On Two Flexible Methods of 2-Dimensional Regression Analysis

Petr Volf

Technical University of Liberec
Faculty of Science, Humanities and Education
Studentská 2, 461 17, Liberec, Czech Republic

*Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Prague 8, Czech Republic
petr.volf@tul.cz

Abstract

The paper deals with the problem of non-parametric statistical modeling of 2-dimensional surfaces from observed data, i.e. the regression analysis. In general, the model is constructed from a set of basal functions, as are the splines, gaussians and others. However, such modeling means to estimate a large set of parameters (locations of functional units and parameters of their combination). We shall present two approaches allowing reduction of the number of needed parameters. Namely, a well known method of projection pursuit, and the less known method of Gordon surface. Further, we shall analyze possible serious consequences of sparse data to precision of model and uncertainty of prediction. Methods will be illustrated in artificial examples.

Keywords: Statistics; regression analysis; splines; projection pursuit; Gordon surface; prediction error.

Introduction

Though the main concern of the paper is two-dimensional regression analysis, we shall start, in Part 1, with a 1-dimensional case. We think that it is the best way how to show the way of modeling curves from functional units and the problems connected with such an approach. The use of localized units (as B-splines or gaussians) is convenient when we wish to describe a non-regularly varying function (a signal, for instance). However, as we shall see, the use of combination of localized units requires also sufficiently 'localized' (i.e. dense) measurements, to avoid unexpected non-precision of model performance.

This problem will be illustrated first in the case of a 1-dimensional curve, variance of prediction will be computed, and its relationship with measurement design shown. Then, in Part 2, we shall devote to the 2-dimensional regression models. We shall recall some approaches how the number of involved parameters can be reduced. Then, after brief overview of the projection pursuit method, we concentrate, in Part 3, to the model construction via so called Gordon surface. Even here, we shall analyze the relation between data design and model (and prediction) uncertainty.

In Part 1 we shall employ also the MCMC (Markov chain Monte Carlo) procedures in the framework of the Bayes statistical inference. The reason is that MCMC generates a representation of posterior distribution of estimated model. With its aid it is possible to visualize the variability of estimates. Simultaneously, estimated distribution of predicted values is obtained. More details on the MCMC methods can be found elsewhere, for instance in [5] and also in [8].

1 Nonparametric Regression

Let us consider first a pair of one-dimensional random variables X (input variable, predictor) and Y (output variable, response) and a general response model defined by a density $f(y;r(x))$ of conditional probability of Y for given $X = x$, where $r(x)$ is a smooth non-parametrized response function. This definition involves, as a special case, the standard regression model $Y = r(X) + \varepsilon$, where ε is a random Gauss noise with zero mean and an unknown variance σ^2.

There are essentially two different ways how to estimate unknown function r. The first consists in the local (e.g. kernel) smoothing. The other approach, studied here, employs the approximation of $r(x)$ by a combination of functions from some functional basis. For instance radial basis functions (gaussians), polynomial splines, goniometric functions or wavelets are popular choices. Hence, the model of response function has the form

$$r_M(x) = \boldsymbol{\alpha}' \boldsymbol{B}(x;\boldsymbol{\beta}) = \sum_{j=1}^{M} \alpha_j B_j(x;\boldsymbol{\beta}), \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)'$ is a vector of linear parameters, B_j are basis functions and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)'$ is a vector of parameters of the basis functions (e. g. knots of splines, centers and scales of radial functions). While the estimates of $\boldsymbol{\alpha}$ can be obtained directly from linear regression context, estimation of $\boldsymbol{\beta}$ is a difficult optimization problem. As a solution to the nonlinear problem for coefficients $\boldsymbol{\beta}$ as well as to optimal choice of M, it is possible to use the Bayes methodology in combination with the Markov chain Monte Carlo (MCMC) algorithms. In this framework, the parameter $\boldsymbol{\beta}$ is considered to be a multi-dimensional random vector, with a prior distribution satisfying certain constraint. Simultaneously, M is also regarded as a random variable, with some prior on $\{0, 1, 2, \ldots, M_{max}\}$.

1.1 Modeling via B–splines

Polynomial splines are constructed from piece-wise polynomials which are joined together in the 'knots'. At these points, continuity conditions are fulfilled. We mostly deal with the cubic splines which have continuous two first derivatives. There are several variants of functional bases creating the spline, we prefer the B–splines as they are localized. Let us consider an interval $[a,b]$ and a set of M different inner knots, $\beta_0 = a < \beta_1 < \ \ldots\ \beta_M < \beta_{M+1} = b$, let us add six other 'dummy' knots $\beta_{-j} = a - j(\beta_1 - a),\ \beta_{M+1+j} = b + j(b - \beta_m),\ j = 1,2,3$. One way how to define the B–spline function, following for instance [9], employs divided differences:

$$B_j(x,\boldsymbol{\beta}) = \sum_{k=j-2}^{j+2} \left\{ (x-\beta_k)_+^3 \Big/ \prod_{s=j-2, s\neq k}^{j+2} (\beta_k - \beta_s) \right\},$$

