Exercise 1.1

(a) The advantages of a hard model are:

- A model can be constructed which reflect the current knowledge of the system (which can be substantial).
- A hard model is **causal** and thus, it is easier to propose testing hypotheses and explanations.
- A hard model is easier to interpret and understand than a soft model. It typically provides a deeper understanding of the system than a soft model.

The main disadvantage of the hard modeling approach is that it is often very difficult to have sufficient information about a system to set up a realistic model. In addition, the complexity of the problem is so high that a hard model may become unrealistic to formulate.

The advantages of a soft model are:

- The approach can handle problems where we have very little information about the inner workings of a system.
- It is easy to add more variables, also those that may (at a later stage) be irrelevant to the problem.

The main disadvantages of soft models are:

- They are only local, i.e. extrapolation becomes riskier.
- They do not provide a very deep understanding of a problem as a hard model can do.

It is indeed possible to combine a hard and soft modeling approach. For instance, the soft modeling approach can be applied to residual variation in the observed data which cannot be explained by a hard model. This suggests that the hard model is not correct. Further, a soft model may give inspiration for suggesting a hard model.

(b) Other possible variables influencing the distance traveled are wind speed, wind direction, and the size and shape of the person in the cannon.

(c) A set of experiments would be needed where controllable factors such as cannon angle and initial velocity are varied. It is also possible for the person to wear different types of clothes to simulate different degrees of drag during flight. The wind can also be measured during the experiments. A matrix is collected where one row describes a single experiment with all the variables. The dependent variable will be the distance the person landed from the cannon.

(d) As mentioned in the previous question more than the cannon angle (θ) and speed (v_0) are important. Therefore, the Newtonian equation involved here,

$$\hat{y}_{\text{Newton}} = \frac{v_0^2}{g} \sin\theta,$$

where $g = 9.81$ m/s^2 is not sufficient to explain all the variation of y, i.e. the observed distance traveled. So we could suggest a model of the following kind:

$$y = \hat{y}_{\text{Newton}} + \hat{y}_{\text{soft}},$$

where $\hat{y}_{\text{soft}} = f(\text{wind speed}, \text{drag}, \ldots)$ and f is some empirical-based model, most likely a linear model.

Exercise 1.2

(a) As row vectors. Mathematically it does not matter, but in chemometrics, it is often the case that sample objects are stored as rows and the variables as columns.

(b) A Fourier transform can be performed for each recording and then use the intensities for each wavelength. Or we could perform statistical analysis of each sound track and store variables about the mean, standard deviation, kurtosis, etc. Many different transforms can be used to generate variables for such a time series.

(c) The length of the recordings should be the same. If actual time series are used directly as variables, the profiles **must** be aligned properly.

Exercise 1.3

(a) Example of variables that can be used: melting and boiling point, van der Waals volume, charge and HOMO/LUMO energy levels.

(b) The first problem is whether the relative orientation of the molecules is comparable. If not the values will be completely useless. The other problem is comparability between atoms in different molecules. Are the atoms comparable chemically? If not, wrong results will emerge.

(c) One approach is to align the molecules according to a principle or reference. The other is to make use of orientation independent variables such as distances and angles.

Exercise 1.4

Considering the different cases:

(a) In this case, we expect no problems.

(b) In this case, we can encounter problems if the peaks have widely different width.

(c) In this case, we can encounter problems. Change of positions may cause a problem, and this will be dependent on the degree of shift of the peaks. If the shifts are small and the peaks are broad this is not a problem. But if the shifts are larger and the peaks are more narrow, the comparability problem will be serious.

In summary we expect to encounter problems for the last two cases.

Exercise 1.5

(a) Since each spectrum, s_j, is represented as the following function:

$$s_j(x) = \sum_{i=1}^{3} a_i \mathrm{e}^{-(x-b_i)^2/c_i},$$

we can make a vector of the form $[a_1, b_1, c_1, a_2, b_2, c_2, a_3, b_3, c_3]$ which represents the spectrum. This representation is also independent of the relative and absolute shifts of the peaks as this is now contained in the b_i variable which is by definition comparable for the different samples. This is called the peak parameter representation (PPR).

(b) PPR can only handle a situation where the relative positions of the peaks are constant, i.e. peak 1 is before peak 2 which is before peak 3. If they swap positions, the assumption behind the PPR is no longer valid.

Exercise 1.6

If the peaks are really narrow, a small shift in peak position would result in either measuring nothing or the full peak at the SPR measuring point. However, for broader peaks, the difference in measured intensity due to peak shift would be much less extreme. E.g there is a higher chance that some of the peak intensity will still overlap with the SPR measuring point for broader peaks.

Exercise 1.7

Since the product must be very cheap, we are limited to something which has low production cost and low cost in terms of transportation. The problem is trying to find an instrument which an ordinary customer is able to carry around. A solution is to use a mobile phone with a camera. The product FoodInspect can then sell an app which performs image analysis and chemometric analysis. A chemometric model can be made for each different type of

fruit or vegetable. An experiment for each group must be performed where freshness can be interpreted as time after harvesting and related to the different image descriptors.