Modelling plant height data with scaled and shifted prototype curves

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

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

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

Abstract: In agricultural research phenotypic data are mainly collected through field and greenhouse experiments. Often a whole population of plants is monitored at different time points during the growing season. Here we are analyzing time series of plant height data in potato. The plant-specific data is modelled inspired by a model that was originally developed for a study of growth of children. We are using P-splines for a smooth curve and introduce a vertical as well as a horizontal shift per plant.

Keywords: growth modelling; smooth curve; shift; scale; P-splines

1 Introduction

Measurements on growing plants often show a characteristic shape: a monotonically increasing curve. In agricultural trials many (up to hundreds) of such curves are being collected for one genetic population. Data are collected either by hand or completely automatically. To combine such measurements with genomic data, meaningful summaries have to be developed. A number of methods have already been proposed to estimate good characteristics of plant growth curves. They include classical parametric approaches based on the logistic curve (Malosetti et al., 2006), a semi-parametric approach based on splines (Hurtado et al., 2012) and a survival analysis approach for phenotypic data series on an ordinal scale (Schnabel et al., 2010).

In this contribution we study models that assume that there is one prototype curve, which has been stretched and shifted on both the horizontal (time) and the vertical axes. The transformation parameters as well as the prototype curve itself are to be estimated from the data. In our example the response variable has been log-transformed. Therefore a shift along the vertical axis translates into a rescaling on the original response scale.

The inspiration for our model comes from a publication by Cole et al. (2010) based on work by Beath (2007). Cole and co-authors call their model

SITAR, which stands for *SuperImposition by Translation And Rotation*. This acronym is not obvious, as the rotation property is not immediately clear from the model. However, it shows an excellent fit to growth data of children as demonstrated in an example for height. It is relatively parsimonious, because the curve for any individual is summarized with only three parameters in addition to the spline coefficients common to all curves.
The model has one parameter for shifting along the vertical scale. Given the prototype curve it can be estimated by linear regression. However, the two other parameters occur in the argument of the curve and lead to a non-linear problem see (1). Beath used natural B-splines to model the prototype curve and non-linear mixed models to estimate the parameters. After some experiments with the software provided in his paper, we decided to start from scratch. We model the curves with P-splines. In addition because of the rather precise and detailed data, we drop the mixed model approach. A fixed model is sufficient. We also experiment with simplifications using less parameters and initially include only two parameters.
In the next section we present briefly the original model and explain our approach and its estimation procedure. Finally we apply it to plant height measurements from a potato field experiment.

2 Method

Epidemiological studies often deal with longitudinal data for developmental characteristics of the cohort under study. Cole et al. (2010) presented the SITAR model based on an earlier paper (Beath, 2007). Both publications propose a shape invariant model with a single fitted curve. There are three different mechanisms that drive the shape of an individual curve in relation to the mean curve for the whole population: a curve can be shifted up or down, left or right, or the scale on the x-axis can be shrunk or stretched. The original SITAR model uses three subject-specific random effects for the characterization of a response y_{ij} for subject i at age j:

$$y_{ij} = \alpha_i + h\left(\frac{t - \beta_i}{\exp(-\gamma_i)}\right) \tag{1}$$

with α a random intercept adjusting for height, β a random shift along the x-axis and γ a random scaling factor. The three parameters are termed *size*, *tempo* and *velocity* respectively by Cole et al.. In the original notation $h(\cdot)$ is a natural cubic spline of the response over age t. This model formulation has the advantage that the parameters are directly interpretable in a biological context.
Inspired by this model we propose a simplification and adaptation of it. Instead of the mixed model we are sticking to a fixed model using P-splines for the functional form (Eilers and Marx, 1996). The response is log-transformed in our context.