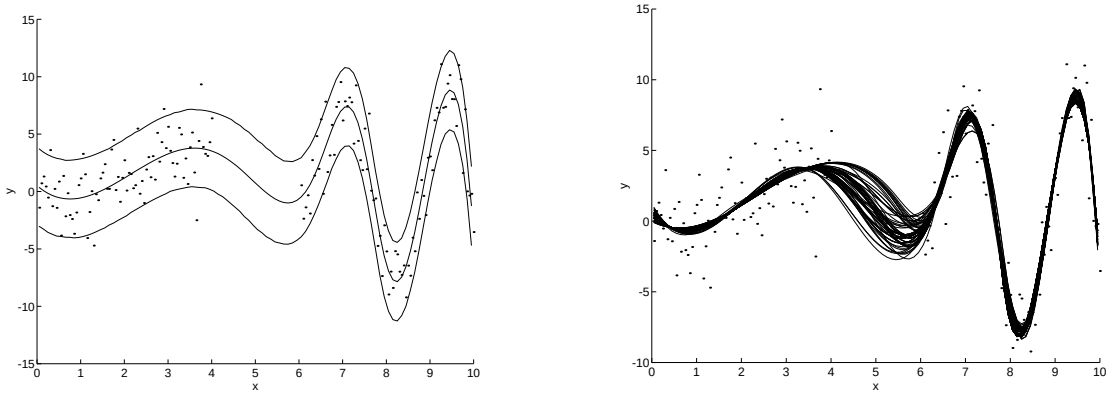
for $j = -1, 0, 1,\ \ldots\ M, M+1, M+2$. Here, function $(u)_+$ means $u \cdot 1[u > 0]$. Thus, M inner knots define $M+4$ basal cubic units. Each unit B_j is zero outside the interval $(\beta_{j-2}, \beta_{j+2})$. The interference of two neighboring units depends on position of the knots, a change of one knot has effect on several nearest units only (e. g. in the case of cubic B–splines, five units have to be updated when one knot is changed). It means that in model (1) a change of one β_k results in updating of only five $\alpha_j,\ j = k-2, k-1, k, k+1, k+2$. This leads to the reduction of necessary computations.

1.2 Optimal Number of Units

It is expected that the model with more units decreases residual variance (or increases the likelihood). Therefore we should examine whether the addition of one unit from corresponding

functional basis improves the fit of the model 'sufficiently'. In a non–Bayes setting, this is often measured by a penalty criterion, e.g. Akaike's AIC, Schwarz's BIC, GCV (see also Friedman [4]). Similar is the criterion $\hat{\sigma}^2_M exp(\frac{M}{N^\gamma})$, where γ is a number from $(0.5,1)$, $\hat{\sigma}^2_M$ is the estimate of residual variance σ^2, M denotes the number of used units, N is the extent of data sample.

Equivalently, we can obtain the penalty as a part of the acceptance probability in the MCMC algorithm. Let us assume that the prior for variable M is specified in such a way that the proportion $Q_0(M^*)/Q_0(M) < 1$ if $M^* > M$. For instance, if prior is proportional to $\exp(-M/N^\gamma)$. Such a choice decreases the chance to accept a model with higher number of units, if the gain of that model (measured by likelihood ratio) is low. The addition of new units can be complementary controlled by a rule guaranteeing a reasonable minimal distance between them and by prescribing maximal number of units.



Source: Own

Fig. 1. *Left: Data, estimated regression curve (central) and $\pm 2\hat{\sigma}$ bands. Right: Variability of MCMC generated posterior representation $r^{(m)}(x)$*

Example 1. 160 uniformly distributed points x_i were sampled in (0,4) ∪ (6,10), and output values $y_i = r(x_i) + \varepsilon_i$ were then generated, where

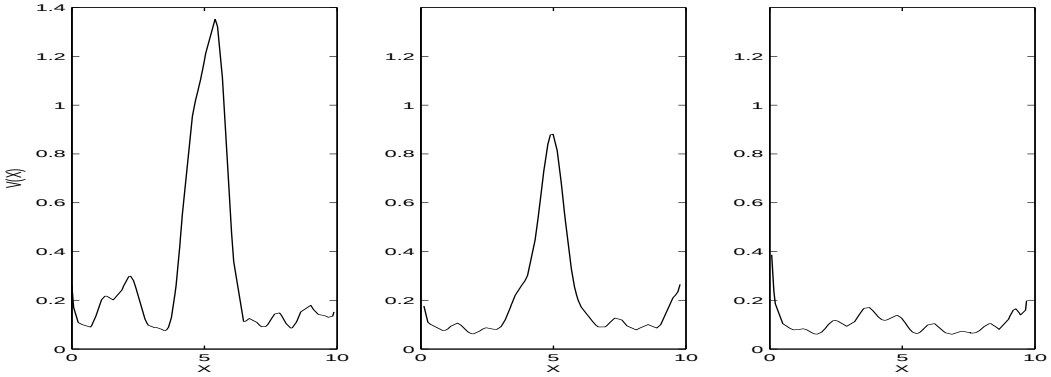
$$r(x) = x \sin\left(\left(\frac{x}{0.25}\right)^2\right)$$

and ε_i were independent identically distributed Gauss random variables with mean zero and variance $\sigma^2 = 4$. For estimation of function $r(x)$, the cubic B-splines were used. As regards the prior for their knots, we used uniform distribution on the set $\{0 < \beta_1 < \beta_2 < \ \ldots \ < \beta_M < 10\}$. 100 loops of the Markov chain generation were performed. One loop updated sequentially all components of $\boldsymbol{\beta}$, with possible change of M. It means that it contains between 20 – 50 iterations of model, depending on actual number M.

