Research Article

Esther Herberich, Christine Hassler and Torsten Hothorn*

Multiple Curve Comparisons with an Application to the Formation of the Dorsal Funiculus of Mutant Mice

Abstract: Much biological experimental data are represented as curves, including measurements of growth, hormone, or enzyme levels, and physical structures. Here we consider the multiple testing problem of comparing two or more nonlinear curves. We model smooth curves of unknown form nonparametrically using penalized splines. We use random effects to model subject-specific deviations from the group-level curve. We present an approach that allows examination of overall differences between the curves of multiple groups and detection of sections in which the curves differ. Adjusted p-values for each single comparison can be obtained by exploiting the connection between semiparametric mixed models and linear mixed models and employing an approach for multiple testing in general parametric models. In simulations, we show that the probability of false-positive findings of differences between any two curves in at least one position can be controlled by a pre-specified error level. We apply our method to compare curves describing the form of the mouse dorsal funiculus – a morphological curved structure in the spinal cord – in mice wild-type for the gene encoding EphA4 or heterozygous with one of two mutations in the gene.

Keywords: equality of functions, growth curve, mixed model, multiple comparisons, semiparametric regression

DOI 10.1515/ijb-2013-0003

1 Introduction

In biology, experimental data are often presented as curves, e.g. growth curves [1], hormone level profiles [2], drug concentration profiles [3], antigen trajectories [4], and viral load profiles [5–7]. Other curves can be those formed by a physical structure, such as the dorsal funiculus, the white substance of the spinal cord that forms a characteristic nonlinear curve over the length of the spinal cord. Reduction of the dorsal funiculus and a modified shape of its curve along the length of the spinal cord are observed in mice carrying mutations in ephrin type-A receptor 4 (EphA4) in which different domains of the EphA4 protein are knocked out.

Such biological curves, where we exclusively mean a smooth function, obtained under different conditions have been compared using various approaches. Zhang et al. [2] used nonparametric functions to model smooth time effects on hormone data and proposed a scaled χ^2-test statistic based on the fitted group-level curves to examine the overall difference between the curves of two groups. The procedure was extended by Kong and Yan [8] to the overall comparison of more than two groups. The authors suggest that after detection of an overall difference between any curves, groups should be compared pairwise with multiplicity adjustment using the Bonferroni method. Behseta and Chenouri [9] modeled smooth intensity functions of groups of neurons using Bayesian adaptive regression splines. To compare the curves obtained using two different experimental conditions, they developed a parametric approach using a modified Hotelling T^2 statistic and a nonparametric approach based on a signed-rank test statistic. For comparisons

***Corresponding author: Torsten Hothorn,** Institut für Sozial- und Präventivmedizin, Universität Zürich, Zürich, Switzerland, E-mail: Torsten.Hothorn@R-project.org

Esther Herberich, Institut für Statistik, LMU München, München, Germany, E-mail: Esther.Herberich@stat.uni-muenchen.de

Christine Hassler, Max-Planck-Institut für Neurobiologie, Martinsried, Germany, E-mail: Hassler@neuro.mpg.de

of curves described by a parametric model tests can be based on the model parameters [10]. However, none of these approaches provide information on the positions at which the curves differ if an overall difference exists.

A parametric model describing the form of the dorsal funiculus along the spinal cord does not exist. Smooth curves of unknown form can be nonparametrically modeled using penalized splines, and within-subject correlation arising from repeated measurements on the same subject can be accounted for by subject-specific random effects [11, 12].

In this paper, we describe a method for multiple comparisons of nonlinear curves that allows the assessment of the positions of the curves which differ. We used this approach to study the biological function of knocked-out domains of EphA4 on the formation of the mouse dorsal funiculus and compared the dorsal funiculus curves of two EphA4 mutants and the EphA4 wild-type. We aimed at detecting the overall difference between each set of two curves and at identifying at which positions the curves of the dorsal funiculus of the EphA4 mutants differ from that of the wild-type, i.e. which regions of the spinal cord are sensitive to the lack of certain EphA4 domains.

We refer to this testing, in which several group-specific curves are compared two at a time along the length of the curves on a grid, as "multiple curve comparisons". These comparisons can be Dunnett-type comparisons, where the curve of a control group is compared to the curves of several other groups; Tukey-type comparisons, where all possible pairs of groups are compared; or any other kind of multiple comparisons. Pairwise comparisons of several curves on a grid along the length of the curves result in multiple testing, with the total number of tests equal to the number of pairwise comparisons multiplied by the number of positions at which the curves are compared. Multiplicity adjustment is therefore required to prevent an increase of the probability of false-positive findings above the nominal level α.

Our multiple curve comparisons combine two frameworks, each implemented in standard statistical software. The first framework exploits the connection between semiparametric mixed models and linear mixed models [13]. Smooth curves of unknown form for several groups are nonparametrically modeled using penalized splines to describe a smooth curve for each group. Random effects are used to model the subject-specific deviation from the group-level curve, leading to a semiparametric mixed model. Asymptotic normal parameter estimates can be obtained by first representing the semiparametric mixed model as a linear mixed model and then using best linear unbiased prediction (BLUP). The second framework allows for simultaneous inference in general parametric models [14]. For multiple curve comparisons, multiple contrasts of parameters from the linear mixed model are built such that each contrast represents the difference of two curves at a particular position over the curve, with the set of contrasts defining all necessary single comparisons. Adjusted p-values for each single comparison can be calculated based on the asymptotic normality of the estimated contrasts following Hothorn et al. [14].

In Section 2, we describe our proposed semiparametric mixed model and how to obtain parameter estimates from the linear mixed model representation of our model. We specify the hypotheses of interest in Section 3 and obtain the asymptotic distribution of parameter estimates on which the calculation of adjusted p-values is based. We demonstrate the performance of our proposed method in a simulation study in Section 4. In Section 5, we apply the method to compare curves describing the shape of the mouse dorsal funiculus in two EphA4 mutants and the wild-type and to detect the regions of the spinal cord affected by the lack of certain EphA4 domains. Section 6 provides details on how the method can be applied using the R [15] add-on packages **mgcv** [16] and **multcomp** [14].

