A Penalized Spline Estimator For Fixed Effects Panel Data Models

Peter Pütz*, Thomas Kneib†

Abstract

Estimating nonlinear effects of continuous covariates by penalized splines is well established for regressions with cross-sectional data as well as for panel data regressions with random effects. Penalized splines are particularly advantageous since they enable both the estimation of unknown nonlinear covariate effects and inferential statements about these effects. The latter are based, for example, on simultaneous confidence bands that provide a simultaneous uncertainty assessment for the whole estimated functions. In this paper, we consider fixed effects panel data models instead of random effects specifications and develop a first-difference approach for the inclusion of penalized splines in this case. We take the resulting dependence structure into account and adapt the construction of simultaneous confidence bands accordingly. In addition, the penalized spline estimates as well as the confidence bands are also made available for derivatives of the estimated effects which are of considerable interest in many application areas. As an empirical illustration, we analyze the dynamics of life satisfaction over the life span based on data from the German Socio-Economic Panel (SOEP). An open source software implementation of our methods is available in the R package *pamfe*.

Keywords: first-difference estimator; life satisfaction; panel data; penalized splines; simultaneous confidence bands.

1 Introduction

Nonparametric and semiparametric regression methods are extremely popular in statistics and econometrics when studying the impact of one or more continuous covariates on a response variable. Their main advantage is that they do not impose strong prior assumptions on the functional shape of the covariate effects but rather let the data speak for themselves such that a data-driven amount of nonlinearity is identified. In this paper, our interest lies in estimating regression models with flexible covariate effects for panel data. We therefore think of N persons observed at T points in time and consider an additive panel data model of the form

$$y_{it} = \gamma_i + \sum_{h=1}^{p} f_h(x_{hit}) + u_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T,$$

where y_{it} is the response variable of interest, $f_1(x_{1it}), \ldots, f_p(x_{pit})$ represent the nonlinear effects of p continuous covariates, u_{it} are independent and identically distributed normal error terms with constant variance and γ_i are individual-specific, time-invariant effects either allowed (fixed effects model) or not allowed (random effects model) to be correlated with the covariates. For the specification of the nonlinear effects, we rely on penalized B-splines (Eilers and Marx, 1996) which approximate a nonlinear effect of interest by a rich B-spline basis while adding a penalty to the penalized least squares criterion to regularize estimation. In addition to their computational attractiveness, penalized splines are also easily combined with parametric effects to obtain partially nonlinear models and allow for easy access to uncertainty measures.

So far, penalized splines have mostly been used for either cross-sectional data or in combination with random effects specifications for panel data. The main reason for this is the fact that the penalty considered for penalized splines fits nicely together with the "penalty" imposed by the random effects and in fact penalized splines can be considered a special type of random effects model as well, see for example Ruppert and Wand (2003) or Fahrmeir et al. (2013). However, when utilizing a random effects specification for panel data, one has to critically evaluate whether correlations between the random effects and the regression covariates are present. Fixed effects specifications loosen this crucial assumption and are particularly popular in econometrics. To avoid the incidental parameter

*Georg-August-Universität Göttingen, Centre for Statistics, e-mail: ppuetz@uni-goettingen.de.

†Georg-August-Universität Göttingen, Faculty of Economic Sciences, Chairs of Statistics and Econometrics.

An online supplement and the R package *pamfe* can be found at `https://www.uni-goettingen.de/de//511092.html`.

problem that arises when including fixed effects, estimation is then typically based on first order differenced or demeaned data. For nonparametric and semiparametric panel data models with fixed effects, a growing strand of literature has emerged during the last years, including Baltagi and Li (2002), Su and Ullah (2006), Henderson et al. (2008) and Mammen et al. (2009). Extensive literature reviews are provided by Su and Ullah (2011) and Chen et al. (2013). While having different concepts to handle the fixed effects and strictly parametric effects, all discussed methods have in common that they rely on some kind of kernel estimator to estimate the nonparametric model components. This makes inference on the nonlinear effects challenging or at least computationally demanding in cases of large sample sizes and many nonparametric model components since uncertainty assessments for kernel estimators are typically based on bootstrapping techniques (Claeskens and Van Keilegom, 2003; Li et al., 2013).

To overcome this difficulty, we consider a penalized spline specification for the nonlinear model components and apply first order differences to the model. This basically implies a differenced basis function approximation of the nonparametric effects while relying on the same parameterization of the penalized spline as the original model. To account for the serial correlation induced by first differencing, we use a generalized least squares (GLS) criterion. Utilizing the mixed model representation of penalized splines, we develop a fast way of inference for first-difference penalized spline estimates via simultaneous confidence bands building on the ideas of Wiesenfarth et al. (2012) for cross-sectional data. This also allows us to derive simultaneous confidence bands for the derivatives of the nonlinear effects.

In terms of the model specification, our approach is closely related to Hajargasht (2009) who also proposed a penalized spline estimator for fixed effects panel data, based on the within-transformation, i.e., demeaned data. However, our approach differs from the one by Hajargasht (2009) with respect to the following important aspects: (i) we use the mixed model representation of penalized splines not only to obtain a data-driven estimate for the smoothing parameter but also simultaneous confidence bands, (ii) we develop and investigate inferences for the nonlinear effects directly and for the derivatives, and (iii) we provide an open source implementation in the accompanying R package *pamfe* that enables practitioners to apply the proposed method which is capable to handle partially linear models and models with multiple nonlinear components.

To illustrate the applicability of our methods, we use the information from the German Socio-Economic Panel (SOEP) database[1] on the dynamics of life satisfaction over the life span. So far, there has not been reached a consensus on the functional form of the relationship between age and life satisfaction. Typically, it is modeled via a strictly parametric specification, which might be too restrictive and is therefore likely to affect the results adversely. Our more flexible approach avoids this issue and also accounts for individual heterogeneity among the survey respondents by including fixed effects.

The remainder of this paper is organized as follows: First-difference penalized spline estimation for panel data models is introduced in Section 2. Inference via simultaneous confidence bands is considered in Section 3. In Section 4, the performance of our approach is tested in a simulation study while the empirical investigation of the dynamics of life satisfaction is described in Section 5. Section 6 summarizes our conclusions and discusses directions for future research.

2 Penalized splines for cross-sectional and panel data

2.1 Penalized splines in the cross-sectional context

We start our considerations by discussing penalized spline specifications for cross-sectional data. Consider the additive regression model

$$y_i = \beta_0 + \sum_{h=1}^{p} f_h\left(x_{hi}\right) + u_i,\ u_i \sim N\left(0, \sigma_u^2\right),\ i = 1, \ldots, n, \tag{1}$$

where y_i is the response variable of interest, β_0 is an overall intercept term, $f_1(x_{1i}), \ldots, f_p(x_{pi})$ represent the nonlinear effects of p deterministically observed covariates and u_i are independent and identically distributed normal error terms with variance σ_u^2.[2] To approximate the nonlinear effects f_h, we use the weighted sum of d_h B-spline

[1] Socio-Economic Panel (SOEP), data of the years 1984-2011, version 28, SOEP, 2012, doi: 10.5684/soep.v28.

