Multivariable model-building with continuous covariates: 2. Comparison between splines and fractional polynomials

Harald Binder, Willi Sauerbrei & Patrick Royston

Universität Freiburg i. Br.

Nr. 106 July 2011

Zentrum für Datenanalyse und Modellbildung

Universität Freiburg

Eckerstraße 1

D–79104 Freiburg im Breisgau

und

Institut für Medizinische Biometrie und Medizinische Informatik

Universitätsklinikum Freiburg

Stefan-Meier-Straße 26

D–79104 Freiburg im Breisgau

und

MRC Clinical Trials Unit

222 Euston Road

London NW1 2DA, UK

binderh@imbi.uni-freiburg.de

Abstract

In observational studies, many continuous or categorical covariates may be related to a response of interest. Analyses based on splines as well as the multivariable fractional polynomial (MFP) approach can be applied to identify important variables and appropriate functional forms for continuous covariates. Whereas MFP is one well-defined procedure, many strategies based on splines have been suggested, and we chose to study two of them. The aim of an analysis often guides the level of complexity that is deemed acceptable for the final model. Spline-based strategies and MFP have tuning parameters for choosing the required level of complexity. However, it is unclear whether the strategies can equally well provide simple as well as complex models. Furthermore, a 'reasonable' level of complexity may depend on the specific data situation. Therefore, we perform a comprehensive simulation study predicated on an underlying ('true') structure that realistically reflects biomedical contexts. We vary the amount of information (signal-to-noise ratio) in the data and the complexity levels for model selection. We consider prediction performance, Type I and Type II error rates, costs at the covariate level, and quantitative as well as qualitative criteria for judging selected functional forms. No one procedure performs best in all scenarios, but overall, MFP shows better performance than the multivariable spline strategies we investigated.

Keywords: continuous covariates; fractional polynomials; model selection; non-linear effects; simulation; splines.

1 Introduction

For building a multivariable regression model with continuous covariates, any statistical technique needs to meet two primary challenges: selecting covariates that are related to the response of interest, and selecting a suitable functional form for the continuous covariates. Sometimes it may be advisable to ignore the second task; linear regression

models (in which the functional form is assumed linear) have been quite successful (see (1), for example). However, when the focus is on interpretability of the selected models, a wider class of functional forms needs to be allowed for. Non-linear effects of individual covariates have often been modeled using spline techniques (see (2), for example, for B-splines), where local polynomials are employed for fitting non-linear effects. Several multivariable approaches have been developed based on this. Examples are Harrell's use of restricted cubic splines (3), or penalized splines as described in (4). On the other hand, the multivariable fractional polynomial (MFP) procedure, proposed in (5), formalizes the approach of systematically testing for deviations from linearity using (global) fractional polynomials for modeling potential non-linear effects, while performing variable selection at the same time.

Before deciding on the technique for selecting variables and functional form, the particular aim of a study raises a fundamental question: What level of complexity is wanted or still deemed acceptable for the selected model (6). For example, when good prediction performance is the only objective, complex models that potentially are too big, i.e., that contain several covariates without true effect, might be advantageous (7). Most of the techniques for variable and function selection offer a tuning parameter for choosing complexity, e.g., the level α employed for variable selection and for testing for linearity in the MFP procedure. Correspondingly, such complexity tuning parameters might have to be set to a relaxed level for optimizing prediction performance. If interpretability is required, it is more important to identify covariates that really have an effect and to avoid erroneously calling a covariate influential that has no effect or at most a weak effect. This consideration might imply an intermediate level for the complexity tuning parameter. Finally, if a model is to be applied, e.g., for future patients, it might pay off to select models with only few covariates, corresponding to a stringent significance level, as each additional covariate increases future measurement costs. Therefore, we broadly distinguish between complexity levels resulting from *relaxed*, *intermediate*, or *stringent*

model selection criteria, each of which might be the most satisfactory, depending on the application.

The level of complexity that can reasonably be chosen may be restricted by the situation at hand. For example, with a smaller number of observations, a relaxed selection level could be problematic, as there might not be enough observations for reliably estimating model components for many covariates. The strategies differ in many aspects and they may have advantages and disadvantages in different scenarios depending on the significance levels for selecting variables and functional form, the sample size or the amount of information in the data. Finally, the choice of a 'best' approach depends on the definition of 'performance'. For example, when wiggly function estimates are to be avoided, more stringent selection of functional form might be preferred, compared with a setting where good prediction performance is the primary aim.

Unfortunately, little is known about the properties of multivariable model-building techniques based on splines, and only limited results are available for MFP (see (8), for example). In (9), some comparison of the two types of approaches is provided, but a comprehensive evaluation is missing. To address this, we perform an extensive simulation study, taking the design from the companion paper (10). We do not expect to identify one approach that performs best in all situations, regardless of the performance measure considered. Such a result would at least critically depend on the specific simulation design employed, as certainly another design could be devised where the identified 'best' approach performs less well. Instead, our aim is to provide guidance for selecting an appropriate technique if a certain level of complexity is wanted, with a focus on biomedical settings with a moderate number of covariates and no interactions (see (9) for a more detailed description of such settings). The results also indicate when a particular complexity level might be problematic, regardless of the technique used.

The techniques to be evaluated, i.e., MFP and two approaches based on splines, are

described in Section 2. The simulation design, as well as the performance measures employed for evaluation, are only briefly given in Section 3, as many details and some background for the necessary choices are provided in a companion paper (10). Section 4 presents results on prediction performance, Type I and Type II error levels, costs at the covariate level, and functional form. Concluding remarks are given in Section 5.

2 Approaches for model selection

Given observations $(y_i, x_i), i = 1, \ldots, n$, with a continuous response y_i and covariates $x_i = (x_{i1}, \ldots, x_{ip})'$, the objective is to fit an additive model

$$y_i = \beta_0 + \sum_{j \in J_{lin}} \beta_j x_{ij} + \sum_{j \in J_{nonlin}} f_j(x_{ij}) + \epsilon_i,$$

with error term $\epsilon_i \sim N(0, \sigma^2)$. Besides estimating the intercept term β_0, the main modeling task is to determine which covariates should enter as linear terms (by definition also including ordinal and dummy-coded categorical covariates that have an effect), i.e., to identify $J_{lin} \subset \{1, \ldots, p\}$, and which of the continuous covariates should enter as non linear terms, i.e., determining $J_{nonlin} \subset \{1, \ldots, p\}$, where in addition the functional form f_j has to be selected. Note that model selection in particular means identifying the set $J_{noe} = \overline{J_{lin} \cup J_{nonlin}}$ of covariates that are not included in the model.

We consider several ways to select variables and functional forms. For modeling the functions $f_j(x)$, evaluation is restricted to methods that can, in principle, provide an explicit model equation. For example, kernel smoothing methods are excluded. For methods that provide model equations, the functions $f_i(x)$ can typically be written as some type of expression involving polynomials. We distinguish between *global* functions, for example those with global polynomial terms, or *local* models that use local polynomial segments, as in splines.

As a global polynomial technique, we consider MFP. MFP formalizes and extends traditional model-building in which polynomial terms are added 'manually' to a regression model according to the value of some test statistic. (A key difference is that MFP starts from a model of predefined complexity and attempts to simplify it according to a type of backward elimination procedure which preserves the familywise error rate.) Many model-building techniques based on splines have been suggested, but there is no obvious 'best' choice; see (4) for a comprehensive overview. As a criterion, we decided to adopt what would be available to a typical user in the R (11) statistical environment. We therefore consider the penalized spline technique offered there as a default. As a second approach, we consider restricted cubic splines as popularized by (3) and provided by the widely used R package `Design` (12). Wherever there are choices to be made with respect to details such as tuning parameters, we either take the default provided by the implementation, or we attempt to imitate what a typical user would probably do, based on the documentation available. This is especially challenging when variable selection is required for the spline models, because no particular method has been widely recommended. To address this issue, we perform variable selection for spline functions like a typical user would do it manually. An overview of the variants of the model-building techniques we use is given in Table 1. For two of them, the significance level is a key tuning parameter which determines the complexity of the selected model and it must be prespecified by the user. We choose 0.01, 0.05 and 0.157 as stringent, intermediate and relaxed selection criteria, respectively. The last of these roughly imitates selection according to a minimal AIC criterion (6). A brief description follows.

2.1 MFP

MFP uses (fractional) polynomial terms for non-linear effects. It was initially developed in a univariate context with a less than ideal approach to multivariable model selection

Table 1: Techniques for selecting variables and functional form, to be considered in the simulation study.

Underlying technique	Selection criterion		
	stringent	intermediate	relaxed
MFP approach	mfp.01	mfp.05	mfp.157
Restricted cubic splines	rcs.01	rcs.05	rcs.157
Penalized splines	-	-	gamm.step

(13). An improved algorithm for the selection of variables and functional forms in a multivariable context was described by (5). For a comprehensive overview, see (8).

