Model selection and smoothing of mean and variance functions in nonparametric heteroscedastic regression

Samir Touzani, Daniel Busby

Model selection and smoothing of mean and variance functions in nonparametric heteroscedastic regression

Samir Touzani[a], Daniel Busby[a]

[a] *IFP Energies nouvelles, 92852 Rueil-Malmaison, France*

Abstract

In this paper we propose a new multivariate nonparametric heteroscedastic regression procedure in the framework of smoothing spline analysis of variance (SS-ANOVA). This penalized joint modelling estimators of the mean and variance functions is based on COSSO like penalty. The extended COSSO model performs simultaneously the estimation and the variable selection in the mean and variance ANOVA components. This allows to discover the sparse representation of the mean and the variance function when such sparsity exists. An efficient iterative algorithm is also introduced. The procedure is illustrated on several analytical examples and on an application from petroleum reservoir engineering.

Keywords: Joint Modelling, COSSO, Heteroscedasticity, Nonparametric Regression, SS-ANOVA

1. Introduction

In this article, we consider the multivariate nonparametric heteroscedastic regression problem, which can be mathematically formulated, for a given observation set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, as

$$y_i = \mu(\mathbf{x}_i) + \sigma(\mathbf{x}_i)\epsilon_i, \qquad i = 1, \ldots, n \tag{1}$$

where $\mathbf{x}_i = (x^{(1)}, \ldots, x^{(d)})$ are d dimensional vectors of input variables, $\epsilon_i \sim N(0, 1)$ are independent Gaussian noise with mean 0 and variance 1, μ and σ are unknown functions to be estimated, which correspond to the mean and the variance function. Note that in contrast with the classical homoscedastic regression model, the heteroscedastic regression model (1) has a non-constant input-dependent variance.

In a wide range of scientific and engineering applications, the joint modelling of mean and variance functions is an important problem. Indeed, in such applications it is important to model the local variability, which is described by variance function. For example, variance functions estimation is important for: measuring the volatility or risk in finance (Andersen and Lund, 1997; Gallant and Tauchen, 1997), quality control of experimental design (Box, 1988), industrial quality improvement experiments (Fan, 2000), detecting segmental genomic alterations (Huang and Pan, 2002; Wang and Guo, 2004; Liu et al., 2007), and for approximation of stochastic computer codes (Zabalza-Mezghani et al., 2001; Marrel et al., 2012). In a

Email address: samirtouzani.phd@gmail.com (Samir Touzani)

nutshell, the heteroscedastic regression is a common statistical method for analyzing unreplicated experiments.

In statistical literature various nonparametric heteroscedastic regression methods have been proposed to estimate the mean and variance functions (Carroll, 1982; Hall and Carroll, 1989; Fan and Yao, 1998; Yuan and Wahba, 2004; Cawley et al., 2004; Liu et al., 2007; Gijbels et al., 2010; Wang, 2011). These methods are based on local polynomial smoothers or smoothing splines, and most of them are based on a two step procedure. First, the mean function is estimated. Then the variance function is estimated using the squared regression residuals as observations on the variance function. There is also an active research on parametric joint modelling, for example we can cite Nelder and Lee (1998); Smyth and Verbyla (1999, 2009), these methods used two coupled generalized linear models, which allows to use the well established generalized linear model diagnostics and concepts (McCullagh and Nelder, 1989).

A well established and popular approach to the nonparametric homoscedastic estimation for high dimensional problems is the smoothing spline analysis of variance (SS-ANOVA) model (Wahba, 1990; Wang, 2011; Gu, 2013). This approach generalize the additive model and is based on ANOVA expansion, which is defined as

$$f(\mathbf{x}) = f_0 + \sum_{j=1}^{d} f_j(x^{(j)}) + \sum_{j<l} f_{jl}(x^{(j)}, x^{(l)}) + ... + f_{1,...,d}(x^{(1)}, ..., x^{(d)}) \tag{2}$$

where $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})$, f_0 is a constant, f_j's are univariate functions representing the main effects, f_{jl}'s are bivariate functions representing the two way interactions, and so on. It is important to determine which ANOVA components should be included in the model. Recently Lin and Zhang (2006) proposed Component Selection and Smoothing Operator method (COSSO), which is a new nonparametric regression method based on a penalized least square estimation with the penalty functional being the sum of component norms instead of the sum of component squared norms, which characterize the classical SS-ANOVA method. This method can be seen as an extension of the LASSO (Tibshirani, 1996) variable selection method to nonparametric models. Thus, COSSO regression method achieves sparse solutions, which mean that it estimates some functional components to be zero. This property can help to increase efficiency in functions estimation.

The aim of this article is to develop a new method to jointly estimate mean and variance functions in the framework of smoothing spline analysis of variance (SS-ANOVA, see Wahba (1990)) based on COSSO like (Component Selection and Smoothing Operator) penalty. Our extended COSSO model performs simultaneously the estimation and the variable selection in the mean and variance ANOVA components. This allows us to discover the sparse representation of the mean and the variance function when such sparsity exists. Our work is closely similar to Yuan and Wahba (2004) and Cawley et al. (2004), who independently proposed a doubly penalized likelihood kernel method, which estimates both the mean and variance functions simultaneously. In particualr, the doubly penalized likelihood estimator of Yuan and Wahba (2004) can be seen as a generalization of the SS-ANOVA method to deal with problems, which are characterized by a heteroscedastic Gaussian noise process.

We organized this article as follows, we first introduce the extended COSSO method. Then we present algorithm of the proposed procedure. In Section 4 a discussion on evaluating the prediction performance is provided as well as a simulation study investigating the performance of the joint modelling COSSO. Finally, an application to an approximation of a stochastic

computer code in the framework of reservoir engineering is presented.

2. COSSO-like Joint modelling of mean and variance procedure

By assuming the heteroscedastic regression problem (1), the conditional probability density of output y_i given input $\mathbf{x}_i$, is given by

$$p(y_i|\mathbf{x}_i) = \frac{1}{\sqrt{2\pi}\sigma(\mathbf{x}_i)} \exp\left\{-\frac{(\mu(\mathbf{x}_i) - y_i)^2}{2\sigma^2(\mathbf{x}_i)}\right\} \tag{3}$$

Therefore we can write (by omitting the constant term which is not depending on μ and σ) the average negative log likelihood of (1) as

$$L(\mu, \sigma) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(y_i - \mu(\mathbf{x}_i))^2}{2\sigma^2(\mathbf{x}_i)} + \frac{1}{2}\log\sigma^2(\mathbf{x}_i)\right) \tag{4}$$

μ and σ^2 are the unknown functions that we must estimate. However, to ensure the positivity constraint of σ^2 we estimate g function instead of σ^2. This g function is defined as

$$\sigma^2(\mathbf{x}) = \exp\{g(\mathbf{x})\} \tag{5}$$

2.1. Smoothing spline ANOVA model in heteroscedastic regression

For simplicity, we assume in what follows that $\mathbf{x}_i \in [0,1]^d$. In the framework of smoothing spline analysis of variance (SS-ANOVA) we suppose that μ and g reside in some Hilbert spaces $\mathcal{H}_\mu$ and $\mathcal{H}_g$ of smooth functions. Then the functional ANOVA decompositions of multivariate functions μ and g are defined as

$$\mu(\mathbf{x}) = b_\mu + \sum_{j=1}^{d}\mu_j(x^{(j)}) + \sum_{j<k}\mu_{jk}(x^{(j)}, x^{(k)}) + \dots \tag{6}$$

$$g(\mathbf{x}) = b_g + \sum_{j=1}^{d} g_j(x^{(j)}) + \sum_{j<k} g_{jk}(x^{(j)}, x^{(k)}) + \dots \tag{7}$$

where b_μ and b_g are constant, μ_j and g_j are the main effects, μ_{jk} and g_{jk} are the two-way interactions, and so on. In the SS-ANOVA model we assume $\mu_j \in \bar{\mathcal{H}}_\mu^{(j)}$ and $g_j \in \bar{\mathcal{H}}_g^{(j)}$, where $\bar{\mathcal{H}}_\mu^{(j)}$ and $\bar{\mathcal{H}}_g^{(j)}$ are a reproducing kernel Hilbert spaces (RKHS) such that

$$\mathcal{H}_\mu^{(j)} = \{1\} \oplus \bar{\mathcal{H}}_\mu^{(j)} \tag{8}$$

$$\mathcal{H}_g^{(j)} = \{1\} \oplus \bar{\mathcal{H}}_g^{(j)} \tag{9}$$

Thereby the full functions spaces $\mathcal{H}_\mu$ and $\mathcal{H}_g$ are defined as the tensor products

$$\mathcal{H}_\mu = \bigotimes_{j=1}^{d}\mathcal{H}_\mu^{(j)} = \{1\} \oplus \left[\bigoplus_{j=1}^{d}\bar{\mathcal{H}}_\mu^{(j)}\right] \oplus \left[\bigoplus_{j<k}^{d}(\bar{\mathcal{H}}_\mu^{(j)} \otimes \bar{\mathcal{H}}_\mu^{(k)})\right] \oplus \dots \tag{10}$$

$$\mathcal{H}_g = \bigotimes_{j=1}^{d} \mathcal{H}_g^{(j)} = \{1\} \oplus \left[\bigoplus_{j=1}^{d} \bar{\mathcal{H}}_g^{(j)} \right] \oplus \left[\bigoplus_{j<k}^{d} (\bar{\mathcal{H}}_g^{(j)} \otimes \bar{\mathcal{H}}_g^{(k)}) \right] \oplus \dots \tag{11}$$

