for Optimized Potential for Liquid Simulation<sup>66</sup> and as its name implies it is intended for simulation of organic molecules in water, and CHARMM<sup>67</sup>

---

\*Since two atoms cannot occupy the same position in space, most of the  $10^{300}$  configurations have infinite energy and do not contribute to the integral. Only  $\binom{1000}{100}$  configurations out of  $1000!$  have no more than one atom at each space coordinate. This smaller number of configurations does not make the computations feasible:  $\binom{1000}{100} > 9^{100}$ ; it is still a prohibitively large number.



**Figure 3.1.** The bond harmonic bond potential (solid line) and the Morse potential (dashed).

and AMBER<sup>68</sup> were developed for the simulation of biological systems with proteins and nucleic acids.

A typical forcefield may look like:

$$U = U^{\text{bonds}} + U^{\text{angles}} + U^{\text{torsions}} + U^{\text{improper torsions}} + U^{\text{non-bonded}}. \quad (3.5)$$

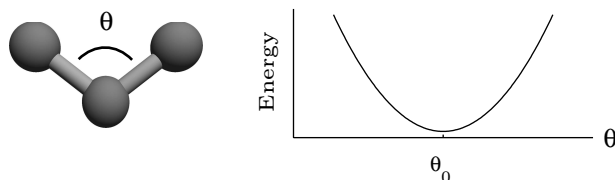
The specification of different forcefields may differ in detail but Eq. (3.5) contains the most important terms. Some forcefields, like MM2 and MM3, use cross terms, *e.g.*, between angles and the bonds that constitute the angle; the cross terms are used to obtain a more accurate description of the internal dynamics of the molecule and to reproduce the frequencies of its vibrational spectrum.<sup>64, 65, 69</sup> Cross terms are less important for structural properties.

### 3.2.1 Bonds

Bond potentials are often modelled as harmonic springs using a second-order polynomial,  $u_{ij}^{\text{bond}} = k(r_{ij} - r_0)^2$  (see Figure 3.1.) If the harmonic approximation is insufficient, anharmonicity can be modelled by third- or fourth-order polynomials, or by the so-called Morse potential.

### 3.2.2 Angles

Angles between three bound atoms (see Figure 3.2) are also modelled by a second-degree polynomial,  $u_{ijk}^{\text{angle}}(\theta - \theta_0)^2$ , where  $\theta$  is the current value of the angle and  $\theta_0$  its value at equilibrium.  $u_{ijk}^{\text{angle}}$  determines the stiffness of the angle and its value depends on the atoms defining the angle. The three atoms define the local chemical environment by their atomic numbers and hybridization states. The  $u_{ijk}^{\text{angle}}$  values are insensitive to atoms further away in a molecule and are therefore transferable to other molecules with the same three atoms, bound in the same order.



**Figure 3.2.** Angle bending potential modelled by a second order polynomial.

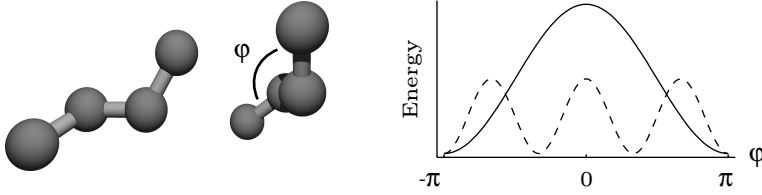
### 3.2.3 Torsional angles

Torsions, or dihedrals, are defined as the rotation of a bond where the two bound atoms have one or more extra bonds to other atoms, see Figure 3.3. As this potential is inherently periodic, it is modelled by trigonometric functions:  $\sum_n u_{ijkl}^{\text{tors}} \cos(n\varphi - \varphi_0)$ . These potentials are often parameterized from small model structures where the bond rotation potential is determined by quantum calculation at different angles. The periodicity depends on the substituents and the hybridization of the central atoms. Ethane, where the carbons are  $sp^3$ -hybridized, have a threefold symmetry,  $n = 3$ . *n*-butane might have an extra term with  $n = 1$  to describe the repulsion when the two methyl groups are in eclipsed position. Some authors use the term *Fourier expansion* to implicitly state the periodicity and use of trigonometric functions for the torsional potential.

Improper torsions are used to maintain the planar structure of, for example, phenylic compounds. These are called torsions as they are defined by four atoms and improper because the atoms are not bound as on a string, as are the atoms of a regular torsion. How improper torsions are defined differs between forcefields, but it is common to use a cosine or second-degree polynomial for the functional form.

### 3.2.4 Non-bonded interactions

The non-bonded potential is usually a pair potential, although three-body potentials exist. The pair potential is, however, much more tractable due to its lower complexity. The number of pairs of atoms is proportional to  $N^2$ , while the number of three-body combinations are in the order of  $N^3$ . A three-body



**Figure 3.3.** Torsional cosine potentials for  $n = 1$  (solid) and  $n = 3$  (dashed).

potential might be needed if polarization effects are significant, although these potentials are rare in practice. The pair interaction between atoms  $j$  and  $k$  is usually a combination of a Lennard–Jones potential and a Coulomb potential,

$$u_{jk}^{\text{non-bonded}} = 4\epsilon_{jk} \left( \left( \frac{\sigma_{jk}}{r_{jk}} \right)^{12} - \left( \frac{\sigma_{jk}}{r_{jk}} \right)^6 \right) + \frac{1}{4\pi\epsilon_0} \frac{q_j q_k}{r_{jk}^2}. \quad (3.6)$$

### Lennard–Jones potential

The Lennard–Jones potential is empirical in its origin, but the terms can be interpreted as follows: the  $r_{jk}^{-12}$  term is the repulsion when the electron clouds around each atom start to overlap, the  $-r_{jk}^{-6}$  term is the dispersive interaction due to polarization.  $\epsilon_{jk}$  determines the strength of the dispersive attraction, *i.e.*, the depth of the potential well, and  $\sigma_{jk}$  is a combination of the two atom sizes (diameters),  $\sigma_{jk} = \frac{1}{2}(\sigma_j + \sigma_k)$ .

