Molecular dynamics of sense and sensibility in processing and analysis of data

Wassenaar, Tsjerk Andrys

Chapter 2

Data Processing and Analysis of Results using Statistical Methods

1 Introduction

The major objective of molecular dynamics simulations is to understand physicochemical, physiological or biophysical processes at the atomic level. Analysis of data usually involves either the investigation of the successive configurations and the structure of atomic motions or the comparison of properties to evaluate the effect of differences in the systems. Concerning the latter, one can think of the influence of mutations on the behaviour of a protein, or the effect of ligand binding. In this chapter methods are presented for both analysis of the structure of atomic fluctuations and the statistical comparison of simulation data. Most of the methods in this chapter find their origin in multivariate statistics, of which a brief introduction is given. Many of these methods are used in the following chapters or follow from the same principles.

There are several reasons to combine the different methods of analysis into this chapter. First, the aim of this chapter is to introduce the methods used subsequently, starting from the common reference frame of statistical analysis. Second, this chapter provides a toolbox of methods for the analysis and processing of data from molecular simulations. Third, the chapter highlights the assumptions and background of the methods for the interested reader. Fourth, while some of the techniques described in the following have been applied in molecular dynamics before, others are presented here for the first time.

This chapter is divided into several sections. First, some key elements of statistics are presented and the structure of the data is explored. Several methods for estimating the variance within a set of data are presented. In the second section, several methods are presented for the comparison of averages and fluctuations in properties obtained from molecular simulations. In the third section, multivariate analysis is discussed, including the principles of principal component analysis (PCA). These first three sections essentially form a brief review of (multivariate) statistical methods, with emphasis on the possible application to molecular simulations. The fourth section contains a more detailed discussion of PCA, including a number of new methods derived from PCA. These new methods were specifically developed for the analysis of interacting components (subsystems) in molecular simulations. They are given together with their mathematical derivation and justification, as appropriate. In the fifth section, a further extension of these methods is given, which are based on a combination of principal component analysis and multiple regression. These methods are developed for the investigation of relations between observables and atomic motions or fluctuations.

Finally, a number of methods are given for data reduction, based on the statistical techniques given in the other sections. These methods comprise a technique for the investigation of the nature and importance of rigid body motions in a macromolecular system. Besides, the use of principal component analysis as a means of data reduction is briefly discussed.
Note that throughout this chapter (and the thesis) vectors are represented by lowercase boldface, whereas matrices are represented as uppercase boldface. Normal variables are given in italics. The prime indicates that the transpose is taken and if not explicitly stated otherwise, the term vector refers to a column vector.

2 Statistics

The term statistics refers to the collection of quantitative data as well as to the branch of mathematics dealing with the collection, analysis, interpretation, and presentation of numerical data. In particular, it is used to refer to the study of likelihood and probability, often inferred from limited sample sizes. In this section some elementary concepts of statistics are introduced.

2.1 The moments of a distribution

Most distributions can be characterized in terms of their *moments*. The *k*th moment of the distribution is defined as the mean of the *k*th power of the deviations of the observed values from a fixed value:

$$\mu_k = \int_{-\infty}^{\infty} x^k p(x) dx \tag{2.1}$$

Here *p*(*x*) denotes the probability of *x*. The first moment is the *mean* of a distribution, also termed the *expectation* or the *expected value*, which is given by

$$\mathrm{E}(x) = \int_{-\infty}^{\infty} x \, p(x) dx = \mu \tag{2.2}$$

and corresponds to the centre of mass of the distribution. E is the expectation operator. Often the higher order moments are calculated with respect to the mean. Such moments are called the *central moments* of the distribution. The most important is the *second central moment*, which is the expected value of the square of the deviations about the mean, referred to as the *variance*

$$\begin{aligned} \mathrm{var}(x) &= \int_{-\infty}^{\infty} \left[x - \mathrm{E}(x)\right]^2 p(x) dx \\ &= \int_{-\infty}^{\infty} x^2 p(x) dx - \left[\mathrm{E}(x)\right]^2 \\ &= \mathrm{E}(x^2) - \left[\mathrm{E}(x)\right]^2 \\ &= \sigma_x^2 \end{aligned} \tag{2.3}$$

2.2 Statistical distributions

Among the most important concepts in statistics is the *Gaussian* or *normal distribution*, the density or *probability function* of which is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad (2.4)$$

and shown in Figure 2.1A.

One reason why this distribution is so important is related to the *central limit theorem*, which states that variates which are the sum of many independent effects tend to be normally distributed as the number of effects becomes large. One consequence of this theorem is that a variate, which is the average of a number of original variables, has a distribution which is more normal than the original distribution and is closer to the mean of the population.

Another important property of the normal distribution is that the sum of a set of normally distributed variates is itself normally distributed with a mean equal to $\sum_{i=1}^{n} a_i\mu_i$ and a variance of $\sum_{i=1}^{n} a_i^2\sigma_i^2$. This property is important for the methods used for data processing and analysis presented in the later sections.

Three other distributions are important to mention here. Each of these arises from a transformation of a set of normal variates. The first is the *chi-squared distribution*, which is the distribution of squared values. Let a set of variates $x_1,\ldots,x_n$ be independent samples from a normal distribution with a mean of zero and unit variance. In this case the chi-squared variate with *n* degrees of freedom is given by

$$\chi^2 = x_1^2 + x_2^2 + \ldots + x_n^2 \qquad (2.5)$$

which has a skewed distribution as shown in Figure 2.1B. The 100*α* percentage point of the chi-squared distribution with *n* degrees of freedom is denoted $\chi^2_{\alpha;n}$, where

$$\alpha = P(\chi^2 > \chi^2_{\alpha;n}) \qquad (2.6)$$

denotes the probability that the observed value exceeds the given percentage point.

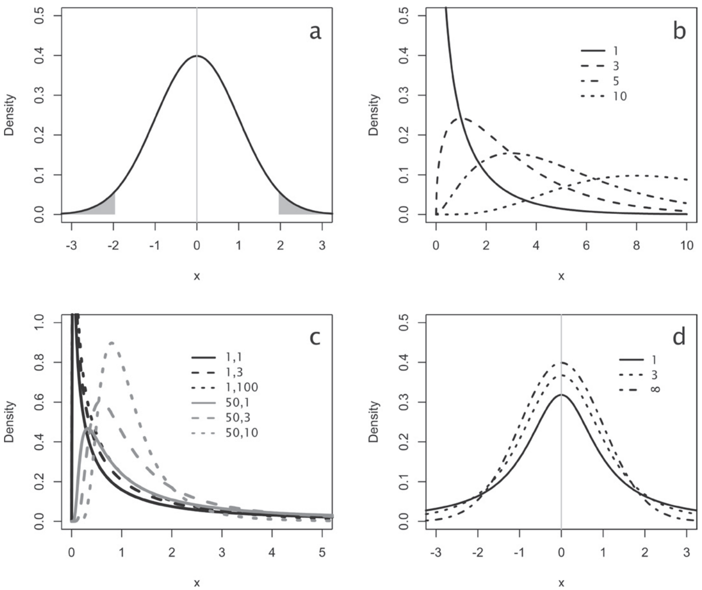

Figure 2.1: Statistical distributions. The densities of four important theoretical statistical distributions are shown. A. The normal distribution, B. The χ^2 distribution with degrees of freedom 1, 3, 5 and 10, C. The variance-ratio (F) distribution with degrees of freedom (numerator, denominator) 1, 1; 1, 3; 1, 100; 50, 1; 50, 3 and 50, 10, D. The t-distribution with degrees of freedom 1, 3, 5 and infinite

Another important distribution is obtained from the ratio of two independent chi-squared variates, divided by their respective degrees of freedom:

$$F = \frac{\chi_1^2 / n_1}{\chi_2^2 / n_2} \qquad (2.7)$$

This distribution is called the *variance-ratio* or F distribution. This distribution is shown in Figure 2.1C for a number of combinations of degrees of freedom. The 100α upper percentage point of the F-distribution with n_1, n_2 degrees of freedom is denoted $F_{\alpha;n1,n2}$, where $\alpha = P(F > F_{\alpha;n_1,n_2})$.
The third distribution to be introduced here is the *Student's t-distribution*, which is the distribution of a random variable t with n degrees of freedom, defined as the quotient of a standard normal variate z and the square root of an independent chi-squared variate divided by its n degrees of freedom.

$$t = \frac{z}{\sqrt{\chi^2 / n}} \qquad (2.8)$$

t is a dimensionless quantity and its distribution depends on the degrees of freedom parameter n. This distribution is particularly important when dealing with samples for which the variance s^2 is unknown and has to be estimated from the sample itself. The upper 100α percentage point is denoted $t_{\alpha;n}$ where $\alpha = P(t > t_{\alpha;n})$. For smaller sample sizes the t-distribution has more density in the tails, whereas the distribution tends to a normal distribution if the number of degrees of freedom is large. This can be seen in Figure 2.1D, where the Student's t-distribution is plotted for different degrees of freedom.

2.3 Confidence intervals and hypothesis testing

2.3.1 Hypotheses

An important area of statistical inference is concerned with the problem of testing the validity of a hypothesis regarding distribution functions and their parameters, or the parameters or components of a mathematical model. A hypothesis in statistical theory is generally a set of statements which are mutually exclusive and complementary. Usually, hypotheses are formulated such that the original hypothesis or *null hypothesis* reflects a situation of no effect or no difference. For example, when testing whether two samples originate from a single underlying distribution the null hypothesis, H_0, and the alternative hypothesis, H_1, can be formulated as

$$H_0 : \mu_1 = \mu_2 \qquad (2.9)$$

and

$$H_1 : \mu_1 \neq \mu_2 \qquad (2.10)$$

2.3.2 Errors

If one of two hypotheses reflects the true state, there are two possible types of error in the decision. An *error of the first kind*, or *Type I error*, is made when H_1 is declared true, when in fact H_0 is true. When H_0 is accepted, while H_1 reflects the true state, the error made is called an *error of the second kind*, or a

Type II error. The probability of making an error of the first kind is denoted α and is called the size of the test or the confidence level. The probability of an error of the second kind is denoted β, and its complement $1 - \beta$, is called the *power* of the test or decision rule. Power analysis is important for the design of experiments, but falls out of the scope of this work.

3 Data structure

3.1 Variables

3.1.1 Positions and momenta

For the analysis of results obtained from an MD experiment, and notably for statistical analysis, it is necessary to understand the structure of the data. The primary data obtained from a molecular dynamics simulation are a series of configurations of the system as a function of time, called the trajectory. The configuration of the system at a given time is defined by the positions **p** and momenta **q**, which are both sets of 3*N* coordinates. Each configuration can thus be thought of as a point in a 6*N*-dimensional space, called *phase-space*. The trajectory, given as a series of successive configurations, can consequently be thought of as a path through phase space. Note that the path obtained is only one of an infinite number of possible paths.

The trajectory can be represented as a two-dimensional matrix **X**, whose columns correspond to the configurations sampled during the simulation

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_M \end{pmatrix} \tag{2.11}$$

$$\mathbf{x}'_i = \begin{pmatrix} p_{1,x,i} & p_{1,y,i} & p_{1,z,i} & \cdots & p_{N,z,i} & q_{1,x,i} & \cdots & q_{N,x,i} & q_{N,y,i} & q_{N,z,i} \end{pmatrix} \tag{2.12}$$

Though the momenta may be useful, in particular when dealing with non-equilibrium simulations, in the following we focus primarily on the positions of the particles. Thus **x** will generally be used to denote the vector of size 3*N* of particle-coordinates. This vector will often be referred to as the *conformation* and the *conformational space* is used to refer to the region of phase-space corresponding to these coordinates.

A system in a molecular dynamics simulation has an underlying *probability density function* or simply *density function*, which describes the probability of finding a given configuration or conformation. This probability is linked to the energy of the configuration. The set of all possible configurations under a fixed set of external conditions is called the *ensemble*.

The ensemble can be seen as a *stationary process*. The distributions of such processes do not depend on the time or position from which the sampling started, but remain the same for all points in space and time.

According to the *ergodic hypothesis* the distribution obtained by sampling a single system over sufficiently long time properly reflects the distribution of the ensemble. The ergodic hypothesis is often given in a more casual formulation, namely that the time average of a single system over a sufficiently long period of time is equal to the ensemble average or the mean of the underlying distribution. However, ergodicity is not limited to the averages and it can be generalized to state that the moments of the distribution obtained from a single system over time are defined and equal to the moments of the probability density function of the ensemble

$$\mu_n = \int_{-\infty}^{\infty} x^n p(x) dx = \int_0^{\infty} x^n dt \tag{2.13}$$

In particular, it follows that the system has a mean configuration and a defined covariance matrix. Accordingly, these are the expectation values for the time average and sample covariance matrix obtained from a simulation, respectively.

3.1.2 Instantaneous properties

It is possible to interpret a given configuration of a system according to a set of rules and to obtain a value characterizing the state of the system in terms of a certain property of that system. For example, one can take the distance from the system to a reference system according to

$$d(t) = \sqrt{\sum_{i=1}^{3N} \left(p_i(t) - p_{ref,i}\right)^2} = \sqrt{\left(\mathbf{p}(t) - \mathbf{p}_{ref}\right)' \left(\mathbf{p}(t) - \mathbf{p}_{ref}\right)} \tag{2.14}$$

or alternatively as

$$\mathrm{RMSD}(t) = \sqrt{\frac{1}{3N} \sum_{i=1}^{3N} \left(p_i(t) - p_{ref}\right)^2} = \sqrt{\frac{1}{3N} \left(\mathbf{p}(t) - \mathbf{p}_{ref}\right)' \left(\mathbf{p}(t) - \mathbf{p}_{ref}\right)} \tag{2.15}$$

which is called the *root mean square deviation*. Other properties one can think of are the radius of gyration of a solute, the number of hydrogen bonds, according to given criteria for the donor – hydrogen – acceptor distance and angle, etcetera. These properties are characterized by the dependence on the conformation. As such, they are likely to change over time and for that reason will be referred to as *instantaneous properties* or sometimes as dependent properties. On the other hand the term *characteristic property* will be used for those properties describing an intrinsic property of the system, such as the number of each type of residue in the case of a protein. The experimentally observed value for an instantaneous property can sometimes be used as a characteristic property.

