TECHNICAL REPORT

0535

SIMPLE AND MULTIPLE P-SPLINES REGRESSION WITH SHAPES CONSTRAINTS

BOLLAERTS, K., EILERS , P., and I. VAN MECHELEN

IAP STATISTICS NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Simple and Multiple P-splines regression with Shape Constraints

Kaatje Bollaerts

Katholieke Universiteit Leuven

Paul H.C. Eilers

Leiden University Medical Centre

Iven Van Mechelen

Katholieke Universiteit Leuven

Author notes: The authors are grateful to Patrick Groenen for his helpful suggestions on convexity and to Han van der Maas for placing his data at our disposal. All correspondence concerning this paper should be sent to Iven Van Mechelen, Department of psychology 102, B-3000. E-mail: iven.vanmechelen@psy.kuleuven.ac.be

Abstract

In many research areas, especially within social and behavioral sciences, the relationship between predictor and criterion variables is often assumed to be of a particular non-parametric functional form, such as monotone, single-peaked or U-shaped. Such assumptions can be transformed into (local or global) constraints on the sign of the n^{th} order derivative of the functional form. To check for such assumptions, we present a non-parametric regression method, P-splines regression with additional asymmetric discrete penalties enforcing the constraints. We show that the corresponding loss function is convex and present a Newton-Raphson algorithm to optimize. Constrained P-splines will be illustrated with an application on monotonicity constrained regression with both one and two predictor variables using data from research on cognitive development of children.

Keywords: non-parametric regression, P-splines, shape constraints, monotonicity, one-dimension, two-dimensions

1 Introduction

In many research areas much effort is made to investigate the relationship between a set of predictor variables (also called independent variables) X_i and a criterion or dependent variable Y. Such a relationship is mostly denoted as a function f in $Y = f(X_1, \dots, X_i, \dots, X_n)$. Often, salient features of f are assumed without presupposing any parametric functional form such as monotonicity, U-shapedness and single-peakedness. In the following we will discuss some substantive psychological theories in which these non-parametric functional forms play a central role.

Monotonicity means that f is either monotone decreasing or monotone increasing. As an example, consider the function f which expresses the relationship between some measure of cognitive performance of children and age. In this case, it is common to assume that f is a monotone increasing function; note in this respect that assuming monotonicity is more plausible than assuming linearity or an exponential function. Another example stems from the trait approach to personality (McCrae & Costa, 1987). Within this approach, it is assumed that the ordering of persons with respect to a particular behavior, such as fighting in a specific situation, corresponds to the ordering of persons with respect to some relevant underlying trait (e.g., aggressiveness). Hence, behavior of persons is assumed to be a monotone increasing function of the relevant underlying trait, even though no specific parametric assumption is made.

A second non-parametric functional form we may consider is *U-shapedness* as put forward in many theories concerning growth and development. Strauss (1982) defines U-shaped behavioral growth curves as curves indicating the initial appearance of a behaviour, a later dropping out of that behavior and subsequently, its reappearance.

A third non-parametric functional form we consider is *single-peakedness*. A single-peaked function is a function that increases up to some point and then decreases. The importance of non-parametric single-peaked functions has been underscored by Coombs (1977). In particular, preference and psychophysical functions are frequently observed to be single-peaked.

So far, we only considered examples of assumptions on non-parametric functional forms with regard to a single predictor variable. However, non-parametric functional forms with respect to two or more predictor variables can be assumed as well. For instance, one may consider the assumed monotone increasing functional relationship between some measure of cognitive performance of children and both age and amount of training, or between aggressiveness of persons and both trait anger and the anger-eliciting power of the situation (McCrae & Costa, 1987). Still another example concerns the singlepeaked preference functions for options varying along two dimensions like preference for beer as a function of its alcohol level and its bitterness (Coombs, 1977).

The examples above illustrate the versatility of assumptions that imply non-parametric functional forms. In this paper we will introduce a statistical tool by which such assumptions can be dealt with, namely constrained P-splines regression. P-splines regression as such is introduced in a target article by Eilers & Marx (1996) with illustrations on data containing a single predictor variable. An illustration of P-spline regression with two predictor variables is given in Durban et al. (2002). In the present paper the main focus is on P-splines regression with shape constraints, which is briefly introduced in Eilers (1994). Both illustrations with one and two predictor variables will be given.

The remainder of this paper is organised as follows: In Section 2 , we will discuss simple as well as multivariate unconstrained and constrained P-splines regression. In Section 3, an application of simple and multiple P-splines regression with monotonicity constraints is given with data from research on cognitive development of children. In Section 4 we present some concluding remarks. The condition and optimization of the loss function are discussed in Appendix 1 and 2 respectively.

2 Method

In this section,we successively discuss unconstrained and constrained P-splines regression. In each case, we will discuss regression with one and two predictor variables. The discussion of unconstrained P-splines regression with a single predictor variable is mainly based on Eilers & Marx (1996).

2.1 Unconstrained P-splines regression

2.1.1 Simple regression

Eilers and Marx (1996) introduced non-parametric regression with P-splines, which is essentially least squares regression with an excessive number of univariate B-splines (De Boor, 1978; Dierckx, 1993) and an additional discrete penalty to correct for overfitting.

Univariate B-splines are piecewise linear functions with local support. A B-spline of degree q consists of q+1 polynomial pieces of degree q joined smoothly (i.e., differentiable in order $q-1$) at q points λ_i (called interior knots) between boundaries λ_{min} and λ_{max} (called exterior knots) and with a positive value between and a value of zero outside these boundaries. An example of a B-spline of the first degree is given in Figure 1a; it is clear that this B-spline consists of two linear pieces joined smoothly at one interior knot. An example of a B-spline of the third degree is given in Figure 1b; this B-spline consists of four cubic pieces joined smoothly at three interior knots. Note that the B-splines shown in Figure 1 both have equally-spaced knots. B-splines with unequally-spaced knots exist as well, but are not considered in this paper.

Insert Figure 1 about here

In order to use B-splines for non-parametric regression, a basis of r overlapping B-splines is constructed, which is such that

$$\forall x : \sum_{j=1}^{r} B_j(x, q) = 1 \tag{1}$$

with $B_j(x, q)$ denoting a B-spline of degree q with left most knot j. In Figure 2a, one can see an example of a basis of B-splines of the third degree, which is the most commonly used degree in B-splines regression.

The B-splines of a B-spline basis act as the predictors in spline regression. Given m observations (x_i, y_i), least squares regression with B-splines of Y on the basis of X comes down to minimizing the following loss function:

$$S = \sum_{i=1}^{m} (y_i - \widehat{y}_{(\alpha)_i})^2 \tag{2}$$

with

$$\widehat{y}_{(\alpha)_i} = \sum_{j=1}^{r} \alpha_j B_j(x_i, q) \tag{3}$$

and the α_j's being the coefficients (or amplitudes) of the corresponding B-splines. In Figure 2b, spline regression with B-splines of the third degree is illustrated.

