Katholieke Universiteit Leuven

Departement Elektrotechniek ESAT-SISTA/TR 2004-229

Regularized semiparametric model identification

with application to NMR signal quantification with unknown macromolecular baseline[1]

Diana M. Sima[2] Sabine Van Huffel[2]

July 2005
Accepted for publication in
Journal of the Royal Statistical Society: Series B (Statistical Methodology).

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.ac.be* in the directory *pub/sista/dsima/reports/SISTA4-229.pdf*

[2]ESAT-SISTA, K.U. Leuven, Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, Belgium, (`diana.sima,sabine.vanhuffel@esat.kuleuven.ac.be`). This work was supported by **Research Council KUL:** GOA-Mefisto 666, IDO/99/003 and IDO/02/009 (Predictive computer models for medical classification problems using patient data and expert knowledge), several PhD/postdoc & fellow grants; **Flemish Government: FWO:** PhD/postdoc grants, projects, G.0078.01 (structured matrices), G.0407.02 (support vector machines), G.0269.02 (magnetic resonance spectroscopic imaging), G.0270.02 (nonlinear Lp approximation), research communities (ICCoS, ANMMM); **IWT:** PhD Grants, **Belgian Federal Science Policy Office:** IUAP P5-22 ("Dynamical Systems and Control: Computation, Identification & Modelling"); **EU:** BIOPATTERN, eTUMOUR.

Abstract

In this paper, we formulate and solve a semiparametric fitting problem with regularization constraints. The model that we focus on is composed of a parametric nonlinear part and a nonparametric part that can be reconstructed via splines. Regularization is employed in order to impose a certain degree of smoothness on the nonparametric part.

Semiparametric regression is presented in this paper as a generalization of nonlinear regression, and all important differences that arise from the statistical and computational points of view are highlighted. We motivate the problem formulation with a biomedical signal processing application.

Keywords: nonlinear regression, regularization, semiparametric, smoothing, splines.

Regularized semiparametric model identification
with application to NMR signal quantification with unknown macromolecular baseline

Diana M. Sima†

Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium.

Sabine Van Huffel

Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium.

Summary. In this paper, we formulate and solve a semiparametric fitting problem with regularization constraints. The model that we focus on is composed of a parametric nonlinear part and a nonparametric part that can be reconstructed via splines. Regularization is employed in order to impose a certain degree of smoothness on the nonparametric part.
Semiparametric regression is presented in this paper as a generalization of nonlinear regression, and all important differences that arise from the statistical and computational points of view are highlighted. We motivate the problem formulation with a biomedical signal processing application.

Keywords: nonlinear regression, regularization, semiparametric, smoothing, splines.

1. Introduction

This paper is a survey dedicated to a semiparametric fitting problem with regularization constraints. We consider nonlinear models that are partially known, partially unknown. From given noisy measurements, we are interested in estimating regression parameters of the known nonlinear part, as well as estimating the nuisance nonparametric part.

Our motivation for studying this problem comes from an application in nuclear magnetic resonance (NMR) spectroscopy, namely the problem of quantifying metabolite concentrations from short-echo time *in vivo* NMR measurements, *e.g.*, measured proton spectra from the human brain (Provencher, 1993; Lemmerling et al., 2002). The goal of this application is to compute the parameters of a certain model function, which give information about the concentrations of chemical substances in a region of the brain. With improved NMR scanner resolution nowadays, more and more information is captured in the measured NMR signal. This means that not only the most relevant metabolites in the brain, but also a nuisance 'baseline' for which no model function is available must be taken into account in automated quantification methods. For this reason, this application requires semiparametric modeling.

In the context of semiparametric modeling, much work has been devoted to partially linear stochastic models of the form

$$y_i = F(t_i)^\top \theta + g(t_i) + \varepsilon, \qquad (i = 1, 2, \ldots, m),$$

†*Address for correspondence:* Diana Sima, ESAT-SCD, Dept. Electrical Engineering, K.U. Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium.
E-mail: diana.sima@esat.kuleuven.be

where $y_1,\ldots,y_m \in \mathbb{R}$, as well as $F(t_1),\ldots,F(t_m) \in \mathbb{R}^p$ are measured quantities, ε_i denotes the measurement noise; $\theta \in \mathbb{R}^p$ is the linear regression vector to be estimated, and $g : \mathbb{R} \to \mathbb{R}$ is a nonlinear nonparametric part. Usually, the goal is to find a consistent estimate of θ, while considering the nonparametric part $g(\cdot)$ as nuisance (see, *e.g.*, a very recent comprehensive study of semiparametric models presented in the book (Ruppert et al., 2003); see also (Bhattacharya and Zhao, 1997)). The nonlinearity $g(\cdot)$ is often modeled by smoothing splines (Wahba, 1990). Smoothing splines are sometimes employed together with a nonlinear model as a way of testing if the nonlinear model is adequate (Wahba, 1990, Chapter 9): an almost zero reconstructed spline means that the nonlinear model is the right one. Another example of semiparametric modeling is presented by Carter and Kohn (1997), which employ a Bayesian framework in order to discriminate nonparametrically between a deterministic regression function and a noise term with smooth spectral density. In (Ke and Wang, 2004), a very general spline framework is developed such that nonlinear relationships are allowed between several nonparametric functions.

The semiparametric regression problem treated in this paper combines additively parametric nonlinear regression with nonparametric spline smoothing. It is shown that it is adequate to solve such a semiparametric fitting problem using nonlinear least squares plus a penalty on the spline smoothness. Although this semiparametric problem formulation can be seen as a special case of the nonlinear nonparametric regression models in (Ke and Wang, 2004), the goal of this presentation deviates from the focus in (Ke and Wang, 2004). There, the main issue gravitates around the estimation of nonparametric functions, while here we mainly consider the nonparametric part as nuisance and we are interested in obtaining estimates for the nonlinear regression parameters of the known part of the model, as precisely as possible. As in nonlinear regression, we can infer asymptotic properties of the semiparametric regression estimates. Under Gaussian noise assumption, normality of the estimates is recovered; however, because of the regularization term, the computed parameters will be biased from the "true" values. We develop detailed bias and covariance formulas, which allow derivation of other statistically relevant information, such as confidence intervals.

We give an algorithmic outline of regularized semiparametric regression, with emphasis on efficient computation. One of the main issues in this context is the choice of the *regularization parameter* that controls the trade-off between nonlinear misfit minimization and effective regularization. We propose an automated iterative selection method that is based on the classical generalized cross-validation criterion. The method is data-driven and does not need prior estimates for the noise statistics.

The outline of this paper is as follows. We start with the description of our motivating application, explaining the challenges that this problem offers and the need for developing specialized techniques for approaching it in a correct way. Then in Section 3 we switch to theory, with a compressed background material on the topics of nonlinear regression and spline smoothing. The semiparametric problem formulation that we propose, its theoretical solution, as well as computational issues are discussed in Section 4. In Section 5, its statistical properties are developed, and in Section 6 we discuss to which types of problems we can safely apply this framework without having to deal with identifiability and non-uniqueness issues. Finally, in Section 7 we show through simulation examples the performance as well as the limitations of the method.

A word about notation: We shall use the bold-face notation for some vectors, for example: $\mathbf{y}$ is the vector $(y_1,\ldots,y_m)^\top$, and $\mathbf{F}(\theta) = (F(t_1,\theta),\ldots,F(t_m,\theta))^\top$.

2. Motivating application: NMR spectroscopy data quantification with unknown macromolecular baseline

The NMR spectrometer outputs complex-valued time-domain signals that have a decaying pattern. Figure 1 shows a typical preprocessed time-domain signal obtained *in vivo* from a healthy selected small volume in a human brain. For visual interpretation, the Fourier transformed signal is usually plotted in the frequency domain (see Figure 2), because the position and magnitude of each peak gives information about the chemicals present in the sample.

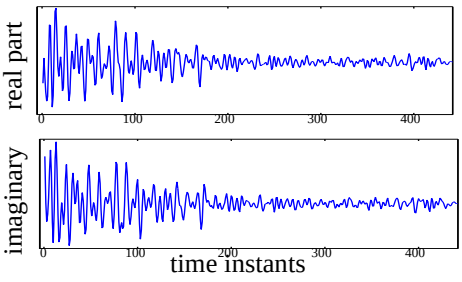


Fig. 1. The time domain NMR signal measured from a selected volume in the human brain.

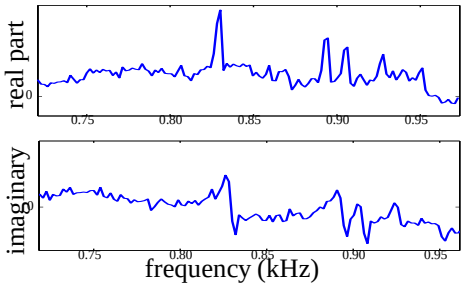


Fig. 2. The frequency domain spectrum measured from a selected volume in the human brain.

Each pure metabolite (chemical substance) has a peculiar time response that depends on the number and position of hydrogen protons in the molecular structure; in theory, the time response is a sum of complex damped exponentials, which yield the typical Lorentzian peaks in the frequency domain signals. Spectra of metabolites that are known to be relevantly present in the human brain can be also measured *in vitro*. Such measurements can be grouped together in a database of metabolite signals, see Figure 3. An *in vivo* signal can be modeled as a combination of metabolites

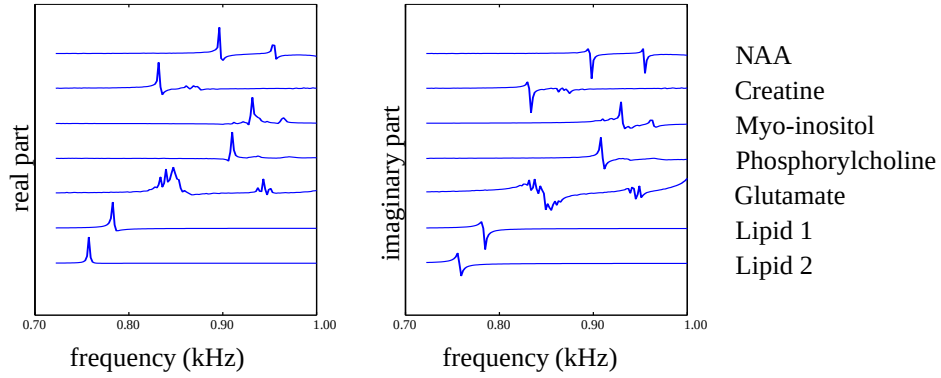


Fig. 3. The frequency domain spectrum profile for several typical metabolites in the human brain.

in the database. In this way, the weighting coefficients (amplitudes) in the combination yield the *concentrations* of the metabolites. Estimating these concentrations is the main goal of the NMR spectroscopy data quantification problem. However, this combination cannot simply be a linear combination; one should allow small corrections in other spectral parameters (frequency shifts,

damping corrections, phase shifts, etc), since these parameters may slightly vary from measurement to measurement (Ratiney et al., 2004).

Denote by $\{v_k, \text{ for } k = 1,\ldots,K\}$ the K given complex-valued time series of length T, representing *in vitro* measured metabolites, and by w the *in vivo* measured NMR signal. The model that allows spectral corrections in the combination of metabolites is:

$$w(t) = \widehat{w}(t) + \varepsilon_t := \sum_{k=1}^{K} \alpha_k \, \zeta_k^t \, v_k(t) + \varepsilon_t, \quad t = 0,\ldots,T-1, \tag{1}$$

where ε_t is an unknown noise perturbation with zero mean, for all t's from 0 to $T-1$ and $\alpha_k, \zeta_k \in \mathbb{C}$ can also be written in real coordinates as $\alpha_k = r_k \exp(j\phi_k)$, $\zeta_k = \exp(-d_k + 2\pi j f_k)$, with $r_k, \phi_k, d_k, f_k \in \mathbb{R}$ representing the amplitude, the phase correction, the damping correction and, respectively, the frequency correction for each metabolite signal, and $j = \sqrt{-1}$.

The identification of complex amplitudes α_k, and complex ζ_k's (or, equivalently, the real parameters r_k, ϕ_k, d_k, f_k), for $k = 1,\ldots,K$, can be accomplished by minimizing the least squares criterion: $\|w - \widehat{w}\|_2^2$. Possibly, additional constraints, such as linear bounds on the real parameters, are required; for instance, it is expected that the damping and frequency shifts are bounded within a small interval around zero, since otherwise some metabolite patterns can become interchangeable.