If an instantaneous property is defined for all configurations of an ensemble, the ergodic hypothesis holds for this property as it does for the system itself. Therefore, each of these properties has a defined population mean and variance, which are the expectation values of the property average and sample variance obtained from a trajectory.

3.2 Estimation of the mean and variance from simulation data

A primary goal of molecular dynamics or Monte Carlo simulations is to probe the statistical mechanical or thermodynamic ensemble. The ergodic hypothesis provides a basis for the use of the moments from the observed distribution as estimators for the distribution of the ensemble. In particular, in accordance with the casual formulation of the ergodic hypothesis, the time average

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.16}$$

is used as an estimate for the ensemble average or population mean μ (2.2). The time average is an *unbiased estimator*, since the expectation value $\mathrm{E}(\langle x \rangle)$ equals μ. Nevertheless, in practice the time scales of a simulation are limited and different simulations yield different time averages, reflecting the spread in the underlying distribution. In order to make correct inferences from an observed time average, it is necessary to have an estimate of the natural variation or variance (2.3). From the variance it is possible to derive the sampling error in the time average from a simulation.

The most obvious and intuitive estimator for the variance is $\hat{\sigma}^2$, which is given by

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2\right] \tag{2.17}$$

However, inspection of the expectation value shows that $\hat{\sigma}^2$ underestimates the true variance and is therefore a biased estimate. The proper estimate for the variance from a series of independent samples is given by

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2\right] \tag{2.18}$$

The factor $N - 1$ is introduced to correct for the use of the sample average, which is estimated from the same data, rather than the independent mean of the distribution. This factor is referred to as the number of degrees of freedom.

In the case of correlated data, such as typically obtained from molecular simulations, s^2 in the definition given above is also biased and underestimates the real variance. Because of correlations between subsequent sample points, the number of degrees of freedom is substantially smaller than $N - 1$. To obtain a proper, unbiased estimate it is necessary to compensate for such correlations.

Several methods have been proposed to deal with this problem. Here, three methods will be discussed. The first two of these are established methods, one of which is based on compensating for the correlation lengths and the other uses block averaging of data. These methods yield an estimate for the variance of the time average rather than for the population variance. The third method, based on a so-called bootstrap test, while not new, to the best of my knowledge has not been previously applied to obtain robust error estimates for data obtained from molecular dynamics simulations.

3.2.1 Variance estimation on correlated data

The first method here discussed was introduced in molecular dynamics independently by Schiferl[1] and by Straatsma *et al*.[2], but was originally proposed by Jenkins and Watts[3]. They considered a correlated time series x_i, $i = 1, \ldots, n$, with constant spacing and calculated the variance in the obtained time average by compensating for the correlations up to a certain correlation length. For the derivations of the estimates given in this section, the reader is referred to the original papers. The estimate of Straatsma *et al*. for the variance in the time average $\bar{x}$ is given by

$$\hat{\sigma}_{\bar{x}}^2 = \frac{1+2\tau}{n}\hat{\sigma}_x^2 \tag{2.19}$$

where $\hat{\sigma}_x^2$ is the (biased) estimator for the variance given by equation 2.17 and τ is the correlation length estimated by Straatsma *et al*. using

$$\tau = \sum_{k=1}^{K}\frac{c'_k}{\hat{\sigma}_x^2} \tag{2.20}$$

where

$$c'_k = \frac{1}{n-k}\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x}) \tag{2.21}$$

s the sum over the time correlations with lags running from 1 through *k*. Morales *et al*.[4] proposed a different estimator, which was proven by Dietrich and Dette[5] to always equal -0.5 when all possible correlation times were take into account. In turn, they proposed two other estimates, which were shown to be more robust than the one given by Straatsma *et al*. According to their results, the best estimator for the correlation length is

$$\hat{\tau} = \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) \hat{r}_k \tag{2.22}$$

where

$$\hat{r}_k = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x}_1)(x_{i+k} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n-k} (x_i - \bar{x}_1)} \sqrt{\sum_{i=1}^{n-k} (x_{i+k} - \bar{x}_2)}} \tag{2.23}$$

In this expression, which was first given by Jenkins and Watts[3], and by Kendall[6], $\bar{x}_1$ denotes the mean of the first *n* – *k* observations and $\bar{x}_2$ denotes the average over the last *n* – *k* observations of the series.

Apart from providing an estimate for the variance, the correlation length can also be used to estimate the number of independent data points in the series. This is not *n*, but is equal to $n/(1+2\tau)$. The factor 1 + 2τ has previously been termed the sampling ratio.

3.2.2 Block averaging in variance estimation

A different approach to estimate the sampling error or the variance in a time average obtained from simulation data was introduced in molecular simulations by Flyvbjerg and Petersen[7]. Rather than estimating the correlation length, they proposed the use of a block averaging method.

For the derivation and details of the method the reader is referred to the original paper. Note, the proposed method avoids the calculation of time correlations and in addition gives information about the quality of the estimate for the variance in the time average. The essence of the method is that the original data $x_1, \ldots, x_n$ is transformed into a new data set $y_1, \ldots, y_m$ of half the original size, such that

$$m = \tfrac{1}{2} n \tag{2.24}$$

and

$$y_i = \tfrac{1}{2}(x_{2i-1} + x_{2i}) \tag{2.25}$$

Then from the new data set, the variance can be calculated according to

$$s_y^2 = \frac{1}{m} \sum_{i=1}^{m} (y_i - \bar{y})^2 \tag{2.26}$$

and the estimate of the variance of the time average is given by

$$s_{\bar{x}}^2 = \frac{s_y^2}{m-1} \tag{2.27}$$

This procedure is repeated on the transformed data set, replacing $x_1, \ldots, x_n$ with $y_1, \ldots, y_m$ and *n* with *m*, until *m* = 2. The series of values for $s_{\bar{x}}^2$ will increase until a fixed plateau value is reached, which is the correct estimate for the variance in the time average. Flyvbjerg and Peterson have suggested using the confidence interval on $s_{\bar{x}}^2$ to determine whether the fixed point has been reached. For this they give the confidence interval as

$$\sigma_{\bar{x}}^2 \approx \frac{s_y^2}{m-1} \pm \sqrt{\frac{2}{m-1}} \frac{s_y^2}{m-1} = \frac{s_y^2}{m-1}\left(1 \pm \frac{1}{\sqrt{2(m-1)}}\right) \tag{2.28}$$

However, this is not the correct confidence interval. The variance has a χ^2 distribution, which is the distribution of sums of squares, with degrees of freedom equal to the number of independent data points minus one. For smaller samples sizes (i.e. for a smaller number of blocks) this distribution is notably asymmetric, and the correct (asymmetric) confidence interval is given by

$$\frac{(m-1)s_{\bar{x}}^2}{\chi_{1-\alpha/2;m-1}^2} \leq \sigma_{\bar{x}}^2 \leq \frac{(m-1)s_{\bar{x}}^2}{\chi_{\alpha/2;m-1}^2} \tag{2.29}$$

such that

$$\frac{s_y^2}{\chi_{1-\alpha/2;m-1}^2} \leq \sigma_{\bar{x}}^2 \leq \frac{s_y^2}{\chi_{\alpha/2;m-1}^2} \tag{2.30}$$

where α is the desired confidence level. For a large number of blocks the χ^2 distribution converges towards a normal distribution according to the central limit theorem.

The best approach to obtain an estimate for the variance using this method is by plotting the estimates together with the intervals against the number of transformations and infer at which point the fixed value is reached, taking the confidence interval into account.

3.2.3 Lifting by the bootstrap for estimation of the variance

In addition to the previous methods, it is also possible to obtain a robust estimate of the variance using a technique called bootstrapping[8], or similar methods such as the jackknife[9, 10]. Bootstrapping was originally proposed to make inferences about distributions from small samples. It exploits the similarity of the sample to the population. From the available sample an approximate or *bootstrap* population is reconstructed by replicating it a large number of times, typically thousands to millions. From the bootstrap population, a large number of samples can be drawn to estimate e.g. the mean of the population. Then the averages from the bootstrap samples can be displayed as a *bootstrap sampling distribution* of which the central 95% provides the desired confidence interval for the population mean. In the case of (correlated) time series care has to be taken, since the neglect of dependence between original observations can lead to incorrect answers[11]. The use of bootstrap techniques, as well as the jackknife, in the case of stationary processes has been discussed by Künsch[12].

Though not widely used in the field, bootstrapping is not new in molecular dynamics. For example, Knecht and Grubmüller have used the technique to make inferences about the energy necessary to tilt an α-helix into a presumed orientation[13]. The versatility and robustness of bootstrapping make this technique also very suitable to make inferences about the moments of the ensemble.

The only assumption made in the application of the bootstrap is that the distribution sampled in the simulation reflects the distribution of the ensemble. Then, each sample randomly drawn from the simulation data with replacement, meaning that each observation is put back in the pool such that it can be drawn multiple times, corresponds to a random sample from the underlying distribution. The average and variance obtained from such a sample are thus unbiased estimates of the mean and variance of the distribution. Repeating the random sampling from the simulation data a large number of times leads to distributions for the average and variance obtained from such a sample, from which a 95% confidence interval can be constructed, and the estimate of maximum likelihood, the mode of the distribution, can be inferred.

3.2.4 Degrees of freedom in correlated data

The degrees of freedom, denoted ν, are a measure for the number of independent observations contributing to an estimate, e.g. of the variance. For making statistical inferences and notably for the comparisons and the construction of confidence intervals, it is necessary to have a good estimate of the number of degrees of freedom. In the method of Straatsma *et al*.[2] or Jenkins and Watts[3], the degrees of freedom follow from the correlation length and should be taken as

$$\nu = \frac{n-1}{1+2\tau} \tag{2.31}$$

where n is the number of observations and τ is calculated from equation 2.20 or 2.22. The number of degrees of freedom according to the method of Flyvbjerg and Petersen[7] follows from the number of blocks at which the estimate of the variance levels off. Finally, if the bootstrap method was used, the number of degrees of freedom could be estimated using the correlation length according to 2.31. However, it is in general also possible to use bootstrap methods directly for comparing data sets or for the construction of confidence intervals. Two examples of such methods are given in the following section.

4 Comparing simulations

4.1 Comparing two simulations or comparing the results from a simulation with experiment

The objective of many studies using molecular dynamics is the comparison of the results from simulations with experimental results and/or the comparison of results obtained from simulations performed under different conditions with each other. Such comparisons between simulations are often made by looking at the average values obtained from single trajectories, whereas the comparison with experiments usually involved looking whether the average value from the simulation was consistent with an experimentally defined confidence interval. These approaches implicitly assume that the time average is a sufficiently good estimator of the ensemble average. However, when a simulation only covers a limited time, it is necessary to explicitly take the sampling error of the time average into account. In addition, the source of the sampling error in an experiment is quite different from the source of the sampling error in a simulation. The former is often not appropriate to be applied to the simulation results. To make a proper comparison with the results from simulations it is necessary to explicitly take the sampling error or the natural variance in the simulation into account as well as the degrees of freedom in the sample.

4.1.1 The Student's *t*-test

The most obvious way to account for the variance is to estimate the sampling error from the simulation data, according to any of the methods given in the previous section, together with the number of degrees of freedom ν. If it is assumed that the samples are normally distributed and the variances are equal, the averages can be compared using the well known Student's *t*-test. Let $\bar{x}$ and $\bar{y}$ denote the time averages from two different simulations and let s_x^2 and s_y^2 denote the estimates for the variance for each of the simulations, corresponding to *m* and *n* independent samples. Then the test statistic

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_p^2 \left(\frac{1}{m} + \frac{1}{n} \right)}} \tag{2.32}$$

has the Student's *t*-distribution with $\nu = m + n - 2$ degrees of freedom. s_p^2 denotes the pooled variance, which is the estimate of the population variance obtained from both samples taken together, and is given by

$$s_p^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m + n - 2} \tag{2.33}$$

For a two-sided test of size α the null hypothesis $H_0 : \mu_1 = \mu_2$ is rejected in favour of the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ if

$$|t| > t_{\alpha/2;n+m-2} \tag{2.34}$$

If there is *a priori* information that the first mean, if different, will be larger than the second, it is also possible to perform a one-sided test by evaluating

$$t > t_{\alpha;n+m-2} \tag{2.35}$$

4.1.2 The Welch *t*-test

Sometimes there are indications that the variances are different between the simulations. In particular, when comparing the results from a simulation to experiments, the difference in the source of the sampling error or variance will generally cause the variances to be different. In that case, a correction can be applied to the Student's *t*-test, as suggested by Welch[14] to allow assessment of the equality of means from distributions with unequal variances. The correction involves a modification of the degrees of freedom, which is taken to be

$$\nu = \frac{\left(s_1^2/m + s_2^2/n \right)^2}{\left(s_1^2/m \right)^2/(m-1) + \left(s_2^2/n \right)^2/(n-1)} \tag{2.36}$$

and a change in the calculation of the *t*-statistic

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} \tag{2.37}$$

which is evaluated for a two-sided test of size α against $t_{\alpha/2;\nu}$. The test is then referred to as a Welch *t*-test. Both the Student's *t*-test and the Welch *t*-test assume normal distributions. A basis for this assumption is the central limit theorem, which implies that with longer time scales a time average will be more narrowly distributed around the population mean (the ensemble average) and that this distribution will be more normal. However, when the time scale is limited, the distribution of time averages can be distinctly non-normal, rendering the above tests for comparisons invalid. In that case, one can turn to rank tests, such as the Wilcoxon-Mann-Whitney test for the comparison of two samples or to bootstrap tests for comparing samples.

4.1.3 The Wilcoxon-Mann-Whitney rank test

The Wilcoxon-Mann-Whitney test, or WMW test for short, is a rank-based test. For these tests the observations from the original samples are ranked, and the test is performed on the ranks, rather than on the original data. Instead of testing the equality of the means, these tests assess the equality of the medians of the distributions. If two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ are to be compared, the test statistic can be calculated as

$$U = \sum_{x_i < y_j} 1 + \sum_{x_i = y_j} \tfrac{1}{2} \tag{2.38}$$

For large sample sizes, such as the series of values obtained from a simulation, the *U*-statistic is approximately normally distributed under the null-hypothesis of equal medians, even if the distributions underlying the samples themselves are not. The mean of the distribution of the *U*-

$$\mu_U = \frac{nm}{2} \tag{2.39}$$

statistic is

$$\sigma_U^2 = \frac{nm(n+m+1)}{12} \tag{2.40}$$

and the variance is given by

$$z = \frac{U - \mu_U}{\sigma_U} \tag{2.41}$$

From these parameters an approximate *z*-score can be calculated using

$$|z| \geq z_{\frac{1}{2}\alpha} \tag{2.42}$$

which can be related to the desired percentage point of the normal distribution *N*~(0,1), such that the null hypothesis of equal medians is rejected for a test of size *α* if

In the case of smaller sample sizes, the statistic *U* can be checked against tables of the statistic with the given sample sizes.