Insert Figure 2 about here

A major problem in B-spline regression is the choice of the optimal number of B-splines. An insufficient number of B-splines leads to underfitting such that the fitted curve is too coarse and, hence, relevant information is neglected. On the other hand, too many B-splines leads to overfitting such that the fitted curve is too flexible and hence, random fluctuations are modelled. To overcome this problem, O'Sullivan (1988) suggested to use an excessive number of B-splines with, in order to correct for overfitting, a smoothness penalty consisting of the integrated squared second order derivative of the fitted curve; this approach has become standard in spline literature. Eilers and Marx (1996) propose to use a discrete smoothness penalty based on (second) order differences of the coefficients of adjacent B-splines; this approach, also called P-splines regression, is very similar to O'Sullivan's. Furthermore, P-splines regression is easy to implement, has no boundary effects, conserves moments and has polynomial curve fits as limits. For a penalty on second order differences, the corresponding least squares loss function equals:

$$S = \sum_{i=1}^{m} (y_i - \widehat{y}_{(\alpha)_i})^2 + \lambda \sum_{j=3}^{r} (\triangle^2 \alpha_j)^2 \tag{4}$$

with $\triangle^2 \alpha_j$ being the second order differences, that is

$$\triangle^2 \alpha_j = \triangle^1(\triangle^1 \alpha_j) = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$$

The amount of smoothness can be controlled for by means of λ, which is a user-defined smoothness parameter. If $\lambda \to \infty$, then, for regression with a smoothness penalty on m^{th} order differences, the fitted function will approach a polynomial of degee $m-1$. To choose an optimal value for λ, Eilers and Marx (1996) propose to use Akaike's information criterion:

$$AIC(\lambda) = -2L(\alpha, \lambda) + 2dim(\alpha, \lambda) \tag{5}$$

with $L(\alpha, \lambda)$ denoting the loglikelihood of the data and $dim(\alpha, \lambda)$ the effective dimension of the vector of parameters. The determination of the latter requires some extra attention. Indeed, since the rationale behind P-spline regression is to use an excessive number of B-splines with a penalty to correct for overfitting, the total number of parameters of the P-spline model is an overestimation of the effective dimension of the vector of parameters. This problem can be solved by using the trace of the hat matrix **H** as an approximation of the effective dimension of the vector of parameters (Hastie & Tibshirani, 1990). Then, under the assumption of normally distributed errors, $y_i \sim \mathcal{N}(\widehat{y}_i, \sigma^2)$, Akaike's information criterion equals:

$$AIC(\lambda) = \sum_{i=1}^{m} \frac{(y_i - \hat{y}_i)^2}{\hat{\sigma}^2} + 2m\ln(\hat{\sigma}) + m\ln(2\pi) + 2\text{tr}(H) \tag{6}$$

As an estimate of the nuisance parameter $\hat{\sigma}^2$, Eilers & Marx (1996) propose to use the variance of the residuals computed for an optimal value for λ chosen on the basis of (generalized) cross-validation.

2.1.2 Multiple regression with two predictor variables

P-splines regression with two predictor variables is a straightforward extension of P-spline regression with one predictor variable as introduced in the previous section. The constitutive elements of P-splines regression with two predictor variables are bivariate B-splines, illustrated in Figure 3a. A bivariate B-spline of degree q is a tensor product of two univariate B-splines of degree q, that is

$$B_{jk}(X_1, X_2, q) = B_j(X_1, q) \otimes \check{B}_k(X_2, q) \tag{7}$$

Insert Figure 3 about here

A basis of bivariate B-splines of degree three is displayed in Figure 3b. Then, for m observations (X_{1i}, X_{2i}, Y_i) regression of Y on X_1 and X_2 with a basis of $r \times r'$ overlapping bivariate B-splines comes down to minimizing the following least squares loss function:

$$S = \sum_{i=1}^{m} (y_i - \widehat{y}_{(\alpha)_i})^2, \tag{8}$$

with

$$\widehat{y}_{(\alpha)_i} = \sum_{j=1}^{r} \sum_{k=1}^{r'} \alpha_{jk} B_j(X_{1i}, q) \otimes \check{B}_k(X_{2i}, q) \tag{9}$$

where α_{jk}'s are the coefficients of the corresponding bivariate B-splines.

In P-splines regression with two predictor variables, two smoothness penalties are used, one for each predictor variable. For penalties on second order differences, the loss function to be minimized is

$$S = \sum_{i=1}^{m} (y_i - \widehat{y}_{(\alpha)_i})^2 + \lambda_{x_1} \sum_{j=3}^{r} \sum_{k=1}^{r'} (\triangle_{x_1}^2 \alpha_{jk})^2 + \lambda_{x_2} \sum_{k=1}^{r} \sum_{j=3}^{r'} (\triangle_{x_2}^2 \alpha_{jk})^2 \tag{10}$$

with $\triangle_{x_1}^2$ being a columnwise and $\triangle_{x_2}^2$ a rowwise smoothness penalty on the matrix $\mathbf{A} = [\alpha_{jk}]$, and with λ_{x_1} and λ_{x_2} being the smoothness parameters of X_1 and X_2 respectively.

2.2 Constrained P-splines regression

2.2.1 Simple regression

As indicated in the introduction, assumptions of non-parametric functional forms that can be transformed into local or global constraints on the sign of the n^{th} order derivative, can be checked using constrained P-splines regression. This is P-splines regression with an asymmetric discrete penalty on n^{th} order differences, reflecting the assumed non-parametric functional form. This penalty is asymmetric since it differentially penalizes positive and negative n^{th} order differences, in order to restrict the sign of the n^{th} order differences to be positive (resp. negative). The latter implies a positive (resp. negative) n^{th} order derivative of the fitted function. Indeed, the first order derivative of a B-splines function with equally spaced knots equals

$$f^{(1)}(x) = \frac{\partial f(x)}{\partial x} = \frac{\partial}{\partial x} \sum_{j=1}^{r} \alpha_j B_j(x, q) = (qh)^{-1} q \sum_{j=1}^{r+1} \triangle_{\alpha_j}^{1} B_j(x, q-1) \tag{11}$$

with h denoting the distance between two adjacent knots (De Boor, 1978). By induction, the n^{th} order derivative of a B-splines function is

$$f^{(n)}(x) = \frac{\partial f^{(n-1)}(x)}{\partial x} = \prod_{l=1}^{n} [(q+1-l)h]^{-1} (q+1-l) \sum_{j=1}^{r+n} \triangle_{\alpha_j}^{n} B_j(x, q-n) \tag{12}$$

Then, since h, $q+1-l$ and $B_j(x, q-n)$ are all positive by definition, restricting $\triangle_{\alpha_j}^{n}$ to be positive (resp. negative) is a sufficient condition for the $f^{(n)}(x)$ to be positive (resp. negative). In addition, for $q-n=0$ and $q-n=1$, these sufficient conditions are necessary as well since in that case $f^{(n)}(x)$ is piecewise constant (resp. piecewise linear).