First we consider a single continuous covariate x, where the straight line model, $\beta_1 x$ (for simplicity, we suppress the constant term, β_0) is a suitable starting point. Often it is an adequate description of the relationship, but other models must be investigated for possible improvements in fit. A simple extension of the straight line is a power transformation model, $\beta_1 x^p$. Royston and Altman (13) formalized the model slightly and called it a first-degree fractional polynomial or FP1 function. The power p is chosen from a pragmatically chosen restricted set $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where x^0 denotes $\log x$.

As with polynomial regression, an extension from one-term FP1 functions to the more complex and flexible two-term FP2 functions follows immediately. Instead of $\beta_1 x^1 + \beta_2 x^2$, FP2 functions with powers (p_1, p_2) are defined as $\beta_1 x^{p_1} + \beta_2 x^{p_2}$ with p_1 and p_2 taken from S. If $p_1 = p_2$, Royston and Altman (13) proposed $\beta_1 x^{p_1} + \beta_2 x^{p_1} \log x$, a so-called repeated-powers FP2 model.

With the set S of powers as just given, there are 8 FP1 transformations, 28 FP2 transformations with distinct powers ($p_1 \neq p_2$) and 8 FP2 transformations with equal powers ($p_1 = p_2$). The best fit among the combinations of powers from S is defined as that with the highest likelihood.

Choosing the best FP1 or FP2 function by mininizing the deviance (minus twice the

maximized log likelihood) is straightforward. However, having a sensible default function is important for increasing the parsimony, stability and general usefulness of selected functions.

For our situation the linear function is a natural choice. Therefore, unless the data support a more complex FP function, a straight line model is chosen. As a function selection procedure (FSP) a closed test procedure with 3 steps is proposed. First, it investigates whether the continuous variable has any influence on the outcome (comparing the best FP2 model with the null model at a chosen significance level). If significant, it tests FP2 versus a straight line model, and, if also significant, a final step compares the best FP2 and FP1 models. For more details see (8).

In many studies, a larger number of predictors is available and the aim is to derive an interpretable multivariable model which captures the important features of the data: the stronger predictors are included and plausible functional forms are found for continuous variables.

As a pragmatic strategy for building such models, a systematic search for possible non-linearity (provided by the FSP) is added to a backward elimination (BE) procedure. The extension is feasible with any type of regression model to which BE is applicable. Sauerbrei and Royston (5) called it the multivariable fractional polynomial (MFP) procedure, or simply MFP.

The nominal significance level is the main tuning parameter required by MFP. Actually, two significance levels are needed: α_1 for selecting variables with BE, and α_2 for comparing the fit of functions within the FSP. Often, $\alpha_1 = \alpha_2$ is a good choice. The extension of FPm for degree greater than 2 ($m > 2$) is obvious, but rarely if ever needed in a multivariable context and will not be considered here. Since the model is derived data-dependently, parameter estimates are likely to be somewhat biased.

2.2 Restricted cubic splines

In methods based on splines, the function $f_j(x)$ for a covariate x can be expressed as a sum of M basis functions

$$f_j(x) = \sum_{k=1}^{M} \beta_{jk} B_{jk}(x),$$

where the basis functions $B_{jk}(x)$ depend on the particular spline used and are fixed. The parameters β_k are estimated from the data. Truncated polynomial terms are often used for the basis functions, or the latter can be re-expressed in such a form. Typically, cubic terms $(x-\xi)_+^3$ are used, where $()_+$ returns the positive part of its argument or 0. The basis functions are parametrized via the positions ξ of knots that cover the range of the covariate values. Restricted cubic splines (RCS) (14) are parameterized such that the functions are linear in the tails of the distribution.

Specifically, we consider the approach provided in the R package `Design` (12), which has been illustrated in (3). As a starting point, a model containing all covariates is fitted, with an RCS component for each continuous covariate. Five knots are used in the RCS, located at corresponding quantiles of the covariate values (the default in the R implementation).

Starting from this model, backward elimination is performed, using some suitable significance level as stopping criterion. This procedure is implemented for variable selection and is also described in the software documentation (15). For the continuous covariates remaining in the model after backward elimination, a test is applied to decide whether to replace the spline component with a linear function. This is performed in a stepwise manner. The spline component with the largest p-value is the first to be checked for replacement by a linear function. The model is then refitted and the next spline component is checked. This is similar to the MFP procedure.

2.3 Penalized splines

To avoid overfitting with regression splines, only a small number of basis functions are used. However, the positions of the knots are very influential. While placing the knots at quantiles can be expected to result in reasonable performance, the amount of local structure that can be modeled between two knots is limited. Methods based on penalized splines (PS) avoid such problems by using a large number of knots. Overfitting is avoided by penalized estimation, which discourages complex fits. The complexity is controlled by applying a penalty parameter for each covariate. Satisfactory selection of the penalty parameter is therefore critical. For a comprehensive discussion of penalized splines, see (4).

To represent the functions $f_j(x)$ we use a thin plate regression spline basis (16), the default in the package `mgcv` distributed with R. Fitting is done through a linear mixed model representation, thus allowing simultaneous estimation of regression and penalty parameters (see for example the appendix of (17), or (4; 18) for a more comprehensive discussion). Such an approach has been shown to be competitive in terms of prediction performance and to produce relatively smooth fitted functions (19). Nevertheless, graphing the function is the main way to present results, as it is infeasible to express the spline function as a simple formula.

To select variables and functional forms, we adapt the approach of (20). Simplified complexity levels are defined for each covariate. Starting from a model that contains linear components for all covariates, stepwise selection of complexity is carried out, guided by the marginal AIC. For binary and categorical covariates, we consider the complexity levels 'not included' and 'included'. For continuous covariates, three complexity levels are considered: 'not included', 'included as a linear term', and 'included as a non-linear term'. In each step, the complexity of the component for a given covariate is either decreased or increased, such that the marginal AIC decreases the most. The procedure

continues until no further improvement is possible. For more details see (21).

Model selection by AIC roughly corresponds to selection at a signficance level of $\alpha = 0.157$ (6), i.e., to a relaxed model selection criterion. In principle, AIC could for example be replaced by BIC to obtain a different level of complexity. However, there is no theory underpinning model selection by such other criteria, in the framework of the penalized spline representation based on the mixed model. For the PS approach, therefore, we evaluate only models corresponding to selection at relaxed complexity levels.

3 Simulation design

3.1 True structure

We use the simulation design described in detail in the companion paper (10). For $n \in \{200, 500, 1000\}$ observations, 15 underlying variables are generated from a standard normal distribution, with correlation structure given in Figure 1. Covariates are constructed from these underlying variables. Skewness is introduced for continuous covariates and categorical covariates are obtained by categorizing their underlying continuous variables. This results in a total of 17 covariates. For the eight covariates that have an effect on the response, x_1, x_3, x_{4a}, x_5, x_6, x_8, x_{10}, and x_{11}, the related component in the true regression model is given in Figure 1. The functions for the continuous covariates can be seen in Figure 1 of the companion paper (10), or also in Figures 5 and 6.

To obtain a continuous response variable, normally distributed error terms $\epsilon_i \sim N(0, \sigma^2)$ are added to the linear predictors. We take $\sigma^2 \in \{3.47, 0.868, 0.217, 0.096\}$, corresponding to a signal-to-noise ratio $\in \{0.25, 1, 4, 9\}$, and an explained variation of $R^2 \in \{0.2, 0.5, 0.8, 0.9\}$ for the true model. We number the scenarios from 1 to 12, sorting

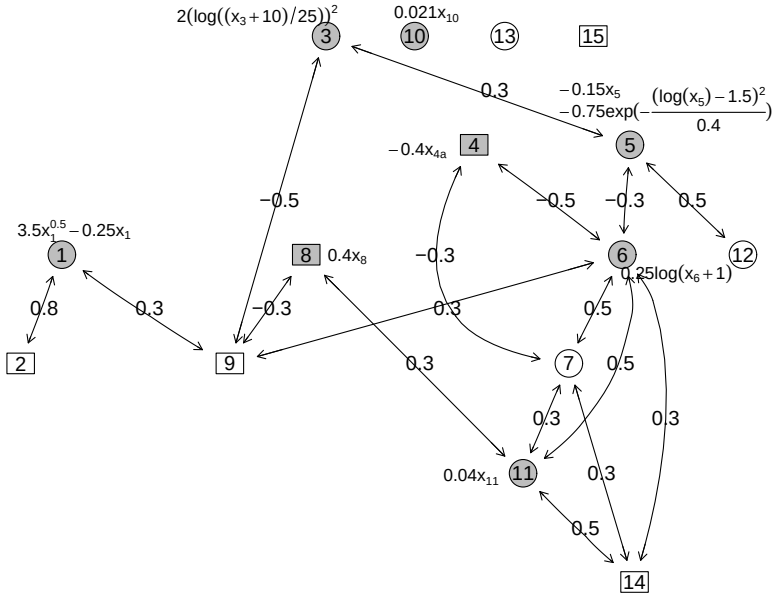

Figure 1: Simulation design: Correlation structure of the underlying variables is indicated by arrows. If the covariate built from an underlying variable is continuous, the latter is indicated by a circle, otherwise by a rectangle. If the covariate has an effect on the response, the circle/rectangle has grey shading, and the corresponding model component is given besides the symbol. Note that some of the underlying variables correspond to several variables/model components, e.g., variable 4 corresponds to x_{4a} and x_4b, where only the former has a non-zero effect.

them by R^2 and by the number of observations within each level of R^2. The smallest amount of information is present in Scenario 1, with $R^2 = 0.2$ and $n = 100$. In such a scenario, multivariable model-building with selection of functional form is barely feasible. From Scenario 5 ($R^2 = 0.5$, $n = 500$) on, an intermediate amount of information is available; here, differences in multivariable model-building strategies might become apparent.

