Université catholique de Louvain
Institut de statistique, biostatistique et sciences actuarielles

Thèse de doctorat en co-tutelle avec

Ruprecht-Karls-Universität Heidelberg
Naturwissenschaftlich-Mathematische Gesamtfakultät

Non parametric estimation in the presence of noise with unknown distribution

Membres du jury:

Prof. Rainer Dahlhaus
Prof. Christine De Mol
Prof. Jan Johannes
Prof. Sébastien Van Bellegem
Prof. Ingrid Van Keilegom
Prof. Pierre Van Moerbeke

Thèse présentée en vue de l'obtention du grade de Docteur en Sciences (orientation statistique) par:

Maik Schwarz

Louvain-la-Neuve
Juillet 2011

Acknowledgements

In the first place I am much obliged to my advisors and co-authors Jan Johannes and Sébastien Van Bellegem for the numerous (not only) scientific discussions in Louvain-la-Neuve and Heidelberg from which I learned a lot. I am grateful to my senior advisors Ingrid Van Keilegom and Rainer Dahlhaus for their support and advice, and to Christine De Mol and Pierre Van Moerbeke for their critical questions and helpful suggestions. I cordially thank Anne Vanhems and Jean-Pierre Florens for their hospitality and for stimulating discussions in Toulouse. I would like to say thanks to my fellow PhD students for many conversations and the good time; and to the administrative team of the ISBA for their friendliness and helpfulness. Meinen Eltern und meiner Schwester danke ich von Herzen für all die Unterstützung. Τελικά ευχαριστώ την Μάνια μου που μ'άντεχε και υποστήριζε με πολλή αγάπη.

I gratefully acknowledge the financial support from the Belgian Science Policy of the Belgian government (IAP Network grant no. P6/03) and the Fonds Spéciaux de Recherche of the Université catholique de Louvain.

Contents

Introduction

In this thesis we develop identification and non parametric estimation techniques in the context of statistical ill-posed inverse problems with unknown operator. Non parametric statistics is concerned with the estimation of an unknown function g in a functional class which cannot be indexed by a finite-dimensional parameter space. Typical non parametric estimation problems arise in density deconvolution and regression. An introduction to classical non parametric estimation methods and their statistical properties is provided in the textbook by Tsybakov (2004).

An *inverse problem* arises if we are not interested in the function g itself, but in the solution f of the equation $g = Tf$, where T denotes some transform. If g and T are known, the inverse problem may be solved by numerical methods. In contrast to this, we are going to deal with *statistical inverse problems with unknown operator*. This means that in the settings we are going to consider, both g and T are unknown and need to be estimated in the context of a statistical model.

We organize this introduction as follows. After a brief paragraph on non parametric estimation, we give some background information on inverse problems in a deterministic context before discussing statistical inverse problems in general. Then, we explain the objectives we will be guided by throughout this thesis, and we present the particular models treated in Chapters 1 to 4. Finally, we give a résumé of our contributions.

Non parametric estimation

Non parametric estimation is concerned with the reconstruction of an unknown function belonging to some infinite-dimensional vector space $\mathbb{G}$, based for example on an independent and identically distributed (iid.) sample from a random variable Y. More precisely, we will denote by Y a real valued random variable with a probability distribution P (abbreviated by $Y \sim P$) eventually belonging

to a collection $\mathcal{P}_{\mathbb{G}} = \{P_g \mid g \in \mathbb{G}\}$ indexed by $\mathbb{G}$. Such a collection is called a *statistical model*. The model is said to be *correctly specified* if there exists $g \in \mathbb{G}$ such that $Y \sim P_g$. It is further called *identifiable* if $P_g = P_{g'}$ implies $g = g'$ for any elements P_g and $P_{g'}$ of $\mathcal{P}_{\mathbb{G}}$. The function g could be a probability density or a regression function, for example. Classical non parametric estimation methods are kernel estimators (e.g. Rosenblatt, 1956) or estimation by projection (e.g. Tsybakov, 2004).

Background: Inverse problems

Suppose now that we are not actually interested in finding $g \in \mathbb{G}$ itself, but a function f belonging to another vector space $\mathbb{H}$, given implicitly by the relation $g = Tf$, where T denotes some transform. The reconstruction of f from an approximation of g is called an *inverse problem,* because it necessitates the inversion of the operation T. According to Hadamard (1902), an inverse problem $g = Tf$ is called *well-posed* if the following three conditions are satisfied; otherwise it is called *ill-posed*. In what follows, we will suppose that $\mathbb{H}$ and $\mathbb{G}$ are Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_{\mathbb{H}}$, $\langle \cdot, \cdot \rangle_{\mathbb{G}}$ and associated norms $\|\cdot\|_{\mathbb{H}}$ and $\|\cdot\|_{\mathbb{G}}$, respectively.

Condition 1: Existence of a solution Firstly, a solution must exist, meaning that g belongs to the image of T in $\mathbb{G}$. If this is not the case, one can consider the least squares solution if it exists. It is defined as the minimizer of the distance $\|g - Tf\|_{\mathbb{G}}$ over all $f \in \mathbb{H}$.

Condition 2: Uniqueness of the solution Secondly, the solution has to be unique, which is the case if and only if T is injective. If this condition is violated, one cannot find a well-defined solution f even if g is fully known or can be approximated with arbitrarily high precision. However, the operator may be injective on subsets of the space $\mathbb{H}$, and uniqueness can be recovered by restricting the class of potential solutions f to such a subset. In order not to be too restrictive, it is desirable to find classes which on the one hand include a large variety of possible solutions and on the other hand can be defined by conditions which one can verify in practice.

Condition 3: Continuous inverse operator Hadamard's third condition demands that the solution depend on g in a continuous way, or in other words that the inverse operator T^{-1} be continuous. Suppose that we do not know g exactly, but that for any given $\delta > 0$, we can compute an approximation g^{δ} satisfying $\|g - g^{\delta}\|_{\mathbb{G}} < \delta$. If the inverse problem is well-posed, this enables us to approximate the solution f as well. Indeed, by the definition of continuity, for every $\varepsilon > 0$ there is a $\delta > 0$ such that $\|T^{-1}g^{\delta} - f\|_{\mathbb{H}} < \varepsilon$. If however T^{-1} is discontinuous, we cannot control the approximation error in $\mathbb{H}$.

Example Let us illustrate Hadamard's conditions with an example. Denote by $\mathcal{L}^2[0,1]$ the space of square integrable complex-valued functions on the unit interval with respect to the Lebesgue measure λ, endowed with the inner product $\langle f_1, f_2\rangle_{\mathcal{L}^2} = \int_{[0,1]} f_1\overline{f_2}\,\mathrm{d}\lambda$. We consider the quotient spaces $\mathbb{H} = \mathbb{G} = L^2[0,1] := \mathcal{L}^2[0,1] \,/\, \doteq$ with respect to the equivalence relation $f_1 \doteq f_2$ which is true if and only if $f_1 = f_2$ almost everywhere with respect to the Lebesgue measure. As the integral is unique up to a change of the integrand on a null set, the inner product of the space $L^2[0,1]$ can also be written $\langle f_1, f_2\rangle_{L^2} = \int_{[0,1]} f_1\overline{f_2}\,\mathrm{d}\lambda$, and by slight abuse of language we use the term «function» for the elements of $L^2[0,1]$ as well.

Define now a linear operator from $\mathbb{H}$ to $\mathbb{G}$ by $g(x) = (Tf)(x) = \int_0^x f(t)\mathrm{d}t$. Its inverse is the differential operator, that is $T^{-1}g = g'$. Obviously, a unique solution to the equation $g = Tf$ exists if g is a differentiable function. Thus, the first and the second Hadamard conditions are satisfied if we restrict the image space $\mathbb{G}$ to a class of differentiable functions.

Consider Hadamard's third condition and let us illustrate how it depends on the topological structures of the particular spaces $\mathbb{H}$ and $\mathbb{G}$ under consideration. An important example of differentiability classes is the *Sobolev class* W_p^{per} of periodic functions on $[0,1]$. It contains the p-times differentiable periodic functions whose first $(p-1)$ derivatives are absolutely continuous. It can be written as

$$W_p^{\mathrm{per}} = \left\{ h \in L^2[0,1] \;\middle|\; \|h\|_{W_p}^2 := \sum_{j\in\mathbb{Z}} |j|^{2p}\, \langle h, e_j\rangle_{L^2}^2 < \infty \right\},$$

where $\{e_j \,|\, j \in \mathbb{Z}\}$ denotes the exponential basis of $L^2[0,1]$ which is defined by $e_j(x) = \exp(-i2\pi jx)$ for all $j \in \mathbb{Z}$ and $x \in [0,1]$. The p-th derivative $g^{(p)}$ of a p-times differentiable function $g \in L^2[0,1]$ can be written in this basis as

$$g^{(p)} = \sum_{j\in\mathbb{Z}} (2i\pi j)^p\; [g]_j\, e_j,$$

where $[g]_j := \langle g, e_j\rangle_{L^2}$ denotes the j-th coefficient of g (e.g. Neubauer, 1988b). It follows that for a differentiable $g \in L^2[0,1]$, we have

$$\|T^{-1}g\|_{L^2}^2 = \sum_{j\in\mathbb{Z}} 4\pi^2 j^2\; |[g]_j|^2 = 4\pi^2 \|g\|_{W_1}^2.$$

This implies immediately that the inverse T^{-1} is continuous if we equip $\mathbb{H}$ with the L^2 norm and $\mathbb{G}$ with the W_1 norm. If we consider the L^2 norm in both spaces, however, T^{-1} is discontinuous. Indeed, the sequence of functions $g_n := n^{-1}e_n$ in $\mathbb{G}$ clearly converges to zero in the L^2 norm as n tends to infinity, but we have $\|T^{-1}g_n\|_{L^2}^2 = 4\pi^2$ for all $n \in \mathbb{N}$.

This example shows in particular that the well-posedness of an inverse problem actually depends on the topological structure of the spaces $\mathbb{H}$ and $\mathbb{G}$.

Regularization

Suppose that we want to evaluate the approximation of the solution of an inverse problem by means of the integrated squared error. In this case, we have no alternative to the L^2 structure of the spaces. If the inverse operator T^{-1} is not continuous with respect to the L^2 norms, one may *regularize* it, that is, approximate it by a continuous operator. The monograph by Engl et al. (1996) gives an introduction to the regularization of inverse problems and provides a collection of different techniques. Let us mention two examples here.

Tikhonov regularization Given an inverse problem $g = Tf$ and an approximation g^δ of g with $\|g - g^\delta\|_{\mathbb{G}} < \delta$, the regularized solution is defined as the minimizer of the *Tikhonov functional*, that is

$$f_\alpha^\delta := \operatorname*{argmin}_{f \in \mathbb{H}} \left\{ \|Tf - g^\delta\|_{\mathbb{G}}^2 + \alpha \, \|f\|_{\mathbb{H}}^2 \right\}$$

for some $\alpha > 0$ which is called *regularization parameter*. The role of this parameter is crucial. On the one hand, the vulnerability of the approximation f_α^δ to perturbations of g decreases as α increases. On the other hand, if considering the true g instead of g^δ in the Tikhonov functional, the approximation will deteriorate with increasing α. This systematic error is the price for stability. See Engl et al. (1996) for a discussion of how α has to be chosen subject to δ such that f_α^δ converges to the solution as δ tends to zero.

In certain cases, the Tikhonov regularization has an alternative, more explicit representation. Suppose that the operator T is compact and that its image is infinite-dimensional, which implies that T^{-1} is not continuous. Denote the adjoint of T by T^*. Then, the symmetric operator T^*T admits a spectral decomposition. This means that there is an orthonormal basis of eigenfunctions $\{u_j\} \subset \mathbb{H}$ and a corresponding sequence of eigenvalues $\{\tau_j^2\}$ such that $T^*Tu_j = \tau_j^2 u_j$ for all j. The eigenvalues tend to zero as j tends to infinity. It is easy to verify that the definition $v_j := Tu_j \, \|Tu_j\|_{\mathbb{G}}^{-1}$ yields an orthonormal system in $\mathbb{G}$. Furthermore, we have $Tu_j = \tau_j v_j$ and $T^*v_j = \tau_j u_j$. The collection $(\{u_j\}, \{v_j\}, \tau_j)$ is called the *singular value decomposition* of the operator T. We deduce that the solution of the inverse problem $g = Tf$ can be written as

$$f = \sum_{j=1}^{\infty} \frac{1}{\tau_j} \, [g]_j \, u_j.$$

Note that the sequence τ_j^{-1} is unbounded. Thus, small perturbations of the coefficients $[g]_j = \langle g, v_j \rangle_{\mathbb{G}}$ are greatly amplified, and replacing g with the approximation g^δ from above, we may end up far off the true solution. This mechanism reflects the discontinuity of the inverse operator. In this situation,

Tikhonov's regularization can be written as

$$f_\alpha^\delta = \sum_{j=1}^{\infty} \frac{\tau_j}{\tau_j^2 + \alpha} [g^\delta]_j \, u_j, \tag{1}$$

where $\alpha > 0$ is the regularization parameter that appears in the Tikhonov functional. Due to the regularization, errors in the j-th coefficient of g are only amplified by the factor $\tau_j/(\tau_j^2+\alpha)$ which remains bounded as j tends to infinity. Heuristically speaking, we have replaced T^{-1} with a continuous operator T_α^{-1} described by (1).