### The Coulomb potential and partial atomic charges

The Coulomb potential has two parameters, the partial atomic charges,  $q_j$  and  $q_k$ ;  $\epsilon_0$  is the dielectric constant in vacuo. The potential falls off as  $r^{-2}$  and is a so-called long-range potential, which means that a significant contribution to the potential energy of the system comes from the periodic images of the simulation box. If the Coulomb potential is evaluated directly as a pair potential, the computational complexity is high,  $\mathcal{O}(N^2)$ , because of the contribution from the periodic images. Fortunately, the sum may be partitioned into a short-range and a long-range part, where the latter may be evaluated in Fourier space with a complexity of  $\mathcal{O}(N)$ . This technique is called Ewald summation<sup>70,71</sup> and is used in simulations with charged particles.

The values of the parameters in the Coulomb potential are often debated. Since atomic charge is not a quantum mechanical observable or experimentally observable, there is an element of arbitrariness in their determination. A number of computational schemes exist that assign partial charges to the atomic centers of a molecule, *e.g.*, Mulliken charges, charges derived from the electrostatic potential (ESP), and restrained ESP charges (RESP).<sup>69</sup> These charges are based on quantum mechanical calculations.

Mulliken charges comes from population analysis. The electron density in an orbital is assigned to the atom on which the orbital resides. The electron density that is associated with overlap population is partitioned equally between participating atoms. A problem with Mulliken charges is that they tend to be unstable with respect to the basis set, especially when the latter includes orbitals with an electron density far from the nucleus (*p*, *d*, and *f* orbitals). The advantages are that they are easy to compute and are consistent with, what may be called, “chemical intuition.”

Charges derived from the electrostatic potential are usually less dependent on the basis set. They are determined by positioning charges on each atom and minimizing the residual between the charge-induced and the quantum-chemical electrostatic potential using a least squares fit. The drawbacks are that these charges are sensitive to, for example, molecular configuration and how the electrostatic potential is sampled. The electrostatic potential is usually sampled outside the molecule, where other molecules can be expected to be present in simulation. This gives rise to problems with the numerical stability of the solution, *i.e.*, the set of partial atomic charges; the charge on buried atoms may be poorly determined. The calculated values become highly dependent on the molecular configuration and how the electrostatic potential is sampled. To overcome this limitation, Bayly *et al.*<sup>72</sup> developed the RESP charges where, a hyperbolic restraint is used to keep charges closer to zero and lower the statistical uncertainty for buried atoms. The RESP charges are used in conjunction with the AMBER forcefield.<sup>68</sup>

### 3.3 Monte Carlo

In Monte Carlo, the partition function is factorized into an ideal gas part and a configurational part. The objective is to integrate the partition function to determine the value of a property *A*. Monte Carlo methods use random numbers to generate configurations that follow the Boltzmann distribution:

$$p(\mathbf{r}^N, V) \propto \exp(-U(\mathbf{r}^N, V)/k_B T). \quad (3.7)$$

A basic Monte Carlo algorithm consists of the following steps:<sup>†</sup>

---

<sup>†</sup>This algorithm largely follows the original Monte Carlo algorithm by Metropolis *et al.*,<sup>61</sup> as described in the book by Frenkel.<sup>63</sup>

1. Generate a new configuration (1) by random displacement of one or more atoms in the previous configuration (0). If the *NPT* ensemble is simulated, the volume,  $V$ , must also be changed once in a while.
2. Evaluate the energy of the new configuration ( $U_1$ ).
3. Compare the energies  $U_1$  and  $U_0$  to determine whether the new configuration is accepted or whether the previous configuration is kept. The new configuration is accepted if  $U_1 \leq U_0$ . If  $U_1 > U_0$  it is accepted with a probability,  $p_{acc} = \exp(-(U_1 - U_0)/k_B T)$ .
4. Compute value of the property ( $A$ ) of interest.

The steps of the algorithm are repeated  $M$  times and the value of the thermodynamic property  $A$  is computed as the arithmetic mean:  $\langle A \rangle = 1/M \sum_{j=1}^M A_j$ . The uncertainty of the property can be estimated as  $\hat{\sigma}_A = \langle A^2 \rangle - \langle A \rangle^2$ . To get a representative value of  $\hat{\sigma}_A$ , the simulation must have reached equilibrium and be sufficiently long. A better estimate can be obtained by repeated simulations from different starting configurations.

Monte Carlo integration of the configurational partition function rests on the so-called ergodic hypothesis, which declares that all states of the systems must be reachable from any other state with a finite number of steps. This hypothesis imposes limitations on the first step of the algorithm. The way in which new configurations are generated must be compatible with the hypothesis.

To reproduce the correct (Boltzmann) distribution, Monte Carlo algorithms are often designed to satisfy a condition called detailed balance, whereby the number of moves from one state A to another state B must be equal to the number of reverse moves, when  $N$  tends to infinity. Detailed balance imposes restrictions on how new configurations are accepted (step 3). The condition of detailed balance is sufficient but not necessary; it is too strong. A necessary condition is that the number of moves from any state to state A should be counterbalanced by the number of moves from state A (at infinite sampling). If an algorithm satisfies detailed balance, it follows that it also satisfies the necessary condition. The reason why detailed balance is used is that violations of the condition are easy to prove, while it is difficult to ascertain whether an integration scheme fulfils or violates the necessary condition.

The algorithm above satisfies both the ergodic hypothesis and detailed balance and is guaranteed to produce the correct distribution of phase space points at infinite sampling, thereby yielding correct values for every equilibrium property. The values are correct in the sense that they depend only on the interactions in our model.

## 3.4 Molecular dynamics

Molecular dynamics is the integration of Newton's equations of motion for atoms with interactions described by a forcefield. Molecular dynamics is also a form

of importance sampling to integrate the partition function. The difference from Monte Carlo is that it samples from the full partition function. The dimensionality of the integral doubles, which may seem counterproductive. It appears, however, that molecular dynamics is a good way of using the momenta to create new configurations with the correct distribution (Boltzmann). Like Monte Carlo, molecular dynamics is guaranteed to produce correct results. In addition to the thermodynamic properties determinable by Monte Carlo, molecular dynamics can give information about dynamic properties, *e.g.*, diffusion coefficients.