Hence, a penalty reflecting the constraint of a positive (resp. negative) n^{th} order derivative within a range as defined by indicator variable v_j is

$$\sum_{j=n+1}^{r} v_j w(\alpha)_j (\triangle^n \alpha_j)^2 \tag{13}$$

with

$$v_j = \begin{cases} 1 & \text{if the constraint on } \partial^{\,n} f(x)/\partial x^n \text{ is to hold on at least part of the support of } B_j \\ 0 & \text{otherwise} \end{cases}$$

and with

$$w(\alpha)_j = \begin{cases} 0 & \text{if } \triangle^n_{\alpha_j} \geq 0 \text{ (resp. } \triangle^n_{\alpha_j} \leq 0) \\ 1 & \text{otherwise} \end{cases}$$

being asymmetric weights. As can be easily seen, negative (resp. positive) values for $\triangle^n \alpha_j$ are penalized whereas non-negative (resp. non-positive) are not. Then, with κ being a user-defined constraint parameter, the overall loss function reads as:

$$S = \sum_{i=1}^{m} (y_i - \widehat{y}(\alpha)_i)^2 + \lambda \sum_{j=3}^{r} (\triangle^2 \alpha_j)^2 + \kappa \sum_{j=n+1}^{r} v_j w(\alpha)_j (\triangle^n \alpha_j)^2 \tag{14}$$

which is convex in α. [For a proof, see Appendix 1.]

2.2.2 Multiple regression with two predictor variables

Again, it is straightforward to extend constrained P-splines regression with one predictor variable to constrained P-splines regression with two predictor variables. In the latter case, two constraint penalties are used, such that the corresponding loss function equals

$$
\begin{aligned}
S &= \sum_{i=1}^{m}(y_i - \widehat{y}(\alpha)_i)^2 + \lambda_{x_1}\sum_{j=3}^{r}\sum_{k=1}^{r'}(\triangle_{x_1}^2\alpha_{jk})^2 + \lambda_{x_2}\sum_{j=1}^{r}\sum_{k=3}^{r'}(\triangle_{x_2}^2\alpha_{jk})^2 \\
&+\kappa_{x_1}\sum_{j=n+1}^{r}\sum_{k=1}^{r'} v_{x_1.jk}w(\alpha)_{x_1.jk}(\triangle_{x_1}^n\alpha_{jk})^2 + \kappa_{x_2}\sum_{j=1}^{r}\sum_{k=n+1}^{r'} v_{x_2.jk}w(\alpha)_{x_2.jk}(\triangle_{x_2}^n\alpha_{jk})^2 \quad (15)
\end{aligned}
$$

with $\triangle_{x_1}^n\alpha_{jk}$ being a columnwise and $\triangle_{x_2}^n\alpha_{jk}$ a rowwise constraint penalty on the matrix $\mathbf{A} = [\alpha_{jk}]$, $v_{x_1.jk}$ and $v_{x_2.jk}$ being indicator variables defining the range for which the constraints should hold, with

$$
v_{x_1.jk} = \begin{cases} 1 & \text{if the constraint on } \partial\, ^n f(x_1,x_2)/\partial x_1^n \text{ is to hold on at least part of the support of } B_{jk} \\ 0 & \text{otherwise} \end{cases}
$$

$$
v_{x_2.jk} = \begin{cases} 1 & \text{if the constraint on } \partial\, ^n f(x_1,x_2)/\partial x_2^n \text{ is to hold on at least part of the the support of } B_{jk} \\ 0 & \text{otherwise} \end{cases}
$$

and with $w(\alpha)_{x_1.jk}$ and $w_{(\alpha)_{x_2.jk}}$ denoting asymmetric weights defined as

$$
w(\alpha)_{x_1.jk} = \begin{cases} 0 & \text{if } \triangle_{x_1}^n\alpha_j \geq 0 \text{ (resp. } \triangle_{x_1}^n\alpha_j \leq 0) \\ 1 & \text{otherwise} \end{cases}
$$

$$
w_{(\alpha)_{x_2.jk}} = \begin{cases} 0 & \text{if } \triangle_{x_2}^n\alpha_j \geq 0 \text{ (resp. } \triangle_{x_2}^n\alpha_j \leq 0) \\ 1 & \text{otherwise} \end{cases}
$$

for the constraint of a positive (resp. negative) n^{th} order derivative with respect to x_1 and x_2 respectively, and with κ_{x_1} and κ_{x_2} being user-defined constraint parameters. The loss function in (15) is convex, of which the proof is a straightforward extension of the one given in Appendix 1.

2.3 Algorithm

We will make use of a Newton-Raphson procedure in order to find an optimal solution of the loss functions described in (14) and (15). An iteration of this procedure comes down to calculating $w(\alpha)$ on α as estimated in the previous iteration and calculating the new estimates, α', conditional on $w(\alpha)$. A schematic presentation of the algorithm reads as follows.

1. l $\leftarrow$ 0
2. set initial weights $\mathbf{W}_{(\alpha)}^{(l)} = [\mathbf{0}]$
3. l $\leftarrow$ l + 1
4. estimate $\boldsymbol{\alpha}^{(l)}$ on $\mathbf{W}_{(\alpha)}^{(l-1)}$
5. calculate $\mathbf{W}_{(\alpha)}^{(l)}$ on $\boldsymbol{\alpha}^{(l)}$
6. repeat step 3, 4 and 5 until $\mathbf{W}_{(\alpha)}^{(l)} = \mathbf{W}_{(\boldsymbol{\alpha})}^{(l+1)}$.
7. if $\mathbf{W}_{(\alpha)}^{(l)} = \mathbf{W}_{(\alpha)}^{(l+1)}$, $\boldsymbol{\alpha}^{(l)}$ is the optimal solution looked for

For a more indepth discussion the reader is referred to Appendix 2. The corresponding MATLAB software is available upon request.

3 Application

In this section we discuss an application on monotonicity constrained P-splines regression, which is P-splines regression with an additional discrete penalty, forcing the first order differences to be positive. The data we use to illustrate come from a study of van der Maas (1992, 1993) on cognitive development of children. In this study the understanding that an amount of liquid remains the same when you pour it into another container, that is conservation of liquid, is investigated. Conservation of liquid has been introduced by Piaget (1960), the well-kown pioneer of stagewise developmental evolution. Piaget distinguished between three acquisition stages: a nonconserving equilibrium stage, a transitional disequilibrium stage and a conserving equilibrium stage. In the nonconserving equilibrium stage, children believe that the amount of liquid may increase or decrease when it is poured from one container to another. In the transitional disequilibrium stage, children start to realize that pouring liquid from one container to another does not change quantity; however, this insight is not yet consolidated. From the conserving equilibrium stage only, children truly understand conservation of liquid. Based on this theory, Han van der Maas (1993) developed a computer test to measure conservation understanding. We will now give a description of this test.

3.1 Computer test of liquid conservation

The computer test of liquid conservation consists of three different parts (van der Maas, 1993). Since we only use data with respect to the first part, we restrict our description of the test to this part. The latter consists of eight items. Each item contains: (1) an initial situation consisting of two identical containers filled with liquid, (2) a transformation which comes down to pouring the liquid of one of the two containers into an empty container with a different shape, and (3) the resulting situation. Both the initial and the resulting situation are to be judged by the respondent using three response alternatives: more liquid in the left container, the same amount of liquid in both containers, more liquid in the right container. Furthermore, three different types of items are included in this part of the test: three standard equality items, three standard inequality items and two guess items. An example of each type is shown in Figure 4. In a standard equality item, though the containers of the initial situation are both filled with a same quantity of liquid, the height of the liquid differs in the two containers of the resulting situation. In a standard inequality item, the containers of the initial situation are not filled with a same quantity of liquid; however, the transformation results in equal heights of the liquid in both containers. Finally, in guess items the containers of the initial situation are not filled with a same quantity of liquid nor are the heights equal in the final situation.