Galerkin solution Another regularization approach consists in approximating the solution f of the inverse problem $g = Tf$ in a finite dimensional subspace $\mathbb{H}_k$ of $\mathbb{H}$ generated by linearly independent vectors $\{w_1, \ldots, w_k\}$. The Galerkin solution is defined as

$$f_k^\delta := \operatorname*{argmin}_{f \in \mathbb{H}_k} \|Tf - g^\delta\|_{\mathbb{G}},$$

where g^δ denotes the approximation of g from above. The dimension k acts as the regularization parameter. Like the Tikhonov regularization, the Galerkin solution is defined implicitly. However, if we choose the elements w_j generating the space $\mathbb{H}_k$ be the eigenfunctions u_j of T^*T with corresponding eigenvalues τ_j^2, then the Galerkin solution takes the explicit form

$$f_k^\delta = \sum_{j=1}^{k} \frac{1}{\tau_j} [g^\delta]_j \, u_j.$$

Of course, this requires that the eigenfunctions of T^*T are known. For example, the eigenfunctions of a convolution operator on a compact space are given by the exponential basis. In many situations, however, no information about the eigenfunctions is available. Obviously, the restriction to the finite-dimensional space causes a systematic approximation error while stabilizing the Galerkin solution at the same time, so the same remarks as in the case of the Tikhonov regularization hold in regard to the choice of the regularization parameter.

Statistical ill-posed inverse problems with unknown operator

In a classical statistical inverse problem, the left hand side g is estimated by some $\widehat{g}$ in the framework of a specific statistical model $\{P_g \mid g \in \mathbb{G}\}$. Suppose that this model is correctly specified and identified with respect to the function g. If the first two Hadamard conditions are satisfied, that is if T is injective and g belongs to its image, then the model $\{P_{Tf} \mid f \in \mathbb{H}\}$ is obviously correctly specified and identified for the solution f. We will assume this throughout this work.

Consider Hadamard's third condition. If it is satisfied, that is if the inverse operator is continuous, the natural estimator $\widehat{f} := T^{-1}\widehat{g}$ of f eventually inherits desirable properties such as consistency or asymptotic normality from $\widehat{g}$ by virtue of the delta method. Otherwise the inverse problem is ill-posed and we need to regularize T^{-1}. To this end, one can for instance use the regularization methods discussed in the previous paragraph. An estimator is then obtained by replacing the deterministic approximation g^δ by the estimator $\widehat{g}$. In this thesis, we use the general Galerkin approach and its special case, the orthogonal series estimator.

Additional problems arise when the operator T is unknown. For example, the solution may fail to be identifiable even if the unknown operator is assumed to be injective. Indeed, the same image g could possibly be represented by combinations $g = T_1 f_1 = T_2 f_2$ of different solutions f_1, f_2 and operators T_1, T_2.

We will discuss two approaches of recovering identifiability in this situation. On the one hand, we will develop identifiability conditions that involve both the solution and the operator simultaneously. More precisely, we will assume that T belongs to a parametric class. In a second step, we define a functional class of possible solutions such that f is identifiable in spite of the uncertainty about the operator. On the other hand, we will consider settings where additional observations allow for a preliminary estimation of the operator, which also ensures the identifiability of the solution. As far as estimation is concerned, the methods discussed in the paragraph on regularization remain available. An estimator of f can be constructed by replacing the approximation g^δ of g and the operator T by their respective estimates $\widehat{g}$ and $\widehat{T}$.

Objectives of this thesis

Within the general framework of non parametric estimation in statistical inverse problems with unknown operator, this thesis comprises four chapters treating various specific models by technically rather different approaches. A leitmotif of the present work lies in the development and evaluation of the proposed methods along the thread defined by the four criteria identifiability, consistency, minimax optimality and adaptation, which we will now discuss in more detail.

Identifiability Identifiability in the sense of injectivity of the operator T is our minimal requirement concerning the model specification. Though, as the operator will be unknown in our models, we need further assumptions. Smoothness conditions on f and T are a common way of dealing with this problem. However, we will also develop results allowing for identification when no information about the solution's smoothness is available beforehand.

Consistency This is a minimal quality criterion for statistical estimators. Roughly speaking, consistency means convergence of the estimator to the estimated object as the sample size tends to infinity: the more information is available, the more precise should the estimate become. Consequently, there are as many notions of consistency as there are of convergence. In a well-posed inverse problem, the consistency of the estimator $\widehat{f} = T^{-1}\widehat{g}$ follows by the delta method if $\widehat{g}$ is consistent. Being confronted with inverse problems that are ill-posed with respect to the initial topological structures of $\mathbb{H}$ and $\mathbb{G}$, we will either have to regularize T^{-1} or to choose the notion of convergence in a way as to ensure consistency.

Minimax One way of benchmarking the performance of a consistent estimator $\widehat{f}$ of f is minimax theory with respect to some risk $\mathcal{R}(f, \widehat{f})$. For example, if $\mathbb{H} = L^2$, one could consider the mean integrated squared error $\mathcal{R}(f, \widehat{f}) = \mathbf{E}\|f - \widehat{f}\|_{L^2}^2$. As the value of the risk depends on the target density f, one considers the supremum of the risk over a class $\mathcal{F} \subset \mathbb{H}$. Note that the estimator $\widehat{f}$ and hence the value of $\mathcal{R}(f, \widehat{f})$ also depend on the unknown operator T. Therefore, we take another supremum over a class $\mathcal{T}$ of possible operators. The performance of a given estimator $\widehat{f}$ of f with respect to the classes $\mathcal{F}$ and $\mathcal{T}$ is then measured by its *maximal risk*

$$\sup_{f \in \mathcal{F}} \sup_{T \in \mathcal{T}} \mathcal{R}(f, \widehat{f}).$$

An estimator is called *minimax optimal* if its maximal risk coincides, up to a constant, with the *minimax risk*

$$\inf_{\widetilde{f}} \sup_{f \in \mathcal{F}} \sup_{T \in \mathcal{T}} \mathcal{R}(f, \widetilde{f}),$$

where the infimum is taken over all possible estimators $\widetilde{f}$ of f. In ill-posed inverse problems, the estimator generally depends on some regularization parameter as for example the dimension of the orthogonal series estimator mentioned above. Minimax optimality of the estimator can only be achieved if this parameter is chosen in an optimal way. The optimal choice generally depends on characteristics of both the solution and the estimator via the classes $\mathcal{F}$ and $\mathcal{T}$. For an introduction to minimax theory in non parametric estimation, see also Tsybakov (2004).

Adaptation The difficulty of choosing the regularization parameter in ill-posed statistical inverse problems can be overcome by *adaptation.* By adaptive, we mean estimators which depend exclusively on the data but which are nevertheless minimax optimal with respect to a wide range of classes $\mathcal{F}$ and $\mathcal{T}$.

Density deconvolution, frontiers, and instrumental regression

Having explained the general theoretical framework of this thesis, let us now introduce the specific statistical models we are going to work with. More detailed descriptions of the models and motivation by applications are provided in the introductions to the respective chapters.

Density deconvolution on $\mathbb{R}$ Suppose we want to estimate non parametrically a probability density $f \in L^2(\mathbb{R})$ of a real-valued random variable X. When an iid. sample from f is available, the density can be estimated directly by means of a kernel estimator, for example. We will however drop the assumption that the available data are an exact sample from the density of interest. Instead, we are going to suppose that we observe an sample of iid. copies of the variable

$$Y = X + \varepsilon. \tag{2}$$

The real-valued random variable ε is supposed to be independent of X. It models a stochastic measurement error present in the observed data and is referred to as the *error* for short. Assuming that ε has a density $h \in L^2(\mathbb{R})$, the variable Y is distributed according to the density $h*f = g \in L^2(\mathbb{R})$, where $*$ denotes *convolution*, meaning that $g(x) = \int_{\mathbb{R}} f(x-t)h(t)\mathrm{d}t$ for all $x \in \mathbb{R}$. In other words, the density g of the observations is the image of the density f under a convolution operator T_h depending on the error density h. Recovering f from the observations is hence an inverse problem. In Section 1.1 of the first chapter, we provide a literature review of some classical and more recent estimation methods in deconvolution models. A quite exhaustive overview can be found in the monograph by Meister (2009).

Note that the model is correctly specified and identifiable when the error density is known and its characteristic function is strictly positive on the whole real line. However, we will investigate under which conditions we can obtain identification and consistent estimation when the operator is not fully known.

Density deconvolution on the circle Again, consider the model described by (2), but suppose that the random variables are defined on the circle instead of the real line. Identifying the circle with the half open unit interval $[0,1)$, we assume $f, h \in L^2[0,1]$. Using the floor function $\lfloor\cdot\rfloor$, the model equation can be rewritten as $Y = X + \varepsilon - \lfloor X + \varepsilon \rfloor$ and the circular convolution operator as $(T_h f)(x) = g(x) = \int_{[0,1)} h((x-s) - \lfloor x-s \rfloor)\, f(s)\,\mathrm{d}s$ for all $x \in [0,1)$. The model is correctly specified and identifiable when the error density h is known and its Fourier coefficients are strictly positive. However, we will not assume any a priori knowledge on h. Instead, we suppose that an extra sample is observed that allows for a preliminary estimation of h. Our objective in this model is adaptive minimax optimal estimation of f.

Frontier estimation Interestingly, a similar deconvolution problem as in (2) arises in the context of frontier estimation. A *frontier* is the boundary point of the support of a probability distribution. In economic models, the support boundary of a production distribution represents the maximal production output that can be achieved for a given input (investment).

More precisely, suppose that a positive real output $z \in \mathbb{R}_+$ can be produced using an input $x \in \mathbb{R}_+$ if and only if the pair (x, z) belongs to the set of *production possibilities* Φ which is a subset of the quadrant $\mathbb{R}_+^2$. The frontier is defined as the boundary of this set. Under appropriate assumptions on Φ, it can be described by a continuous function φ mapping $\mathbb{R}_+$ to itself. One objective in this model is the non parametric estimation of φ based on an iid. sample from the joint in- and output (X, Z) associated to individual production units.

However, a sample from the exact production distribution may not be available. Assume for example that the input data is contaminated with a centered normally distributed measurement error of unknown variance. This amounts to observing an iid. sample from (Y, Z) rather than from (X, Z), where $Y = X + \varepsilon$ is a version of X which is subject to the independent normal error ε. This relation lets us expect that a similar deconvolution problem as above is involved in the estimation of the frontier function φ, although we are not primarily interested in reconstructing the density of X. Indeed, we will see that a preliminary step in the estimation of the boundary consists in reconstructing not the density, but the survival function of X. This can be expressed as a deconvolution problem as well.

Instrumental regression Regression is a classical model in non parametric statistics: the dependence of a real-valued response variable Y on a regressor $Z \in \mathbb{R}^p$ is modeled by

$$Y = f(Z) + U, \tag{3}$$

where f is the regression function and the random variable U the error with zero mean. The objective is to estimate f based on an iid. sample from (Y, Z). The classical estimation methods require that the regressor Z be stochastically independent of the error or at least that the conditional expectation of U given Z satisfies $\mathbf{E}[U|Z] = 0$. This is an assumption which does not always hold in applications.

If Z and U are dependent, the regression function f is no longer identifiable by the observation of (Y, Z). However, if an *instrumental variable* is available, identification and estimation are possible. An instrumental variable, or *instrument* for short, is a random variable $W \in \mathbb{R}^q$ which is correlated to the regressor Z on the one hand but satisfying $\mathbf{E}[U|W] = 0$ on the other hand. The regression function f is to be estimated based on an iid. sample from (Y, Z, W). This is a statistical inverse problem. Indeed, taking the conditional expectation with respect to the instruments W in (3), we obtain a regression problem

expressed by

$$g = \mathbf{E}[Y|W] = \mathbf{E}[f(Z)|W] = Tf,$$

where T denotes the conditional expectation operator. This operator is obviously determined by the joint distribution of the couple (Z, W) which is unknown in general. We are thus dealing with an inverse problem with unknown operator T which has to be estimated based on the sample from (Z, W). The left hand side g can be estimated using the sample from the couple (Y, W). For a methodological overview regarding instrumental regression, we refer to Darolles et al. (2001) and Carrasco et al. (2007).