The WMW test can be regarded the equivalent of performing a Student's *t*-test on the ranks of the pooled samples, rather than on the original samples.

4.1.4 Bootstrap hypothesis test on the equality of means

The equality of means or medians can also be assessed by using bootstrap methods as was briefly mentioned before. Two tests are particularly worth mentioning, namely the *bootstrap hypothesis test on the equality of means*[15] and the *bootstrap rank Welch test*[16]. A key element in both of these tests is the initial transformation of the data sets to make these satisfy the null hypothesis. Then from the transformed data sets a large number of samples are drawn and used to build the distribution of the desired test statistic. Finally, the probability that the observed value belongs to the obtained bootstrap distribution is assessed.

Consider two samples $x_1,\ldots,x_n$ and $y_1,\ldots,y_m$. These samples are shifted to satisfy the hypothesis according to

$$x_i^* = x_i - \bar{x} + c \tag{2.43}$$

and

$$y_i^* = y_i - \bar{y} + c \tag{2.44}$$

where $x_1,\ldots,x_n$ and $y_1,\ldots,y_m$ are the averages of each sample and *c* is the average of the combined samples or the common location parameter. From these transformed samples a large number of bootstrap samples are drawn and the corresponding *t*-distribution is built using

$$t(x^*, y^*) = \frac{\bar{x}^* - \bar{y}^*}{\sqrt{\frac{\sigma_{x^*}^2}{m^*} + \frac{\sigma_{y^*}^2}{n^*}}} \tag{2.45}$$

If the observed *t*-value from the original samples lies beyond the desired percentage point of this distribution, the difference between the means of the samples is considered to be statistically significant.

4.1.5 Bootstrap rank Welch test for stochastic equality

In addition to the previous tests, it is worth mentioning that recently a new test was introduced by Reiczigel *et al*. for the purpose of comparing stochastic distributions, which are often distinctly non-normal[16]. This test is also based on rank-testing, but uses bootstrapping to evaluate the hypothesis of equal medians. It requires no *a priori* information or assumptions with regards to the shape of the distribution. The test was specifically designed to test stochastic equality, reflected by the null hypothesis

$$H_0: \quad P(x < y) = P(x > y) \tag{2.46}$$

against the alternative

$$H_1: \quad P(x < y) \neq P(x > y) \tag{2.47}$$

First, given two samples **x** = ($x_1,\ldots,x_n$) and **y** = ($y_1,\ldots,y_m$), let $\mathbf{r}_x$ and $\mathbf{r}_y$ denote the sets of the ranks of the values of the pooled samples. Then let $\bar{r}_x$, $\bar{r}_y$ and s_x^2 , s_y^2 denote the averages and sample variances of the sets of ranks. Then the Welch statistic can be calculated for the ranked samples, according to 2.37, giving t_{RW} .

The distribution of t_{RW} under the null hypothesis is obtained by transforming the original data sets **x** and **y**, in a way similar to that previously given. However, rather than shifting both samples with respect to the sample means, the data is transformed by shifting one sample, say **y**, according to the median difference between the samples, such that

$$y_i^* = y_i + c \tag{2.48}$$

where *c* is the median of all combinations $x_i - y_j$.

Then from the two distributions **x** and **y*** a large number of samples are drawn, equal in size to the original samples, and the test statistic t^*_{RW} is calculated for each sample pair. Analogous to the previous method, if the observed value for t_{RW} from the original samples lies beyond the desired percentage point of this distribution, the difference between the medians of the samples is considered to be statistically significant and thus the underlying distributions are regarded stochastically different.

4.1.6 Confidence intervals for the difference between two simulations

$$\bar{x} - \bar{y} + s_p \sqrt{\frac{1}{n} + \frac{1}{m}}\, t_{a/2;n+m-2} \leq \mu_x - \mu_y \leq \bar{x} - \bar{y} + s_p \sqrt{\frac{1}{n} + \frac{1}{m}}\, t_{a/2;n+m-2} \qquad (2.49)$$

In addition to knowing whether two means are likely to be equal or not, it is often desired to know what the range of reasonable values is for the difference between the means of the simulations. To this end one can construct a *confidence interval*. Assuming a normal distribution, the confidence interval for the difference between the time averages $\bar{x}$ and $\bar{y}$ on a certain level α is given by

4.2 Comparing two sets of simulations

The tests described above to assess the similarity or statistical equality of two simulations are valid if the simulations have reached convergence. However, in most cases the time scales of the simulations are too limited to allow convergence to the degree necessary and it is not possible to make a robust estimation of the ensemble average and/or the sampling error. If two simulations start from slightly different starting configurations and sample for a limited time, they can end up in different localized regions of phase space and thus yield different estimators for the ensemble average and possibly for the sampling error, despite the fact that the two simulations sample from the same ensemble. In this case, one can not rely on the estimators and tests described above, since these will lead to a large error of the first kind; a false rejection of the null hypothesis that the samples are from the same distribution.

In this work it is proposed to take a different approach to improve the robustness of the test when convergence is not reached. Imagine a starting structure as a point on a high dimensional energy landscape. Usually, the starting structure will lie in a region with a certain slope, such that the landscape more or less determines which route a simulation takes. If a second simulation is started from a slightly different configuration, it will take a different path. But, as the underlying landscape is the same, these paths are likely to be similar. On the other hand, differences in the simulation conditions themselves can affect the underlying energy landscape. In this case a simulation started from a slightly different configuration will take a consistently different path over the landscape. This assumption is the basis for the tests suggested here and used in chapter 4.

From this perspective, the observed trajectory of a given length *l* is considered to be a sample from the population of all possible paths of length *l*, starting from a pool of similar starting configurations. This population will have a defined set of moments and it is thus possible to define a mean path and determine the variance of paths around this mean. Each simulation starting from a random configuration from within the 'starting pool' may be considered to be an independent sample from the distribution of possible paths. This sample can be characterized by choosing an instantaneous property of interest and taking the average over a pre-defined time window. From a sample of a number, e.g. five, independent simulations it is then possible to estimate the mean of the population and the variance. The number of degrees of freedom follows directly from the number of simulations. If the starting configurations are selected randomly from the starting pool, the individual simulations may be considered as independent.

If one has two samples obtained using this approach, originating from a common pool of starting configurations but with different simulation conditions, it is possible to assess the probability that these samples reflect a common underlying distribution of possible paths and should therefore be considered equal. If the starting conditions affect the outcome of the simulation, the local energy landscape is perturbed and there will be a consistent difference between the distributions of paths.

To assess the similarity or otherwise of samples of simulations thus obtained, any of the tests presented in the previous section can be applied, depending on assumptions and *a priori* or *a posteriori* knowledge of the distributions of paths.

4.3 Comparing two simulations with regards to fluctuations

The mean or the median is a logical choice to use as a basis for the assessment of equality of two simulations or sets of simulations. However, when dealing with dynamic systems, as is typically the case in molecular dynamics, the difference between two simulations performed under different conditions may well be in the rate and the amplitude of fluctuations, rather than in mean values. It can even be imagined that for some proteins, the difference between the active and the inactive state is the result of altered fluctuations, whereas the average structure remains more or less the same[17]. In such cases, it is obviously desirable to test the equality of the fluctuations in a certain property obtained from a molecular dynamics simulation.

Consider two sample variances, s_x^2 and s_y^2, which are obtained from two simulations or from two sets of simulations obtained according to any of the methods above. Let m and n correspond to the number of independent samples used to obtain these estimates. Then it is desired to test the null hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 \tag{2.50}$$

against the alternative

$$H_1 : \sigma_1^2 \neq \sigma_2^2 \tag{2.51}$$

at a certain level of confidence or test size α. To this end one can calculate the variance-ratio or F-statistic

$$F = \frac{s_1^2}{s_2^2} \tag{2.52}$$

If the null hypothesis is true, this statistic has the F-distribution (variance-ratio distribution) with $m-1$, $n-1$ degrees of freedom and the null hypothesis of equal variances is rejected for a two-sided test of size α if

$$F > F_{\alpha/2;m-1,n-1} \qquad \text{or} \qquad F < \frac{1}{F_{\alpha/2;n-1,m-1}} \tag{2.53}$$

From equation 2.52 it can be seen that for the calculation of *F* and the comparison of two trajectories it is also possible to use the total fluctuation rather than an unbiased estimate of the real sampling error if the number of degrees of freedom is the same. In that case, these cancel in the equation. However, one still needs to have an estimate of the degrees of freedom, to compare the calculated *F*-value with the desired percentage point from the appropriate distribution. The test is usually referred to as the *F*-test for equality of variances. It is worth noting that there have been a number of recent studies, in which the *F*-statistic was used to investigate atomic fluctuations in molecular dynamics simulations[18-20].

4.4 Comparing *k* sets of simulations

There are many situations in which it is desirable to compare multiple simulations or even multiple sets of simulations at once. Practical examples of such cases are the comparison of three different force fields by Price and Brooks[21] and the comparison of three different Generalized-Born Models by Fan and Mark[22]. Two other examples are given in Chapter 4 of this thesis, namely the comparison of simulations performed in different box types and the comparison of three different GROMOS force fields. Note, the first of these studies, from Price and Brooks, compared single trajectories, without considering the sampling error or the spread of the distribution. In other words, no robust statistical assessment was made of the validity of the conclusions.

In order to make a statistically robust comparison of *k* samples, either simulations or sets of simulations, with regard to several discrete levels of a certain condition, most of the tests and procedures outlined above can be generalized. For example, one might wish to test whether the samples are drawn from the same distribution with a common mean. In other words, to test the hypothesis

$$H_0 : \mu_1 = \ldots = \mu_k \tag{2.54}$$

against the alternative that some of the means are different. Under the assumption that the variances of the different samples are equal, this hypothesis can be tested using a technique called analysis of variance (ANOVA). The name appears counterintuitive, but indicates the principle of the method, i.e. comparing the variance between samples to the variance within samples.

4.4.1 Analysis of Variance

For ANOVA it is assumed that the observed data can be described by the mathematical model

$$x_{ij} = \mu + \tau_i + \varepsilon_{ij} \tag{2.55}$$

with $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, where μ is a location parameter common to all observations, τ_i is the additive effect peculiar to the *i*th treatment or condition and ε_{ij} is a normally distributed random variable with mean zero and variance σ^2. n_i is the number of observations of sample *i*. This model is an example of a general linear model underlying statistical design. Using this model, the hypothesis of equal means can now be written as

$$H_0 : \tau_1 = \ldots = \tau_k = 0 \tag{2.56}$$

and the alternative hypothesis is the general model for the observations.

To test the hypothesis, the total sum of squares (*SS*) obtained is decomposed into separate terms, corresponding to the terms in the model.

$$SS_{\text{Total}} = SS_{\text{Treatment}} + SS_{\text{Residual}} \tag{2.57}$$

which can be expanded to

$$\sum_{i=1}^{k}\sum_{j=i}^{n_i}(x_{ij} - \bar{\bar{x}})^2 = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 \tag{2.58}$$

Here $\bar{x}_i$ denotes the average for sample *i*, given by

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \tag{2.59}$$

and $\bar{\bar{x}}$ denotes the overall or grand average

$$\bar{\bar{x}} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} \tag{2.60}$$

Division of the sums of squares by their respective degrees of freedom gives the means of squares (MS), corresponding to the variance contributions due to the model components. Under the null-hypothesis of equal means, it is expected that the variance between treatments is equal to the variance within treatments. To test this, the *F*-statistic is calculated according to

$$F = \frac{MS_{\text{Treatment}}}{MS_{\text{Residuals}}} = \frac{N-k}{k-1}\frac{SS_{\text{Treatment}}}{SS_{\text{Residuals}}} \tag{2.61}$$

and H_0 is rejected for a test of size α if

$$F > F_{\alpha;k-1,N-k} \tag{2.62}$$

The calculations outlined above are usually summarized in an ANOVA table (Table 2.1). The comparison of different levels of a single treatment or condition (generally referred to as a factor) is called One-Way ANOVA. It is also possible to include more factors, each with a number of different levels. As an example, consider the general linear model for a series of experiments in which the effects of two factors, e.g. the boxshape and the application or not of rotational constraints (Chapter 4), were simultaneously assessed:

$$x_{ijk} = \mu + \tau_i + \nu_j + \gamma_{ij} + \varepsilon_{ijk} \tag{2.63}$$

Table 2.1: One-Way ANOVA. The meaning of the different elements is given in the text.

Source	*SS*	*df*	*MS*	*F*
Treatment	$SS_T = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2$	$k-1$	${}^{SS_T}\!/\!{}_{df}$	${}^{MS_T}\!/\!{}_{MS_{Res}}$
Residuals	$SS_{Res} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$	$\sum_{i=1}^{k}(n_i - 1)$	${}^{SS_{Res}}\!/\!{}_{df}$	
Total	$SS_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{\bar{x}})^2$	$\sum_{i=1}^{k} n_i$		

In this model, the meaning of μ, τ_i and ε_{ijk} is equal to their meaning in (2.55). However, the inclusion of one additional factor gives rise to two extra terms, an additive effect ν_j peculiar to the *j*th treatment of the second factor and an interaction term γ_{ij} denoting an additive effect specific for the combination of the *i*th treatment of the first factor and the *j*th treatment of the second factor under consideration. Accordingly, the null-hypothesis changes and now is comprised of three partial hypotheses:

$$\begin{aligned} &H_{0,a} : \tau_1 = \ldots = \tau_p \\ &H_{0,b} : \nu_1 = \ldots = \nu_q \\ &H_{0,c} : \gamma_{11} = \ldots = \gamma_{1q} = \gamma_{21} = \ldots = \gamma_{pq} \end{aligned} \qquad (2.64)$$

This is a Two-Way ANOVA model with interaction. The components and test statistics for the partial hypotheses are given in Table 2. It should be noted that the first effect to be tested in such a model is the interaction. When the interaction effect is not statistically significant at the given level α, the sum of squares of the interaction and the residual sum of squares have to be combined before testing the partial hypotheses $H_{0,\mathrm{a}}$ and $H_{0,\mathrm{b}}$.