Insert Figure 4 about here

With the help of these three types of items, differentiation between conservers, non-conservers and guessers is possible. Conservers, who understand that transformation does not change quantity, are expected to answer correctly that standard equality items contain a same quantity of liquid, whereas standard inequality and guess items do not. Non-conservers are assumed to focus on height and to conclude that two containers are filled with a same quantity of liquid if the heights are equal; this should result into correct answers for guess items and incorrect answers for standard items. Finally, guessers are expected to score at chance level irrespective of the item type.

3.2 Data

Participants were 101 children with ages ranging from 6.2 to 10.6 (van der Maas, 1996). The computer test was administered at 11 consecutive moments in time (i.e., 11 test sessions). The time between two successive sessions varied from two to four weeks. Many test sessions (265 out of 1111) were missing, however.

3.3 Monotonicity constrained simple regression

According to Piaget's theory, the performance of children on the liquid conservation test is expected to be a monotone non-decreasing function of time except during the transitional disequilibrium stage during which relapses are possible. With respect to the latter, however, sparse empirical results show that relapses primarily occur when countersuggestions or completely unfamiliar items are given (Inhelder, Sinclair, Bovet & Wedgwood, 1974), which is not the case in the study of van der Maas. Therefore, monotonicity can generally be assumed to hold for this study's data.

We check this assumption for four children with different overall performance levels on the liquid conservation test. Overall performance levels, which range from 1.67 to 8.00, are simply computed by averaging the child's scores across the sessions. The levels of the children selected are 3.1, 4.8, 5.9 and 7.5, respectively. For this analysis, we opt for a regression with 12 B-splines of the third degree and a second-order smoothness penalty. Regarding the smoothness weight λ, a value of 0.28 is choosen by making use of Akaike's information criterion assuming independence of the four children and with the variance $\widehat{\sigma}$ of the residuals in (6) being estimated on the basis of generalized cross validation. Regarding the monotonicity weight κ, we choose a value as high as 10^6 to assure that violations of the monotoncity assumption are negligible. The results, for both unconstrained and constrained regression, are graphically represented in Figure 5.

Insert Figure 5 about here

As a goodness-of-fit measure, we compute squared correlations between observed and predicted scores. If the data are approximately monotone increasing, unconstrained and constrained regression are expected to have a comparable fit, if not, constrained regression is expected to yield a much lower fit than unconstrained regression. The results are summarized in Table 1. Only for child 1, a significant discrepancy in fit-values is observed, which can be explained by the fact that this child seems to be purely guessing; additional support for this explanation is the child's overall performance level of 3.1, which comes close to the chance level of 2.7. For the other three children, who mainly differ with respect to the time period during which the transition from non-conserving to conserving occurs, the assumption of monotonicity seems justified.

Insert Table 1 about here

3.4 Monotonicity constrained regression with two predictor variables

Following Piaget's theory (1960), it can be hypothesized that performance of children on the liquid conservation test is a monotone non-decreasing function of both time and overall level of performance of the children. This assumption of double monotonicity can be taken apart in two separate assumptions. First, the performance on the liquid conservation test is expected to be a monotone non-decreasing function of time conditional on the overall level of performance, in line with the argument and results of the previous section. Second, the performance on the liquid conservation test is expected to be a monotone non-decreasing function of overall level of performance conditional on time. The latter is implied by Piaget's assumption of a stepwise transition from the non-conserving to the conserving stage, with individual differences only occurring with respect to the moment of transition. Hence, at any moment of time, children with a high level of overall performance (i.e., children who are faster in understanding conservation) are expected to perform at least as good as children with a lower level of overall performance (i.e.,children who are slower in understanding conservation).

The assumption of double monotonicity is checked for all 101 children simultaneously. We opt for a regression with 20 bivariate B-splines of the third degree and second-order smoothness penalties. Regarding the smoothness weights λ_x and λ_y, with X referring to the overall performance level of the children and Y to the 11 different monents in time, values of respectively 0.1 and 44.4 are choosen by making use of Akaike's information criterion with the variance $\widehat{\sigma}$ of the residuals in (6) being estimated on the basis of generalized cross-validation. Both monotonicity weights κ_x

and κ_y are set at either 0 or 10^6 resulting in 4 different analyses: (1) unconstrained regression, (2) monotonicty-constrained regression with respect to overall performance level, (3) monotonicity-constrained regression with respect to time and (4) double monotonicity-constrained regression. The results are graphically displayed as surface plots in Figure 6 and as contour plots in Figure 7. The contour plots clearly reveal whether monotonicity is imposed. Figure 7(a) displays the model without monotonicity restrictions. Indeed, it can be seen that in both directions violations of the rank order as displayed in the color bar occur. Figure 7(b) displays the model with monotonicity imposed on the overall level of performance. In this case, only in the vertical direction, violations of the rank order as displayed in the color bar occur. On the other hand, only violations in the horizontal direction can be seen in Figure 7(c), which displays the model with monotonicity imposed on time. Finally, in Figure 7(d), which displays the model with monotonicity restrictions in both dimensions, no violations occur at all.

Insert Figure 6 about here

Insert Figure 7 about here

For each of the four regressions, a goodness-of-fit measure, that is the squared correlation between observed and predicted scores, is computed. The results are summarized in Table 2. Since no significant discrepancies in fit are observed, the assumption of double monotonicity seems justified.

Insert Table 2 about here

It is interesting to note that monotonicity constrained models as those represented in Figure 6 and 7 can be further explored by computing derivatives of the fitted function. The latter can be easily done making use of expression (12). We illustrate this by computing the first order derivative with respect to time of the results of the P-splines regression with monotonicity restrictions on time [see Figure 6(c) and 7(c)]. Figure 8 displays the surface plot of the computed first order derivative. Evidently, at any point, the first order derivative is non-negative which is as to be expected due to the monotonicity restrictions; values of zero for the first order derivative indicate stagnation in learning, positive values indicate learning with the higher these values the faster the learning. As it is clear from Figure 8, children with a lower overall level of performance learn later and slower as compared to children with a higher overall level of performance.

Insert Figure 8 about here

4 Concluding remarks

To check non-parametric functional forms that can be formalized as local either global constraints on the sign of an n^{th} order derivative, we presented constrained P-splines regression. This is essentially non-parametric regression with additional asymmetric discrete penalties reflecting the assumed functional form. In particular, these penalties restrict the sign of the n^{th} order differences and, as such, the sign of the n^{th} order derivative. In this paper, only applications on monotonicity constrained P-splines are given, but other interesting applications could be considered as well. For instance, consider the assumption of an ideal point of temperature. This means that a particular temperature (e.g., the ideal point) is judged as most pleasant whereas temperatures deviating from the ideal point are judged as less pleasant, with the larger the deviation the less pleasant the temperatures are judged. This assumption can be checked with the help of P-splines regression with two additional asymmetric penalties; a first one which constrains the first order derivative to be positive up to the ideal point, and a second one which constrains the first order derivative to be negative after that ideal point. As a consequence, the ranges of the constraints need to be determined, which can be done either on the basis of prior theoretical considerations, or by making use of model selection techniques. The latter can be achieved by fitting several models with different ranges of the constraints and by subsequently selecting the best model using Akaike's information criterion. As such, constrained P-splines regression may be useful in checking the assumption of a single-peaked functional form.