Contributions of this thesis

To conclude the introduction, let us now summarize the results that we contribute in this thesis on the subject of statistical inverse problems with unknown operator.

Chapter 1
Identification & consistent density deconvolution

In the framework of the density deconvolution model on $\mathbb{R}$, we assume that the error density belongs to a parametric class. More precisely, we show our results under the assumption that the error is normally distributed with mean zero and unknown variance σ^2, but we discuss other choices of distribution classes that are possible as well. In this setting, we prove a sufficient condition under which the solution f and the error variance σ^2 are identifiable from the observation of the noisy sample alone. The novelty of this condition lies in its being independent of the solution's smoothness properties – its definition does not involve any Fourier coefficients. Instead, the condition uses an easily interpretable property in the time domain. Finally, we propose a simultaneous minimum contrast estimator of the solution and the error variance. We prove its consistency by showing that the underlying inverse problem is well-posed with respect to almost sure weak convergence. The results of this chapter have been published in Schwarz and Van Bellegem (2010).

Chapter 2
Frontier estimation from noisy observations

In the context of a frontier model, the purpose of this chapter is the development of a consistent estimator of the frontier function φ when the input variable is observed with a centered normal measurement error. Reviewing the literature, we show that classical frontier estimators fail to be consistent in this situation. By solving first the underlying deconvolution problem using the techniques from the first chapter, we are able to define a new robust frontier estimator and to prove a sufficient condition for its consistency. The results of this chapter have been published in Schwarz et al. (2011).

Chapter 3
Adaptive circular density deconvolution

In the circular density deconvolution model, we suppose that in addition to the iid. sample of size n from $Y = X + \varepsilon$, an independent iid. sample of size m from the error ε is available. First, we develop the minimax theory for this problem by deriving a lower bound for the maximal risk over certain density classes $\mathcal{F}$ and $\mathcal{E}$ of possible solutions and error densities, respectively. This lower bound consists of two terms. Each of these terms depends on one of the two sample sizes but not on the other. In this way, the influence of the sample sizes on the difficulty of the estimation problem is well characterized.

The circular structure of the model allows for the representation of f as a discrete series. Therefore, it is natural to regularize the underlying inverse problem by projection. Thus, we define an orthogonal series estimator $\widehat{f}_k$ and show that for an appropriate choice $k = k_n^*$, it is minimax optimal over a wide range of classes $\mathcal{F}$ and $\mathcal{E}$. However, k_n^* depends on characteristics of the classes $\mathcal{F}$ and $\mathcal{E}$.

The main contribution of this chapter is the development of a fully data-driven choice $\widehat{k}$ of the regularization parameter k that mimics the behavior of k_n^*. This is done using the *model selection* approach developed in Barron et al. (1999). We derive an upper risk bound for the adaptive estimator $\widehat{f}_{\widehat{k}}$ and show that it is minimax optimal over a wide range of density classes $\mathcal{F}$ and $\mathcal{E}$ including in particular classical smoothness classes. The results of this chapter are also available in Johannes and Schwarz (2009).

Chapter 4
Adaptive non parametric instrumental regression

We derive the minimax rate for the non parametric instrumental regression problem with respect to certain classes $\mathcal{F}$ and $\mathcal{T}$ of regression functions and conditional expectation operators, respectively. The operator is supposed to be compact but otherwise unknown. We define an estimator of the regression function based on projection and additional thresholding and show that it is minimax optimal, provided the dimension of the projection space is chosen optimally. We are confronted with the difficulty that the optimal choice of this dimension depends on characteristics of the classes.

This problem is solved by defining a fully data driven choice of k following the model selection approach. In order to do so, we need to assume, however, that the eigenbasis of the conditional expectation operator is known. We show that the adaptive estimator is minimax optimal over classical smoothness classes.

Finally, we briefly sketch an approach to how the risk of the adaptive estimator could possibly be controlled for a completely unknown operator, that is without assuming its eigenfunctions to be known. The results of this chapter are also available in Johannes and Schwarz (2010).

Résumé The most important contributions of this thesis are the time-domain identification condition for the density in a deconvolution problem on the real line and its consistent estimation under weakest assumptions in Chapter 1; the identification and consistent estimation of a stochastic frontier in the presence of noise in the data in Chapter 2; a minimax theory for the circular density deconvolution problem taking into account the two sample sizes, and an adaptive estimator which is minimax optimal with respect to a wide range of classical smoothness classes (Chapter 3); finally, a minimax theory for the non parametric instrumental regression model and, in the case where the eigenfunctions of the operator are known, an adaptive estimator attaining minimax optimal rates of convergence under classical smoothness assumptions.

Chapter 1

Consistent deconvolution on the real line under partially known error distribution

A classical problem in non parametric statistics is the consistent estimation of the distribution of some real random variable X based on a statistical sample which is subject to an independent additive measurement error ε. Formally, one usually assumes iid. observations from a random variable $Y = X + \varepsilon$. In the case where the cumulative distribution function (cdf) of ε is known, a vast literature focuses on the accuracy of estimation of the cdf of X (e.g. Carroll and Hall, 1988; Fan, 1991). The full knowledge of the cdf of ε is a strong assumption that one is rarely able to impose in real data analysis.

In order to deal with uncertainty about the distribution of ε, two approaches suggest themselves: Firstly, one could consider different sampling processes providing empirical information about the unknown distribution. This would allow for a preliminary estimation of the error distribution before proceeding to the estimation of the target density itself. Secondly, one may restrict the sets of possible distributions of X and ε in a way that allows for identifiability even in the case where only the contaminated sample is observed. In this chapter, we proceed according to the second approach.

This chapter is organized as follows: We begin with a literature review on the density deconvolution problem. First, we consider the case of a known error distribution, then the case with incertainty about this distribution. In particular, we discuss the two aforementioned approaches to dealing with this uncertainty. At the end of the review, we outline our own approach, comparing it to methods available in the literature. In Section 1.2, we address the

identification issue and in Section 1.3 we define a minimum contrast estimator and prove its consistency. This chapter is based mainly on Schwarz and Van Bellegem (2010).

1.1 A review on density estimation from noisy observations

1.1.1 The case of a known error distribution

Assume that we have a sample of iid. copies or the variable Y in the deconvolution model $Y = X + \varepsilon$ with independent error. Many research papers focus on the accurate estimation of the cumulative distribution function (cdf) of X under the assumption that the cdf of ε is known. The independent additive measurement error implies that if we assume that densities exist, then the density of Y is the convolution of the density of ε with the one of X:

$$f^Y(y) = (f^\varepsilon * f^X)(y) := \int_{-\infty}^{\infty} f^\varepsilon(y-s) f^X(s) \mathrm{d}s \; .$$

A key observation for solving the deconvolution problem, that is the problem of reconstructing f^X, is the following well-known result which can be found in Walker (1988), for example.

Theorem 1.1 (Convolution Theorem) *Let f and g be two probability densities defined on $\mathbb{R}$. If $\mathcal{F}[f](\,\cdot\,) = \int_{\mathbb{R}} f(t) \exp(-2\pi i t \,\cdot\,) \mathrm{d}t$ denotes the Fourier transform of f, then we have*

$$\mathcal{F}[f * g] = \mathcal{F}[f]\, \mathcal{F}[g].$$

Similarly, if X and Y are independent random variables and ψ^X and ψ^Y their characteristic functions, then $\psi^{X+Y} = \psi^X \, \psi^Y$.

In view of this result, most estimators of f^X studied in the literature use the characteristic function of the involved random variables. Such methods are said to work *in the Fourier domain.* Hypotheses on the characteristic functions or on the Fourier transforms of the densities are called *assumptions in the Fourier domain.* A natural estimator of ψ^Y is the empirical characteristic function based on the sample $(Y_1, \ldots, Y_n)$, i.e.

$$\widehat{\psi}^Y(t) := \frac{1}{n} \sum_{k=1}^{n} \exp(itY_k), \qquad t \in \mathbb{R}.$$

Stefanski and Carroll (1990) remark that in order to estimate f^X, it is not sufficient, however, to consider the inverse Fourier transform of $(\widehat{\psi}^Y / \psi^\varepsilon)$, which need not even exist due to the decay of ψ^ε. Instead, they consider a kernel

estimator $\widehat{g}_K$ of g for some kernel K. The characteristic function of $\widehat{g}_K$ writes $\widehat{\psi_{g_K}} = \widehat{\psi}^Y \psi_K$, where ψ_K denotes the Fourier transform of K. The authors then propose the inverse Fourier transform of $(\widehat{\psi_{g_K}}/\psi^\varepsilon)$. Under some integrability conditions on K, this estimator is consistent.

Estimators following similar construction principles have been studied by Carroll and Hall (1988) and Fan (1991), for example. Alternative estimators using wavelets, for instance, have been developed by Pensky and Vidakovic (1999); Johnstone et al. (2004); Bigot and Van Bellegem (2009).

The exact knowledge of the characteristic function and hence the distribution of the error is however not realistic in many empirical studies. If we want to relax the condition that the distribution of the error is known, one major obstacle is that the distribution of X is no longer identifiable. At least three approaches to overcome this problem can be found in the literature.

1.1.2 The case of an unknown error distribution

A first approach assumes that an independent sample from the measurement error ε is available in addition to the sample of Y. From the independent observation of ε, the density f^ε is identified and consequently, the target density f^X is identified as well. Based on the sample of the ε's, a non parametric estimator of f^ε can be constructed and then be used in the construction of the estimator of f^X (Neumann, 1997; Johannes and Schwarz, 2009; Johannes et al., 2011). If this approach may be realistic in a number of practical situations (e.g. some problems in biostatistics or astrophysics), it is hardly applicable in the context of production frontier estimation, for example, which we are going to consider in the next chapter.

Similarly, one can consider various sampling processes which allow for identification. Li and Vuong (1998) suppose that repeated measurements for one single value of X are available, such as $Y_j = X + \varepsilon_j$ for $j = 1, 2$. Assuming further that X, ε_1, and ε_2 are independent, $\mathbf{E}[\varepsilon_j] = 0$, and that the characteristic functions of X and ε are non-zero everywhere, they show how these characteristic functions can be expressed as functions of the joint characteristic function of (Y_1, Y_2). From this representation it follows that the distributions of both X and ε can be identified from the observation of the couple (Y_1, Y_2). The joint characteristic function of (Y_1, Y_2) can be estimated from a sample of (Y_1, Y_2) and then used to derive an estimator of f^X. The characteristic functions of X and ε, denoted by ψ^X and ψ^ε, can then be computed using the above-mentioned representation. Delaigle et al. (2008) have considered this setting and present modified kernel estimators which, if the number of repeated measurements is large enough, can perform as well as they would under known error distribution.

The assumption of repeated measurements of X in a multilevel model provides a similar setting. Neumann (2007) supposes that $Y_{ij} = X_i + \varepsilon_{ij}$ are

observed for $j = 1, \dots, N$ and $i = 1, \dots, n$ (see also Meister et al., 2010). In this sampling process, the identification of the cdf of X is ensured by a condition on the zero-sets of the characteristic functions of X and ε. Let $\mathcal{Y} = (Y_{i1}, \dots, Y_{iN})'$, $\psi^{\mathcal{Y}}$ its characteristic function, and $\widehat{\psi^{\mathcal{Y}}}$ the empirical characteristic function of $\mathcal{Y}$. A consistent estimator of the density of X is obtained by minimizing the discrepancy

$$\int_{\mathbb{R}^n} \left|\psi^X(t_1 + \cdots + t_n)\psi^\varepsilon(t_1)\cdots\psi^\varepsilon(t_n) - \hat{\psi}_n^{\mathcal{Y}}(t_1, \dots, t_n)\right| K(t_1, \dots, t_n)\mathrm{d}t_1 \dots \mathrm{d}t_n$$

over certain classes of possible characteristic functions ψ^X and ψ^ε of X and ε, respectively. The function K is some positive kernel ensuring the existence of the integral. Repeated measurements of multilevel sampling arise in some economic situations, for instance when production units are observed over time (Park et al., 2003; Daskovska et al., 2010).

A second approach to recover the identification of X in spite of the noise ε consists in assuming that the distribution of ε is only partially unknown. A realistic case for practical purposes is to assume that ε is normally distributed, but the variance of ε is unknown. The distribution of X is not identified in general under such assumptions, and it is necessary to restrict the class of possible distributions in order to recover identification.