The model can be further extended by the inclusion of additional factors. However, the interaction terms can become complicated and difficult to interpret. In addition, systematic testing of the effect of several factors often leads to a complicated and costly experimental setup.

As mentioned previously, ANOVA is applied under the assumption that the variances of the different samples are equal. The equality of variances is called homoscedasticity, the opposite of which is heteroscedasticity. This requirement can be tested after performing ANOVA, by means of diagnostic tests. For example, the residuals can be plotted to examine whether these follow a normal distribution. On the other hand, it is also possible to test the equality of the variances obtained from different samples before applying ANOVA, which is discussed later.

If the samples are (expected to be) heteroscedastic, there are several alternatives to ANOVA. For example, the Welch correction[14] (2.36) can be used (Welch ANOVA) or it is possible to choose a rank based alternative to the parametric ANOVA. In particular, the Kruskal Wallis test[23] is often

Table 2.2: Two-Way ANOVA with interaction. The meaning of the different elements is given in the text.

Source	*SS*	*df*	*MS*	*F*
Treatment A	$bn\sum_{i=1}^{a}(\bar{x}_i - \bar{\bar{x}})^2$	$a-1$	SS_A/df	MS_A/MS_{Res}
Treatment B	$an\sum_{j=1}^{b}(\bar{x}_j - \bar{\bar{x}})^2$	$b-1$	SS_B/df	MS_B/MS_{Res}
Interaction	$n\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2$	$(a-1)(b-1)$	SS_I/df	MS_I/MS_{Res}
Residuals	$\sum_{i=1}^{a}\sum_{j=i}^{b}\sum_{k=1}^{n}(x_{ijk} - x_{ij})^2$	$ab(n-1)$	SS_{Res}/df	
Total	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(x_{ijk} - \bar{\bar{x}})^2$	$abn-1$		

used as a rank based variant of one-way ANOVA, and can be regarded the equivalent of this test performed on the ranks of the pooled samples rather than on the original data. Likewise, a Friedman test[24] can be used as a rank based alternative to two-way ANOVA.

4.4.2 Simultaneous confidence intervals for sets of means: multiple comparisons

When for a set of samples the hypothesis of equal means is rejected, it is usually desired to know which of the sets lead to the rejection. This problem is referred to as that of *multiple comparisons*, which are aimed at making inferences about the members of a family of hypotheses. Such tests are constructed in such a way that the probability of making an error of the first kind will be at most α for the entire family.

There are several methods available to construct a set of simultaneous confidence intervals for multiple comparisons. The technique given here is due to Scheffé[25, 26]. Define a *contrast* of the parameters τ_i of the one-way model (2.55) as any linear function

$$\sum_{i=1}^{k} c_i \tau_i \tag{2.65}$$

the coefficients of which have the property

$$\sum_{i=1}^{k} c_i = 0 \tag{2.66}$$

Note that $\tau_1 - \tau_2$ and $3\tau_1 - (\tau_2 + \tau_3 + \tau_4)$ are contrasts, whereas $\tau_2 - (\tau_3 + \tau_4)$ is not. In the case of the one-way analysis of variance model the simultaneous confidence intervals with the joint coefficient $1 - \alpha$ for all contrasts of the τ_i have the form

$$\sum_{i=1}^{k} c_i \bar{x}_i - s_p \sqrt{(k-1) F_{\alpha;k-1,N-k} \sum_{i=1}^{k} \frac{c_i^2}{n_i}} \le \sum_{i=1}^{k} c_i \tau_i \le \sum_{i=1}^{k} c_i \bar{x}_i + s_p \sqrt{(k-1) F_{\alpha;k-1,N-k} \sum_{i=1}^{k} \frac{c_i^2}{n_i}} \tag{2.67}$$

where $\sum_{i=1}^{k} c_i \bar{x}_i$ is the sample estimate of the contrast $\sum_{i=1}^{k} c_i \tau_i$ and

$$s_p = \sqrt{\frac{SS_{\text{Residuals}}}{N-k}} \tag{2.68}$$

The null hypothesis

$$H_0 : \sum_{i=1}^{k} c_i \tau_i = 0 \tag{2.69}$$

is accepted at the level α if the simultaneous confidence interval for that contrast includes zero. Alternatives to the test for multiple comparisons (also called multiple contrasts) given above are the Bonferroni method[27] and the method of Tukey's honest significant differences[27, 28].

4.4.3 Testing the equality of variances from several samples

To assess whether several samples come from distributions with equal variances, there are two tests which are commonly applied. The first is Bartlett's test[29], which has the better performance, but is sensitive to departures from normality. The second, Levene's test[30], is less sensitive to departures from normality. Here, let it suffice to give the test statistic for Bartlett's test

$$T = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(n_i-1)\ln s_i^2}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1}-\frac{1}{N-k}\right)} \tag{2.70}$$

where $s_p{}^2$ is the pooled variance given by

$$s_p^2 = \frac{\sum_{i=1}^{k}(n_i-1)s_i^2}{N-k} \tag{2.71}$$

and which is rejected for a test of size α if

$$T > \chi^2_{\alpha;k-1} \tag{2.72}$$

5 Multivariate observations

Rather than looking at a single observable, characterizing a simulation over time, it is also possible to regard a number of observables simultaneously. In that case, each frame from a simulation yields a multidimensional observation vector, resulting in a more complete description of the system. The main difference with the univariate analysis is that the common source of the observables will generally lead to dependencies or correlation among the different dimensions. By taking the correlation structure into account in the analysis, the power of the tests can be increased. This is the basis for multivariate statistics. In this section a brief overview is given of multivariate statistics, starting from the distribution of an observation vector. Subsequently, several multivariate analogues or generalizations of the univariate tests given before will be presented and discussed in the context of molecular simulations.

Let **x** denote a p-dimensional observation vector, defined as

$$\mathbf{x}' = (x_1 \quad \cdots \quad x_p) \tag{2.73}$$

The elements of the vector **x** are assumed to be continuous unidimensional variables with density functions $f_1(x_1)$, …, $f_p(x_p)$ and distribution functions $F_1(x_1)$, …, $F_p(x_p)$. The joint distribution function of **x** is given by

$$F(x_1,\ldots,x_p) = P(X_1 \le x_1,\ldots,X_p \le x_p) \tag{2.74}$$

If this function is absolutely continuous it is possible to write

$$F(x_1,\ldots,x_p) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(u_1,\ldots,u_p)du_1\cdots du_p \tag{2.75}$$

where $f(u_1,\ldots,u_p)$ is the *joint density function* of the elements of **x**.

5.1 The moments of multivariate distributions

The moment generating functions for one-dimensional variables can also be applied to distributions of vectors or multivariate distributions. The first moment of the distribution of an observation vector **x** is simply the vector of the expectation values of the elements

$$E(\mathbf{x}') = [E(x_1),\ldots,E(x_p)] = [\mu_1,\ldots,\mu_p] = \boldsymbol{\mu}' \tag{2.76}$$

for which the estimate is the sample average vector

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k \tag{2.77}$$

This is an unbiased estimator since each individual element is an unbiased estimator for the mean of the univariate observable.

The second (central) moment of a multivariate distribution is the variance-covariance matrix or covariance matrix for short. The population covariance matrix is denoted **Σ**, and is given by

$$\boldsymbol{\Sigma} = \mathrm{E}\left\{[\mathbf{x}-\mathrm{E}(\mathbf{x})][\mathbf{x}-\mathrm{E}(\mathbf{x})]'\right\} = \mathrm{E}\left\{[\mathbf{x}-\boldsymbol{\mu}][\mathbf{x}-\boldsymbol{\mu}]'\right\} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_p^2 \end{pmatrix} \tag{2.78}$$

where σ_i^2 is the variance of variate *i* and σ_{ij} denotes the covariance between variates *i* and *j*, defined as

$$\begin{aligned}\sigma_{ij} &= \mathrm{cov}(x_i,x_j) \\ &= \mathrm{E}\left\{[x_i-\mu_i][x_j-\mu_j]\right\} \\ &= \frac{1}{N}\sum_{k=1}^{N}(x_{ik}-\mu_i)(x_{jk}-\mu_j) \\ &= \frac{1}{N}\sum_{k=1}^{N}x_{ik}x_{jk}-\mu_i\mu_j\end{aligned} \tag{2.79}$$

The sample covariance matrix, **S**, is obtained from the matrix of sums of squares and crossproducts, **A**, given by

$$\mathbf{A} = \sum_{k=1}^{N}(\mathbf{x}_k-\bar{\mathbf{x}})(\mathbf{x}_k-\bar{\mathbf{x}})' = \sum_{k=1}^{N}\mathbf{x}_k\mathbf{x}_k' - N\bar{\mathbf{x}}\bar{\mathbf{x}}' \tag{2.80}$$

This matrix is divided by the number of degrees of freedom to yield **S**

$$\mathbf{S} = \frac{1}{N-1}\mathbf{A} \tag{2.81}$$

The maximum likelihood estimate of the covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\mathbf{A} \tag{2.82}$$

is a biased estimator, analogous to the estimate for the population variance in the univariate case. Also note that the correlation between successive sampling points again adds more bias to the estimator and to obtain a better estimate from a single simulation the methods for error estimation as given in the previous section can be generalized.

If the (fluctuations in the) variates differ by an order of magnitude, it may be better to look at the correlation rather than at the covariance. The *correlation coefficient* of two variables x_i and x_j is defined as

$$\rho_{ij} = \frac{\operatorname{cov}(x_i, x_j)}{\sqrt{\operatorname{var}(x_i)\operatorname{var}(x_j)}} \tag{2.83}$$

and is a measure for the interdependence between the variables, which is invariant under changes of scale and origin. Accordingly, the population correlation matrix **P** is given by

$$\mathbf{P} = \mathbf{D}\left(\frac{1}{\sigma_i}\right)\mathbf{D}\left(\frac{1}{\sigma_i}\right) \tag{2.84}$$

where $\mathbf{D}\left(\frac{1}{\sigma_i}\right)$ denotes the diagonal matrix of the standard deviations of the variables.

5.2 The multivariate normal distribution

When the nature of the distribution of the multivariate data is important, notably for the testing of a hypothesis, it is generally assumed that the data is drawn from an approximately multinormal or multivariate normal distribution. There are several reasons for this assumption, which have been explained elsewhere. Do note that the assumption has to be rationalized and that in the case of obvious deviations from (multi)normality caution must be used.

The density of the multinormal distribution is given by

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{1}{2}p}|\mathbf{\Sigma}|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{2.85}$$

where $\mathbf{\Sigma}$ is a $p \times p$ symmetric positive definite matrix. To illustrate the idea of the multinormal distribution, the density function of a bivariate normal distribution is shown in Figure 2.2.

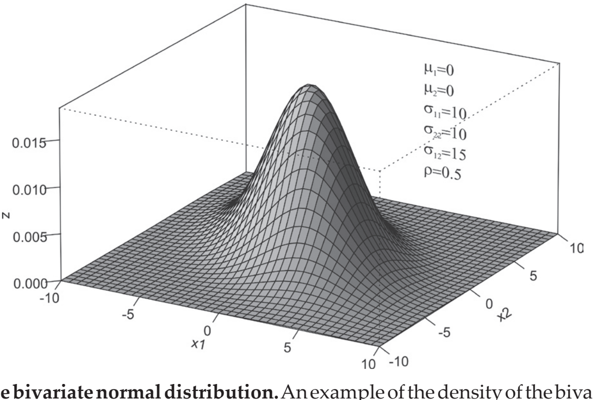

Figure 2.2: The bivariate normal distribution. An example of the density of the bivariate normal distribution is shown with mean value $\boldsymbol{\mu}' = (0 \quad 0)$, variances $\sigma_{11}^2 = 10$ and $\sigma_{22}^2 = 10$, covariance $\sigma_{12} = 15$ and correlation coefficient $\rho = 0.5$. The density is calculated according to (c.f. 2.85):

$$f(x) = \frac{1}{2\pi\sqrt{\sigma_{11}^2\sigma_{22}^2(1-\rho^2)}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_{11}^2}+\frac{(x_2-\mu_2)^2}{\sigma_{22}^2}-2\rho\frac{x_1-\mu_1}{\sqrt{\sigma_{11}^2}}\frac{x_2-\mu_2}{\sqrt{\sigma_{22}^2}}\right]\right)$$

5.3 Testing the equality of two mean vectors

Without going into detail in regards to the derivation, we here give the multivariate analogue of the Student's *t*-test for the comparison of two (multinormally distributed) multivariate samples[31]. The test statistic is Hotelling's T^2, which is given by

$$T^2 = \frac{N_1 N_2}{N_1 + N_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{2.86}$$

The quantity

$$F = \frac{N_1 + N_2 - p - 1}{(N_1 + N_2 - 2)p} T^2 \tag{2.87}$$

then has the variance ratio F distribution with degrees of freedom p and $N_1 + N_2 - p - 1$. Note that in the expression of T^2, **S** is the pooled covariance matrix given by

$$\mathbf{S} = \frac{1}{N_1 + N_2 - 2}(\mathbf{A}_1 + \mathbf{A}_2) \tag{2.88}$$

where **A** is the matrix of sums of squares and cross-products. The null hypothesis of equal mean vectors is accepted for a test of size α if

$$T^2 \le \frac{(N_1 + N_2 - 2)p}{N_1 + N_2 - p - 1} F_{\alpha;p,N_1+N_2-p-1} \tag{2.89}$$

5.4 The multivariate general linear model: MANOVA

When the objective is to test the equality of several sets of multivariate observations, assumed to be drawn from a set of multivariate normal distributions, these can be described by a (multivariate) general linear model[32-34], similar to the case of univariate observations. This forms a basis for the multivariate analysis of variance or MANOVA, which is used for the comparison of simulations in Chapter 4. To illustrate the linear model underlying MANOVA, consider the general linear model for one-way ANOVA (2.55) rewritten in matrix notation

$$\mathbf{x} = \mathbf{A}\boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{2.90}$$

where **x** is the vector of observations and $\boldsymbol{\varepsilon}$ is the vector of random errors. $\boldsymbol{\mu}$ is the parameter vector, given by

$$\boldsymbol{\mu}' = [\tau_1, \ldots, \tau_k, \mu] \tag{2.91}$$

and **A** is the design matrix, which assures that the *ij*th observation only involves the constant $\boldsymbol{\mu} + \boldsymbol{\tau}_i$. The design matrix is partitioned into k $n_i \times (k+1)$ submatrices. Although for multivariate observations the model becomes more complicated the principles are the same.