Figure 4 shows the linear fit, as well as the nonlinear fit that uses spectral corrections, of the same *in vivo* signal as plotted before with the signals from the metabolite database. (In practice, we might have up to 25 metabolites in the database, but here, for illustration, we consider only the 7 metabolites given in Figure 3.) It is noticeable that the residuals are highly biased from the zero line, in both fits. This happens because the pure metabolite frequency-domain profiles are zero outside the range of resonant frequencies, but this is different in what concerns an *in vivo* signal. The database

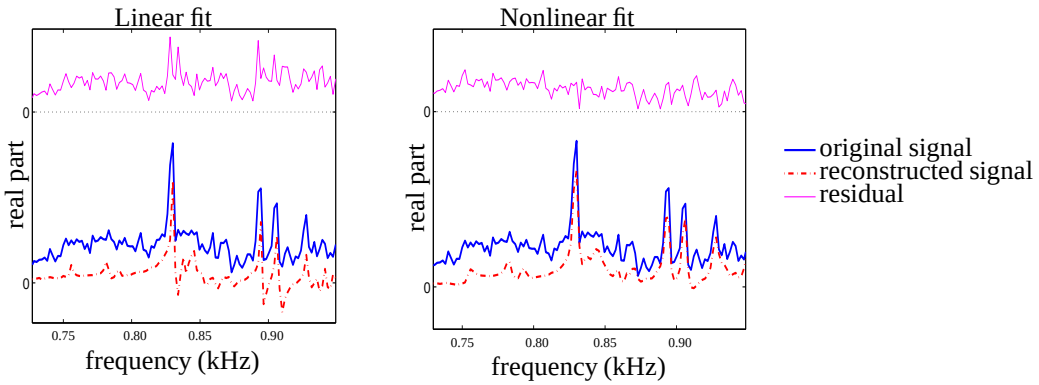


Fig. 4. Real part of the original frequency domain spectrum from an *in vivo* signal, together with the optimal reconstructed spectrum using a combination of metabolites from the database. Left: linear combination. Right: linear combination with nonlinear spectral corrections.

of metabolite profiles only contains the most prominent and relevant metabolites, but the *in vivo* signal also holds information that originates from some unknown 'macromolecules'. This nuisance part is known as the *baseline*. The baseline is a highly damped component; seen in the frequency domain, it is a smooth, broadband, low amplitude spectrum, that gives an underlying trend to the *in vivo* signal. (See Figure 5.) Its shape can vary and it is in general unpredictable, especially in pathological cases. In mathematical terms, we can say that the baseline is characterized by the fact

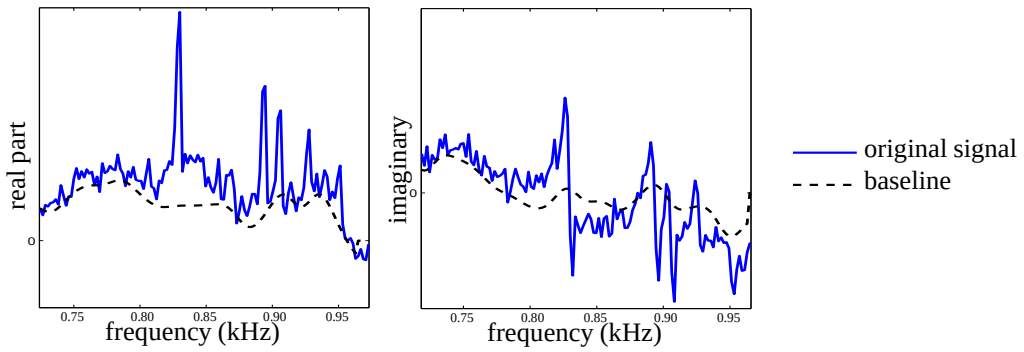


Fig. 5. The original frequency domain spectrum from an *in vivo* signal, together with a baseline constructed with splines.

that its Fourier transformation should be a smooth function‡. For this reason, a good choice is found in identifying it as a smooth curve using splines in the frequency domain. We denote the baseline with $b(\cdot)$ and the model becomes

$$w(t) = \widehat{w}(t) + \varepsilon_t := \sum_{k=1}^{K} \alpha_k \zeta_k^t v_k(t) + b(t) + \varepsilon_t, \quad t = 0, \ldots, T-1.$$

We fit the exemplified data using this semiparametric model and we apply the method to be described in Section 4. The result is depicted in Figure 6 (only the real part, for more clarity). Observe that the bias that was visible in the residuals of Figure 4 is now removed.

Introducing the baseline term into the model proves to be an adequate solution for fitting the given data. But does this also imply that a more reliable estimation of the parameters of interest (and, finally, of the metabolite concentrations) can be obtained? In the assumption that the reconstructed baseline doesn't model the part of the data that is characteristic for the metabolites of interest, the answer is yes. An important problem is deciding how many features of the data could be allowed to be fitted with the baseline. The trade-off between the parametric part of the model and the baseline can be viewed, in this application, as a trade-off in the smoothness of the frequency domain baseline. Choosing the right order of smoothness will be dealt with in Section 4, in a general setting.

3. Theoretical background

In order to introduce more easily the semiparametric setting, we start by shortly reviewing some elements of the nonlinear regression and the nonparametric regression theories.

3.1. Nonlinear regression

When observed variables exhibit nonlinear relations among each other, estimation methods are provided by the nonlinear regression theory. Nonlinear least squares, maximum likelihood, quasi likelihood, or Bayesian estimation methods are typical techniques that can be defined as well in a

‡There is an abuse of terms here: actually, the Fourier transform of the unknown *continuous-time* baseline should be a smooth function. In practice, we work with discrete time instants and with the discrete Fourier transform.

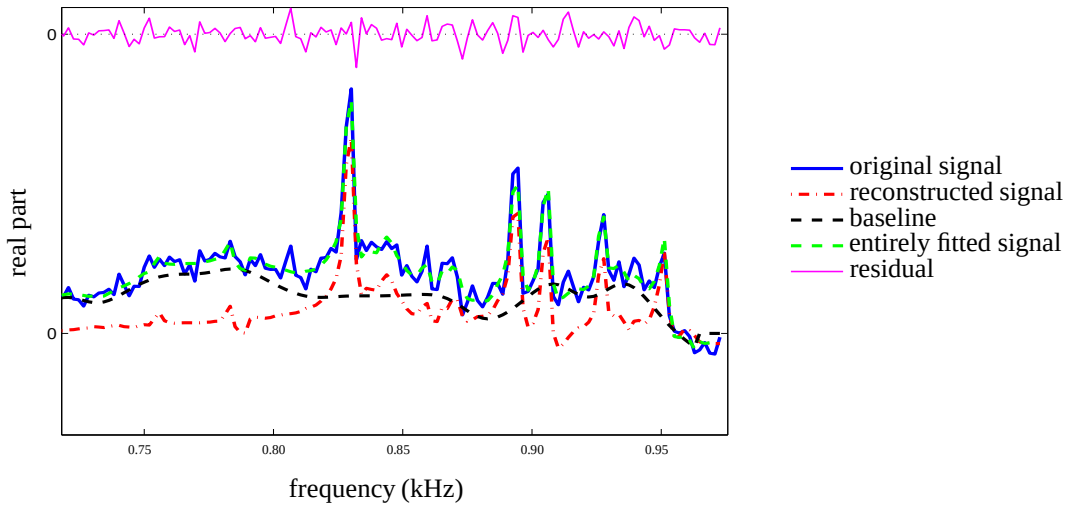


Fig. 6. Real part of the original frequency domain spectrum from an *in vivo* signal, together with the optimal reconstructed spectrum using the metabolites from the database, the optimal baseline constructed with penalized splines and the entirely reconstructed signal (metabolites plus baseline).

nonlinear setting, similarly to the case of linear regression (Seber and Wild, 1989). However, more difficult problems with identifiability, ill-conditioning, convergence of numerical algorithms, design of confidence intervals, etc., are encountered in nonlinear regression than in the linear case.

Under regularity assumptions§, nonlinear regression has also desirable (asymptotic) properties. The next statement is related to the nonlinear least squares estimation in the Gaussian noise case:

THEOREM 1 (ADAPTED FROM (SEBER AND WILD, 1989), CHAPTER 1). *Given m observations $(\mathbf{t}_i, y_i) \in \mathbb{R}^q \times \mathbb{R}$, $i = 1,2,\ldots,m$, from a nonlinear model with known functional relationship $F : \mathbb{R}^q \to \mathbb{R}$,*

$$y_i = F(\mathbf{t}_i, \theta^\star) + \varepsilon_i \qquad (i = 1,2,\ldots,m),$$

where $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$ and $\theta^\star$ denotes the true value of the unknown parameter $\theta \in \Theta \subset \mathbb{R}^p$, then the least squares estimate of $\theta^\star$, i.e., *the global minimizer $\widehat{\theta}$ of*

$$E(\theta) := \sum_{i=1}^{m} (y_i - F(\mathbf{t}_i, \theta))^2$$

over $\theta \in \Theta$, satisfies:

(a) $\widehat{\theta}$ is a consistent estimate of $\theta^\star$ with $\widehat{\theta} - \theta^\star \sim \mathcal{N}(0, \sigma^2(\mathscr{A}^\top \mathscr{A})^{-1})$ where $\mathscr{A} := \left[\frac{\partial F}{\partial \theta}(\mathbf{t}, \theta^\star)\right]$;

(b) $s^2 := E(\widehat{\theta})/(m-p)$ is a consistent estimate of the variance σ^2.

§To ease the exposition, we don't enumerate all assumptions that are needed to prove the asymptotic properties of nonlinear least squares estimation. A complete discussion can be found in (Seber and Wild, 1989, Chapter 12). We keep the same lack of completeness in the next section on semiparametric models, because the classical regularity assumptions are usually rather technical and not essential for the goal of this presentation.

Normality of ε is not required to prove that $\widehat{\theta}$ is consistent; the zero mean condition $\mathscr{E}(\varepsilon) = 0$ is sufficient. This theorem gives the necessary justification of using nonlinear least squares when fitting nonlinear models – under zero-mean noise assumption.

3.2. Spline fitting for nonparametric regression

Nonparametric regression is concerned with modeling a nonlinear function from data. The nonlinear function is not dependent on some regression variables; instead, it may be constrained to belong to a certain set of functions (*e.g.*, the set of smooth functions of a given order d, *i.e.*, functions that have continuous derivatives up to the order d).

Denote by $\mathscr{H}$ a Hilbert space of model functions defined on a given interval $\mathscr{I}$ and by $\langle ,\rangle_{\mathscr{H}}$ its inner product. Let $\mathscr{P}$ be a linear operator from $\mathscr{H}$ to another given space $\mathscr{H}'$, with its inner product $\langle ,\rangle_{\mathscr{H}'}$. $\mathscr{P}$ is chosen such that an inequality of the form $\|\mathscr{P}g\|_{\mathscr{H}'} \leq \delta$ ($\delta > 0$) reflects a desired "smoothness constraint".

Then, given scalar abscissas $t_1, \ldots, t_m \in \mathscr{I}$ and corresponding scalar measurements $y_1, \ldots, y_m$, a *template spline* (Sima and Van Huffel, 2004) is defined as the optimal solution of the minimization problem

$$\min_{g \in \mathscr{H}} \frac{1}{m} \sum_{i=1}^{m} (y_i - g(t_i))^2 + \lambda \, \|\mathscr{P}g\|^2_{\mathscr{H}'}, \tag{2}$$

where $\lambda > 0$ is a regularization parameter that controls the balance between fitting and smoothing via the norm constraint. The formulation (2) is general, but it encompasses the smoothing splines (Wahba, 1990), the penalized splines (Eilers and Marx, 1996), and others.