As an initial step we formulate the curves for genotype i at time j as

$$\log(y_{ij}) = \alpha_i + f(t_j) \tag{2}$$

including a subject-specific vertical shift α (as intercept of the model or *size* in Cole's terminology). In our case we use P-splines for the functional form f, therefore:

$$\log(y_{ij}) = \alpha_i + \sum_k b_{jk}\beta_k \tag{3}$$

with $B = [b_{jk}]$ a B-spline basis with a generous number of splines and β the associated coefficients. In the implementation two parameters λ –for a difference penalty– and κ –for a ridge penalty– are included to ensure smoothness as well as numerical stability.
In order to correct for the horizontal shifts that the different curves undergo, we introduce a transformation of the horizontal axis. To this end we rewrite and substitute in (3):

$$\begin{aligned} b_{jk} &= B_k(t_j) \\ b_{ijk} &= B_k(t_j + \delta_i) \approx B_k(t_j) + \delta_i \dot{B}_k(t_j) \end{aligned} \tag{4}$$

where $\dot{B}_k(t_j)$ is the first derivative of the kth spline evaluated at time t_j. δ is the so-called *tempo* effect inducing a non-linear transformation of the horizontal scale.
In the model above we assume that for all subjects i the measurements are taken at the same time points t_j. However, this might not be the case in all applications. This can be easily included by using subject-specific time points t_{ij}.

3 Application

We analyze data from a field experiment with a diploid potato mapping population with more than 150 different genotypes. Over the course of the growing season different characteristics are measured. Plant height (in cm) has been assessed at nine time points over the course of three months. Figure 1(a) shows the log-transformed data of this heterogeneous potato population. We apply the model including a vertical and horizontal shift as explained above. The fitted curves and its residuals are depicted in Figures 1(b) and 1(d). Figure 1(c) shows a scatterplot of the results on the shifted scales $\log(y_{ij}) - \alpha_i$ versus $t_j + \delta_i$. The colors are ordered according to the order of the individual *tempo* effects δ_i.

4 Conclusion and Discussion

Our model, inspired by Cole et al. (2010), was successfully applied to longitudinal phenotypic data generated in a field experiment. We can estimate

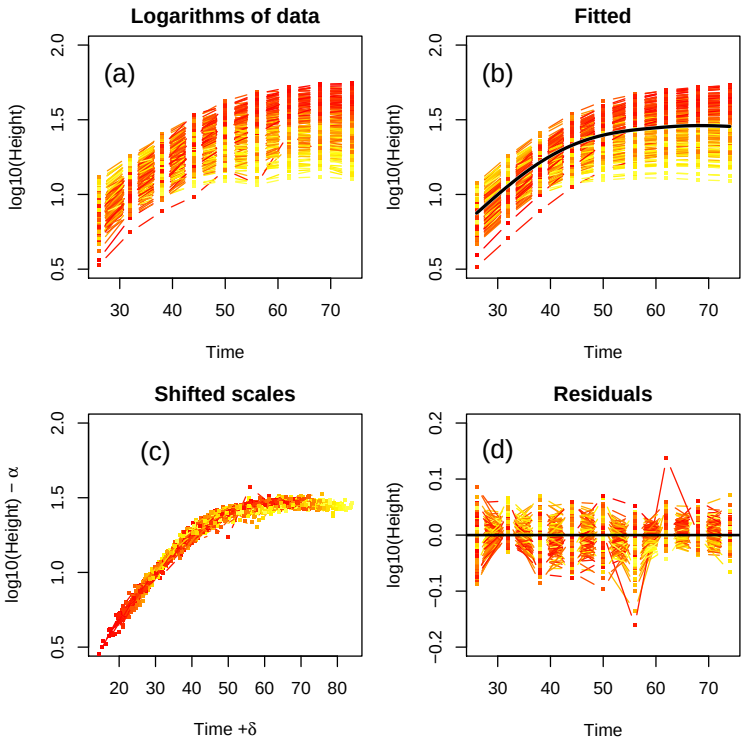

FIGURE 1. (a) Log-transformed data for the whole population, (b) fitted curves using vertical and horizontal shifts with the mean curve, (c) scatterplot on shifted scales, (d) residuals .