The procedure for MANOVA is quite similar to that for ANOVA. The outline given here corresponds to the analysis of a set of samples obtained with variations in two conditions, including the assessment of the statistical significance of interaction between the conditions. For convenience, the results are assumed to be tabulated with the columns corresponding to the different levels of the first condition and the rows corresponding to the different levels of the second condition. Then each cell contains the samples obtained under a specific combination of conditions.

Similar to the univariate two-way ANOVA with interaction, MANOVA starts with the calculation of the total sums of each cell (*C*), row (*A*) and column (*B*) as well as the grand total (*G*). The elements of each sum are given by

$$C_{ijh} = \sum_{k=1}^{n} x_{ijkh} \tag{2.92}$$

$$A_{ih} = \sum_{j=1}^{b}\sum_{k=1}^{n} x_{ijkh} \tag{2.93}$$

$$B_{jh} = \sum_{i=1}^{a}\sum_{k=1}^{n} x_{ijkh} \tag{2.94}$$

$$G_{h} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} x_{ijkh} \tag{2.95}$$

where *a* and *b* denote the number of columns and rows, respectively, and *n* denotes the number of observations per cell. The subscript *h* denotes the observable. These sums are then used to construct the *hypothesis matrices* $\mathbf{H}_A$, $\mathbf{H}_B$ and $\mathbf{H}_I$, according to the equations given in Table 2.3. $\mathbf{H}_A$, $\mathbf{H}_B$ and $\mathbf{H}_I$ are the matrices of sums of squares and cross products (SSCP) for the column treatments, row treatments and the interaction, respectively. These matrices correspond to the sums of squares SS_A, SS_B and SS_I calculated in univariate ANOVA and given in Table 2. In addition, the *residual covariance matrix* **E** is constructed, which corresponds to the residual sums of squares. In effect, this comes down to a decomposition of the total matrix of sums of squares and cross products $\mathbf{A}_{Tot}$ similar to the decomposition of the total sums of squares

$$\mathbf{A}_{Tot} = \mathbf{H}_A + \mathbf{H}_B + \mathbf{H}_I + \mathbf{E} \tag{2.96}$$

The matrix **E** is inverted and for each of the hypothesis matrices the product with the inverted matrix is taken, giving three matrices $\mathbf{H}_A\mathbf{E}^{-1}$, $\mathbf{H}_B\mathbf{E}^{-1}$ and $\mathbf{H}_I\mathbf{E}^{-1}$. Note the analogy with the ratio between the sums of squares due to a given source and the residual sums of squares in the univariate ANOVA. The evaluation of the equality is typically based on the characteristic roots from the resulting matrices.

Table 2.3: Two-Way MANOVA with interaction. The meaning of the different elements and the variables used in the equations are given in the text.

Source	*Matrix*	*General Element*
Treatments A (rows)	$\mathbf{H}_\mathrm{A}$	$h_{Auv} = \frac{1}{bn}\sum_{i=1}^{a} A_{iu}A_{iv} - \frac{G_u G_v}{abn}$
Treatments B (columns)	$\mathbf{H}_\mathrm{B}$	$h_{Buv} = \frac{1}{an}\sum_{j=1}^{b} B_{ju}B_{jv} - \frac{G_u G_v}{abn}$
Interaction	$\mathbf{H}_\mathrm{I}$	$h_{Iuv} = t_{uv} - h_{Auv} - h_{Buv} - e_{uv}$
Residuals	**E**	$e_{uv} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} x_{ijku}x_{ijkv} - \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} C_{iju}C_{ijv}$
Total	**T**	$t_{uv} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} x_{ijku}x_{ijkv} - \frac{G_u G_v}{abn}$

Common examples of test statistics available for the evaluation of MANOVA results are Roy's greatest root[35, 36], the Hotelling-Lawley trace statistic[37], the Pillai trace statistic[38] and Wilks lambda criterion[39]. In Chapter 4, Wilks lambda criterion is used for the evaluation of the MANOVA models. This statistic considers all characteristic roots and is easy to calculate. Wilks lambda is calculated according to

$$\Lambda = \frac{1}{\left|\mathbf{HE}^{-1} - \mathbf{I}\right|} \tag{2.97}$$

which is approximately distributed as a χ^2 variate.

5.5 Principal Component Analysis

A particularly useful technique in multivariate statistics is principal component analysis. The objective of this technique is to extract from a given data set a new set of latent or hidden variables, based on the dependence structure of the original variables. Especially when the observables are symmetric or when there is no *a priori* information regarding causality, principal component analysis can help to interpret and understand complex multivariate data.

Principal component analysis was first introduced as a method to fit planes by orthogonal mean squares[40]. Later it was recognized that the method is particularly useful for the analysis of correlation structures[41]. Principal component analysis is widely applicable and is used in a broad range of scientific fields, including molecular dynamics.

The use of principal component analysis in molecular dynamics focuses on revealing the structure of atomic fluctuations. This will be more thoroughly discussed in the next section of this chapter.

The aim of principal component analysis is to describe the original data in terms of new variables which are linear combinations of the original ones. Given the vector of p observables $\mathbf{x}$, the first principal component y_1 is the linear combination of the elements of $\mathbf{x}$

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{21}x_2 + \ldots + a_{p1}x_p \\ &= \mathbf{a}_1'\mathbf{x} \end{aligned} \tag{2.98}$$

for which the sample variance or *mean square displacement*

$$\begin{aligned} s_{y_i}^2 &= \sum_{i=1}^{p}\sum_{j=1}^{p} a_{i1}a_{j1}s_{ij} \\ &= \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 \\ &= l_1 \end{aligned} \tag{2.99}$$

is greatest for all coefficient vectors under the constraint that $\mathbf{a}_1'\mathbf{a}_1 = 1$. The vector $\mathbf{a}_1$ is referred to as the first eigenvector and the quantity l_1 is called the associated eigenvalue and is the largest characteristic root of the covariance matrix.

The definition of the *j*th principal component of a sample of *p*-variate observations is the linear construct

$$y_j = a_{1j}x_1 + a_{2j}x_2 + \ldots + a_{pj}x_p \tag{2.100}$$

whose coefficients are the elements of the characteristic vector of the sample covariance matrix $\mathbf{S}$ corresponding to the *j*th largest characteristic root l_j. If $l_i \neq l_j$, the coefficients of the *i*th and *j*th component are necessarily orthogonal; if $l_i = l_j$, the elements can be chosen orthogonal, although an infinity of such orthogonal vectors exists. The sample variance of the *j*th component is l_j, and the total variance in the system is thus equal to

$$\sum_{i=1}^{p} l_i = \operatorname{tr}\mathbf{S} \qquad (2.101)$$

The importance of the *j*th component can be measured by

$$\frac{l_j}{\operatorname{tr}\mathbf{S}} \qquad (2.102)$$

which indicates the fraction of the total variability in the system explained.

The *eigen decomposition* of the covariance matrix **S** into a matrix of eigenvectors **P** and a diagonal matrix of eigenvalues **D** can be written in matrix notation as

$$\mathbf{S} = \mathbf{PDP}' \qquad (2.103)$$

This can be rewritten to

$$\mathbf{S} = \mathbf{PD}^{\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}\mathbf{P}' = \mathbf{LL}' \qquad (2.104)$$

where $\mathbf{D}^{\frac{1}{2}}$ is the diagonal matrix of the square roots of the eigenvalues. **L** is referred to as the matrix of *loadings*, the columns of which indicate how much a variate from the original set is related to the different eigenvectors.

6 The structure of atomic fluctuations

In this section, the use of principal component analysis is discussed in the context of molecular simulations. First, the general method is explained in some detail, applied to a trajectory obtained from a simulation. Then an extension of this general form of principal component analysis is presented, which was derived in the course of this work to deal with characteristic motions and interactions between subunits in multimeric systems. At the end of this section, a brief discussion is given of a related technique called *maximum covariance analysis* and its possible application in molecular simulations.

The fluctuations of particles in a molecular dynamics simulation are by definition correlated due to interactions between the particles. The degree of correlation will vary and notably particles which are directly connected through bonds or lie in the vicinity of each other will move in a concerted manner. The correlations between the motions of the particles give rise to structure in the total fluctuations in the system and for a macromolecule this structure is often directly related to its function or (bio)physical properties. Therefore, the study of the structure of the atomic fluctuations can give insight in the behaviour of such macromolecules.

The application of methods to reveal and study the structure of atomic fluctuations commenced in 1991[42]. The original method and all of the methods now available are based on principal component analysis.

6.1 Principal component analysis in molecular simulations

When applied to the configurations obtained from a molecular dynamics simulation, the new variates or principal components correspond to linear combinations of individual atomic motions. In other words, a principal component extracted from a trajectory describes a collective motion of a set of particles in Cartesian space. It has been shown that in general a limited number of these collective motions account for the main motional features of a solute in a molecular dynamics simulation[43]. Furthermore, the smaller principal components correspond to further uncorrelated thermal motions of individual particles, which are usually not of interest. These smaller principal components can in general be disregarded and it has been suggested that the region of conformational space where these thermal motions occur can be separated from the region spanned by the limited set of eigenvectors describing the collective motions which are of importance. This latter region of conformational space has been termed the *essential subspace*, and it has been suggested that this is a unique feature of a given system. Hence, the study of the collective motions defining that space is sometimes referred to as *essential dynamics*.

The term essential is in a sense misleading, since there is no statistical ground on which such a separation can be based. The division is in fact largely arbitrary and is the equivalent of a scree plot.

To extract the principal components from a trajectory obtained from a molecular simulation the frames of the trajectory are first fitted onto a reference structure using a method of least-squares. This is done in order to remove overall translational and rotational motion.

6.1.1 Extracting the principal components from a simulation

Given a trajectory, represented as the *m* times *n* data matrix **X**, the covariance matrix is obtained by subtracting the average value for each row from each observation and taking the inner product of the resulting matrix

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = \frac{1}{N}\mathbf{YY}' \qquad (2.105)$$

where

$$\mathbf{Y} = (\mathbf{I} - \frac{1}{N}\mathbf{E})\mathbf{X} \qquad (2.106)$$

Note, this again underestimates the population covariance matrix by some scaling factor. This does not influence the nature and order of the principal components though, which are obtained from the covariance matrix. However, the sample eigenvalues do underestimate the theoretical eigenvalues from the population by the same scaling factor.

6.1.2 The interpretation and use of principal components

In most sciences the interpretation of principal components is difficult and sometimes impossible due to the nature of the original observables. Fortunately, the principal components obtained from the configurations from a molecular dynamics simulation have a distinct physical meaning and can be interpreted in terms of particles moving more or less in concert. In addition, the eigenvectors can be used to represent the part of the conformational space in which the events of greatest interest, namely the large scale global motions, take place. Projection of a trajectory onto these eigenvectors in turn gives an indication of the sampling of conformational space. Notably, the sampling along a single eigenvector can be followed by investigating the time evolution of the projection of the trajectory according to

$$y_i(t) = \mathbf{a}_i'(\mathbf{x}(t) - \overline{\mathbf{x}}) \tag{2.107}$$

where $y_i(t)$ is called the score of the frame at time t on component i. In molecular dynamics, the score thus obtained is often simply referred to as the projection.

For the first few eigenvectors, the projection often shows a time evolution resembling a cosine similar to that found for high-dimensional random diffusion[44]. In that regard, the cosine content of the projection can be used as a measure of the convergence of the collective motions of a system in simulation. By definition, the first few principal components describe the largest scale motions observed, which take the longest to converge in their sampling.

In addition to projecting a trajectory to obtain the scores, a trajectory can also be filtered to allow a visual inspection of the motion associated with a single component or a limited set of components. Note that the use of a set of components rather than individual components may be particularly useful in order to include possible non-linear correlations between motions along two or three eigenvectors. The filtered frame is then given by

$$\mathbf{x}^*(t) = \overline{\mathbf{x}} + y_i(t)\mathbf{a}_i + y_j(t)\mathbf{a}_j + \ldots = \overline{\mathbf{x}} + \mathbf{P}^*\mathbf{y}(t) \tag{2.108}$$

where $\mathbf{P}^*$ denotes the N×r matrix formed by r selected eigenvectors and $\mathbf{y}(t)$ is the vector of size r of the scores of the frame for the selected eigenvectors.

It is also possible to calculate degree to which the motion of a particular atom along a certain axis (atom-coordinate i) is correlated with a given principal component j. This correlation is given by the relation

$$r_{ij} = \frac{\mathrm{cov}(x_i, y_j)}{\mathrm{var}(x_i)\,\mathrm{var}(y_j)} = \frac{l_j a_{ij}}{s_i\sqrt{l_j}} = \frac{a_{ij}\sqrt{l_j}}{s_i} \tag{2.109}$$

and is often termed the *loading* of variable i onto component j. The matrix of loadings **L** is determined according to (cf. 2.104)

$$\mathbf{L} = \mathbf{P}\mathbf{D}^{\frac{1}{2}} \tag{2.110}$$

In molecular dynamics, colouring or arrows can be used to visualize the loadings for a certain component. For an example of this, the reader is referred to Figure 5.5.

6.1.3 Fitting a trajectory prior to principal component analysis

As mentioned, the frames of a trajectory are commonly fitted to a reference structure to remove overall translational and rotational motion in order to highlight intramolecular motions. In most cases the fit is performed according to the method of least-squares on all atoms to be used for further analysis. The fit can either be mass-weighted or with equal weights for all atoms.

When there is *a priori* information in regard to the nature of the collective motions of interest, it is also possible to prepare the data in such a way that these motions are enhanced by choosing an appropriate subset of atoms for the least-squares fit.

If the purpose of a study is to investigate the relation between two domains, linked by a mechanical hinge, then by fitting the trajectory to just one of the domains, the interdomain motions or the fluctuations of atoms belonging to the second domain become exaggerated. This will be reflected in the principal components with the first few primarily describing the motion of the second domain relative to the first. An example of this is given in Chapter 5, where principal component analysis is used to investigate changes in collective motions of a modular receptor (Death Receptor 5) in the presence and absence of ligand (TRAIL).

6.2 Principal component analysis applied to partitioned systems

6.2.1 General approach

Principal component analysis can also be extended to investigate relations between subsystems in a molecular dynamics simulation. The extension presented here involves expressing the observed motions in terms of the characteristic motions of the subsystems and the interaction or covariance of these collective motions between the subsystems.

Let A and B be two non-overlapping subsystems of a larger simulation system with *p* and *q* variates (atom-coordinates) respectively. Then from the respective trajectories of the two subsystems, *p* and *q* eigenvectors can be extracted, which represent collective motions of particles in subsystems A and B. Since these are part of a larger system, they influence each other and the collective motions of the subsystems need not be independent. Though in general dependencies limit the sampling obtained in a simulation, in many cases they are vital for understanding the system under study and reflect a physiological function.

The relationship between the two subsystems in terms of collective motions of the parts is here extracted as follows:

Let **S** be the covariance matrix from the variates of A + B. Then **S** is a $p + q \times p + q$ partitioned matrix

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} \tag{2.111}$$

Assuming that $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ are positive definite, i.e. for all non null **a** it holds that $\mathbf{a}'\mathbf{S}\mathbf{a} > 0$, there exist matrices $\mathbf{P}_{11}$ and $\mathbf{P}_{22}$ such that

$$\mathbf{S}_{11} = \mathbf{P}_{11}\mathbf{D}_{11}\mathbf{P}'_{11} \tag{2.112}$$

and

$$\mathbf{S}_{22} = \mathbf{P}_{22}\mathbf{D}_{22}\mathbf{P}'_{22} \tag{2.113}$$

where $\mathbf{P}_{ii}$ is the matrix of eigenvectors of $\mathbf{S}_{ii}$ and $\mathbf{D}_{ii}$ is the corresponding diagonal matrix of eigenvalues. Then the total covariance matrix is processed such that the diagonal blocks $\mathbf{S}_{ii}$ are diagonalized and the off-diagonal blocks reveal the covariances between the principal components of the subsystems

$$\begin{aligned}\mathbf{P}'\mathbf{S}\mathbf{P} &= \begin{pmatrix} \mathbf{P}'_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}'_{22} \end{pmatrix}\begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}\begin{pmatrix} \mathbf{P}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{22} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{P}'_{11}\mathbf{S}_{11}\mathbf{P}_{11} & \mathbf{P}'_{11}\mathbf{S}_{12}\mathbf{P}_{22} \\ \mathbf{P}'_{22}\mathbf{S}_{21}\mathbf{P}_{11} & \mathbf{P}'_{22}\mathbf{S}_{22}\mathbf{P}_{22} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{D}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{D}_{22} \end{pmatrix}\end{aligned} \tag{2.114}$$

Division of the elements of $\mathbf{C}_{12}$ by the square roots of the corresponding eigenvalues gives the correlation matrix $\mathbf{R}_{12}$

$$r_{12,ij} = \frac{c_{12,ij}}{\sqrt{\lambda_{11,i}\lambda_{22,j}}} \tag{2.115}$$

$$\mathbf{R}_{12} = \mathbf{D}_{11}^{-\frac{1}{2}}\mathbf{C}_{12}\mathbf{D}_{22}^{-\frac{1}{2}} \tag{2.116}$$

Here the matrix $\mathbf{D}_{ii}^{-\frac{1}{2}}$ denotes the diagonal matrix of the roots of the reciprocal eigenvalues.
In molecular dynamics, the matrices $\mathbf{S}_{12}$ and $\mathbf{R}_{12}$ give the covariances and correlations between collective motions from subsystem A and subsystem B. These can be valuable for the investigation of the effect of either subsystem on the other. Practical applications of this method are the study of the effect of two protein domains on each other, or the effect of a specific cofactor or ligand on the characteristic motions of a protein. With regards to the correlation, it should be noted that a high value need not be indicative of a significant relation, as the fluctuations involved can be small.
The procedure outlined above can easily be extended to investigate relationship between larger sets of subsystems. This involves extending the 2 × 2 partitioned matrices to a set of $n \times n$ matrices, with n being the number of subsystems to include.

6.2.2 Identical subsystems

In many cases, e.g. haemoglobin, GroES/GroEL and MscL, the functional multimeric protein consists of a number of identical subunits, which are indistinguishable with respect to their environment (Figure 2.3). The same is true for molecules of bacteriorhodopsin within the purple membrane (Chapter 7). In this case it is possible to transform the system such that all of the subunits sample the same configurational space. In other words, each subsystem is sampling from its own distribution with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$, but these mean vectors and covariance matrices are equal among the different components, except for their orientation with respect to the common coordinate system. This means that each of the subsystems can be transformed such that each of the instances of a subsystem is an observation from a common distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
Let A and B denote the two subunits of a homodimeric, symmetric protein. Then let **a** and **b** denote the configurational observation vectors from these systems. Because the subunits are equal except for their orientation, the mean vectors can be interrelated according to

$$\boldsymbol{\mu}_a = \mathbf{M}\boldsymbol{\mu}_b + \mathbf{d} \tag{2.117}$$

with **M** a rotation matrix and **d** a shift vector**.** The covariance matrices are related accordingly

$$\begin{aligned}
\boldsymbol{\Sigma}_a &= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{a}_i - \boldsymbol{\mu}_a)(\mathbf{a}_i - \boldsymbol{\mu}_a)' \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{Mb}_i + \mathbf{d} - \mathbf{M}\boldsymbol{\mu}_b - \mathbf{d})(\mathbf{Mb}_i + \mathbf{d} - \mathbf{M}\boldsymbol{\mu}_b - \mathbf{d})' \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbf{M}(\mathbf{b}_i - \boldsymbol{\mu}_b)(\mathbf{b}_i - \boldsymbol{\mu}_b)'\mathbf{M}' \\
&= \mathbf{M}\left(\frac{1}{N}\sum_{i=1}^{N}(\mathbf{b}_i - \boldsymbol{\mu}_b)(\mathbf{b}_i - \boldsymbol{\mu}_b)'\right)\mathbf{M}' \\
&= \mathbf{M}\boldsymbol{\Sigma}_b\mathbf{M}'
\end{aligned} \tag{2.118}$$

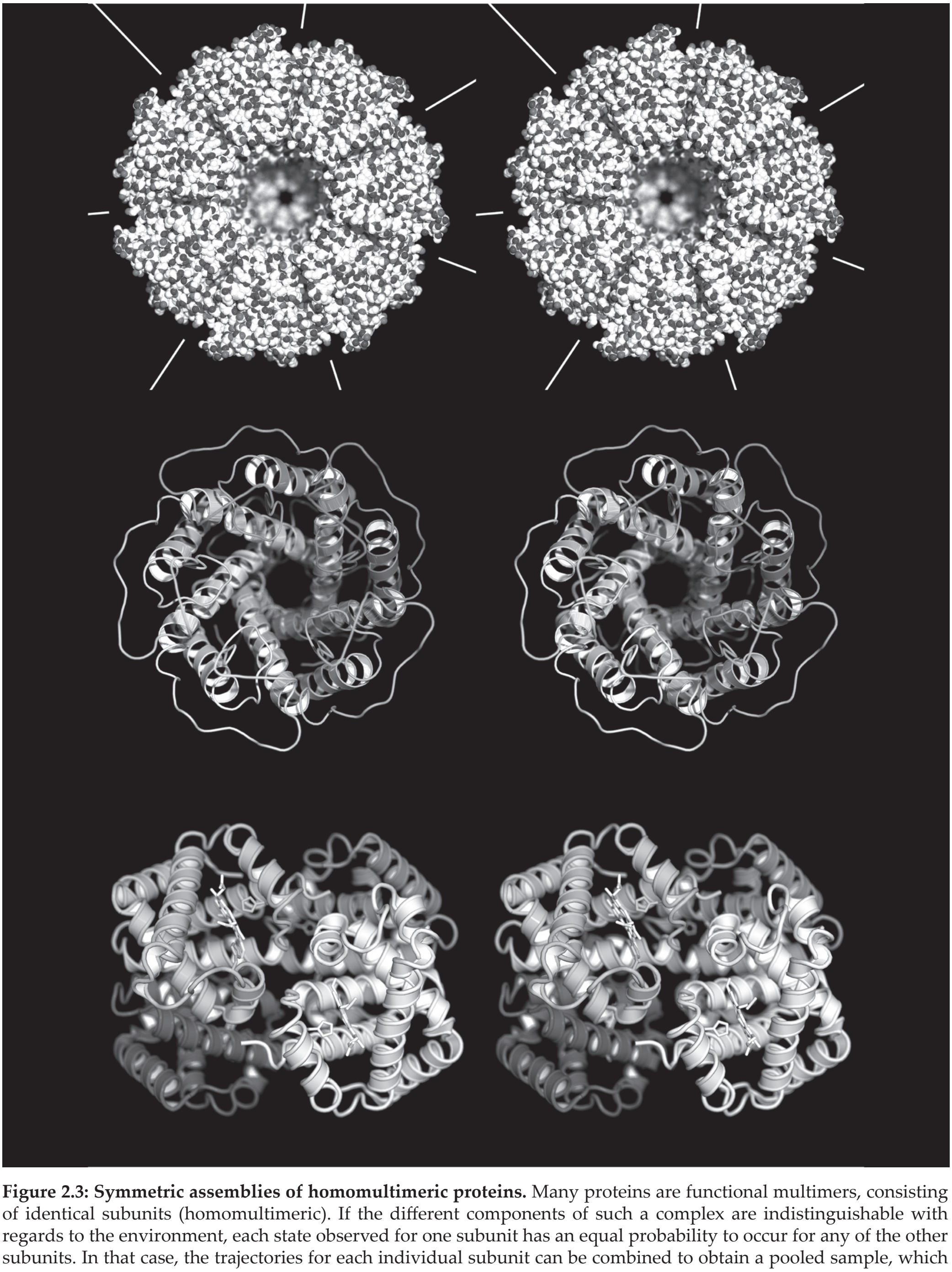

Figure 2.3: Symmetric assemblies of homomultimeric proteins. Many proteins are functional multimers, consisting of identical subunits (homomultimeric). If the different components of such a complex are indistinguishable with regards to the environment, each state observed for one subunit has an equal probability to occur for any of the other subunits. In that case, the trajectories for each individual subunit can be combined to obtain a pooled sample, which gives a better estimate for the ensemble. These properties can be used explicitly in the design of principal component analysis aimed at investigating inter domain motions as well as global motions. Three examples of multimeric systems to which this applies are top: the heptameric chaperone GroEL/GroES (lines indicate the symmetry), middle: the pentameric mechanosensitive gating channel MscL and bottom: the tetrameric oxygen carrier hemoglobin.

In practice, as neither $\boldsymbol{\mu}_\mathrm{a}$ nor $\boldsymbol{\mu}_\mathrm{b}$ is known, the best way to assure that they are transformed to have the same orientation is by fitting the frames of the subsystems onto a common reference structure. Let $\mathbf{M}_a$ and $\mathbf{M}_b$ denote the matrices which relate both subunits to a reference structure, such that

$$\mathbf{a} = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_{3N} \end{pmatrix} = \mathbf{M}_a\mathbf{r} \tag{2.119}$$

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_{3N} \end{pmatrix} = \mathbf{M}_b\mathbf{r} \tag{2.120}$$

Then let **c** denoted the compounded vector consisting of **a** and **b**, such that

$$\mathbf{c} = \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{M}_a\mathbf{r}_a \\ \mathbf{M}_b\mathbf{r}_b \end{pmatrix} = \begin{pmatrix} \mathbf{M}_a & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_b \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \mathbf{M}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \tag{2.121}$$

where the subscripts of **r** indicate that the subsystems do not need to be equal for a given time *t* or observation *i*. However, in both cases they originate from the same distribution with mean vector $\boldsymbol{\mu}_r$ and covariance matrix $\boldsymbol{\Sigma}_r$. The covariance matrix for **c** is defined by

$$\boldsymbol{\Sigma}_c = \frac{1}{K}\sum_{i=1}^{K}(\mathbf{c}_i - \boldsymbol{\mu}_c)(\mathbf{c}_i - \boldsymbol{\mu}_c)' \tag{2.122}$$

and can be expressed in terms of the subsystems to give

$$\begin{aligned}
\boldsymbol{\Sigma}_c &= \frac{1}{K}\sum_{i=1}^{K}\mathbf{M}\left(\begin{pmatrix} \mathbf{r}_a \\ \mathbf{r}_b \end{pmatrix}_i - \begin{pmatrix} \boldsymbol{\mu}_r \\ \boldsymbol{\mu}_r \end{pmatrix}\right)\left(\begin{pmatrix} \mathbf{r}_a \\ \mathbf{r}_b \end{pmatrix}_i - \begin{pmatrix} \boldsymbol{\mu}_r \\ \boldsymbol{\mu}_r \end{pmatrix}\right)'\mathbf{M}' \\
&= \mathbf{M}\left(\frac{1}{K}\sum_{i=1}^{K}\begin{pmatrix} \mathbf{r}_a - \boldsymbol{\mu}_r \\ \mathbf{r}_b - \boldsymbol{\mu}_r \end{pmatrix}\begin{pmatrix} \mathbf{r}_a - \boldsymbol{\mu}_r \\ \mathbf{r}_b - \boldsymbol{\mu}_r \end{pmatrix}'\right)\mathbf{M}' \\
&= \mathbf{M}\begin{pmatrix} \frac{1}{K}\sum_{i=1}^{K}(\mathbf{r}_a - \boldsymbol{\mu}_r)(\mathbf{r}_a - \boldsymbol{\mu}_r)' & \frac{1}{K}\sum_{i=1}^{K}(\mathbf{r}_a - \boldsymbol{\mu}_r)(\mathbf{r}_b - \boldsymbol{\mu}_r)' \\ \frac{1}{K}\sum_{i=1}^{K}(\mathbf{r}_b - \boldsymbol{\mu}_r)(\mathbf{r}_a - \boldsymbol{\mu}_r)' & \frac{1}{K}\sum_{i=1}^{K}(\mathbf{r}_b - \boldsymbol{\mu}_r)(\mathbf{r}_b - \boldsymbol{\mu}_r)' \end{pmatrix}\mathbf{M}'
\end{aligned} \tag{2.123}$$

Since by definition $\mathbf{r}_a$ and $\mathbf{r}_b$ have the same distribution, the expectation values for the diagonal submatrices are equal. The off diagonal submatrices relate motions of both subsystems to each other

$$\boldsymbol{\Sigma}_c = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}'_{ab} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} = \mathbf{M}\begin{pmatrix} \boldsymbol{\Sigma}_r & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}'_{12} & \boldsymbol{\Sigma}_r \end{pmatrix}\mathbf{M}' \tag{2.124}$$

Since both the sample covariance matrices for $\mathbf{r}_a$ and $\mathbf{r}_b$ are estimates for the same population covariance matrix $\boldsymbol{\Sigma}_r$, a better estimate for the latter can be obtained by combining the data sets to construct the pooled covariance matrix

$$\mathbf{S}_r = \frac{1}{2K-2}\left(\sum_{i=1}^{K}(\mathbf{r}_{ai} - \bar{\mathbf{r}})(\mathbf{r}_{ai} - \bar{\mathbf{r}})' + \sum_{i=1}^{K}(\mathbf{r}_{bi} - \bar{\mathbf{r}})(\mathbf{r}_{bi} - \bar{\mathbf{r}})'\right) \tag{2.125}$$

where

$$\bar{\mathbf{r}} = \frac{1}{2K}\left(\sum_{i=1}^{K}\mathbf{r}_{ai} + \sum_{i=1}^{K}\mathbf{r}_{bi}\right) \tag{2.126}$$

is used as the estimate for the population mean.