The main property of (2) is that for a whole range of spaces $\mathscr{H}$, the optimal solution can be obtained by solving a linear least squares problem. In particular, if $\mathscr{H}$ is finite dimensional, then any $g \in \mathscr{H}$ can be represented as $g = \sum_{k=1}^{n} a_k \phi_k$, for some real coefficients $a_1, \ldots, a_n$ and given elements $\phi_1, \ldots, \phi_n \in \mathscr{H}$, which we call "basis functions"; then the problem (2) can be written as:

$$\min_{a_1,\ldots,a_n \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \sum_{k=1}^{n} a_k \phi_k(t_i) \right)^2 + \lambda \left\| \mathscr{P} \sum_{k=1}^{n} a_k \phi_k \right\|^2_{\mathscr{H}'}. \tag{3}$$

The problem (3) is easily transformed to a regularized linear least squares problem (Tikhonov type of regularization):

$$\min_{\mathbf{a} \in \mathbb{R}^n} \frac{1}{m} \|A\mathbf{a} - \mathbf{y}\|_2^2 + \lambda \, \mathbf{a}^\top B \mathbf{a}, \tag{4}$$

with A being the $m \times n$ matrix that has as elements $\phi_k(t_i)$, for k from 1 to n and i from 1 to m and B the $n \times n$ matrix having the elements $\langle \mathscr{P}\phi_k, \mathscr{P}\phi_l \rangle_{\mathscr{H}'}$, for k, l from 1 to n, and with the shorthand notations $\mathbf{y} := (y_1, \ldots, y_m)^\top$ and $\mathbf{a} := (a_1, \ldots, a_n)^\top$. The closed-form optimal solution for a fixed λ is $\widehat{\mathbf{a}}_\lambda = \left(A^\top A + m\lambda B\right)^{-1} A^\top \mathbf{y}$. This solution gives thus a reconstructed model of the form $\widehat{\mathbf{y}}_\lambda := A \left(A^\top A + m\lambda B\right)^{-1} A^\top \mathbf{y}$.

3.2.1. Choosing smoothing parameter λ

Many methods for choosing "optimal" regularization parameter λ can be employed in the context of splines. For nonparametric regression using smoothing splines, a recent simulation study (Lee, 2003) compares several parameter selection procedures (leave-one-out cross-validation, generalized cross-validation, Mallows' C_p, Akaike's information criterion, and risk estimation methods); the conclusion is that none of the methods is overall the best.

A widely-used technique for choosing the λ parameter in the context of smoothing splines is the generalized cross-validation (GCV) criterion of Craven and Wahba (1979). It is a rotation-invariant modification of the classical leave-one-out cross-validation criterion. In mathematical terms, the criterion to be minimized for choosing λ is

$$G(\lambda) := \frac{\|(I_m - A(\lambda))\mathbf{y}\|_2^2}{[\mathrm{Tr}(I_m - A(\lambda))]^2} = \frac{\|\mathbf{y} - \widehat{\mathbf{y}}_\lambda\|_2^2}{[\mathrm{Tr}(I_m - A(\lambda))]^2}, \tag{5}$$

where $A(\lambda)$ is the *influence matrix* that makes the link between the estimated solution $\widehat{g}_\lambda$ of (2) and the vector of measurements, $\mathbf{y} := (y_1, \ldots, y_m)^\top$, through

$$A(\lambda)\mathbf{y} = (\widehat{g}_\lambda(t_1), \ldots, \widehat{g}_\lambda(t_m))^\top.$$

In the case when the solution $\widehat{g}_\lambda$ is expressed as a linear combination of basis functions, then the influence matrix is defined in terms of the optimal coefficient vector $\widehat{\mathbf{a}}_\lambda$ (optimal solution of (4)), and has the formulation $A(\lambda) = A(A^\top A + m\lambda B)^{-1}A^\top$.

4. Semiparametric model with smoothness constraint

4.1. Model formulation

Let $\mathscr{H}$ be a Hilbert space of functions defined on an interval $\mathscr{I} \subset \mathbb{R}$, endowed with an inner product $\langle , \rangle_{\mathscr{H}}$ and an induced norm $\|\cdot\|_{\mathscr{H}}$. Let $\mathscr{P}$ denote an operator from $\mathscr{H}$ to a space $\mathscr{H}'$ (which could be $\mathscr{H}$ itself, $\mathbb{R}^q$ for some $q \in \mathbb{N}$, etc); assuming that $\mathscr{P}$ has a finite dimensional null space is sufficient in general, but here, for simplicity, we assume that even the space $\mathscr{H}$ is finite dimensional.

We consider the following semiparametric model:

$$y_i = F(t_i, \theta^\star) + g^\star(t_i) + \varepsilon_i, \qquad (i = 1, 2, \ldots, m), \tag{6}$$

where $y_1, \ldots, y_m$ are scalar measurements, $F : \mathscr{I} \times \Theta \to \mathbb{R}$ is a known nonlinear model function, parameterized by a vector $\theta \in \Theta \subset \mathbb{R}^p$, $\theta^\star$ is the *true* but unknown value of θ, and $g^\star \in \mathscr{H}$ is a true but unknown nonlinearity.

Given the measurements $y_1, \ldots, y_m$, we are interested in finding optimal approximation $\widehat{\theta}$ of $\theta^\star$ and of the function $g^\star \in \mathscr{H}$, under the criterion:

$$\min_{\theta \in \Theta,\ g \in \mathscr{H}} \frac{1}{m} \sum_{i=1}^{m} (y_i - F(t_i, \theta) - g(t_i))^2 \quad \text{such that} \quad g \in \mathscr{H}^\delta, \tag{7}$$

where $\mathscr{H}^\delta := \{g \in \mathscr{H} : \|\mathscr{P}g\|_{\mathscr{H}'} \leq \delta\}$. The constraint $g \in \mathscr{H}^\delta$ will be referred to as the *smoothing constraint*, since in general the subspace $\mathscr{H}^\delta$ will be defined so that it contains smooth functions. The smoothing constraint can be easily imposed by adding a penalty term to the nonlinear least squares objective function. The constrained least squares problem (7) becomes:

$$\min_{\theta \in \Theta,\ g \in \mathscr{H}} \frac{1}{m} \sum_{i=1}^{m} (y_i - F(t_i, \theta) - g(t_i))^2 + \lambda \|\mathscr{P}g\|_{\mathscr{H}'}^2. \tag{8}$$

4.2. Spline fitting for the nonparametric part

The regularized nonlinear least squares criterion (8) can be rewritten as a double minimization

$$\min_{\theta\in\Theta}\left(\min_{g\in\mathscr{H}}\sum_{i=1}^{m}(y_i-F(t_i,\theta)-g(t_i))^2+\lambda\|\mathscr{P}g\|^2_{\mathscr{H}'}\right),$$

thus it is possible to apply the theory of spline fitting for the nonparametric part g, in order to solve the inner minimization problem.

We use a parameterization of g in terms of a family of generators for $\mathscr{H}$, *i.e.*, $g=\sum_{k=1}^{n}a_k\phi_k$, where $a_1,\ldots,a_n$ are free coefficients, and $\phi_1,\ldots,\phi_n$ are basis functions or generators that span the space $\mathscr{H}$, assumed to be finite dimensional. Thus, the nonparametric part of the model is simply reduced to a linearly parameterized submodel, subject to a regularization constraint. For further reference, here is the *completely parameterized* formulation of the model (6):

$$y_i=F(t_i,\theta^\star)+\sum_{j=1}^{n}A_{ij}a_j+\varepsilon_i,\qquad(i=1,2,\ldots,m),\tag{9}$$

where A is the $m\times n$ matrix that has as elements $\phi_k(t_i)$, for k from 1 to n and i from 1 to m. Using vector notation, the model in problem (9) is written as $\mathbf{y}:=\widehat{\mathbf{y}}+\varepsilon:=\mathbf{F}(\theta)+A\mathbf{a}+\varepsilon$. This semiparametric model will be fitted using the following regularized nonlinear least squares formulation:

$$\min_{\theta\in\Theta,\,\mathbf{a}\in\mathbb{R}^n}\frac{1}{m}\|\mathbf{y}-\mathbf{F}(\theta)-A\mathbf{a}\|^2+\lambda\,\mathbf{a}^\top B\mathbf{a},\tag{10}$$

with B the $n\times n$ matrix having the elements $\langle\mathscr{P}\phi_k,\mathscr{P}\phi_l\rangle_{\mathscr{H}'}$, for k,l from 1 to n, as in subsection 3.2.

For a fixed value of $\lambda\geq 0$, we denote by $\widehat{\theta}_\lambda$ and $\widehat{\mathbf{a}}_\lambda$ the globally optimal solution of the minimization (10). Moreover, we denote with $\widehat{\mathbf{y}}_\lambda$ the optimal model obtained for a fixed λ, that is $\widehat{\mathbf{y}}_\lambda:=\mathbf{F}(\widehat{\theta}_\lambda)+A\widehat{\mathbf{a}}_\lambda$.

REMARK 2. *Choosing the space $\mathscr{H}$ and the penalty term is a separate issue and will not be dealt with here. In particular, we don't put emphasis on the link between the number of regression points and the dimension n of the spline matrix A. This is a typical subject in spline literature. It depends on the type of splines that are used for the nonparametric part whether m and n are of the same magnitude or $n\ll m$.*

4.3. Computationally efficient method using the Levenberg-Marquardt algorithm

Denote by $\mathbf{y}(\theta)$ the m-dimensional vector with elements $y_i-F(t_i,\theta)$, for $i=1,\ldots,m$. Thus, $\mathbf{y}(\theta)=\mathbf{y}-\mathbf{F}(\theta)$.

For a fixed value of the parameter θ, the original nonlinear minimization (10) becomes a regularized linear least squares problem only in $\mathbf{a}\in\mathbb{R}^n$. Its closed-form solution (see subsection 3.2) is $\mathbf{a}(\theta,\lambda)=\left(A^\top A+m\lambda B\right)^{-1}A^\top\mathbf{y}(\theta)$. Plugging-in this formula in the original optimization problem, we get a nonlinear least squares problem in the variable θ alone:

$$\min_{\theta\in\Theta}\frac{1}{m}\|\mathbf{y}(\theta)-A\mathbf{a}(\theta,\lambda)\|_2^2+\lambda\,\mathbf{a}(\theta,\lambda)^\top B\mathbf{a}(\theta,\lambda)\Leftrightarrow$$

$$\min_{\theta\in\Theta}\frac{1}{m}\left\|\left(I_m-A(A^\top A+m\lambda B)^{-1}A^\top\right)\mathbf{y}(\theta)\right\|_2^2+\lambda\left\|D(A^\top A+m\lambda B)^{-1}A^\top\mathbf{y}(\theta)\right\|_2^2\Leftrightarrow$$

$$\min_{\theta\in\Theta}\left\|\begin{bmatrix} I_m - A(A^\top A + m\lambda B)^{-1}A^\top \\ \sqrt{m\lambda}\,D(A^\top A + m\lambda B)^{-1}A^\top \end{bmatrix}\mathbf{y}(\theta)\right\|_2^2, \tag{11}$$

where D denotes a Cholesky factor of B, *i.e.*, $D^\top D = B$. Denote the coefficient matrix under the norm in (11) by $B(\lambda)$.

The minimization problem (11) can be solved using a nonlinear least squares solver, such as the Levenberg-Marquardt (LM) algorithm. We consider the cases when the parameter set Θ is either the full $\mathbb{R}^p$ or it is defined by linear constraints, for which good implementations of the (modified) LM algorithm are available (Moré, 1978; Coleman and Li, 1996).

A nonlinear least squares solver requires at each new θ the evaluation of the function $f(\theta,\lambda) := B(\lambda)\mathbf{y}(\theta)$ and of the Jacobian $J(\theta,\lambda) := B(\lambda)\nabla\mathbf{y}(\theta)$. Efficient computations of these two ingredients are essential for the overall computational time. See Appendix A for more details on how to use generalized singular value decomposition (Golub and Van Loan, 1996) as a preprocessing step and to increase the efficiency of the computations in every iteration.

4.4. Choice of regularization parameter

Although many methods (from the model selection literature) can be adapted to our setting, we concentrate herein only on the derivation of the generalized cross-validation criterion.

To explain the GCV criterion, we start as in (Wahba, 1990) from the leave-one-out cross-validation, which chooses a λ that minimizes the function

$$C(\lambda) := \frac{1}{m}\sum_{i=1}^{m}\left(y_i - F\left(t_i,\widehat{\theta}_\lambda^{[-i]}\right) - A_i\widehat{\mathbf{a}}_\lambda^{[-i]}\right)^2,$$

where A_i is the i^{th} line of A and where $\widehat{\theta}_\lambda^{[-i]}$ and $\widehat{\mathbf{a}}_\lambda^{[-i]}$ denote the solution of problem (10), when the data point (t_i,y_i) is omitted from $(\mathbf{t},\mathbf{y})$. This formulation is inconvenient since it involves solving m problems of the type (10), one for each deleted data point. We show that also in the nonlinear semiparametric setting it is possible, as in the classical smoothing splines case, to simplify this formulation such that only the solution of the total problem (10) is needed for evaluating the cross-validation function.

Firstly, we emphasize the influence that the i^{th} measured output y_i has on the optimal solution of (10), for a fixed value of λ and for fixed values of the other data points in $y_1,\ldots,y_m$. We denote the direct link between y_i and the corresponding component of the optimal model $\widehat{y}_\lambda$ by a function h such that $h(y_i) = (\widehat{y}_\lambda)_i = F\left(t_i,\widehat{\theta}_\lambda\right) + A_i\widehat{\mathbf{a}}_\lambda$.

Secondly, we note that the leaving-one-out lemma that was proved in the context of smoothing splines (Craven and Wahba, 1979) still holds trivially for our semiparametric problem. It ensures that if the measured y_i was by any chance equal to the function value predicted by the solution computed without the i^{th} measurement, *i.e.*, $y_i = (\widehat{y}_\lambda^{[-i]})_i := F\left(t_i,\widehat{\theta}_\lambda^{[-i]}\right) + A_i\widehat{\mathbf{a}}_\lambda^{[-i]}$, then the vectors $\widehat{\theta}_\lambda^{[-i]}$ and $\widehat{\mathbf{a}}_\lambda^{[-i]}$ would be also the optimal solution for the complete problem (10). We can write this observation in terms of the function h as $h((\widehat{y}_\lambda^{[-i]})_i) = (\widehat{y}_\lambda^{[-i]})_i$.

Using a similar trick as in (Wahba, 1990) for smoothing splines and in (Ke and Wang, 2004) for nonlinear nonparametric regression, we have:

$$y_i - F\left(t_i,\widehat{\theta}_\lambda^{[-i]}\right) - A_i\widehat{\mathbf{a}}_\lambda^{[-i]} = \frac{y_i - F\left(t_i,\widehat{\theta}_\lambda\right) - A_i\widehat{\mathbf{a}}_\lambda}{1-\Delta_i(\lambda)},$$

where

$$\Delta_i(\lambda) := \frac{\left(F\left(t_i,\widehat{\theta}_\lambda\right)+A_i\widehat{\mathbf{a}}_\lambda\right)-\left(F\left(t_i,\widehat{\theta}_\lambda^{[-i]}\right)+A_i\widehat{\mathbf{a}}_\lambda^{[-i]}\right)}{y_i-F\left(t_i,\widehat{\theta}_\lambda^{[-i]}\right)-A_i\widehat{\mathbf{a}}_\lambda^{[-i]}} = \frac{h(y_i)-h((\widehat{y}_\lambda^{[-i]})_i)}{y_i-(\widehat{y}_\lambda^{[-i]})_i},$$

which holds whenever $y_i \neq (\widehat{y}_\lambda^{[-i]})_i$. $\Delta_i(\lambda)$ is a divided difference for the function h, which can be approximated with the derivative of h. This leads to the definition of the following generalized influence (or smoother or hat) matrix $S(\lambda)$, which agrees with the definition given by Moody (1992) for the generalized influence matrix in a nonlinear context:

$$S(\lambda)_{ij} := \frac{\partial\left(F\left(t_i,\widehat{\theta}_\lambda\right)+A_i\widehat{\mathbf{a}}_\lambda\right)}{\partial y_j} = \frac{\partial(\widehat{\mathbf{y}}_\lambda)_i}{\partial y_j},$$

for which $S(\lambda)_{ii} \approx \Delta_i(\lambda)$, as a first order approximation.

The ordinary cross-validation can be approximated by

$$C(\lambda) \approx \frac{1}{m}\sum_{i=1}^{m}\left(y_i-F\left(t_i,\widehat{\theta}_\lambda\right)-A_i\widehat{\mathbf{a}}_\lambda\right)^2/(1-S(\lambda)_{ii})^2,$$

and the generalized cross-validation is its "rotation-invariant" version:

$$G(\lambda) = \frac{1}{m}\sum_{i=1}^{m}\left(y_i-F\left(t_i,\widehat{\theta}_\lambda\right)-A_i\widehat{\mathbf{a}}_\lambda\right)^2 / \left[\frac{1}{m}\mathrm{Tr}(I_m-S(\lambda))\right]^2 = \frac{m\|\widehat{\mathbf{y}}_\lambda-\mathbf{y}\|_2^2}{[\mathrm{Tr}(I_m-S(\lambda))]^2}. \tag{12}$$

In section 3.2.1 we have seen that for spline smoothing (which is a linear regularization problem), the influence matrix is $A(\lambda) = A(A^\top A+m\lambda B)^{-1}A^\top$.

For our semiparametric model, we derive the following result.

LEMMA 3. *The generalized influence matrix for the semiparametric model (9) is given by*

$$S(\lambda) \approx \begin{bmatrix} \nabla\mathbf{F}(\widehat{\theta}_\lambda) & A \end{bmatrix} \begin{bmatrix} \nabla\mathbf{F}(\widehat{\theta}_\lambda)^\top\nabla\mathbf{F}(\widehat{\theta}_\lambda) & \nabla\mathbf{F}(\widehat{\theta}_\lambda)^\top A \\ A^\top\nabla\mathbf{F}(\widehat{\theta}_\lambda) & A^\top A+m\lambda B \end{bmatrix}^{-1} \begin{bmatrix} \nabla\mathbf{F}(\widehat{\theta}_\lambda)^\top \\ A^\top \end{bmatrix}, \tag{13}$$

where the approximation sign indicates that some high order terms are ignored.

PROOF. We compute explicitly

$$S(\lambda) = \frac{\partial(\widehat{\mathbf{y}}_\lambda)}{\partial\mathbf{y}} = \frac{\partial\left(\mathbf{F}\left(\widehat{\theta}_\lambda\right)+A\widehat{\mathbf{a}}_\lambda\right)}{\partial\mathbf{y}} = \frac{\partial\left(\mathbf{F}\left(\widehat{\theta}_\lambda\right)+A\widehat{\mathbf{a}}_\lambda\right)}{\partial(\theta,\mathbf{a})}\frac{\partial(\widehat{\theta}_\lambda,\widehat{\mathbf{a}}_\lambda)}{\partial\mathbf{y}}. \tag{14}$$

Denote by $\mathbf{E}$ the regularized least squares objective function, but with the dependence on the measured $\mathbf{y}$ explicitly marked:

$$\mathbf{E}(\mathbf{y},\theta,\mathbf{a};\lambda) := \|\mathbf{y}-\mathbf{F}(\theta)-A\mathbf{a}\|_2^2+m\lambda\mathbf{a}^\top B\mathbf{a}.$$

Since $\widehat{\theta}_\lambda, \widehat{\mathbf{a}}_\lambda$ are optimal for given data $\mathbf{y}$, it means that $\partial\mathbf{E}(\mathbf{y},\widehat{\theta}_\lambda,\widehat{\mathbf{a}}_\lambda;\lambda)/\partial(\theta,\mathbf{a}) = 0$. On the other hand, if $\mathbf{y}$ is 'perturbed' and the data becomes $\mathbf{y}+d\mathbf{y}$, the optimum also changes. We denote the new optimal solution by $(\widehat{\theta}_\lambda+d\theta,\widehat{\mathbf{a}}_\lambda+d\mathbf{a})$; it satisfies

$$\partial\mathbf{E}(\mathbf{y}+d\mathbf{y},\widehat{\theta}_\lambda+d\theta,\widehat{\mathbf{a}}_\lambda+d\mathbf{a};\lambda)/\partial(\theta,\mathbf{a}) = 0.$$

Using a first order Taylor approximation, it implies that

$$\frac{\partial^2 \mathbf{E}(\mathbf{y},\widehat{\theta}_\lambda,\widehat{\mathbf{a}}_\lambda;\lambda)}{\partial(\theta,\mathbf{a})\partial\mathbf{y}^\top}d\mathbf{y}+\frac{\partial^2 \mathbf{E}(\mathbf{y},\widehat{\theta}_\lambda,\widehat{\mathbf{a}}_\lambda;\lambda)}{\partial(\theta,\mathbf{a})\partial(\theta,\mathbf{a})^\top}d(\theta,\mathbf{a})\approx 0. \tag{15}$$

It is easy to see from the formula of $\mathbf{E}$ that

$$\frac{\partial^2 \mathbf{E}(\mathbf{y},\widehat{\theta}_\lambda,\widehat{\mathbf{a}}_\lambda;\lambda)}{\partial(\theta,\mathbf{a})\partial\mathbf{y}^\top}=-2\frac{\partial\left(\mathbf{F}\left(\widehat{\theta}_\lambda\right)+A\widehat{\mathbf{a}}_\lambda\right)^\top}{\partial(\theta,\mathbf{a})}=-2\begin{bmatrix}\nabla\mathbf{F}(\widehat{\theta}_\lambda) & A\end{bmatrix}^\top$$

and the block in the Hessian of $\mathbf{E}$ corresponding to $(\theta,\mathbf{a})$ is

$$2\widehat{H}(\lambda)=\begin{bmatrix}\frac{\partial^2}{\partial\theta\partial\theta^\top}E(\mathbf{y},\widehat{\theta}_\lambda,\widehat{\mathbf{a}}_\lambda;\lambda) & \frac{\partial^2}{\partial\theta\partial\mathbf{a}^\top}E(\mathbf{y},\widehat{\theta}_\lambda,\widehat{\mathbf{a}}_\lambda;\lambda)\\ \frac{\partial^2}{\partial\mathbf{a}\partial\theta^\top}E(\mathbf{y},\widehat{\theta}_\lambda,\widehat{\mathbf{a}}_\lambda;\lambda) & \frac{\partial^2}{\partial\mathbf{a}\partial\mathbf{a}^\top}E(\mathbf{y},\widehat{\theta}_\lambda,\widehat{\mathbf{a}}_\lambda;\lambda)\end{bmatrix}\approx 2\begin{bmatrix}\nabla\mathbf{F}(\widehat{\theta}_\lambda)^\top\nabla\mathbf{F}(\widehat{\theta}_\lambda) & \nabla\mathbf{F}(\widehat{\theta}_\lambda)^\top A\\ A^\top\nabla\mathbf{F}(\widehat{\theta}_\lambda) & A^\top A+m\lambda B\end{bmatrix},$$

where the approximation appears only in the (1,1) block of the matrix $\widehat{H}(\lambda)$ and is related to ignoring a term containing the second order differential of $\mathbf{F}$ (which is a tensor) and the residual vector $\mathbf{y}-\mathbf{F}(\widehat{\theta}_\lambda)-A\widehat{\mathbf{a}}_\lambda$. It is customary to ignore such a difficult-to-compute term, and thus use a *pseudo-Hessian*, since this approximation is made at converged solutions, where the residual is usually small.

Since $d\mathbf{y}$ and $d(\theta,\mathbf{a})$ represent small perturbations on $\mathbf{y}$ and $(\widehat{\theta}_\lambda,\widehat{\mathbf{a}}_\lambda)$, we can use the relation (15) and set the matrix $\widehat{H}(\lambda)^{-1}\begin{bmatrix}\nabla\mathbf{F}(\widehat{\theta}_\lambda) & A\end{bmatrix}^\top$ as approximation for the differential $\frac{\partial(\widehat{\theta}_\lambda,\widehat{\mathbf{a}}_\lambda)}{\partial\mathbf{y}}$ in (14). The formula (13) of $S(\lambda)$ now follows. □

The evaluation and minimization of the GCV function (12) is more difficult than in classical linear spline fitting, because there is no closed form expression for the optimal regularized estimate $\widehat{\theta}_\lambda$, which enters implicitly into the computation of $G(\lambda)$ through the formula of $\widehat{\mathbf{y}}_\lambda$ and of $S(\lambda)$.

In Algorithm 1, a scalar minimization algorithm (*e.g.*, a golden section method) is used for minimizing the GCV function, while at each iteration a nonlinear least squares minimization is carried out for estimating $\widehat{\theta}_\lambda$, at the current value of λ. (Remember from subsection 4.3 that once we have an estimated θ we have a corresponding $\mathbf{a}$ in closed form.) Often in applications, the GCV function is unimodal thus it is natural to search for a globally optimal λ via a scalar minimization method. However, there is no proof for such an observation. Any scalar optimization method could be used (Algorithm 1 exemplifies a variant with golden section search method), or improved combinations, *e.g.*, steps of a golden section search can be alternated with a parabolic search. All methods require one GCV function evaluation per iteration, and correspondingly, one nonlinear least squares solution. Both the (nonlinear least squares) inner minimization and the outer scalar minimization over λ in Algorithm 1 are globally convergent; but different initial values of the parameters might give different converged local solutions, since we deal with nonlinear (nonconvex) optimization.

At every iteration a nonlinear optimization problems is solved, and this operation should be made as efficient as possible. For the optimization on θ (step 4: in 1), one should use efficient function and Jacobian computations as described in Appendix A. For the evaluation of the GCV criterion (step 5:), efficient computations are described in Appendix B.

REMARK 4. *We have also experimented with the procedure shortly outlined in §9.1 of (Wahba, 1990), which suggested to employ an iterative nonlinear minimization method where at each iteration a new λ, optimal for the GCV of the linearized nonlinear function at the current iterates, is used.*

Algorithm 1 Algorithm for regularized nonlinear least squares with adaptive choice of the regularization parameter; golden section variant

Input: Measurements vector $\mathbf{y}$, deterministic function $\mathbf{F}$, matrix A for spline basis and matrix B for smoothing constraint. Initial approximation $\theta_0 \in \Theta$ and bounds $0 \leq \lambda_{\min} < \lambda_{\max}$. Convergence tolerance *tol*. Scalar $\gamma \in (0, 0.5)$, *e.g.*, $\gamma \cong 0.3819$ defined by the golden section

Initialize: $\theta \leftarrow \theta_0$, $\lambda_{\text{left}} \leftarrow \lambda_{\min}$, $\lambda_{\text{right}} \leftarrow \lambda_{\max}$
repeat
  $\lambda_1 \leftarrow \lambda_{\text{left}}(1-\gamma) + \lambda_{\text{right}}\gamma$, $\lambda_2 \leftarrow \lambda_{\text{left}}\gamma + \lambda_{\text{right}}(1-\gamma)$
  compute / update θ_{λ_1} and θ_{λ_2} as the optimal solutions of the NLS criterion (11), for the fixed values λ_1 and λ_2, respectively
  evaluate / update the GCV function values at λ_{left}, λ_{right}, λ_1 and λ_2, using formula (12)
  if $G(\lambda_1) > G(\lambda_2)$ **then**
    $\lambda_{\text{left}} \leftarrow \lambda_1$
  else
    $\lambda_{\text{right}} \leftarrow \lambda_2$
  end if
until $\lambda_2 - \lambda_1 < tol$
$\widehat{\lambda} \leftarrow \lambda_1$, $\widehat{\theta} \leftarrow \theta_{\widehat{\lambda}}$, $\widehat{\mathbf{a}} \leftarrow \mathbf{a}(\widehat{\theta}, \widehat{\lambda})$.

Output: Solution $(\widehat{\theta}, \widehat{\mathbf{a}})$ and smoothing parameter $\widehat{\lambda}$.

However, this method didn't exhibit a good convergence behavior compared to Algorithm 1. This suggests that applying the linearization in early iterations, where we might be far from the optimal solution, is not appropriate for highly nonlinear functions.

5. Asymptotic properties of semiparametric regression

5.1. Asymptotic normality

Although we don't impose additional regularization constraints on θ, the smoothness constraint that is imposed on the nonparametric part g also influences the computation of the parameter of interest, θ. We describe precisely what this influence is on the bias and variance of the estimator $\widehat{\theta}_\lambda$ in the following theorem.

THEOREM 5. *Let (t_i, y_i), $i = 1, 2, \ldots, m$, denote m observations from a semiparametric model with partially known functional relationship,*

$$y_i = F(t_i, \theta^\star) + g^\star(t_i) + \varepsilon_i, \qquad (i = 1, 2, \ldots, m),$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $t_1, \ldots, t_m$ are regression abscissas from a real interval $\mathcal{I}$, $\theta^\star$ denotes the true value of the unknown parameter $\theta \in \Theta \subset \mathbb{R}^p$, and $g^\star$ is the true 'baseline' function. Assume that:

(A1). $g^\star \in \mathcal{H}$, for a certain given finite dimensional Hilbert space $\mathcal{H}$;

(A2). $\phi_1, \ldots, \phi_n \in \mathcal{H}$ is a family of generators in $\mathcal{H}$, A denotes the $m \times n$ matrix that has as elements $\phi_k(t_i)$, for k from 1 to n and i from 1 to m, and B denotes the $n \times n$ matrix having the elements $\langle \mathscr{P}\phi_k, \mathscr{P}\phi_l \rangle_{\mathcal{H}'}$, for k, l from 1 to n, where $\mathscr{P}$ and $\mathcal{H}'$ have been introduced in subsection 4.1.

Then, for any $\lambda \geq 0$, the estimates of $\theta^\star$ and $\mathbf{a}^\star$ (where $\mathbf{a}^\star$ satisfies $A\mathbf{a}^\star = g^\star$) defined as the global minimizer $(\widehat{\theta}_\lambda, \widehat{\mathbf{a}}_\lambda)$ of (10) over $\theta \in \Theta$, $\mathbf{a} \in \mathbb{R}^n$, satisfy, when $m \to \infty$:

$$\begin{bmatrix} \widehat{\theta}_\lambda \\ \widehat{\mathbf{a}}_\lambda \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \theta^\star \\ \mathbf{a}^\star \end{bmatrix} - H(\lambda)^{-1} \begin{bmatrix} 0 \\ m\lambda B\mathbf{a}^\star \end{bmatrix}, \sigma^2 H(\lambda)^{-1} \left(H(\lambda) - \begin{bmatrix} 0 & 0 \\ 0 & m\lambda B \end{bmatrix} \right) H(\lambda)^{-1} \right), \tag{16}$$

where $H(\lambda) := \begin{bmatrix} \nabla\mathbf{F}(\theta^\star)^\top \nabla\mathbf{F}(\theta^\star) & \nabla\mathbf{F}(\theta^\star)^\top A \\ A^\top \nabla\mathbf{F}(\theta^\star) & A^\top A + m\lambda B \end{bmatrix}$.

PROOF. When $m \to \infty$, we can use first order approximation of the nonlinear function $\mathbf{F}$:

$$\mathbf{F}(\theta) \approx \mathbf{F}(\theta^\star) + \nabla\mathbf{F}(\theta^\star)(\theta - \theta^\star),$$

since we expect that at convergence the bias between $\widehat{\theta}$ and $\theta^\star$ is not so big.

Replacing $\mathbf{y} = \mathbf{F}(\theta^\star) + A\mathbf{a}^\star + \varepsilon$ into the minimization (10) leads to:

$$\min_{\theta\in\Theta, \mathbf{a}\in\mathbb{R}^n} \frac{1}{m} \|J\theta^\star + A\mathbf{a}^\star + \varepsilon - J\theta - A\mathbf{a}\|^2 + \lambda \mathbf{a}^\top B\mathbf{a}, \tag{17}$$

where we used the short-hand notation $J := \nabla\mathbf{F}(\theta^\star)$. From the first order optimality conditions, we get

$$\begin{bmatrix} \widehat{\theta}_\lambda \\ \widehat{\mathbf{a}}_\lambda \end{bmatrix} = \begin{bmatrix} J & A \\ 0 & \sqrt{m\lambda}D \end{bmatrix}^\dagger \begin{bmatrix} J\theta^\star + A\mathbf{a}^\star + \varepsilon \\ 0 \end{bmatrix} = \begin{bmatrix} \theta^\star \\ \mathbf{a}^\star \end{bmatrix} - H(\lambda)^{-1} \begin{bmatrix} 0 \\ m\lambda B\mathbf{a}^\star \end{bmatrix} + H(\lambda)^{-1} \begin{bmatrix} J^\top \\ A^\top \end{bmatrix} \varepsilon,$$

where D is a Cholesky factor of B, *i.e.*, $D^\top D = B$.

The assumptions on the noise term ε imply the normality of the estimator, with the bias from the true values and the covariance matrix given in the conclusion of the theorem. □

REMARK 6. *The assumptions (A1-2) can be relaxed, at the expense of a more technical proof. More specifically, the true nonlinearity $g^\star$ might violate (A1) by not being in $\mathscr{H}$, but then we should consider an asymptotic case with respect to the number of spline basis functions n such that when n goes to infinity $g^\star$ should become arbitrarily close to a member of $\mathscr{H}$. (A2) can be replaced by the condition that the null space of the operator $\mathscr{P}$ in $\mathscr{H}$ is finite dimensional, in the case when $\mathscr{H}$ itself is infinite dimensional.*

Theorem 5 gives, for the case $\lambda = 0$, similar conclusion as the results on nuisance parameter analysis of Spall and Garner (1990): the estimator is unbiased and its covariance depends on the nuisance term (in our case, the baseline parameter $\mathbf{a}^\star$). The difference in the $\lambda \neq 0$ case is, however, that we must also incorporate the contribution of the regularization term; thus, there will be a bias, but the covariance can be of smaller magnitude. An adequate choice of λ should provide an optimal bias-variance trade-off. The GCV criterion provides such a trade-off. We observe next that the bias and covariance formulas are very much linked to the formula of the generalized smoother matrix $S(\lambda)$ in (14), where instead of the estimated $(\widehat{\theta}_\lambda, \widehat{\mathbf{a}}_\lambda)$ we plug in the values of the true parameters $(\theta^\star, \mathbf{a}^\star)$.

Denote by T the gradient $\begin{bmatrix} J & A \end{bmatrix}$, where as before $J = \nabla\mathbf{F}(\theta^\star)$. Then the smoother matrix at the true values becomes $S(\lambda) = TH(\lambda)^{-1}T^\top$.

Whenever the linear approximation $\mathbf{F}(\theta) \approx \mathbf{F}(\theta^\star) + J(\theta - \theta^\star)$ holds, we have as in the proof of Theorem 5 that

$$\widehat{\mathbf{y}}_\lambda := \mathbf{F}(\widehat{\theta}_\lambda) + A\widehat{\mathbf{a}}_\lambda \approx T \begin{bmatrix} J & A \\ 0 & \sqrt{m\lambda}D \end{bmatrix}^\dagger \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} = TH(\lambda)^{-1}T^\top \mathbf{y} = S(\lambda)\mathbf{y}.$$

Thus, we obtain the approximation of a classical relation, $\mathbf{y} - \widehat{\mathbf{y}}_\lambda \approx (I_m - S(\lambda))\mathbf{y}$. This means that the GCV criterion for the regularized nonlinear problem can be approximated by the GCV criterion for the regularized linear problem (17), where the linear approximation is plugged in. In the asymptotic linear case (see Wahba, 1990, §4.4), the optimal λ given by GCV approaches the value of λ minimizing the *predictive mean square error* criterion $\|J\theta^\star + A\mathbf{a}^\star - J\widehat{\theta}_\lambda - A\widehat{\mathbf{a}}_\lambda\|_2^2$. This implies bounds on the forward error $\|\theta^\star - \widehat{\theta}_\lambda\|_2^2 + \|\mathbf{a}^\star - \widehat{\mathbf{a}}_\lambda\|_2^2$, as well.

Moreover, note that the bias in (16) can be rewritten as

$$-H(\lambda)^{-1}\begin{bmatrix} 0 \\ m\lambda B\mathbf{a}^\star \end{bmatrix} = -H(\lambda)^{-1}(H(\lambda) - T^\top T)\begin{bmatrix} \theta^\star \\ \mathbf{a}^\star \end{bmatrix} = -(I_{p+n} - H(\lambda)^{-1}T^\top T)\begin{bmatrix} \theta^\star \\ \mathbf{a}^\star \end{bmatrix}$$

and the covariance matrix in (16) satisfies $\mathscr{C} = \sigma^2 H(\lambda)^{-1}T^\top T H(\lambda)^{-1}$, and $T^\top \mathscr{C} T = \sigma^2 S(\lambda)^2$.

The magnitude of the matrix $I_{p+n} - H(\lambda)^{-1}T^\top T$ influences the bias term. One measure of this magnitude could be the trace, and $\mathrm{Tr}(I_{p+n} - H(\lambda)^{-1}T^\top T) = p + n - \mathrm{Tr}S(\lambda)$. In nonlinear or regularized models, the quantity $p_{\mathrm{eff}} := \mathrm{Tr}S(\lambda)$, the trace of the influence matrix, has the meaning of the *effective number of parameters* (Moody, 1992); in other words, the number p_{eff} replaces $p+n$, the number of components in the regression variable θ and $\mathbf{a}$. This quantity comes into play when we estimate the noise variance in the semiparametric model (generalizing (Wahba, 1990, Equation (5.1.3))):

$$\widehat{\sigma}^2 := \frac{\|\mathbf{y} - \mathbf{F}(\widehat{\theta}_\lambda) - A\widehat{\mathbf{a}}_\lambda\|_2^2}{m - p_{\mathrm{eff}}} \approx \frac{\|(I_m - S(\lambda))\mathbf{y}\|_2^2}{\mathrm{Tr}(I_m - S(\lambda))}. \tag{18}$$

5.2. Asymptotic confidence intervals

In this section, we derive practical statistical information, as corollaries of Theorem 5. In essence, one is interested in finding confidence intervals for the parameter θ. As usual when dealing with Fisher information matrices, one cannot compute the exact covariance matrix (since it depends on the true unknown parameter); thus, it is common to replace the true parameters (in our case, $\theta^\star$ and $\mathbf{a}^\star$) by their estimated values (*i.e.*, $\widehat{\theta}_\lambda$ and $\widehat{\mathbf{a}}_\lambda$). Due to the same reason, we can only perform an approximate bias correction, which follows by replacing $H(\lambda)$ with $\widehat{H}(\lambda)$ in the following corollary.

COROLLARY 7. *With the notation of Theorem 5, an unbiased estimate of* $\left[\theta^{\star\top},\ \mathbf{a}^{\star\top}\right]^\top$ *is given by*

$$\begin{bmatrix} \widehat{\widehat{\theta}}_\lambda \\ \widehat{\widehat{\mathbf{a}}}_\lambda \end{bmatrix} := \begin{bmatrix} \widehat{\theta}_\lambda \\ \widehat{\mathbf{a}}_\lambda \end{bmatrix} + \left(H(\lambda) - \begin{bmatrix} 0 & 0 \\ 0 & m\lambda B \end{bmatrix}\right)^{-1}\begin{bmatrix} 0 \\ m\lambda B\widehat{\mathbf{a}}_\lambda \end{bmatrix}.$$

PROOF. From Theorem 5, we see that $\left[\widehat{\theta}_\lambda^\top,\ \widehat{\mathbf{a}}_\lambda^\top\right]^\top$ is centered around

$$\begin{bmatrix} \theta^\star \\ \mathbf{a}^\star \end{bmatrix} - H(\lambda)^{-1}\begin{bmatrix} 0 \\ m\lambda B\mathbf{a}^\star \end{bmatrix} = \left(I - H(\lambda)^{-1}\begin{bmatrix} 0 & 0 \\ 0 & m\lambda B \end{bmatrix}\right)\begin{bmatrix} \theta^\star \\ \mathbf{a}^\star \end{bmatrix}.$$

Thus, using the matrix inversion lemma,

$$\left(I - H(\lambda)^{-1}\begin{bmatrix} 0 & 0 \\ 0 & m\lambda B \end{bmatrix}\right)^{-1}\begin{bmatrix} \widehat{\theta}_\lambda \\ \widehat{\mathbf{a}}_\lambda \end{bmatrix} = \left(I + \left(H(\lambda) - \begin{bmatrix} 0 & 0 \\ 0 & m\lambda B \end{bmatrix}\right)^{-1}\begin{bmatrix} 0 & 0 \\ 0 & m\lambda B \end{bmatrix}\right)\begin{bmatrix} \widehat{\theta}_\lambda \\ \widehat{\mathbf{a}}_\lambda \end{bmatrix}$$

is an unbiased estimate of $\left[\theta^{\star\top},\ \mathbf{a}^{\star\top}\right]^\top$. □

At this point, we are ready to give formulas for $100(1-\alpha)\%$ confidence intervals¶ for each of the parameters θ_i, $i = 1, \ldots, p$.

COROLLARY 8. *An approximate* $100(1-\alpha)\%$ *confidence interval for the parameter* θ_i *is given by*

$$(\widehat{\widehat{\theta}}_\lambda)_i \pm t_{m-p_{eff}}^{\alpha/2} \cdot (\widehat{\sigma}^2 \widehat{\mathscr{C}}_{ii}^{\theta})^{1/2},$$

where t_k^α *denotes the* α *quantile of the Student t distribution with k degrees of freedom,* $\widehat{\sigma}^2$ *is given in (18), and* $\widehat{\mathscr{C}}_{ii}^{\theta}$ *is the* i^{th} *diagonal element of the covariance matrix of* $\widehat{\theta}_\lambda$*,*

$$\widehat{\mathscr{C}}^{\theta} := \widehat{\mathscr{G}}^\top \widehat{\mathscr{G}}, \qquad \textit{with} \quad \widehat{\mathscr{G}} := (I_m - A(\lambda))\nabla\mathbf{F}(\widehat{\theta}_\lambda)\left(\nabla\mathbf{F}(\widehat{\theta}_\lambda)^\top (I_m - A(\lambda))\nabla\mathbf{F}(\widehat{\theta}_\lambda)\right)^{-1}$$
$$\textit{and} \quad A(\lambda) = A(A^\top A + m\lambda B)^{-1}A^\top.$$

PROOF. We further simplify the covariance formula in Theorem 5,

$$\mathscr{C} = H(\lambda)^{-1}\left(H(\lambda) - \begin{bmatrix} 0 & 0 \\ 0 & m\lambda B \end{bmatrix}\right)H(\lambda)^{-1} = H(\lambda)^{-1}\begin{bmatrix} \nabla\mathbf{F}^\top \\ A^\top \end{bmatrix}\begin{bmatrix} \nabla\mathbf{F} & A \end{bmatrix}H(\lambda)^{-1},$$

focusing only on $\mathscr{C}^\theta$, the upper-left $p \times p$ block of $\mathscr{C}$, which gives the covariance information for the parameter of interest, θ. We use a block matrix inversion formula involving the Schur complement of the (2,2) block, and partition

$$H(\lambda)^{-1} = \begin{bmatrix} \nabla\mathbf{F}^\top\nabla\mathbf{F} & \nabla\mathbf{F}^\top A \\ A^\top\nabla\mathbf{F} & A^\top A + m\lambda B \end{bmatrix}^{-1} \qquad \text{as} \qquad \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^\top & \star \end{bmatrix},$$

where $H_{11} = \left(\nabla\mathbf{F}^\top(I_m - A(\lambda))\nabla\mathbf{F}\right)^{-1}$, and $H_{12} = -H_{11}\nabla\mathbf{F}^\top A(A^\top A + m\lambda B)^{-1}$. We see that

$$\mathscr{C}^\theta = \mathscr{G}^\top\mathscr{G}, \quad \text{with} \quad \mathscr{G} = \nabla\mathbf{F}H_{11} + AH_{12}^\top = (I_m - A(\lambda))\nabla\mathbf{F}H_{11} = (I_m - A(\lambda))\nabla\mathbf{F}\left(\nabla\mathbf{F}^\top(I_m - A(\lambda))\nabla\mathbf{F}\right)^{-1}.$$

(For simplicity, we left out the argument of $\nabla\mathbf{F}$, which in exact formulas should be $\theta^\star$, but can be replaced by $\widehat{\theta}_\lambda$, yielding the computable "hat" formulas.)

Since the estimator $\widehat{\widehat{\theta}}_\lambda$ is unbiased and normally distributed, we plug-in the approximate estimated variance (18) and we get that, approximately and for large m,

$$\frac{(\widehat{\widehat{\theta}}_\lambda)_i - \theta_i^\star}{\sqrt{\widehat{\sigma}^2\widehat{\mathscr{C}}_{ii}^{\theta}}} \sim t_{m-p_{\text{eff}}},$$

which gives the announced confidence interval. □

Taking a careful look at the specialized confidence intervals that we obtained, and observing in particular the formula of $\widehat{\mathscr{C}}^\theta$, the covariance matrix of the estimated parameter, we note the following differences from the classical confidence intervals in nonlinear regression (Seber and Wild, 1989):

- we need to use a bias-corrected estimate;

¶In a straightforward manner, we can design *confidence regions* (in the $\mathbb{R}^p$ space) instead of individual confidence intervals for each parameter; see (Seber and Wild, 1989, Chapter 5) for more details.

- the number of regression parameters p is replaced by the effective number of parameters p_{eff}, both in the variance formula (18) and in the degrees of freedom of the Student t distribution;

- if λ goes to infinity, then the penalty on the baseline is very strong. Under some simple conditions on the penalty matrix B, it means that the baseline goes to zero and, thus, only the parametric part of the model is used in the fitting. In this case, the 'linear smoother matrix' $A(\lambda)$ becomes zero as well and we get as covariance matrix $\widehat{\mathscr{C}^\theta}$ the classical inverse of the Fisher information matrix as covariance matrix for the estimate of θ; indeed:

$$\text{if} \quad \widehat{\mathscr{G}} = I_m \nabla\mathbf{F}(\nabla\mathbf{F}^\top I_m \nabla\mathbf{F})^{-1} \quad \text{then} \quad \widehat{\mathscr{C}^\theta} = \widehat{\mathscr{G}}^\top \widehat{\mathscr{G}} = (\nabla\mathbf{F}^\top \nabla\mathbf{F})^{-1}.$$

6. Discussion on identifiability, redundancy and uniqueness

The *identifiability* concept involves proving or disproving that the parameters of the exact model can be exactly recovered from noiseless measurements. Some parameters might be *redundant*; this happens when the same (noisy) data can be explained by at least two different parameter values. Even for an identifiable model, ill-conditioning and noise can make some parameters non-uniquely determined. Other *uniqueness* issues are related to the optimization objective function (*e.g.*, nonlinear least squares) that might not be unimodal, thus several locally optimal solutions could be computed, depending on the starting point of the iterative optimization algorithm.

In linear regression, the uniqueness of the least squares solution, the identifiability and the parameter redundancy questions are all related to the rank deficiency (or nearly rank deficiency) of the coefficient matrix X of the linear model $\mathbf{y} = X\theta + \varepsilon$.

For nonlinear regression, one often translates the results of linear regression by using an estimate of the matrix $\mathscr{A} = \left[\frac{\partial F}{\partial \theta}(\mathbf{t}, \theta^\star)\right]$ (see Theorem 1) instead of the linear coefficient X. However, the asymptotic linearization is not always appropriate, thus the problems become much more involved. A whole detailed discussion on these issues for nonlinear regression is found Chapter 3 of (Seber and Wild, 1989).

In this section, we focus on the combination of a nonlinear parametric part (which we assume identifiable and not prone to non-uniqueness and parameter redundancy problems) and the nonparametric part (the part that is actually linearly *parameterized* by splines).

In order to avoid identifiability problems that come from partly modeling the nonlinear F with splines, one can first think of a strategy of designing an appropriate spline space $\mathscr{H}$ such that it doesn't interfere with the nonlinear function of the parametric part. If such an $\mathscr{H}$ is possible to construct, then a pure semiparametric regression setting without any penalty term can be used. A sufficient identifiability condition would simply require that the given nonlinear function $F(\cdot, \theta)$ is **not** in the space $\mathscr{H}$ for any θ. This restriction forbids the possibility that the output data $\mathbf{y}^\star$ can be interpolated in $\mathscr{H}$. Thus, the nonlinear F and the nonparametric part would be completely disentangled.

In practice, when the space $\mathscr{H}$ is generated by splines, it can be quite rich. Even if the nonlinear F is not a member of $\mathscr{H}$, it might happen that (parts of) F can be very well approximated in $\mathscr{H}$. In this situation, we would like to ensure that the problem is still identifiable, at least for an appropriate penalty term and for an appropriate value of the regularization parameter λ.

The semiparametric regression problem is not identifiable if the parametric part doesn't have at least one *distinctive feature* that can separate it from what could be considered as background. Since we are focusing on the case when we require that the nonparametric part is fitted with a certain degree of smoothness, it means that our condition for identifiability requires that the nonlinear F

should contain a certain non-smooth feature. This will definitely exclude the cases when F is a linear function or a low order polynomial.

The NMR spectroscopy data quantification problem is a good example for applying our semi-parametric modeling method because the frequency domain metabolite profiles contain those specific Lorentzian peaks, which cannot be fitted by a smooth baseline. Moreover, in this application, we have in general good convergence of the numerical optimization, because the starting values for the nonlinear parameters ζ_k (see (1)) can be set to one, since the frequency shifts f_k and the damping corrections d_k should be close to zero.

Indicative information about the quality of the estimates can be obtained by looking at the estimated confidence intervals.

7. Numerical examples

In this section, we illustrate with a few examples the use of Algorithm 1, and of the statistical information extracted with the procedures of Sections 5.1 and 5.2.

We choose a simulation scenario that approaches in a simplified way the design of the NMR spectroscopy data quantification problem. Extensive numerical simulations for the NMR spectroscopy data quantification problem, with emphasis on clinical performance, are currently being described in a report on the AQSES software package for automated quantification of short echo-time NMR signals (available from http://www.esat.kuleuven.be/sista/members/biomed/new/).

7.1. Description of a simulation example and results

Inspired by the shape of the frequency domain transformation of metabolite signals in the NMR experiment (*i.e.*, Lorentzian lineshapes), we consider as a simplification the rational function

$$F(t,(\theta_1^{(1)},\ldots,\theta_1^{(K)},\theta_2^{(1)},\ldots,\theta_2^{(K)})) := \sum_{k=1}^{K} \frac{\theta_1^{(k)}}{\left(t-\theta_2^{(k)}\right)^2+0.1},$$

where t represents a scalar abscissa in the interval $\mathscr{I} = [-10,10]$. We use as true values for the parameter $\theta^\star$ the following data: $\theta_1^{(1,\ldots,5)} = [5,12,8,20,15]$, $\theta_2^{(1,\ldots,5)} = [-7,-4,-3,1,5]$. Note that $\theta_1^{(1,\ldots,5)}$, which appear linearly in the model function F, determine the amplitude of the Lorentzian shapes, while $\theta_2^{(1,\ldots,5)}$, appearing in the denominator, determine the positions for each peak. (See middle plot in Figure 7.)

We consider the following simple baseline function:

$$g(t) := \sin(t/2).$$

Then we simulate $m = 200$ measured outputs $y_1,\ldots,y_m$ at equidistant abscissas $t_1,\ldots,t_m$ in $\mathscr{I}$, by adding up the simulation function values (at fixed parameter $\theta^\star$), the values of the baseline g, and Gaussian noise with given variance σ^2, which represents a noise-to-signal ratio of approximately 30%, taking into account the magnitude of the considered simulated function.

For computing the nonparametric part, we use "penalized splines" (Eilers and Marx, 1996). Thus, the matrix A is formed from a B-spline basis (de Boor, 1978) of degree 3; we set the size of A to $m \times n = 200 \times 40$. Moreover, instead of the regularization operator $\mathscr{P}$, we compute the matrix B as $B = D^\top D$, where D is the second order derivative matrix, *i.e.*, the $(n-2)\times n$ tridiagonal Toeplitz matrix, with 2 on the main diagonal and -1 on the first super- and subdiagonal.

Table 1. Monte Carlo simulation results for the example problem. The noise-to-signal ratio (N/S) is varied from 10% to 50%, while the problem dimensions $m = 200$, $n = 40$ and $p = 10$ remain constant. Averaged values of the true and estimated noise variances, and averaged relative errors in the regression parameter θ (*i.e.*, $\|\widehat{\theta}_\lambda - \theta^\star\|_2/\|\theta^\star\|_2$), in the function fit (*i.e.*, $\|\mathbf{F}(\widehat{\theta}_\lambda) - \mathbf{F}(\theta^\star)\|_2/\|\mathbf{F}(\theta^\star)\|_2$), and, respectively, in the fit of the baseline (*i.e.*, $\|A\widehat{\mathbf{a}}_\lambda - \mathbf{g}^\star\|_2/\|\mathbf{g}^\star\|_2$) are presented.

N/S	*true variance*	*estimated variance*	*error in* θ	*error in* $\mathbf{F}$	*error in baseline*
10 %	32.61	33.18	0.024	0.028	0.052
20 %	131.4	133.1	0.044	0.052	0.095
30 %	292.2	296.1	0.060	0.074	0.123
40 %	530.5	526.5	0.065	0.086	0.151
50 %	826.5	823.2	0.071	0.098	0.176

Figure 7 shows the signal $\mathbf{y}$ obtained as described above, together with the fit returned by our Matlab implementation of Algorithm 1. As initial values for the parameters we choose random values with the constraint that $\theta_1^{(1,\ldots,5)}$ are positive and scaled to reflect the magnitude of the simulated signal, and $\theta_2^{(1,\ldots,5)}$, appearing in the denominators, are taken from non-overlapping intervals in $[-10, 10]$ that contain the true values.

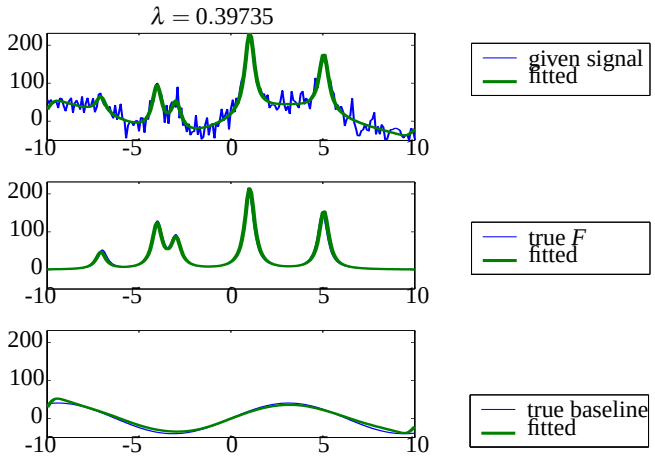


Fig. 7. Simulated signal $\mathbf{y}$: the bigger peaks come from the function of interest F, the baseline gives a smooth trend, and the Gaussian noise has a noise-to-signal ratio of 30%. With a GCV choice for λ and with reasonably good initial values for the nonlinear parameters of F, we obtain excellent fit of the model and of the baseline.

We performed a Monte Carlo simulation study to assess statistical information on the converged estimates given by Algorithm 1. Table 1 presents averaged results after 100 simulations for several noise levels. The averaged relative errors for the regression parameter θ, as well as for the reconstructed nonlinear function $\mathbf{F}(\theta)$ and baseline $A\mathbf{a}$ are reported in the last three columns of Table 1. These small errors confirm that the regularized nonlinear least squares algorithm, with the GCV choice for the smoothing parameter λ, performs well. Moreover, in the left part of Table 1, next to the noise-to-signal ratio, we have the corresponding noise variance used in the simulations, as well as the estimated noise variance using formula (18). The values agree well in all experiments, thus confirming the validity of formula (18) in practice.

An illustration of the estimation results for each individual parameter in θ is shown in Figure 8. The mean squared errors obtained in the Monte Carlo study are plotted together with the Cramer-

Table 2. Percentages of confidence intervals that contain the true parameter value.

Method	$\theta_1^{(1)}$	$\theta_1^{(2)}$	$\theta_1^{(3)}$	$\theta_1^{(4)}$	$\theta_1^{(5)}$	$\theta_2^{(1)}$	$\theta_2^{(2)}$	$\theta_2^{(3)}$	$\theta_2^{(4)}$	$\theta_2^{(5)}$
Classical	79%	77%	75%	76%	80%	91%	96%	91%	93%	93%
Specialized	97%	94%	93%	99%	97%	94%	99%	96%	97%	96%

Rao lower bounds computed from the Fisher information matrix corresponding to the true parameter values of the nonlinear function F. Note that these Cramer-Rao bounds ignore the incorporation of a possible baseline in the model. As a result, we see that the mean squared error for the estimate of the linear parameters $\theta_1^{(1,\ldots,5)}$ are further away from the Cramer-Rao bounds than it is the case for the nonlinear parameters $\theta_2^{(1,\ldots,5)}$. This means that the latter group of parameters is less affected by the nuisance part of the semiparametric model.

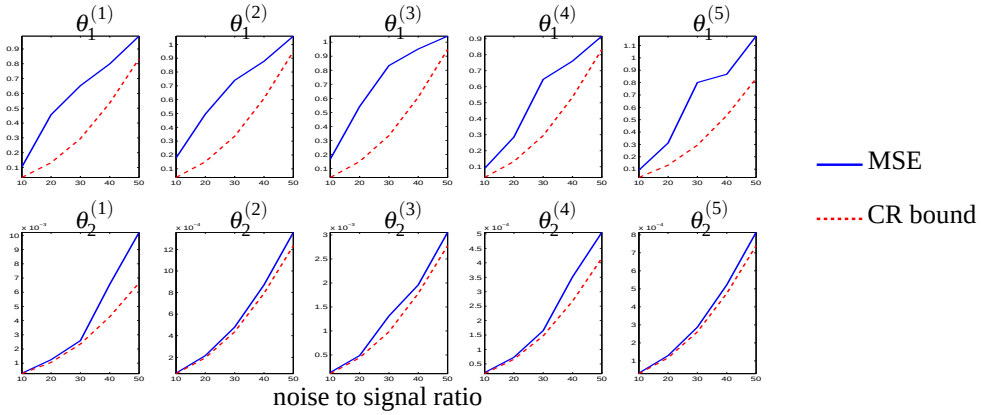


Fig. 8. Each plot corresponds to an individual variable in θ; the horizontal axis corresponds to several noise levels, and the vertical axis corresponds to squared errors in each parameter. In every plot, the full (blue) line shows the mean squared errors of the estimated parameter to its true value, and the interrupted (red) line is the Cramer-Rao lower bound.

7.2. Comparison between classical confidence intervals and specialized confidence intervals

At the end of Section 5.2 we observed that the confidence intervals taking into account the presence of the baseline term (see Corollary 8) are a generalization of the classical confidence intervals from nonlinear regression. Here we compare experimentally the new specialized confidence intervals with the classical ones.

With the same example setting as above, we use a noise-to-signal ratio of 30% and generate 100 random simulations (*i.e.*, 100 noise realizations); we set the required confidence level to 95% and we count the number of successful outcomes for the condition: "$\theta_i^\star$ is inside the computed confidence interval", for every individual parameter in $\theta^\star$.

Table 2 gives the percentages of successes for both the classical confidence intervals and the new specialized confidence intervals. Note that the specialized method gives percentages closer to the required confidence level of 95%, while the classical method, which ignores some aspects of the

problem, can give as low as 75% successful counts, for the same required level of 95%. This means that in up to 20% of the outcomes, the true value $\theta_i^\star$ is inside the specialized confidence interval, but outside the classical one.

The new confidence intervals are a bit wider than the classical ones. In Figure 9 we plot the *averaged* confidence intervals in the 100 simulations next to the boxplots of the computed θ parameters, for the noise level of 30%. We note that the length difference is just between 5 to 50%, and thus the specialized confidence intervals are still tight enough for giving useful information on the estimated parameters.

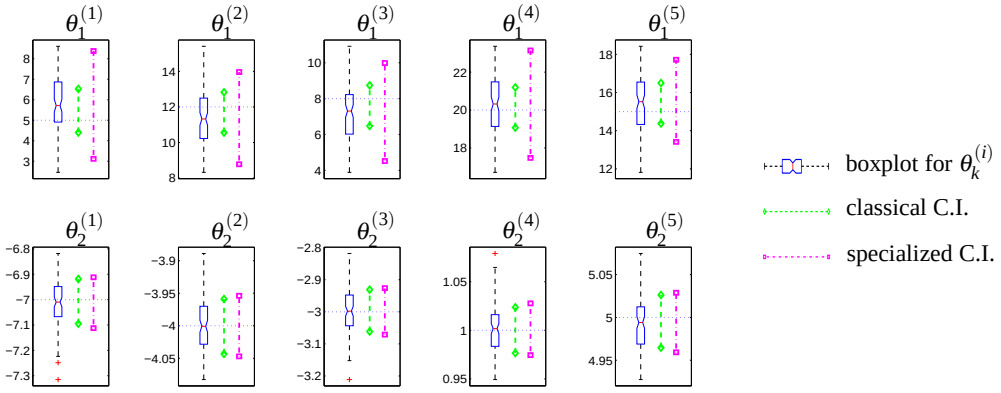


Fig. 9. Boxplots for the θ parameters in 100 simulations; averaged confidence intervals, where the shorter are the classical and the wider are the specialized confidence intervals. Horizontal dotted line marks the true value of the parameters.