Each component in the ANOVA decompositions (6) and (7) are associated to a corresponding subspace in the orthogonal decompositions (10) and (11). In practice, for computational reason we assume that only two way interactions are considered in the ANOVA decomposition and an expansion to the second order generally provides a satisfactory description of the model. Let's consider the index $\alpha \equiv j$ for $\alpha = 1, \dots, d$ with $j = 1, \dots, d$ and $\alpha \equiv (j, l)$ for $\alpha = d+1, \dots, d(d+1)/2$ with $1 \leq j < l \leq d$. Thus the truncated function spaces assumed for the SS-ANOVA model of μ and g are

$$\mathcal{H}_\mu = \{1\} \bigoplus_{\alpha=1}^{q} \mathcal{F}_\mu^\alpha \tag{12}$$

$$\mathcal{H}_g = \{1\} \bigoplus_{\alpha=1}^{q} \mathcal{F}_g^\alpha \tag{13}$$

where $\mathcal{F}_\mu^1, \dots \mathcal{F}_\mu^q$ and $\mathcal{F}_g^1, \dots \mathcal{F}_g^q$ are q orthogonal subspaces of $\mathcal{H}_\mu$ and $\mathcal{H}_g$ and $q = d(d+1)/2$ corresponds to the number of ANOVA components for the two-way interaction model.

2.2. Doubly penalized COSSO likelihood method

In this work we consider the COSSO-like Lin and Zhang (2006) penalized likelihood strategy, which conducts simultaneous model selection and estimation. The COSSO penalizes on the sum of the norms which achieves sparse solutions (some estimated ANOVA components are equal to zero), and hence this method can be seen as an extension of the LASSO Tibshirani (1996) variable selection method in parametric models to nonparametric models. Thus, the estimators $\hat{\mu}$ and $\hat{g}$ are defined to be the minimizer of the following doubly penalized negative log likelihood function

$$L(\mu, g) + \lambda_\mu \sum_{\alpha=1}^{q} ||P^\alpha \mu|| + \lambda_g \sum_{\alpha=1}^{q} ||P^\alpha g|| \tag{14}$$

where λ_μ and λ_g are the nonnegative smoothing parameters, $P^\alpha \mu$ and $P^\alpha g$ are the orthogonal projection of μ and g onto $\mathcal{F}_\mu^\alpha$ and $\mathcal{F}_\sigma^\alpha$. Thus the doubly penalized negative likelihood (14) is re-expressed as

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ (y_i - \mu(\mathbf{x}_i))^2 \exp\{-g(\mathbf{x}_i)\} + g(\mathbf{x}_i) \right\} + \lambda_\mu \sum_{\alpha=1}^{q} ||P^\alpha \mu|| + \lambda_g \sum_{\alpha=1}^{q} ||P^\alpha g|| \tag{15}$$

An equivalent formulation of (15) that is easier to compute using an iterative algorithm is given by

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} \left\{ (y_i - \mu(\mathbf{x}_i))^2 \exp\{-g(\mathbf{x}_i)\} + g(\mathbf{x}_i) \right\} \\ &\qquad + \lambda_\mu^0 \sum_{\alpha=1}^{q} (\theta_\alpha^\mu)^{-1} ||P^\alpha \mu||^2 + \tau_\mu \sum_{\alpha=1}^{q} \theta_\alpha^\mu + \lambda_g^0 \sum_{\alpha=1}^{q} (\theta_\alpha^g)^{-1} ||P^\alpha g||^2 + \tau_g \sum_{\alpha=1}^{q} \theta_\alpha^g \end{aligned} \tag{16}$$

subject to $\theta_\alpha^\mu, \theta_\alpha^g \geq 0$, where $\lambda_\mu^0, \lambda_g^0 > 0$ are constant and τ_μ, τ_g are the smoothing parameters. If $\theta_\alpha^\mu = \theta_\alpha^g = 0$ (We take the convention $0/0 = 0$), then the minimizer is taken to satisfy $||P^\alpha \mu|| = ||P^\alpha g|| = 0$.
The penalty terms $\sum_{\alpha=1}^q \theta_\alpha^\mu$ and $\sum_{\alpha=1}^q \theta_\alpha^g$ control the sparsity in the ANOVA components of μ and g. Indeed, the sparsity of each component μ_α and g_α is controlled by θ_α^μ and respectively by θ_α^g. The functional (16) is similar to the doubly penalized likelihood estimator in heteroscedastic regression introduced by Yuan and Wahba (2004) except that in (16) we have introduced model selection parts represented by the penalty terms on θ^μ's and θ^g's.

3. Algorithm

We assume that $\mathcal{F}^\alpha = \mathcal{F}_\mu^\alpha = \mathcal{F}_g^\alpha$ and K_α is the reproducing kernel of $\mathcal{F}^\alpha$. We also assume that the considered input parameters are continuous, then a typical RKHS $\mathcal{F}^j$ is the second-order Sobolev Hilbert space and the corresponding reproducing kernel (Wahba, 1990) is defined as

$$K_\alpha(x, x') = K_j(x, x') = k_1(x)k_1(x') + k_2(x)k_2(x') - k_4(|x - x'|) \tag{17}$$

where $k_l(x) = B_l(x)/l!$ and B_l is the lth Bernoulli polynomial. Thus, for $x \in [0, 1]$

$$\begin{aligned} k_1(x) &= x - \frac{1}{2} \\ k_2(x) &= \frac{1}{2}(k_1^2(x) - 1/12) \\ k_4(x) &= \frac{1}{24}(k_1^4(x) - \frac{k_1^2(x)}{2} + \frac{7}{240}) \end{aligned} \tag{18}$$

Moreover, the reproducing kernel K_α for the RKHS $\mathcal{F}^\alpha$ such as $\mathcal{F}^\alpha \equiv \bar{\mathcal{H}}_\mu^{(j)} \otimes \bar{\mathcal{H}}_\mu^{(l)}$ (respectively $\mathcal{F}^\alpha \equiv \bar{\mathcal{H}}_g^{(j)} \otimes \bar{\mathcal{H}}_g^{(l)}$), are given by the following tensor products

$$K_\alpha(\mathbf{x}, \mathbf{x}') = K_j(x^{(j)}, x^{(j)'}) K_l(x^{(l)}, x^{(l)'})$$

The mean function μ and the function g can be estimated by alternatively minimizing (16) with respect to μ and g via an iterative procedure. Indeed, it is easy to see that (16) is a convex function of each μ and g by fixing the other.