In general, constrained P-splines regression can be situated in between an exploratory and a confirmatory data-analytic approach, shifting more towards a confirmatory approach when higher values of the penalty weights are chosen. As such, the penalty weights constitute an interesting source of flexibility by which the method of constrained P-splines regression can be easily fine-tuned according to the researcher's purposes. In this regard, a possible interesting approach consists of determining the minimal values for the constraint weights such that the corresponding assumptions are not violated. These values then indicate the extent to which the constraints fit the data, with lower values indicating a better fit. In this regard, it may also be of interest to investigate reference distributions under the null-model assumption that the assumed functional form perfectly holds, using, for instance, a non-parametric bootstrap type of procedure (Efron & Tibshirani, 1993).

Taken together, constrained P-splines regression constitutes a useful method that can be adapted easily in order to optimally investigate a broad range of substantively guided assumptions on functional forms. Constrained P-splines regression is not to be applied blindly but requires choices regarding the weight and the type of constraints, the latter not being restricted to the constraint of monotonicity. As such, constrained P-splines is a more general approach than Integrated B-splines (Winsberg & Ramsay, 1980), which has been developed to deal with monotone transformations of some or all of the variables in a regression model.

References

Coombs, C. H. (1977). Single-peaked functions and the theory of preference. *Psychological Review, 84(2)*, 216–230.

De Boor, C. (1978). *A practical guide to splines.* Berlin: Springer.

Dierckx, P. (1993). *Curve and surface fitting with splines.* Oxford: Clarendon.

Durban, M., Currie, I. D. & Eilers, P. H. C. (2002). Using p-splines to smooth two-dimensional data. In *Proceedings of* 17^{th} *International Workshop on Statistical Modelling.* (pp. 207–214). Crete.

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap.* New-York: Chapman and Hall.

Eilers, P. H. C. (1994). Sign-constrained, monotone and convex nonparametric regression with asymmetric penalties. In *Proceedings of the* 9^{th} *International Workshop on Statistical Modelling.* Exeter.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing using b-splines and penalized likelihood (with comments and rejoinder). *Statistical Science, 11(2)*, 89–121.

Hastie, T. & Tibshirani, R. (1990). *Generalized additive models.* London: Chapman and Hall.

Inhelder, B., Sinclair, H., Bovet, M. & Wedgwood, S. (1974). *Learning and development of cognition.* London: Routledge and Kegan Paul.

Van der Maas, H. L. J. (1993). *Catastrophe anlaysis of stagewise congitive development: Model, method and applications.* Unpublished doctoral dissertation, University of Amsterdam.

Van der Maas, H. L. J. & Molenaar, P. C. M. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review*, *99*, 395–417.

McCrae, R. R. & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, *52*, 81–90.

O'Sullivan, F. (1988). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, *1*, 505–527.

Piaget, J. (1960). *The psychology of intelligence.* Paterson, New Jersey: Littlefield, Adams & Co.

Strauss, S. (1982). *U-shaped behavioral growth.* New York: Academic Press.

Winsberg, S. & Ramsay, J. O. (1980). Monotonic transformations to additivity using splines. *Biometrika*, *67*, 669–674.

4.1 Appendix 1

To show that the loss function given in (14) is convex, we first rewrite (14) in matrix notation:

$$L(\boldsymbol{\alpha}) = (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha})^T(\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) + \lambda(\mathbf{D_2}\boldsymbol{\alpha})^T(\mathbf{D_2}\boldsymbol{\alpha}) + \kappa(\mathbf{D_n}\boldsymbol{\alpha})^T\mathbf{I}\mathbf{W}_{(\alpha)}(\mathbf{D_n}\boldsymbol{\alpha}) \tag{16}$$

or, equivalently,

$$L(\boldsymbol{\alpha}) = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}\mathbf{B}\boldsymbol{\alpha} + \boldsymbol{\alpha}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\alpha} + \lambda\boldsymbol{\alpha}^T\mathbf{D_2}^T\mathbf{D_2}\boldsymbol{\alpha} + \kappa\boldsymbol{\alpha}^T\mathbf{D_n}^T\mathbf{I}\mathbf{W}_{(\alpha)}\mathbf{D_n}\boldsymbol{\alpha} \tag{17}$$

with $b_{ij} = B_j(x_i)$ being the elements of $\mathbf{B}$, $\mathbf{D_2}$ and $\mathbf{D_n}$ being the matrix representations of the difference operators $\triangle^2$ and $\triangle^n$, respectively, $\mathbf{I}$ an $r \times r$ diagonal matrix containing the indicator variables v_j and $\mathbf{W}_{(\alpha)}$ an $r \times r$ diagonal matrix containing the weights w_{α_j}.

The Hessian of the first term at the righthand side of (16) equals $2\mathbf{B^T B}$, which is positive semi-definite; hence, this term is convex in $\boldsymbol{\alpha}$. Similarly, the second term at the righthand side of (16) is convex in $\boldsymbol{\alpha}$ since its Hessian equals $2\lambda\mathbf{D_2^T D_2}$, which is positive semi-definite as well. With respect to the third term, some additional explanation is needed. We first note that:

$$\kappa\boldsymbol{\alpha}^T\mathbf{D_n}^T\mathbf{I}\mathbf{W}_{(\alpha)}\mathbf{D_n}\boldsymbol{\alpha} = \sum_{j=n+1}^{r} \kappa(\mathbf{D_n}\boldsymbol{\alpha})_j^T v_j w_{\alpha_j}(\mathbf{D_n}\boldsymbol{\alpha})_j \tag{18}$$

Then, for $v_j = 0$, $\kappa(\boldsymbol{\alpha}^T\mathbf{D_n}^T)_j v_j w_{\alpha_j}(\mathbf{D_n}\boldsymbol{\alpha})_j = 0$, which is, of course, convex in $\boldsymbol{\alpha}$. For $v_j = 1$, the assignment of the weights w_{α_j} as in (13) implies that $\kappa(\boldsymbol{\alpha}^T\mathbf{D_n}^T)_j v_j w_{\alpha_j}(\mathbf{D_n}\boldsymbol{\alpha})_j$ is a truncated power function of second degree with argument $A = \triangle^n_{\alpha_j}$ of the form

$$\kappa(\boldsymbol{\alpha}^T\mathbf{D_n}^T)_j v_j w_{\alpha_j}(\mathbf{D_n}\boldsymbol{\alpha})_j = \begin{cases} 0 & \text{if } A \geq 0 \text{ (resp. } A \leq 0) \\ \kappa A^2 & \text{otherwise} \end{cases} \tag{19}$$

which is convex in A. The latter expression is convex in $\boldsymbol{\alpha}$ as well since a convex function of a linear combination of elements of a vector is convex in that vector too; that is, if $f(\mathbf{x})$ is a convex function of $\mathbf{x}$, then $g(\mathbf{y}) = f(\mathbf{Ay})$ is a convex function of $\mathbf{y}$.