[2] For notational simplicity, we refrain from adding stochastic covariates and covariates with strictly parametric effects. However, as can be seen in Section 5, partially linear models can also be easily handled within our framework.

basis functions, $B_{h1}, \ldots, B_{hd_h}$, such that

$$f_h(x_{hi}) = \sum_{j=1}^{d_h} B_{hj}(x_{hi})\beta_{hj} = \mathbf{z}_h^T(x_{hi})\boldsymbol{\beta}_h, \tag{2}$$

where $\boldsymbol{\beta}_h$ is a d_h-dimensonal column vector of basis coefficients and $\mathbf{z}_h(x_{hi})$ is the d_h-dimensonal column vector containing the evaluations of the basis functions at the observed covariate value x_{hi}. In compact matrix notation we drop this dependency on the covariate and write

$$f_h(\mathbf{x}_h) = \mathbf{Z}_h\boldsymbol{\beta}_h,$$

where $\mathbf{Z}_h$ is a design matrix of dimension $n \times d_h$. In order to avoid an overfit to the data, a matrix $\mathbf{K}_h$ penalizing to much variability of adjacent coefficients in the coefficient vector $\boldsymbol{\beta}_h$ is assigned to each smooth function resulting in the penalized least squares criterion

$$\left(\mathbf{y} - \beta_0\mathbf{1}_n - \sum_{h=1}^{p}\mathbf{Z}_h\boldsymbol{\beta}_h\right)^T\left(\mathbf{y} - \beta_0\mathbf{1}_n - \sum_{h=1}^{p}\mathbf{Z}_h\boldsymbol{\beta}_h\right) + \sum_{h=1}^{p}\lambda_h\boldsymbol{\beta}_h^T\mathbf{K}_h\boldsymbol{\beta}_h, \tag{3}$$

where $\mathbf{y}$ denotes the n-dimensional response vector, $\mathbf{1}_n$ is an n-dimensonal column vector of ones and λ_h is a smoothing parameter determining the impact of the penalty on the minimization criterion. The $d_h \times d_h$-dimensional matrix $\mathbf{K_h}$ of first-order differences, i.e. penalizing differences of directly contiguous coefficients, has the form

$$\boldsymbol{K_h} = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}.$$

Difference matrices of higher orders can be easily constructed. Details on such penalties for B-spline functions can be found, for example, in Eilers and Marx (1996).

Let now $x_0^{(h)}$ be an arbitrary value on the domain of $\mathbf{x}_h$. Defining the smoothing matrix $\mathbf{L}_h(x_0^{(h)})$ as

$$\mathbf{L}_h(x_0^{(h)}) = (\mathbf{I}_n - \mathbf{S}_{-h})\,\mathbf{Z}_h\left[\mathbf{Z}_h^T(\mathbf{I}_n - \mathbf{S}_{-h})\mathbf{Z}_h + \lambda_h\mathbf{K}_h\right]^{-1}\mathbf{z}_h^T(x_0^{(h)}) \tag{4}$$

with $\mathbf{I}_n$ denoting the identity matrix of dimension $n \times n$, $\mathbf{S}_{-h} = \mathbf{Z}_{-h}(\mathbf{Z}_{-h}^T\mathbf{Z}_{-h} + \boldsymbol{\lambda}_{-h}\mathbf{K}_{-h})^{-1}\mathbf{Z}_{-h}^T$ with $\boldsymbol{\lambda}_{-h} = (\lambda_1, \ldots, \lambda_{h-1}, \lambda_{h+1}, \ldots, \lambda_p)$, $\mathbf{Z}_{-h} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{h-1}, \mathbf{Z}_{h+1}, \ldots, \mathbf{Z}_p)$, $\mathbf{K}_{-h} = (\mathbf{K}_1, \ldots, \mathbf{K}_{h-1}, \mathbf{K}_{h+1}, \ldots, \mathbf{K}_p)$, and $\mathbf{z}_h(x_0^{(h)})$ defined as in (2), the estimator of each $f_h(x_0^{(h)})$ can be written as

$$\hat{f}_h(x_0^{(h)}) = \mathbf{L}_h^T(x_0^{(h)})\mathbf{y}.$$

It follows that

$$\text{Var}\left[\hat{f}_h(x_0^{(h)})\right] = \text{Var}\left[\mathbf{L}_h^T(x_0^{(h)})\mathbf{y}\right] = \mathbf{L}_h^T(x_0^{(h)})\text{Var}(\mathbf{y})\mathbf{L}_h(x_0^{(h)}) = \mathbf{L}_h^T(x_0^{(h)})\sigma_u^2\mathbf{I}_n\mathbf{L}_h(x_0^{(h)}) \tag{5}$$

holds for homoscedastic and independent errors.

Naturally, the smoothing parameters λ_h are unknown. One way to estimate them is to exploit the mixed model representation of penalized splines. In particular, it is possible to rewrite

$$\mathbf{Z}_h\boldsymbol{\beta}_h = \mathbf{Z}_h(\mathbf{F}_{hf}\boldsymbol{\alpha}_{hf} + \mathbf{F}_{hr}\boldsymbol{\alpha}_{hr}) = \mathbf{X}_{hf}\boldsymbol{\alpha}_{hf} + \mathbf{X}_{hr}\boldsymbol{\alpha}_{hr}$$

such that $\boldsymbol{\alpha}_{hf}$ contains fixed coefficients and $\boldsymbol{\alpha}_{hr}$ is a vector of i.i.d. random coefficients with variance σ_{re}^2, which are assumed to be independent from the errors u_i.[3] In this mixed model formulation, we obtain estimates both for the coefficients (fixed and random) and smoothing parameters by a single (restricted) maximum likelihood estimation. For details we refer the reader to Ruppert and Wand (2003) or Fahrmeir et al. (2013). In Section 3 we will make use of the the mixed model formulation to construct confidence bands.

[3] One way to obtain the mixed model representation is described in Wood (2006, pp. 316-317).

It should be noted that each row in the initial design matrix $\mathbf{Z}_h$ (i.e., before applying the mixed model reformulation) for each covariate sums up to one, i.e., $\sum_{j=1}^{d} B_{hj}(x_{hi}) = 1 \; \forall \, i = 1, \ldots, n$. Obviously, this leads to an identification problem in an additive model with an intercept or multiple smooth components. This issue can be solved by imposing a centering constraint on each function f_h such that

$$\sum_{i=1}^{n} f_h(x_{hi}) = \sum_{i=1}^{n} \mathbf{z}_h^T(x_{hi})\boldsymbol{\beta}_h = 0$$

holds for all $h = 1, \ldots, p$. Following the ideas of Wood (2006, pp. 167-168), this can be achieved by constructing appropriate matrices $\mathbf{W}_h$ of dimension $d_h \times (d_h - 1)$ with orthogonal columns, leading to a reparameterized model with design matrices $\tilde{\mathbf{Z}}_h = \mathbf{Z}_h\mathbf{W}_h$ and penalty matrices $\tilde{\mathbf{K}}_h = \mathbf{W}_h^T\mathbf{K}_h\mathbf{W}_h$. If the mixed model framework is used to determine the smoothing parameters as described above, the reparameterizing procedure to ensure identifiability is done before the mixed model reformulation of the model.