3.1. Step A: Estimating the mean function

By fixing g, the functional (16) is equivalent to a penalized weighted least squares procedure

$$\frac{1}{n}\sum_{i=1}^n \exp\{-g(\mathbf{x}_i)\}(y_i - \mu(\mathbf{x}_i))^2 + \lambda_\mu^0 \sum_{\alpha=1}^q (\theta_\alpha^\mu)^{-1} ||P_\mu^\alpha||^2 + \tau_\mu \sum_{\alpha=1}^q \theta_\alpha^\mu \tag{19}$$

For fixed $\boldsymbol{\theta}^\mu$ the representer theorem for the smooting spline states that the minimizer of (19) has the folowing form

$$\mu(\mathbf{x}) = b_\mu + \sum_{i=1}^n c_i^\mu \sum_{\alpha=1}^q \theta_\alpha^\mu K_\alpha(\mathbf{x}, \mathbf{x}_i) \tag{20}$$

Let $K_{\theta^\mu} = \sum_{\alpha=1}^q \theta_\alpha^\mu K_\alpha$ with K_α stand for the $n \times n$ matrix $\{K_\alpha(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$ $\mathbf{c}^\mu = (c_1^\mu, \ldots, c_n^\mu)^T$, D a $n \times n$ diagonal matrix with the i-th diagonal elements equals to $\exp\{-g(\mathbf{x}_i)\}$, $\mathbf{Y} =$

$(y_1, \ldots, y_n)^T$ and $\mathbf{1}_n$ be the vector of ones of length n. Then (19) is equivalent in matrix notation to

$$(\mathbf{Y} - b_\mu \mathbf{1}_n - K_{\theta^\mu}\mathbf{c}^\mu)^T D(\mathbf{Y} - b_\mu \mathbf{1}_n - K_{\theta^\mu}\mathbf{c}^\mu) + n\lambda_\mu^0 (\mathbf{c}^\mu)^T K_{\theta^\mu}\mathbf{c}^\mu + \tau_\mu \mathbf{1}_q^T \boldsymbol{\theta}^\mu \tag{21}$$

If $\boldsymbol{\theta}_\mu = (\theta_1^\mu, \ldots, \theta_q^\mu)^T$ is fixed, then (19) becomes

$$\min_{\mathbf{c}^\mu, b_\mu} (\mathbf{Y} - b_\mu \mathbf{1}_n - K_{\theta^\mu}\mathbf{c}^\mu)^T D(\mathbf{Y} - b_\mu \mathbf{1}_n - K_{\theta^\mu}\mathbf{c}^\mu) + n\lambda_\mu^0 (\mathbf{c}^\mu)^T K_{\theta^\mu}\mathbf{c}^\mu \tag{22}$$

which is equivalent to

$$\min_{\mathbf{c}_\mu^d, b_\mu} (\mathbf{Y}_d - b_\mu \mathbf{1}_n^d - K_{\theta^\mu}^d \mathbf{c}_d^\mu)^T (\mathbf{Y}^d - b_\mu \mathbf{1}_n^d - K_{\theta^\mu}^d \mathbf{c}_d^\mu) + n\lambda_\mu^0 (\mathbf{c}_d^\mu)^T K_{\theta^\mu}^d \mathbf{c}_d^\mu \tag{23}$$

where $\mathbf{Y}_d = D^{1/2}\mathbf{Y}$, $\mathbf{1}_n^d = D^{1/2}\mathbf{1}_n$, $K_{\theta^\mu}^d = D^{1/2}K_{\theta^\mu}$ and $\mathbf{c}_d^\mu = D^{-1/2}\mathbf{c}^\mu$. Thus (23) is equivalent to the standard smoothing spline ANOVA problem, and so (23) is minimized by the solution of the following linear equations

$$\begin{aligned} (K_{\theta^\mu} + n\lambda_\mu^0 D^{-1})\mathbf{c}^\mu + b_\mu \mathbf{1}_n &= \mathbf{Y} \\ \mathbf{1}_n \mathbf{c}^\mu &= 0 \end{aligned} \tag{24}$$

For fixed $\mathbf{c}^\mu$ and b_μ (21) can be written as

$$\min_{\boldsymbol{\theta}^\mu} (\mathbf{Y} - b_\mu \mathbf{1}_n - W\boldsymbol{\theta}^\mu)^T D(\mathbf{Y} - b_\mu \mathbf{1}_n - W\boldsymbol{\theta}^\mu) + n\lambda_\mu^0 (\mathbf{c}^\mu)^T W\boldsymbol{\theta}^\mu + n\tau_\mu \mathbf{1}_q \boldsymbol{\theta}^\mu \tag{25}$$

where $\theta_\alpha^\mu \geq 0$ for $\alpha = 1, \ldots, q$, W is a $n \times q$ matrix with the α-th column $\mathbf{w}_\alpha = K_\alpha \mathbf{c}_\mu$. Let $W_d = D^{1/2}W$ and $z_d = \mathbf{Y}_d - b_\mu \mathbf{1}_n^d - \frac{1}{n} n\lambda_\mu^0 \mathbf{c}_d^\mu$. Then (25) is equivalent to the following nonnegative garrot Breiman (1995) optimization problem

$$(z_d - W_d\theta^\mu)^T (z_d - W_d\theta^\mu) + n\tau_\mu \sum_{\alpha=1}^{q} \boldsymbol{\theta}^\mu \quad \text{subject to} \quad \theta_\alpha^\mu \geq 0 \tag{26}$$

An equivalent form of (26) is given by

$$(z_d - W_d\theta^\mu)^T (z_d - W_d\theta^\mu) \quad \text{subject to} \quad \theta_\alpha^\mu \geq 0 \quad \text{and} \quad \sum_{\alpha=1}^{q} \boldsymbol{\theta}^\mu \leq M_\mu \tag{27}$$

for some $M \geq 0$. Lin and Zhang (2006) noted that the optimal M_μ seems to be close to the number of important components. Depending on the method used to solve the nonnegative garrot problem we use one or other of (26) and (27), for more details we refer to Touzani and Busby (2013).

Thus the algorithm to estimate the mean function (the COSSO algorithm) is presented as a one step uptade procedure

Algorithm 1

1: Initialization: Fix $\theta_\alpha^\mu = 1$, $\alpha = 1, ..., q$

2: For each fixed λ_μ^0 in a chosen range solve for $\mathbf{c}^\mu$ and b_μ with (23). Tune λ_μ^0 using v-fold-cross-validation. Set $\mathbf{c}_0^\mu$ and b_μ^0 the solution of (23) for the best value of λ_μ^0. Fix λ_μ^0 at the chosen value.

3: For each fixed M_μ in a chosen range, apply the following procedure:

3.1: Solve for $\mathbf{c}^\mu$ and b_μ with (23).

3.2: For $\mathbf{c}^\mu$ and b_μ obtained in step 3.1, solve for $\boldsymbol{\theta}^\mu$ with (27).

3.3: For $\boldsymbol{\theta}^\mu$ obtained in step 3.2, solve for $\mathbf{c}^\mu$ and b_μ with (23).

Tune M_μ using v-fold-cross-validation. Fix M_μ at the chosen value.

4: For λ_μ^0, $\mathbf{c}_0^\mu$ and b_μ^0 from steps 2, and with M_μ from steps 3, solve for $\boldsymbol{\theta}^\mu$ with (27). The solution corresponds to the final estimate of $\boldsymbol{\theta}^\mu$.

5: For $\boldsymbol{\theta}^\mu$ from steps 4 and λ_μ^0 from steps 2, solve for $\mathbf{c}^\mu$ and b_μ with (23). The solution corresponds to the final estimate of $\mathbf{c}^\mu$ and b_μ.

A good choice of the smoothing parameters λ_μ^0 and M_μ is important to the predictivity performance of the estimate $\hat{\mu}$ of the mean function. We use the v-folds Cross Validation to minimize the following weighted least squares (WLS) criterion

$$WLS(\hat{\mu}) = \frac{1}{n_{test}} \sum_{i \in v_{test}} d_{ii}(y_i - \tilde{\mu}(\mathbf{x}_i))^2 \tag{28}$$

where v_{test} is the cross validation set of the n_{test} points and $\tilde{\mu}$ is an estimation of μ using the cross validation test points.

