Modeling latent curves for genotype by environment interaction

Sabine K. Schnabel[1,2], Fred A. van Eeuwijk[1,2] and Paul H.C. Eilers[1,3]

[1] Biometris, Wageningen University and Research Centre, The Netherlands
[2] Centre for BioSystems Genomics, Wageningen, The Netherlands
[3] Erasmus Medical Center, Rotterdam, The Netherlands

E-mail for correspondence: sabine.schnabel@wur.nl

Abstract: In plant research data for a population of plants is often collected through repeated field and greenhouse trials under different environmental conditions. An appropriate model for this type of data should also deal with genotype by environment interaction. While the environments in field trials can frequently be characterized by geographic or meteorological conditions, it is equally likely that they are ordered according to a latent property. Therefore the order of the environments can often be unknown or unclear. We propose to use smooth latent curves to estimate this underlying order. The method is illustrated with simulated and empirical data.

Keywords: Latent curves; genotype by environment interaction; smoothing; P-splines

1 Introduction

A typical result of breeding trials with plants is a table in which for each genotype (G) a characteristic property (a so-called phenotype) is recorded for a number of environments (E). Generally, this kind of table cannot be fitted well by an additive model with effects for G and E. Hence we speak of genotype-by-environment (GxE) interaction. In general, the resulting data is given in a GxE table of phenotypic means. Most methods use centered data: row-wise, column-wise or double-centered.
In the literature one finds several ways to proceed which are essentially all based on the addition of multiplicative components. References to this model go back as far as Fisher and MacKenzie in 1923. Here, we explore another approach that assumes a smooth latent environmental gradient.
Suppose we have a set of n smooth curves –possibly with added noise– and we sample all of them at m positions given by the vector x. In the following we collect the data in an m by n matrix Y. Given this matrix and x, it is easy to estimate the curves by any smoothing method that works on each

column of Y separately. Permuting the rows randomly can do no harm, as long as the elements of x are permuted in the same way.
Imagine that the rows are permuted indeed and that we lost the vector x. How can we reconstruct it in a decent way? We propose a model that can perform this task. For a given GxE table it provides a latent gradient, the estimated x, that can be interpreted as an unknown environmental characteristic.

2 The model

Assume that x is given and we will smooth one column of Y denoted as y. An attractive choice is to use P-splines, which minimize

$$S_j = \sum_i (y_i - \sum_k a_k B_k(x_i))^2 + \lambda \sum_k (\Delta^2 a_k)^2 \tag{1}$$

Here, $B_k(x_i)$ is one of a set of (cubic) splines. The set is large and based on equally spaced knots. To tune smoothness a difference penalty is applied to the coefficients – see Eilers and Marx (1996) for details. If we consider all columns of Y, the objective function is the same for each of them, but we have to index the coefficients for the columns, to get a_{jk}. The overall objective function becomes

$$S = \sum_j \sum_i (y_{ij} - \sum_k a_{jk} B_k(x_i))^2 + \lambda \sum_j \sum_k (\Delta^2 a_{jk})^2. \tag{2}$$

Notice that the function is separable, i.e. we can smooth each column of Y separately.
Now assume that the coefficients are given as $A = [a_{jk}]$. We do not know x, but an approximation $\tilde{x}$ to it. Therefore we want to compute a vector of corrections u minimizing

$$S^* = \sum_j \sum_i (y_{ij} - \sum_k a_{jk} B_k(\tilde{x}_i + u_i))^2. \tag{3}$$

We try to get as good a fit as possible given the coefficients A by shifting the positions of the rows of Y.
If the corrections u are small, we can use the following first order approximation:

$$B_k(\tilde{x}_i + u_i) \approx B_k(\tilde{x}_i) + u_i B'_k(\tilde{x}_i), \tag{4}$$

where $B'_k(\tilde{x}_i)$ is the first derivative of the kth spline evaluated at $\tilde{x}_i$. It is easy to see that minimization of S^* in (3) leads to regression of the residuals

$$r_{ij} = y_{ij} - \sum_k a_{jk} B_k(\tilde{x}_i) \tag{5}$$

on the derivatives

$$g_{ij} = \sum_k a_{jk} B'_k(\tilde{x}_i) \tag{6}$$

This is again a separable problem leading, for each i, to

$$u_i = \sum_j r_{ij} g_{ij} / \sum_j g_{ij}^2. \tag{7}$$

Derivatives of cubic B-splines are easily computed by combining quadratic B-splines (using the same knots), differences of the coefficients and a correction for the knot distance.
Now we have the building blocks for an iterative algorithm:

- Fit the P-spline using the current estimates $\tilde{x}$.
- Update $\tilde{x}$.
- Repeat until convergence.

The scale and location of x is arbitrary, so we have to choose a normalization. Our choice is to scale and shift it after each iteration, so that the minimum is 0 and the maximum is 1.
The iterative procedure is straightforward, but the crucial step is the choice of the starting values for $\tilde{x}$. We have experimented with two approaches. One is to compute the singular value decomposition of Y and use the singular vector connected to the largest singular value (after proper normalization). In simulations this approach gave mixed results. An alternative is to use a random start. This seems to work well, but tens of trials may be needed. The random start vector minimizing the final S^* is chosen.
Here, the role of the penalty with smoothing parameter λ is minor, in the sense that it prevents singularities when fitting the splines. Without the penalty, an unfortunate choice of $\tilde{x}$ might lead to missing support for one or more B-splines.