Several recent research papers have proposed identification restrictions on the class of X given a partial knowledge about the cdf of the noise. Butucea and Matias (2005) assume that the error density is «s-exponential», meaning that its Fourier transform, ψ^ε, satisfies

$$b\exp(-|u|^s) \leqslant |\psi^\varepsilon(u)| \leqslant B\exp(-|u|^s)$$

for some constants b, B, s, and for $|u|$ large enough. In their approach the error density is supposed to be known up to its scale σ (called *noise level*). As for the density f^X, both polynomial and exponential decay of its Fourier transform are shown to lead to a fully identified model. For $\tau > 0$, let ψ_τ^ε denote the Fourier transform of (τf^ε). The key to the estimation of σ is the observation that the function $|F(\tau, u)| = |\psi^Y(u)|/|\psi_\tau^\varepsilon(u)|$ diverges as $u \to \infty$ if $\tau > \sigma$ and that it converges to 0 otherwise. Let $\widehat{F}(\tau, u) = |\widehat{\psi}^Y(u)|/|\psi_\tau^\varepsilon(u)|$. Then Butucea and Matias (2005) show that

$$\widehat{\sigma}_n := \inf\{\tau > 0 \mid |\hat{F}(\tau, u_n)| \geqslant 1\}$$

yields a consistent estimator of σ for some well balanced sequence $(u_n)_{n\geqslant 1}$. This estimator is then used to deconvolve the empirical density of Y and to obtain an estimator of the density of X. Some extensions are proposed in Butucea et al. (2008), where the error density is assumed to have a stable symmetric distribution with $\psi^\varepsilon(u) = \exp(-|\gamma u|^s)$ in which γ represents some known scale parameter and s is an unknown parameter, called the self-similarity index.

A similar setting is considered in Meister (2006). In this paper, the error is supposed to be normally distributed with an unknown variance parameter. Identification is recovered by assuming that ψ^X lies in

$$\{\psi \mid c_1|u|^{-\beta} \leqslant |\psi(u)| \leqslant c_2|u|^{-\beta} \text{ for all } u \gg 0\} \tag{1.1}$$

for some strictly positive constants c_1, c_2. In Meister (2007), the author assumes that ψ^ε is known on some arbitrarily small interval $[-\nu, \nu]$ and that f^ε belongs to some class

$$\mathcal{G}_{\mu,\nu} = \{f \text{ is a density such that } \|f\|_\infty \leqslant C, |\mathcal{F}[f](t)| \geqslant \mu \quad \forall |t| \geqslant \nu\}.$$

The target density f^X is assumed to belong to

$$\mathcal{F}_{S,C,\beta} = \left\{ f \text{ a density} \;\middle|\; \int_{-S}^{S} f(u)\mathrm{d}u = 1 \text{ and } \int |\mathcal{F}[f](t)|^2(1+t^2)^\beta \mathrm{d}t \leqslant C \right\},$$

that is in the class of densities with compact support that are uniformly bounded in the Sobolev norm. Empirically the direct access to ψ^X via Fourier deconvolution is restricted to the interval $[-\nu, \nu]$. However, it is shown using a Taylor expansion that ψ^X is uniquely determined by its restriction to $[-\nu, \nu]$, and therefore is everywhere identified.

We conclude this review mentioning the work by Horrace and Parmeter (2011) who consider a *composed error model,* where the error ε in a regression equation $Y = X\beta + \varepsilon$ can on its part be decomposed into $\varepsilon = U + V$, where U is the original error with unknown distribution and V a centered normally distributed additional error with unknown variance σ^2, independent of U. The aim in this situation is to reconstruct the density of U based on a sample from (X, Y). Identification is ensured by assuming that the characteristic function ψ^U of U has polynomial decay, that is, it lies in the class defined in (1.1).

Note that these models and approaches require either additional observations – which we do not assume in our model – or assumptions in the Fourier domain. Such assumptions have the disadvantage of not being easily interpretable in an heuristic and illustrative way. In fact, they may be hard to verify in practice. Let us now briefly outline our strategy in this chapter.

1.1.3 The approach of this chapter

In the present chapter our goal consists in identifying and consistently estimating the distribution of X using the noisy Y-sample only, and this under as few restrictions as possible on the class of possible cdfs of X. Our strategy in this setting is the assumption of partial knowledge about the error distribution: we assume that ε is normally distributed with mean zero and unknown

variance σ^2. The cdf of X is of course still not identified from the observation of Y. Identification is achieved by restricting the class of possible cdfs of X. However, in Remark 1.7 we show by giving a counter-example that this restriction is not sufficient in order to ensure the consistency of the minimum distance estimator we will be considering. Consistency is obtained restricting the density class further. Thus, the procedure in this chapter illustrates well how we can obtain additional properties (identification, consistency) by gradually adding hypotheses. The next step beyond the scope of this chapter could be the derivation of convergence rates. It seems to us that to this end another refinement of the hypotheses would be necessary.

In Hall and Simar (2002), a similar setting is considered, but under the additional assumption that σ^2 depends on the sample size n in such a way that σ^2 tends to zero as n tends to infinity. Matias (2002) and Butucea and Matias (2005) also consider the consistent estimation of σ^2 and of the cdf of X, but under strong restrictions on the characteristic function of X. Both of these assumptions have a drawback: As far as the noise level is concerned, it is not clear in which situations it tends to zero with increasing sample size. And as for assumptions in the Fourier domain, they are rather hard to verify in practice.

In Section 1.2, we define the above-mentioned class of probability distributions within which the cdf of X is identified from a Y-sample. This class is characterized by a simple condition in the time domain. Restricting this class slightly further, we prove the consistency of a minimum penalized contrast estimator of the cdf of X and of the variance σ^2 in Section 1.3. We illustrate by a counterexample that this restriction is indeed necessary in order to obtain a consistent estimator. The estimation procedure presented in this chapter is inspired by a similar estimator suggested by Neumann (2007) in the context of panel data. However, Neumann uses an identification condition in the Fourier domain which to avoid is our purpose.

The results in this chapter show that in the deconvolution model with partially known error distribution, identification and consistency can be obtained under fairly weak and easy to interpret assumptions in the time domain.

1.2 Identification

Suppose we want to recover the probability distribution P^X of a random variable X that is observed with an additive and independent random contamination error ε. One might argue that in practical settings, the independence assumption can hardly be verified. We discuss a possible relaxation of this assumption in Section 1.4 on further research. The measurement error is assumed to be normally distributed with mean zero and unknown variance σ^2. The resulting observational model is

$$Y = X + \varepsilon. \tag{1.2}$$

The distribution of Y is the convolution $P^Y = P^X * \mathcal{N}_\sigma$, where $\mathcal{N}_\sigma$ denotes the centered normal distribution with variance σ^2. Writing ψ^X, ψ^Y and ψ_σ for the characteristic functions of P^X, P^Y and $\mathcal{N}_\sigma$, respectively, the convolution equation is equivalent to $\psi^Y = \psi^X \psi_\sigma$ by virtue of the convolution Theorem 1.1. Because of the uncertainty with regard to the variance of the measurement error, not all probability distributions can be recovered from the model. Define the set of distributions

$$\mathcal{P}_0 = \left\{ P \in \mathcal{P} \;\middle|\; \exists\, A \in \mathcal{B}(\mathbb{R}) \,:\, |A| > 0 \;\wedge\; P(A) = 0 \right\},$$

where $\mathcal{B}(\mathbb{R})$ denotes the set of Borel sets in $\mathbb{R}$ and $\mathcal{P}$ the set of all probability distributions on $\mathbb{R}$, and $|A|$ the Lebesgue measure of A. In the following theorem, we show that all distributions belonging to $\mathcal{P}_0$ are identifiable from the observational model.

Theorem 1.2 (Identification) *The model defined by (1.2) is identifiable for the parameter space $\mathcal{P}_0 \times (0, \infty)$, that is, for any two probability measures $P^1, P^2 \in \mathcal{P}_0$ and $\sigma_1, \sigma_2 > 0$, we have that $P^1 * \mathcal{N}_{\sigma_1} = P^2 * \mathcal{N}_{\sigma_2}$ implies $P^1 = P^2$ and $\sigma_1 = \sigma_2$.*

In contrast to the assumptions made in Butucea and Matias (2005), the hypothesis that the solution P^X lies in $\mathcal{P}_0$ does not involve the characteristic function of X. The assumption is natural insofar as it has an obvious interpretation: there has to be some interval in which X does almost surely not fall. This interval may be arbitrarily small and located anywhere on the real axis. Note that we do not need to know neither the length nor the location of this interval in order that P^X be identifiable. However, we will see below that more information is needed in order to be able to estimate P^X consistently.

The proof of the identification theorem is based on the following lemma.

Lemma 1.3 *Let P^1 and P^2 be probability distributions and $0 < \sigma_1 < \sigma_2$. Then,*

$$P^1 * \mathcal{N}_{\sigma_1} = P^2 * \mathcal{N}_{\sigma_2} \implies P^1 = P^2 * \mathcal{N}_{\sigma_3}, \qquad \text{where } \sigma_3 = \sqrt{\sigma_2^2 - \sigma_1^2}.$$

Proof. First, apply the convolution theorem on both sides of the equation, then divide by ψ_{σ_1} which is positive everywhere. To conclude, it suffices to remark that $\psi_{\sigma_3} = (\psi_{\sigma_2} / \psi_{\sigma_1})$. □

Proof of Theorem 1.2. Suppose that $(P^1, \sigma_1), (P^2, \sigma_2) \in \mathcal{P}_0 \times (0, \infty)$. We have to show that

$$P^1 * \mathcal{N}_{\sigma_1} = P^2 * \mathcal{N}_{\sigma_2} \implies (P^1, \sigma_1) = (P^2, \sigma_2).$$

First, we prove by contradiction that $\sigma_1 = \sigma_2$. Suppose that $\sigma_1 \neq \sigma_2$. Without loss of generality, say $\sigma_1 < \sigma_2$. By virtue of Lemma 1.3, this implies $P^1 = P^2 * \mathcal{N}_{\sigma_3}$.

We show now that this is only possible if P^1 is not in $\mathcal{P}_0$ which contradicts the assumption. Indeed, let $A = [a_1, a_2]$ be some interval of positive length $|A| = a_2 - a_1$ and $B = [b_1, b_2]$ another interval with $|B| < |A|$ and $P^2(B) > 0$. By definition of the convolution and in view of the independence of X and ε in our model, we can write

$$(P^2 * \mathcal{N}_{\sigma_3})(A) = (P^2 \otimes \mathcal{N}_{\sigma_3})(S_A), \text{ where } S_A = \{(x, y) \in \mathbb{R}^2 \mid x + y \in A\}$$

and $\otimes$ denotes the product measure. We have that $a_1 - b_1 < a_2 - b_2$ because of $|B| < |A|$. It is easily verified that $B \times [a_1 - b_1, a_2 - b_2] \subset S_A$ and hence

$$P^1(A) = (P^2 * \mathcal{N}_{\sigma_3})(A) \geqslant P^2(B)\, \mathcal{N}_{\sigma_3}([a_1 - b_1, a_2 - b_2]) > 0.$$

This contradicts the assumption that $P^1 \in \mathcal{P}_0$, showing that $\sigma_1 = \sigma_2$. The characteristic function of the normal distribution being positive everywhere, an application of the convolution theorem completes the proof. □

Remark 1.4 The identification theorem assumes that the measurement error ε is normally distributed with an unknown variance σ^2. Although this is a most natural assumption from a practical point of view, it should be noticed that the proof essentially exploits the infiniteness of the support of ε. Therefore, the identification result may be extended to other scale families of error distributions, provided some counterpart of Lemma 1.3 holds. This is the case for Cauchy distributions with location parameter $\mu = 0$ or, more generally, for stable distributions with fixed exponent $\alpha \in (0, 2]$, skewness parameter $\beta = 0$, and location $\mu = 0$, for example.

Note that one could also exchange the roles of X and ε in the model: The distribution of the error would have to lie in the class $\mathcal{P}_0$ (e.g. a bounded error) and the distribution of X would belong to one of the parametric classes described above. By symmetry, the distribution of ε and the parameter of the distribution of X would be identified and the estimation techniques developed below would apply. □

It is worth noticing that if in Theorem 1.2 we do not suppose both P^1 and P^2 to belong to $\mathcal{P}_0$, the conclusion is false in general as the following counterexample illustrates. Let P^1 be the uniform distribution on the unit interval and ψ^1 its characteristic function. Clearly, $P^1 \in \mathcal{P}_0$. If we let P^2 be the probability distribution with characteristic function $\psi^2 := \psi^1\, \psi_\sigma / \psi_{(\sigma/2)}$, then, in view of the convolution theorem, we have $P^1 * \mathcal{N}_\sigma = P^2 * \mathcal{N}_{(\sigma/2)}$, but $P^1 \neq P^2$.

We conclude this section by remarking that the probability distributions in question are not required to have densities. For those having one, the identification condition can be equivalently expressed by requiring that the density has to vanish on a set of positive Lebesgue measure.