To prove the latter, assume $0 \leq \theta \leq 1$; then

$$\begin{aligned} g[\theta \mathbf{y_1} + (1-\theta)\mathbf{y_2}] &= f[\mathbf{A}[\theta \mathbf{y_1} + (1-\theta)\mathbf{y_2}] \\ &= f[\theta \mathbf{Ay_1} + (1-\theta)\mathbf{Ay_2}] \\ &\leq \theta f(\mathbf{Ay_1}) + (1-\theta) f(\mathbf{Ay_2}) \\ &= \theta g(\mathbf{y_1}) + (1-\theta) g(\mathbf{y_2}) \end{aligned}$$

making use, in the third step, of the convexity of f.

Furthermore, since a sum of convex functions is convex, (18) is also convex in $\boldsymbol{\alpha}$. Finally, as we showed that all three terms of (16) are convex in $\boldsymbol{\alpha}$, (16) is convex in $\boldsymbol{\alpha}$ as well.

4.2 Appendix 2

To find an optimal solution of (14), we will make use of a Newton-Raphson procedure. Hereby, at each iteration l, $\boldsymbol{\alpha}^{(l+1)}$ is computed such that

$$\mathbf{g}(\boldsymbol{\alpha}^{(l)}) + \mathbf{H}(\boldsymbol{\alpha}^{(l)})(\boldsymbol{\alpha}^{(l+1)} - \boldsymbol{\alpha}^{(l)}) = 0 \tag{20}$$

with $\mathbf{g}$ being the gradient and $\mathbf{H}$ the Hessian of $L(\alpha)$. The gradient of $L(\alpha)$ equals

$$\mathbf{g}(\boldsymbol{\alpha}) = -2\mathbf{B}^T\mathbf{y} + 2[\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}_2^T\mathbf{D}_2 + \kappa\mathbf{D}_n^T\mathbf{I}\mathbf{W}_{(\boldsymbol{\alpha})}\mathbf{D}_n]\boldsymbol{\alpha} + \kappa\boldsymbol{\alpha}^T\mathbf{D_n}^T\mathbf{I}\frac{\partial\mathbf{W}(\alpha)}{\partial\boldsymbol{\alpha}}\mathbf{D_n}\boldsymbol{\alpha} \tag{21}$$

which can be simplified to

$$-2\mathbf{B}^T\mathbf{y} + 2[\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}_2^T\mathbf{D}_2 + \kappa\mathbf{D}_n^T\mathbf{I}\mathbf{W}_{(\boldsymbol{\alpha})}\mathbf{D}_n]\boldsymbol{\alpha} \tag{22}$$

To see this, consider the following two cases for every j; (1) α_j^*, being values for α_j such that $\triangle^n_{\alpha'_j} = 0$ and (2) values for $\alpha_j \neq \alpha_j^*$. Regarding case (1), both the left and the right limit of $g(\boldsymbol{\alpha})$ for α_j going to α_j^* equals (22). Regarding case (2), an infinite small change in α_j will not change the sign of $\triangle^n_{\alpha'_j}$, and as such, will not change the corresponding weight. Hence, $\frac{\partial\mathbf{W}(\alpha)}{\partial\boldsymbol{\alpha}}$ in (21) equals zero and as such, in this case, $g(\boldsymbol{\alpha})$ can be simplified to (22) as well.

Expression (22) shows that the gradient is a piecewise linear function of $\boldsymbol{\alpha}$. This implies that the Hessian is a step function of $\boldsymbol{\alpha}$ with discontinuities at α_j^*. Indeed, for $\alpha_j \neq \alpha_j^*$ the Hessian equals

$$H(\boldsymbol{\alpha}) = 2\mathbf{B}^T\mathbf{B} + 2\lambda\mathbf{D_2}^T\mathbf{D_2} + 2\kappa\mathbf{D_n}^T\mathbf{I}\mathbf{W}_{(\alpha)}\mathbf{D_n} + 2\kappa\mathbf{D_n}^T\mathbf{I}\frac{\partial\mathbf{W}_{(\alpha)}}{\partial\boldsymbol{\alpha}}\mathbf{D_n} \tag{23}$$

As argued earlier, for $\alpha_j \neq \alpha_j^*$, $\frac{\partial\mathbf{W}_{(\alpha)}}{\partial\boldsymbol{\alpha}}$ equals zero and as such, in this case, the Hessian $H(\boldsymbol{\alpha})$ can be simplified to

$$H(\boldsymbol{\alpha}) = 2\mathbf{B}^T\mathbf{B} + 2\lambda\mathbf{D_2}^T\mathbf{D_2} + 2\kappa\mathbf{D_n}^T\mathbf{I}\mathbf{W}_{(\alpha)}\mathbf{D_n} \tag{24}$$

Furthermore, the function values of $H(\boldsymbol{\alpha})$ at α_j^* are uniquely defined as in (24) since the allocation of the weights.

Then, substituting (22) and (24) in (20) yields

$$\begin{aligned} -2\mathbf{B}^T\mathbf{y} + 2[\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}_2^T\mathbf{D}_2 + \kappa\mathbf{D}_n^T\mathbf{I}\mathbf{W}^{(l)}_{(\boldsymbol{\alpha})}\mathbf{D}_n]\boldsymbol{\alpha}^{(l)} & \\ +2[\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D_2}^T\mathbf{D_2} + \kappa\mathbf{D_n}^T\mathbf{I}\mathbf{W}^{(l)}_{(\boldsymbol{\alpha})}\mathbf{D_n}](\boldsymbol{\alpha}^{(l+1)} - \boldsymbol{\alpha}^{(l)}) &= 0 \end{aligned} \tag{25}$$

With some algebra, (25) can be simplified to

$$-\mathbf{B^T}\mathbf{y} + [\mathbf{B^T}\mathbf{B} + \lambda\mathbf{D_2^T}\mathbf{D_2} + \kappa\mathbf{D_n}^T\mathbf{I}\mathbf{W}^{(l)}_{(\boldsymbol{\alpha})}\mathbf{D_n}]\boldsymbol{\alpha}^{(l+1)} = 0 \tag{26}$$

and, hence,

$$\boldsymbol{\alpha}^{(l+1)} = (\mathbf{B^T}\mathbf{B} + \lambda\mathbf{D_2^T}\mathbf{D_2} + \kappa\mathbf{D_n}^T\mathbf{I}\mathbf{W}^{(l)}_{(\boldsymbol{\alpha})}\mathbf{D_n})^{-1}\mathbf{B^T}\mathbf{y}; \tag{27}$$

the first two terms at the righthand side have the same form as the normal equations for a least squares linear model, whereas the third term has the same form as the normal equations for a weigthed least squares linear model, except that it has to be solved iteratively since $\mathbf{W}$ depends on $\boldsymbol{\alpha}$.

FIGURE CAPTIONS

Figure 1. Single B-splines of first and third degree.

Figure 2. Spline regression with B-splines of third degree.

Figure 3. Bivariate B-splines of third degree.

Figure 4. Three different types of items of the liquid conservation test. (a) Standard equality item. (b) Standard inequality item. (c) Guess item. The top two containers constitute the initial situation, the bottom three the resulting situation.

Figure 5. Results of unconstrained (dotted lines) and monotonicity constrained (full lines) P-splines regression applied to performance on liquid conservation across sessions for 4 children with different overall performance levels (child 1 : 3.1, child 2 : 4.8, child 3 : 5.9 and child 4 : 7.5).