A simple molecular dynamics integration scheme is:<sup>‡</sup>

$$\mathbf{r}_i(t+1) = \mathbf{r}_i(t) + \frac{\mathbf{p}_i(t)}{m_i}\Delta t + \frac{\mathbf{f}_i(t)}{2m_i}(\Delta t)^2 \quad (3.8)$$

$$\mathbf{p}_i(t+1) = \mathbf{p}_i(t) + \mathbf{f}_i(t)\Delta t \quad (3.9)$$

where  $\mathbf{f}_i(t)$  is the force acting on particle  $i$  at time  $t$ . The force is computed as  $\mathbf{f}_i = -\nabla_i U(\mathbf{r}^N)$ ; the gradient is evaluated with respect to the spatial coordinates of particle  $i$ . The time step,  $\Delta t$ , used with the integration scheme is crucial to the stability and efficiency of the method. Too long a time step yields unstable trajectories that impart unphysical behavior to the system. The shorter the time step, the longer the computational time required to achieve the same accuracy for the averages, so there is a desire to use the longest time step possible. Hydrogens have the lowest mass and will thus have the fastest motion and thereby limit the length of the time step. Three common remedies to this limitation are: the avoidance of explicit hydrogens, to use of constrained molecules where the bonds have constant length,<sup>73</sup> or the employment of the multiple time step algorithm by Tuckerman *et al.*<sup>74</sup> The OPLS forcefield has so-called *united atoms*, where carbon, nitrogen, oxygen, *etc.* have increased diameter and adjusted charge to compensate for the exclusion of hydrogens.

### 3.5 Expanded ensemble molecular dynamics

Expanded ensemble molecular dynamics is a hybrid Monte Carlo–molecular dynamics method for computation of solvation free energies. The method uses Monte Carlo steps to gradually insert and delete a solute from the bulk medium, while the configurational space is sampled using molecular dynamics. The difference in free energy between two states of a system can be computed by Monte Carlo. The free energy difference is computed from the relative probabilities of finding the system in the two states:

$$\Delta G = G_A - G_B \propto -\ln \frac{p_A}{p_B} \quad (3.10)$$

To compute solvation free energies, the two states are represented by the absence of a solute and the presence of a solute. In order to determine the

---

<sup>‡</sup>For integration schemes of practical use and discussions about numerical stability and other issues, reference should be made to the books by Allen<sup>62</sup> and Frenkel.<sup>63</sup>

(excess) free energy of solvation, the system must be able to move between these two states. Direct (Monte Carlo) insertion of a molecule into a dense medium has a very low probability of being accepted. The low acceptance ratio leads to low statistical accuracy. Higher acceptance probabilities can be achieved by introducing a coupling parameter,  $\alpha$ , which scales the solute interaction with the surrounding bulk medium:

$$\mathcal{H}_\alpha = \mathcal{H}^{\text{solvent}} + \mathcal{H}^{\text{solute}} + U_\alpha^{\text{solvent/solute}} \quad (3.11)$$

It is common to use linear scaling,  $U_\alpha^{\text{solvent/solute}} = \alpha U^{\text{solvent/solute}}$ . There is no interaction at  $\alpha = 0$ , and the solute interacts fully with the solvent at  $\alpha = 1$ . The partition function is expanded in terms of a set of  $\alpha$ -values,  $\alpha_0 = 0 < \alpha_1 < \dots < \alpha_M = 1$ :

$$Q = \sum_{m=0}^M \iiint \exp(-(\mathcal{H}_\alpha(\mathbf{r}^N, \mathbf{p}^N, \alpha_m) + PV)/k_B T + \eta_m) d\mathbf{p}^N d\mathbf{r}^N dV, \quad (3.12)$$

where  $\eta_m$  is the balancing factor of the biasing potential. Every so many MD steps a change of sub-ensemble is attempted according to Monte Carlo rules, *i.e.*,  $m \rightarrow m'$ , where  $m' \in \{m-1, m+1\}$ . The excess free energy is determined by a random walk along the  $\alpha$ -dimension and is computed from the relative probabilities of finding the system at  $\alpha = 0$  and  $\alpha = 1$  as:

$$\mu_{\text{ex}} = -k_B T \left( \ln \frac{p_M}{p_0} + \eta_0 - \eta_M \right). \quad (3.13)$$

The efficiency of the random walk is determined by the acceptance probabilities for moves between sub-ensembles and by their number. The relative probabilities converge faster if the number of sub-ensembles is low. Increased acceptance probability also enhances the convergence rate. There is a conflict between increasing the acceptance probability and decreasing the number of sub-ensembles, since the former increases with an increase in the latter. In practice, one strives to achieve roughly equal acceptance ratios in the range of 0.3 to 0.5 for all transitions.

The energy difference,  $\Delta E = \mathcal{H}_{\alpha_{m'}} - \mathcal{H}_{\alpha_m}$ , together with the difference in biasing potential,  $\Delta\eta = \eta_{m'} - \eta_m$ , determines the acceptance probability as  $p_{\text{acc}} = \exp(-\Delta E/k_B T - \Delta\eta)$ . The acceptance of such moves would be infinitesimally small were it not for the use of the biasing potential. As soon as  $\Delta E$  exceeds a few  $k_B T$ , the acceptance probability becomes very low.<sup>§</sup> However, if the biasing potential is roughly equal to  $-\langle\Delta E\rangle$ , then the acceptance probability is only governed by the spread in the distribution of  $\Delta E$ .

Initially, we do not know what values of the balancing factors to use. The values must be determined by means of simulation. The original scheme for

---

<sup>§</sup>A free energy barrier of 100 kJ/mol or more is not uncommon when a molecule is inserted in a dense liquid. This barrier corresponds to about 40  $k_B T$  at 25 °C. The probability of crossing such a barrier is  $4 \cdot 10^{-18}$  and a crossing will never be observed in practical simulations.

computing the balancing factors is to use a series of simulations where the factors are updated gradually until convergence:<sup>75, 76</sup>

$$\eta_k^{(i+1)} = \eta_k^{(i)} - \ln(p_k/p_0)^{(i)}, \quad (3.14)$$

where  $i$  is the number of the simulation. This works well if transitions are cheap and one can afford many simulation steps. With larger molecules and the use of molecular dynamics, simulations tend to equilibrate too slowly and it becomes difficult to judge whether one needs to insert an extra sub-ensemble where the acceptance probability is low, or whether it is just the value of the biasing factor that is off. A solution to the problem of finding balancing factors is presented in Paper II.