1.3 Consistent estimation

Now suppose we have an iid. sample $\{Y_1, \dots, Y_n\}$ from the model (1.2). Let $\xrightarrow{\mathcal{D}}$ denote convergence in distribution. An estimator $(\widehat{P}_n^X, \widehat{\sigma}_n)$ of (P^X, σ) is called *consistent* if, almost surely, $\widehat{P}_n^X \xrightarrow{\mathcal{D}} P^X$ and $\widehat{\sigma}_n \to \sigma$ as $n \to \infty$. For a consistent estimator, we always have $\widehat{P}_n^X * \mathcal{N}_{\widehat{\sigma}_n} \xrightarrow{\mathcal{D}} P^Y$, which is hence a necessary condition of consistency. We call an estimator satisfying this condition *admissible*.

1.3.1 Minimum distance estimation

In this section we develop a minimum distance estimator which is inspired by a method proposed in Neumann (1997). Let $\widehat{\psi}_n^Y(t) = \frac{1}{n}\sum_{k=1}^n \exp(itY_k)$ be the empirical characteristic function of the observations. By the Glivenko-Cantelli theorem, it converges almost surely uniformly to the true characteristic function ψ^Y. For characteristic functions $\widetilde{\psi}^X, \widetilde{\psi}_\sigma$, and $\widetilde{\psi}^Y$ let us define a distance ρ,

$$\rho(\widetilde{\psi}^X, \widetilde{\psi}_\sigma; \widetilde{\psi}^Y) := \int_{\mathbb{R}} |\widetilde{\psi}^X(t)\, \widetilde{\psi}_\sigma(t) - \widetilde{\psi}^Y(t)|\, K(t)\, \mathrm{d}t, \tag{1.3}$$

where K is some continuous and strictly positive probability density ensuring the existence of the integral. Under slight abuse of notation we do not make the dependence of the distance on K explicit, as it does not have any influence on the results derived in this work. The estimation consists in choosing $\widehat{P}_n^X$ and $\mathcal{N}_{\widehat{\sigma}_n}$ such that their characteristic functions $\widehat{\psi}_n^X$ and $\psi_{\widehat{\sigma}_n}$ minimize the distance $\rho(\cdot, \cdot; \widehat{\psi}_n^Y)$.

Definition 1.5 *Let $(\delta_n)_{n\in\mathbb{N}}$ be a vanishing sequence of positive real numbers and $\mathcal{C}$ a set of probability distributions. We call a random sequence $(\widehat{P}_n^X, \widehat{\sigma}_n)$ depending on the observations $\{Y_1, \dots, Y_n\}$ a* minimum distance estimator on $\mathcal{C}$ *if it is such that the corresponding characteristic functions yield*

$$\rho(\widehat{\psi}_n^X, \psi_{\widehat{\sigma}_n}; \widehat{\psi}_n^Y) \leqslant \inf_{\substack{\widetilde{\psi}^X \in \Psi_\mathcal{C} \\ \widetilde{\sigma} \geqslant 0}} \rho(\widetilde{\psi}^X, \psi_{\widetilde{\sigma}}; \widehat{\psi}_n^Y) + \delta_n, \tag{1.4}$$

where we denote by $\Psi_\mathcal{C}$ the set of the characteristic functions of all the distributions in $\mathcal{C}$. Let further $\widehat{P}_n^{X+\varepsilon} := \widehat{P}_n^X * \mathcal{N}_{\widehat{\sigma}_n}$, the characteristic function of which is $\widehat{\psi}_n^X\, \psi_{\widehat{\sigma}_n}$.

Our aim is to prove the consistency of this estimator. Obviously, this requires further assumptions on the class $\mathcal{C}$. In the first instance, we show that the minimum distance estimator is always admissible.

Lemma 1.6 (Admissibility) *Every minimum distance estimator $(\widehat{P}_n^X, \widehat{\sigma}_n)$ on the set $\mathcal{P}$ of all probability distributions is admissible.*

Proof. The empirical characteristic function $\widehat{\psi}_n^Y$ converges almost surely pointwise to ψ^Y. By Lebesgue's Theorem, this implies $\rho(\psi^X, \psi_\sigma; \widehat{\psi}_n^Y) \to 0$ almost surely. Applying the triangle inequality and using $\psi^X \psi_\sigma = \psi^Y$, we obtain

$$\rho(\widehat{\psi}_n^X, \psi_{\widehat{\sigma}_n}; \psi^Y) \leqslant \rho(\widehat{\psi}_n^X, \psi_{\widehat{\sigma}_n}; \widehat{\psi}_n^Y) + \rho(\psi^X, \psi_\sigma; \widehat{\psi}_n^Y).$$

Because of (1.4), we can write $\rho(\widehat{\psi}_n^X, \psi_{\widehat{\sigma}_n}; \widehat{\psi}_n^Y) \leqslant \rho(\psi^X, \psi_\sigma; \widehat{\psi}_n^Y) + \delta_n$, so that we conclude that, almost surely,

$$\rho(\widehat{\psi}_n^X, \psi_{\widehat{\sigma}_n}; \psi^Y) \leqslant 2\rho(\psi^X, \psi_\sigma; \widehat{\psi}_n^Y) + \delta_n \to 0.$$

We choose an element $\omega \in \Omega$ of the underlying probability space such that this convergence holds. As the integrand in (1.3) is non-negative, it follows that

$$\int_0^a \widehat{\psi}_n^X(t) \psi_{\widehat{\sigma}_n}(t)\, \mathrm{d}t \longrightarrow \int_0^a \psi^Y(t)\, \mathrm{d}t$$

for all $a \in \mathbb{R}$ as $n \to \infty$. We have shown that the integrated characteristic functions of the measures $\widehat{P}_n^{X+\varepsilon}$ converge to the integrated characteristic function of the probability measure P^Y. Theorem 6.3.3 from Chung (1968) states that in this case, we have $\widehat{P}_n^{X+\varepsilon} \xrightarrow{\mathcal{V}} P^Y$, where $\xrightarrow{\mathcal{V}}$ denotes vague convergence. Recall that a sequence of probability distributions Q_n on $\mathbb{R}$ is said to converge *vaguely* to a limiting sub-probability measure Q if $\int g\, \mathrm{d}Q_n$ converges to $\int g\, \mathrm{d}Q$ as $n \to \infty$ for all functions $g : \mathbb{R} \to \mathbb{R}$ with compact support. Note that the limit measure may have a total mass strictly smaller than 1. In our situation, however, the vague limit P^Y is a probability distribution. The Portmanteau Theorem states that that in this case we have indeed weak convergence, which means that the estimator is admissible. □

Remark 1.7 We have seen that the minimum distance estimator is always admissible. Next, we determine classes of distributions on which it is also consistent. One might wonder if the identification condition alone is sufficient to guarantee consistency, that is, if minimum distance estimators on $\mathcal{P}_0$ are consistent. This is not the case, as the following counterexample illustrates. Let

$$\widehat{P}_n^X(A) := \widehat{P}_n^Y(A \cap [-n, n]) / \widehat{P}_n^Y([-n, n])$$

for any Borel set A, where $\widehat{P}_n^Y$ is any consistent estimator of P^Y, and denote by $\widehat{\psi}_n^X$ its characteristic function. Note that $\widehat{P}_n^X \in \mathcal{P}_0$ for every $n \geqslant 1$, and let further $\widehat{\sigma}_n := n^{-1}$. Then, $(\widehat{P}_n^X, \widehat{\sigma}_n)$ is a minimum distance estimator. Indeed, $\widehat{P}_n^X$ converges to P^Y in distribution almost surely by construction. Using

Lévy's continuity theorem, we deduce that $|\widehat{\psi}_n^X(t)\psi_{\widehat{\sigma}_n}(t) - \widehat{\psi}_n^Y(t)| \to 0$ for all t almost surely and hence $\rho(\widehat{\psi}_n^X, \psi_{\widehat{\sigma}_n}; \widehat{\psi}_n^Y) \to 0$ almost surely by Lebesgue's dominated convergence theorem. Thus, the sequence $(\widehat{P}_n^X, \widehat{\sigma}_n)$ is a candidate for a minimum distance estimator when δ_n decreases sufficiently slow. It is easily verified that this estimator is admissible but not consistent. The following consideration shows in which way we have to restrict the class $\mathcal{P}_0$ in order to obtain consistency. Assume that $P^X \in \mathcal{P}_0$ and let $(\widehat{P}_n^X, \widehat{\sigma}_n)$ be an admissible estimator. By virtue of Lemma 1.10 below, admissibility implies the existence of an increasing sequence $(n_k)_{k\in\mathbb{N}}$, some probability measure P_∞^X, and $\sigma_\infty \geqslant 0$ such that

$$\widehat{P}_{n_k}^X \xrightarrow{\mathcal{D}} P_\infty^X \quad \text{and} \quad \widehat{\sigma}_{n_k} \longrightarrow \sigma_\infty \tag{1.5}$$

as $n \to \infty$, which implies $\widehat{P}_{n_k}^X * \mathcal{N}_{\widehat{\sigma}_{n_k}} \xrightarrow{\mathcal{D}} P_\infty^X * \mathcal{N}_{\sigma_\infty}$, and hence, due to admissibility and by uniqueness of the weak limit,

$$P_\infty^X * \mathcal{N}_{\sigma_\infty} = P^X * \mathcal{N}_\sigma. \tag{1.6}$$

It follows from (1.5) that a necessary condition for $\widehat{P}_n^X$ to be consistent is $P_\infty^X = P^X$. In view of (1.6) and Theorem 1.2, this is equivalent to $P_\infty^X \in \mathcal{P}_0$. But this may not be the case in spite of all $\widehat{P}_{n_k}^X$ lying in $\mathcal{P}_0$ as this class is not closed under convergence in distribution as the above counterexample shows.□

1.3.2 Consistency

In Remark 1.7 we have seen that in order to show consistency of the minimum distance estimator, we need to restrict the set of considered distributions further than to the identifiable set $\mathcal{P}_0$. For every $R, \eta > 0$, let

$$\mathcal{P}_R^\eta := \{P \in \mathcal{P} \mid \exists A = (a_1, a_2) \subset [-R, R] \ : \ |A| \geqslant \eta \ \wedge \ P(A) = 0\}.$$

This choice avoids the problem of possibly obtaining a sequence of estimators the weak limit of which lies outside the identified class $\mathcal{P}_0$. Indeed, the following lemma shows that the closure of $\mathcal{P}_R^\eta$ with respect to weak convergence is a subset of $\mathcal{P}_0$. This allows us to show in Theorem 1.9 that the minimum distance estimator is consistent when we restrict the class of solutions to $\mathcal{P}_R^\eta \subset \mathcal{P}_0$ for some $\eta, R > 0$. The class $\mathcal{P}_0$ imposes that the distribution P^X assign zero mass to some Borel set of positive Lebesgue measure. In addition, the class $\mathcal{P}_R^\eta$ involves knowledge about size and location of this Borel set. In general, such knowledge may not be available in practice. Note however that for a positive random variable X for example, we have $P^X \in \mathcal{P}_R^\eta$ for any choice of η and R.

Lemma 1.8 *For any $R, \eta > 0$, weakly convergent sequences in $\mathcal{P}_R^\eta$ have their limit in $\mathcal{P}_0$.*

Proof. Let $(P_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}_R^\eta$. Then, we have

$$\forall n \in \mathbb{N}\ \exists \text{ interval } A_n \subset [-R,R]\ :\ P_n(A_n) = 0 \wedge |A_n| \geqslant \eta. \tag{1.7}$$

Suppose further that $(P_n)_{n\in\mathbb{N}}$ converges in distribution to some P_∞. We have to show that there is some $A_\infty \in \mathcal{B}([-R,R])$ of positive Lebesgue measure such that $P_\infty(A_\infty) = 0$, that is $P_\infty \in \mathcal{P}_0$. Firstly, we deduce from (1.7) that there exists an $x_0 \in [-R,R]$ which lies in infinitely many A_n, or in other words,

$$\exists x_0 \in [-R,R]\ \exists \text{ strictly monotone } (n_k)_{k\in\mathbb{N}}\ \forall k \in \mathbb{N}\ :\ x_0 \in A_{n_k}.$$

As all A_{n_k} are intervals of length at least η, there is an interval containing x_0 which is a null set for infinitely many measures of the sequence P_{n_k}. More precisely, there is a subsequence n'_k of n_k such that

$$(x_0 - \eta/2, x_0] \subset \bigcap_{k\in\mathbb{N}} A_{n'_k} \quad \text{or} \quad [x_0, x_0+\eta/2) \subset \bigcap_{k\in\mathbb{N}} A_{n'_k}.$$

Hence, we can choose $A_\infty = (x_0 - \eta/2, x_0)$ or $A_\infty = (x_0, x_0 + \eta/2)$ such that $|A_\infty| = \eta/2 > 0$ and $P_{n'_k}(A_\infty) = 0$ for all $k \in \mathbb{N}$. The latter assertion implies that $\liminf_{n\to\infty} P_n(A_\infty) = 0$. Recall that the P_n converge weakly to P_∞ and that A_∞ is an open set. Therefore, the Portmanteau theorem allows us to conclude that $P_\infty(A_\infty) = 0$. □

Before proving consistency, recall the definition of the Lévy distance: For probability distributions P^1, P^2 with cumulative distribution functions F^1, F^2, define

$$d(P^1,P^2) := \inf\{\delta > 0 \mid F^1(x-\delta) - \delta \leqslant F^2(x) \leqslant F^1(x+\delta) + \delta \quad \forall x \in \mathbb{R}\}. \tag{1.8}$$

For a sequence P_n of probability distributions, one has that $P_n \xrightarrow{\mathcal{D}} P$ if and only if $d(F_n, F) \to 0$ as $n \to \infty$. In other words, d metrizes the weak convergence. Now define, for probability distributions $\widetilde{P}^X$ and real numbers $\widetilde{\sigma}$,

$$\Delta(\widetilde{P}^X, \widetilde{\sigma}; P^X, \sigma) = d(\widetilde{P}^X, P^X) + |\widetilde{\sigma} - \sigma|.$$

Remark that $\Delta(P_n^X, \sigma_n; P^X, \sigma) \to 0$ if and only if $P_n^X \xrightarrow{\mathcal{D}} P^X$ and $\sigma_n \to \sigma$ (and hence $\mathcal{N}_{\sigma_n} \xrightarrow{\mathcal{D}} \mathcal{N}_\sigma$).

