More factors than observations: what can we do?

Dr.ir. Erwin Stinstra

In the age of internet of things, connected systems and smart phone apps, the size of data sets is increasing enormously. The size of a dataset can typically be split up into two dimensions: the number of observations and the number of factors (a.k.a. features, variables, parameters, items). Increasing the number of observations typically makes life easier: we get more information, which helps to make valid transfer functions. Increasing the number of factors (we are also measuring/recording more information per observation) could make life better, since the likelihood of recording the useful information also increases. However, this increase may introduce a statistical issue, if we have even more factors than observations.

	Factor 1		...		Factor p
Observation 1	2.867				3.124
...			DATA		
Observation n	3.142				2.719

Figure 1 - The size of a dataset is two dimensional: the number of factors and the number of observations

It is easy to see why this is a problem. Think of trying to fit a quadratic line through only 2 points (observations). The quadratic model contains 3 coefficients (a constant, a linear and a quadratic coefficient) that we need to fit, which is more than the number of observations. The problem now is that there are infinitely many possible quadratic curves that fit equally well through the data, and we cannot select the best based on prediction error. The same is true for a dataset that has more factors than observations. We cannot determine which model is best, because there are infinite combinations of coefficient values that fit exactly through the data.

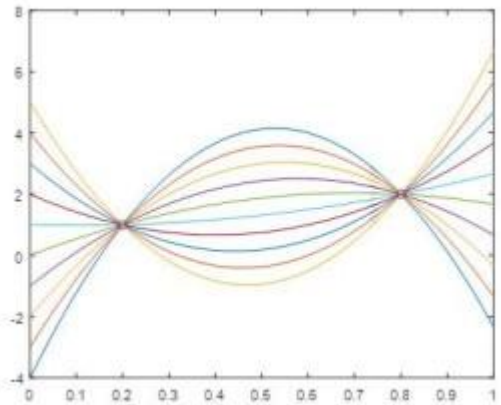

Figure 2 - All these quadratic curves fit equally well through the 2 data points

In these situations, the traditional regression model fails. Fortunately, there are other options. First of all, we could inspect the structure in the observed data and try to describe all recorded items with fewer dimensions (dimension reduction). This is the idea of, e.g., Principal Component Analysis. However, the structure in the data may not always be described well by a limited number of components. Further, the interpretation of the results may be difficult.

Another option is variable selection: if we can assume that only a few factors influence the response, it makes sense to build a regression model on only those few factors. Of course, the difficulty is to find those factors. Many variable selection techniques exist, like stepwise regression (start with a few factors, add or delete factors from the model in a structured way), all-subsets regression (try all possible regression models containing only k factors) or selection via lasso regularization (limiting the sum of absolute values of the coefficients, thereby forcing many coefficients to become 0).

Stepwise regression is very start-point dependent, and may end up with a predictive model that is not very good. All subsets regression obtains better predictive models, but may take enormous amounts of computation time, especially if the number of possible factors is large (over 50). In our experience, Lasso regularization works better for selection of factors, since it is not start-point dependent and does not require enormous computation time on typical datasets.

Note however, that all aforementioned techniques are exploratory (we try to find a mathematical relation based on observed data, we do not try to confirm a hypothesis). The interpretation of outcomes should therefore be done with care. Selection of factors does not imply that these factors actually caused the response to change; they may be markers for other effects that were the true cause.

For more information, please contact CQM
T +31 40 750 23 23
info@cqm.nl