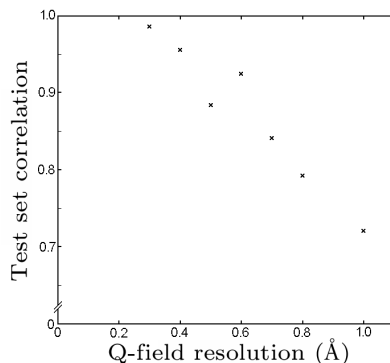
# Chapter 4

## Discussion

### 4.1 Paper I: The PCNN as a preprocessor for QSRR

The main part of the work on this paper was carried out at AstraZeneca in Södertälje for my Masters thesis. The scope of the work was to investigate whether the time series from a three-dimensional pulse-coupled neural network could be used for modelling quantitative structure-retention relationships. The main conclusion in Paper I is that it is possible, at least for the set of 24 steroid structures that were used for the investigation. Previous studies on the PCNN had not used the time series for regression modelling; it had only been used for classification. The fact that the time series can be used to classify objects in images does not necessarily mean that the time series contains information that is relevant for regression purposes. The extension of the neural network to three dimensions was trivial: it simply involves adding an extra index to the equations and implementing them in code. A very similar neural network model called the unified cortical model (UCM) had been presented to the chemometrics community about one year earlier.<sup>54</sup> That paper showed the segmentation ability of the network as an image processor on images of medical and chemical relevance. It also presented a series of images of the steroid  $17\beta$ -estradiol produced by a three-dimensional UCM. At the time I was hardly aware of that paper, mostly because it was in the peer-review process during the entire time span of my Masters thesis work. Nor did I realize its strengths. The UCM is basically a simplified version of the PCNN and is obtained by removing the so-called linking input from the PCNN.

The study was a success from the start. The very first PLS regression model showed good results from cross-validation and on the test set. After playing around with the PCNN in two dimensions, learning how its parameters affected the binary output images, the three-dimensional version was implemented. The parameters were studied once more, this time on Q-field images of steroids. The



**Figure 4.1.** The test set correlation for PLS models obtained by varying the resolution of the Q-field images.

Q-field had previously been used for QSRR studies using the same steroid data set.<sup>77</sup> In that study the regression model had been fitted directly on unfolded Q-field images using PLS, with the response being the logarithm of the capacity factor,  $\lg k'$ , for a particular LC separation. (The capacity factor is directly related to the chromatographic retention of a compound.) Our PLS regression models used the PCNN time series to predict  $\lg k'$ . At the time I was not fully aware of the implications and details of model selection and used the test set several times. The results in Paper I are therefore somewhat biased. In my defense, I would like to stress that the first regression model showed RMSEP values and test set correlations in parity with those presented in the paper. Thus, the bias not likely to be so large as to affect the conclusions of the study.

There are a lot of entirely different methods available that can model these relationships well, see, for instance, Ref.<sup>78</sup> The research field is known as QSAR for quantitative structure-activity relationships. The ‘A’ in QSAR can be replaced by almost any other letter, *e.g.*, P for property or R for retention as in our case. Free-Wilson analysis is based on functional group contributions to the free energy. Other common molecular descriptors may come from quantum chemical calculations. The solvent accessible surface area (SASA)<sup>57,58</sup> is one of the most important descriptors. The Figure 4.1 shows a trend where the test set predictions become better, the higher the resolution of the grid. At a given time step, the active neurons describe what can be interpreted almost as an iso-surface. With higher resolution this surface becomes better described. The almost-iso-surfaces create “positive” and “negative” analogues of solvent accessible surfaces and the number of active neurons corresponds to the area of the surface. The time series can be interpreted as a series of different SASA values. And since SASA is a good descriptor in QSAR, the PCNN time series can be expected to hold information that can be used for QSRR modelling.

More accurate results could most likely have been obtained with an existing

method, had prediction accuracy been the target of the study. As this was not the case, there is an opening for comparing the PCNN method for QSRR with existing state-of-the-art methods.

## 4.2 Paper II: Adaptive expanded ensembles

As stated in the introduction, the project started out with a much larger scope: we intended to determine octanol–water partition coefficients ( $\log P_{o/w}$ ) from expanded ensemble simulations for about thirty or forty drug-related molecules. The preliminary results reported by Lyubartsev *et al.*<sup>79</sup> were promising. The project seemed fairly straightforward and it seemed that it would only be a matter of putting many hours of hard work into it. Naturally, this was not the case. We ran into problems almost immediately, partly because I was inexperienced in molecular dynamics simulations in general and expanded ensemble simulations in particular, and partly due to limitations in the algorithms that we used.

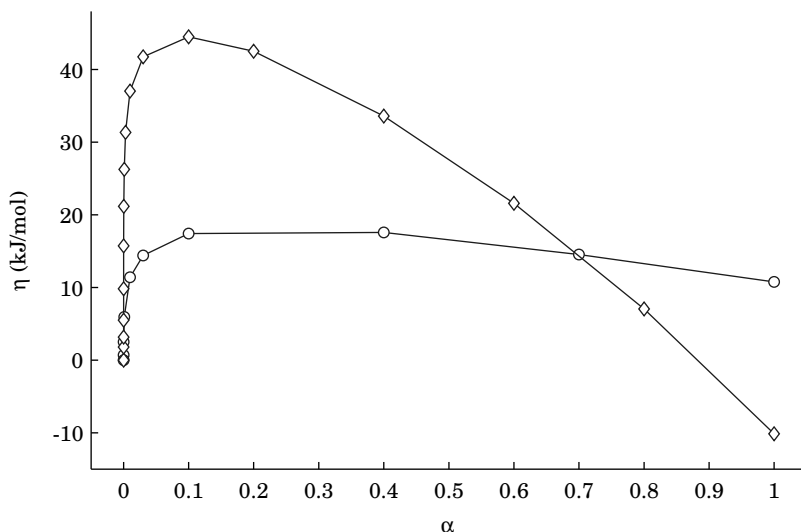
Looking back at the project from a wider perspective I have reached the conclusions that it is difficult to accurately determine free energies using molecular dynamics and/or Monte Carlo simulations, and that the particle insertion method that we have used may lead to convergence problems in the simulations and possibly even bias the results. Nevertheless, I hold Paper II to be perhaps my best work (with the possible exception being Paper V). The method presented in Paper II is a considerable improvement, which makes the expanded ensemble simulations both faster and easier to perform.