3.2. Step B: Estimating the variance function

With μ fixed (16) is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n}(z_i \exp\{-g(\mathbf{x}_i)\} + g(\mathbf{x}_i)) + \lambda_g^0 \sum_{\alpha=1}^{q}(\theta_\alpha^g)^{-1}||P^\alpha g||^2 + \tau_g \sum_{\alpha=1}^{q}\theta_\alpha^g \tag{29}$$

where $z_i = (y_i - \tilde{\mu}(\mathbf{x}_i))^2$ and $\tilde{\mu}$ is the estimate of μ given by the step A. By considering z_i, $i = 1, \ldots, n$ independent samples from Gamma distributions, the functional (29) has the form of a penalized Gamma likelihood. For fixed $\boldsymbol{\theta}^g$ the representer theorem for the smooting spline states that the minimizer of (29) has the following form

$$g(\mathbf{x}) = b_g + \sum_{i=1}^{n} c_i^g \sum_{\alpha=1}^{q} \theta_\alpha^g K_\alpha(\mathbf{x}, \mathbf{x}_i) \tag{30}$$

The solution of (29) can then be estimated via Newton-Raphson iteration algorithm (Zhang and Lin, 2006). Given an initial solution g^0, the second order Taylor expansion of $z_i \exp\{-g(\mathbf{x}_i)\} + g(\mathbf{x}_i)$ at g^0 is

$$\begin{aligned} z_i \exp\{-g^0(\mathbf{x}_i)\} + g^0(\mathbf{x}_i) + (g(\mathbf{x}_i) - g^0(\mathbf{x}_i))(-z_i \exp\{-g^0(\mathbf{x}_i)\} + 1) \\ + \frac{1}{n} z_i \exp\{-g^0(\mathbf{x}_i)\}(g(\mathbf{x}_i) - g^0(\mathbf{x}_i))^2 \end{aligned} \tag{31}$$

which is equal to

$$\frac{1}{2} z_i \exp\{-g^0(\mathbf{x}_i)\} \left[g(\mathbf{x}_i) - g^0(\mathbf{x}_i) + \frac{(-z_i \exp\{-g^0(\mathbf{x}_i)\} + 1)}{z_i \exp\{-g^0(\mathbf{x}_i)\}} \right]^2 + \beta_i \tag{32}$$

where β_i is independent of $g(\mathbf{x}_i)$. Let $\xi_i = g^0(\mathbf{x}_i) + \frac{(z_i \exp\{-g^0(\mathbf{x}_i)\}-1)}{z_i \exp\{-g^0(\mathbf{x}_i)\}}$ and the weight matrix D_g which is a $n \times n$ diagonal matrix with the i-th diagonal elements equals to $z_i \exp\{-g^0(\mathbf{x}_i)\}$. Then at each iteration of the Newton-Raphson algorithm we solve

$$(\boldsymbol{\xi} - b_g \mathbf{1}_n - K_{\theta^g} \mathbf{c}^g)^T D_g (\boldsymbol{\xi} - b_g \mathbf{1}_n - K_{\theta^g} \mathbf{c}^g) + n\lambda_g^0 (\mathbf{c}^g)^T K_{\theta^g} \mathbf{c}^g + \tau_g \mathbf{1}_q^T \boldsymbol{\theta}^g \tag{33}$$

where $\boldsymbol{\xi} = \xi_1, \ldots, \xi_n$, $K_{\theta^g} = \sum_{\alpha=1}^q \theta_\alpha^g K_\alpha$ and $\mathbf{c}^g = (c_1^g, \ldots, c_n^g)^T$. The functional (33) is a penalized weighted least squares with the weights changing at each iteration. As proposed by Zhang and Lin (2006) we minimize (33) by alternatively solving $(b_g, \mathbf{c}^g)$ with $\boldsymbol{\theta}^g$ fixed and solving $\boldsymbol{\theta}^g$ with $(b_g, \mathbf{c}^g)$ fixed, which is similar to the algorithm 1.
Thus, for fixed $(\theta)^g$ solving (33) is equivalent to

$$\min_{\mathbf{c}_g^d, b_g} (\boldsymbol{\xi}_d - b_g \mathbf{1}_n^d - K_{\theta^g}^d \mathbf{c}_d^g)^T (\boldsymbol{\xi}^d - b_g \mathbf{1}_n^d - K_{\theta^g}^d \mathbf{c}_d^g) + n\lambda_g^0 (\mathbf{c}_d^g)^T K_{\theta^g}^d \mathbf{c}_d^g \tag{34}$$

where $\boldsymbol{\xi}_d = D^{1/2}\boldsymbol{\xi}$, $\mathbf{1}_n^d = D^{1/2}\mathbf{1}_n$, $K_{\theta^g}^d = D^{1/2} K_{\theta^g}$ and $\mathbf{c}_d^g = D^{-1/2}\mathbf{c}^g$. Then (34) is equivalent to the SS-ANOVA problem and then minimized by the solution of

$$\begin{aligned} (K_{\theta^g} + n\lambda_g^0 D_g^{-1})\mathbf{c}^g + b_g \mathbf{1}_n &= \boldsymbol{\xi} \\ \mathbf{1}_n \mathbf{c}^g &= 0 \end{aligned} \tag{35}$$

For fixed $\mathbf{c}^g$ and b_g (33) is equivalent to the following nonnegative garrot optimization problem

$$(u_d - W_d \theta^g)^T (u_d - W_d \theta^g) + n\tau_g \sum_{\alpha=1}^q \boldsymbol{\theta}^g \quad \text{subject to} \quad \theta_\alpha^g \geq 0 \tag{36}$$

where $W_g^d = D^{1/2} W_g$ with W_g is a $n \times q$ matrix with the α-th column $\mathbf{w}_\alpha = K_\alpha \mathbf{c_g}$ and $u_d = \boldsymbol{\xi}_d - b_g \mathbf{1}_n^d - \frac{1}{n} n\lambda_g^0 \mathbf{c}_d^g$. As previously seen, (36) is equivalent to

$$(u_d - W_d \theta^g)^T (u_d - W_d \theta^g) \quad \text{subject to} \quad \theta_\alpha^g \geq 0 \quad \text{and} \quad \sum_{\alpha=1}^q \boldsymbol{\theta}^g \leq M_g \tag{37}$$

The algorithm to estimate the variance function is presented as the following one step uptade procedure

Algorithm 2

1: Initialization: set $g^0 = \mathbf{0}_n$
2: Fix $\theta_\alpha^g = 1$, $\alpha = 1, ..., q$
3: For each fixed λ_g^0 in a chosen range solve for $\mathbf{c}^g$ and b_g with (34). Tune λ_g^0 using v-fold-cross-validation. Set $\mathbf{c}_0^g$ and b_g^0 the solution of (34) for the best value of λ_g^0. Fix λ_g^0 at the chosen value.
4: For each fixed M_g in a chosen range, apply the following procedure:

3.1: Solve for $\mathbf{c}^g$ and b_g with (34).
3.2: For $\mathbf{c}^g$ and b_g obtained in step 3.1, solve for $\boldsymbol{\theta}^g$ with (37).
3.3: For $\boldsymbol{\theta}^g$ obtained in step 3.2, solve for $\mathbf{c}^g$ and b_g with (34).

Tune M_g using v-fold-cross-validation. Fix M_g at the chosen value.
5: For λ_g^0, $\mathbf{c}_0^g$ and b_g^0 from steps 2, and with M_g from steps 3, solve for $\boldsymbol{\theta}^g$ with (37). The solution corresponds to the final estimate of $\boldsymbol{\theta}^g$.
6: For $\boldsymbol{\theta}^g$ from steps 4 and λ_g^0 from steps 2, solve for $\mathbf{c}^g$ and b_g with (34). The solution corresponds to the final estimate of $\mathbf{c}^g$ and b_g.
7: Calculate $\mathbf{g} = b_g \mathbf{1}_n + K_{\theta^g} \mathbf{c}^g$.
8: Go to the step 2 with $\mathbf{g}^0 = \mathbf{g}$ until the convergence criterion is satisfied.

The parameters λ_g^0 and M_g are tuned by v-folds Cross Validation by minimizing the comparative Kullback-Leiber criterion

$$CKL(\tilde{g}, g) = \frac{1}{n_{test}} \sum_{i \in v_{test}} [z_i \exp\{-\tilde{g}(\mathbf{x}_i)\} + \hat{g}(\mathbf{x}_i)] \tag{38}$$

where $\tilde{\mu}$ and $\tilde{g}$ the estimation omitting all members of v_{test} and since the function g is unknown we approximate it by $\hat{g}$, which is the estimation of g using all observations.
The complete algorithm for the doubly penalized COSSO likelihood method, which we name joint modelling COSSO (JM COSSO) in what follows, is presented as

Algorithm 3

1: Initialization: Fix $g = \mathbf{0}_n$.
2: Estimate $\mathbf{c}^\mu$, b_μ and $\boldsymbol{\theta}^\mu$ using algorithm 1.
3: Set $\tilde{\boldsymbol{\mu}} = b_\mu \mathbf{1}_n + K_{\theta^\mu} \mathbf{c}^\mu$.
4: Estimate $\mathbf{c}^g$, b_g and $\boldsymbol{\theta}^g$ using algorithm 2 and $\tilde{\boldsymbol{\mu}}$.
5: Set $\tilde{\boldsymbol{g}} = b_g \mathbf{1}_n + K_{\theta^g} \mathbf{c}^g$.
6: Estimate $\mathbf{c}^\mu$, b_μ and $\boldsymbol{\theta}^\mu$ using algorithm 1 and $\tilde{\boldsymbol{g}}$.