3 Some results

Figure 1 shows the results of a simulation with eight curves in 15 hypothetical environments. Some of the simulated curves are almost linear, others show curvature. We permuted the rows in the originally simulated data set. A set of 50 random starts was tried. The knot distance for the B-spline basis is 0.05 and $\lambda = 1$. Convergence is linear and not too slow: after 50 iterations the size of the updates in u is of the order of 10^{-5}.
Gregorius and Namkoong (1986) use a small data set with five environments and information about the stem strength of six different types of pines. The results are presented in Figure 2. In these data the environments

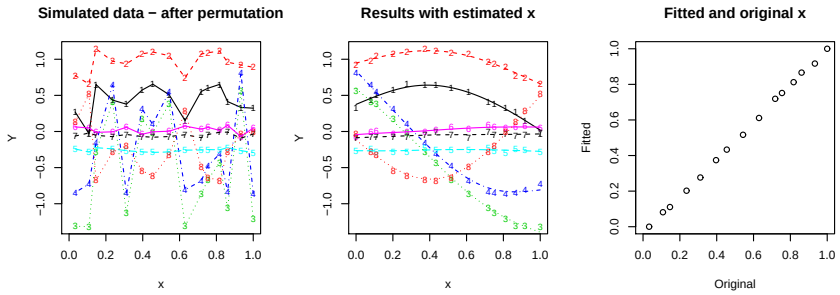

FIGURE 1. Simulated data with permuted x-axis (left), results from latent curve modeling (middle), original versus fitted x (right).

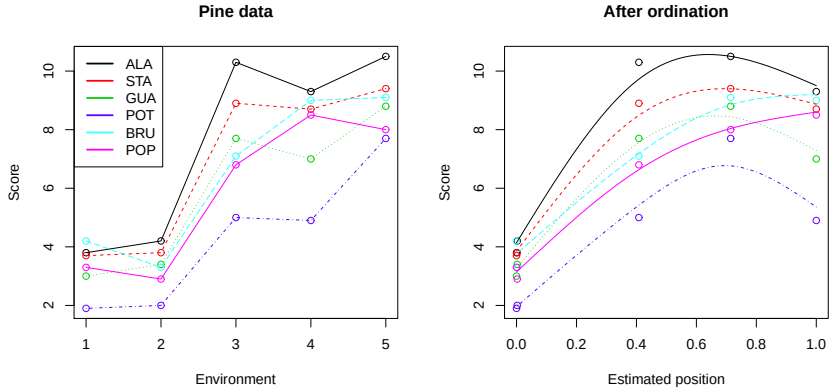

FIGURE 2. Data from Gregorius and Namkoong (1986). Original data (left), results with ordinated environments (right).

were already quite ordered, but the principle of the method is illustrated nevertheless.

Additionally we applied the technique to well-known textbook data from the plant breeding literature (Kleinhofs et al. 1993). A double haploid barley cross with 150 lines has been evaluated in 16 different environments and years. First analyses show promising results. In order to provide more guidelines for breeders for choosing well performing genotypes we suggest to estimate the relative performance of the genotypes. This second step can be done using using performance measurement based on expectiles as suggested by Schnabel and Eilers in 2009.

4 Conclusion

By using smooth latent curves to describe an unknown environmental gradient we propose a new approach to model genotype by environment interaction in plant breeding trials. The order of the environments can be unknown when they are ordered along a latent gradient. The result of our proposed method is an order of the environments which was initially unknown. These complete data can be used in further analysis to gain more insight about the relationship of the phenotypic trait and the environmental conditions. Additionally characteristics of the fitted curves might be associated with the genetic background of the plants. The model can also be extended to accommodate missing data through a weighting scheme. Results will be reported elsewhere.

References

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89 – 121.

Fisher, R.A. and MacKenzie, W.A. (1923). Studies in variation II. The manurial response in different potato varieties. *Journal of Agricultural Science*, **13**, 311 – 320.

Gregorius, H-R. and Namkoong, G. (1986). Joint analysis of genotypic and environmental effects. *Theoretical and Applied Genetics*, **72**, 413 – 422.

Kleinhofs, A., Kilian, A., Maroof, M.S., Biyashev, R., Hayes, P., Chen, F., Lapitan, N., Fenwick, A., Blake, T., Kanazin, V., Ananiev, E., Dahleen, L., Kudrna, D., Bollinger, J., Knapp, S., Liu, B., Sorrells, M., Heun, M., Franckowiak, J., Hoffman, D., Skadse, R., and Steffenson, B. (1993). A molecular, isozyme, and morphological map of the barley (hordeum vulgare) genome. *Theoretical and Applied Genetics*, **86**, 705 – 712.

Schnabel, S.K. and Eilers, P.H.C. (2009). An analysis of life expectancy and economic production using expectile frontier zones. *Demographic Research*, **21**, 109 – 134.