Only the final result after each loop was registered as a new member of the chain, $r^{(m)}$. The chain obtained in such a manner has a rather low autocorrelation. The average of this sequence of functions, after skipping first $s = 20$ of them,

$$\hat{r}(x) = \frac{1}{S-s} \sum_{m=s+1}^{S} r^{(m)}(x), \tag{2}$$

serves then as the final estimate of $r(x)$, empirical variance or the quantiles of the set $\{r^{(m)}(x), m = s+1, ..., S\}$ yield information about the uncertainty of the estimate at given x. In the case of Gauss random noise, unknown parameter σ is estimated from the averaged squares of residuals of the preceding iteration. The procedure started from 7 initial units defined by 3 inner knots located equidistantly inside (0,10). It converged to final 13 units, with final estimate $\hat{\sigma}^2 = 4.18$. Figure 1, left plot, shows the points $\{x_i, y_i\}$, the estimate $\hat{r}(x)$ and $\hat{r}(x) \pm 2\hat{\sigma}$ intervals connected to two bands. However, the right plot shows the variability of last 80 members of chain $r^{(m)}(x)$. It can be taken as a representation of posterior distribution of $r(x)$, quite sufficient for illustration of certain important features. Namely, it is seen how the variability of estimate increases in the region with sparse data. It also means that there increases uncertainty of prediction of true values of $r(x)$ (compare also discussion in [2], Ch. 10). A vertical cut at a given x represents Bayes prediction distribution of corresponding $r(x)$.



Source: Own

Fig. 2. *Evaluation of function $V(x)$, in the case without data in (4,6) (left), with one measurement added to $x = 5$ (center), from data distributed uniformly through whole (0,10) (right)*

1.3 Variance of Prediction

In the present part we shall recall some well known results quantifying the variance of prediction in linear regression model, see for instance [1], and adapt them to our case. For simplicity, let us assume that the functional units, i. e. their number and inner parameters, are fixed. Hence, we solve the linear regression case

$$y_i = \boldsymbol{B}^T(x_i) \cdot \boldsymbol{\alpha} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where ε_i are the i.i.d. normal random variables $\mathcal{N}(0, \sigma^2)$, $\boldsymbol{B}(x_i) = (B_1(x_i), \ldots, B_M(x_i))^T$ are functional units (e. g. B-splines) evaluated at data-points $\boldsymbol{x} = (x_1, \ldots, x_n)^T$. Denote $\boldsymbol{B}$ the $n \times M$ matrix with rows $\boldsymbol{B}^T(x_i)$, $\boldsymbol{A} = (\boldsymbol{B}^T \cdot \boldsymbol{B})^{-1}$, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. Then the least squares method yields the estimate

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{A} \cdot \boldsymbol{B}^T \cdot \boldsymbol{y}, \quad \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{O}, \sigma^2 \cdot \boldsymbol{A}),$$

where $\boldsymbol{O}$ is the null vector. Further, at a selected point z the prediction of $y(z)$ is $\hat{y}(z) = \boldsymbol{B}^T(z) \cdot \hat{\boldsymbol{\alpha}}$. Its expectation is $r(z)$, while its variance equals

$$\mathrm{var}\big(\hat{y}(z)\big) = \boldsymbol{B}^T(z) \cdot \boldsymbol{A} \cdot \boldsymbol{B}(z) \cdot \sigma^2.$$

We see that it depends both on data ($\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{x})$) from which the model was estimated, and on position of prediction point z.

On this basis we can study interaction of data design and prediction error, particularly the influence of additional measurements to model precision. For illustration, let us return to Example 1. Let the model be constructed from B-splines defined by 7 inner knots placed equidistantly inside $(0,10)$, so that the model contains $M = 11$ B-spline functions. Figure 2 shows function $V(z) = \boldsymbol{B}^T(z)\boldsymbol{A}\boldsymbol{B}(z)$ in 3 cases. The left plot corresponds to data shown in Figure 1, i. e. without measurements in interval $(4,6)$. The central plot corresponds to the situation when one measurement was added to $x = 5$. Finally, the right-hand plot shows function $V(z)$ computed from data distributed uniformly over whole interval $(0,10)$, without any gap.

2 Multi-Dimensional Case

Now we assume that the input variable $\boldsymbol{X} = (X_1, \ldots, X_p)$ is a p-component vector. In general, the multivariate regression modeling has to deal with functions of interactions of several predictors. Such a function is as a rule modeled by a tensor product of one-dimensional units. The problem with multivariate units is not only that their number grows (exponentially) with dimension, but that there also grows a number of units (and parameters) which are influenced by updating of one component of 'inner' parameter (e. g. of one knot). For instance, a function $r(\boldsymbol{x})$ in R^2 can be modeled as

$$r(\boldsymbol{x}) = \alpha_{00} + \sum_{j=1}^{M_1} \alpha_{j0} B_j(x_1, \boldsymbol{\beta}) + \sum_{k=1}^{M_2} \alpha_{0k} C_k(x_2, \boldsymbol{\gamma}) + \sum\sum \alpha_{jk} B_j(x_1, \boldsymbol{\beta}) C_k(x_2, \boldsymbol{\gamma}). \qquad (3)$$