4. Simulations

In the present section we investigate the performance of the JM COSSO procedure via two numerical simulation studies, involving normal data. The empirical performance of mean and variance estimators will be measured in terms of prediction accuracy and model selection. For this end we compute several quantities on a test set of size $n_{test} = 10000$. The tuning parameters are choosed using 5-fold Cross Validation. All computing has been done using the R-programming language.

4.1. Evaluating the prediction performance

A common approach for comparing the prediction performance of different heteroscedastic regression methods is based on the negative log-likelihood of the models (Juutilainen and Roning, 2008; Cawley et al., 2004). However, it is difficult to evaluate the goodness of the prediction accuracy using the log-likelihood quantity. Thus there is an advantage in using as a prediction performance criterion, not the log-likelihood but a scaled deviance (Juutilainen and Roning, 2010). By considering the squared residuals as observations and following the assumption that the residuals $(y_i - \hat{\mu}_i)$ are normally distributed, the squared residuals $(y_i - \hat{\mu}_i)^2$ are gamma distributed. Thus in this case the deviance is defined as twice the difference between the log-likelihood of the model and the log-likelihood that would be achieved if variance prediction $\hat{\sigma}_i^2$ is equal to the observed squared residual.

$$D = 2\sum_{i=1}^{N}\left(\log\left(\frac{\hat{\sigma}_i^2}{(y_i-\hat{\mu}_i)^2}\right) + \frac{(y_i-\hat{\mu}_i)^2 - \hat{\sigma}_i^2}{\hat{\sigma}_i^2}\right) \tag{39}$$

where N is the number of elements of the test sample set. A better prediction accuracy has smaller deviance. Therefore, if the estimation of the variance and the mean is correct, $\hat{\sigma}_i^2 = \sigma_i^2$ and $\hat{\mu}_i^2 = \mu_i^2$ for $i = 1, \ldots, n_{test}$, then the expected value of deviance is defined as

$$\mathbb{E}(D) = 2\mathbb{E}\left[\sum_{i=1}^{N}\left(\log\left(\frac{\hat{\sigma}_i^2}{\sigma_i^2 z^2}\right) + \frac{\sigma_i^2 z^2 - \hat{\sigma}_i^2}{\hat{\sigma}_i^2}\right)\right] = -2n_{test}E[\log(z^2)] \approx 2.541 n_{test} \tag{40}$$

where $z \sim N(0,1)$ is a standard normal random variable and the expected logarithm of χ_1^2 distribution $E[\log(z^2)]$ is approximately equal to -1.2704. If D is much higher than $2.541 n_{test}$ then this result should be interpreted as a weakness of the prediction accuracy or as erroneous Gaussian assumption. The departure of the calculated deviance from the expected deviance of an optimal model can be used as a qualitative criterion of model prediction accuracy. In what follows D will be normalized by n_{test}.

In addition to the deviance measure, and since we have the analytical form of mean and variance in these studies we can also evaluate the prediction accuracy of each mean and variance estimators via Q_2 quantities. The Q_2 for mean and variance are defined as

$$Q_2^{\mu} = 1 - \frac{\sum_{i=1}^{n_{test}}(y_i^{\mu} - \widehat{\mu}(\mathbf{x}_i))^2}{\sum_{i=1}^{n_{test}}(y_i^{\mu} - \bar{y}^{\mu})^2} \quad ; \quad Q_2^{\sigma} = 1 - \frac{\sum_{i=1}^{n_{test}}(y_i^{\sigma} - \widehat{\sigma}(\mathbf{x}_i))^2}{\sum_{i=1}^{n_{test}}(y_i^{\sigma} - \bar{y}^{\sigma})^2} \tag{41}$$

The Q_2 is defined as the coefficient of determination R^2 computed on a test set and provide a measure of what proportion of the empirical variance in the output data is accounted for by the estimator. Note that more the Q_2 is close to one and more the the estimator is accurate.

4.2. Example 1

Let consider a 5 dimensional normal model, where the mean function is defined as

$$\mu(\mathbf{x}) = 5g_1(x^{(1)}) + 3g_2(x^{(2)}) + 4g_3(x^{(3)}) + 6g_4(x^{(4)}) \tag{42}$$

and the variance function is defined as

$$\sigma^2(\mathbf{x}) = 15g_5(x^{(1)})g_6(x^{(2)}) \tag{43}$$

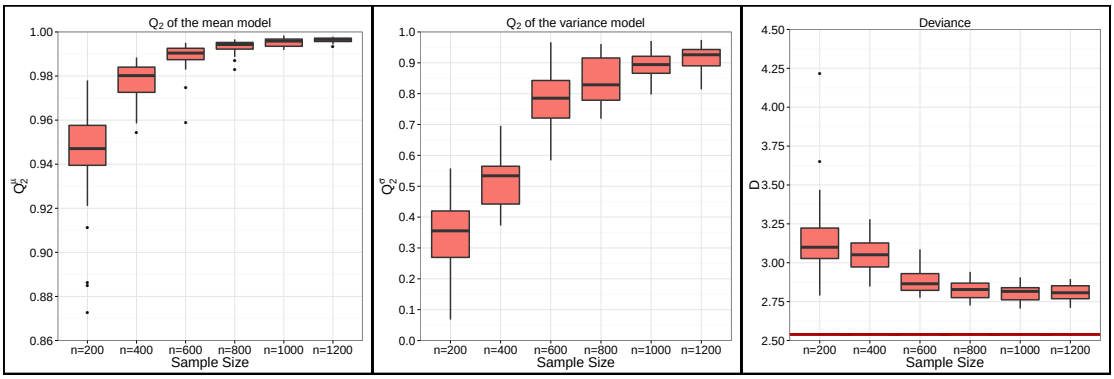

Figure 1: Boxplots of the predicivity results from example 1

where

$$g_1(t) = t; \quad g_2(t) = (2t-1)^2; \quad g_3(t) = \frac{\sin(2\pi t)}{2-\sin(2\pi t)};$$
$$g_4(t) = 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin^2(2\pi t) + 0.4\cos^3(2\pi t) + 0.5\sin^3(2\pi t);$$
$$g_5(t) = 0.8t^2 + \sin^2(2\pi t); \quad g_6(t) = \cos^2(\pi t);$$

Therefore $x^{(5)}$ is uninformative for the mean function and $x^{(3)}$, $x^{(4)}$ and $x^{(5)}$ are uninformative for the variance function. Notice that for this example we considered an additive model.
Figure 1 summarizes the results for 100 learning samples using 6 different sizes (n=200, n=400, n=600, n=800, n=1000 and n=1200). The learning points are sampled $\sim \mathbb{U}[0;1]$ by using the Latin Hypercube Design procedure (McKay et al., 1979) with maximin criterion (Santner et al., 2003) (maximinLHD). The JM COSSO run with the additive model. It appears that the estimation of the variance function is more complicated than the estimation of the mean. Indeed, by looking at Q_2 measures we can see that the mean estimators are accurate even for small size of the learning sample, while the estimation of the variance function requires a higher number of simulations to obtain a good accuracy. We can also note from Figure 1 that more the joint model is accurate and more the deviance is close to the expected value of an optimal model (represented by horizontal red lines). This result shows the relevance of using deviance measures as criterion to evaluate the goodness of the prediction accuracy in the framework of the joint modelling. Indeed, even if Q_2^μ and Q_2^σ provides more quantitative information about the prediction accuracy, in practice we can not use this measures because the observations of the mean and of the variance are not provided.
Table 1 and Table 2 respectively show for the mean and variance models the number of times each variables appears in the 100 models for each learning sample size. We state that the variables do not appear in the models if theirs θ's are smaller than 10^{-5}. Starting from the size $n = 400$ for the mean model and from $n = 600$ for the variance function the JM COSSO do not miss any influential variables in the chosen model. For the variance model the JM COSSO tend to include non-influential variables. However, the frequency of inclusion of non-influential variables decrease when the learning sample size increases.