2 Statistical model and estimation

Let K be the number of genotype groups (in our application: wild-type genotype and two mutant genotypes) with $N(k)$ mice in group k, $k = 1, \dots, K$. For the ith animal in the kth group, we have measurements y_{jik} (standardized width of the dorsal funiculus) taken at positions x_{jik}, $j = 1, \dots, J(ik)$, which are equally spaced along the lumbar region of the spinal cord (see Figure 6). The measurements of y_{jik} sum up to $N = \sum_{k=1}^{K} \sum_{i=1}^{N(k)} J(ik)$ observations in total. We assume that for each genotype k, the width of the dorsal funiculus follows a smooth, unknown function $f_k(x)$ along the length of the spinal cord. We specify a semiparametric mixed model

$$y_{jik} = f_k(x_{jik}) + \alpha_{ik} + \varepsilon_{jik}, \tag{1}$$

where the curve of the ith animal in the kth genotype group is shifted from the group-level effect function f_k by a random, animal-specific value α_{ik}. The homoscedastic errors $\varepsilon_{jik} \sim N(0, \sigma_\varepsilon^2)$ are normal at each position x_{jik} along the spinal cord.

We approximate the smooth functions $f_k(x)$ by a spline, i.e. a linear combination of L basis functions $B_l : \mathbb{R} \to \mathbb{R}^+$. The model now reads

$$y_{jik} = \sum_{l=1}^{L} B_l(x_{jik})\beta_{kl} + \alpha_{ik} + \varepsilon_{jik},$$

or in matrix notation

$$\boldsymbol{y} = \boldsymbol{B\beta} + \boldsymbol{\alpha} + \boldsymbol{\varepsilon}. \tag{2}$$

The response vector $\boldsymbol{y} = (y_{jik}) \in \mathbb{R}^{N \times 1}$ contains the dorsal funiculus measurements of all animals at all positions, and the matrix

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}^* & & \\ & \ddots & \\ & & \boldsymbol{B}^* \end{pmatrix} \in \mathbb{R}^{N \times (KL)}$$

is a block-diagonal B-spline design matrix with block matrices

$$\boldsymbol{B}^* = \begin{pmatrix} B_1(x_{11k}) & \cdots & B_L(x_{11k}) \\ \vdots & \vdots & \vdots \\ B_1(x_{J(N(k),k),N(k),k}) & \cdots & B_L(x_{J(N(k),k),N(k),k}) \end{pmatrix} \in \mathbb{R}^{\left(\sum_{i=1}^{N(k)} J(ik)\right) \times L}.$$

The vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{KL \times 1}$ with $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kL}) \in \mathbb{R}^{L \times 1}$ contains the spline effects such that

$$\sum_{l=1}^{L} B_l(x)\beta_{kl} \approx f_k(x);$$

the vector $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{11}, \ldots, \boldsymbol{\alpha}_{N(k)K}) \in \mathbb{R}^{N \times 1}$ with $\boldsymbol{\alpha}_{ik} = (\alpha_{ik}) \in \mathbb{R}^{J(ik) \times 1}$ contains the random, animal-specific deviations from the group-level curve for all animals; and the vector $\boldsymbol{\varepsilon} = (\varepsilon_{jik}) \in \mathbb{R}^{N \times 1}$ contains the normal, homoscedastic errors of all measurements.

Smoothness of the functions is ensured by introducing a penalty on the spline coefficients $\boldsymbol{\beta}$, which leads to the penalized least-squares criterion

$$\underset{\boldsymbol{\beta},\boldsymbol{\alpha}}{\operatorname{argmin}} \left(||\boldsymbol{y} - (\boldsymbol{B\beta} + \boldsymbol{\alpha})||^2 + \sum_{k=1}^{K} \lambda_k \boldsymbol{\beta}^\top \boldsymbol{P}_k \boldsymbol{\beta} + \lambda_{K+1} \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \right), \tag{3}$$

where

$$\boldsymbol{P}_k = \begin{pmatrix} \mathbf{0} & & & & \\ & \ddots & & & \\ & & \boldsymbol{P} & & \\ & & & \ddots & \\ & & & & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{KL \times KL}$$

are block-diagonal penalty matrices with the kth block equal to $\boldsymbol{P} = \boldsymbol{K}^{\mathrm{T}}\boldsymbol{K} \quad \forall k = 1, \ldots, K$, and $\boldsymbol{K} \in \mathbb{R}^{(L-2)\times L}$ is the second-order differences matrix [17].

We now reparameterize the semiparametric mixed model following Fahrmeir et al. [13] by decomposing the spline coefficients $\boldsymbol{\beta}_k$ of each smooth function f_k into an unpenalized part and a penalized part. The decomposition

$$\boldsymbol{\beta}_k = \boldsymbol{U}^{*}\boldsymbol{y}_k + \boldsymbol{V}^{*}\boldsymbol{\delta}_k$$

with unpenalized coefficients $\boldsymbol{y}_k$ and penalized coefficients $\boldsymbol{\delta}_k$ can be defined by

$$\boldsymbol{U}^{*} = \begin{pmatrix} 1 & \kappa_1 \\ \vdots & \vdots \\ 1 & \kappa_L \end{pmatrix},$$

where $\kappa_1, \ldots, \kappa_L$ are the B-spline knots and

$$\boldsymbol{V}^{*} = \boldsymbol{K}^{\top}(\boldsymbol{K}\boldsymbol{K}^{\top})^{-1,}$$

where $\boldsymbol{K}$ is the second-order differences matrix [13]. Model (2) can be reformulated as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{B\beta} + \boldsymbol{\alpha} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{B}(\boldsymbol{Uy} + \boldsymbol{V\delta}) + \boldsymbol{I}_N\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \\ &= \underbrace{\boldsymbol{BU}}_{:=\boldsymbol{X}}\boldsymbol{y} + \underbrace{(\boldsymbol{BV}\,|\,\boldsymbol{I}_N)}_{:=\boldsymbol{Z}}(\boldsymbol{\delta}, \boldsymbol{\alpha}) + \boldsymbol{\varepsilon} \qquad (4) \\ &= \boldsymbol{Xy} + \boldsymbol{Z\xi} + \boldsymbol{\varepsilon} \end{aligned}$$

with identity matrix $\mathbf{I}_N \in \mathbb{R}^{N\times N}$; block-diagonal matrices $\boldsymbol{U} \in \mathbb{R}^{KL\times 2K}$ and $\boldsymbol{V} \in \mathbb{R}^{KL\times K(L-2)}$ where $\boldsymbol{U}^{*}$ and $\boldsymbol{V}^{*}$ are the block entries; $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K)$; and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{K+1})$ with entries $\boldsymbol{\xi}_k = \boldsymbol{\delta}_k, k = 1, \ldots, K$, and $\boldsymbol{\xi}_{K+1} = \boldsymbol{\alpha}$.