Given the rotation matrices $\mathbf{M}_a$ and $\mathbf{M}_b$, the matrix of sample covariances $\mathbf{S}_{12}$ can be written as

$$\mathbf{S}_{12} = \mathbf{M}_a(\mathbf{r}_{ai} - \bar{\mathbf{r}})(\mathbf{r}_{bi} - \bar{\mathbf{r}})'\mathbf{M}_b' \tag{2.127}$$

showing how the covariance matrix relating subsystems A and B is related to the covariance matrix of the rotated subsystems. Note that the rotation is not needed for the qualitative evaluation of the relations, but is needed when it is desired to visually inspect the interactions in the original coordinate system.

The extraction of the collective motions follows the same procedure as given before. However, in the case of equal subsystems, it is sufficient to determine the eigenvectors from $\mathbf{S}_r$.

This method can easily be generalized for the case of k subsystems, which are indifferent except for a rotation (and possibly translation). If each observation of each subsystem is repositioned onto a reference system by a least-squares fit, the system can be represented by a partitioned vector **r** with subvectors $\mathbf{r}_1$ through $\mathbf{r}_k$, which all have the same distribution with mean vector μ_r and covariance matrix $\mathbf{\Sigma}_r$. in that case the sample covariance matrix has a distinct pattern and is given by

$$\mathbf{S}_r = \begin{pmatrix} \mathbf{S}_r^* & \mathbf{S}_{12} & \cdots & \cdots & \mathbf{S}_{1k} \\ \mathbf{S}_{21} & \mathbf{S}_r^* & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \mathbf{S}_r^* & \mathbf{S}_{12} \\ \mathbf{S}_{k1} & \cdots & \cdots & \mathbf{S}_{21} & \mathbf{S}_r^* \end{pmatrix} \tag{2.128}$$

with

$$\mathbf{S}_r^* = \frac{1}{k}\sum_{i=1}^{k}\mathbf{S}_{ii} \tag{2.129}$$

and

$$\mathbf{S}_{ij} = \frac{1}{N}\sum_{t=1}^{N}(\mathbf{r}_{it} - \bar{\mathbf{r}})(\mathbf{r}_{jt} - \bar{\mathbf{r}})' \tag{2.130}$$

This covariance matrix is processed, such that the diagonal submatrices are diagonalized and the off-diagonal submatrices show the covariances between the principal components of the subsystems for each distinct pair of orientations

$$\mathbf{P}\mathbf{S}_r\mathbf{P}' = \begin{pmatrix} \mathbf{D}_r^* & \mathbf{C}_{12} & \cdots & \cdots & \mathbf{C}_{1k} \\ \mathbf{C}_{21} & \mathbf{D}_r^* & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \mathbf{D}_r^* & \mathbf{C}_{12} \\ \mathbf{C}_{k1} & \cdots & \cdots & \mathbf{C}_{21} & \mathbf{D}_r^* \end{pmatrix} \tag{2.131}$$

It is easy to verify that the matrix **P** is a partitioned matrix with off-diagonal submatrices equal to **0** and submatrices on the diagonal all equal to $\mathbf{P}^*$, which is the matrix of eigenvectors of $\mathbf{S}_r^*$

$$\mathbf{P}^*\mathbf{S}_r^*\mathbf{P}^{*\prime} = \mathbf{D}_r^* \tag{2.132}$$

6.3 Maximum Covariance Analysis

The principal components of the subsystems obtained using the approach above may have a complicated covariance pattern $\mathbf{C}_{12}$. Notably, a single component from subsystem A may correlate to several components of subsystem B and vice versa. Instead, the fluctuations in the system can also be expressed in terms of orthogonal modes for A and B in such a way that the covariance between principal components $\mathbf{a}_1$ and $\mathbf{b}_1$ is maximized and the covariance between any two components $\mathbf{a}_i$ and $\mathbf{b}_j$ is zero for all $i \neq j$. This can be done using *Maximum Covariance Analysis*[45].

In the method of principal components applied to interacting systems outlined above the covariance matrices of the subsystems were diagonalized and the covariance pattern of these components was revealed. In contrast, the objective of maximum covariance analysis is to diagonalize the covariance matrix $\mathbf{S}_{12}$. This is done by *singular value decomposition* (SVD) of that matrix according to

$$\mathbf{S}_{12} = \mathbf{U}_1 \mathbf{\Lambda} \mathbf{U}_2' \qquad (2.133)$$

such that $\mathbf{U}_1$ and $\mathbf{U}_2$ are the matrices of the components for A and B respectively and $\mathbf{\Lambda}$ is the diagonal matrix of singular values. The matrix $\mathbf{\Lambda}$ has size $r \times r$, where $r \leq \min(p,q,n-1)$ is the rank of $\mathbf{S}_{12}$. $\mathbf{U}_1$ and $\mathbf{U}_2$ are column-orthonormal matrices of size $p \times r$ and $q \times r$, respectively. The total covariance of $\mathbf{S}_{12}$ is equal to the sum of the square of the diagonal values of $\mathbf{\Lambda}$ and the relative importance of the kth mode from A and B is given by

$$\frac{\lambda_k^2}{\sum_{i=1}^{r} \lambda_i^2} \qquad (2.134)$$

The components given by $\mathbf{U}_1$ and $\mathbf{U}_2$ correspond to the collective motions of particles in subsystems A and B, which are column wise linearly correlated between the subsystems and (linearly) uncorrelated with other modes from both A and B. Thus the first of these components is that collective motion which best describes the interaction between the two subsystems.

7 Relations between atomic fluctuations and instantaneous properties

The collective motions identified from the structure in the atomic fluctuations give insight into the mechanical behaviour of a macromolecule. However, the physical characteristics are often better understood by examining various instantaneous properties, such as the number of hydrogen bonds or the presence or absence of specific elements of secondary structure. In many cases, it is of interest to know how specific collective motions correlate with these instantaneous properties and which collective motions best describe the changes in these properties. The techniques presented in the previous section provide several ways to investigate such relations, which are described here.

The methods described in this section were derived during the work for this dissertation to assess specific questions regarding the simulations.

7.1 Multiple regression relations between instantaneous properties and atomic fluctuations

Consider a system X in simulation, of which the positional vector at time t is given by $\mathbf{x}(t)$ and the trajectory is represented as the data matrix $\mathbf{X}$. From the trajectory a instantaneous property q is retrieved, denoted $q(t)$ for the value at time t. This property is expected to be correlated with the atomic fluctuations according to a certain pattern.
Let $\mathbf{c}$ denote the compounded vector, consisting of q and $\mathbf{x}$:

$$\mathbf{c} = \begin{pmatrix} q \\ \mathbf{x} \end{pmatrix} \qquad (2.135)$$

Then the partitioned sample covariance matrix is given by

$$\mathbf{S}_c = \begin{pmatrix} s_q^2 & \mathbf{s}_{qx} \\ \mathbf{s}'_{qx} & \mathbf{S}_x \end{pmatrix} \qquad (2.136)$$

As a first approach, it is desired to find the linear compound

$$Y = \boldsymbol{\beta}'\mathbf{x} \qquad (2.137)$$

of the atomic positions having the greatest correlation with q. The vector $\boldsymbol{\beta}$ consists of the *regression coefficients* of q upon the elements of $\mathbf{x}$. The sample estimate of $\boldsymbol{\beta}$ is $\mathbf{b}$ and is obtained from the relation

$$\mathbf{b} = \mathbf{S}_x^{-1}\mathbf{s}_{qx} \qquad (2.138)$$

The maximum correlation between q and $\mathbf{x}$ is given by the *multiple correlation coefficient*

$$r_{1.2\ldots(p+1)} = \frac{\sqrt{\mathbf{s}'_{12}\mathbf{S}_{22}^{-1}\mathbf{s}_{12}}}{s_1} = \frac{\sqrt{\mathbf{s}'_{12}\mathbf{b}}}{s_1} \qquad (2.139)$$

In this way, the relations between an instantaneous property q and the atomic positions $\mathbf{x}$ are obtained. Likewise, it is possible to obtain a matrix of regression coefficients $\mathbf{B}$, relating the positional vector $\mathbf{x}$ to a number of instantaneous properties $\mathbf{q}$.

7.2 Relations between instantaneous properties and collective motions

Rather than obtaining the coefficients for individual atoms, one may look at the relationship between a property q or a vector of properties $\mathbf{q}$ and the collective motions of a macromolecule. For this one can perform regression of q on the projections or scores of the trajectory on a selected number of principal components according to the following method.
Given the matrix of scores $\mathbf{Y}$ of r selected principal components,

$$\mathbf{Y} = \mathbf{A}'_r\mathbf{X}\left(\mathbf{I} - \frac{1}{N}\mathbf{E}\right) \qquad (2.140)$$

one can construct a combined trajectory data matrix $\mathbf{Z}$, defined as

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Q} \\ \mathbf{Y} \end{pmatrix} \qquad (2.141)$$

where **Q** is the time series of q or **q**. Then from the covariance matrix of **Z** the regression and correlation coefficients can be obtain in a manner similar to that given above.

7.3 Relations between instantaneous properties and collective motions starting from original data

An alternative approach is to start from a data matrix **Z**, consisting of the instantaneous property time series and the original trajectory:

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Q} \\ \mathbf{X} \end{pmatrix} \tag{2.142}$$

The covariance matrix of **Z** can be processed in such a way that the submatrix $\mathbf{S}_{22}$ corresponding to the atom coordinates is diagonalized, while the submatrix $\mathbf{S}_{11}$ remains unaltered. The submatrix $\mathbf{S}_{12}$, giving the covariances between the instantaneous properties and the atom coordinates, is then processed such that the relations between the former and the principal components are revealed.

$$\mathbf{P'SP} = \mathbf{P'}\begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}\mathbf{P} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}'_{22} \end{pmatrix}\begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{D}_{22} \end{pmatrix} \tag{2.143}$$

This method is equivalent to that given in 2.114 for interacting subsystems except that it leaves $\mathbf{S}_{11}$ unaltered.

Note that in certain cases it also makes sense to diagonalize $\mathbf{S}_\mathbf{q}$ and perform the analysis as in 2.114. In particular, with the proper choice of the variables in **q**, together forming the phase space of a set of reaction coordinates, diagonalization of $\mathbf{S}_\mathbf{q}$ yields a new set of reaction coordinates of which the first corresponds to the largest change in the system. As mentioned before, if the scales of the observables differ by order of magnitude, it is possibly better to use the correlation matrix of **q** rather than the covariance matrix.

7.4 Using maximum covariance analysis

The previous techniques allow the investigation of which collective motions are linked to one or a set of given instantaneous properties. Alternatively, it is also possible to determine which mode has the highest covariance with a certain instantaneous property. This is also the objective of maximum covariance analysis; i.e. to extract from two sets of data the components which maximize the covariance. To this purpose the singular value decomposition was introduced in section 6.3. When performing the decomposition of the covariance vector from a univariate instantaneous property and a full trajectory, only one component will be defined, corresponding to that linear combination of atomic fluctuations which is maximally correlated with the instantaneous property

$$\mathbf{s}_{12} = u\lambda\mathbf{v}' = \lambda\mathbf{v}' \tag{2.144}$$

where u is one by definition of the constraints. The correlation between the component and the univariate property is given by

$$r = \frac{\lambda^2}{s_{11}\sqrt{\mathbf{v}'\mathbf{S}_{22}\mathbf{v}}} \tag{2.145}$$

7.4.1 Inclusion of latent variables for the instantaneous properties

Applying the previous method to a covariance matrix of a set of instantaneous properties and atom coordinates, one obtains a set of components for the former as well as for the latter. Though it would be possible to keep the instantaneous properties fixed, it should be noted that these may well be better described in terms of a new set of latent variables. This is because the instantaneous properties at a given time are different projections of the same configuration and are likely to be correlated. However, one should normalize the data of the instantaneous properties if the values of these differ by an order of magnitude or more.

8 Data reduction

To complete this chapter, three methods for data reduction are presented, which find their basis in the statistical methods given in the previous sections. These methods are not directly used in the work described in the following chapters, except that the decomposition of the mean square displacement into rigid body and residual contributions, used in Chapters 5 and 6, follows from the method for structure reduction presented here.

The size and complexity of data collected in molecular simulations calls for methods to reduce the dimensionality of the data set, while retaining as much information as possible. Data from molecular simulations can be processed in several ways to reduce its complexity. One may express the trajectory in terms of a number of instantaneous properties, which sufficiently characterize the system. Alternatively, one may express the total fluctuation in the system in terms of a new set of variates, describing successively smaller portions of the fluctuation. This generally allows one to capture the greater proportion of fluctuations in a limited number of new variates, typically in the order of ten to fifteen. Finally, this technique can also be used to reduce the “dimensionality” of a molecular structure, by reducing the number of coordinates to consider for analysis. In particular, *a priori* information with regards to semi-rigid bodies can be used to express the structure with fewer coordinates, while retaining information on the most important events. All three techniques will be discussed more extensively in the following paragraphs.

