"Penalized Spline Additive Models: An application to study the wheat grain filling under different conditions"

Scopetta, Ivana Florencia
E-mail: ivanascopetta@hotmail.com

Cuesta, Cristina
E-mail: cbcuesta@gmail.com

ABSTRACT

Additive models under the approach of using P-splines belong to a research area that is booming nowadays. They can be used to identify and characterize the effect of more than one explanatory variable on a response and also give the possibility of including nonlinear effects of the covariates. To this end, additive models use smoothing functions that do not necessarily have any predetermined functional form before the fitting.

The revolution of the P-Splines occurred in recent years is due mainly to the possibility of writing a penalized spline regression model as a linear mixed model. The advantage of this approach is that allow you, first, to use the methodology developed for mixed models, and secondly, to use standard computer programs to estimate them.

We show the use of these techniques to model some data from a wheat field experiment. Particularly, we are looking forward to study the relationship between the green weight of the wheat grain, obtained from the filling stage, depending on the moisture content of the grain and the accumulated thermal sum in the cultivars.

Key words: *Additive models, P-Splines, Smoothing functions, Wheat grain*

ADDITIVE MODELS

They can be used to identify and characterize the effect of more than one explanatory variable on a response variable, but avoiding the assumption of linearity. In fact, with these smoothing models there is not any a priori assumption of the functional form with which you describe this relationship.

There are two continuous explanatory variables x_1 and x_2, so the additive model can be expressed as:

$$y_i = \beta_0 + f(x_{1i}) + g(x_{2i}) + \varepsilon_i$$

where f and g are smoothing functions. This model can be extended to the case of having more than two explanatory variables.

The estimation of the functions f and g can be performed using different smoothing techniques, among which stand out the ones based on splines, those known as spline regressions. They consist of a piecewise regression, where each piece is a variation region of the explanatory variable, in which it fits a polynomial regression model. These sections are joined at the endpoints, known as the "knots" to give continuity to the curve.

Spline regression models mostly depend on the number of regions considered and the respective range of them. To overcome this disadvantage, it is suggested to define a large number of regions and then weigh the importance of the regions that are considered different. This is done with the penalized spline regression models (or P-Splines).

The matrix representation of a P-Spline regression model is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} \quad \text{with } E(\boldsymbol{\varepsilon}) = 0 \text{ and } \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_{x_1} \\ \beta_{x_2} \end{bmatrix} \qquad \boldsymbol{\beta}^* = \begin{bmatrix} \beta_1^{*x_1} \\ \vdots \\ \beta_{K_1}^{*x_1} \\ \beta_1^{*x_2} \\ \vdots \\ \beta_{K_2}^{*x_2} \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{bmatrix}$$

$$\mathbf{Z} = \begin{bmatrix} \left(x_{11} - c_1^{x_1}\right)_+ & \dots & \left(x_{11} - c_{K_1}^{x_1}\right)_+ & \left(x_{21} - c_1^{x_2}\right)_+ & \dots & \left(x_{21} - c_{K_2}^{x_2}\right)_+ \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \left(x_{1n} - c_n^{x_1}\right)_+ & \dots & \left(x_{1n} - c_{K_1}^{x_1}\right)_+ & \left(x_{2n} - c_n^{x_2}\right)_+ & \dots & \left(x_{2n} - c_{K_2}^{x_2}\right)_+ \end{bmatrix}$$

where $\left(x_{ij} - c_k^{x_1}\right)_+$ allow the piecewise linear fit, being $c_1^{x_1}, \dots, c_{K_1}^{x_1}$ and $c_1^{x_2}, \dots, c_{K_2}^{x_2}$ the respective knots in x_1 and x_2 directions. They can be chosen using different rules, such as equispaced knots within the range of interest in each explanatory variable, or knots located in equispaced percentiles determined by the data set.

The vector of fitted values using penalized least squares and the coefficients associated with the knots, is given by:

$$\hat{Y} = \mathbf{A}\left(\mathbf{A}'\mathbf{A} + \mathbf{D}\right)^{-1}\mathbf{A}'\mathbf{Y}$$

with $\mathbf{A} = (\mathbf{X} \mid \mathbf{Z})$ and $\mathbf{D} = \text{diag}\left(0, 0, 0, \lambda_1^2 \mathbf{1}_{K_1 \times 1}, \lambda_2^2 \mathbf{1}_{K_2 \times 1}\right)$, where $\mathbf{D}$ is the matrix that controls the influence of the knots. λ_1 and λ_2, are the smoothing parameters.

In particular, $\beta_k^{*x_1}$ and $\beta_k^{*x_2}$ can be treated as random effects in a mixed model. Ruppert et al. (2003) and Ngo et al. (2004) have shown that least squares estimators are equivalent to the best linear unbiased predictor (BLUP) in the mixed models, with $\lambda_1 = \sigma_\varepsilon/\sigma_1$ and $\lambda_2 = \sigma_\varepsilon/\sigma_2$. Thus, the representation of the additive model as a mixed model can be used to facilitate the fitting, inference and model selection.