The penalized least-squares criterion (3) then becomes

$$\underset{\boldsymbol{y},\boldsymbol{\xi}}{\operatorname{argmin}}\left(||\boldsymbol{y} - (\boldsymbol{Xy} + \boldsymbol{Z\xi})||^2 + \sum_{k=1}^{K+1} \lambda_k \boldsymbol{\xi}_k^{\top}\boldsymbol{\xi}_k \right).$$

According to Ruppert et al. [18] the solution of this minimization problem is equivalent to the BLUP estimation of $\boldsymbol{y}$ and $\boldsymbol{\xi}$ in the linear mixed model representation (4) with fixed effects $\boldsymbol{y}$; random effects $\boldsymbol{\xi} \sim N(\mathbf{0}, \operatorname{diag}(\sigma^2_{\xi_1}\boldsymbol{I}_L, \ldots, \sigma^2_{\xi_K}\boldsymbol{I}_L, \sigma^2_{\xi_{K+1}}\boldsymbol{I}_N))$ with fixed variances $\sigma^2_{\xi_k} = \sigma^2_{\varepsilon}/\lambda_k$, $k = 1, \ldots, K+1$; and errors $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2_{\varepsilon}\boldsymbol{I}_N)$ for given $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{K+1})$. Estimates of $\boldsymbol{\beta}$ can then be obtained via BLUP estimation in the linear mixed model (4), and the smoothing parameters λ_k can be chosen as estimates of the variance ratios $\sigma^2_{\varepsilon}/\sigma^2_{\xi_k}$ obtained via ML or REML methodology.

For multiple tests of hypotheses on linear combinations of the parameters of a linear mixed model, the simultaneous inference procedure of Hothorn et al. [14] can be used. The application of the method for multiple comparisons of curves fitted by model (1) is described in the following section.

3 Multiple curve comparisons

We are looking at M pairwise comparisons of group-level curves, where two genotype groups k and k' are compared in the mth hypothesis

$$H_0^m : \sup_{x\in\mathbb{R}} |f_k(x) - f_{k'}(x)| = 0, \quad 1 \le k < k' \le K, \quad m = 1, \ldots, M.$$

We approximate these hypotheses by comparing the associated splines on a grid $\{x^1, \ldots, x^S\}$

$$H_0^{m,x} : (B_1(x), \ldots, B_L(x))(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k'}) = 0 \quad \forall x \in \{x^1, \ldots, x^S\}, \quad m = 1, \ldots, M,$$

with the grid values being the positions of the measurements. These hypotheses can be reformulated to

$$H_0^{m,x} : \boldsymbol{C}_{m,x}\boldsymbol{\beta} = 0 \quad \forall x \in \{x^1, \ldots, x^S\}, \quad m = 1, \ldots, M,$$

using

$$\begin{aligned}(B_1(x), \ldots, B_L(x))(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k'}) = \\ \underbrace{(B_1(x), \ldots, B_L(x))\boldsymbol{D}_m}_{=:\boldsymbol{C}_{m,x} \in \mathbb{R}^{1\times KL}} \boldsymbol{\beta} = \boldsymbol{C}_{m,x}\boldsymbol{\beta},\end{aligned}$$

with $\boldsymbol{D_m} = (\boldsymbol{0}| \overbrace{\boldsymbol{I}_L}^{k\text{th block}} |\boldsymbol{0}| \overbrace{-\boldsymbol{I}_L}^{k'\text{th block}} |\boldsymbol{0}) \in \mathbb{R}^{L\times KL}$ [12].

The hypotheses for the M pairwise comparisons of curves over all positions $x^1, \ldots, x^S$ can then be specified by

$$H_0 : \boldsymbol{C\beta} = 0,$$

with $\boldsymbol{C} \in \mathbb{R}^{MS\times KL}$ denoting the row stack of $\boldsymbol{C}_{m,x}$, $x = x^1, \ldots, x^S$, $m = 1, \ldots, M$.

The BLUP estimates $(\hat{\boldsymbol{y}}, \hat{\boldsymbol{\delta}})$ asymptotically follow a multivariate normal distribution

$$\sqrt{n}\Big((\hat{\boldsymbol{y}}, \hat{\boldsymbol{\delta}}) - \mathbb{E}((\hat{\boldsymbol{y}}, \hat{\boldsymbol{\delta}}))\Big) \xrightarrow{d} N(0, \boldsymbol{\Sigma})$$

with $\boldsymbol{\Sigma} = \mathbb{V}((\hat{\boldsymbol{y}}, \hat{\boldsymbol{\delta}}))$[18]. The covariance of $\hat{\boldsymbol{\beta}} = \boldsymbol{U}\hat{\boldsymbol{y}} + \boldsymbol{V}\hat{\boldsymbol{\delta}}$ can be calculated as

$$\mathbb{V}(\hat{\boldsymbol{\beta}}) = \mathbb{V}(\boldsymbol{U}\hat{\boldsymbol{y}} + \boldsymbol{V}\hat{\boldsymbol{\delta}}) = \mathbb{V}(\underbrace{[(\boldsymbol{U}|\boldsymbol{V})]}_{=:\boldsymbol{W}}(\hat{\boldsymbol{y}}, \hat{\boldsymbol{\delta}}))= \boldsymbol{W}\mathbb{V}((\hat{\boldsymbol{y}}, \hat{\boldsymbol{\delta}}))\boldsymbol{W}^\top = \boldsymbol{W\Sigma W}^\top$$

following Fahrmeir et al. [13]. Therefore, we get

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \underbrace{\mathbb{E}(\hat{\boldsymbol{\beta}})}_{}) \xrightarrow{d} N(0, \boldsymbol{W\Sigma W}^\top),$$

$$\sqrt{n}(\boldsymbol{D}_m\hat{\boldsymbol{\beta}} - \underbrace{\boldsymbol{D}_m\mathbb{E}(\hat{\boldsymbol{\beta}})}_{=0 \text{ under } H_0}) \xrightarrow{d} N\big(0, \boldsymbol{D}_m\boldsymbol{W\Sigma W}^\top\boldsymbol{D}_m^\top\big),$$

and

$$\sqrt{n}\, \boldsymbol{C}\hat{\boldsymbol{\beta}} \xrightarrow{d} N\big(0, \boldsymbol{CW\Sigma W}^\top\boldsymbol{C}^\top\big).$$

The covariance matrix $\boldsymbol{\Sigma}$ of the estimates $(\hat{\boldsymbol{y}}, \hat{\boldsymbol{\delta}})$ can be estimated by a Bayesian posterior covariance matrix $\hat{\mathbb{V}}((\hat{\boldsymbol{y}}, \hat{\boldsymbol{\delta}}))$[19].