Let's now study just one estimator built on a learning sample of size $n = 1200$. The predictivity (using the test sample of size $n_{test} = 10000$) measures of this estimator are: $Q_2^\mu = 0.998$,

	1	2	3	4	5
$n = 200$	100	92	100	100	0
$n = 400$	100	100	100	100	1
$n = 600$	100	100	100	100	0
$n = 800$	100	100	100	100	0
$n = 1000$	100	100	100	100	0
$n = 1200$	100	100	100	100	0

Table 1: Frequency of the appearance of the inputs in the mean models from example 1

	1	2	3	4	5
$n = 200$	68	100	28	40	20
$n = 400$	88	100	8	24	12
$n = 600$	100	100	10	16	4
$n = 800$	100	100	8	8	8
$n = 1000$	100	100	1	3	0
$n = 1200$	100	100	0	2	0

Table 2: Frequency of the appearance of the inputs in the variance models from example 1

$Q_2^g = 0.96$ and deviance $D = 2.707$. In figure 2(a) we can see the true standardized residuals versus the true mean, which are computed using the analytical formula of the mean and variance function. While in figure 2(b) the predicted standardized residuals versus the predicted mean are shown. These figures show that the predicted standardized residuals have the same dispersion structure as the true residuals. The cross plot of the figure 2(c) confirms the goodness of the mean model. In figure 2(d) normal quantile-quantile plots (QQ-plot) of the predicted standardized residuals are represented. It is obvious that the predicted residuals have a good distribution result. This is confirmed by the figure 2(e), which represents a QQ-plot of the true standardized residuals versus the predicted standardized residuals. Figure 3 depicts the data from the previous realization along with the true components curves and the estimated ones of the mean model. It can be seen that the JM COSSO fits very well the functional components of the mean function. Figure 4 displays the true components curves and the estimated ones of the variance model. Here we see that the JM COSSO procedure estimates reasonably well the functional of the variance model. Note that in Figure 3 and Figure 4 the components are centered according to the ANOVA decomposition.

4.3. Example 2

Let us consider this 5 dimensional regression problem, where the mean function is defined as

$$\mu(\mathbf{x}) = 5g_1(x^{(1)}) + 3g_2(x^{(2)}) + 4g_3(x^{(3)}) + 6g_4(x^{(4)}) + 3g_1(x^{(3)}x^{(4)}) + 4g_2(\frac{x^{(1)} + x^{(3)}}{2}) \quad (44)$$

and the variance function is defined as

$$\sigma^2(\mathbf{x}) = 25g_5(x^{(1)})g_6(x^{(2)})g_6(x^{(1)}x^{(2)}) \quad (45)$$

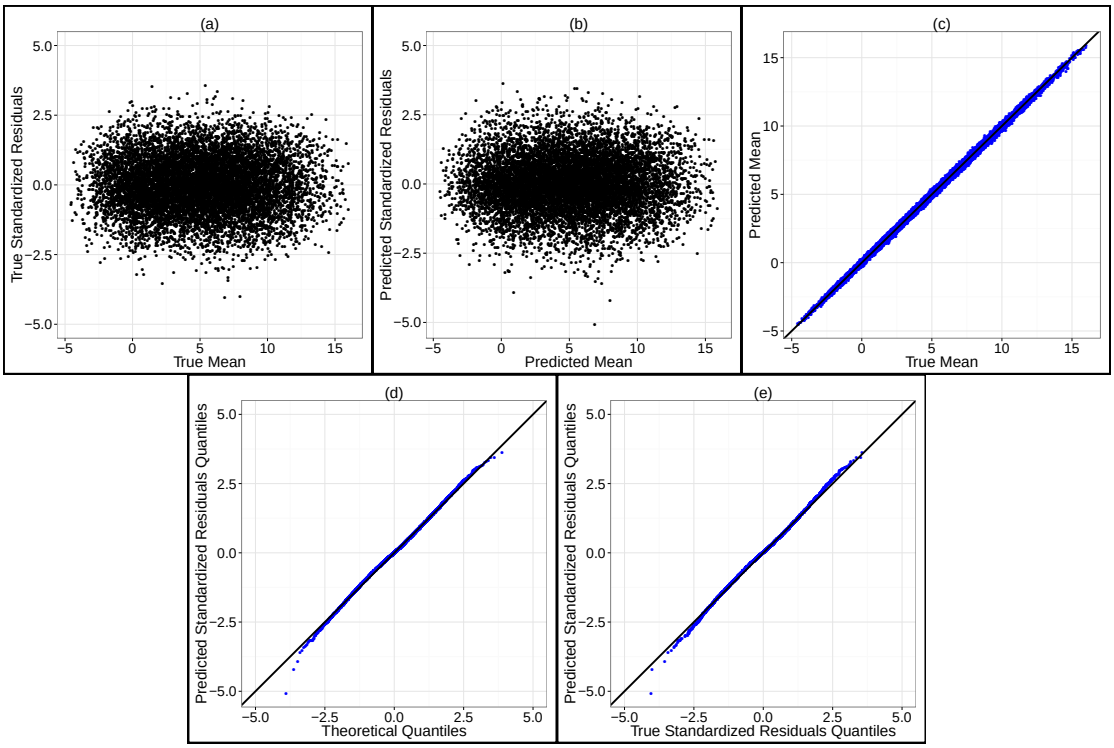

Figure 2: (a) True standardized residuals versus true means (computed using the analytical form of the example 1). (b) Predicted standardized residuals versus predicted means. (c) Cross plot of the predicted mean versus the true mean. (d) Normal quantile-quantile plot of predicted standardized residuals. (e) quantile-quantile plot of predicted standardized residuals versus true tandardized residuals

Variable $x^{(5)}$ is uninformative for the mean function and $x^{(3)}$, $x^{(4)}$ and $x^{(5)}$ are uninformative for the variance function. Notice that in this example several two way interactions are present. As for the previous example we consider 100 learning samples using 6 different sizes (n=200, n=400, n=600, n=800, n=1000 and n=1200). The learning points are sampled $\sim \mathbb{U}[0;1]$ by using the maximinLHD. Figure 5 summarizes the predictivity results of the considered realizations. Note that as for the previous example the estimation of the variance function is more complicated than the estimation of the mean. Indeed, the JM COSSO does suffer from small sample size to estimate the variance function. The deviance measure decreases following the same trend as the Q_2^σ. The lowest value for the deviance that we obtain is approximately equal to 3, which seems to be relatively good result.
Table 3 and Table 4 respectively show for the mean and variance function the performance of the JM COSSO in terms of model selection. For the mean function we can notice that the two way interaction between the inputs 3 and 4 is much harder to detect than the other effects. This is due to the small impact of this interaction effect. For the variance function the JM COSSO procedure does not miss any influential inputs starting from $n = 600$. However, as remarked in the previous example the JM COSSO tends to include non-influential components, but the frequency of this inclusion decreases as the sample size increases.

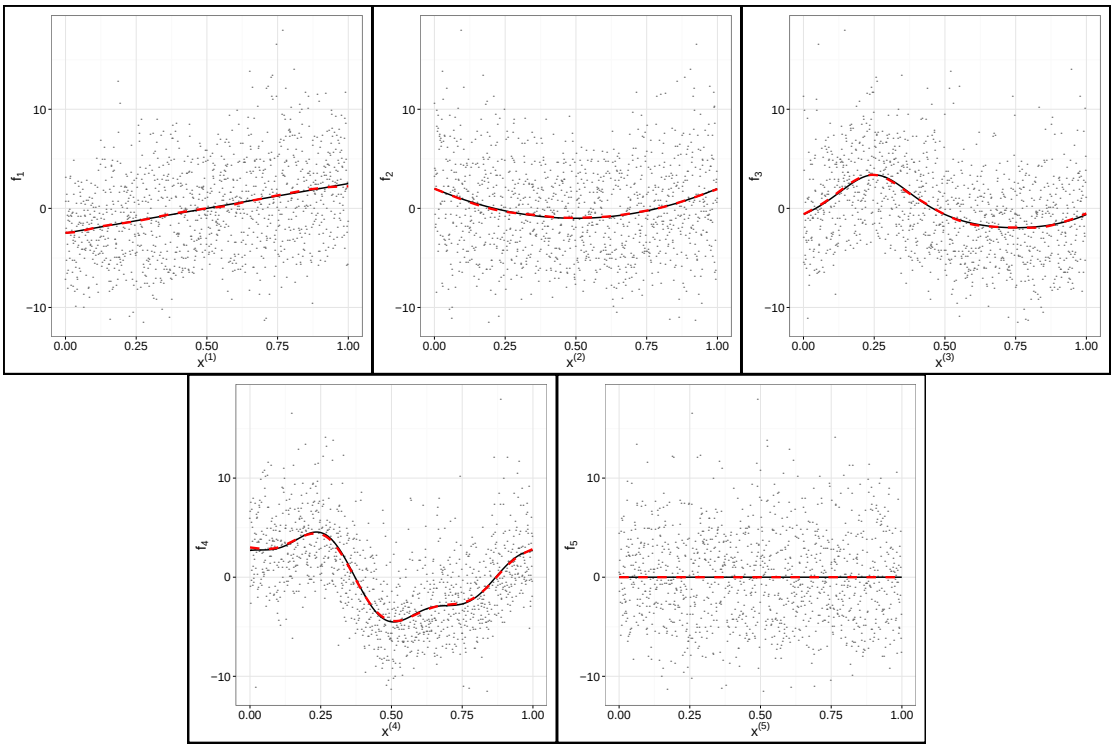

Figure 3: Scatterplots of the data generated from example 1 along with the true components curves (solid) and the estimated ones (dashed) of the mean function across each of the five inputs