Such a function contains $M_1 + M_2$ inner parameters β_j, γ_k, but $1 + M_1 + M_2 + M_1 M_2$ 'linear' parameters α_{jk}. The high number of parameters leads naturally to the high time needed for (iterative, as a rule) computations. That is why there are attempts to reduce the dimensionality of 'decision space' of the model construction. Especially the methods based on the idea of decision tree are successful. The most known one is the CART (Classification and Regression Tree), giving a histogram-like result, and its modification giving a continuous function, the MARS (Multi-dimensional Adaptive Regression Splines) [4].

In the simplest scenario the multi-dimensional model has an additive form. The response function $r(\boldsymbol{x})$ is then a sum of p functions of one variable. In our context

$$r(\boldsymbol{x}) = \sum_{k=1}^{p} \sum_{j=1}^{M_k} \alpha_{jk} B_{jk}(x_k, \boldsymbol{\beta}_k) \;=\; \sum_{k=1}^{p} \boldsymbol{\alpha}_k' \boldsymbol{B}_k(x_k, \boldsymbol{\beta}_k),$$

vectors $\boldsymbol{\beta}_k$ and values M_k, $k = 1, \ldots, p$, are optimized iteratively. Most of algorithms innovate sequentially one component after another. Naturally, flexibility of such a model is rather limited.

2.1 Projection Pursuit

This is one of the approaches how to model the multi-dimensional interactions of input variables [7]. The PP estimator has the following form

$$r^*(\boldsymbol{x}) = \sum_{j=1}^{K} s_j(\boldsymbol{\beta}'_j \boldsymbol{x}),$$

i.e. it is the sum of 1-d functions of linear combinations (rotations) of covariates. Now, the objective is to find an optimal K and optimal p-dimensional vectors $\boldsymbol{\beta}_j$. The problem of estimation of non-parametrized functions s_j is then solved in the framework of additive model. The space

of $\boldsymbol{\beta}$'s is limited to such that $|\boldsymbol{\beta}_j| = 1$. Notice also that the rotation in R^p is fully described by an angle having $p-1$ components, each with values in $[0,\pi]$. So that the dimension of nonlinear parameter is actually $p-1$. For instance, in R^2 the angles of rotations can be ordered to a sequence $0 \le \gamma_1 < \gamma_2 < ... < \gamma_K < \pi$, so that the process of solution is quite similar to that dealing with optimal location of knots in R^1. Thus, the problem of construction of 2-dimensional model is changed to two nested 1-dimensional problems.

On the other hand, it is well known that the projection pursuit is highly sensitive, that the dependence on $\boldsymbol{\beta}$'s is rather non-smooth. The method is implemented to several popular statistical software packets (S-plus, R).

2.2 Variance of Prediction in PP

We shall study now the error or prediction in a similar manner as in part 2.4. Again, let us assume that the angles of rotations γ_k, $k = 1,\dots,K$, are selected, and that also functional units $B_{jk}(z)$ are already fixed, $j = 1,2,\dots,M_k$ for k-th projection. Thus, we shall deal with the model

$$y(\boldsymbol{x}) = \sum_{k=1}^{K} \sum_{j=1}^{M_k} \alpha_{jk} B_{jk}(z_k),$$

where we denoted $z_k = \gamma_k \star \boldsymbol{x} = \cos\gamma_k \cdot x_1 + \sin\gamma_k \cdot x_2$ the projection of point $\boldsymbol{x} = (x_1, x_2)$ to direction γ_k. Thus, we again deal with the linear regression scheme

$$y_i = \boldsymbol{B}^T(\boldsymbol{x}_i) \cdot \boldsymbol{\alpha} + \varepsilon_i,$$

where $\boldsymbol{B}(\boldsymbol{x}_i) = \left\{B_{jk}(z_{ki}),\ j = 1,\dots,M_k,\ k = 1,\dots,K\right\}$, $z_{ki} = \gamma_k \star \boldsymbol{x}_i$, $\boldsymbol{\alpha} = \{\alpha_{jk}\}$, its dimension is $M = \sum_{k=1}^{K} M_k$, and $(y_i, \boldsymbol{x}_i)$, $i = 1,\dots,n$, are measurements. Formally the case is the same as the case discussed in 2.3. A numerical illustration is a part of Example 2 in the next section.

3 A Model of Gordon Surface

The Gordon surface [6] is one of constructions of smooth surfaces used mostly in engineering applications, some others from this set are for instance Extrusion Surface, Ruled Surface or Coons Patch. We shall adapt it here to the needs of 2-dimensional statistical regression analysis.