With

$$\sqrt{n}\, \hat{\mathbb{V}}((\hat{\boldsymbol{y}}, \hat{\boldsymbol{\delta}})) \xrightarrow{\mathbb{P}} \boldsymbol{\Sigma},$$

we get

$$\sqrt{n}\, \boldsymbol{CW}\hat{\mathbb{V}}((\hat{\boldsymbol{y}}, \hat{\boldsymbol{\delta}}))\boldsymbol{W}^\top\boldsymbol{C}^\top \xrightarrow{\mathbb{P}} \sqrt{n}\, \boldsymbol{CW\Sigma W}^\top\boldsymbol{C}^\top.$$

Adjusted p-values for all hypotheses $H_0^{m,x}$, $x \in \{x^1, \ldots, x^S\}$, $m = 1, \ldots, M$, i.e. comparisons of two curves at all positions, can then be computed based on the distribution

$$\sqrt{n}\, \boldsymbol{C}\hat{\boldsymbol{\beta}} \xrightarrow{d} N\Big(0, \boldsymbol{CW}\hat{\mathbb{V}}((\hat{\boldsymbol{y}}, \hat{\boldsymbol{\delta}}))\boldsymbol{W}^\top\boldsymbol{C}^\top\Big)$$

as described in Hothorn et al. [14].

4 Simulations

We ran simulations to investigate the performance of the presented approach and estimated the size and power of the testing procedure for Dunnett- and Tukey-type comparisons of three curves.

For $N(k)$ subjects in each group k, observations at $J(ik)$ values $x \in [0,1]$ were simulated from the "true" function

$$f(x) = x^{11} \cdot (10 \cdot (1-x))^6 + 10 \cdot (10x)^3 \cdot (1-x)^{10} \tag{5}$$

scaled to the interval $[0,1]$, with a subject-specific error $\alpha_{ik} \sim N(0, 0.004)$ and a random error $\varepsilon_{jik} \sim N(0, 0.004)$ added to each observation:

$$y_{jik} = f(x_{jik}) + \alpha_{ik} + \varepsilon_{jik}. \tag{6}$$

The function f was taken from the simulations in Wood [20] and is displayed in Figure 1 (black curve).

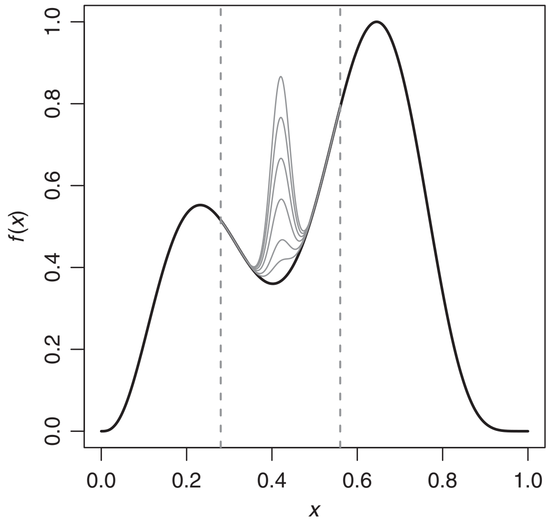

Figure 1 True smooth function used for estimating the size of the testing procedure (black line) and the smooth function of group 3 for varying values of a used for estimating the power of the testing procedure (gray lines)

Three different grid patterns for x were considered:
(a) equally spaced on $[0,1]$,
(b) continuous uniformly distributed on $[0,1]$ with different positions for different subjects,
(c) decreasing density of x (positions at the quantiles of the Beta(1,3) distribution).

We investigated scenarios with 15, 20, or 25 positions and 5, 10, 15, or 20 subjects per group.

To estimate the size of the testing procedure for Dunnett- and Tukey-type comparisons of three curves, we simulated observations from the "null model" (6) for all three groups. We fitted the curves using the semiparametric mixed model (1) and approximated the smooth terms by a linear combination of B-splines basis functions [17]. We compared the fitted curves at each position for settings (a) and (c) and at $N(k)$ equally spaced positions for setting (b). We estimated the size as the portion of 1,000 datasets in which at least one difference was found among all comparisons made, and we used the same datasets for both Dunnett- and Tukey-type comparisons.

Additionally, we examined settings (a), (b), and (c) when the observations following rather small measurements were truncated. In practice, if $\alpha_{ik} + \varepsilon_{jik} < -sd(\alpha_{ik} + \varepsilon_{jik})$ for any measurement at the fourth or higher position, i.e. x_{jik}, $j>3$, the observations on this subject were truncated, i.e. this observation and all following observations from the same subject were missing until the proportion of missing observations was ~25%. The nominal level was chosen as $\alpha = 0.05$.

The estimated size for all simulation scenarios is shown in Figure 2. The results of Dunnett- and Tukey-type comparisons are similar. In setting (b), where the grid points in which the curves were compared did not equal the measurement points, the procedure is liberal. In settings (a) and (c) without truncated observations, the estimated size is close to 0.05 for almost all combinations of measurements per subject and number of subjects per group. In the presence of truncated observations, the procedure is liberal in setting (c), where measurements are taken more frequently at the beginning.