a mean curve that provides general information about the growth of the potato plants in the population as a whole. More importantly we also determine characteristics for the individual genotypes that can be used in further genetic analyses. A scatterplot of the vertical and horizontal shifts per genotype can be found in Figure 2.
In future work we plan to extend the model in different ways. At the moment our model includes a so-called *size* as well as a *tempo* effect. In a next step we plan to extend it with a *velocity* effect γ:

$$b_{ijk} = B_k(\gamma_i t_j + \delta_i). \tag{5}$$

A preliminary analysis with the example data this extension did not seem to improve the estimation results, but it is important for future applications. Although the data presented in this manuscript are typical for plant breeding trials, they are still a simplification of the real data situation. To complicate matters data are often measured for several replications of the same

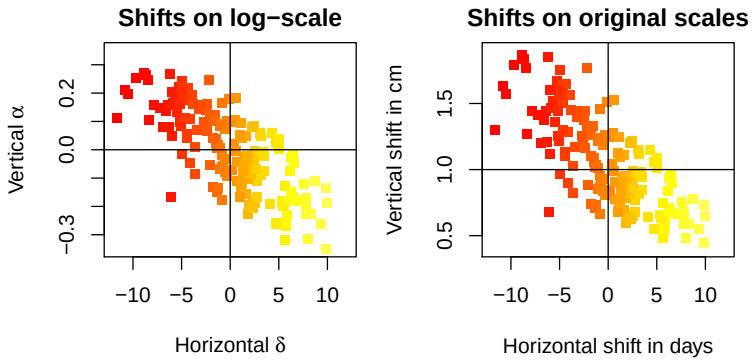

FIGURE 2. Right: horizontal shift (δ) versus vertical shift (α) on log scale for the fitted curves. Left: horizontal shifts (in days) versus vertical shifts (in cm). Colour codes in increasing size of the tempo effect δ.

genetically identical plant. Additionally in some trials the data might include a mixture of cross-sectional and longitudinal data due to intermediate harvest of parts of the experimental field. For replicates a direct solution is at hand by extending the model to accommodate replicates as a random effect within the genotype. Mixture of data through different collection methods need a more theoretical treatment before this situation can be integrated in the current context. Last but not least the estimated individual characteristics of the genotypes will be used in further genetic analysis to associate these with regions on the chromosomes. In order to offer more powerful tools for the plant research community these topics will be treated in future work and reported elsewhere.

Acknowledgments: The data set used in the application has been collected at the Holetta Agricultural Research Center in Holetta, Ethiopia, by Biructawit Bekele Tessema who is financed by a grant from NUFFIC and is part of the Laboratory of Plant Breeding at Wageningen University and Research Centre.

References

Beath, K.J. (2007). Infant growth modelling using a shape invariant model with random effects. *Statistics in Medicine*, **26**, 2547 – 2564.

Cole, T.J., Donaldson, M.D.C., and Y. Ben-Shlomo (2010). SITAR — a useful instrument for growth curve analysis. *International Journal of Epidemiology*, **39**, 1558 – 1566.

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

Hurtado, P., Schnabel, S., Zaban, A., Veteläinen, M., Virtainen, E. , Eilers, P., van Eeuwijk, F., Visser, R., and Maliepaard, C. (2010). Dynamics of senescence-related QTL in potato. *Euphytica*, **183**, 289 – 302.

Malosetti, M., Visser, R.G.F., Celis-Gamboa, C., and van Eeuwijk, F.A. (2006). QTL methodology for response curves on the basis of non-linear mixed models, with an illustration to senescence in potato. *Theoretical Applied Genetics*, **113**, 288 – 300.

Schnabel, S.K., Eilers, P.H.C., Hurtado López, P., Visser, R.G.F., and van Eeuwijk, F.A. (2010). Haulm senescence in potatoes and semi-parametric survival models. In: *Proceedings of the 25th International Workshop on Statistical Modelling*, Glasgow, UK, pp. 489 – 494.