2.2 Penalized splines for panel data: A first-difference estimator

In comparison to cross-sectional data leading to model (1) introduced in the previous section, we now consider individuals (e.g., persons) observed at T consecutive points of time.[4] We therefore consider an additive panel data model

$$y_{it} = \gamma_i + \sum_{h=1}^{p} f_h(x_{hit}) + u_{it}, \; i = 1, \ldots, N, \; t = 1, \ldots, T, \tag{6}$$

where u_{it} are assumed to be independent and normally distributed errors with constant variance and γ_i are individual-specific, time-invariant fixed effects allowed to be correlated with other covariates. As model (6) holds for each point of time, we obtain

$$y_{i,t-1} = \gamma_i + \sum_{h=1}^{p} f_h(x_{hi,t-1}) + u_{i,t-1} \tag{7}$$

for a one period time lag. To cancel out the individual-specific effects γ_i, we subtract (7) from (6) and obtain

$$\begin{aligned}
\Delta y_{it} = y_{it} - y_{i,t-1} &= \gamma_i - \gamma_i + \sum_{h=1}^{p} \left[f_h(x_{hit}) - f_h(x_{hi,t-1}) \right] + u_{it} - u_{i,t-1} \qquad (8) \\
&= \sum_{h=1}^{p} \left[\sum_{j=1}^{d_h} B_{hj}(x_{hit})\beta_{hj} - \sum_{j=1}^{d_h} B_{hj}(x_{hi,t-1})\beta_{hj} \right] + \Delta u_{it} \\
&= \sum_{h=1}^{p} \left[\mathbf{z}_h(x_{hit}) - \mathbf{z}_h(x_{hi,t-1}) \right]^T \boldsymbol{\beta}_h + \Delta u_{it} \\
&= \sum_{h=1}^{p} \left[\Delta \mathbf{z}_h(x_{hit}) \right]^T \boldsymbol{\beta}_h + \Delta u_{it}, \qquad (9)
\end{aligned}$$

where equation (2) is used for the second and third equality and Δ denotes the first-difference operator over time. Note that only $T - 1$ observations per individual are retained after differencing. Accordingly, as the $NT \times d_h$-

[4] The only reason to refrain from incorporating different observation horizons between persons is notational convenience. As can be seen in Section 4 and Section 5, unbalanced panels can be handled without any difficulties in our framework.

dimensional design matrix $\mathbf{Z}_h$ of the evaluated basis functions is given by

$$\mathbf{Z}_h = \begin{pmatrix} B_{h1}(x_{h11}) & \dots & B_{hd_h}(x_{h11}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{h1T}) & \dots & B_{hd_h}(x_{h1T}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{hN1}) & \dots & B_{hd_h}(x_{hN1}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{hNT}) & \dots & B_{hd_h}(x_{hNT}) \end{pmatrix}, \tag{10}$$

we obtain

$$\Delta\mathbf{y} = \sum_{h=1}^{p} \Delta\mathbf{Z}_h \boldsymbol{\beta}_h + \Delta\mathbf{u} \tag{11}$$

in compact matrix notation, where $\Delta\mathbf{y} = (y_{12} - y_{11}, \dots, y_{1T} - y_{1,T-1}, \dots, y_{N2} - y_{N1}, \dots, y_{NT} - y_{N,T-1})^T$ is a $N(T-1)$-dimensional column vector, $\Delta\mathbf{u}$ is defined analogously and the $N(T-1) \times d_h$-dimensional matrix $\Delta\mathbf{Z}_h$ is obtained by building the difference between matrix $\mathbf{Z}_h$ in (10) and its one period lagged counterpart:

$$\Delta\mathbf{Z}_h = \begin{pmatrix} B_{h1}(x_{h12}) & \dots & B_{hd_h}(x_{h12}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{h1T}) & \dots & B_{hd_h}(x_{h1T}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{hN2}) & \dots & B_{hd_h}(x_{hN2}) \\ & \ddots & \\ B_{h1}(x_{hNT}) & \dots & B_{hd_h}(x_{hNT}) \end{pmatrix} - \begin{pmatrix} B_{h1}(x_{h11}) & \dots & B_{hd_h}(x_{h11}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{h,T-1}) & \dots & B_{hd_h}(x_{h,T-1}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{hN1}) & \dots & B_{hd_h}(x_{hN1}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{hN,T-1}) & \dots & B_{hd_h}(x_{hN,T-1}) \end{pmatrix}.$$

Additionally taking into account penalization, a first-difference penalized spline estimator for all $\boldsymbol{\beta}_h$ can be obtained by minimizing the penalized least squares criterion

$$\left[\Delta\mathbf{y} - \sum_{h=1}^{p} (\Delta\mathbf{Z}_h)\boldsymbol{\beta}_h\right]^T \left[\Delta\mathbf{y} - \sum_{h=1}^{p} (\Delta\mathbf{Z}_h)\boldsymbol{\beta}_h\right] + \sum_{h=1}^{p} \lambda_h {\boldsymbol{\beta}_h}^T \mathbf{K}_h \boldsymbol{\beta}_h. \tag{12}$$

Since the smoothing parameters are unknown, one can again exploit the mixed model representation and using (restricted) maximum likelihood estimation as discussed in the previous subsection.

We briefly have to refer to the identification problem in case of multiple smooth model components: Our aim is to estimate the functions f_h, $h = 1, \dots, p$. Hence, model (6) should be identified such that

$$\sum_{i=1}^{N} \sum_{t=1}^{T} f_h(x_{hit}) = \mathbf{Z}_h \boldsymbol{\beta}_h = 0$$

holds for all $h = 1, \dots, p$. Therefore, we rewrite the design matrices of the evaluated basis function given in (10) and the penalty matrices such that $\tilde{\mathbf{Z}}_h = \mathbf{Z}_h \mathbf{W}_h$ and $\tilde{\mathbf{K}}_h = \mathbf{W}_h^T \mathbf{K}_h \mathbf{W}_h$, proceeding as described in the previous subsection. Furthermore, the identification restriction also implies that a one period lagged design matrix is then constructed directly from $\tilde{\mathbf{Z}}_h$ by taking its one-period-lagged rows. After building the difference between each $\tilde{\mathbf{Z}}_h$ and its respective lagged counterpart, the resulting matrices $\Delta\tilde{\mathbf{Z}}_h$ and the penalty matrices $\tilde{\mathbf{K}}_h$ are plugged into (12) to obtain estimators for $\boldsymbol{\beta}_h$ and thus for f_h.

Another common approach in fixed effects panel data models is time-demeaning, i.e., removing the individual-specific effects γ_i by building the mean over time for each individual in equation (6) and subtracting the resulting equation from (6). Using the information above, this variant is straightforward to derive.