Let us consider a surface S given by a smooth function $y(\boldsymbol{x})$, $\boldsymbol{x} \in X \subset R_2$. We assume that S is given in a form of the Gordon surface, namely: Consider a set of K lines L_k $(k = 1,\dots,K)$ crossing the domain X. Cuts of surface S above lines L_k are smooth functions $y_k(z)$, $z \in L_k$. For each point $\boldsymbol{x} \in X$, let $\boldsymbol{x}(k)$ be its projection to L_k. Further, let $R_k(\boldsymbol{x})$ be a weight of point $\boldsymbol{x}$ w.r. to line L_k. We assume that $R_\ell(\boldsymbol{x}) = 1$ if $\boldsymbol{x} \in L_\ell$ and for points lying between lines L_k the weight is given by a convenient (sufficiently smooth) kernel function, e. g. by a Gauss kernel, with property $\sum_{k=1}^{K} R_k(\boldsymbol{x}) = 1$. Then let us define

$$r(\boldsymbol{x}) = \sum_{k=1}^{K} R_k(\boldsymbol{x}) \cdot r_k(\boldsymbol{x}(k)) \tag{4}$$

as a surface value at $\boldsymbol{x}$. Regarding the functions $r_k(z)$, it is assumed that they are linear combinations of 1-dimensional functional units (e. g. B-splines or radial-basic functions) $\varphi_{km}(z)$

$$r_k(z) = \sum_{m=1}^{M_k} w_{km}\, \varphi_{km}(z) = \boldsymbol{w}'_k \cdot \boldsymbol{\varphi}_k(z).$$

Hence, each such function contains M_k unknown parameters w_{km} and localized units $\varphi_{km}(z)$. The units, as well as blending (mixing) weights $R_k(\boldsymbol{x})$, should be selected by an analyst. It is seen that the model has a similar feature as the projection pursuit, instead of beams given by rotations the core is given by a net of lines, and the projection of data points is weighted. The simplest choice of lines L_k are equidistant vertical (or horizontal) lines crossing X.

3.1 Statistical Model and Estimation

Let us now imagine a set of measurements $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$ of the surface S at points $\boldsymbol{x}_i$ with errors ε_i (i. e. a regression model)

$$y_i = r(\boldsymbol{x}_i) + \varepsilon_i,$$

ε_i are centered independent Gauss variables with $\operatorname{var}\varepsilon_i = \sigma^2$. The objective of statistical analysis is to estimate unknown parameters w_{km}, compute the variance of estimates and also of prediction $r(\boldsymbol{x})$ at a point $\boldsymbol{x}$, based on these estimates. We shall use the following approach: First, functions r_k will be estimated. Then $\widehat{r}_k(z)$ will be propagated to whole X by blending given in expression (4).

Again, denote by $\boldsymbol{x}_i(k)$ projections of points $\boldsymbol{x}_i$ to lines L_k. Let P_{ki} be again the weight of $\boldsymbol{x}_i$ w.r. to line L_k. Further, denote $\boldsymbol{y} = (y_1, \ldots, y_n)'$, for each k denote Φ_k a matrix $n \times M_k$ with elements $\varphi_{mk}(\boldsymbol{x}_i(k))$, P_k the $\operatorname{diag}\{P_{ki}\}$ matrix $n \times n$ and G_k $\operatorname{diag}\left\{(P_{ki})^{-1/2}\right\}$ matrix. Let us eventually omit indices i such that $P_{ki} = 0$. A natural estimator of parameters $\boldsymbol{w}_k$ is then the result of weighted least squares method,

$$\widehat{\boldsymbol{w}}_k = \left(\Phi_k' P_k \Phi_k\right)^{-1} \Phi_k' P_k \boldsymbol{y},$$

corresponding to a regression model $\boldsymbol{y} = \Phi_k \boldsymbol{w}_k + G_k \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. It follows that $\widehat{\boldsymbol{w}}_k = \boldsymbol{w}_k + \boldsymbol{A}_k \cdot \Phi_k' P_k G_k \boldsymbol{\varepsilon}$, with $\boldsymbol{A}_k = \left(\Phi_k' P_k \Phi_k\right)^{-1}$, therefore $\mathrm{E}\,\widehat{\boldsymbol{w}}_k = \boldsymbol{w}_k$ and

$$\operatorname{var}(\widehat{\boldsymbol{w}}_k) = \boldsymbol{A}_k \cdot \sigma^2.$$

In general, we can consider one set of weights, $R_k(\boldsymbol{x})$, for blending functions y_k, and a different set, $P_k(\boldsymbol{x})$, as weights used in the estimation procedure. Naturally, it is possible to set $P_k(\boldsymbol{x}) \equiv R_k(\boldsymbol{x})$. As regards the large sample (asymptotic) behavior of the procedure, it is natural to adopt the approach common in statistical smoothing methods. Namely, as n increases, the width of kernels is controlled by a decreasing window-width, while simultaneously the number of lines, K, should increase, both with a proper rate. In such a way, the consistency of estimation procedure can be guaranteed.