Figure 6. Surface plots of P-splines regression applied to performance of liquid conservation with 'sessions' and 'overall level of performance' as predictor variables. (a) unconstrained regression, (b) monotonicity-constrained regression with respect to overall level of performance, (c) monotonicity-constrained regression with respect to time and (d) double monotonicity-constrained regression. Each individual child's overall level of performance is indicated with a black horizontal line.

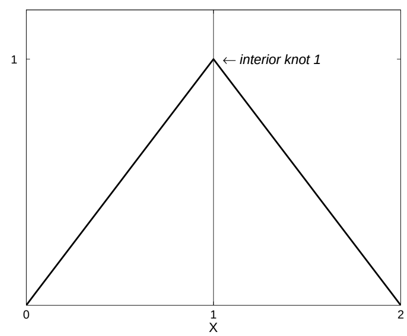
Figure 7. Contour plots of P-splines regression applied to performance of liquid conservation with 'sessions' and 'overall level of performance' as predictor variables. (a) unconstrained regression, (b) monotonicity-constrained regression with respect to overall level of performance, (c) monotonicity-constrained regression with respect to time and (d) double monotonicity-constrained regression. At the right side of each figure, a color bar indicating the level of performance of liquid conservation is displayed.

Figure 8. Surface plot of first order derivative with respect to time of the results of P-splines regression with monotonicity restrictions on time.

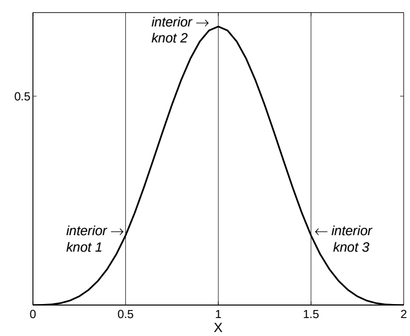
TABLE CAPTIONS

Table 1. Squared correlations per child for unconstraineded ($\kappa = 0$) and constrained regression ($\kappa = 10^6$).

Table 2. Squared correlations for 2×2 different bivariate regression models.

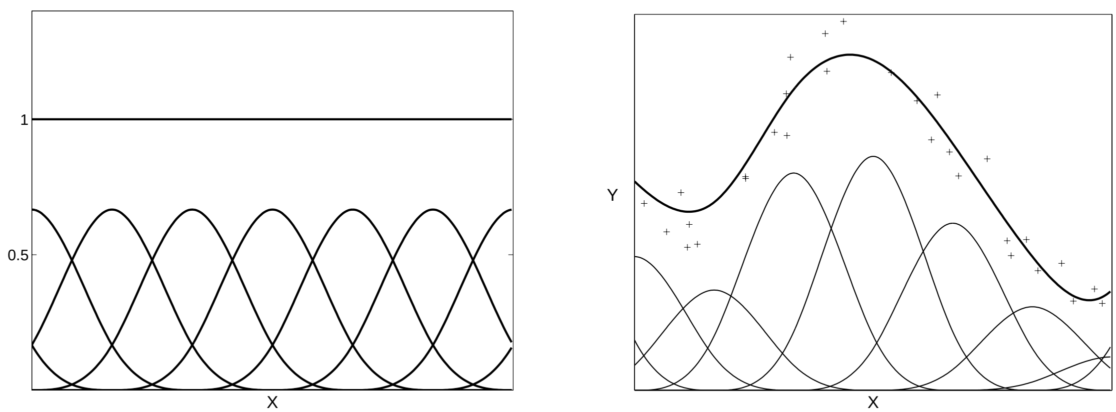


(a) first degree



(b) second degree

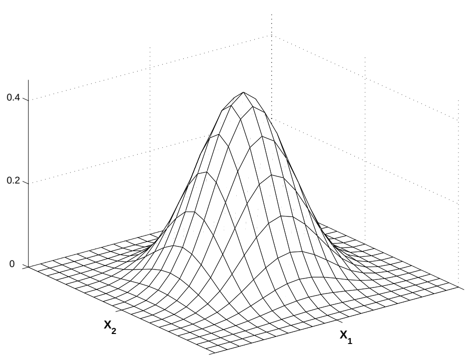
Figure 1:



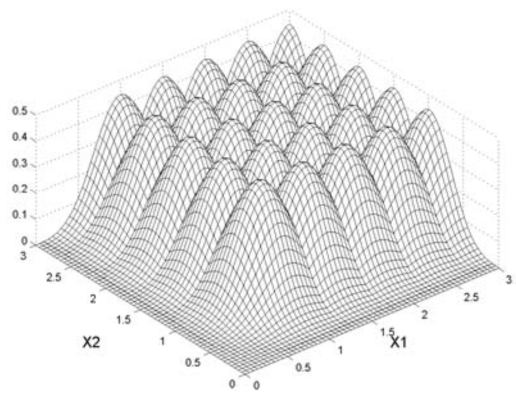
(a) B-splines basis

(b) Splines regression

Figure 2:



(a) Single P-spline



(b) Basis of B-splines

Figure 3:

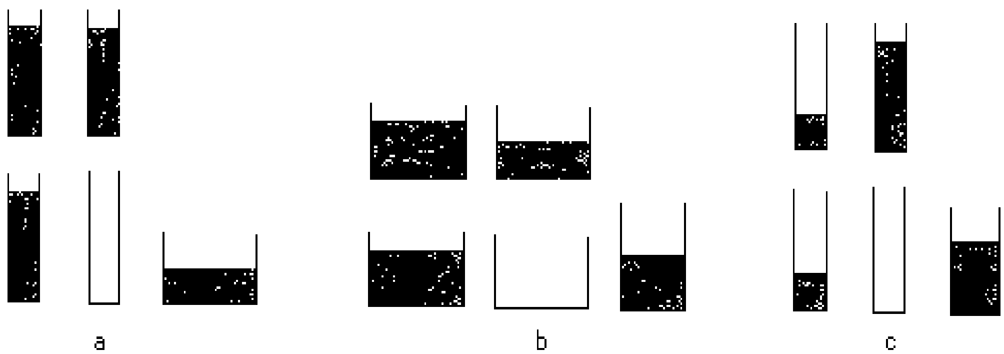
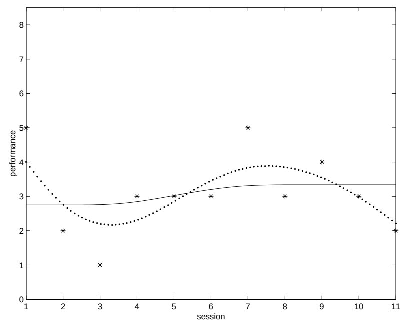
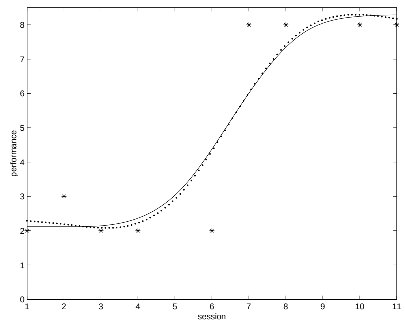