Inference

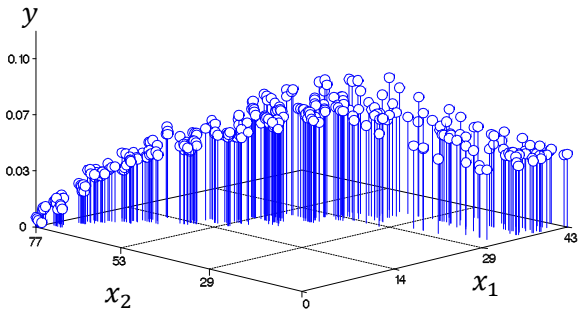
The overall effect of adding a continuous predictor to explain a continuous response variable, as well as the fact of whether or not the effect is linear, quadratic or if its shape cannot be explained with a polynomial, is of great interest when additive models are built. Research into corresponding hypothesis tests is still barely sufficient for this models and has been limited mainly to likelihood ratio tests, and F tests, possible to calculate with the available tools in statistical software. These tests give some indication of the amount of evidence regarding overall effect or non-linearity, but it must be stressed that they are approximated.

APPLICATION

The methodology is illustrated using data from a field experiment conducted at the Agricultural Experimental Station (EEA) of the National Institute of Agricultural Technology (INTA), in Marcos Juárez, Córdoba, Argentina in 2007.

We consider the following variables and we can see the join scatter plot:

- *Green weight(y):* is the wet weight of 12 grains for each plot (3-pin and each 4 grains).
- *Thermal sum(x_1):* is the accumulated mean temperature from anthesis in degree days (°Cd).
- *Moisture content(x_2):* is the moisture content of 12 grains for each plot; before the drying time.



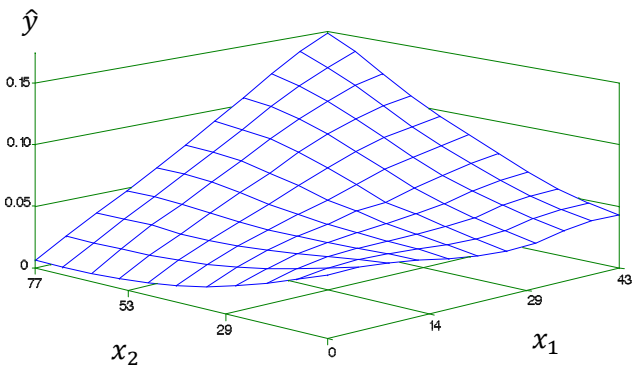
In order to find a functional relationship that can describe the data behavior it is proposed to set the following P-spline additive regression model with truncate bases.

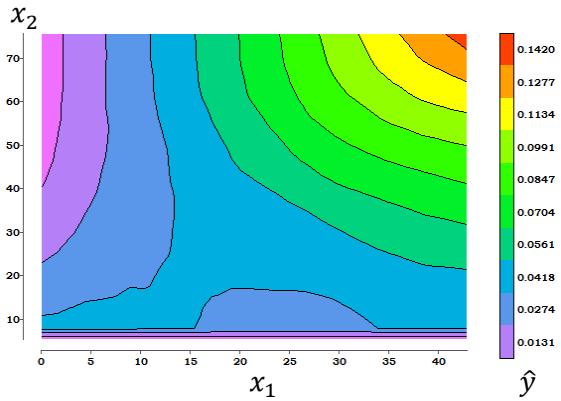
$$y_i = \beta_0 + \beta_{x_1} x_{1i} + \sum_{k=1}^{10} u_k^{x_1}\left(x_{1i} - c_k^{x_1}\right)_+ + \beta_{x_2} x_{2i} + \sum_{k=1}^{35} u_k^{x_2}\left(x_{2i} - c_k^{x_2}\right)_+ + \varepsilon_i$$

where $\left(x_{1i} - c_k^{x_1}\right)_+$ and $\left(x_{2i} - c_k^{x_2}\right)_+$ are the respective truncated bases for x_1 and x_2, representing piecewise linear fit, and $c_1^{x_1}, \ldots, c_{10}^{x_1}$ and $c_1^{x_2}, \ldots, c_{35}^{x_2}$ are the knots of each of variable.

From the results of model fitting and hypothesis testing, it is clear that to explain the grain green weight of wheat both variables, thermal sum and moisture content, are needed. Moreover, the relationship between the explanatory variables and the response should be preferably explained by smoothing curves, which justifies and encourages the use of additive models.

To conclude the analysis, we have the 3D graphic and the contour plot of the fitted additive model which displays the functional relationship of the variables, obtaining a more explicit idea of the represented shape.





DISCUSSION

P-spline smoothing techniques proposed to estimate the additive model functions have many advantages, mainly because they can be written as linear mixed models. Thus, it has all the methodology already developed in this area and, in turn, access to computer software available to work with this type of models.

About the additive model finally adopted to detail the grain green weight in terms of the thermal sum and humidity, there are some features that could be considered in future scans, which are not of easy theoretical resolution:

- Possible interactions between the explanatory variables (in this case should not work with truncated bases, we recommend the use of B-spline bases.)
- Eventually heteroscedasticity, apparently corresponding to the thermal sum, since the marginal scatter plot shows that the variability of the data increases significantly for high levels of degree days. (Suggested to consider in the covariance matrix of the errors).
- Potential need of including other explanatory variables in the model that may be related to grain green weight of the wheat grain, and therefore the shape of the surface that describes it.

Finally, it is interesting to note that the additive models under the approach of using P-splines belong to a research area that is booming nowadays, and presents a wide range of options for development and application in different areas of knowledge.