We perform 500 repetitions for each simulation scenario. When convergence problems occur (mostly for $n = 200$ and never for $n = 1000$), only those repetitions are considered where results for all approaches are available. This problem affects less than 1% of the repetitions in any scenario.

3.2 Quantifying performance

For quantifying prediction performance, we consider the prediction error

$$\text{PE}(\hat{y}) = E[(y - \hat{y})^2]$$

for predictions $\hat{y}$ on new data. This is evaluated in $n_{new} = 5000$ newly drawn observations.

For judging the selected models, inclusion vs. exclusion is considered for a given covariate. Depending on whether the covariate does or does not truly have an effect, this characteristic relates to the power or Type I error rate, respectively. For combining these two measures, power and Type I error rate, costs for erroneous inclusion/exclusion of covariates are assigned and summed for each selected model. As described in (10), the definition of "cost" for a covariate is based on the absolute value of the marginal correlation of the corresponding true model component with the response, as determined on a test set of size $n = 5000$. Erroneously excluding a covariate costs more than erroneously

Table 2: Costs assigned for erroneous inclusion/exclusion of covariates. Smallest (for exclusion) respectively largest (for inclusion) values in bold.

cost for erroneous exclusion								
x_1	x_3	x_5	x_6	x_{10}	x_{11}	x_{4a}	x_8	
0.152	0.179	0.342	0.539	0.119	0.384	0.170	**0.093**	
cost for erroneous inclusion								
x_7	x_{12}	x_{13}	x_2	x_{4b}	x_{9a}	x_{9b}	x_{14}	x_{15}
0	0.021	0.082	0.052	0.029	**0.093**	0.079	0.001	0.086

including one. The costs depend on the correlation structure and are given in Table 2.

For evaluating the shape of the selected functions, the mean squared difference of the first derivatives

$$\text{PED}_1(\hat{f}; f) = E[(\hat{f}(x) - f(x))^2]$$

is evaluated for each function, again on $n_{new} = 5000$ new observations. Here $\hat{f}$ is the selected function and f is the true function. This measure favors functions that have a shape similar to the true function and penalizes excessive 'wiggliness' (22).

In addition to this quantitative measure for evaluating the selected functions, we also use qualitative criteria, enabling us to investigate Type II errors with respect to identification of functional form. The criteria applied in our simulation study are given in Table 3. For details see (10).

4 Results

4.1 Prediction performance

The prediction error for all of the model-building methods and the different complexity levels is shown in Figure 2 for various scenarios. For each scenario, the prediction error from the unknown true model (true error variance) is used as the minimum on the y-

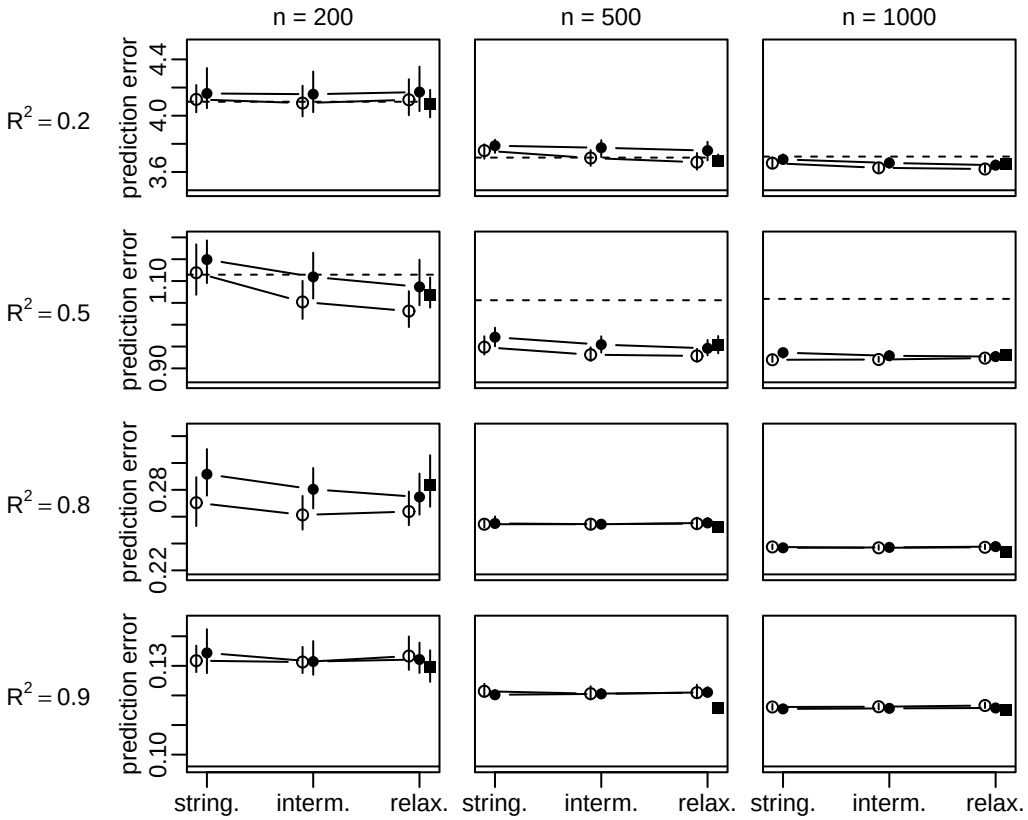

Figure 2: Prediction error (median and interquartile range) for MFP (hollow circles), RCS (filled circles) and PS (filled squares) at different complexity levels in scenarios with different numbers of observations n and R^2. The minimal value of the y-axis is the true error variance (solid horizontal line). The median for a 'full linear' model is indicated by a dashed horizontal line. For $R^2 \in \{0.8, 0.9\}$ values are larger than the y-axis. For details see text.

Table 3: Properties of the true functions that have to be met by a selected function for not being considered a Type II error.

covariate	monotonicity	extrema	slope	inflection point
x_{i1}	-	exactly one maximum between 45 and 55	decreasing	-
x_{i3}	strictly increasing for > 20	exactly one minimum between 12 and 20	-	not more than one, for > 20
x_{i5}	strictly decreasing for < 3 and > 9	-	absolute value for 7 smaller than for 2 and 10	exactly one, between 3 and 9
x_{i6}	strictly increasing	-	decreasing	-
x_{i10}	strictly increasing	-	constant	-
x_{i11}	strictly increasing	-	constant	-

axis. The median prediction error from a model including all 17 covariates, assuming linearity for continuous variables (we call it a 'full linear' model), is shown as a horizontal dashed line. Whereas the former is a lower bound for prediction error, the latter may be considered an upper bound in our situation. For the two larger R^2 values it is much larger than the chosen maximum value of the scale and therefore not visible. The prediction error from the 'full linear' model would be located at 0.419, 0.410, and 0.376 for $R^2 = 0.8$, and at 0.288, 0.270, and 0.263 for $R^2 = 0.9$, for the scenarios with $n = 200$, $n = 500$, and $n = 1000$ observations respectively.

For $R^2 = 0.2$, all approaches have values similar to the full linear model, irrespective of the complexity level. With increasing sample size, prediction error decreases from about 4.15 for n=200 to about 3.7. Regarding prediction error, nothing seems to be lost by excluding variables and nothing seems to be gained by permitting non-linear functions.

When there is little information in the data, e.g., Scenario 1, the PS approach performs slightly better, also exhibiting the least variability. As more information becomes available, PS is outperformed by both RCS and MFP. The performance of MFP is consistently better or at least as good as that of RCS. For small sample size and $R^2 = 0.5$ or $R^2 = 0.8$, the prediction error is much smaller, but for a large amount of information,

e.g., $R^2 = 0.8$ and $n = 500$, the differences are negligible compared with the gains with respect to the full linear model. For a very large signal-to-noise ratio and a large number of observations (e.g., $R^2 = 0.8, n = 500$ or $R^2 = 0.9, n = 500$), the PS approach again has a slightly better prediction performance.

4.2 Type I error rates and power

For MFP and RCS, different complexity levels are realized by varying the nominal significance level $\alpha \in \{0.01, 0.05, 0.157\}$. The actual Type I error levels, calculated across all covariates that have no effect — some are correlated and some are uncorrelated — are shown in Figure 3. Also shown is the power, i.e., the proportion of covariates correctly identified as having an effect. See Table 4 for more details of individual variables in a scenario with an intermediate amount of information (Scenario 5). As expected, choosing a larger significance level results in larger Type I error rates for all approaches.