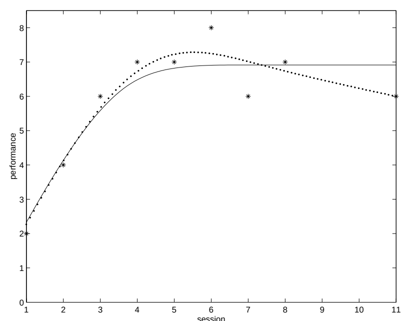
Figure 4:



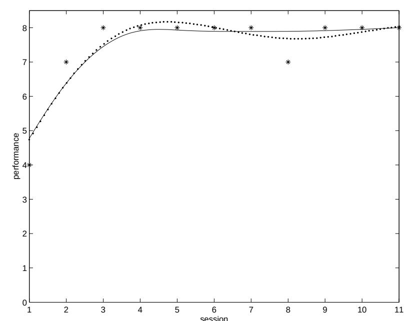
(a) child 1



(b) child 2

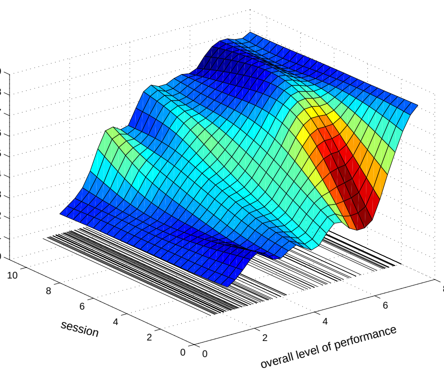


(c) child 3

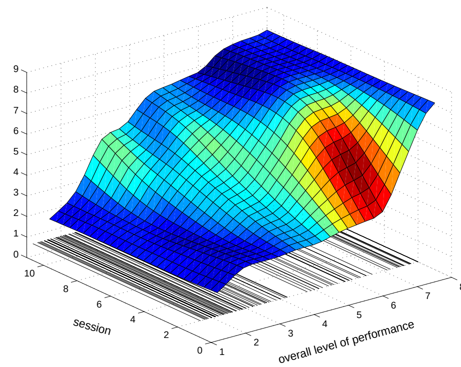


(d) child 4

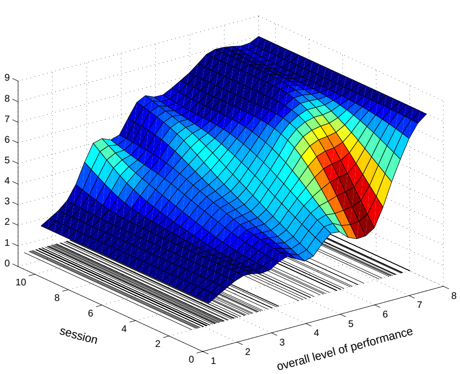
Figure 5:



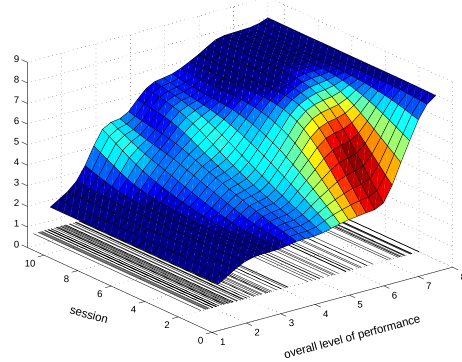
(a) unconstrained



(b) monotonicity of overall performance level

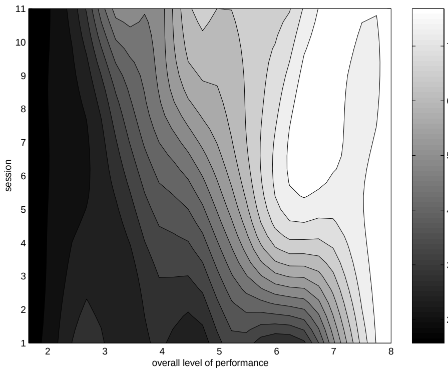


(c) monotonicity of time

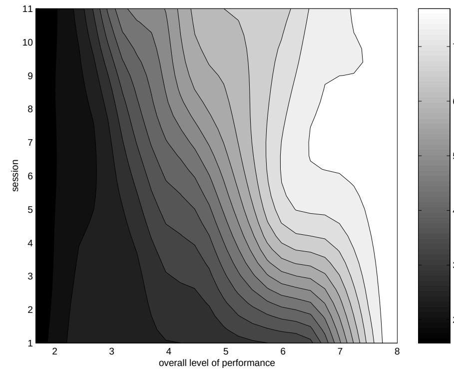


(d) double monotonicity

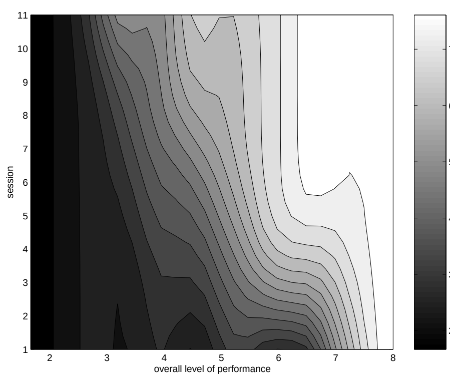
Figure 6:



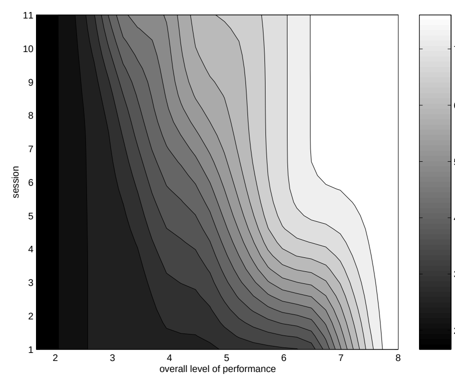
(a) unconstrained



(b) monotoncity of overall performance



(c) monotonicity of time



(d) double monotonicity

Figure 7:

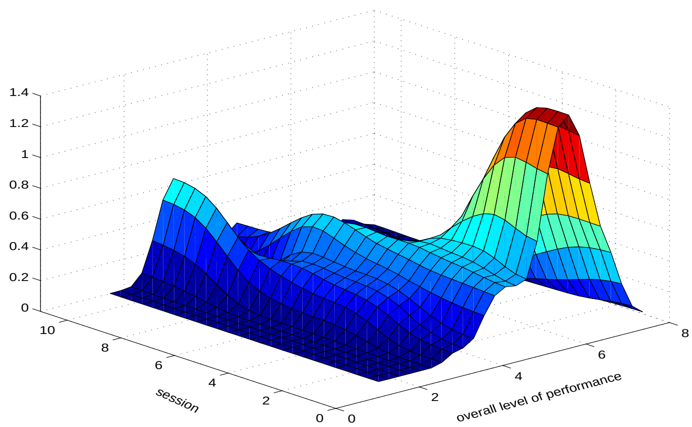

1.4
1.2
1
0.8
0.6
0.4
0.2
0
10
8
6
4
2
0
session
0
2
4
6
8
overall level of performance

Figure 8:

	child			
κ	1	2	3	4
0	0.66	0.86	0.93	0.89
10^6	0.055	0.855	0.875	0.87

Table 1:

	contraints on X	
constraints on Y	no	*yes*
no	0.79	0.78
yes	0.78	0.78

Table 2: