Polynomial surfaces in low dimension

When working with optical instruments we sometimes see a slight curvature in data. An example is high fat content in fish measured by a NIR instrument. If the curvature is parabolic, a second order surface in score variables may be appropriate. If the curvature is S-shaped, a third order surface may be needed. The estimation of polynomial surfaces in score variables is briefly reviewed in the case of optimal response estimation.

A linear and quadratic model is given by

(17) $$y = \mathbf{b}^T\mathbf{x} + \mathbf{x}^T \mathbf{F} \mathbf{x} + \text{residual}$$

Here $\mathbf{x}=(x_1,x_2,...)$ are the numerical variables that are measured. The transformation from sample space to score space is $\mathbf{t}=\mathbf{R}^T\mathbf{x}$. Inserting this in (17) we get,

(18) $$\begin{aligned} y &= \mathbf{b}^T (\mathbf{R}^T)^{-1} \mathbf{R}^T \mathbf{x} + \mathbf{x}^T \mathbf{R} \mathbf{R}^{-1} \mathbf{F} (\mathbf{R}^T)^{-1} \mathbf{R}^T \mathbf{x} \\ &= (\mathbf{R}^{-1} \mathbf{c})^T (\mathbf{R}^T \mathbf{x}) + (\mathbf{R}^T \mathbf{x})^T \mathbf{R}^{-1} \mathbf{F} (\mathbf{R}^T)^{-1} (\mathbf{R}^T \mathbf{x}) \\ &= \mathbf{c}^T \mathbf{t} + \mathbf{t}^T \mathbf{E} \mathbf{t}, \end{aligned}$$

where

$$\mathbf{PD} = (\mathbf{R}^T)^{-1}, \ \mathbf{c} = (\mathbf{PD})^T \mathbf{b}, \ \text{and} \ \ \mathbf{E} = (\mathbf{PD})^T \mathbf{F} \, \mathbf{PD}$$

Consider now the estimation task for this type of model. If all score vectors are selected, there is a 1-1 relationship between these two models. Parameters of (18) are estimated and using the transformation $\mathbf{t}=\mathbf{R}^T\mathbf{x}$, parameters of (17) are obtained. But usually only A score vectors are selected. We can still go from (18) to (17) by using $\mathbf{t}=\mathbf{R}_A{}^T\mathbf{x}$, but now the value of $\mathbf{x}$ that will give the optimal value of (17) will not be unique. We shall first consider how to determine an optimal value of the score vector.

By differentiating y with respect to $\mathbf{t}$, we find that the optimal $\mathbf{t}$-value is $\mathbf{t} = -½\mathbf{E}^{-1}\mathbf{c}$. Thus we need both $\mathbf{c}$ to be different from zero and $\mathbf{E}$ to be invertible. In Gauss elimination, when computing the inverse, on can create zeros row-wise. This means that zeros below the diagonal are created variable-wise. The estimation procedure that fits to this way of computing the inverse can be carried out as described in the following.

In the linear model a score vector was sought that maximised $|\mathbf{y}^T\mathbf{t}|^2=|\mathbf{y}^T\mathbf{X}\mathbf{w}|^2$. A natural extension of this criterion is to find $\mathbf{w}$ that maximises

(19) $$|\mathbf{y}^T\mathbf{t}|^2 + |\mathbf{y}^T\mathbf{t}^{\oplus 2}|^2,$$

where the notation $\mathbf{t}^{\oplus n}$ for the n-th power of $\mathbf{t}$, $\mathbf{t}^{\oplus n}=(t_1{}^n, t_2{}^n, ...)$, is used. Similarly, the coordinate wise product of two vectors, $\mathbf{t}$ and $\mathbf{s}$, is denoted by $\mathbf{t}\oplus\mathbf{s}$, $\mathbf{t}\oplus\mathbf{s}=(t_1s_1, t_2s_2, ...)$. (In a MATLAB notation the product is $\mathbf{t}.*\mathbf{s}$.). When such $\mathbf{w}$ has been found, the regression procedure gives us the coefficients c_1 and e_{11}, $\mathbf{E}=(e_{ij})$. Then $\mathbf{X}$ is adjusted for this score vector, $\mathbf{t}_1$. Also $\mathbf{y}$ is adjusted for what has been found, $c_1\mathbf{t}_1+e_{11}\mathbf{t}_1{}^{\oplus 2}$. The next task is to find $\mathbf{t}=\mathbf{t}_2$, which is done by maximising the term,

$$|\mathbf{y}^T\mathbf{t}|^2 + |\mathbf{y}^T\mathbf{t}^{\oplus 2}|^2 + |\mathbf{y}^T(\mathbf{t}\oplus\mathbf{t}_1)|^2,$$

where $\mathbf{y}$ is now the reduced $\mathbf{y}$. When $\mathbf{w}$ has been found, $\mathbf{t}_2$ is computed, $\mathbf{t}_2=\mathbf{Xw}$, and $\mathbf{X}$ is adjusted by $\mathbf{t}_2$, and $\mathbf{y}$ is adjusted by $c_2\mathbf{t}_2+e_{22}\mathbf{t}_2^{\oplus 2}+2e_{21}\mathbf{t}_1\oplus\mathbf{t}_2$. ($\mathbf{E}$ is assumed symmetric). Next task is to find $\mathbf{t}=\mathbf{t}_3$. It is done by maximising,

$$|\mathbf{y}^T\mathbf{t}|^2 + |\mathbf{y}^T\mathbf{t}^{\oplus 2}|^2 + |\mathbf{y}^T(\mathbf{t}\oplus\mathbf{t}_1)|^2+ |\mathbf{y}^T(\mathbf{t}\oplus\mathbf{t}_2)|^2.$$

Like described above, the results are used to estimate the parameters c_3 and (e_{31},e_{32},e_{33}). Using the first three rows and columns of $\mathbf{E}$, the preliminary optimal values of first three score values, (t_1,t_2,t_3), can be computed. In order to obtain good model results it is important not to include score variables that do not improve the prediction of the model. The linear and squared terms are always included in the model even if they are not significant. This is done in order to obtain the optimal solution. The cross product terms are excluded, if they are not significant. The selection of score vectors stops, when linear, squared and cross-product terms are all not significant.

The result is a unique score vector, $\mathbf{t}_{opt}$, which is a solution to the linear equation. The eigen values of $\mathbf{E}$ will show the type of solution obtained. If all eigen values are negative, it is a maximum. The corresponding sample vector is computed as $\mathbf{x}_{opt}=(\mathbf{PD})\mathbf{t}_{opt}$. It is studied closer, where $\mathbf{t}_{opt}$ is located in score space. If it is far away from other score values, one should be careful in using the results. Similarly, $\mathbf{x}_{opt}$ can be compared to the present samples.

The estimation procedure that has been reviewed here is presented in details in Ref 6. Here models of linear and second order polynomials in score vectors have been presented. In Ref 5 it is formulated for polynomials in score vectors of any order.

It is also important here to exclude x-variables that do not contribute to the modelling task. The method of section 9 is used to remove variables that do not have significant weights.

In summary, we have presented a method of finding optimal polynomial in score vectors. It has been used to show how optimal responses can be determined, when data is of reduced rank, which is commonly the situation in nature and industry. If there is detected curvature in data, then it is important to account for it in the modelling task. The numerical procedure is fast and efficient in finding optimal polynomials in score vectors.