Except for a smaller amount of information, nominal and actual significance level agree well for MFP whereas the actual significance level is often smaller than the nominal level for RCS. PS has the largest Type I error rate (between 16% and 20%) in all scenarios. However, the power of PS dominates that of RCS in nearly all scenarios, indicative of the general trade-off between Type I error and power. Overall, the power of PS is similar to that of MFP with $\alpha = 0.157$ in most scenarios. For a larger amount of information ($R^2 = 0.8$ or 0.9 and $n \geq 500$; $n = 1000$ and $R^2 \geq 0.5$), the Type II error is negligible for all approaches and all nominal significance levels.

Table 4 shows the Type I error rates separately for each uninfluential covariate in a scenario with an intermediate amount of information ($R^2 = 0.5$, $n = 500$). For all approaches, the rates vary widely across covariates, with lower values mostly for the two uncorrelated covariates x_{13} and x_{15}. In principle, estimates for these two covariates may be considered to provide the only 'true' Type I error estimates, as uninfluential correlated

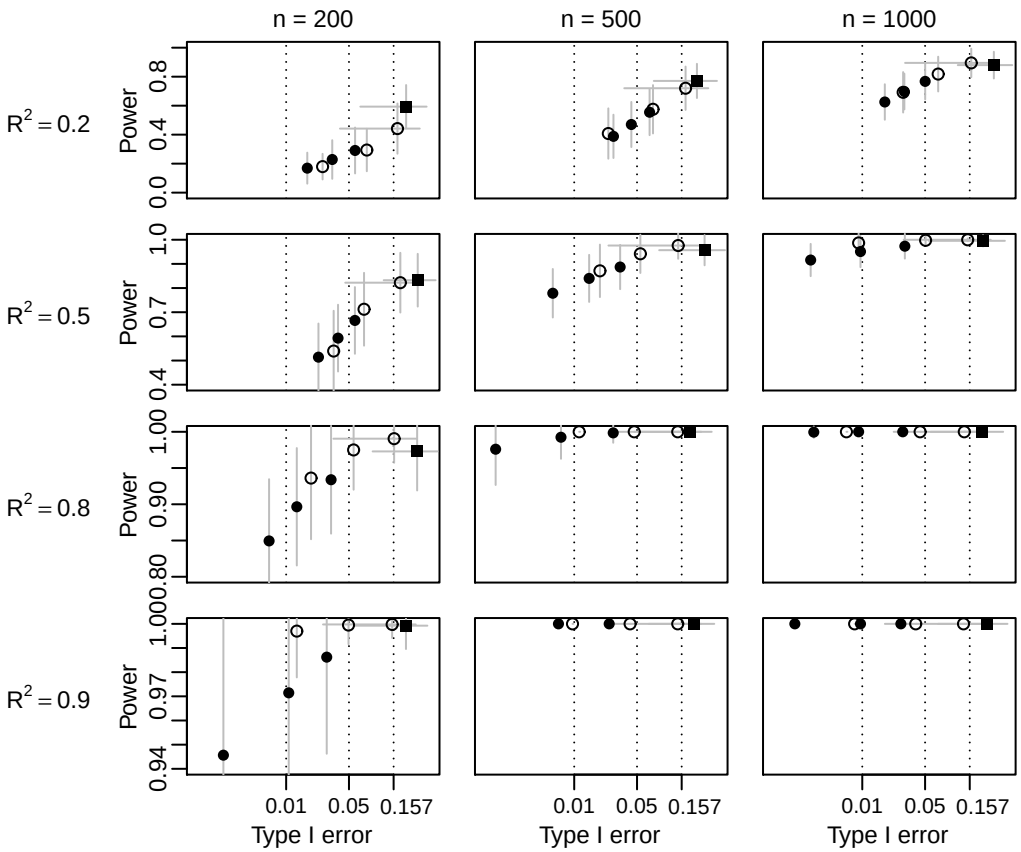

Figure 3: Type I error rates and power (mean +/- SD), calculated across all covariates without and with effect respectively, for MFP (hollow circles), RCS (filled circles), and PS (filled squares) at different complexity levels in scenarios with different numbers of observations n and R^2. The x-axis uses a log-scale. Nominal levels of $\alpha \in \{0.01, 0.05, 0.157\}$ are indicted by dashed vertical lines.

Table 4: Type I error rates for the covariates that have no effect (continuous on the left, binary on the right), in a scenario with $n = 500$ observations and $R^2 = 0.5$. Note that x_{13} and x_{15} are the two uncorrelated covariates that have no effect.

	x_7	x_{12}	x_{13}	x_2	x_{4b}	x_{9a}	x_{9b}	x_{14}	x_{15}
mfp.01	0.004	0.006	0.004	0.086	0.014	0.020	0.010	0.018	0.012
rcs.01	0.006	0.002	0.010	0.018	0.002	< 0.001	0.006	0.008	< 0.001
mfp.05	0.038	0.016	0.014	0.098	0.050	0.082	0.082	0.064	0.046
rcs.05	0.016	0.010	0.022	0.024	0.014	0.006	0.012	0.016	0.012
mfp.157	0.126	0.064	0.042	0.174	0.166	0.200	0.202	0.178	0.140
rcs.157	0.038	0.030	0.046	0.044	0.034	0.022	0.022	0.034	0.022
gamm.step	0.254	0.304	0.150	0.240	0.412	0.410	0.438	0.178	0.154

variables may sometimes be selected instead of the correlated influential partner. For MFP, the nominal and the actual significance level for the binary covariate x_{15} agree well, whereas the actual level of the continuous covariate x_{13} is much smaller than the nominal level. In our design, x_2 has the strongest correlation with a covariate that has a large effect (x_1), and MFP selects x_2 in some repetitions instead of x_1. RCS seems to be less prone to such erroneous selection, and Type I errors are always far below the nominal significance level. For PS, the two uncorrelated variables have a Type I error around 15%, but correlated covariates without an effect are often selected. Six of these seven are selected in 24% to 44% of the repetitions.

4.3 Variable selection costs

Figure 4 shows costs assigned to selected models according to the scheme in Table 2, giving a measure that provides a trade-off between erroneous inclusion and exclusion of covariates. It is related to the Type I error/power consideration presented in Figure 3, but also incorporates the correlation structure of the design. A loss in power directly adds to the costs, potentially counterbalancing gains due to a decreased Type I error rate. PS performs best for a very low level of information (Scenario 1) and has costs comparable to those of MFP and RCS with $\alpha = 0.157$ for Scenarios 2 to 6. The large

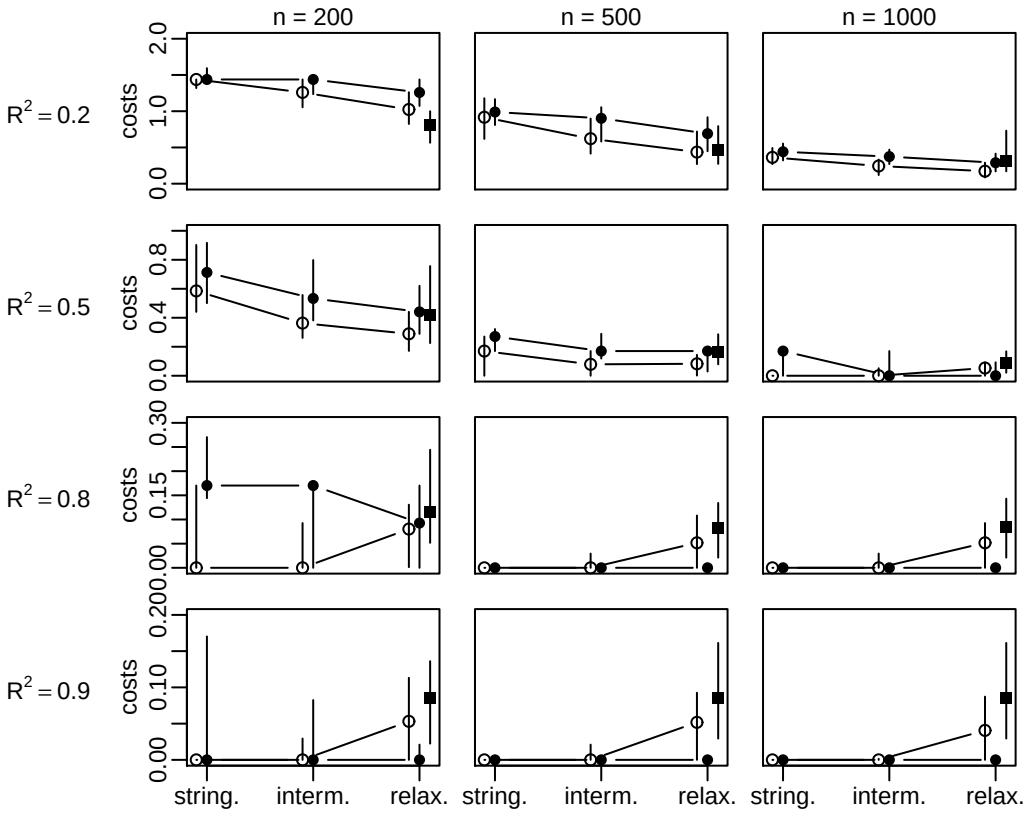

Figure 4: Costs (median and interquartile ranges) incurred by selected models, for MFP (hollow circles), RCS (filled circles), and PS (filled squares) at different complexity levels in scenarios with different numbers of observations n and R^2.

Type I error rates result in the highest costs if $R^2 = 0.8$ or larger.

Increasing the significance level for MFP or RCS decreases the costs in all scenarios with a lower level of information (Scenarios 1 to 5). The costs for MFP are always lower than the cost for the corresponding RCS model.

For a stringent and intermediate significance level ($\alpha \leq 0.05$), MFP always costs little in scenarios with much information ($R^2 \geq 0.8$). Costs of RCS are higher in two scenarios with an intermediate amount of information (Scenarios 6 and 7), but are similar otherwise. With the relaxed significance level and much information, the Type I error increases for MFP and dominates the cost calculations. The Type I error is less serious for RCS and the corresponding costs are much lower.

4.4 Functional form

Figure 6 shows selected functions for a random set of 20 repetitions from the scenario with $R^2 = 0.5$ and $n = 500$ observations, i.e., a medium amount of information. We show functions for those continuous covariates that have linear (x_{10} and x_{11}) or no effect (x_{12} and x_{13}). For each type of covariate, one pair is uncorrelated (x_{10} and x_{13}) and the other (x_{11} and x_{12}) is correlated with other covariates that have a non-linear effect. As PS selects models at a more relaxed level, we also show graphs for the two other approaches using 0.157 as the significance level. Figures with the more typical level $\alpha = 0.05$ are given in the web appendix (with some comments provided below).

Ideally, uninfluential covariates would be excluded and covariates with a linear effect would be assigned a linear term in a model. However, even in the random set of 20 repetitions, there are several instances where a non-linear functional form has been selected by all the approaches. This is due to the large selection level ($\alpha \approx 0.157$). As expected due to its local character, RCS selects functions that fluctuate randomly

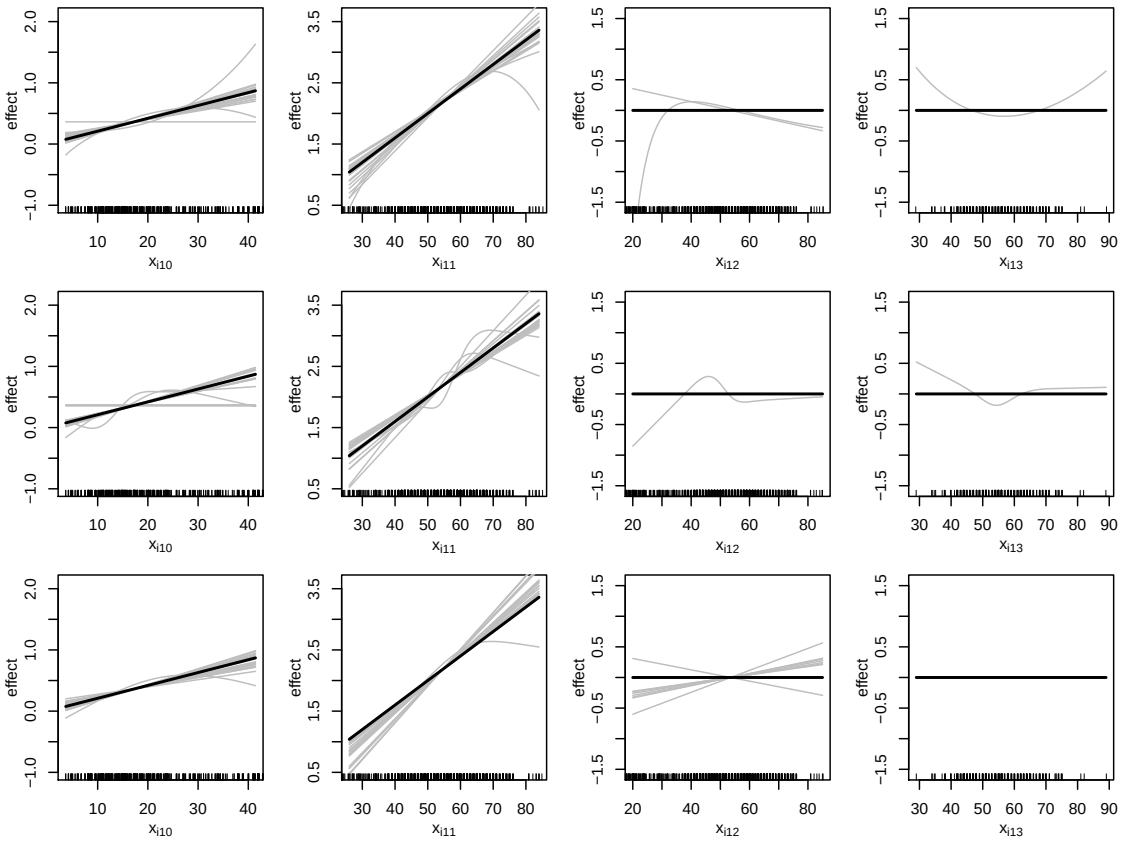

Figure 5: True functions for the four covariates with linear (left two columns) or no effect (right two columns) and fitted functions (top: mfp.157; middle: rcs.157; bottom: gamm.step) for a random sample of 20 repetitions for $R^2 = 0.5$ and $n = 500$.

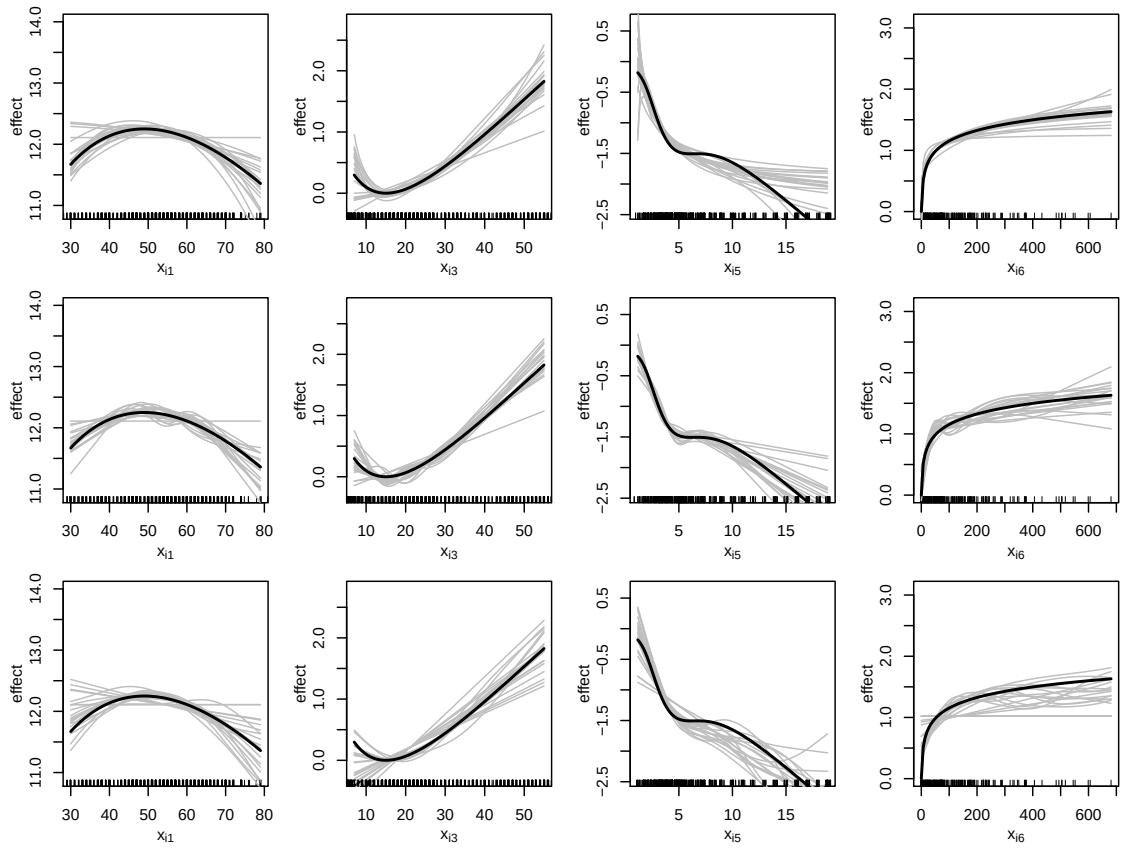

Figure 6: True functions for the four covariates with non-linear effects and fitted functions (top: mfp.157; middle: rcs.157; bottom: gamm.step) for a random sample of 20 repetitions for $R^2 = 0.5$ and $n = 500$.

around the true functions. In contrast, MFP selects functions that are still smooth but which deviate with respect to global shape. A similar tendency is seen for PS; it selects rather smooth functions, despite being based on local polynomials. For covariate x_{12} (no effect), PS often selects a linear function, because of the large Type I error rate in Table 4.

Figure 6 shows selected functions for the covariates with a non-linear true function. While RCS sometimes results in functions with local minima, e.g., around value 45 for x_1, the other two approaches provide smoother functions. In several cases a linear function is erroneously selected for x_1 and x_3. For MFP, this is caused by insufficient

power to select a non-linear function, but PS misses the non-linear parts even more often, which might also be due to low power. For x_5, a very challenging function for a non-local approach, large deviations from the true function are seen for MFP. The global character of the function results in major differences for very low and very high values. The other two approaches come closer, but at the price of larger variability. A detrimental effect of the local flexibility offered by splines is seen for the true function with logarithmic shape. While MFP fits the data well in most replications and always preserves the monotonic structure, the functions selected by the other two approaches exhibit local minima or decreasing slope in several instances. Using MFP and the intermediate significance level (see Figures 1 and 2 the web appendix) gives improved results for the covariates without an effect and the two linear functions, but slightly worse results for the true non-linear functions. The function selection procedure in MFP more often selects linear functions. In contrast, RCS sometimes selects even more wiggly functions with the intermediate significance level.

More thorough investigation of the shapes of the selected functions based on example plots is difficult. For more closely inspecting the shape of the functions, we consider the first derivatives, following (22). Figure 7 shows the first derivatives of the selected functions for some of the continuous covariates in 20 repetitions where all approaches selected a non-zero effect for the considered covariates. For MFP, the smoothness of the selected functions is seen to carry over to the first derivatives. For the spline approaches, there is large variability that is not easily seen from the plots of the original functions in Figures 5 and 6. For RCS, the linearity restriction for large and small covariate values seems to be a considerable source of misfitting for all covariates that have a non-linear effect.

For quantifying how close the shape of the selected functions is to that of the true functions, we consider the mean squared difference between the first derivative of the

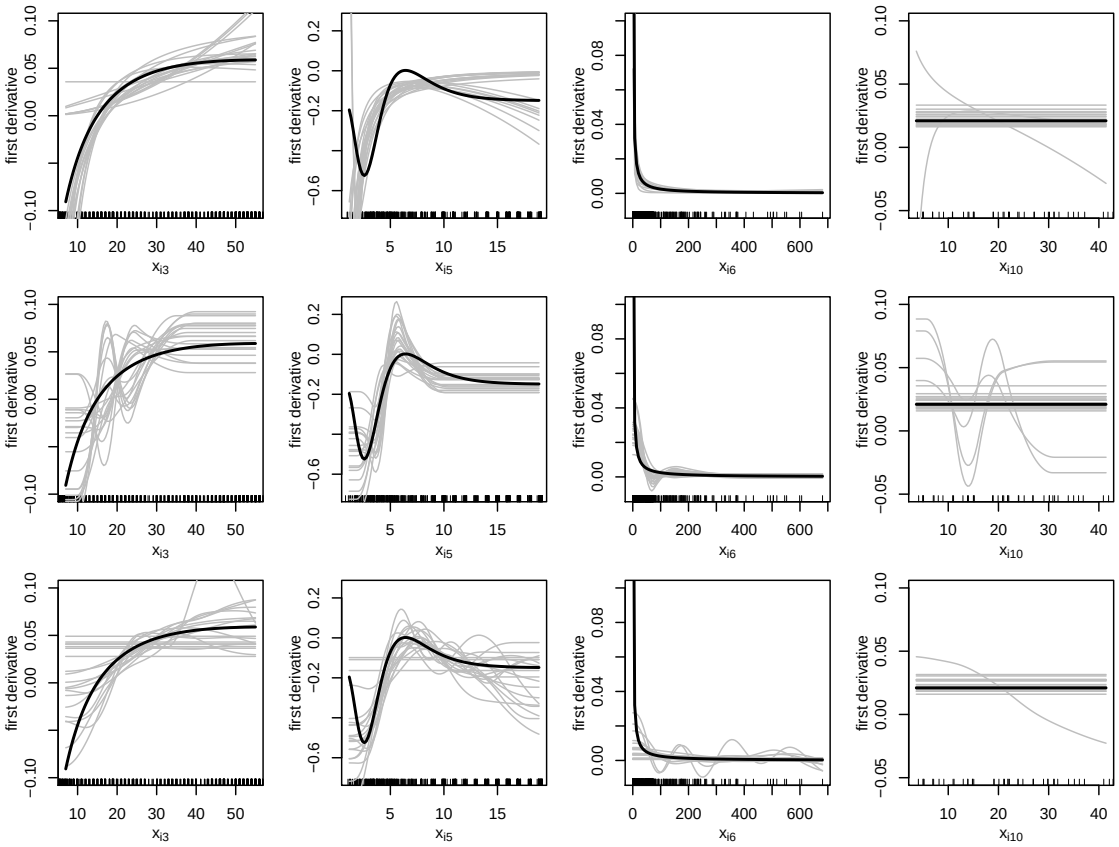

Figure 7: First derivatives of functions (grey curves) selected by MFP (top row), RCS (middle row), and PS (bottom row) for a random sample of 20 repetitions where all of the approaches selected a non-zero effect for each of the shown covariates in the scenario with $R^2 = 0.5$ and $n = 500$. The derivatives of the true functions are indicated by black curves.

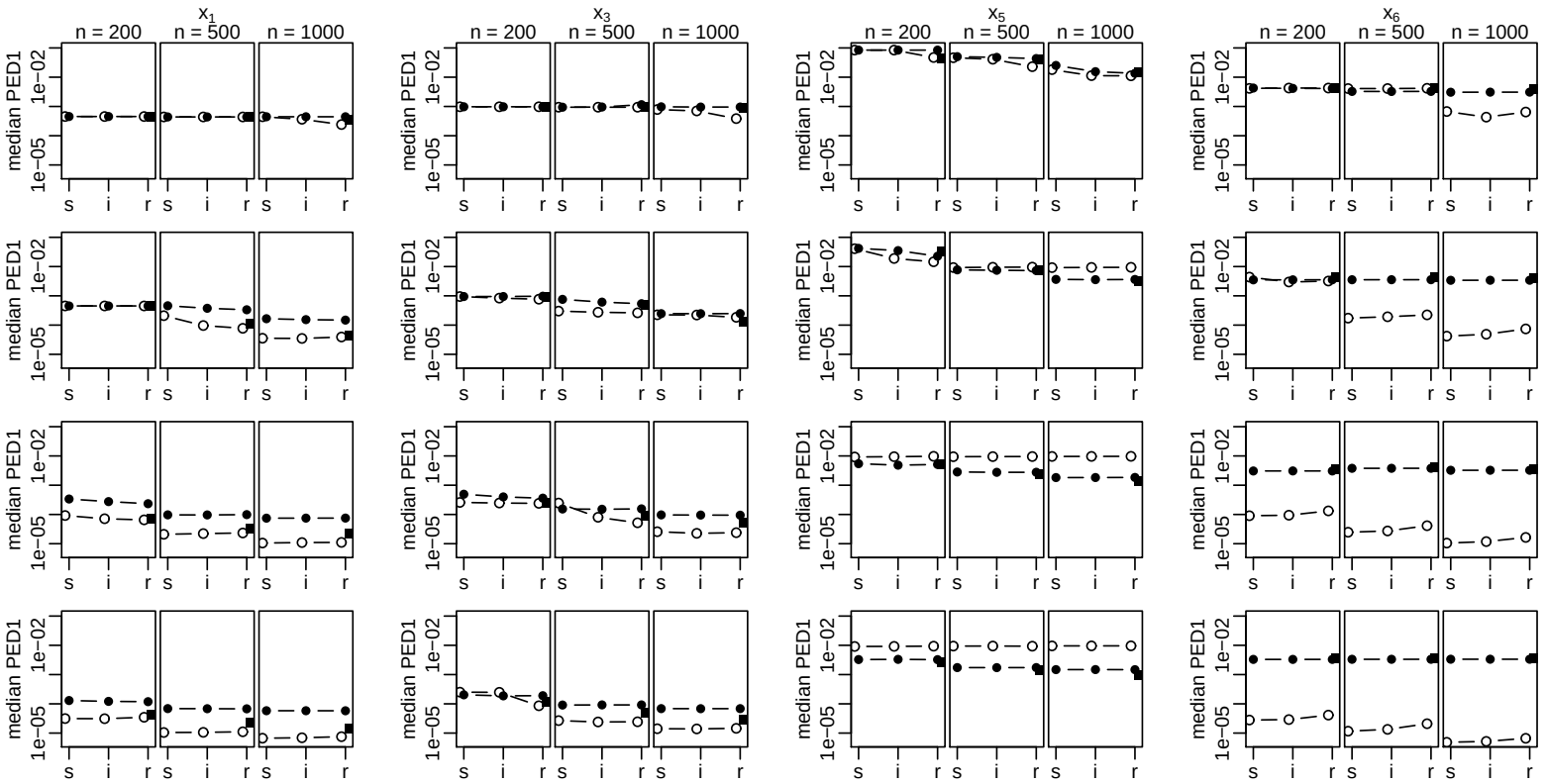

Figure 8: Median of PED_1 for functions selected by MFP (hollow circles), RCS (filled circles), and PS (filled squares) at different complexity levels (stringent: 's'; intermediate: 'i'; relaxed: 'r') in scenarios with different numbers of observations n and different R^2 (0.2: first row; 0.5: second row; 0.8: third row; 0.9: fourth row).