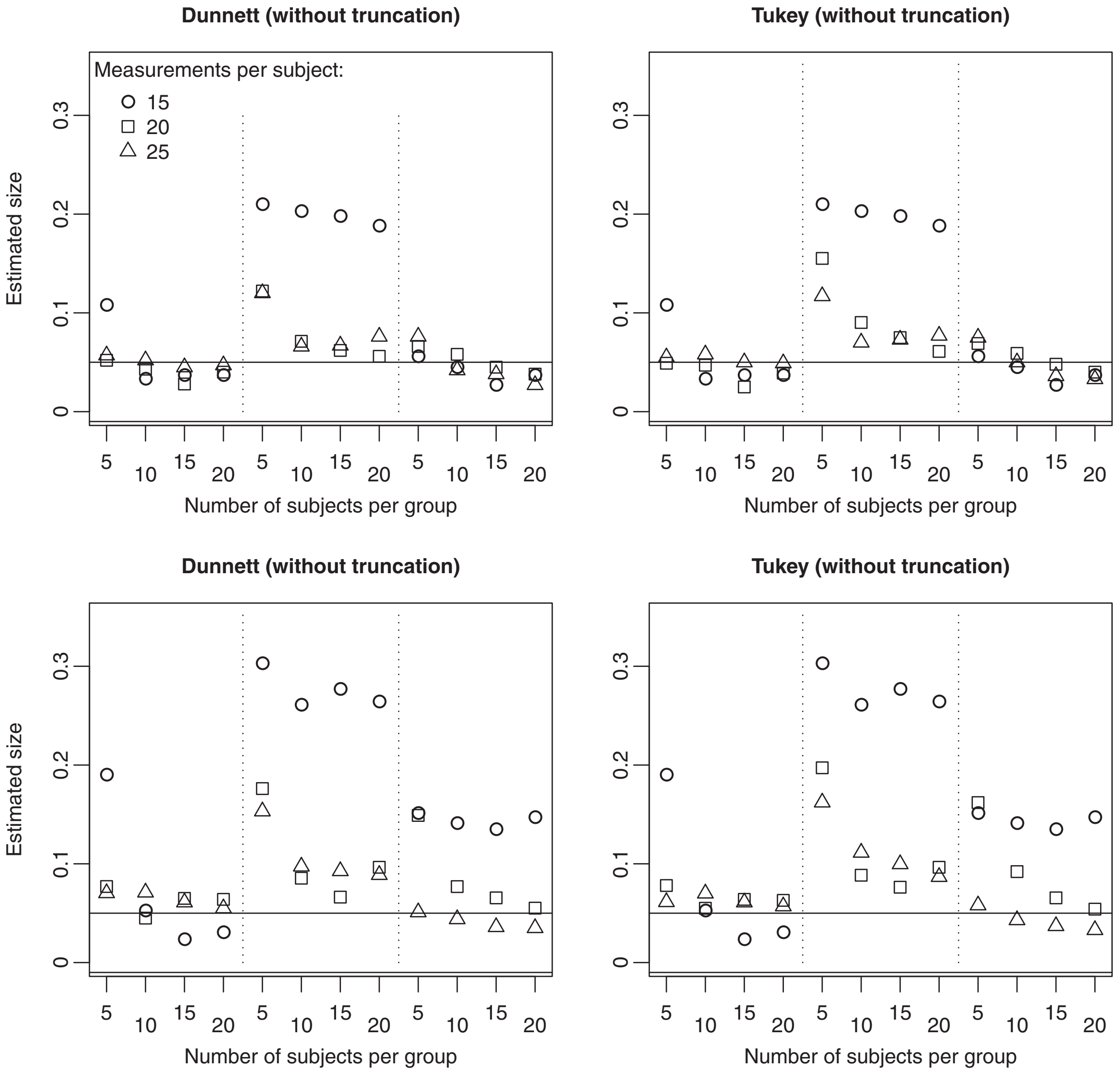

Figure 2 Estimated size of the testing procedure for Dunnett-type (left column) and Tukey-type (right column) comparisons of three curves for setting (a) (left section of each graph), (b) (middle section of each graph), and (c) (right section of each graph) each estimated from 1,000 datasets without truncated observations (top row) and with truncated observations (bottom row)

One may wonder how the number of grid points affects the p-value for a specific dataset. We sampled 100 instances from model (6), performed Tukey-type comparisons, and computed an overall p-value as the minimal adjusted p-value over all differences for $S = 5$, 10, 15, 20, 30, 40, and 50 equally spaced grid points. The p-value trajectories as a function of the number of grid points are given in Figure 3. Except for two instances, where the p-value decreases extremely between $S = 20$ and $S = 50$, the p-values are rather stable starting at $S = 20$ grid points.

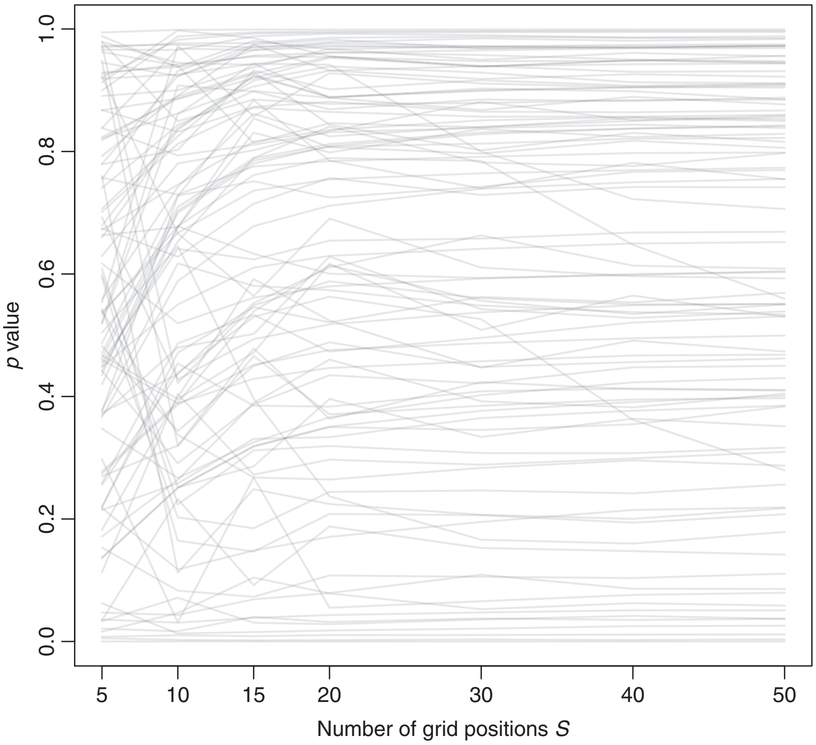

Figure 3 Trajectories of minimal adjusted p-values from Tukey-type comparisons under the null model for increasing number of grid points S. Each line corresponds to one of 100 simulated datasets from model (6)

To investigate the power of the procedure, the observations of group 3 were simulated from a function f_3, which differed from function (5) for x values in the interval $[0.28, 0.56]$:

$$f_3(x) = f(x) + \exp\left(-\frac{(x-0.42)^2}{0.01}\right) \cdot a \cdot I_{[0.28,0.56]}(x), \tag{7}$$

with a ranging from 0.05 to 0.5 in steps of 0.05 to increase the difference between f_3 and f (see Figure 1). For equally spaced positions on $[0, 1]$, the number of positions with values differing between f and f_3 are 3, 4, 5, or 7 for 15, 20, or 25 positions in total.

The power of the procedure was estimated by the portion of 1,000 simulated datasets in which at least one significant difference between two differing functions was found (not necessarily for positions in which the values of the underlying functions truly differed).

The estimated power curves for setting (a), 15 subjects per group, and 15, 20, or 25 positions are shown in Figure 4. The power is slightly higher for curves fitted from measurements taken at fewer positions compatible to the estimated size for 15 subjects per group, where the procedure becomes conservative with increasing number of positions. The power is rather low over a wide range of the parameter a, which controls how the curve of the third group differs from the other curves. A considerable difference in the curves is needed for the procedure to detect a difference.

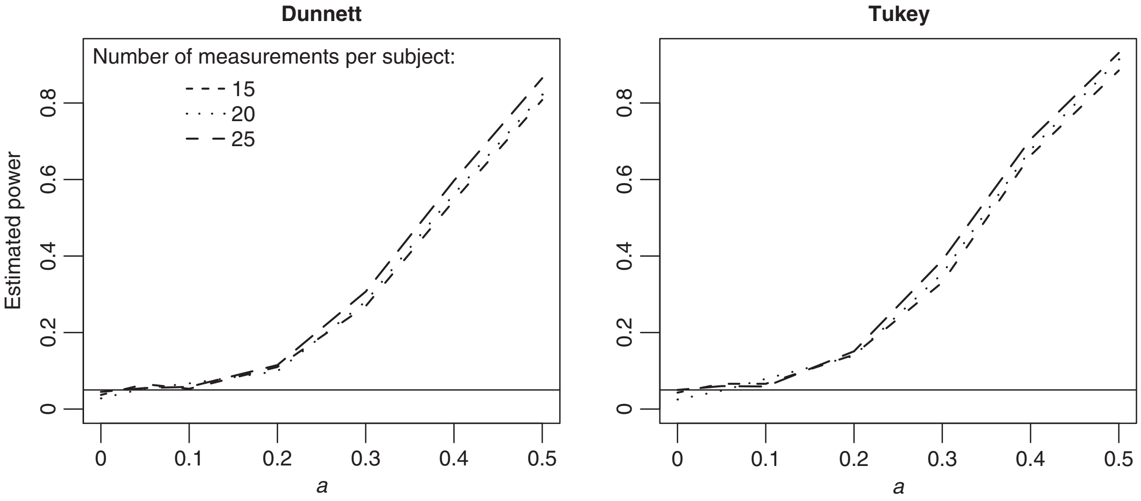

Figure 4 Power of the testing procedure for Dunnett- and Tukey-type comparisons of three curves, where the curve of one group differs from the others according to eq. (7) for varying values of *a*

5 Application: comparisons of the mouse dorsal funiculus of wild-type and EphA4 mutants

The protein EphA4 plays a major role in the development of the central nervous system. The absence of EphA4 leads to neuronal axons not finding their target cell and neural networks not properly connecting. EphA4 is also required for the development of the so-called dorsal funiculus, a morphological structure in the spinal cord comprised major axon bundles. When the EphA4 gene is knocked out or is enzymatically inactive, formation of the dorsal funiculus is impaired [21, 22].

In wild-type mice with a completely conserved EphA4 protein, the width of the dorsal funiculus forms a characteristic nonlinear curve over a subsection of the spinal cord. Neurobiologists of the Max-Planck-Institute in Martinsried, Germany have studied the role of EphA4 mutations in the formation of the dorsal funiculus by comparing the dorsal funiculus curve of a wild-type control group with those of two different groups of EphA4 mutant mice. Our analysis is based on their results. The homozygous wild-type control group had EphA4 completely conserved (genotype EphA4$^{\mathrm{KI/KI}}$), and each of the two heterozygous mutant groups had one wild-type EphA4 allele and one mutant EphA4 allele. In one mutant mouse line (genotype EphA4$^{\mathrm{KD/KI}}$), the mutant allele harbored a point mutation in the encoded tyrosine kinase domain located in the C-terminus of EphA4, which renders EphA4 enzymatically inactive. In the other mutant mouse line (genotype EphA4$^{\mathrm{GFP/KI}}$), the mutant allele encoded a protein lacking the complete cytoplasmic region of the C-terminus, which was replaced by the green fluorescent protein (GFP) (Figure 5).

The standardized width of the dorsal funiculus was measured on 25 cross-sections along the lumbar spinal cord (Figure 6) of five mice with genotype EphA4$^{\mathrm{KI/KI}}$, six mice with genotype EphA4$^{\mathrm{KD/KI}}$, and four mice with genotype EphA4$^{\mathrm{GFP/KI}}$. The measurements from all mice are displayed in Figure 7.

We modeled the curves of each group of mice in a semiparametric mixed model as described in Section 2:

$$y_{jik} = f_k(x_{jik}) + \alpha_{ik} + \varepsilon_{jik}, \tag{8}$$

with $\alpha_{ik} \sim N(0, \sigma_\alpha^2)$ and $\varepsilon_{jik} \sim N(0, \sigma_\varepsilon^2)$ for $K = 3$ groups $k = 1, \ldots, K$, mice $i = 1, \ldots, N(k)$ in the kth group, and $N(ik) = 25$ measurements $j = 1, \ldots, N(ik)$ for each animal. The number of mice $N(k)$ in the kth group are $N(1) = 5$, $N(2) = 6$, and $N(3) = 4$ with $k = 1$ corresponding to genotype EphA4$^{\mathrm{KI/KI}}$, $k = 2$ corresponding to genotype EphA4$^{\mathrm{KD/KI}}$, and $k = 3$ corresponding to genotype EphA4$^{\mathrm{GFP/KI}}$. This leads to $N = 375$ observations y_{jik} in total.

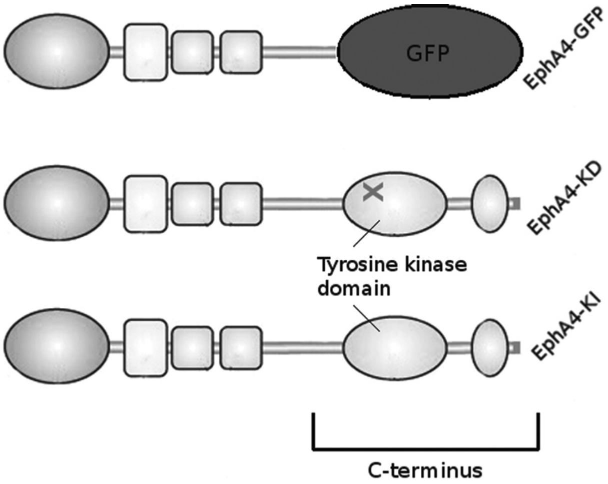

Figure 5 Schematic diagram showing the C-terminus of EphA4 encoded by the wild-type allele KI (bottom) and the mutant alleles KD (middle) and GFP (top)

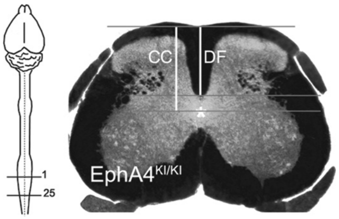

Figure 6 Range of the 25 lumbar spinal cord cross-sections (left) and cross-section of a wild-type mouse (right). The standardized width of the dorsal funiculus is the ratio between the total width of the dorsal funiculus (DF) and the width of the dorsal part of the cord to the level of the central canal (CC)

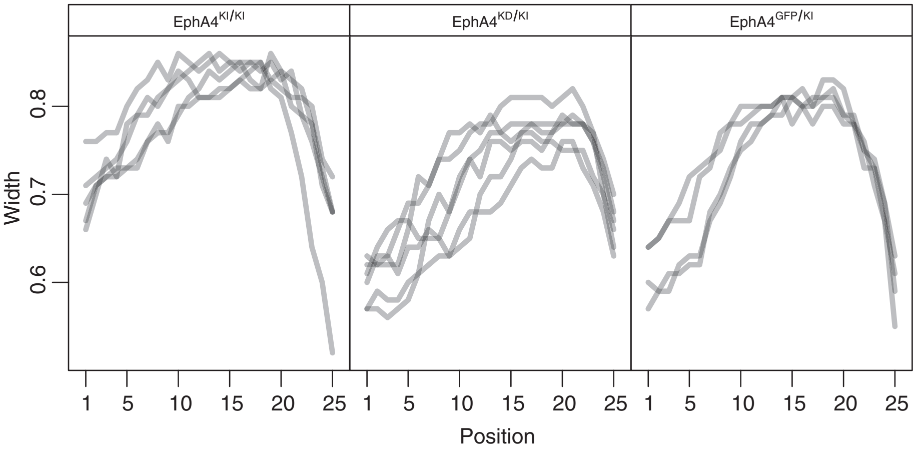

Figure 7 Standardized width of the dorsal funiculus measured at 25 positions along its length in five mice with genotype EphA4$^{KI/KI}$ (left), six mice with genotype EphA4$^{KD/KI}$ (center), and four mice with genotype EphA4$^{GFP/KI}$ (right). Each line corresponds to one mouse

The fitted groupwise curves are shown in Figure 8. All-pairwise comparisons of the three groups were conducted, and each pair of curves was compared at each of the 25 positions. Significant differences were found at positions 1–20 when the wild-type control was compared to the mutant with genotype EphA4$^{KD/KI}$, and at positions 1–9 when the wild-type control was compared to the mutant with genotype EphA4$^{GFP/KI}$.

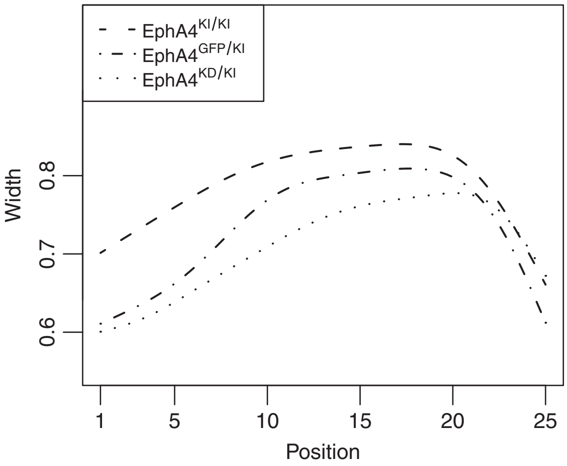

Figure 8 Fitted curves of the width of the dorsal funiculus for the three genotypes

Hence, these results indicated that the kinase domain of the C-terminus is required for the development of the complete dorsal funiculus and that one allele encoding an inactive kinase domain or lacking the kinase domain affects the shape of the dorsal funiculus compared to that in the homozygous wild-type. Specifically, the absence of the EphA4 C-terminus including the kinase domain in the heterozygous EphA4$^{GFP/KI}$ mutant led to a reduction of the dorsal funiculus in the lower positions, and inactivation of the EphA4 kinase domain in the heterozygous EphA4$^{KD/KI}$ mutant led to a reduction of the dorsal funiculus in almost the entire region inspected. Significant differences in the dorsal funiculus of the two heterozygous mutant mouse lines were found in the middle region (positions 9–13), which implied that even though the tyrosine kinase domain is required for the development of the complete dorsal funiculus, other cytoplasmic regions of the C-terminus are involved in the formation as well.

6 Computational details

In this section, we provide details on how multiple curve comparisons of several groups can be performed using the software R [15]. The **multcomp** package provides a general implementation of the framework for simultaneous inference in parametric models according to Hothorn et al. [14]. In the web-based Supplementary material a simulated dataset DorsalFuniculus is available, whose observations were generated according to the structure of the data presented in Section 5. The dataset contains the dependent variable y (width of the dorsal funiculus), the position variabe x, the grouping variable group, and the subject-specific identifier id. The semiparametric mixed models studied in this paper were fitted by BLUP estimation in the linear mixed model representation using the function *gamm()* provided in the package **mgcv** [16]:

```
mod <- gamm(y ~ - 1 + s(x, by = group, bs = "ps") + group,
      random = list(id= ~ 1), data = DorsalFuniculus)$gam
```

Alternatively, one can use the gam function

```
mod <- gam(y ~ - 1 + s(x, by = group, bs = "ps") +
      group + s(id, bs = "re"),
      data = DorsalFuniculus, method = "REML")
```

The latter call also allows shared smoothing parameters for the curves fitted to the different groups. Estimates of the unpenalized and penalized spline coefficients $(\hat{\boldsymbol{y}}, \hat{\boldsymbol{\delta}})$ and the posterior covariance matrix for these parameter estimates can be extracted from the returned object via `coef(mod)` and `vcov(mod)`, but are automatically extracted when multiple comparisons of linear combinations of the model parameters are performed using the tools provided by the package **multcomp**. The function *glht()* takes the fitted model mod and sets up the linear combinations to be tested for the associated contrast matrix `K` specified by the argument `linfct`:

```
glht_mod <- glht(model = mod, linfct = K)
```

The matrix `K` needs to be user defined such that it corresponds to $\boldsymbol{CW}$ in Section 3. `K %*% coef(mod)` then corresponds to $\boldsymbol{C\beta}$ and defines the multiple curve comparisons of interest on a certain grid. The specification of `K` for Dunnett-type or Tukey-type multiple curve comparisons is given in the R Code provided in the web-based Supplementary material. Adjusted p-values for each single comparison are returned via `summary(glht_mod)`. Further details are available in Bretz et al. [23].

7 Discussion

In this paper, we developed a procedure for multiple comparisons of curves fitted by a semiparametric mixed model. Previously existing approaches only perform overall comparisons and do not provide information on the grid points at which the curves differ between two groups. The method we propose allows comparisons of two or more groups over a grid along the length of the curves, with control of the probability of at least one false-positive finding among all comparisons made. Our simulations showed that the overall error level of multiple comparisons of several curves fitted from a reasonable number of observations per subject and per group can be controlled when curves are compared at the positions at which the measurements were taken. Nevertheless, it is possible to use alternative grid points for defining hypotheses. For example, differences in a certain domain of the curves might be not very interesting *a priori*, so a more powerful procedure can be set up by placing the grid points in the domain of interest. As this grid becomes denser, the correlations between the test statistics increase. In terms of multiplicity correction, the price of a dense grid is very small but the computational aspects become more challenging [24].

The procedure is based on the Bayesian posterior covariance matrix and the asymptotic normality of penalized estimates. Therefore, the small-sample performance might be problematic. Our simulations showed considerable size distortions for certain configurations and especially for datasets with a small number of replications per observation. Also, the power of the procedure is somewhat limited, so larger sample sizes are required to actually detect interesting differences. Wood [25] introduced a correction for Wald statistics on penalized functions. A similar correction for the max-type statistics applied to the multiple comparisons studied here putting more emphasis on the less penalized parts of the estimated curves promises to improve the procedure.

When we compared the dorsal funiculus curves of groups of mice with different EphA4 genotypes, we gained information about which region of the dorsal funiculus along the length of the spinal cord is sensitive to the lack of certain EphA4 domains. The proposed method could also be applied to comparisons of growth curves or hormone level profiles, or in pharmacogenetics, when several groups are compared over time.

We limited our attention to position and group as covariates, but the methodology can be extended to multiple comparisons of group-level curves with adjustment for further covariates. In our application of dorsal funiculus formation, measurements were taken at regularly spaced positions in all mice, but the method can also be applied when individual curves are measured at variable and irregularly spaced points. For the dorsal funiculus data, it seems that subject-specific deviations from the group-average curve can be adequately modeled by a parametric random effect since the differences between the measurements of mice belonging to the same group were regular over all positions. To be more flexible in the case of irregular subject-specific deviations, nonparametric functions could be used for both the group-level and the subject-level curves.

The methodology can be extended to multiple comparisons of areas under curves in settings where a parametric model to fit the curves and estimate the areas thereunder does not exist. Using a semiparametric mixed model to fit the curves and applying the trapezoidal rule to estimate the area under the curves, hypotheses on the equality of two or more areas can be set up as linear combinations of the model parameters, and the procedure by Hothorn et al. [14] can be applied to perform multiple comparisons of the areas under the curves.

Supplementary materials

A Web Appendix containing the R code referenced in Section 6 is available.

References

1. Villandr L, Hutcheon JA, Trejo ME, Abenhaim H, Jacobsen G, Platt RW. Modeling fetal weight for gestational age: a comparison of a flexible multi-level spline-based model with other approaches. Int J Biostat 2011;7:article 32.
2. Zhang D, Lin X, Sowers M. Semiparametric regression for periodic longitudinal hormone data from multiple menstrual cycles. Biometrics 2000;56:31–9.
3. Bertrand J, Comets E, Chenel M, Metré F. Some alternatives to asymptotic tests for the analysis of pharmacogenetic data using nonlinear mixed effects models. Biometrics 2012;68:146–55.
4. Bhadra D, Daniels MJ, Kim S, Ghosh M, Mukherjee B. A Bayesian semiparametric approach for incorporating longitudinal information on exposure history for inference in case-control studies. Biometrics 2012;68:361–70.
5. Dagne G, Huang Y. Bayesian inference for a nonlinear mixed-effects Tobit model with multivariate skew-t distributions: application to AIDS studies. Int J Biostat 2012;8:article 27.
6. Huang Y, Chen J, Yan C. Mixed-effects joint models with skew-normal distribution for HIV dynamic response with missing and mismeasured time-varying covariate. Int J Biostat 2012;8:article 34.
7. Huang Y, Chen R, Dagne G. Simultaneous Bayesian inference for linear, nonlinear and semiparametric mixed-effects models with skew-normality and measurement errors in covariates. Int J Biostat 2011;7:article 8.
8. Kong M, Yan J. Modeling and testing treated tumor growth using cubic smoothing splines. Biometrical J 2011; 53:1–19.
9. Behseta S, Chenouri S. Comparison of two populations of curves with an application in neuronal data analysis. Stat Med 2011;30:1441–54.
10. Dasgupta N, Shaffer MJ. Many-to-one comparisons of nonlinear growth curves for Washington’s red delicious apple. J Appl Stat 2012;39:1781–95.
11. Chen H, Wang Y. A penalized spline approach to functional mixed effects model analysis. Biometrics 2011;67:861–70.
12. Thilakarathne PJ, Clement L, Lin D, Shkedy Z, Kasim A, Talloen W, et al. The use of semiparametric mixed models to analyze PamCHIP peptide array data: an application to an oncology experiment. Bioinformatics 2011;27:2859–65.
13. Fahrmeir L, Kneib T, Lang S. Penalized structured additive regression for space-time data: a Bayesian perspective. Stat Sin 2004;14:715–45.
14. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. Biometrical J 2008;50: 346–63.
15. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. Available at: http://www.R-project.org/, ISBN 3-900051-07-0.
16. Wood S. Generalized additive models: an introduction with R. Boca Raton, FL: Chapman and Hall/CRC, 2006.
17. Eilers PH, Marx BD. Flexible smoothing with B-splines and penalties. Stat Sci 1996;11:89–121.
18. Ruppert D, Wand MP, Carroll RJ. Semiparametric regression, Cambridge series in statistical and probabilistic mathematics. Cambridge: Cambridge University Press, 2003.
19. Lin X, Zhang D. Inference in generalized additive mixed models by using smoothing splines. J R Stat Soc Ser B Stat Meth 1999;61:381–400.
20. Wood S. On confidence intervals for generalized additive models based on penalized regression splines. Aust New Zealand J Stat 2006;48:445–64.

21. Egea J, Klein R. Bidirectional eph-ephrin signaling during axon guidance. Trends Cell Biol 2007;17:230–8.
22. Kullander K, Mather NK, Diella F, Dottori M, Boyd AW, Klein R. Kinase-dependent and kinase-independent functions of EphA4 receptors in major axon tract formation in vivo. Neuron 2001;29:73–84.
23. Bretz F, Hothorn T, Westfall P. Multiple comparisons using R. Boca Raton, FL: Chapman & Hall/CRC Press, 2010.
24. Genz A, Bretz F. Computation of multivariate normal and t probabilities, Lecture notes in statistics. Heidelberg: Springer, 2009.
25. Wood SN. On p-values for smooth components of an extended generalized additive model. Biometrika 2013;100:221–8.