Theorem 1.9 (Consistency) *Let $R, \eta > 0$ and suppose that in the deconvolution model (1.2), we have $P^X \in \mathcal{P}_R^\eta$. Then, any minimum distance estimator $(\widehat{P}_n^X, \widehat{\sigma}_n)$ on $\mathcal{P}_R^\eta$ is consistent, that is, we have $\Delta(\widehat{P}_n^X, \widehat{\sigma}_n; P^X, \sigma) \to 0$ almost surely.*

Proof. We have seen in Lemma 1.6 that the estimator is admissible. Now we show that under the assumptions of this theorem, the admissibility already implies

$$\Delta(\widehat{P}_n^X, \widehat{\sigma}_n; P^X, \sigma) \to 0.$$

The proof is by contradiction. Assume there is a $\delta > 0$ and an increasing sequence $(n_k)_{k\in\mathbb{N}}$ in $\mathbb{N}$ such that $\Delta(\widehat{P}_{n_k}^X, \widehat{\sigma}_{n_k}; P^X, \sigma) \geqslant \delta \quad \forall\, k \in \mathbb{N}$. Lemma 1.10 furnishes a subsequence $(n'_k)_{k\in\mathbb{N}}$ of $(n_k)_{k\in\mathbb{N}}$, a probability measure P_∞^X and a constant $\sigma_\infty \geqslant 0$ such that

$$\widehat{P}_{n'_k}^X \xrightarrow{\mathcal{D}} P_\infty^X \qquad \text{and} \qquad \mathcal{N}_{\widehat{\sigma}_{n'_k}} \xrightarrow{\mathcal{D}} \mathcal{N}_{\sigma_\infty}. \tag{1.9}$$

Denote the characteristic functions of P_∞^X and $\mathcal{N}_{\sigma_\infty}$ by ψ_∞^X and ψ_{σ_∞}, respectively. Since weak convergence implies point-wise convergence of the corresponding characteristic functions, we obtain by Fatou's Lemma that

$$\rho(\psi_\infty^X, \psi_{\sigma_\infty}; \psi^Y) \leqslant \liminf_{k\to\infty} \rho(\widehat{\psi}_{n'_k}^X, \psi_{\widehat{\sigma}_{n'_k}}; \psi^Y) = 0,$$

that is, $\int_{\mathbb{R}} |\psi_\infty^X(t)\, \psi_{\sigma_\infty}(t) - \psi^Y(t)|\; K(t)\; dt = 0$. As K is strictly positive and characteristic functions are uniformly continuous on $\mathbb{R}$, we conclude that

$$\psi_\infty^X(t)\, \psi_{\sigma_\infty}(t) = \psi^Y(t) = \psi^X(t)\, \psi_\sigma(t) \qquad \forall\, t \in \mathbb{R}.$$

Lemma 1.8 allows us to deduce from (1.9) that $P_\infty^X \in \mathcal{P}_0$. Consequently, Theorem 1.2 ensures that $P_\infty^X = P^X$ and $\sigma_\infty = \sigma$. Together with (1.9), this implies $\Delta(\widehat{P}_{n'_k}^X, \widehat{\sigma}_{n'_k}; P^X, \sigma) \to 0$. This contradicts the assumption and thus completes the proof. □

To complete the arguments of this chapter, it remains to show the following technical lemma which we have used in Remark 1.7 and in the main result Theorem 1.9.

Lemma 1.10 *Let Q_n be a sequence of probability distributions and σ_n a sequence of positive real numbers. Suppose further that $(Q_n * \mathcal{N}_{\sigma_n})_{n\in\mathbb{N}}$ converges weakly to some probability distribution. Then, there exist an increasing sequence $(n_k)_{k\in\mathbb{N}}$, a probability distribution Q_∞, and a constant $\sigma_\infty \geqslant 0$ such that*

$$Q_{n_k} \xrightarrow{\mathcal{D}} Q_\infty \quad \textit{and} \quad \mathcal{N}_{\sigma_{n_k}} \xrightarrow{\mathcal{D}} \mathcal{N}_{\sigma_\infty}$$

as $n \to \infty$, where $\mathcal{N}_0 := \delta_0$ denotes the Dirac measure by convention.

Proof. By Helly's selection theorem, there is a subsequence $(n_k)_{k\in\mathbb{N}}$ and a sub-probability measure Q_∞ such that $Q_{n_k} \xrightarrow{\mathcal{V}} Q_\infty$, where $\xrightarrow{\mathcal{V}}$ denotes vague convergence (e.g. Chung (1968)). We show below that the σ_{n_k} are bounded

from above such that they have a convergent subsequence; without loss of generality, say $\sigma_{n_k} \to \sigma_\infty$ for some $\sigma_\infty \geqslant 0$. Proposition 3.1 from Jain and Orey (1979) states that if $R_n \xrightarrow{\mathcal{V}} R$ and $S_n \xrightarrow{\mathcal{D}} S$, then $R_n * S_n \xrightarrow{\mathcal{V}} R * S$, so we have $Q_{n_k} * \mathcal{N}_{\sigma_{n_k}} \xrightarrow{\mathcal{V}} Q_\infty * \mathcal{N}_{\sigma_\infty}$. By assumption, the same sequence converges weakly, and hence vaguely, to some distribution, so the uniqueness of the vague limit of measures on locally compact spaces implies $Q_\infty(\mathbb{R}) = 1$ because $(\mu{*}\nu)(\mathbb{R}) = \mu(\mathbb{R})\,\nu(\mathbb{R})$ for any two finite measures μ and ν on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The Portmanteau Theorem then implies $Q_{n_k} \xrightarrow{\mathcal{D}} Q_\infty$, which was our claim.

It remains to prove that σ_{n_k} is bounded from above. We show that otherwise the sequence $(Q_{n_k} * \mathcal{N}_{\sigma_{n_k}})_{k\in\mathbb{N}}$ would not be tight, which contradicts its weak convergence. Random variable notation is more convenient for this argument, so let $U_k \sim Q_{n_k}$ and $V_k \sim \mathcal{N}_{\sigma_{n_k}}$ be i.i.d. random variables and $W_k := U_k + V_k$. We have to show the non-tightness of the distributions of $\{W_k\}_{k\in\mathbb{N}}$, that is

$$\exists\, \delta \in (0,1)\ \forall\, J > 0\ \exists\, k \in \mathbb{N}\ :\ \mathbf{P}\big[W_k \in [-J, J]\big] < 1 - \delta.$$

Fix $\delta = (1/12)$ and $J > 0$, and let $\mathcal{J} = [-J, J]$. Put $I_j^+ = [3jJ, (3j+1)J]$ and $I_j^- = [-(3j+2)J, -(3j-1)J]$, and let $\mathcal{I} := \biguplus_{j\geqslant 0} I_j^+$ be the disjoint union of the I_j^+. Because of the monotony of the normal density on $[0,\infty)$, we have $\mathbf{P}[V_k \in I_j^+] > (1/3)\mathbf{P}\big\{V_k \in [3jJ, 3(j+1)J)\big\}$. The disjoint union over $j \geqslant 0$ of the intervals on the right hand side of this inequality is $[0,\infty)$, and $\mathbf{P}[V_k \geqslant 0] = (1/2)$. Thus, we have $\mathbf{P}[V_k \in \mathcal{I}] > (1/6)$. We can now write $\mathbf{P}[W_k \in \mathcal{J}] < (5/6) + (1/6)\mathbf{P}[W_k \in \mathcal{J} \mid V_k \in \mathcal{I}]$, and it is sufficient to prove that the conditional probability appearing in this inequality is less than $(1/2)$ for some k.

It is easy to check that

$$\mathbf{P}[W_k \in \mathcal{J} \mid V_k \in \mathcal{I}] = \sum_{j=0}^{\infty} \mathbf{P}[W_k \in \mathcal{J} \mid V_k \in I_j^+]\,\mathbf{P}[V_k \in I_j^+ \mid V_k \in \mathcal{I}].$$

By construction, $V_k \in I_j^+$ and $W_k \in \mathcal{J}$ together imply $U_k \in I_j^-$. Using further the monotony of the normal density on $[0,\infty)$, we deduce that

$$\mathbf{P}[W_k \in \mathcal{J} \mid V_k \in \mathcal{I}] \leqslant 6\mathbf{P}[V_k \in I_0^+] \sum_{j=0}^{\infty} \mathbf{P}[U_k \in I_j^-].$$

As the I_j^- are pairwise disjoint, the sum is bounded from above by 1, and hence $\mathbf{P}[W_k \in \mathcal{J} \mid V_k \in \mathcal{I}] \leqslant 6\mathbf{P}[V_k \in I_0^+]$. If σ_{n_k} is unbounded, k can be chosen in such a way that the right hand side of this inequality is less than $(1/2)$, which completes the proof. □

1.4 Conclusion

In this chapter, we have considered the problem of density deconvolution from one single sample which is contaminated with some independent additive normal error whose variance is unknown. We have explained the identification problem that arises in this situation, and discussing several known methods of coping with this issue, we have noticed that most of them are based on identification conditions in the Fourier domain. Such conditions are often difficult to verify in practice.

Our purpose in this chapter was hence to propose a new type of assumption which ensures identification of the deconvolution density and allows for its consistent estimation without referring to its properties in the Fourier domain. Accordingly, we have first shown identification under a weak and easy to interpret condition in the time domain. One could be interested in different time-domain conditions ensuring identifiability. The key step in the identifiability section was the construction of a distribution class $\mathcal{P}_0$ such that the convolution of any distribution $P \in \mathcal{P}_0$ with a centered normal distribution does not belong to $\mathcal{P}_0$. Alternatives to the class defined in this chapter could possibly be defined by conditions involving the decay of the densities. This could be subject to further research.

The condition defined by $\mathcal{P}_0$, although ensuring identifiability, was not strong enough as to ensure the consistency of a minimum contrast estimator, as we have illustrated by a counterexample.

Sharpening the identification condition slightly, we have been able to show that consistency is indeed achieved. However, we do not have convergence rates for the estimator. Kneip et al. (2010) have derived asymptotic results under similar assumptions in the case of boundary estimation. Their interesting approach of a maximum penalized profile likelihood could be explored in the context of the model discussed here.

Another interesting question concerns the independence between X and the error ε. One could be interested in relaxing this hypothesis by allowing for some heteroscedasticity. As a first step, one could let depend the unknown noise level σ^2 on X in a two-level way:

$$\sigma^2 = \begin{cases} \sigma_+^2 & (X \geqslant \xi) \\ \sigma_-^2 & (X < \xi) \end{cases}$$

for some unknown $\xi \in \mathbb{R}$. In a second step, one could try to cope with step functions σ_x^2 in x. Obviously, the identification problem becomes more difficult when one dispenses with the independence assumption. Some computations in the two-level model have shown that the techniques of this chapter might be generalized to this case in order to establish identification, although some technical difficulties remain. This seems an interesting problem worth investigating further.

Chapter 2

Consistent robust frontier estimation from noisy observations

The modelling and estimation of production functions have been the topic of many research papers on economic activity. The objects of study are *production units* to which one associates a vector of inputs (cost) $x \in \mathbb{R}_+^p$ and a vector of outputs (production) $y \in \mathbb{R}_+^q$. The set of *production possibilities* is denoted by Φ. It is a subset of $\mathbb{R}_+^{p+q}$ on which the inputs x can produce the outputs y. Following Shephard (1970), several assumptions are often imposed on Φ: convexity, free disposability, or strong disposability. Free disposability means that if (x, y) belongs to Φ and if x', y' are such that $x' \geqslant x$ and $y' \leqslant y$ then $(x', y') \in \Phi$. Strong disposability requires that one can always produce a smaller amount of outputs using the same inputs.