3.2 Prediction

Now, let $\boldsymbol{x}$ be an arbitrary point from X, different from all $\boldsymbol{x}_i$. Predicted value suggested by (4), where parameters $\boldsymbol{w}_k$ are substituted by their estimates, is given as

$$\widehat{r}(\boldsymbol{x}) = \sum_{k=1}^{K} R_k(\boldsymbol{x}) \cdot \widehat{\boldsymbol{w}}_k' \boldsymbol{\varphi}_k(\boldsymbol{x}(k)),$$

while its 'true' value (which we do not know) would be $y(\boldsymbol{x}) = r(\boldsymbol{x}) + \varepsilon_x$, where $r(\boldsymbol{x})$ is given by (4), ε_x is the same random variable like all ε_i, independent on them. We are interested in the difference

$$\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x}) = \sum_{k=1}^{K} R_k(\boldsymbol{x}) \left(\widehat{\boldsymbol{w}}_k - \boldsymbol{x}_k\right)' \boldsymbol{\varphi}_k(\boldsymbol{x}(k)).$$

Let us denote by $\boldsymbol{w} = (\boldsymbol{w}_1', \boldsymbol{w}_2', \ldots, \boldsymbol{w}_K')'$ the vector of length $M \times 1$, where $M = \sum_{k=1}^{K} M_k$ containing all parameters, similarly for $\widehat{\boldsymbol{w}}$. Further,

$$\boldsymbol{b}(\boldsymbol{x}) = \left(R_1(\boldsymbol{x})\,\boldsymbol{\varphi}_1(\boldsymbol{x})', R_2(\boldsymbol{x})\,\boldsymbol{\varphi}_2(\boldsymbol{x})', \ldots, R_K(\boldsymbol{x})\,\boldsymbol{\varphi}_K(\boldsymbol{x})'\right)'$$

is also $M \times 1$ vector, depending on $\boldsymbol{x}$. Then, we obtain that

$$\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x}) = \boldsymbol{b}'(\boldsymbol{x})\,(\widehat{\boldsymbol{w}} - \boldsymbol{w}).$$

Its expectation is zero, while its variance equals

$$\mathrm{var}\,(\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x})) = \boldsymbol{b}'(\boldsymbol{x}) \cdot \mathscr{V} \cdot \boldsymbol{b}(\boldsymbol{x}),$$

where $\mathscr{V}$ is the $M \times M$ covariance matrix $\mathrm{cov}(\widehat{\boldsymbol{w}} - \boldsymbol{w})$. It is, naturally, symmetric, positive definite almost surely, and composed from blocks $V_{k\ell}$, $k, \ell = 1, \ldots, K$, of dimension $M_k \times M_k$:

$$\begin{aligned} V_{k\ell} &= \mathrm{E}\left\{(\widehat{\boldsymbol{w}}_k - \boldsymbol{w}_k) \cdot (\widehat{\boldsymbol{w}}_\ell - \boldsymbol{w}_\ell)'\right\} = \\ &= \mathrm{E}\left\{\boldsymbol{A}_k \Phi_k' P_k G_k\, \boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}'\, G_\ell P_\ell \Phi_\ell \boldsymbol{A}_\ell\right\} = \boldsymbol{A}_k \Phi_k' P_k G_k G_\ell P_\ell \Phi_\ell \boldsymbol{A}_\ell \cdot \sigma^2. \end{aligned}$$

"Diagonal" blocks are then $V_{kk} = \boldsymbol{A}_k \cdot \sigma^2$. Finally, the difference $\widehat{r}(\boldsymbol{x}) - y(\boldsymbol{x})$ has the expectation zero, too, and variance $\boldsymbol{b}'(\boldsymbol{x})\,\mathscr{V}\,\boldsymbol{b}(\boldsymbol{x}) + \sigma^2$.

As the matrix $\mathscr{V}$ depends on the design of observed points $\boldsymbol{x}_i$, the variability of prediction at a point $\boldsymbol{x}$ depends on information in its neighborhood. We thus, in the following example, see the same phenomenon as in Example 1, i.e. a growing uncertainty of model in sparse data regions.

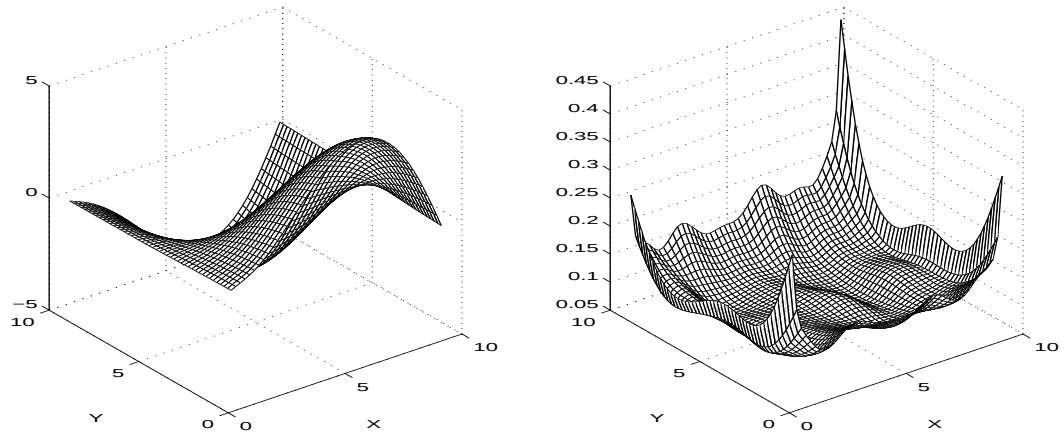
The model considered here contains $M = \sum_{k=1}^{K} M_k$ parameters. In a standard regression setting the number of observations should be $n > M$ to guarantee a reliable estimation. Notice that in the Gordon surface model $n > \max_k M_k$ suffices.