3 Simultaneous confidence bands for penalized splines

In linear regression models, one is typically interested in the uncertainty of the parameter estimates. Confidence intervals are an established tool to make inferential statements. Similarly, inference about entire smooth functions in nonparametric regression models can be obtained by constructing simultaneous confidence bands around the estimated functions:

$$\left\{ \hat{f}_h(x_h) - c_{h,1-\alpha}\sqrt{\text{Var}\left[\hat{f}_h(x_h)\right]}, \hat{f}_h(x_h) + c_{h,1-\alpha}\sqrt{\text{Var}\left[\hat{f}_h(x_h)\right]},\ x_{h,min} \leq x_h \leq x_{h,max} \right\}. \tag{13}$$

The critical value $c_{h,1-\alpha}$ should ensure that the resulting bands (depending on the sample at hand) cover the true function with a prespecified probability $1-\alpha$ in all possible samples, i.e., $c_{1-\alpha}$ is the $(1-\alpha)$-quantile of the random variable

$$\sup_{x_{h,min} \leq x_h \leq x_{h,max}} \frac{|\hat{f}_h(\mathbf{x}_h) - f_h(\mathbf{x}_h)|}{\sqrt{\text{Var}\left[\hat{f}_h(\mathbf{x}_h)\right]}}.$$

The difficulty in the penalized spline framework lies in finding the distribution of this random variable. Due to the introduction of a penalty, the estimators for f_h, obtained for instance by minimizing (3) or (12), are usually not unbiased.[5] Krivobokova et al. (2010) propose a solution that takes this bias into account when constructing the simultaneous confidence bands for penalized splines. They consider univariate models while Wiesenfarth et al. (2012) extend the approach to the multivariate case, also covering heteroscedastic errors and spatially heterogeneous splines. The approach performs very well in simulation studies and offers a fast way of inference without the need for computationally intensive resampling procedures. The basic idea (derived for the cross-sectional case here) is to exploit the mixed model representation of penalized splines as described in Section 2, i.e., we consider smooth functions as mixed models:

$$f_h^m(\mathbf{x}_h) := \mathbf{X}_{hf}\boldsymbol{\alpha}_{hf} + \mathbf{X}_{hr}\boldsymbol{\alpha}_{hr} = \mathbf{Z}_h^m\boldsymbol{\beta}_h^m.$$

Recall that both the the random coefficients in each random coefficients vector $\boldsymbol{\alpha}_{hr}$, $h = 1 \dots p$, and the model errors u_i are assumed to be independent and normally distributed with zero expectation and constant variance. Additionally assuming mutual independence, the marginal distribution of $\mathbf{y}$ is given by

$$\mathbf{y} \sim N\left(\beta_0\mathbf{1}_n + \sum_{h=1}^{p}\mathbf{X}_{hf}\boldsymbol{\alpha}_{hf}, \sigma_u^2\mathbf{I}_n + \sum_{h=1}^{p}\sigma_{re}^2\mathbf{X}_{hr}\mathbf{X}_{hr}^T\right).$$

Since the fixed coefficients estimators are unbiased, we obtain a zero mean Gaussian process

$$G_h(\mathbf{x}_h) = \frac{\mathbf{Z}_h^m(\hat{\boldsymbol{\beta}}_h^m - \boldsymbol{\beta}_h^m)}{\sqrt{\mathbf{Z}_h^m\text{Cov}(\hat{\boldsymbol{\beta}}_h^m - \boldsymbol{\beta}_h^m)(\mathbf{Z}_h^m)^T}} \sim N(0, \boldsymbol{\Sigma})$$

with regard to each covariate. The entries of the covariance matrix $\boldsymbol{\Sigma}$ are then given by

$$\text{Cov}\left[G_h(x_{h1}), G_h(x_{h2})\right] = \left[\frac{\mathbf{L}_{m,h}(x_{h1})}{||\mathbf{L}_{m,h}(x_{h1})||}\right]^T\left[\frac{\mathbf{L}_{m,h}(x_{h2})}{||\mathbf{L}_{m,h}(x_{h2})||}\right] =: \boldsymbol{\eta}_{m,h}^T(x_{h1})\boldsymbol{\eta}_{m,h}(x_{h2}),$$

[5] The spline representation of smooth function introduces an additional bias which converges to zero with growing number of knots, see Claeskens et al. (2009) for details. We assume this bias to be negligible by using sufficiently many knots.

where $\mathbf{L}_{m,h}(\cdot)$ denotes the smoothing matrix from (4) in mixed model formulation. Following Sun and Loader (1994), the tail probability of maxima of such processes is determined by

$$\begin{aligned} \alpha &= P\left(\sup_{x_{h,min}\leq x_h \leq x_{h,max}} |G_j(x)| \geqslant c_{h,1-\alpha}\right) \\ &= \frac{\kappa_{m,h}}{\pi}\mathrm{Cov}\left(\frac{-c_{h,1-\alpha}}{2}\right) + 2\left[1-\Phi(c_{h,1-\alpha})\right] + o\left[\exp\left(\frac{-c_{h,1-\alpha}}{2}\right)\right], \end{aligned} \tag{14}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution and

$$\kappa_{m,h} = \int_{x_{h,min}}^{x_{h,max}} ||\frac{d}{dx}\boldsymbol{\eta} m, h(x)||dx$$

is the length of the mixed model manifold implicitely including the amount of bias which has to be corrected for. Thus, the critical value $c_{h,1-\alpha}$ in (13) can be approximately obtained from (14). For further details of such simultaneous confidence bands see Krivobokova et al. (2010) and Wiesenfarth et al. (2012). Their approach is designed for the cross-sectional case, but directly carries over to the panel data context with fixed effects as described in (6). The simple, but crucial new aspect to contemplate is the serial correlation in the error term Δu_{it} of each individual after applying the first-difference transformation described in (8). Assuming the u_{it} to be serially uncorrelated, Δu_{it} and $\Delta u_{i,t-1}$ exhibit a negative autocorrelation for each individual. In case of a homoscedastic variance, this serial correlation for two consecutive points of time amounts to -0.5, see the appendix for a derivation. We therefore adopt the generalized least squares (GLS) approach and premultiply the differenced model matrix ($\Delta\mathbf{Z}_h$) and the differenced dependent variable $\Delta\mathbf{y}$ in equation (12) by $\boldsymbol{\Psi}$, where

$$\boldsymbol{\Psi}\boldsymbol{\Psi}' = \boldsymbol{\Omega}^{-1} = \begin{pmatrix} \boldsymbol{\Omega}_1^{-1} & 0 & \dots & 0 \\ 0 & \boldsymbol{\Omega}_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\Omega}_N^{-1} \end{pmatrix} \tag{15}$$

is a block diagonal matrix with main diagonal block square matrices

$$\boldsymbol{\Omega}_i^{-1} = \begin{pmatrix} 1 & -0.5 & 0 & \dots & 0 \\ -0.5 & 1 & 0 & \dots & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & 0 & 1 & -0.5 \\ 0 & 0 & 0 & -0.5 & 1 \end{pmatrix}$$

of dimension $(T-1)\times(T-1)$.[6] Note that, when using first differences and GLS, the smoothing matrix in (4) and thus the variance and the confidence bands of the estimated spline curve change accordingly. Likewise, applying the GLS transformation on the respective quantities in the penalized least squares criterion (12) results in a more efficient estimator for the unknown functions.

In practice, panel data often exhibit additional serial correlation. In the rare cases of an exactly known error structure, the matrices in (15) can be adjusted. The more common case is that the correlation structure in the error term is unknown and only minor assumptions are made, e.g., that errors between different individuals are uncorrelated. In such a case, it is recommended to investigate the residuals for all individuals before or after applying the GLS procedure. If the autocorrelation and partial autocorrelation function suggest the occurrence of a certain underlying autoregressive moving average process, the obtained information could be exploited in the subsequent estimation of a feasible GLS estimator, see Hansen (2007) for more details. Another option is the maximum likelihood-type reestimation of the model with included simultaneous estimation of the autoregressive and moving average parameters. This can be done in a mixed model framework which additionally allows for modeling heteroscedasticity, as described in Pinheiro and Bates (2000).

Wiesenfarth et al. (2012) describe the extension how to build simultaneous confidence bands around the deriva-

[6] $\boldsymbol{\Psi}$ can be obtained from $\boldsymbol{\Omega}^{-1}$ with the help of the Cholesky factorization and matrix inversion.

tives. In the case of B-spline basis functions, the derivative of the smoothing matrix in (4) for the cross-sectional case is given by

$$\mathbf{L}'_h(x_{h0}) = (\mathbf{I}_n - \mathbf{S}_{-h})\,\mathbf{Z}_h \left[\mathbf{Z}_h^T (\mathbf{I}_n - \mathbf{S}_{-h})\,\mathbf{Z}_h + \lambda_h \mathbf{Z}_h\right]^{-1} \left[\mathbf{z}'_h(x_{h0})\right]^T, \tag{16}$$

where $[\mathbf{z}'_h(x_{h0})]^T$ denotes the row vector of the derivatives of the initial basis functions, evaluated at some value x_{h0} (see De Boor, 2001, Ch. 10). Thus, derivative estimates are practically obtained with negligible effort once a penalized least squares criterion like (3) has been minimized. Critical values and simultaneous confidence bands for the derivatives, also for panel data settings, can then be obtained by analogy with the steps described above.

4 Simulation studies

We consider data generated from model (6) with the individual-specific fixed effects $\gamma_i = i$ and the $p = 3$ true functions

$$\begin{aligned} f_1(x_{1it}) &= \sin^2\left[2\pi(x_{1it} - 0.5)\right], \\ f_2(x_{2it}) &= 0.6 b_{30,17}(x_{2it}) + 0.4 b_{3,11}(x_{2it}), \\ f_3(x_{3it}) &= x_{3it}(1 - x_{3it}), \end{aligned}$$

with $b_{l,m}(x) = \Gamma(l+m)\left[\Gamma(l)\Gamma(m)\right]^{-1} x^{l-1}(1-x)^{m-1}$, where $\Gamma(r)$ denotes the gamma function. All functions were also considered in Wiesenfarth et al. (2012). They are scaled such that their standard deviations are equal to one. The functions and their derivatives are shown in Figure 3 in the appendix. The errors are generated as i.i.d. Gaussian errors with standard deviation $\sigma_u = 0.5$. We consider an unbalanced panel data design with total sample sizes of $n = (525, 1050, 2100)$, where $N = (75, 150, 300)$ imaginary individuals are observed over different time horizons without breaks, i.e., there are no missing observations between the first and last point of time at which one individual is observed. Note that due to taking first differences according to (11), the sample size used for the estimation decreases by the number of individuals, i.e., we obtain the effective sample sizes $n - N = (450, 900, 2700)$. The covariates for each individual are taken to be distributed over $\{a - 0.04, a - 0.03, \ldots, a, a + 0.01\}$ with

$$P(X = x) = \begin{cases} 0.5, & \text{if } x = a, \\ 0.1 & \text{else,} \end{cases}$$

with a being randomly drawn with equal probability from $\{0.04, 0.05, ..., 0.99\}$ for each individual. This setting is designed to mimic a real-world panel data set where covariate values of individuals are often restricted to a finite set of values and can sometimes remain constant over time. In all settings, we take 40 equidistant knots for all covariates. The results are based on 500 Monte Carlo replicates and a nominal coverage rate of 95%. Note that under the error assumptions stated above, the errors after building first differences are serially correlated (see Section 3). We use B-spline basis function of degree three and impose a penalty on second-order differences of the B-spline coefficients.

In Table 1, the resulting coverage rates with and without using GLS are shown. It can be seen that not taking into account the autocorrelation in the error term leads to substantial undercoverage. In contrast, even for moderate sample size, the confidence bands estimated by GLS generally perform quite accurately, i.e., the nominal coverage is met. These results are in line with those of Wiesenfarth et al. (2012).

Using the same setting, we also examine the coverage rates of the confidence bands for the derivatives. The results in Table 1 show adequate coverage rates for the comparably simple linear derivative $f'_3(x_{3it})$ but not for the two other more complicated functions. Especially the confidence bands for $f'_2(x_{2it})$ perform poorly,[7] even if the sample size is huge ($n = 4200$) or the error variance is low (not shown here for brevity). In further simulations, we also varied the number of knots and the difference orders for the penalty. Although sometimes observing improvements in the coverage rates (with or without the expense of wider confidence bands), we did not find a distinct pattern how to reach the nominal coverage rate. Thus, we can only advise to be careful in making inferential statements about the derivative of potentially sophisticated curves.

In addition, we replicated the simulation studies with non-Gaussian errors and autocorrelated errors. The results are comparable to our simulation setting with independent Gaussian errors. The robustness of the here proposed confidence bands for non-normal symmetric error distributions was also demonstrated by Loader and Sun

[7] We observe similar problems for other functions, e.g., $f(x) = (0.5 - x)^3$, see the online supplement.

Table 1: Coverage rates in simulations, average areas between confidence bands in parentheses. Columns (i) denote estimation with using GLS, columns (ii) without using GLS.

n	f_1		f_2		f_3	
	(i)	(ii)	(i)	(ii)	(i)	(ii)
525	0.95	0.86	0.93	0.85	0.97	0.91
	(3.42)	(3.48)	(3.67)	(3.63)	(3.07)	(3.24)
1050	0.95	0.88	0.95	0.85	0.97	0.88
	(2.45)	(2.51)	(2.45)	(2.44)	(2.12)	(2.14)
2100	0.95	0.84	0.96	0.88	0.97	0.88
	(1.81)	(1.80)	(1.90)	(1.88)	(1.55)	(1.55)
	f_1'		f_2'		f_3'	
	(i)	(ii)	(i)	(ii)	(i)	(ii)
525	0.90	0.75	0.80	0.62	0.94	0.86
	(30.60)	(31.71)	(32.80)	(33.57)	(17.82)	(19.29)
1050	0.90	0.77	0.85	0.66	0.94	0.84
	(23.51)	(24.42)	(25.19)	(25.85)	(14.06)	(14.92)
4200	0.85	0.70	0.73	0.60	0.95	0.83
	(15.23)	(15.69)	(17.01)	(17.26)	(8.77)	(9.19)

(1997). Furthermore, our simulations indicate that slight violations of the serial independence assumption are not too harmful. However, as shown above, disregarding major serial correlation as introduced by the first-difference transformation to uncorrelated errors is problematic. Thus, we advise the practitioner to investigate the residuals and apply, if necessary, more adequate modeling approaches as described in section 3.

As for all fixed effects panel data models, it is also important to ensure that there is sufficient intra-personal variation for all covariates. If this is not the case, the model matrix after applying first differences contains many zeros and thus, there is too little variation to estimate the function adequately.

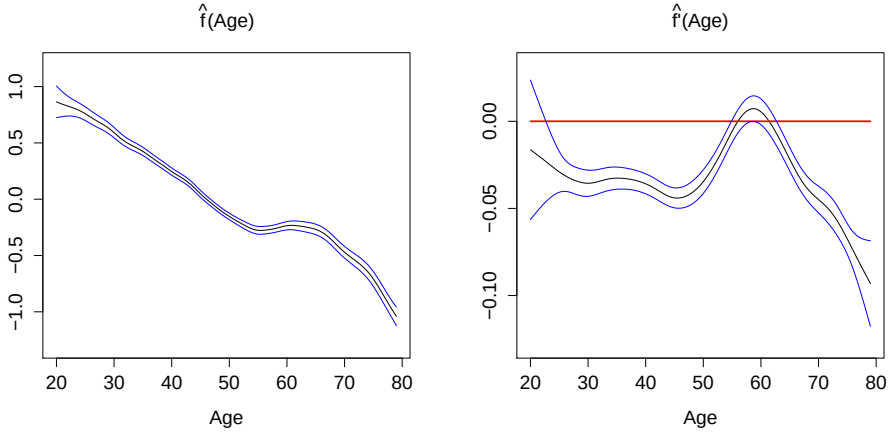
The results for additional simulations not shown but discussed in this section can be found in the online supplement.

5 Studying the relationship between ageing and life satisfaction

There is a considerable strand of literature studying how life satisfaction evolves over the lifespan. So far, there is no broad consensus on the shape of this relationship, as study results differ while applying different methodologies and data sets. A recent overview on this topic is given by López Ulloa et al. (2013). Frequently, an a priori specified U-shaped relationship is tested in a parametric way. One exception is the work of Wunder et al. (2011), who apply a semiparametric random effects model using the SOEP and the British Household Panel Survey. However, they do not address possible endogeneity of time-invariant omitted covariates which can be done by incorporating individual-specific time constant fixed effects. In the context of the relationship between ageing and life satisfaction, the importance of doing so is highlighted by Ferrer-i Carbonell and Frijters (2004). Using fixed effects panel models, Frijters and Beatton (2012) apply a quite flexible step function based on 5-year-intervals for the influence of age on life satisfaction, which is, however, non-continuous and does not allow for uncertainty statements. To the best of our knowledge, we provide the first fully flexible fixed effects panel data approach also allowing for statistical quantification of uncertainty. To illustrate our method, we use SOEP data from 1994 to 2011, see Wagner et al. (2007) for details on the data set. Following the results of Ferrer-i Carbonell and Frijters (2004), we treat the life satisfaction score[8], which is measured on an actually ordinal 11-point scale ranging from 0 (completely dissatisfied) to 10 (completely satisfied), as cardinal. While applying a first-difference estimator, the effects on life satisfaction are assumed to be exclusively instantaneous, i.e., an increase or decrease of an explanatory variable in one year influences life satisfaction solely in the same year. This is questionable especially in the case of certain life events like changes in the marital status, for instance. Therefore, we follow an approach similar to Laporte and Windmeijer

[8] The corresponding question in the SOEP survey ist: "How satisfied are you with your life, all things considered?"

Figure 1: Estimated nonparametric relationship between age and life satisfaction with confidence bands (left panel), corresponding estimated derivative (right panel)



(2005) and add dummy variables for each of the two years before and after a life event,[9] including changes in marital, employment and disability status. Furthermore, we include nonparametric effects for age and net household income (with 60 equidistant knots each) and linear effects for household size and nights stayed in hospital in the previous year. Thus, our model to estimate is

$$\text{Life Satisfaction}_{it} = \gamma_i + f(\text{Age}_{it}) + f(\text{Household Income}_{it}) + \mathbf{c}_{it}^T\delta + u_{it}, \tag{17}$$

where the vector $\mathbf{c}_{it}$ captures the values of all variables (including lags and leads) which are modelled in a parametric fashion. The final sample size after removing missing values amounts to $n = 143,299$.

The results for the nonparametric effect of age on life satisfaction can be found in the left panel of Figure 1. It can be seen that young people tend to become more and more unhappy as they become older. This decrease in life satisfaction is stopped and even slighlty reversed at the age of around 60 for a couple of years. After that, increasing age again goes along with a reduction in life satisfaction. The estimated derivative of this effect and its confidence bands are shown in the right panel of Figure 1. For ages older than about 25 years, the zero line is not covered by the bands over almost the whole life span, indicating a significant negative effect of age on life satisfaction within these ages. This does not hold for the ages around 60 years. There, the confidence bands cover the zero line and the lower band almost crosses the zero line once. With regard to our simulation studies in Section 4, however, these results should be taken with caution.

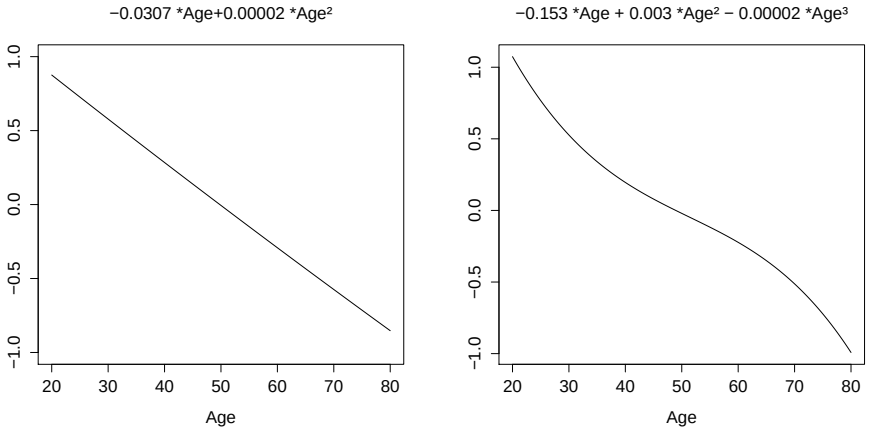
For comparison, we also estimate two simple parametric first-difference panel data models, where the smooth functions of age and household income in equation (17) are replaced by quadratic and cubic polynomials. The results for the age effect can be found in Figure 2. It can be seen that the quadratic fit is a quasi-linear decreasing function, whereas the cubic fit shows some curvature while still exhibiting a clear downward trend over the lifespan. Neither of these estimated functions can capture the stage of constant or even increasing life satisfaction for the ages around 60 years. Thus, it is advisable to use a nonparametric estimator here to estimate the relationship of interest. In our analysis the often found U-shape or any other simple relationship between age and life satisfaction cannot be confirmed. Qualitatively, our results rather resemble those of Wunder et al. (2011). The nonparametric effect of net household income as well as the purely parametric effects are shown in Figure 4 and Table 2 in the appendix.

6 Discussion and conclusions

In this paper, we presented a nonparametric first-difference panel data estimator based on penalized splines together with a corresponding fast way of inference via simultaneous confidence bands. Our approach allows to estimate and

[9] Incorporating leads and lags results in a smaller sample size. In our case, we require an individual to be observed in at least six consecutive periods corresponding to at least one observation for estimation after building two leads and two lags and taking first differences. Albeit the loss of observations, this modeling procedure allows us to investigate whether effects on life satisfaction are long-lasting or just temporary, see for instance Lucas (2007) for a discussion on this issue.

Figure 2: Estimated parametric relationship between age and life satisfaction with sqared (left) and cubic polynomial (right)



draw inferences from fixed effects panel data models in a highly flexible way and without a priori specifications of covariate effects. Furthermore, the derivatives of the estimated effects as well as of their confidence bands are made available with negligible additional effort. Using data from the SOEP, we illustrated our method by modeling the relationship between age and life satisfaction. We found that it is not advisable to model this non-linear relationship in a strictly parametric fashion. Simulation studies showed an overall good performance of our method with the exception of the confidence bands for the derivatives which sometimes failed to hit the nominal coverage rate. A possible explanation is that the smoothing parameters are estimated and optimized for the original functions and not for the derivatives, as pointed out by Ruppert and Wand (2003, Ch. 6.8). It might be an interesting direction for future research to address this problem. The proposed approach is available for practitioners in the R package *pamfe* which enables the fast estimation of partially linear models and models with multiple nonlinear components even for large sample sizes.

Appendix

Serial correlation in the first-difference errors

Consider equation (6): If the error terms $u_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T_i$ are homoscedastic and independent with expectation zero, then $\mathrm{E}(u_{it}u_{i,t-1}) = 0$ and $\mathrm{E}(u_{it}u_{it}) = \sigma_u^2$. It follows for the errors $\triangle u_{it} = u_{it} - u_{i,t-1}$ in equation (8):

$$\mathrm{E}(\Delta u_{it}) = \mathrm{E}(u_{it} - u_{i,t-1}) = 0$$

and

$$\mathrm{Var}(\Delta u_{it}) = \mathrm{Var}(u_{it} - u_{i,t-1}) = \mathrm{Var}(u_{it}) + \mathrm{Var}(u_{i,t-1}) = 2\sigma_u^2.$$

The correlation of two consecutive error terms for the same individual after applying first differences is then given by

$$\begin{aligned}
\mathrm{Cor}(\Delta u_{it}, \Delta u_{i,t-1}) &= \frac{\mathrm{E}\left[(\Delta u_{it})(\Delta u_{i,t-1})\right]}{\sqrt{\mathrm{Var}(\Delta u_{it})\mathrm{Var}(\Delta u_{i,t-1})}} \\
&= \frac{\mathrm{E}\left[(u_{it} - u_{i,t-1})(u_{i,t-1} - u_{i,t-2})\right]}{\sqrt{2\sigma_u^2 2\sigma_u^2}} \\
&= \frac{\mathrm{E}(-u_{i,t-1}^2)}{2\sigma_u^2} = \frac{-\sigma_u^2}{2\sigma_u^2} = -0.5.
\end{aligned}$$

Figures and tables

Figure 3: Simulation studies: True, scaled functions (left) and corresponding derivatives (right).

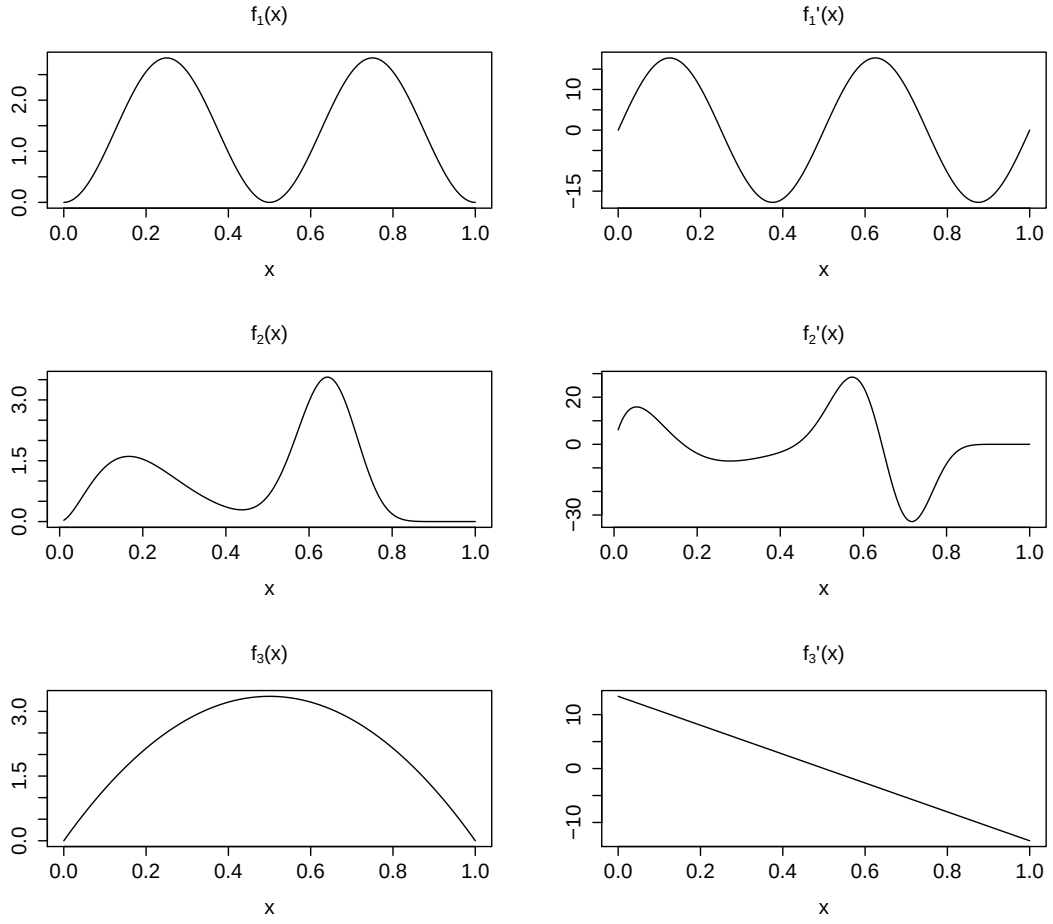


Figure 4: Estimated nonparametric relationship between household income (in 1000 €) and life satisfaction with confidence bands

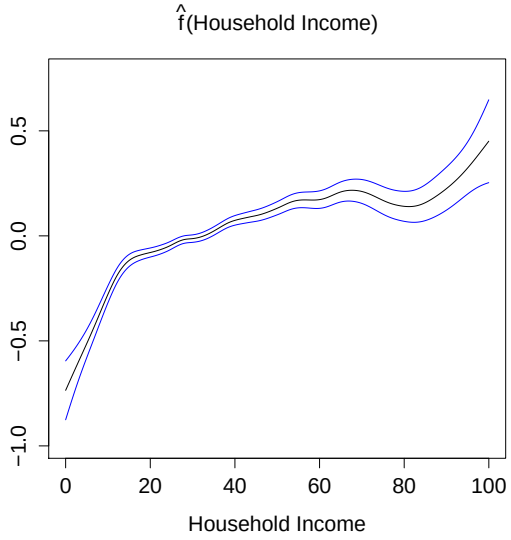


Table 2: Estimation results for strictly parametric components. Note that the reference categories for the marital status and its leads and lags are "single" and its respective leads and lags. For the disability status "not disabled" serves at reference category, so does "non-working" for the employment status.

Variable	Coefficient	P-value
Household size	-0.0048	0.5668
Nights in hospital	-0.0102	0.0000
Disability Status: Disabled + 2 years	-0.0156	0.5107
Disability Status: Disabled + 1 year	0.0334	0.1763
Disability Status: Disabled	-0.1533	0.0000
Disability Status: Disabled - 1 year	-0.2208	0.0000
Disability Status: Disabled - 2 years	-0.1775	0.0000
Divorced + 2 years	0.0482	0.2165
Divorced + 1 year	0.2686	0.0000
Divorced	0.0289	0.5528
Divorced - 1 year	-0.1348	0.1095
Divorced - 2 years	-0.0744	0.3061
Widowed + 2 years	0.2420	0.0000
Widowed + 1 year	0.5067	0.0000
Widowed + 1 year	0.5067	0.0000
Widowed	-0.3942	0.0000
Widowed - 1 year	-0.0820	0.2459
Widowed - 2 years	-0.0935	0.1195
Married + 2 years	-0.1082	0.0006
Married + 1 year	-0.0569	0.1388
Married	0.1143	0.0046
Married - 1 year	0.1463	0.0007
Married - 2 years	0.1418	0.0002
Part time employed	0.0061	0.7807
Full time employed	0.1235	0.0000
Unemployed	-0.4843	0.0000

References

Baltagi, B. H. and Li, D.: 2002, Series Estimation of Partially Linear Panel Data Models with Fixed Effects, *Annals of Economics and Finance* **3**(1995), 103–116.

Chen, J., Li, D. and Gao, J.: 2013, Non- and Semi-Parametric Panel Data Models: A Selective Review.
URL: *http://ecgi.ssrn.com/delivery.php?ID=http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2313431*

Claeskens, G., Krivobokova, T. and Opsomer, J. D.: 2009, Asymptotic Properties of Penalized Spline Estimators, *Biometrika* **96**(3), 529 –544.

Claeskens, G. and Van Keilegom, I.: 2003, Bootstrap Confidence Bands for Regression Curves and their Derivatives, *The Annals of Statistics* **31**(6), 1852–1884.

De Boor, C.: 2001, *A Practical Guide to Splines*, Springer.

Eilers, P. H. C. and Marx, B. D.: 1996, Flexible Smoothing with B-Splines and Penalties, *Statistical Science* **11**(2), 89–102.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B.: 2013, *Regression. Models, Methods and Applications*, Springer.

Ferrer-i Carbonell, A. and Frijters, P.: 2004, How Important is Methodology for the Estimates of the Determinants of Happiness?, *Economic Journal* **114**(497), 641–659.

Frijters, P. and Beatton, T.: 2012, The Mystery of the U-Shaped Relationship between Happiness and Age, *Journal of Economic Behavior & Organization* **82**(2-3), 525–542.

Hajargasht, G.: 2009, Nonparametric Panel Data Models: A Penalized Spline Approach.
URL: *https://sites.google.com/site/ghgasht/NonpPanelSpline.pdf?attredirects=0*

Hansen, C. B.: 2007, Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects, *Journal of Econometrics* **140**(2), 670–694.

Henderson, D. J., Carroll, R. J. and Li, Q.: 2008, Nonparametric estimation and testing of fixed effects panel data models, *Journal of Econometrics* **144**(1), 257–275.

Krivobokova, T., Kneib, T. and Claeskens, G.: 2010, Simultaneous Confidence Bands for Penalized Spline Estimators, *Journal of the American Statistical Association* **105**(490), 852–863.

Laporte, A. and Windmeijer, F.: 2005, Estimation of Panel Data Models with Binary Indicators when Treatment Effects are not Constant over Time, *Economics Letters* **88**(3), 389–396.

Li, G., Peng, H. and Tong, T.: 2013, Simultaneous Confidence Band for Nonparametric Fixed Effects Panel Data Models, *Economics Letters* **119**(3), 229–232.

Loader, C. R. and Sun, J.: 1997, Robustness of Tube Formula Based Confidence Bands, *Journal of Computational and Graphical Statistics* **6**(2), 242–250.

López Ulloa, B. F., Møller, V. and Sousa-Poza, A.: 2013, How Does Subjective Well-Being Evolve with Age? A Literature Review, *Journal of Population Ageing* **6**(3), 227–246.

Lucas, R. E.: 2007, Adaptation and the Set-Point Model of Subjective Well-Being: Does Happiness Change after Major Life Events?, *Current Directions in Psychological Science* **16**(2), 75–79.

Mammen, E., Støve, B. and Tjøstheim, D.: 2009, Nonparametric Additive Models for Panels of Time Series, *Econometric Theory* **25**(02), 442.

Pinheiro, J. C. and Bates, D. M.: 2000, *Mixed Effects Models in S and S-PLUS*, Springer.

Ruppert, D. and Wand, M.: 2003, *Semiparametric Regression*, Cambridge University Press.

Su, L. and Ullah, A.: 2006, Profile Likelihood Estimation of Partially Linear Panel Data Models with Fixed Effects, *Economics Letters* **92**, 75–81.

Su, L. and Ullah, A.: 2011, Nonparametric and Semiparametric Panel Econometric Models: Estimation and Testing, *Handbook of Empirical Economics and Finance*, Taylor & Francis Group, pp. 455–497.

Sun, J. and Loader, C. R.: 1994, Simultaneous Confidence Bands for Linear Regression and Smoothing, *The Annals of Statistics* **22**(3), 1328–1345.

Wagner, G. G., Frick, J. R. and Schupp, J.: 2007, The German Socio-Economic Panel Study (SOEP) - Scope, Evolution, and Enhancements, *Schmollers Jahrbuch* **127**(1), 139–169.

Wiesenfarth, M., Krivobokova, T., Klasen, S. and Sperlich, S.: 2012, Direct Simultaneous Inference in Additive Models and its Application to Model Undernutrition, *Journal of the American Statistical Association* **107**(500), 1286–1296.

Wood, S. N.: 2006, *Generalized Additive Models: An introduction with R*, Chapman & Hall/CRC.

Wunder, C., Wiencierz, A., Schwarze, J. and Küchenhoff, H.: 2011, Well-Being over the Life Span: Semiparametric Evidence from British and German Longitudinal Data, *Review of Economics and Statistics* **95**(1), 154–167.