8. Conclusions

We have presented a semiparametric problem formulation, its statistical properties, its computational solution, a real life application, and simulation examples.

The main points of this study reveal that it is relatively easy to include a nonparametric part into a nonlinear regression problem, in the case when the nonparametric part can be restricted via a weighted norm (*e.g.*, by imposing smoothness of the baseline term). The nonparametric part can then be modeled by using a spline basis, and thus only a linear term is actually added in the nonlinear least squares optimization criterion.

However, special care should be taken when statistical information is retrieved from a regularized semiparametric regression problem, since in general we obtain biased estimates (but the bias can be corrected), and the classical nonlinear regression confidence bounds that use the Fisher information matrix must be adapted to take into account the regularization term.

Future work involves using *regression* splines (Friedman, 1991) in the semiparametric modeling context and optimizing over the number and position of the knots, instead of optimizing over a regularization parameter λ.

A. Efficient computation of function and Jacobian values

If the evaluation of the nonlinear function $F(t,\theta)$ (as a function of θ) and of its gradient (also with respect to θ) are not very computationally demanding, then the evaluation of $f(\theta,\lambda)$ and $J(\theta,\lambda)$, defined at the end of Section 4.3, for any values of the parameters $\theta \in \Theta$ and $\lambda > 0$, can be achieved with linear computational complexity in the problem dimensions m (number of regression points), n (number of basis functions) and p (length of the vector θ).

In the computation of the function and Jacobian values, the inverse of the matrix $A^\top A + m\lambda B$ appears. Dealing with this inverse would be much more efficient if A and B were diagonal matrices, instead of full matrices: then the "inversion" would only involve a diagonal matrix, for any possible value of λ. Algorithm 2 shows the preprocessing operations that should be executed beforehand, *i.e.*, the simultaneous diagonalization of the matrices $A^\top A$ and B, achieved using the generalized singular value decomposition.‖

Algorithm 2 Preprocessing. complexity

Input: matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times n}$.

1: Compute Cholesky decomposition of B, $B = D^\top D$; $n^3/3$

2: Compute GSVD of the pair (A,D); $\mathcal{O}(mn^2) + \mathcal{O}(n^3)$

$$A = UCX^{-1}, \qquad D = VSX^{-1},$$

with $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ orthogonal, $X \in \mathbb{R}^{n\times n}$ invertible, and $C \in \mathbb{R}^{m\times n}$, $S \in \mathbb{R}^{n\times n}$ positive diagonal matrices.

Output: Elements of the GSVD U,V,C,S,X. total: $\mathcal{O}(mn^2) + \mathcal{O}(n^3)$.

Algorithm 3 shows how to employ the preprocessing step in order to compute the function and Jacobian values in a fast way, for any parameters θ and λ. Note that the nonlinear least squares minimization (11) involves the 2-norm of $f(\theta,\lambda)$; therefore, multiplication of $f(\theta,\lambda)$ from the left with an orthogonal matrix does not change the optimization criterion. Algorithm 3 exploits this fact in order to avoid unnecessary matrix multiplications with the orthogonal matrices U and V from Algorithm 2, and provides orthogonally transformed function and Jacobian values.

B. Efficient computation of the GCV function value

The evaluation of the GCV function defined in (12) involves computing the misfit between the data $\mathbf{y}$ and the current model $\widehat{\mathbf{y}}_\lambda$, for the numerator, and computing the trace of the influence matrix $S(\lambda)$ in (13), for the denominator. The estimated model $\widehat{\mathbf{y}}_\lambda$ is easily available from the nonlinear least squares step, thus the numerator is computed in $\mathcal{O}(m)$.

Since the Jacobian matrix $\nabla\mathbf{F}(\widehat{\theta}_\lambda)$ changes at each new evaluation (together with the current iterate $\widehat{\theta}_\lambda = \theta_{\lambda_k}$), it is not possible to compute the influence matrix as efficiently as in the classical linear setting (Golub and von Matt, 1997). However, we propose to use also the preprocessing step described in Algorithm 2 in order to create an easily computable diagonal part in the influence matrix formula. The term $G := U^\top \nabla\mathbf{F}(\widehat{\theta}_\lambda)$ is already available from the nonlinear least squares

‖In practical implementations, we use the "economic" generalized singular value decomposition, but here, for a simplified presentation, we only give formulas using the full decomposition, *i.e.*, with square orthogonal matrix of generalized singular vectors.

Algorithm 3 Compute function value $f(\theta,\lambda)$, and Jacobian $J(\theta,\lambda)$. complexity

$$f(\theta,\lambda) := B(\lambda)\mathbf{y}(\theta), \quad J(\theta,\lambda) := B(\lambda)\nabla\mathbf{y}(\theta),$$
$$\text{where } B(\lambda) = \begin{bmatrix} U^\top & 0 \\ 0 & V^\top \end{bmatrix} \begin{bmatrix} I_m - A(A^\top A + m\lambda B)^{-1}A^\top \\ \sqrt{m\lambda}\, D(A^\top A + m\lambda B)^{-1}A^\top \end{bmatrix}.$$

Input: matrices U,V,C,S, computed from A and B via Algorithm 2, parameters λ and θ.

1: Evaluate nonlinear expressions $v \leftarrow \mathbf{y}(\theta) \in \mathbb{R}^m$ and $G \leftarrow \nabla\mathbf{y}(\theta) \in \mathbb{R}^{m\times p}$. — depends on nonlinear $\mathbf{y}$

2: $v \leftarrow U^\top v, \quad G \leftarrow U^\top G$. — $\mathcal{O}(mn)$, resp. $\mathcal{O}(mnp)$

3: $c_1 \leftarrow \operatorname{diag}(I_m - C(C^\top C + m\lambda S^2)^{-1}C^\top)$ — $\mathcal{O}(n)$

4: $c_2 \leftarrow \sqrt{m\lambda}\operatorname{diag} S(C^\top C + m\lambda S^2)^{-1}C^\top$. — $\mathcal{O}(n)$

5: Set $f(\theta,\lambda) \leftarrow \begin{bmatrix} c_1 \odot v \\ c_2 \odot v_{1:n} \end{bmatrix}$, where $\odot$ denotes element-wise product. — $\mathcal{O}(m+n)$

6: **for** $j = 1,\ldots,p$ **do**

7: $\quad \operatorname{col}_j(J(\theta,\lambda)) \leftarrow \begin{bmatrix} c_1 \odot G_{:,j} \\ c_2 \odot G_{1:n,j} \end{bmatrix}$, — $\mathcal{O}(m+n)$

8: **end for**

Output: $f(\theta,\lambda)$ and $J(\theta,\lambda)$. — total: $\geq \mathcal{O}(mn+m+3n)$, resp. $\geq \mathcal{O}(mnp+np+mp)$

function and Jacobian evaluation, see step 2: in Algorithm 3. The influence matrix becomes (up to an orthogonal similarity transformation, which leaves the trace invariant):

$$S(\lambda) = \begin{bmatrix} G & C \end{bmatrix} \begin{bmatrix} G^\top G & G^\top C \\ C^\top G & C^\top C + m\lambda S^2 \end{bmatrix}^{-1} \begin{bmatrix} G^\top \\ C^\top \end{bmatrix},$$

where U, C and S are matrices provided by Algorithm 2. After some manipulations we get:

$$S(\lambda) = A(\lambda) + (I_m - A(\lambda))G\left[G^\top(I_m - A(\lambda))G\right]^{-1}G^\top(I_m - A(\lambda)),$$

where $A(\lambda) = C(C^\top C + m\lambda S^2)^{-1}C^\top$. Finally, using a QR decomposition of $(I_m - A(\lambda))^{1/2}G = QR$ yields the simplified formula

$$\operatorname{Tr} S(\lambda) = \operatorname{Tr} A(\lambda) + \operatorname{Tr} R(R^\top R)^{-1}R^\top Q^\top(I_m - A(\lambda))Q = \operatorname{Tr} A(\lambda) + p - \operatorname{Tr} R(R^\top R)^{-1}R^\top Q^\top A(\lambda)Q.$$

Note that $\operatorname{Tr} S(\lambda)$ is also used in the computation of confidence intervals (subsection 5.2), since it represents the number of effective parameters.

Acknowledgements

The authors thank the anonymous referees for their detailed comments and suggestions, which helped improving the manuscript.

Diana Sima is a research assistant and Dr. Sabine Van Huffel is a full professor at the Katholieke Universiteit Leuven, Belgium. Research supported by

Research Council KUL: GOA-AMBioRICS, several PhD/postdoc & fellow grants;
Flemish Government: FWO: PhD/postdoc grants, projects, G.0407.02 (support vector machines), G.0269.02 (magnetic resonance spectroscopic imaging), G.0270.02 (nonlinear Lp approximation), G.0360.05 (EEG, Epileptic), research communities (ICCoS, ANMMM); **IWT:** PhD Grants;

Belgian Federal Science Policy Office: IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling');
EU: PDT-COIL, BIOPATTERN, ETUMOUR.

References

Bhattacharya, P. and P.-L. Zhao (1997). Semiparametric inference in a partial linear model. *The Annals of Statistics 25*(1), 244–262.

Carter, C. K. and R. Kohn (1997). Semiparametric Bayesian inference for time series with mixed spectra. *Journal of the Royal Statistical Society: Series B 59*(1), 255–268.

Coleman, T. and Y. Li (1996). An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization 6*, 418–445.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31*, 377–403.

de Boor, C. (1978). *A Practical Guide to Splines*. Berlin: Springer.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*(2), 89–121.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics 19*(1).

Golub, G. H. and C. F. Van Loan (1996). *Matrix Computations* (third ed.). Baltimore, Maryland: Johns Hopkins University Press.

Golub, G. H. and U. von Matt (1997). Generalized cross-validation for large scale problems. In S. Van Huffel (Ed.), *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, pp. 139–148. SIAM.

Ke, C. and Y. Wang (2004). Smoothing spline nonlinear nonparametric regression models. *Journal of the American Statistical Association 99*(468), 1166–1175.

Lee, T. C. M. (2003). Smoothing parameter selection for smoothing splines: A simulation study. *Computational Statistics and Data Analysis 42*(1-2), 139–148.

Lemmerling, P., L. Vanhamme, H. J. A. in't Zandt, S. Van Huffel, and P. Van Hecke (2002). Time-domain quantification of short-echo-time in-vivo proton MRS. *MAGMA 15*, 178–179.

Moody, J. E. (1992). The *effective* number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In Moody, Hanson, and Lippmann (Eds.), *Advances in Neural Information Processing Systems*, Volume 4, pp. 847–854. Palo Alto: Morgan Kaufmann.

Moré, J. J. (1978). The Levenberg-Marquardt algorithm: Implementation and theory. In G. A. Watson (Ed.), *Numerical Analysis: Proceedings of the Biennial Conference held at Dundee, June 28-July 1, 1977 (Lecture Notes in Mathematics #630)*, pp. 104–116. Springer Verlag.

Provencher, S. W. (1993). Estimation of metabolite concentrations from localized *in vivo* proton NMR spectra. *Magnetic Resonance in Medicine 30*(6).

Ratiney, H., M. Sdika, Y. Coenradie, S. Cavassila, D. van Ormondt, and D. Graveron-Demilly (2004). Time-domain semi-parametric estimation based on a metabolite basis set. *NMR in Biomedicine 17*, 1–13.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics 11*, 735–757.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge University Press.

Seber, G. A. F. and C. J. Wild (1989). *Nonlinear Regression*. Probability and Mathematical Statistics. John Wiley & Sons.

Sima, D. M. and S. Van Huffel (2004). A class of template splines. Technical Report 04-228, K.U. Leuven, E.E. Dept. (ESAT-SISTA). To appear in *Computational Statistics and Data Analysis*.

Spall, J. C. and J. P. Garner (1990). Parameter identification of state-space models with nuisance parameters. *IEEE Transactions on Aerospace and Electronic Systems 26*(6), 992–998.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics 59. SIAM.