3.3 Analysis of Prediction Variance

As the direct computation of variance of prediction is here rather complicated (also due the presence of weighting functions) we shall prefer "empirical" illustration utilizing the following example.

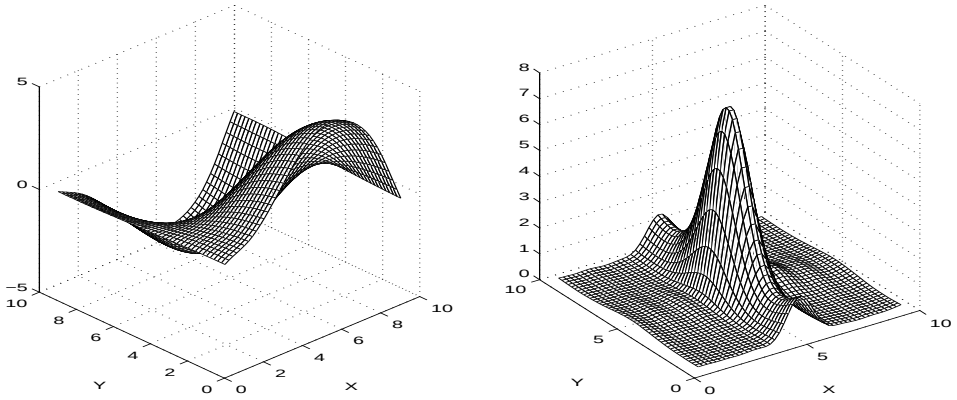
Example 2. The data of size $n = 200$ were generated from the function $r(\boldsymbol{x}) = x_1 \cdot sin(x_1/3) + sin(x_2/2 + 1)$ and additional standard Gauss random noise. For estimation, we fixed all 'nonlinear' components of the Gordon surface, namely: Number of lines, K, was fixed to 7, they were located horizontally, equidistantly in $(0, 10)$. To fit 1-dimensional curves along these lines, the cubic B-splines were used. For them, 5 equidistant knots were placed along each line. The same Gauss kernels were used both for weighting and blending.

In order to evaluate variability of estimates, the analysis was repeated $N = 200$ times. Figure 3 shows the mean and variance from 200 surfaces, estimated from full data covering the whole region $(0, 10) \times (0, 10)$. Then, Figure 4 displays the mean and variance of 200 surfaces estimated from data with missing values in $(4, 6) \times (4, 6)$ square. There is no significant differences between the means, they correspond, more-less, to the 'true' surface $r(\boldsymbol{x})$. However, the large variance in Figure 4 is the warning that one cannot rely just on estimated trend, that the inference has to be accompanied by the analysis of variance (and confidence, concerning the parameters). Notice also that the peak of variance has the vertical direction (in $X \times Y$ plane), i. e. orthogonal to lines L_k. It is caused by mixing of curves constructed along the lines, this mixing propagates the peak in the x direction.



Source: Own

Fig. 3. *The mean (left) and variance (right) estimated from the data covering uniformly the whole region* $(0,10) \times (0,10)$



Source: Own

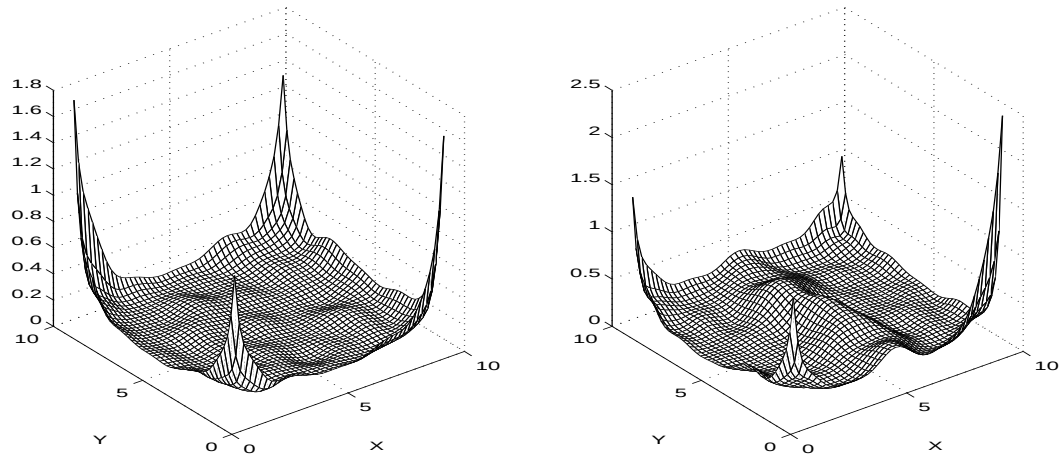
Fig. 4. *The mean (left) and variance (right) estimated from the data with missing values in region* $(4,6) \times (4,6)$

3.4 Comparison with Projection Pursuit

The same data as in Example 2 were repeatedly (with $N = 200$ repetitions) analyzed with the projection pursuit method. Again, the structure of model was fixed, we selected just four angles equidistantly inside $(0,\pi)$, each projection was fitted by B-splines with 5 inner knots located equidistantly between minimum and maximum value of projected $\boldsymbol{x}$-s. Such a selection was quite satisfactory, resulting surface fitted well (not worse than the Gordon surface shown on preceding figures). Figure 5 shows estimated variance of results (i.e. from 200 repetitions), left plot corresponds to full data, the right plot again to data with a gap in the region $(4,6) \times (4,6)$. The variance in the second case is still significant, however much lower compared to results of Gordon model application. It is probably caused by the structure of core lines (parallel lines in Gordon construction, a rosette in projection pursuit).

Conclusion

The objective of the paper was to examine several ways how to construct statistical model of 2-dimensional surfaces and to study advantages and problems of presented methods, from



Source: Own

Fig. 5. *Estimated variance of projection pursuit regression, from full data (left), from the data with missing values in region* $(4,6) \times (4,6)$ *(right)*

the point of their accuracy, flexibility, and also computational effort. The main attention was focused on the method of Gordon surface construction, and to the study of relationship between the design of data (input variables), selected set of localized functional units, and resulting variance of model estimate and prediction.

Acknowledgments

This work was supported by the Grant No. 209/10/2045 of the GA CR.

Literature

[1] ANDĚL, J.: *Foundantions of Mathematical Statistics (in Czech: Základy matematické statistiky)*. Matfyzpress, Praha, 2005.

[2] BISHOP, C.: *Neural Networks for Pattern Recognition*. Cambridge Univ. Press, Cambridge, 1992.

[3] DE BOOR, C.: *A Practical Guide to Splines*. Springer Verlag, Berlin, 1978.

[4] FRIEDMAN, J.H.: Multivariate adaptive regression splines, with Discussion and Rejoinder. *Annals Statist.* 19, 1991, pp. 1–141.

[5] GAMERMAN, D.: *Markov Chain Monte Carlo*. Chapman and Hall, New York, 1997.

[6] GORDON, W.J.: Spline-blended surface interpolation through curve networks. *Journal of Mathematics and Mechanics* 18, 1969, pp. 931–952.

[7] HUBER, P.J.: Projection pursuit. *Annals Statist.* 13, 1985, pp. 435–475.

[8] VOLF, P.: MCMC methods of randomized optimization and data analysis. In: *Proceedings of the ICPM 2007*, TU Liberec, 2007, pp. 123-130.

[9] WOLD, S.: Spline functions in data analysis. *Technometrics* 16, 1974, pp. 1–11.

doc. Petr Volf, CSc.

O DVOU FLEXIBILNÍCH METODÁCH PRO DVOUROZMĚRNOU REGRESNÍ ANALÝZU

Práce je věnována neparametrickému statistickému modelování dvourozměrných ploch na základě pozorovaných dat, tj. dvourozměrné regresní analýze. Model je konstruován jako kombinace jednoduchých bázových funkcí, jako jsou spliny, gaussiány apod. Nevýhodou takového přístupu je potřeba odhadnout velké množství parametrů. Budou ukázány dva přístupy, které počet potřebných parametrů redukují. A to metoda zvaná projection pursuit, která je statistikům dobře známá, a pak méně známá metoda konstrukce tzv. Gordonovy plochy. Zároveň také ukážeme, jak nepřítomnost dat v určité oblasti zvyšuje variabilitu modelu a tedy i neurčitost predikce. Metody budou předvedeny na umělých příkladech.

ÜBER ZWEI FLEXIBLE METHODEN FÜR EINE ZWEIDIMENSIONALE REGRESSANALYSE

Diese Arbeit befasst sich mit einer nichtparametrischen statistischen Modellierung zweidimensionaler Flächen auf Grundlage beobachteter Daten, d. h. mit der zweidimensionalen Regressanalyse. Das Modell ist konstruiert als Kombination einfacher Basenfunktionen wie z. B. Splines, Gaussian'sche Funktionen u. ä. Als Nachteil eines solchen Ansatzes erweist sicht die Notwendigkeit, eine große Menge Parameter zu schätzen. Es werden zwei Ansätze gezeigt, welche die Anzahl der erforderlichen Parameter reduziert, und zwar projection pursuit, welche den Statistikern wohl bekannt ist, und dann die weniger bekannte Konstruktionsmethode der Gordon-Flächen. Gleichzeitig zeigen wir, wie die Abwesenheit von Daten auf einem bestimmten Gebiet die Variabilität des Modells und damit auch die Unbestimmtheit der Prädiktion erhöht. Die Methoden werden an künstlichen Beispielen vorgeführt.

O DWU ELASTYCZNYCH METODACH DO DWUWYMIAROWEJ ANALIZY REGRESJI

Artykuł poświęcony jest nieparametrycznemu statystycznemu modelowaniu powierzchni dwuwymiarowych na podstawie obserwowanych danych, tj. dwuwymiarowej analizie regresji. Model skonstruowano jako połączenie prostych funkcji bazowych, takich jak splajny, gaussiany itp. Wadą takiego podejścia jest konieczność oszacowania dużej liczby parametrów. Pokazano dwa podejścia ograniczające liczbę niezbędnych parametrów. To metoda nazywana projection pursuit, która jest dobrze znana statystykom oraz mniej znana metoda konstrukcji tzw. powierzchni Gordona. Jednocześnie pokazano, jak brak danych w pewnym obszarze zwiększa zmienność modelu, czyli także niepewność prognozowania. Metody zaprezentowano na sztucznych przykładach.