The boundary of the production set is of particular interest in the efficiency analysis of production units. The *efficiency frontier* in the input space is defined as follows. For all $y \in \mathbb{R}_+^q$, consider the set $\rho(y) = \{x \in \mathbb{R}_+^p \mid (x, y) \in \Phi\}$. The radial efficiency boundary is then given by

$$\varphi(y) = \left\{x \in \mathbb{R}_+^p \mid x \in \rho(y) \ \wedge \ \theta x \notin \rho(y) \ \ \forall\, \theta \in (0, 1)\right\}$$

for all y. Similarly, a frontier in the output space may be defined (Färe et al., 1985). In empirical studies, the attainable set Φ is unknown and has to be estimated from observed data. Suppose a random sample of production units

$$\mathcal{X}_n = \{(X_i, Y_i) \in \mathbb{R}_+^{p+q} \mid i = 1, \ldots, n\}$$

is observed. We assume that each unit (X_i, Y_i) is an independent replication of (X, Y) whose joint probability distribution on $\mathbb{R}_+^{p+q}$ describes the production process. The support of this probability measure coincides with Φ. Estimating the efficiency boundary amounts hence to estimating the support of (X, Y).

Non parametric methods are natrual to use in this context, because they do not require restrictive assumptions on the data generating process of $\mathcal{X}_n$. Deprins et al. (1984) have introduced the Free Disposal Hull (FDH) estimator

$$\widehat{\Phi}_{\mathrm{FDH}} = \left\{(x, y) \in \mathbb{R}_+^{p+q} \mid \exists\, i \in \{1, \ldots, n\} \ : \ (y \leqslant Y_i) \ \wedge \ (x \geqslant X_i)\right\}$$

which has become a popular estimation method (De Borger et al., 1994; Leleu, 2006). The convex hull of $\hat{\Phi}_{\mathrm{FDH}}$, called the Data Envelopment Analysis (DEA), is the smallest free disposal convex set covering the data (Seiford and Thrall, 1990). Asymptotic results for these two estimators can be found for example in Kneip et al. (1998) for the DEA and Park et al. (2000) for the FDH.

The FDH estimator and other data envelopment techniques are only consistent when the production units are observed without noise, which implies in particular that (X, Y) belongs to Φ almost surely. However, especially the FDH is very sensitive to the contamination of the data by measurement errors or by outliers (Cazals et al., 2002; Daouia et al., 2009). As measurement errors are frequently encountered in economic data bases, more robust estimation procedures are needed.

Cazals et al. (2002) propose a new non parametric estimator that addresses the problem of contaminated samples in non parametric frontier estimation. For $p = 1$ and under the free disposability assumption, they show that the frontier function $\varphi(y)$ can be represented as

$$\varphi(y) = \inf\{x \in \mathbb{R}_+ \text{ such that } S_{X|Y \geqslant y}(x) < 1\}, \tag{2.1}$$

where $S_{X|Y \geqslant y}(x) = \mathbf{P}(X \geqslant x | Y \geqslant y)$ denotes the conditional survival function. The authors observe that for m independent replications (X_i, Y_i) of the couple (X, Y), the expected minimum input functions

$$\varphi_m(y) := \mathbf{E}\left(\min\{X_1, \ldots, X_m\} | \min\{Y_1, \ldots, Y_m\} \geqslant y\right) \qquad (m \in \mathbb{N}) \tag{2.2}$$

are such that

$$\varphi_m(y) := \int_0^\infty \left\{S_{X|Y \geqslant y}(u)\right\}^m \mathrm{d}u, \tag{2.3}$$

and $\varphi_m(y)$ converges point-wise to the frontier $\varphi(y)$ as m tends to infinity, assuming the existence of $\varphi_m(y)$ for $m = 1$ and hence for all $m \geqslant 1$. This follows from Lebesgue's convergence theorem, because $S_{X|Y \geqslant y}(u) = 1$ for all $u \in [0, \varphi(y)]$ and $S_{X|Y \geqslant y}(u) < 1$ for all $u > \varphi(y)$. The functions $\varphi_m(y)$ are estimated in Cazals et al. (2002) using non parametric estimators of the conditional survival function $S_{X|Y \geqslant y}$. The empirical survival function is defined by

$\hat{S}_{X,Y}(x,y) = n^{-1}\sum_{i=1}^{n} \mathbf{1}(X_i \geqslant x, Y_i \geqslant y)$ and an empirical version of $S_{X|Y\geqslant y}$ is given by

$$\hat{S}_{X|Y\geqslant y}(x) = \frac{\hat{S}_{X,Y}(x,y)}{\hat{S}_Y(y)}, \tag{2.4}$$

where $\hat{S}_Y(y) = n^{-1}\sum_i \mathbf{1}(Y_i \geqslant y)$. Cazals et al. (2002) have studied the asymptotic properties of the *m-frontier estimator*

$$\hat{\varphi}_{m,n}(y) := \int_0^\infty \left\{\hat{S}_{X|Y\geqslant y}(u)\right\}^m \mathrm{d}u \tag{2.5}$$

and they claim that it is less sensitive to extreme values or noise in the sample of production units than FDH- or DEA-type estimators. Our starting point is the observation that when the noise level on the data does not vanish as the sample size n grows, then the m-estimator is no longer consistent.

In this chapter, a new robust estimator of the survival function is studied which copes with an additive error in the inputs X. The error is supposed to be normally distributed with mean zero and unknown variance σ^2. We adapt the density deconvolution techniques from Chapter 1 in order to apply them in this context.

The chapter is organized as follows. In the following section, we consider the problem of estimating the survival function of a random variable from noisy observations. A consistent estimator based on the techniques from the previous chapter is developed. In Section 2.2, we first show that the m-frontier estimator is not consistent in the presence of noise in the data because it is based on the empirical survival function. Finally we show a sufficient condition under which consistency is preserved when plugging in the consistent estimator of the survival function in the m-frontier estimator. Section 2.4 summarizes the results of this chapter and suggests future directions of research. Two technical lemmas are deferred to the end of the chapter. This chapter is mainly based on the article by Schwarz, Van Bellegem, and Florens (2011).

2.1 Estimating the survival function from noisy observations

Suppose we observe a sample $\{Z_1, \ldots, Z_n\}$ of n independent replications of the random variable Z from the model $Z = X + \varepsilon$, where ε is a $\mathcal{N}(0, \sigma^2)$ random variable, independent from the positive random variable X, and with unknown variance $\sigma^2 > 0$. As explained in Chapter 1, the probability density of Z is the convolution $\phi_\sigma * f^X$, where f^X is the probability density of X and ϕ_σ denotes the central normal density with standard deviation σ. As we suppose X to be a positive random variable, it follows immediately from Theorem 1.2 that the density of X and thus its survival function S^X are identifiable from the

observation of Z. Observe that the survival function S^Z of Z can be written as the convolution

$$S^Z(z) = \phi_\sigma * S^X(z).$$

The estimator of S^X is constructed as an approximation in a sieve as follows. For any integers $k, D > 0$, define

$$\Delta^{(k,D)} := \{\delta \in \mathbb{R}^k \mid 0 \leqslant \delta_1 \leqslant \ldots \leqslant \delta_k \leqslant D\},$$

and for $\delta \in \Delta^{(k,D)}$ define

$$S_\delta(t) := \frac{1}{k} \sum_{j=1}^{k} \mathbf{1}(\delta_j > t) \ . \tag{2.6}$$

For any $\delta \in \Delta^{(k,D)}$, denote by P_δ the probability distribution corresponding to the survival function S_δ. The choice of the approximating function is performed minimizing the contrast function

$$\gamma(S, \zeta; T) := \int_{-\infty}^{\infty} \big|(\phi_\zeta * S)(t) - T(t)\big| K(t) \mathrm{d}t,$$

where K is some strictly positive probability density ensuring the existence of the integral. We are now in position to define our estimator of the survival function. Let $(k_n)_{n\in\mathbb{N}}$ and $(D_n)_{n\in\mathbb{N}}$ be two positive, divergent sequences of integers. The estimator $(S_{\hat{\delta}(n)}, \hat{\sigma}_n)$ is defined by

$$(\hat{\delta}(n), \hat{\sigma}_n) := \underset{\substack{\delta \in \Delta^{(k_n, D_n)} \\ \sigma \in [0, D_n]}}{\operatorname{argmin}} \gamma(S_\delta, \sigma; \hat{S}_n^Z) \ , \tag{2.7}$$

where $\hat{S}_n^Z := n^{-1} \sum_{k=1}^{n} \mathbf{1}(Z_k > t)$ is the empirical survival function of Z. Note that the argmin is attained because it is taken over a compact set of parameters, but it is not necessary unique. If it is not, an arbitrary value among the possible solutions may be chosen.

Theorem 2.1 *The estimator $(S_{\hat{\delta}(n)}, \hat{\sigma}_n)$ is consistent in the sense that*

$$P^X_{\hat{\delta}_n} \xrightarrow{\mathcal{D}} P^X \qquad \text{and} \qquad \hat{\sigma}_n \to \sigma$$

almost surely as $n \to \infty$, where $\xrightarrow{\mathcal{D}}$ denotes weak convergence of probability measures.

The proof of this result uses two technical lemmas which can be found in Section 2.3.

Proof. For probability distributions P, P', and positive real numbers σ, σ', define the distance $\Delta(P, \sigma; P', \sigma') := d(P, P') + |\sigma - \sigma'|$, where $d(\cdot, \cdot)$ denotes the Lévy distance, which metrizes weak convergence (see (1.8) for its definition). The theorem is hence equivalent to stating that, almost surely,

$$\Delta(P_{\hat{\delta}(n_k)}, \hat{\sigma}_{n_k}; P^X, \sigma) \to 0$$

as $k \to \infty$. The proof is by contradiction. Suppose that there is some $d > 0$ and an increasing sequence $(n_k)_{k\in\mathbb{N}}$ such that

$$\Delta(P_{\hat{\delta}_{n_k}}, \hat{\sigma}(n_k); P^X, \sigma) > d$$

for all $k \in \mathbb{N}$. By virtue of Lemma 2.4, the distributions given by $(S_{\hat{\delta}(n)} * \phi_{\hat{\sigma}_n})$ converge almost surely weakly to P^Z. Lemma 1.10 from the previous chapter implies that there is a distribution P_∞, some $\sigma_\infty \geqslant 0$, and a sub-sequence $(n'_k)_{k\in\mathbb{N}}$ such that almost surely

$$P_{\hat{\delta}(n'_k)} \xrightarrow{\mathcal{D}} P_\infty \qquad \text{and} \qquad \hat{\sigma}_{n'_k} \to \sigma_\infty,$$

which implies the almost sure point-wise convergence of $S_{\hat{\delta}_{n'_k}}$ to S_∞. Fatou's lemma then implies

$$\gamma(S_\infty, \sigma_\infty; S^Z) \leqslant \liminf_{k\to\infty} \gamma(S_{\hat{\delta}(n'_k)}, \hat{\sigma}_{n'_k}; S^Z) = 0 \qquad \text{almost surely,}$$

where the last equality holds because of Lemma 2.4. Hence, $\gamma(S_\infty, \sigma_\infty; S^Z) = 0$, and using continuity, we conclude that $S_\infty * \phi_{\sigma_\infty} = S^X * \phi_\sigma$. Or equivalently, in terms of distributions, $P_\infty * \phi_{\sigma_\infty} = P^X * \mathcal{N}_\sigma$. As all the distributions $P_{\hat{\delta}(n'_k)}$ have their mass on the positive axis, Lemma 1.8 from the previous chapter implies that $P_\infty \in \mathcal{P}_0$, and hence that $P_\infty = P^X$ and $\sigma_\infty = \sigma$, which is a contradiction to the assumption and thus concludes the proof. □

To illustrate the estimator, we present the result of a Monte Carlo experiment. We consider two designs for the input X. One is uniformly distributed over $[1, 2]$, and the other is a mixture $U[1, 2]+Exp(1)$. In both cases the density of X is zero below 1, and in the second case the support of X is not bounded to the right. For various true values of σ, we calculate the estimators $(\hat{\delta}(n), \hat{\sigma}_n)$ for sample sizes $n = 100$, 200 and 500. No particular optimization over the value of k (appearing in (2.6)) is provided, except that we increase k as the sample size increases. For the considered sample sizes, we set $k = 10\, n^{1/2}$. This choice is surely ad hoc; a better choice could possibly be made by means of a bootstrap procedure. The minimization of the contrast function is calculated using the algorithm `optim` in the R software. For this algorithm, we have chosen the initial values of δ_j to be equi-spaced values over the interval $[0, 3]$ and the initial value of σ is the empirical standard deviation of the sample $Z_1, \ldots, Z_n$.

