Modelling nutrient content and developing control charts using NIR data of insect feed and products

Tamás Baráth - r0738552

Promotor: Bart De Ketelaere
Co-promotor: Eric Schmitt

November 3, 2019

Description of the task

My task is to develop control charts for the nutrient content (protein, fat, ash, and dry matter) of harvested insects. A control chart is a graph used to study how a process changes over time. More specifically, it is meant to measure the variance of a process from its expected value and determine if the variance is only coming from sources common to the process, or if there are corrections needed. In order to develop such charts the nutrient content needs to be modelled using near infrared spectroscopy (NIR) measurements as the explanatory variable, with the nutrient contents being the dependent variables. The greatest difficulty with a regression approach based on NIR data comes from multicollinearity. In many cases the dimensionality of the predictor variables is higher than the number of observations in the sample (training data set), which makes conventional multiple regression approaches impossible. Even when that is not the case, the variables still tend to be highly correlated, inflating the variance of the predictions.

Dimensionality reduction

The common way to handle the multicollinearity problem with NIR regression is by reducing the dimensionality of the regressors. In spectroscopy it is usually not easy to find selective wavelengths for the chemical constituents in the samples, which makes it unfeasible to solve the multicollinearity by selecting only a subset of the predictor variables. Instead, either principal component analysis (PCA) or partial least squares (PLS) regression methods can be used. PCA tries to reduce the dimensionality of the data while preserving as much of the variance in the predictor variables as possible by orthogonal transformation. While this works very well with NIR data, as most of the variance can usually be represented with only a few orthogonal variables (called principal components), there is no guarantee that the variance of the regressors always contains useful information instead of noise. The PLS approach aims to avoid this pitfall by considering the covariance structure with the dependent variables as well.

Principal component analysis

According to Jolliffe (2002) "the central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables." The first principal component is derived as a linear combination of the original variables (also called a principal component score) that has maximal variance with the constraint that weight vector w_1 has to have unit length. For the constrained maximization the technique of Lagrangian multipliers is used. The Lagrange multiplier λ_1 will correspond to the largest eigenvalue of the covariance matrix of the original variables with the corresponding eigenvector being w_1. Once w_1 is known, the first PC can be calculated. In general, the kth component can be found by subtracting the first k - 1 principal components from the original variables and repeating the constrained maximization (finding the unit weight vector w_k which extracts the maximum variance from this new data matrix). As it turns out, the kth component is given by the kth largest eigenvalue and corresponding eigenvector of the original covariance matrix. Formally the process can be described as follows:

Let X be a data matrix, with column-wise zero mean (every variable has been mean centered) and t_k be the PC with the kth largest variance. Then $t_k = X \cdot w_k$ with w_k being the unit weight vector corresponding to the kth PC.

Partial least squares regression

The goal of PLS, as described by Tobias (1996)[2], is to extract latent variables (T and U) from sampled predictors and responses, respectively. The extracted factors T (also referred to as X-scores) are used to predict U (the Y-scores), and then the predicted Y-scores are used to construct predictions for the responses. The X- and Y-scores are chosen so that the relationship between successive pairs

of scores is as strong as possible (meaning that the covariance is maximized), therefore this is not the same as doing PCA on the predictor and response variables. Geladi and Kowalski (1986)[1] describes a PLS regression on NIR data as follows. First, the data (both the regressor and dependent variables) is mean-centred, that is the average of each variable is subtracted from the variable itself. The data can also be rescaled, but no scaling is needed when all the variables in a block are measured in the same units, as in spectrometry. Then the predictor and response variables are decomposed as products of sets of orthogonal factors (the X- and Y-scores) and of sets of specific loadings. The latent T and U matrices are obtained vector-by-vector, in the first iteration only the first pair of t and u latent vectors is obtained, containing the most covariance. The second iteration extracts a pair with the second most covariance and so on. The process goes as follows in the first iteration:

1. Select one of the response variables as an initial estimate of the u vector.
2. Regress the predictor variables X on u by ordinary least squares (OLS), obtaining w, a weight vector.
3. Normalize the euclidean length of w to be equal to 1.
4. Regressing all predictor variables in X on w, obtaining t, a latent vector of X-scores.
5. Regress all response variables in Y on t, obtaining q, a weight vector.
6. Normalize the euclidean length of q to be equal to 1.
7. Regress all response variables in Y on q, obtaining u, a latent vector of Y-scores.

This results in the first pair of latent vectors (t_1 and u_1), along with their factor loadings (w_1 and q_1), defining the first latent component. If there are more components to be extracted, the starting X and Y matrices are deflated by the first component. This means subtracting the product of t_1 and w_1 from X and the product of t_1 and q_1 from Y, removing the variance already explained. Afterwards the above outlined process is repeated for the deflated matrices.

Once we have extracted all the latent components needed, we can regress the T variables on the U variables, obtaining matrix B which contains the coefficients for the nutrient model.

Modeling nutrient content using PLS

First a sample is needed with both NIR data of the insects and their nutrient content measured. The data is mean-centred and both the regressor and dependent variables are decomposed into latent factors T and U with factor loadings W and Q, respectively. Then T is regressed on U by OLS, resulting in B, the coefficients for predicting the nutrient content. From now on the harvested insects' NIR data can be used to predict their nutrient content. To do this, the NIR data has to be mean-centred and transformed onto the latent variables using the factor loadings in W to obtain T. Next, U is estimated by taking the product of T and B. Finally, U is projected back to Y space using Q and the mean-centring is reversed to get the nutrient content estimates.

References

[1] Geladi, P., & Kowalski, B. (1986). Partial least-squares regression: A tutorial. Analytica Chimica Acta, 185(C), 1-17.

[2] Tobias, R.D. (1996). An Introduction to Partial Least Squares Regression. SAS Support