The problem that I encountered, when performing the expanded ensemble simulations, was that the balancing factors never converged to acceptably good values. This frustrating situation triggered thought processes on how to better estimate the balancing factors automatically, rather than to manually adjust them according to the updating scheme described in previous expanded ensemble papers.<sup>75, 76</sup>

The reason why the original updating scheme is not very good\* is that it disregards most of the information from the simulation when updating the balancing factors. The main drawback is that the updated value of a balancing factor is based on very poor statistics if the factor is far from optimal. The convergence rate therefore becomes excruciatingly slow. Another issue that enters the equation is the values of the insertion parameter. Poor values of the insertion parameters further slow down the convergence of the balancing factors. It becomes a practical problem to decide whether to insert an extra sub-ensemble or to just keep on changing the values of the balancing factors; you cannot tell what the reason for the lack of convergence is—it could be either the balancing factors or the insertion parameter, or even both. And since you are left guessing if you are not very experienced, there was a real need to find a better way to determine the balancing factors.

---

\*In defense of the original updating scheme, it must be said that it did work for almost ten years—probably due to a combination of skill and the nature of the studied systems.



**Figure 4.2.** The balancing factors,  $\eta$ , as functions of the insertion parameter,  $\alpha$ . Circles show the free energy profile for the dissolution of methane in water, while diamonds show the same profile for benzylamine in water.

In order to design automatic optimization of the balancing factors, we need to know where to find the information about the factors. Once you realize that the information is in the transition energies (easy) and that there is a fixed distribution of transition energies for given values of the balancing factors (trickier), the rest consists merely of technicalities that need to be solved, and these are discussed in the paper.

There are still problems to be solved with expanded ensemble molecular dynamics. Figure 4.2 shows what I think is *the problem* with expanded ensemble molecular dynamics. The free energy profile is very steep for small values of  $\alpha$  when methane is dissolved in water. This effect becomes much worse when the solute molecule grows bigger. The derivative of the free energy viewed as a function of  $\alpha$  becomes singular at  $\alpha$  equals zero as it tends to infinity. The problem derives from the particle insertion method, and we have ideas on how to correct this problem. This will possibly be addressed in future research.

So far, we have discussed problems specific to expanded ensemble molecular dynamics. Another issue is the validity of the molecular forcefields for computing free energies. The bonds, angles, torsion, improper torsions, etc. are probably good enough. But the use of partial atomic charges is more questionable. It turns out that the free energies are sensitive to the values of the charges. Atomic charges are not an observable and they can be (and are) determined many ways, *e.g.*, by Mulliken or Löwdin population analysis, by partitioning the electron density according to the atoms in molecules theory, or by the presently pop-

ular fitting of the electrostatic potential outside the molecules using restraint optimization. This leads to the conclusion that accurate determination of free energies is difficult, since the set of partial atomic charges is an important element, over which we do not have enough control. I think that there is more work to be done on the treatment of electrostatics in molecular simulation. In the future, I think we will see forcefields with other approximations to electrostatics than partial atomic charges.

## 4.3 Papers III and IV: Solutions to the peak alignment problem

The project on peak alignment started in the autumn 2002 with the development of peak alignment using reduced set mapping (PARS). Paper III presents PARS and compares it to dynamic time warping (DTW) and a complexity-reduced version of DTW. The purpose of this article was to prove the principle that the algorithm performed as well as could be expected. The performance was evaluated by comparing the algorithm with the complexity-reduced DTW and can be summarized as: PARS gave identical alignment in shorter time. The strong points in Paper IV are the proposed dendrogram alignment scheme and the fact that the paper demonstrates how useful PARS can be with chromatographic data in a classification context. The need for peak alignment is demonstrated in Section 2.8.

The discussion of Paper III and Paper IV will attend to the complexity of PARS compared to DTW and the question of target chromatograms/spectra vs. the dendrogram alignment scheme.

### 4.3.1 Complexity of PARS and DTW

In terms of computational complexity it was a major breakthrough to formulate the alignment problem in graph theory. This formulation makes PARS independent of the resolution of the data. A typical sample of 600 MHz FT-NMR data has about 65 000 data points. DTW has a complexity of  $\mathcal{O}(n^2)$ , with  $n$  being the number of data points. The complexity-reduced DTW has a complexity of  $\mathcal{O}(nw)$ , where  $w$  is the window size in data points. If the resolution of the data is increased, *i.e.*,  $n$  is increased while the data otherwise stay the same, the window size,  $w$ , must increase by the same factor as  $n$ . Thus, even the complexity-reduced DTW is  $\mathcal{O}(nw)$ . The complexity of PARS is  $\mathcal{O}(p \log p)$ , where  $p$  is the number of peaks in a sample. There are, of course, many more data points than peaks in a spectrum or chromatogram, so at least an order of magnitude is gained.  $p$  is independent of the resolution. Better instruments or higher sampling frequencies yield more densely spaced data points of identical samples and this will make DTW slower. All methods that work with the original data representation are affected in the same way as DTW.

The breadth first search in its naive form has a complexity of  $\mathcal{O}(2^p)$ . The trick that makes PARS so fast is to keep record of the solutions and eliminate all nonoptimal solutions as soon as they are encountered. This elimination improves the complexity of the breadth first search to  $\mathcal{O}(p)$ . The overall complexity of PARS is governed by a sorting step in the algorithm, which is  $\mathcal{O}(p \log p)$ .

The identical alignments was an expected result. The edge weights in PARS were designed to mimic DTW. The conclusion is that PARS is a significant improvement over DTW.

### 4.3.2 The target question

To align a sample, we need something to align its peaks to: a target. A target can be one of the spectra or chromatograms in the data set. If there are only two samples, we can use either as the target, the result being the same. With more than two samples, the choice of target becomes an intricate problem. It may affect the results in that peak assignments become different and target-dependent.

A sample, that is ideal for use as a target should satisfy two criteria: it should include every peak which occurs in the data set and the position of a peak in the sample should be the position of that peak averaged over the whole data set. This ideal target can normally not be found in the data set and we are forced to choose a less than ideal target from among our samples. A possible choice of target is the sample with the most peaks. This selection minimizes the violation of the first criterion. There is a risk, however, that the positions of the peaks in the target are bad with respect to the second criterion. Another choice of target may be the sample which is closest to the average over all the samples. But because the samples are unaligned this choice may be bad. It is not clear what being closest to the average sample signifies, unless the samples are aligned. A second reason why this may be a bad choice is that the chosen target will almost surely violate the first criterion to an unnecessarily extent. It is very unlikely that a target satisfying the first criterion will be close to the average sample.

Recursive target update (RTU) is a third way to address the target question. Any sample is chosen as the target. As the name implies, this target is updated with new peaks during the alignment. After aligning a sample to the target, any unmatched peaks in the sample are added to the target.

In Paper IV we propose a solution to the question of target selection by circumventing it. Since pairwise alignment is easy, it is used to create a hierarchical alignment scheme that we call “dendrogram alignment scheme.” The scheme is based on the same idea as RTU, but instead of a single alignment, every sample is aligned to intermediate targets several times before it is considered to be fully aligned. The foremost property of the dendrogram alignment scheme is that the solution to the alignment problem becomes unique. However, this desirable property comes at a cost. To align a data set,  $\mathcal{O}(N^2)$  alignments must be performed, instead of  $\mathcal{O}(N)$  if a fixed target or RTU is used. ( $N$  is the number of samples in the data set.)

To conclude, we have reduced the computational complexity of peak alignment by introducing PARS in Paper III. In Paper IV we sacrifice computational complexity to solve an issue of major concern in peak alignment.

## 4.4 Paper V: A measure of class separation

The first seeds of thought about the study presented in Paper V were planted during a discussion in spring 2004 on how to put a figure of merit on peak alignment of  $^1\text{H-NMR}$  data belonging to two classes.<sup>80</sup> The figure should reflect how well the two classes are separated, so that different algorithms could be compared. During the discussion several different measures were tried out. None of them were consistent with the data analyst's view of good and bad class separation. The measures infallibly gave an unintuitive ranking for some cases.

In the paper, the measure is defined from an equal probability criterion. Another possibility is to use equal risk as the criterion to define the measure. The  $\ln|\Sigma_1|$  and  $\ln|\Sigma_2|$  terms vanish from the equation for  $e(\mathbf{x})$  if the equal risk definition is used. Equal risk also makes the measure,  $M$ , more similar to the Fisher criterion,  $F$ , more specifically,  $M \in [\sqrt{F}/2, \sqrt{F}]$ . Another property is that the two Mahalanobis distances, of which the shortest defines the value of the measure, become equal. This property means that the measure has a unique definition and we don't need to fiddle around with choosing the minimum distance.

The most intriguing thing is that the linear classifier that we present in the paper could not be found in the literature. I was certain that it would be described in a book called "Discriminant Analysis and Statistical Pattern Recognition,"<sup>81</sup> but to my surprise, it was not. Reason tells me that somewhere someone has described the classifier already even though I have not been able to find it. If it exists, it must be described using a mathematical notation and with a wording that I cannot understand. Since only the existence and definition of the new linear discriminant are stated in Paper V, it would be interesting to study its properties in detail.



## Chapter 5

# Conclusions

The scope of this thesis is wide. Based on five papers the thesis spans four different subjects. The least common denominator of the papers is summarized in the title of the thesis: variance reduction. The practical work has consisted of developing new numerical methods in chemometrics and in molecular simulation.

Paper I attempts to achieve variance reduction within the field of quantitative structure-retention relationships by the use of pulse-coupled neural networks (PCNN) on three-dimensional images of molecules. In this paper variance reduction is perhaps not really achieved; however, we can conclude that the time series produced by the PCNN do contain quantitative information related to the capacity factor in chromatography. An interpretation of the values of the individual data points in the time series is that these are analogous to solvent-accessible surface areas—a well-reputed molecular descriptor.

In Paper II we reduce the variance of free energy calculations using the method of expanded ensemble molecular dynamics. This project was successful and now the free energy calculations are one step closer to being fully automated—just fire and forget. Not only is the method now easier to use, it also gives answers with higher accuracy in shorter time.

Peak alignment is all about variance reduction. Variance in the measured data that is unrelated to the information is removed and, thus, the information in the data is enhanced. Paper III and Paper IV covers aspects of peak alignment and introduces a new and efficient method to the field.

Paper V, the final and most theoretical, is concerned with quantifying variance reduction in a classification context. In classification one wants to minimize the within-class variation while keeping between-class variation high; the variance difference determines how well the classes are separated and how good a classifier is. In this paper, a new and accurate measure of this variance difference is defined and an algorithm to compute it is presented. Paper IV successfully uses this new measure.



# Acknowledgements

Först vill jag tacka min handledare Sven Jacobsson för ett prestigelöst och entusiasmerande handledarskap. Ett möte med *il Professore* gav alltid nya krafter, både till gamla och nya problem. Även om det ibland har känts som jag har varit ute och cyklat, har resultatet blivit bra och jag har lärt mig mycket under arbetets gång.

Ralf Torgrip för bra samarbete, många kebab-luncher och inte minst för att du korrläst detta manuskript till ögonen blött.

Aatto och Sasha för att ni välkomnat mig till gruppen på Fysikalisk kemi och gett mig möjligheten att samarbeta med er.

Rumskamraterna Jenny, Malin och Stina för att ni gjort A315 till något att se fram emot varje dag. Nu kan ni lugnt återuppta traditionen med godisburk på rummet...

Alla andra doktorander på Analytisk kemi: Petter, Johanna, Thorvald, Stina M, Helena I, Nana, Helena H, Leila, Christoffer, Caroline, Ragnar och nästan-doktoranden Annika, ni gör institutionen till ett toppenställe.

Jonas B och alla utflugna f.d. doktorander: Sindra, Yvonne, Kent, Ove, Magnus E, Anders Ch, Ludde, Magnus A, Gunnar och Kakan—bättre förr, ju förr desto bättre!? Snart sällar jag mig till er skara.

Professor Bo Karlberg, för gott samarbete kring kemometrikursen; det har gett fördjupade kunskaper i ämnets grunder.

Jonas R, för värdefulla layout-tips och alla anekdoter.

Carlo, tack för alla italienska delikatesser du bjudit på.

All övrig personal (Anders C, Anita, Anne-Marie, Björn, Conny, Davide, Håkan, Lena, Leopold, Roger, Ulrika, Ulla) för en öppen och trevlig atmosfär.

Till mamma Evy, pappa Kalle, syster Anna och alla mina vänner—utan er skulle livet bestå av arbete, mat och sömn, och det går ju inte för sig. Nä, fritiden ska vara fylld av aktiviteter: bridge, kilometertempo, fest, friluftsliv, slå-på-liten-boll-med-tillhygge-av-järn, m.m.

Catarina, för att allt är mycket roligare sen jag träffade dig. Nu ska det antligen bli slut på tangentbordsknattret till långt in på småtimmarna.

*Magnus Åberg, 24 oktober 2004.*



# Bibliography

- [1] J. Björklund, P. Tollbäck, C. Hiärne, E. Dyremark, and C. Östman. Influence of the injection technique and the column system on gas chromatographic determination of polybrominated diphenyl ethers. *Journal of Chromatography A* **1041**: 201–210 (2004).
- [2] Capability of detection—part 1: Terms and definitions. ISO 11843-1:1997.
- [3] Capability of detection—part 2: Methodology in the linear calibration case. ISO 11843-2:2000.
- [4] J. Inczedy, T. Lengyel, and A. Ure. *Compendium of Analytical Nomenclature (definitive rules 1997)*. Blackwell Science, 3 ed. (1998).
- [5] L. A. Currie. Detection: International update, and some emerging dilemmas involving calibration, the blank, and multiple detection decisions. *Chemometrics and Intelligent Laboratory Systems* **37**: 151–181 (1997).
- [6] R. G. Brereton. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. Wiley, Chichester (2003).
- [7] D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. de Jong, P. J. Lewi, and J. Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics: Part A*. Elsevier, Amsterdam (1997).
- [8] B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. de Jong, P. J. Lewi, and J. Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier, Amsterdam (1998).
- [9] H. Martens and T. Næs. *Multivariate Calibration*. Wiley, Chichester (1989).
- [10] J. E. Jackson. *A User's Guide to Principal Components*. Wiley, New York (1991).
- [11] R. J. Adcock. Note on the method of least squares. *The Analyst* **4**: 183–184 (1877). \*

---

\*The Analyst was a mathematical journal published by the Annals of Mathematics between 1874 and 1883 and should not be confused with the present-day journal that bears the same name and covers analytical chemistry.

- [12] R. J. Adcock. A problem in least squares. *The Analyst* **5**: 53–54 (1878).
- [13] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal* **6**: 559–572 (1901).
- [14] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**: 417–441, 498–520 (1933).
- [15] R. Tauler and B. Kowalski. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *Journal of Chemometrics* **9**: 31–58 (1995).
- [16] E. R. Mahoney and M. D. Finch. The dimensionality of body-cathexis. *Journal of Psychology* **92**: 277–279 (1976).
- [17] F. Despagne and D. L. Massart. Neural networks in multivariate calibration. *The Analyst* **123**: 157R–178R (1998).
- [18] P. B. Harrington and T. L. Isenhour. Compression of infrared libraries by eigenvector projection. *Applied Spectroscopy* **41**: 449–453 (1987).
- [19] T. A. Lee, L. M. Headley, and J. K. Hardy. Noise reduction of gas chromatography/mass spectrometry data using principal component analysis. *Analytical Chemistry* **63**: 357–360 (1991).
- [20] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 2 ed. (1989).
- [21] American Chemical Society. SciFinder Scholar 2004. A computer program that searches in reference databases Chemical Abstracts (CAS, American Chemical Society) and Medline (U.S. National Library of Medicine).
- [22] H. Wold. Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. In J. Gani (ed.), *Perspectives in Probability and Statistics: Papers in honour of M. S. Bartlett on the occasion of his sixty-fifth birthday*, pp. 117–142. Applied Probability Trust, Sheffield (1975).
- [23] M. Kleijnen, M. Wetzels, and K. de Ruyter. Consumer acceptance of wireless finance. *Journal of Financial Services Marketing* **8**: 206–217 (2004).
- [24] M. Y. Yi and F. D. Davis. Developing and validating an observational learning model of computer software training and skill acquisition. *Information Systems Research* **14**: 146–169 (2003).
- [25] L. Sonesten. Fish mercury levels in lakes—adjusting for Hg and fish-size covariation. *Environmental Pollution* **125**: 255–265 (2003).

- [26] S. de Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **18**: 251–263 (1993).
- [27] F. Lindgren and S. Rännar. Alternative partial least-squares (PLS) algorithms. *Perspectives in Drug Discovery and Design* **12/13/14**: 105–113 (1998).
- [28] M. Stone and R. J. Brooks. Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society B* **52**: 237–269 (1990).
- [29] J. Forshed, I. Schuppe-Koistinen, and S. P. Jacobsson. Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta* **487**: 189–199 (2003).
- [30] P. S. Belton, I. J. Colquhoun, E. K. Kemsley, I. Delgadillo, P. Roma, M. J. Dennis, M. Sharman, E. Holmes, J. K. Nicholson, and M. Spraul. Application of chemometrics to the  $^1\text{H}$  NMR spectra of apple juices: discrimination between apple varieties. *Food Chemistry* **61**: 207–213 (1997).
- [31] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**: 179–188 (1936).
- [32] S. Wold. Pattern recognition by means of disjoint principal components models. *Pattern Recognition* **8**: 127–139 (1976).
- [33] S. Wold and M. Sjöström. SIMCA: A method for analyzing chemical data in terms of similarity and analogy. *ACS Symposium Series* **52**: 243–82 (1977).
- [34] P. C. Mahalanobis. On tests and measures of group divergence. Part I: Theoretical formulæ. *Journal of the Asiatic Society of Bengal* **XXVI**: 541–588 (1930).
- [35] M. Sjöström, S. Wold, Å. Wieslander, and L. Rilfors. Signal peptide amino acid sequences in *Escherichia coli* contain information related to final protein localization. A multivariate data analysis. *EMBO Journal* **6**: 823–831 (1987).
- [36] M. Barker and W. Rayens. Partial least squares for discrimination. *Journal of Chemometrics* **17**: 166–173 (2003).
- [37] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Computer Science and Scientific Computing. Academic Press, San Diego, 2 ed. (1990).
- [38] J. Zupan and J. Gasteiger. *Neural Networks for Chemists*. VCH, Weinheim (1993).

- [39] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, London, 2 ed. (1999).
- [40] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society* **35**: 99–109 (1943).
- [41] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B* **36**: 111–147 (1974).
- [42] B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician* **37**: 36–48 (1983).
- [43] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**: 54–75 (1986).
- [44] C. Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society A* **158**: 419–466 (1995).
- [45] A. J. Miller. *Subset Selection in Regression*. Chapman and Hall, London (1990).
- [46] M. Bentivoglio. Life and discoveries of Santiago Ramón y Cajal. URL = “<http://nobelprize.org/medicine/articles/cajal/index.html>” (2004-10-15).
- [47] M. Bentivoglio. Life and discoveries of Camillo Golgi. URL = “<http://nobelprize.org/medicine/articles/golgi/index.html>” (2004-10-15).
- [48] H. Reichert. *Introduction to Neurobiology*. Thieme, Stuttgart (1992).
- [49] S. Grillner, Örjan Ekeberg, A. El Manira, D. P. Anders Lansner, J. Tegner, and P. Wallen. Intrinsic function of a neuronal network—a vertebrate central pattern generator. *Brain Research Reviews* **26**: 184–197 (1998).
- [50] R. Eckhorn, H. J. Reitboeck, M. Arndt, and P. Dicke. Feature linking via synchronization among distributed assemblies: Simulations of results from cat visual cortex. *Neural Computation* **2**: 293–307 (1990).
- [51] J. L. Johnson. Pulse-coupled neural nets: translation, rotation, scale, distortion, and intensity signal invariance for images. *Applied Optics* **33**: 6239–6253 (1994).
- [52] J. M. Kinser. Pulse-coupled image fusion. *Optical Engineering* **36**: 737–742 (1997).
- [53] H. S. Ranganath and G. Kuntimad. Iterative segmentation using pulse-coupled neural networks. In S. K. Rogers and D. W. Ruck (eds.), *Applications and Science of Artificial Neural Networks II*, vol. 2760, pp. 543–554. SPIE (1996).

- [54] J. M. Kinser, K. Waldemark, T. Lindblad, and S. P. Jacobsson. Multidimensional pulse image processing of chemical structure data. *Chemometrics and Intelligent Laboratory Systems* **51**: 115–124 (2000).
- [55] R. Sedgewick. *Algorithms in Modula-3*. Addison-Wesley, Reading (1993).
- [56] W. G. Richards. Theoretical calculation of partition coefficients. In V. Pliška, B. Testa, and H. van de Waterbeemd (eds.), *Lipophilicity in Drug Action and Toxicology*, vol. 4 of *Methods and Principles in Medicinal Chemistry*, pp. 173–180. VCH, Weinheim (1996).
- [57] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology* **55**: 379–400 (1971).
- [58] T. J. Richmond. Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. *Journal of Molecular Biology* **178**: 63–89 (1984).
- [59] H. Kubinyi. Free–Wilson analysis. Theory, application and its relationship to Hansch analysis. *Quantitative Structure-Activity Relationships* **7**: 121–133 (1988).
- [60] S. Nordholm and R. Penfold. The GvdW theory. A density functional theory of adsorption, surface tension, and screening. *Surfactant Science Series* **95**: 83–103 (2001).
- [61] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculation by fast computing machines. *Journal of Chemical Physics* **21**: 1087–1092 (1953).
- [62] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford Academic Press, Clarendon (1987).
- [63] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, San Diego (1996).
- [64] N. L. Allinger. Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms. *Journal of the American Chemical Society* **99**: 8127–8134 (1977).
- [65] N. L. Allinger, Y. H. Yuh, and J.-H. Lii. Molecular mechanics. The MM3 force field for hydrocarbons. 1. *Journal of the American Chemical Society* **111**: 8551–8566 (1989).
- [66] W. L. Jorgensen and J. Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society* **110**(6): 1657–1666 (1988).

- [67] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **4**: 187–217 (1983).
- [68] P. K. Weiner and P. A. Kollman. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *Journal of Computational Chemistry* **2**: 287–303 (1981).
- [69] A. R. Leach. *Molecular Modelling: Principles and Applications*. Addison Wesley Longman, Harlow (1996).
- [70] P. P. Ewald. Die berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik* **64**: 253–287 (1921).
- [71] C. Sagui and T. A. Darden. Molecular dynamics simulations of biomolecules: Long-range electrostatic effects. *Annual Review of Biophysics and Biomolecular Structure* **28**: 155–79 (1999).
- [72] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *Journal of Physical Chemistry* **97**: 10269–10280 (1993).
- [73] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **23**: 327–341 (1977).
- [74] M. Tuckerman, B. J. Berne, and G. J. Martyna. Reversible multiple time scale molecular dynamics. *Journal of Chemical Physics* **97**: 1990–2001 (1992).
- [75] A. P. Lyubartsev, A. A. Martsinovskii, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov. New approach to Monte Carlo calculation of the free energy: method of expanded ensembles. *Journal of Chemical Physics* **96**: 1776–1783 (1992).
- [76] P. Vorontsov-Velyaminov, D. A. Ivanov, S. D. Ivanov, and A. V. Broukhno. Expanded ensemble Monte Carlo calculations of free energy for closed, stretched and confined lattice polymers. *Colloids and Surfaces A* **148**: 171–177 (1999).
- [77] L. Nord, D. Fransson, and S. P. Jacobsson. Prediction of liquid chromatographic retention times of steroids by three-dimensional structure descriptors and partial least squares modeling. *Chemometrics and Intelligent Laboratory Systems* **44**: 257–269 (1998).

- [78] M. Haeberlein and T. Brinck. Prediction of water–octanol partition coefficients using theoretical descriptors derived from the molecular surface area and the electrostatic potential. *Journal of the Chemical Society, Perkin Transactions 2* (2): 289–294 (1997).
- [79] A. P. Lyubartsev, S. P. Jacobsson, G. Sundholm, and A. Laaksonen. Solubility of organic compounds in water/octanol systems. A expanded ensemble molecular dynamics simulation study of log P parameters. *Journal of Physical Chemistry B* **105**: 7775–7782 (2001).
- [80] J. Forshed, R. J. O. Torgrip, K. M. Åberg, B. Karlberg, J. Lindberg, and S. P. Jacobsson. A comparison of methods for alignment of NMR peaks in the context of cluster analysis. *Journal of Pharmaceutical and Biomedical Analysis* (2004). Submitted.
- [81] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York (1992).