Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees

Mika Sulkava, Jarkko Tikka and Jaakko Hollmén

Helsinki University of Technology, Laboratory of Computer and Information Science,
P.O.Box 5400, FIN-02015 HUT, Finland, Mika.Sulkava@hut.fi, tikka@mail.cis.hut.fi,
Jaakko.Hollmen@hut.fi Tel. +358-9-451 3647, Fax. +358-9-451 3277

Key words: Linear sparse regression, prediction, foliar nutrition

Background and data

Analysis of foliar nutrient concentrations is an important part of environmental monitoring. Understanding and predicting the development of nutrient concentrations based on measurement data of the forest are challenging tasks. In this study sparse regression models were used to represent the relations between different measurements.

The nutrient data used in the analysis consist of needle mass (NM) and 12 element concentrations: Al, B, Ca, Cu, Fe, K, Mg, Mn, N, P, S and Zn. These 13 measurements were made to needles of foliar age classes C and $C + 1$ (the needles that were grown in the measuring year and in the previous year, respectively) in 16 Norway spruce and 20 Scots pine stands located in different parts of Finland between years 1987–2000.

In addition, there were 9 additional measurements available for the stands, namely the geographic coordinates, the total N and S deposition, the average temperature and total precipitation, the deviations of average temperature and precipitation from their long term averages and the age of the forest. All the measurements were done annually.

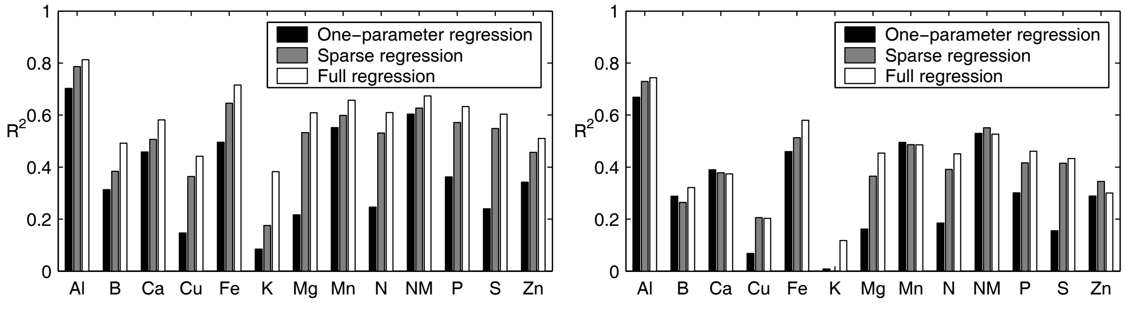
The problem at hand is to predict the nutrient concentrations and needle mass of $C + 1$ needles in year t using the measurements of C needles in year $t - 1$ and the additional measurements in year t. That is, we want to model the effect of the environment and nutrients to the aging of the needles. Also, the models should give an understandable description of the process. The purpose is to use only a few significant regressors of total 22 for each response. The most significant regressors are selected separately for each response, so that differences in dependency relations between the response and regressors in different models can be observed more easily.

Methods

Different multiple linear regression models are used for prediction. The use of linear models is justified by their interpretability and the fact that over short ranges, any process can be well approximated by a linear model. In a linear sparse regression model there are $k < K$ nonzero regression coefficients, where K is the number of regressors in a full model.

Using a sparse regression model instead of the full model is convenient, because reducing the number of coefficients makes the model easier to interpret and at the same time less prone to overfitting. In the models used in this study the most significant regressors are found using the Least Angle Regression model selection algorithm (Efron, 2004). An initial value of k is selected based on the Minimum Description Length information criterion. Subsequently, the final k is obtained by setting statistically insignificant coefficients to zero.

The sparse model is compared to the full linear regression model and a simple linear



(a) Training set

(b) Validation set

Figure 1: *Average R^2-values of the measurements from pine for one-parameter regression, sparse regression and full regression obtained using cross-validation. Results for both (a) training and (b) validation sets.*

one-parameter model that tries to predict the value of a $C + 1$ measurement in year t by only using its C value in year $t - 1$.

Results

The sparse model was found to be more suitable for the problem than the two other models. The quality of prediction was studied using cross-validation. The prediction accuracy of the different models was measured with the coefficient of determination R^2.

The results for pine are shown in Figure 1 for both the training and validation sets. In the right panel of Figure 1 it can be seen that usually the sparse model outperforms the simple one-parameter model, and its results are mainly comparable to the full model. However, the number of parameters in the sparse model is much lower: on an average $k = 5$ coefficients ($K = 22$). This is an important advantage of the sparse models, because it helps finding the important dependencies between the different measurements.

The sparse model fits rather well to the data without any noticeable signs of overfitting. The difference between the R^2 values of the training and validation sets was constantly smaller with the sparse regression model than with the full model.

The quality of the models for spruce is similar, but the dependencies between the measurements were slightly different for the two tree species. The linear sparse regression model proved to be capable of providing rather good and reliable predictions of the development of foliage with a relatively small number of parameters.

Using a permutation test, it was found that virtually always the best possible regressors were chosen to the sparse models. That is, given the number of coefficients, it is extremely difficult to construct a linear model that would better characterize the relations between the measurements.

In addition, using cross validation, relative importance of the regressors was computed, that reveal the strength of the connections between different measurements. The values of relative importance can also be regarded as a discrete probability distribution, that shows, which regressors are likely to be included in the model. Usually, a measurement naturally has the strongest connection to its previous year value. Also other, more interesting dependencies were found between the measurements.

References

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, April 2004.