8.1 Instantaneous properties

A given configuration of atoms can often be described in terms of a small number of characteristic properties. In the case of a protein, one can think of properties such as the number of hydrogen bonds or salt bridges, the solvent accessible surface area, the radius of gyration and the content of secondary structure elements. The interpretation of such properties can be regarded as the projection of the high-dimensional structural data on the uni-dimensional axis of a characteristic property.

A number of instantaneous properties can be combined to constitute a vector of variables characterizing a conformation. The time evolution of this vector will give rise to a trajectory through the parameter space of these properties, and can be regarded a substitute for the original trajectory to the purpose of further analysis.

Since the different properties are obtained from the same configuration, they need not be independent. Rather, these properties will have a distinct correlation pattern, which further characterizes the system.

8.2 Principal component analysis

An established method to reduce the size and complexity of a molecular dynamics trajectory is the use of principal components. In the previous sections this method was discussed in relation to data analysis. The aim of principal component analysis is to express the original data set in terms of a new set of variates, which are linear combinations of the original ones. These new variates are defined such that the first describes the largest portion of the total fluctuation and the following describe successively smaller portions, which are linearly uncorrelated to the previous ones. In general, a limited number of these new variates, called principal components, are needed to describe the greater part of the total fluctuation. Typically ten to fifteen principal components are enough to capture more than 95% of the total motion in molecular dynamics. For further (statistical) analysis the original trajectory can usually be replaced by a trajectory consisting of a limited number of principal components. This gives a significant decrease in the size of the trajectory, but also reduces the complexity, since the principal components are by definition orthogonal.

8.3 Structure reduction

Another possibility to reduce the complexity of a data set obtained from molecular simulations is to regard a macromolecule in terms of a limited set of semi rigid bodies rather than atoms. In many cases, the functionality of a macromolecule can be explained in terms of semi-rigid domains, connected through mechanical hinges. In such cases it is possible to replace each domain with a set of four vectors, representing the position of the centre of mass and the orientation of each domain in Cartesian space. If a typical domain consists of several hundreds of atoms, this will lead to a dramatic decrease in the number of variates.

In Chapter 6 this method is used as a step in the analysis of the active and inactive states of the Erythropoietin Receptor (EPOR). This receptor consists of two subunits, each of which contains ~2300 atoms. Each subunit consists of two distinct domains, which are connected by a linking region. Representing each domain by four points gives 16 points describing the orientations of the domains.

As the first step in the reduction of a structure, the three principal axes of each pre-defined domain in the reference configuration are calculated. Together with the centres of mass, these axes define the positions and orientations of the domains. From the sets of coordinates the distances between the centres of mass and the Euler angles describing the rotational relationships can be calculated.

The principal axes for each domain can be retrieved from all configurations in a trajectory. However, doing so can cause sudden changes in the order of the axes when the domain deforms. Therefore, to obtain the principal axes for a domain at a given time *t*, the corresponding domain from the reference structure is fitted onto the configuration of that domain. This involves applying a rotation and translation. If the same rotation and translation are applied to the four “principal coordinates” defining the orientation of the domain in the reference structure, the new orientation of the domain is obtained, considering all motion to be rigid body motion.

Obviously, this technique leads to a loss of data. In particular information regarding the interactions of individual atoms is lost. This said, it is possible to quantify how much of the information in the trajectory will be lost by decomposition of the root mean-square deviation (RMSD), or rather the mean-square deviation.

The RMSD is a measure for the distance of two configurations in conformational space and is given by

$$\mathrm{RMSD} = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i(\mathbf{r}_i - \mathbf{r}_i^0)'(\mathbf{r}_i - \mathbf{r}_i^0)} \tag{2.146}$$

The MSD is accordingly the second moment of the conformational distribution around the reference configuration $\mathbf{r}^0$

$$\mathrm{MSD} = \frac{1}{M}\sum_{i=1}^{N} m_i(\mathbf{r}_i - \mathbf{r}_i^0)'(\mathbf{r}_i - \mathbf{r}_i^0) \tag{2.147}$$

In the following the reference structure is assumed to be the average configuration. For a molecule consisting of rigid domains, the MSD is considered to consist of two contributions, namely the $\mathrm{MSD}_{\mathrm{RB}}$ due to displacement and reorientation of domains and the $\mathrm{MSD}_{\mathrm{Res}}$

$$\mathrm{MSD}_{\mathrm{Tot}} = \mathrm{MSD}_{\mathrm{RB}} + \mathrm{MSD}_{\mathrm{Res}} \tag{2.148}$$

or

$$\sum_{i=1}^{N} m_i(\mathbf{r}_i - \bar{\mathbf{r}})'(\mathbf{r}_i - \bar{\mathbf{r}}) = \sum_{i=1}^{N} m_i(\mathbf{r}_{i,\mathrm{RB}} - \bar{\mathbf{r}})'(\mathbf{r}_{i,\mathrm{RB}} - \bar{\mathbf{r}}) + \sum_{i=1}^{N} m_i(\mathbf{r}_{i,\mathrm{Re}s} - \bar{\mathbf{r}})'(\mathbf{r}_{i,\mathrm{Re}s} - \bar{\mathbf{r}}) \tag{2.149}$$

where $\mathbf{r}_{\mathrm{RB}}$ is the configuration expected when the domains were true rigid bodies and $\mathbf{r}_{\mathrm{Res}}$ is the difference between the actual configuration and $\mathbf{r}_{\mathrm{RB}}$. $\mathbf{r}_{\mathrm{RB}}$ corresponds to the configuration obtained by performing a least-squares fit of each reference domain on the corresponding domain at time *t*. In practice MSD_{Res} is calculated first as the MSD obtained after performing a least-squares fit of each domain on the corresponding reference domain.

This decomposition of the *MSD* is used in Chapter 5 to investigate the contributions of rigid body and residual motions to the total motility of the Death Receptor 5.

Instead of decomposing the MSD for a given time, it is also possible to express the loss of data or the *goodness of fit* of the data by looking at rigid bodies in terms of the covariance matrix determined from the trajectory or the sums of squares and cross-products (SSCP):

$$\begin{aligned} \mathbf{A} &= \sum_{t=1}^{N}(\mathbf{r}_t - \bar{\mathbf{r}})(\mathbf{r}_t - \bar{\mathbf{r}})' \\ &= \mathbf{A}_{\mathrm{RB}} + \mathbf{A}_{\mathrm{Res}} = \sum_{t=1}^{N}(\mathbf{r}_{t,\mathrm{RB}} - \bar{\mathbf{r}})(\mathbf{r}_{t,\mathrm{RB}} - \bar{\mathbf{r}})' + \mathbf{A}_{\mathrm{Res}} \end{aligned} \tag{2.150}$$

9 References

1. Schiferl, S.K. and D.C. Wallace, *Statistical Errors in Molecular-Dynamics Averages.* Journal of Chemical Physics, 1985. **83**(10): p. 5203-5209.
2. Straatsma, T.P., H.J.C. Berendsen, and A.J. Stam, *Estimation of Statistical Errors in Molecular Simulation Calculations.* Molecular Physics, 1986. **57**(1): p. 89-95.
3. Jenkins, G.M. and D.G. Watts, *Spectral analysis and its applications*. 1968, San Francisco: Holden-Day.
4. Morales, J.J., M.J. Nuevo, and L.F. Rull, *Statistical Error Methods in Computer-Simulations.* Journal of Computational Physics, 1990. **89**(2): p. 432-438.

5. Dietrich, S. and H. Dette, *Correlation Length of Time-Series in Statistical Simulations.* Journal of Computational Physics, 1992. **101**(1): p. 224-226.
6. Kendall, M.G., *Time Series*. 1973, London: Griffin.
7. Flyvbjerg, H. and H.G. Petersen, *Error-Estimates on Averages of Correlated Data.* Journal of Chemical Physics, 1989. **91**(1): p. 461-466.
8. Efron, B., *1977 Rietz Lecture - Bootstrap Methods - Another Look at the Jackknife.* Annals of Statistics, 1979. **7**(1): p. 1-26.
9. Miller, R.G., *Jackknife - Review.* Biometrika, 1974. **61**(1): p. 1-15.
10. Tukey, J.W., *Bias and Confidence in Not-Quite Large Samples.* Annals of Mathematical Statistics, 1958. **29**(2): p. 614-614.
11. Singh, K., *On the Asymptotic Accuracy of Efrons Bootstrap.* Annals of Statistics, 1981. **9**(6): p. 1187-1195.
12. Künsch, H.R., *The Jackknife and the Bootstrap for General Stationary Observations.* Annals of Statistics, 1989. **17**(3): p. 1217-1241.
13. Knecht, V. and H. Grubmuller, *Mechanical coupling via the membrane fusion SNARE protein syntaxin 1A: A molecular dynamics study.* Biophysical Journal, 2003. **84**(3): p. 1527-1547.
14. Welch, B.L., *The Significance of the Difference Between Two Means when the Population Variances are Unequal.* Biometrika, 1938. **29**(3/4): p. 350-362.
15. Efron, B. and R.J. Tibshirani, *An Introduction to the Bootstrap*. 1993, New York: Chapman & Hall.
16. Reiczigel, J., D. Zakarias, and L. Rozsa, *A bootstrap test of stochastic equality of two populations.* American Statistician, 2005. **59**(2): p. 156-161.
17. Cooper, A. and D.T.F. Dryden, *Allostery without Conformational Change - a Plausible Model.* European Biophysics Journal with Biophysics Letters, 1984. **11**(2): p. 103-109.
18. Maragliano, L., *et al.*, *Atomic mean-square displacements in proteins by molecular dynamics: A case for analysis of variance.* Biophysical Journal, 2004. **86**(5): p. 2765-2772.
19. Cottone, G., *et al.*, *Molecular dynamics simulation of sucrose- and trehalose-coated carboxy-myoglobin.* Proteins-Structure Function and Bioinformatics, 2005. **59**(2): p. 291-302.
20. Scharnagl, C., M. Reif, and J. Friedrich, *Local compressibilities of proteins: Comparison of optical experiments and simulations for horse heart cytochrome-c.* Biophysical Journal, 2005. **89**(1): p. 64-75.
21. Price, D.J. and C.L. Brooks, *Modern protein force fields behave comparably in molecular dynamics simulations.* Journal of Computational Chemistry, 2002. **23**(11): p. 1045-1057.
22. Fan, H., *et al.*, *Comparative study of generalized Born models: Protein dynamics.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(19): p. 6760-6764.
23. Kruskal, W.H. and W.A. Wallis, *Use of Ranks in One-Criterion Variance Analysis.* Journal of the American Statistical Association, 1952. **47**(260): p. 583-621.
24. Friedman, M., *The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance.* Journal of the American Statistical Association, 1937. **32**(200): p. 675-701.
25. Scheffe, H., *A Method for Judging All Contrasts in the Analysis of Variance.* Biometrika, 1953. **40**(1-2): p. 87-104.
26. Scheffe, H., *The Analysis of Variance*. 1959, New York: John Wiley and Sons, Inc.
27. Miller, R.G., Jr, *Simultaneous Statistical Inference*. 2 ed. 1981, New York: Springer.
28. Tukey, J., *Multiple Comparisons.* Journal of the American Statistical Association, 1953. **48**(263): p. 624-625.
29. Bartlett, M.S., *Properties of Sufficiency and Statistical Tests.* Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 1937. **160**(901): p. 268-282.
30. Levene, H., in *Contributions to probability and statistics : essays in honor of Harold Hotelling*, I. Olkin, *et al.*, Editors. 1960, Stanford University Press: Stanford, CA.
31. Hotelling, H., *The Generalization of Student's Ratio.* The Annals of Mathematical Statistics, 1931. **2**(3, Six Dollars per Annum): p. 360-378.

32. Dobson, A.J., *An introduction to generalized linear models*. 2 ed. Chapman & Hall texts in statistical science series. 2002, Boca Raton: Chapman and Hall/CRC.
33. McCullagh, P. and J.A. Nelder, *Generalized linear models*. 2 ed. Monographs on statistics and applied probability 1989, London: Chapman and Hall.
34. Morrisson, D.F., *Multivariate Statistical Methods*. McGraw-Hill series in probability and statistics, ed. D. Blackwell and H. Solomon. 1976, Tokyo: McGraw-Hill Kogakusha.
35. Roy, S.N., *p-statistics or some generalizations in analysis of variance appropriate to multivariate problems.* Sankhyā, 1939. **4**: p. 381-396.
36. Roy, S.N., *The sampling distribution of p-statistics and certain allied statistics on the non-null hypothesis.* Sankhyā, 1942. **6**: p. 15-34.
37. Lawley, D.N., *A Generalization of Fisher's z Test*. Biometrika, 1938. **30**(1/2): p. 180-187.
38. Pillai, K.C.S., *Some New Test Criteria in Multivariate Analysis.* Annals of Mathematical Statistics, 1955. **26**(1): p. 117-121.
39. Wilks, S.S., *Certain Generalizations in the Analysis of Variance.* Biometrika, 1932. **24**(3/4): p. 471-494.
40. Pearson, K., *On Lines Planes of Closes Fit to System of Points in Space, London Edinburgh Dublin*, in *Phil. Mag. J. Science*. 1901. p. 559-572.
41. Hotelling, H., *Analysis of a complex of statistical variables into principal components.* Journal of educational psychology, 1933. **24**: p. 417-441,498-520.
42. Ichiye, T. and M. Karplus, *Collective Motions in Proteins - a Covariance Analysis of Atomic Fluctuations in Molecular-Dynamics and Normal Mode Simulations.* Proteins-Structure Function and Genetics, 1991. **11**(3): p. 205-217.
43. Amadei, A., A.B.M. Linssen, and H.J.C. Berendsen, *Essential Dynamics of Proteins.* Proteins-Structure Function and Genetics, 1993. **17**(4): p. 412-425.
44. Hess, B., *Similarities between principal components of protein dynamics and random diffusion.* Physical Review E, 2000. **62**(6): p. 8438-8448.
45. Jolliffe, L.K., *Principal component analysis*. 2 ed. Springer series in statistics. 2002, New York: Springer.