selected and the true function (PED_1). As the true functions are rather smooth, low values indicate less wiggly functions. Figure 8 shows the median value of this measure for the continuous covariates with non-linear effect. For a small R^2 there is little difference between the approaches. The complexity level seems to have little influence on performance with respect to the shape of selected functions in all settings. When more information is available, the performance depends on the shape of the true function. For the true function that has a plateau (third column of plots), MFP is outperformed by the spline approaches. For the other true functions, MFP mostly performs best. An increase of performance with a larger number of observations is seen for all types of true functions, except the function with a logarithmic shape (fourth column of plots). There, only the performance of MFP increases as more observations become available. The spline approaches do not benefit from more information.

Figure 9 shows the proportion of repetitions in which all qualitative criteria are met. Using these criteria as a rough description of the true functional form, the proportions may be interpreted as the probability of selecting the true function. As for the other criteria, the performance depends strongly on the shape of the true function. For x_1, with its quadratic shape, MFP is seen to perform best, with a widening difference as the amount of information in the data increases. The functions selected by the spline approaches rarely meet all criteria for x_1. Concerning the comparison between approaches, the results for x_6 are similar. The main difference is for MFP. In contrast to x_1, the proportion of correct functions hardly increases with increasing R^2. An increase in the selection level results in a slight decrease in the percentage of correct functions. For the functions x_3 and x_5, MFP performs worst. With a large amount of information the functions selected by the spline approaches often meet all the criteria, whereas MFP only performs well when the amount of information is very large. In addition, using a larger significance level improves the results for MFP. PS and RCS also show a definite advantage with respect to x_5. As expected, the functions selected by MFP hardly ever meet

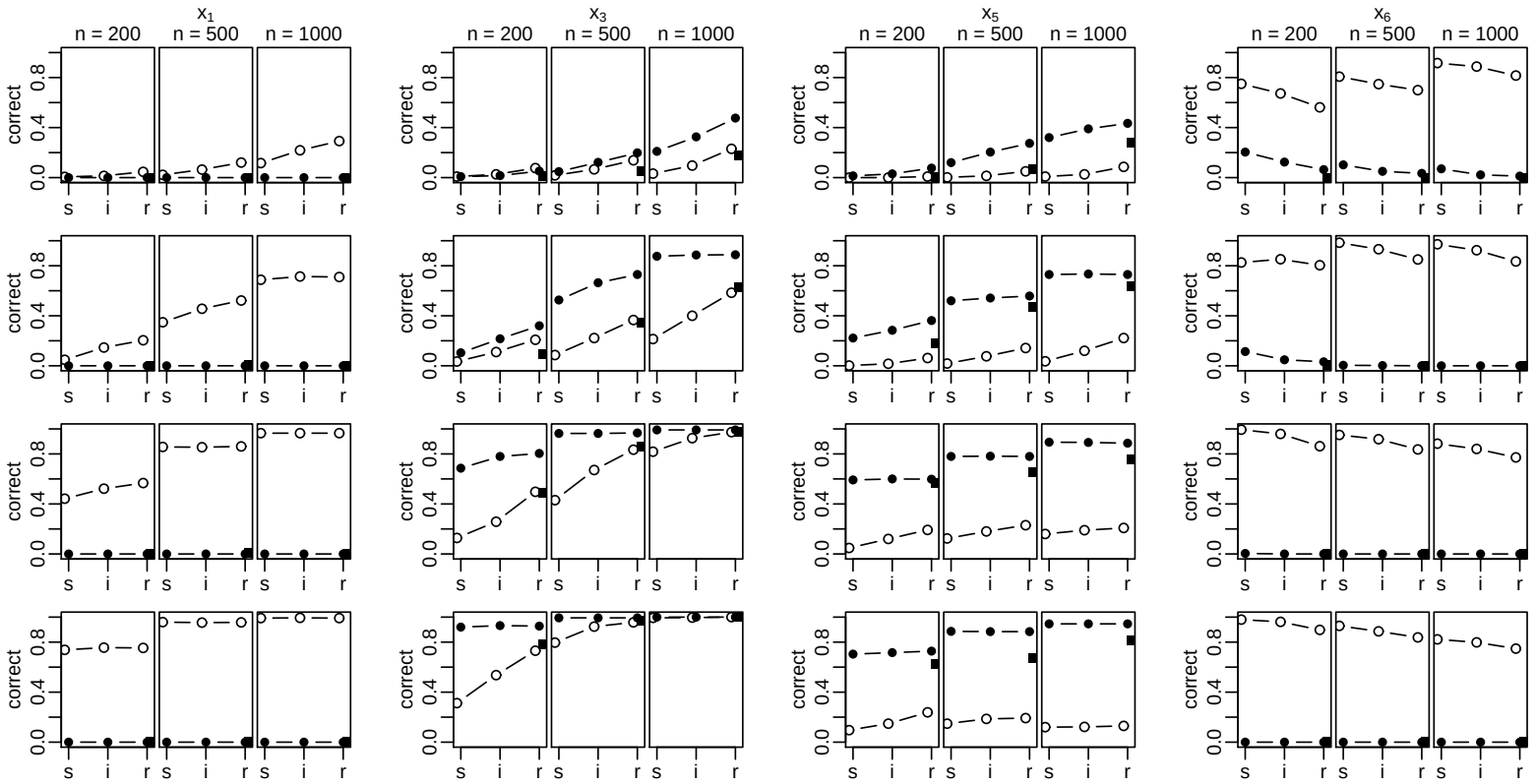

Figure 9: Proportion of repetitions where all qualitative criteria are met (see Table 3) by MFP (hollow circles), RCS (filled circles), and PS (filled squares) at different complexity levels (stringent: 's'; intermediate: 'i'; relaxed: 'r') in scenarios with different numbers of observations n and different R^2 (0.2: first row; 0.5: second row; 0.8: third row; 0.9: fourth row).

all criteria. RCS is better than PS in all scenarios. For low R^2, the significance level has an influence on the performance of RCS, but otherwise the effect seems negligible.

5 Discussion

5.1 General issues in multivariable model building

In developing regression models, data analysts are often faced with many variables that may influence an outcome. If the main aim is to derive a suitable predictor with little consideration of the model structure, the task is relatively easy and the mean square error of prediction is an established and reasonable quantity for comparing the performance of several competing models. However, complexity of the model is often a key issue. Several studies comparing selection strategies in specific examples or simulations have been published in many areas of science (6; 3; 23; 24). There seems to be agreement that no selection procedure does well in small samples (25; 8; 23), and some researchers argue that predictors from more complex models do not perform better in larger samples (6; 23; 24). For continuous variables, these studies usually assume a linear influence on the outcome.

A much more ambitious aim is to derive a suitable explanatory model where it is important to identify influential predictors and for continuous variables to gain insight into the functional relationship with the outcome. Interpretability of the individual components and generalizability are important aims (9). Fractional polynomial functions and various types of splines are the principal competitors for estimating the functional form.

5.2 Issues in comparing splines and FPs in a multivariable setting

Despite enormous practical importance, we are not aware of any simulation study comparing FP and spline modeling in a multivariable context. We think that there are three main reasons that such a task has not so far been addressed. (1) No spline-based procedure for simultaneously selecting variables and functional forms has found wide acceptance. (2) To give sufficient insight into the issues raised, a suitable simulation design must be much more complex than most of the oversimplified designs found in the literature, e.g., a small number of variables or variables uncorrelated. In particular, the choice of true functions is critical in our situation. (3) As the prediction error does not take into account the individual components of a model and also ignores whether a postulated function is smooth or wiggly, it is an insufficient criterion for comparing multivariable explanatory models. Even the concept of a Type II error needs extension if a true and a selected form are to be compared.

To overcome issue (3), we have developed a new measure for judging the shape of an estimated function which penalizes wiggliness (22). In a companion paper (10), we have given details of a simulation design and criteria for comparing multivariable models including functional forms for continuous variables.

Regarding methods based on splines, we have chosen restricted cubic splines (RCS) and penalized splines (PS), and adapted common implementations for a multivariable setting. With these decisions and new developments, we were able to compare multivariable spline and FP modeling with 17 variables. As already mentioned for the case of selecting a predictor, the significance level is also the key tuning parameter for model selection using MFP and RCS. We have chosen the popular 0.05 level and also 0.01 as a more stringent and 0.157 as a more relaxed level. From the simpler case of prediction, it is well known that both the amount of noise in the model and the size of the sample are important (23). Regarding noise, we have chosen a range for R^2 from 0.2 to 0.9. As the aim is

to identify the relevant factors from 17 candidates and to derive a suitable functional form for continuous variables, we decided to use a sample size of $n = 200$ as a lowest level where such complex selections may make sense. With smaller sample sizes it is unrealistic to find any suitable explanatory model. In addition we studied sample sizes of $n = 500$, which we consider as a reasonable choice to select an explanatory model with functional forms from 17 candidate variables, and $n = 1000$, which was chosen to investigate whether and which part of a model can be improved by using a large sample size. In all we investigated 12 scenarios (four levels for R^2, three sample sizes).

5.3 Summary of results from the simulation study

We have considered prediction performance and a cost function as criteria to assess the overall performance of selected models. The cost function was constructed to provide a trade-off between erroneous inclusion and exclusion of covariates, incorporating the correlation structure. For individual variables, we considered Type I error rates, power, agreement with the true functional form with a penalty for wiggliness, and the agreement with certain qualitative criteria which aim to identify important features of the true functional form.

For $R^2 = 0.2$, all methods have a prediction error similar to that of the full linear model, irrespective of the complexity level. Nothing seems to be lost by excluding variables and nothing seems to be gained by permitting non-linear functions to be included. This changes for larger R^2, where the prediction errors of all selection procedures are much smaller. With a medium amount of information (say, scenarios 4 to 7), MFP has smaller prediction error than splines. A larger significance level slightly reduces the prediction error. For a large amount of information (say scenarios 8-12), differences are negligible compared with fitting a linear model, although penalized splines have a slightly smaller prediction error for $R^2 = 0.9$. Models selected with the different approaches and

significance levels vary substantially in the number of variables included, which is not accounted for in the prediction error. To take this issue into account, we defined a cost function as a summary measure to balance Type I and Type II errors and therefore reflect the complexity of models selected. For a medium amount of information, MFP performs best. A more relaxed significance level improves the model, probably because variables with a weaker effect are more often included. With a large amount of information (say, scenarios 6 to 12), the cost functions are dominated by inclusion of variables without an effect. Penalized splines are much worse compared with MFP and RCS. With the two smaller significance levels, inclusion of variables without an effect is hardly critical for MFP and RCS, but for 0.157, MFP includes more variables without an effect than RCS does. Regarding Type I error rate, nominal and actual significance levels agree well for MFP (as for other criteria, scenario 1 is an exception and is implicitly excluded in parts of the discussion), whereas for RCS, the actual significance level is often smaller than the nominal level. Penalized splines have a Type I error rate between 16% and 20% in all scenarios, but they also have a larger power, indicating the usual trade-off between Type I error rate and power. Of course, correlation between variables influences erroneous inclusion and exclusion of variables, and in a stricter sense, the terms Type I error and power are valid only for uncorrelated variables.

Suitability of the functional form is a key criterion with which to judge a selected model. The true functions we chose had very heterogeneous different forms. Due to its local character, RCS often selects functions that fluctuate randomly around the true function, and sometimes selects functions with local minima. In contrast, MFP selects functions that are always smooth but sometimes deviate from the correct global shape. A function with a plateau (the form for x_5) cannot be represented by an FP model. Insufficient power to identify non-linearity is the other critical issue of the FP-based approach. The function selection procedure has a linear function as a default. According to the philosophy of MFP, non-linear functions are only selected if supported by the data,

according to a closed test procedure (8). This results in an obvious price to pay for the intended simpler models. With larger sample sizes, however, MFP often selects a function close to the truth, provided the true function has a global character. The penalized spline method selects smoother functions than RCS. That can be seen from the plots of the first derivatives of the functions and the median of the resulting summary measure, PED1. The proposed qualitative criteria are suitable for assessing how often a function close to the truth is chosen. Naturally, sample size, R^2 and the actual true function have a major influence. In addition, the significance level is important in some cases with a medium amount of information. The spline methods have an advantage in selecting more appropriate functions for x_3 and x_5, whereas selected and true functions for x_1 and x_6 agree more often for MFP.

5.4 Final conclusions

Using extensive experience with selection of variables and functions using fractional polynomials and splines as a starting point (13; 5; 6; 19), we published some recommendations for model-building by selection of variables and functional forms for continuous predictors under some assumptions (9). At that time we had planned to conduct such a simulation study. With a small amount of information (scenario 1), acceptable predictors can be derived with any approach, including the full linear model, but there is no chance of deriving a suitable explanatory model. With a medium amount of information, MFP performs better than RCS and much better than penalized splines on most criteria, with the important exception that a true function like that for x_5 will be seriously mismodeled. This issue should be identified by investigating the residuals, and the function may be improved by adding a local component (26). For a large amount of information, all strategies should select very similar models if outliers and influential points do not harm the model-building process. In the simulation study, we tried to eliminate this

problem by truncation (see companion paper (10)). Besides overall advantages in this simulation study, we consider the relative simplicity in deriving models, interpreting and transporting MFP models and learning how to work with MFP as important components for recommending MFP as being currently the most suitable approach for multivariable model-building with continuous covariates in many scenarios.

Acknowledgements

Harald Binder and Willi Sauerbrei gratefully acknowledge support from Deutsche Forschungsgemeinschaft (SA 580/4-2). Patrick Royston was supported by the UK Medical Research Council (Grant MC_US_A737_0002).

References

[1] Hand DJ. Classifier technology and the illusion of progress. *Statistical Science* 2006; **21**(1):1–14.

[2] de Boor C. *A Practical Guide to Splines (Revised Edition)*. Springer: New York, 2001.

[3] Harrell FE. *Regression Modeling Strategies*. Springer: New York, 2001.

[4] Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press, 2003.

[5] Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: Transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society A* 1999; **162**(1):71–94.

[6] Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *Applied Statistics* 1999; **48**(3):313–329.

[7] Yang Y. Can the strengths of AIC and BIC be shared? a conflict between model indentification and regression estimation. *Biometrika* 2005; **92**(4):937–950.

[8] Royston P, Sauerbrei W. *Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. Wiley: Chichester, UK, 2008.

[9] Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine* 2007; **26**:5512–5528, doi:10.1002/sim.3148.

[10] Binder H, Sauerbrei W, Royston P. Multivariable model-building with continuous covariates: 1. Performance measures and simulation design. FDM-Preprint, University of Freiburg, 2011.

[11] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2009. URL `http://www.R-project.org`, ISBN 3-900051-07-0.

[12] Harrell FE. *Design: Design Package* 2008. URL `http://biostat.mc.vanderbilt.edu/s/Design, http://biostat.mc.vanderbilt.edu/rms`, R package version 2.1-2.

[13] Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics* 1994; **43**(3):429–467.

[14] Stone CJ, Koo CY. Additive splines in statistics. *Proceedings of the Statistical Computing Section ASA*, Washington, DC, USA, 1985.

[15] Alzola C, Harrell FE. *An Introduction to S and the Hmisc and Design Libraries* 2006. URL `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/RS/sintro.pdf`.

[16] Wood SN. Thin plate regression splines. *Journal of the Royal Statistical Society B* 2003; **65**(1):95–114.

[17] Wood SN. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* 2004; **99**(467):673–686.

[18] Wood SN. *Generalized Additive Models. An Introduction with R*. Chapman & Hall/CRC: Boca Raton, 2006.

[19] Binder H, Tutz G. A comparison of methods for the fitting of generalized additive models. *Statistics and Computing* 2008; **18**(1):87–99, doi:10.1007/s11222-007-9040-0.

[20] Wager C, Vaida F, Kauermann G. Model selection for penalized spline smoothing using Akaike information criteria. *Australian & New Zealand Journal of Statistics* 2007; **49**(2):173–190.

[21] Binder H, Sauerbrei W. Stability analysis of an additive spline model for respiratory health data using knot removal. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2009; **58**(5):577–600. Doi: 10.1111/j.1467-9876.2009.00668.x.

[22] Binder H, Sauerbrei W. A new measure for judging the shape of function estimates and penalizing wiggliness 2010. Manuscript.

[23] Raffalovich LE, Deane GD, Armstrong D, Tsao HS. Model selection procedures in social research: Monte-Carlo simulation results. *Journal of Applied Statistics* 2008; **35**(10):1093–1114, doi:10.1080/03081070802203959.

[24] Murtaugh PA. Performance of several variable-selection methods applied to real ecological data. *Ecology Letters* 2009; **12**(10):1061–1068, doi:10.1111/j.1461-0248.2009.01361.x.

[25] Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modeling with logistic regression analysis: In search of a sensible strategy in small data sets. *Medical Decision Making* 2001; **21**(1):45–56.

[26] Binder H, Sauerbrei W. Adding local components to global functions for continuous covariates in multivariable regression modeling. *Statistics in Medicine* 2010; **29**(7-8):808–817, doi:10.1002/sim.3739.