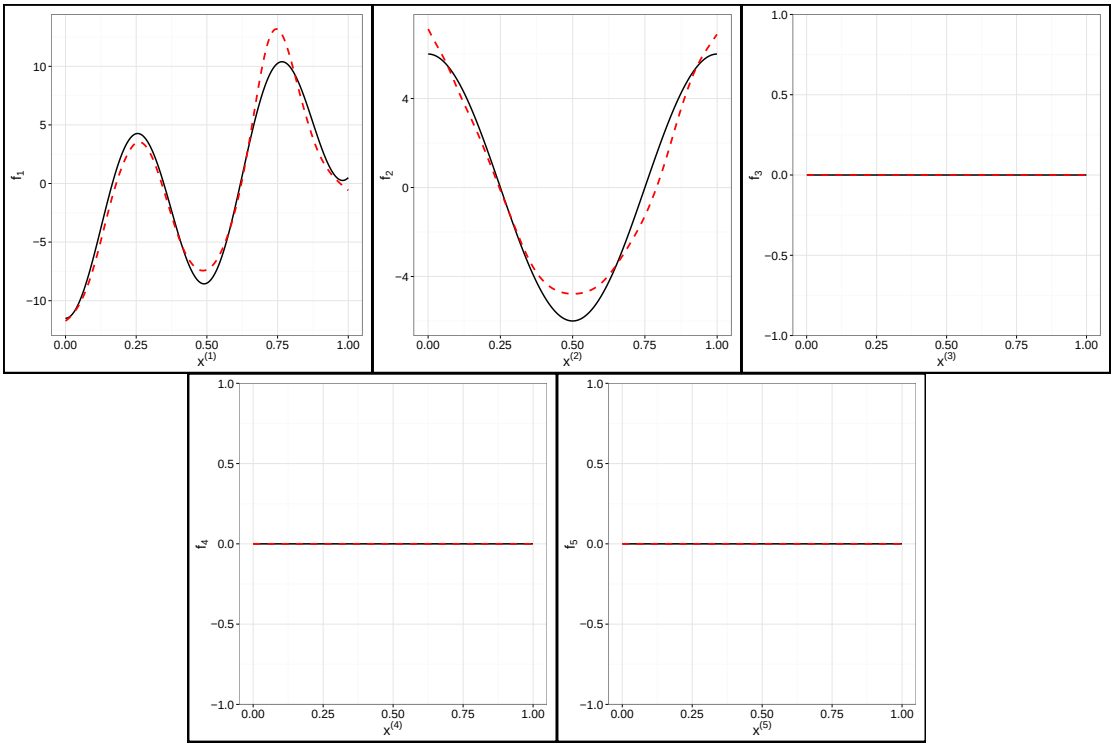

Figure 4: Plots of the the true components curves (solid) and the estimated ones (dashed) of the variance function from example 1 across each of the five inputs

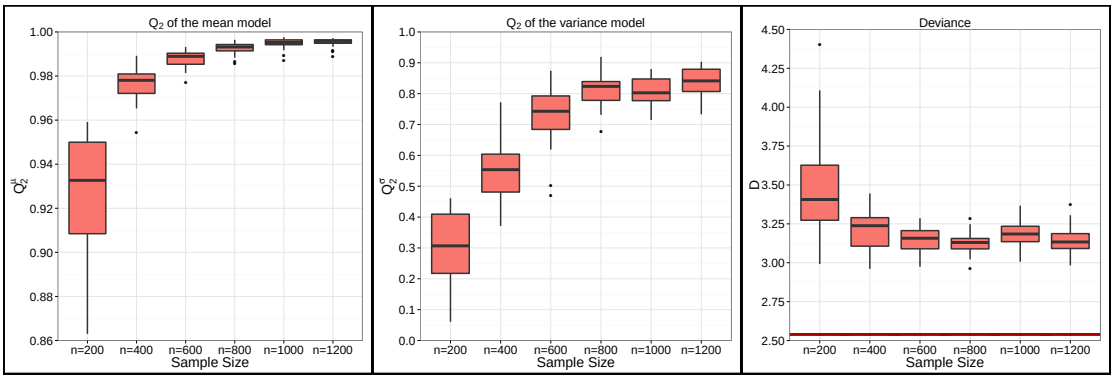

Figure 5: Boxplots of the predicivity results from example 2

	1	2	3	4	5	1:2	1:3	1:4	1:5	2:3	2:4	2:5	3:4	3:5	4:5
$n = 200$	100	92	100	100	5	11	92	2	0	4	0	0	36	4	5
$n = 400$	100	100	100	100	4	0	100	0	0	6	3	0	64	3	0
$n = 600$	100	100	100	100	0	0	100	0	0	3	0	0	80	0	0
$n = 800$	100	100	100	100	2	0	100	1	0	4	1	1	100	1	0
$n = 1000$	100	100	100	100	0	1	100	0	0	0	1	0	100	0	0
$n = 1200$	100	100	100	100	0	0	100	0	0	1	0	0	100	0	0

Table 3: Frequency of the appearance of the inputs in the mean models from example 2

	1	2	3	4	5	1:2	1:3	1:4	1:5	2:3	2:4	2:5	3:4	3:5	4:5
$n = 200$	40	100	20	24	11	20	13	14	7	5	10	7	12	12	13
$n = 400$	83	100	9	30	14	100	7	9	8	5	6	3	8	6	12
$n = 600$	100	100	5	28	8	100	5	0	4	4	4	1	0	1	4
$n = 800$	100	100	7	12	0	100	0	0	2	2	2	1	2	1	1
$n = 1000$	100	100	8	8	0	100	2	1	0	2	3	1	3	0	2
$n = 1200$	100	100	1	6	0	100	0	0	1	1	0	0	2	0	0

Table 4: Frequency of the appearance of the inputs in the variance models from example 2

5. Application

In this section we focus our interest on the reservoir simulators, which models the physical laws governing the recovery process that are mainly modelled by mathematical equations for the phases flow (oil, gas and water) through porous media. These equations are solved by numerical methods over discrete computational grids. In order to get more accurate solutions the reservoir simulators involve higher number of grid cells and an increasing number of reservoir details. The accuracy of the model predictions depends on the input data accuracy, so if there are uncertainties on input parameters, the simulator forecasts will be uncertain. Indeed, input parameters are generally related to the geological properties of the reservoirs and to the production scheme development. The information gathered in such inputs comes from direct measurements, which are clearly very limited and marred by considerable uncertainties. Thus to be able to predict accurately the production, it is important to identify the uncertainty

sources that are relevant to a particular simulator output. It is also important to quantify the amount of the uncertainty. Indeed, once identified one can reduce the complexity of the model by fixing the non-influential inputs on default values, which are defined by experts, and focus the attention on the influential inputs.
To obtain a more realistic reservoir description geostatistical simulation is widely used since it is well suited to model geological information. However, in terms of production forecasts, this modelling induces a large amount of uncertainty since several equiprobable realizations may fit the available geological information, but each realization may produce a different production result. Thereby, relation between the multiplicity of the possible geological models and the studied output must be approximated in order to quantify the corresponding amount of uncertainty. A simple way to treat the uncertainty due to the geostatistic context is to analyze each realization as an independent scenario. In this case the classical meta-modelling techniques can be applied, because for each scenario the reservoir simulator is deterministic. However, in this context the interactions between the classical input parameters and the geostatistical realizations are neglected. Another possible way is to consider the seed corresponding to each realization as a classical input. But in this case the effect due to the variation of the seed is too complex to be described by a reasonable number of scalar. Consequently, a specific approach that combines the impact of both stochastic and deterministic uncertainty must be considered. In this work we consider the reservoir simulator as a stochastic computer model, which means that the same input parameters set leads to different output values. In that case the seed parameters is considered to be a stochastic effect parameters.
We applied the developed joint model COSSO to the PUNQS test case, which is a synthetic reservoir model taken from a real field located in the North Sea. This test case is frequently used as a benchmark reservoir engineering model for uncertainty analysis and for history-matching studies, see PUNQS (1996). The geological model contains $19 \times 28 \times 5$ grid blocks, 1761 of which are active. The reservoir is surrounded by a strong aquifer in the North and the West, and is bounded to the East and South by a fault. A small gas cap is located in the centre of the dome shaped structure. The geological model consists of five independent layers, where the porosity distribution in each layer was modelled by geostatistical simulation. The layers 1, 3, 4 and 5 are assumed to be of good quality, while the layer 2 is of poorer quality. The field contains six production wells located around the gas-oil contact. Due to the strong aquifer, no injection wells are required. Five uncertain parameters uniformly distributed and independent, are considered in this study

- $AQUI \sim \mathbb{U}[0.2; 0.3]$ (adimensional): Analytical porosity of the aquifer strength
- $MPH1 \sim \mathbb{U}[0.8; 1.2]$ (adimensional): horizontal transmissibility multiplier for layers 1
- $SORW \sim \mathbb{U}[0.15; 0.25]$ (adimensional): critical saturation values used in endpoint scaling
- $P1X \sim \mathbb{U}[6; 11]$ (Cell): X coordinate of well PRO-1
- $P1Y \sim \mathbb{U}[21; 23]$ (Cell): Y coordinate of well PRO-1

In addition, the stochastic effect parameter is considered to be petrophysical maps generated using geostatiscs (in this study 10 different maps are generated). The studied output is the reservoir cumulative oil production at surface conditions after 12 years of production.
Each of the input range has been rescaled to the interval $[0, 1]$ and the reservoir simulator is run

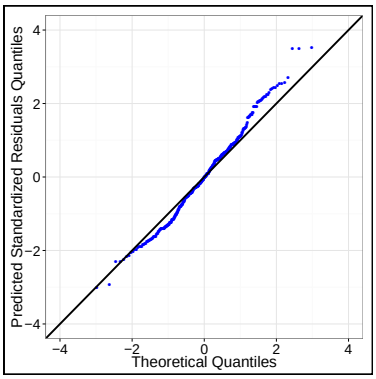

Figure 6: Normal quantile-quantile plot of predicted standardized residuals (PUNQS test case).

on two samples of size $n = 600$ and $n = 350$ that were built using maximinLHD. It is important to note that the samples have been built for a problem of dimension 6. Indeed, for each simulation an independent realization of the porosity map is randomly chosen (with maxminLHD criterion) among the 10 available porosity map realizations. Then, a joint COSSO estimator has been built using as a learning sample the sample of size $n = 600$. To asses the prediction quality, the deviance has been computed using the sample of size $n = 350$. This deviance is equal to 2.641, which is close to the expected deviance of an optimal model (2.541). In figure 6 normal QQ-plot of standardized residuals is presented. It is clear from the figure 6 that the Gaussian assumption provides a satisfactory result.

6. Conclusion

In this article, a new joint modelling of mean and variance method is presented, which is based on a doubly penalized COSSO likelihood. This heteroscedastic regression method performs model fitting and variable selection for mean and variance, which can help to increase efficiency in estimation. Numerical tests performed on analytical examples and on an application from petroleum reservoir engineering, showed that the method gives good results in terms of predictivity and model selection. We have also demonstrated the pertinence of using the deviance as a criterion for evaluating the goodness of the estimated joint model.
To improve the performance of JM COSSO procedure a future research topic will be to use an adaptive weight in the doubly COSSO penalty (Storlie et al., 2011) which may allows for more flexibility to estimate influential functional components and in the same time providing heavier penalty to non-influential functional components.

Acknowledgements

The authors are grateful to Doctor Amandine Marrel and Doctor Sébastien Da-Veiga for many useful suggestions and helpful discussions.

References

Andersen, T. G., Lund, J., 1997. Estimating continuous-time stochastic volatility models of the short-term interest rate. Journal of Econometrics 77 (2), 343 – 377.

Box, G., 1988. Signal-to-noise ratios, performance criteria, and transformations. Technometrics 30 (1), 1–17.

Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37, 373–384.

Carroll, R. J., 1982. Adapting for heteroscedasticity in linear models. The Annals of Statistics 10 (4), 1224–1233.

Cawley, G. C., Talbot, N. L., Foxall, R. J., Dorling, S. R., Mandic, D. P., 2004. Heteroscedastic kernel ridge regression. Neurocomputing 57, 105 – 124.

Fan, J., Yao, Q., 1998. Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85 (3), 645–660.

Fan, S.-K. S., 2000. A generalized global optimization algorithm for dual response systems. Journal of quality technology 32, 444–456.

Gallant, A. R., Tauchen, G., 1997. Estimation of continuous-time models for stock returns and interest rates. Macroeconomic Dynamics 1, 135–168.

Gijbels, I., Prosdocimi, I., Claeskens, G., 2010. Nonparametric estimation of mean and dispersion functions in extended generalized linear models. TEST 19, 580–608.

Gu, C., 2013. Smoothing Spline ANOVA Models. Springer.

Hall, P., Carroll, R. J., 1989. Variance function estimation in regression: the effect of estimating the mean. Journal of the Royal Statistical Society. Series B 51 (1), 3–14.

Huang, X., Pan, W., 2002. Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. Functional and Integrative Genomics 2, 126–133.

Juutilainen, I., Roning, J., 2008. Modelling conditional variance function in industrial data: A case study. Statistical Methodology 5 (6), 564 – 575.

Juutilainen, I., Roning, J., 2010. How to compare interpretatively different models for the conditional variance function. Journal of Applied Statistics 37 (6), 983–997.

Lin, Y., Zhang, H., 2006. Component selection and smoothing in smoothing spline analysis of variance models. Annals of Statistics 34(5), 2272–2297.

Liu, A., Tong, T., Wang, Y., 2007. Smoothing spline estimation of variance functions. Journal of Computational and Graphical Statistics 16 (2), 312–329.

Marrel, A., Iooss, B., Da-Veiga, S., Ribatet, M., 2012. Global sensitivity analysis of stochastic computer models with joint metamodels. Statistics and Computing 22 (3), 833–847.

McCullagh, P., Nelder, J. A., 1989. Generalized linear models (Second edition). London: Chapman & Hall.

McKay, M. D., Beckman, R. J., Conover, W. J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245.

Nelder, J. A., Lee, Y., 1998. Joint modeling of mean and dispersion. Technometrics 40 (2), 168–171.

PUNQS, 1996. Production forecasting with uncertainty quantification. website. URL `http://www.fault-analysis-group.ucd.ie/Projects/PUNQ.html`

Santner, T. J., Williams, B. J., Notz, W. I., 2003. The design and analysis of computer experiments. Springer.

Smyth, G. K., Verbyla, A. P., 1999. Adjusted likelihood methods for modelling dispersion in generalized linear models. Environmetrics 10 (6), 695–709.

Smyth, G. K., Verbyla, A. P., 2009. Leverage adjustments for dispersion modelling in generalized nonlinear models. Australian and New Zealand Journal of Statistics 51 (4), 433–448.

Storlie, C. B., Bondell, H. D., Reich, B. J., Zhang, H., 2011. Surface estimation, variable selection, and the nonparametric oracle property. Statistica Sinica 21, 679–705.

Tibshirani, R. J., 1996. Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society Series B 58, 267–288.

Touzani, S., Busby, D., 2013. Smoothing spline analysis of variance approach for global sensitivity analysis of computer codes. Reliability Engineering and System Safety 112, 67 – 81.

Wahba, G., 1990. Spline models for observational data. SIAM.

Wang, Y., 2011. Smoothing splines: methods and applications. CRC Press Taylor and Francis Group.

Wang, Y., Guo, S.-W., 2004. Statistical methods for detecting genomic alterations through array-based comparative genomic hybridization (CGH). Frontiers in bioscience 9, 540–549.

Yuan, M., Wahba, G., 2004. Doubly penalized likelihood estimator in heteroscedastic regression. Statistics & Probability Letters 69 (1), 11–20.

Zabalza-Mezghani, I., Manceau, E., Roggero, F., 2001. A new approach for quantifying the impact of geostatistical uncertainty on production forecasts: The joint modeling method. Proceedings of IAMG Conference, Cancun, Mexico, Septembre 6-12.

Zhang, H. H., Lin, Y., 2006. Component selection and smoothing for nonparametric regression in exponential families. Statistica Sinica 16, 1021–1041.