	True σ		
n	1	2	5
100		1.30	-1.08
		(1.05)	(0.51)
200	0.91	0.07	-0.38
	(3.84)	(0.45)	(0.45)
500	0.37	0.06	0.14
	(0.30)	(0.44)	(0.49)

Table 2.1: The inputs simulated in this experiment are uniformly distributed over $[1, 2]$. For each sample size and noise level, we compute the mean of $\sigma - \hat{\sigma}_n$ from $B = 2000$ replications (the standard deviation is given between parentheses)

	True σ		
n	1	2	5
100		2.84	-0.92
		(7.80)	(7.15)
200		-0.49	-0.49
		(6.32)	(5.92)
500	1.78	0.029	0.014
	(5.90)	(4.88)	(6.69)

Table 2.2: The inputs simulated in this experiment are a mixture $U[1, 2] + Exp(1)$. For each sample size and noise level, we compute the mean of $\sigma - \hat{\sigma}_n$ from $B = 2000$ replications (the standard deviation is given between parentheses)

Tables 2.1 and 2.2 show the result of the Monte Carlo simulation using $B = 2000$ replications of each design. The mean and standard deviation of $\sigma - \hat{\sigma}_n$ over the B replications are displayed. Some results are not reported for very small sizes, because a stability problem has been observed, especially in the mixture case. In these cases, the `optim` algorithm did not always converge (a similar phenomenon has been observed using the `nlm` algorithm). Better numerical results might be obtained designing an optimization algorithm taking into account the specific properties of the contrast function and the sieve under consideration instead of using the standard procedures. Note further that the simulations have been run for very small sample sizes in order to evaluate the small sample behavior of the estimator. Also, observe that due to condition (2.11) in Theorem 2.2, the convergence of the estimator essentially depends on the rate at which the parameter sequence m_n diverges to infinity. Its choice is crucial for the performance of the estimator as we also discuss at the end of this chapter.

It also has to be mentioned that the procedure is fairly sensitive to the choice of k and to the choice of initial values for δ and σ. For larger sample sizes, or larger values of the noise, the results overall improve with the sample size.

2.2 Robust *m*-frontier estimation in the presence of noise

Let us now consider our initial problem of consistently estimating the production frontier $\varphi(y)$ from a sample of production units (X_i, Y_i), where X_i is the input and Y_i is the output. In this section we assume that the dimension of the input and the output are $p = q = 1$.

First, we discuss briefly the fact that the standard m-frontier estimator is not consistent when applied to noisy data. In the second subsection, we will introduce the plug-in estimator and show a sufficient condition under which it is consistent.

2.2.1 Inconsistency of the *m*-frontier estimator

The m-frontier estimator (cf. (2.5)) is more robust in the presence of noise that the FDH or DEA estimator. In Cazals et al. (2002, Theorem 3.1) it is shown that for any interior point y in the support of the distribution Y and for any $m \geqslant 1$, it holds that

$$\hat{\varphi}_{m,n}(y) \to \varphi_m(y) \qquad \text{almost surely as } n \to \infty \tag{2.8}$$

where $\varphi_m(y)$ is the expected minimum input function of order m given in equation (2.2). When the input of the production units is contaminated by a centered normal additive error, the actually observed inputs are

$$Z_i = X_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

for $i = 1, \ldots, n$, instead of X_i, for some positive, unknown variance parameter σ^2. If σ^2 does not vanish asymptotically, the limit appearing in (2.8) is no longer given by the expected minimum input function (2.2). Instead we get

$$\hat{\varphi}_{m,n}(y) \to \mathbf{E}\left[\min\{Z_1, \ldots, Z_m\} | Y \geqslant y\right] \qquad \text{almost surely as } n \to \infty.$$

The expectation appearing on the right hand side does not coincide with (2.2), because the support of the variable Z is the whole real line. Therefore, the m-frontier estimator does not converge to the desired target function, due to the non vanishing error variance. Note that Hall and Simar (2002) and Simar (2007) assume the noise level to be asymptotically negligible.

The inconsistency of the m-frontier estimator is illustrated in Figures 2.1 and 2.2. The true production frontier in this simulation is given by $\varphi(y) = y^{1/2}$ and is displayed by the dotted line. We have simulated 200 production inputs from the model $X_i = Y_i^2 + E_i$, where $E_i \sim Exp(1)$. The production inputs are then contaminated by an additive noise, so that the observed inputs are $Z_i = X_i + \varepsilon_i$ instead of X_i, where the ε_i are independently generated from a zero mean normal variable with standard error $\sigma = 2$.

The FDH estimator computed in Figure 2.1 is known not to be consistent in this situation, because it is constructed under the assumption that all production units are in the production set Φ with probability one. Figure 2.2 shows the m-frontier of Cazals et al. (2002) for $m = 1$ and 50 respectively (cf. (2.5)). As discussed in Cazals et al. (2002), the appropriate choice of m is delicate and, as far as we know, there is yet no automatic procedure to select it from the data. If m is too low, the m-frontier is not a good estimator of the production function. In the theory of Cazals et al. (2002), m is an increasing parameter with respect to the sample size.

For larger values of m, as shown in Figure 2.2, the estimator is close to the FDH estimator. Because the value of m increases with n in theory, the two estimators will be asymptotically close. This illustrates the inconsistency of the m-frontier in the case where the noise on the data is not vanishing with increasing sample size.

2.2.2 Robust *m*-frontier estimation

In order to recover the consistency of the m-frontier, we need to plug-in a consistent estimator of the conditional survival function in (2.3). The construction of the estimator is easy from the above results if we assume that the additive noise to the inputs is independent from the input X and the output Y. Let y be a point in the output domain where the support of Y is strictly positive. Restricting the data set to $(Z_i | Y_i \geqslant y)$, we can construct the empirical conditional survival function $\hat{S}_{Z|Y\geqslant y}$ using the usual non parametric estimator (2.4). Note that this estimator does not require any regularization parameter such as

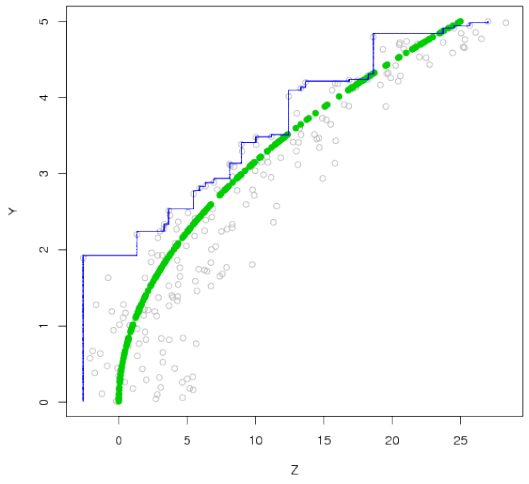

Figure 2.1: The gray points are the simulated production units and the thick line is the true production frontier. The solid line is the Free Disposal Hull (FDH) estimator of the frontier.

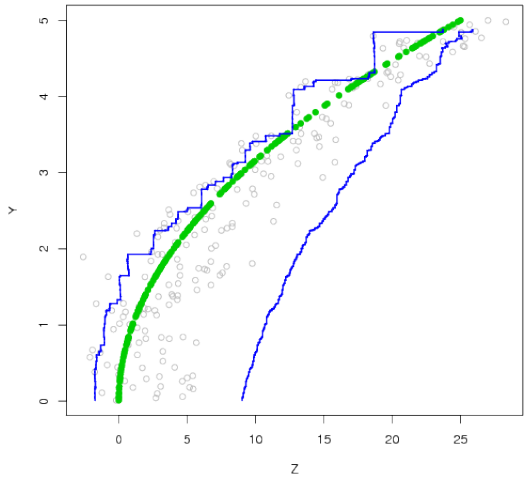

Figure 2.2: Using the same data as in Figure 2.1, the two solid lines are the m-frontier estimator with $m = 1$ and $m = 50$ respectively.

a bandwidth. In analogy to (2.7), we also define

$$(\hat{\delta}(n), \hat{\sigma}_n) := \underset{\substack{\delta \in \Delta^{(k_n, D_n)} \\ \sigma \in [0, D_n]}}{\operatorname{argmin}} \gamma(S_\delta, \sigma; \hat{S}_{Z|Y \geqslant y}) \,. \tag{2.9}$$

The robust m-frontier estimator is then defined by

$$\hat{\varphi}^{rob}_{m,n}(y) := \int_0^\infty \left\{ S_{\hat{\delta}(n)}(u) \right\}^m \mathrm{d}u \,. \tag{2.10}$$

This integral is easy to compute since $S_{\hat{\delta}(n)}$ is a step function. The following result establishes the consistency of this new estimator under an independence assumption and a growth condition on the parameter m. The assumption that there is an error only in the input variable and that this error is further independent of the in- and output may be hard to verify in practice. A next step in the development of robust frontier estimation techniques could be the relaxation of these hypotheses as we discuss at the end of this chapter.

Theorem 2.2 *Suppose that we have n independent observations $(Z_i, Y_i)_{i=1,\ldots,n}$ of production units, where the input data Z_i are a noisy version of the true inputs X_i in the sense that $Z_i = X_i + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is a measurement error that is independent from X_i and Y_i and whose variance σ^2 is unknown. Consider the robust m-frontier estimator given by equations (2.9) and (2.10) and let m_n be a divergent monotone sequence of positive integers such that*

$$\{ S_{\hat{\delta}(n)}(\varphi(y)) \}^{m_n} \to 1 \tag{2.11}$$

almost surely as $n \to \infty$. Then $\hat{\varphi}^{rob}_{m_n,n}(y) \to \varphi(y)$ almost surely as $n \to \infty$.

The proof of the theorem uses two technical lemmas which can be found in Section 2.3 below.

Proof. We begin the proof by plugging-in the sequence m_n into the robust estimator and by splitting up the integral occurring in (2.10) into

$$\begin{aligned} \int_0^\infty \left\{ S_{\hat{\delta}(n)}(u) \right\}^{m_n} \mathrm{d}u = \\ \int_0^{\varphi(y)} \left\{ S_{\hat{\delta}(n)}(u) \right\}^{m_n} \mathrm{d}u + \int_{\varphi(y)}^\infty \left\{ S_{\hat{\delta}(n)}(u) \right\}^{m_n} \mathrm{d}u =: A_n + B_n \end{aligned} \tag{2.12}$$

with obvious definitions for A_n and B_n. We have that $B_n \to 0$ almost surely as n tends to infinity. To see this, let $t_n := \varphi(y) \vee \sup\{t \in \mathbb{R} \mid S_{\hat{\delta}(n)}(t) = 1\}$ and decompose B_n further into

$$\int_{\varphi(y)}^\infty \{ S_{\hat{\delta}(n)}(u) \}^{m_n} \mathrm{d}u = \int_{\varphi(y)}^{t_n} 1 \, \mathrm{d}u + \int_{t_n}^\infty \{ S_{\hat{\delta}(n)}(u) \}^{m_n} \mathrm{d}u. \tag{2.13}$$

Firstly, $t_n \to \varphi(y)$ as $n \to \infty$ because of the consistency of $S_{\hat{\delta}(n)}$. Therefore, the first integral on the right hand side of (2.13) tends to 0 as $n \to \infty$. Secondly, $S_{\hat{\delta}(n)}$ is non increasing and strictly smaller than 1 on (t_n, ∞) for every $n \in \mathbb{N}$. As the sequence $S_{\hat{\delta}(n)}$ is further surely point-wise convergent on $\mathbb{R}$, the other integral of the decomposition in (2.13) also tends to 0.

It remains to show that $A_n \to \varphi(y)$ almost surely as $n \to \infty$. Since $S_{\hat{\delta}(n)}$ is non increasing and $S_{\hat{\delta}(n)}(0) = 1$, we have that $s_n \leqslant \varphi(y)$. On the other hand, $s_n \geqslant \varphi(y) \{S_{\hat{\delta}(n)}(\varphi(y))\}^{m_n}$, which concludes the proof of the result in view of the assumption. □

This result illustrates well the role of the parameter $m = m_n$, which has to tend to infinity at an appropriate rate as $n \to \infty$ in order to achieve consistency of the robust frontier estimator. Indeed, were m_n bounded by some $M > 0$, Fatou's Lemma would imply that almost surely

$$\lim_{n\to\infty} \hat{\varphi}^{rob}_{m_n,n}(y) \geqslant \int_0^\infty \left\{S_{X|Y\geqslant y}(u)\right\}^M \, \mathrm{d}u = \varphi(y) + \int_{\varphi(y)}^\infty \left\{S_{X|Y\geqslant y}(u)\right\}^M \, \mathrm{d}u.$$

Except for the trivial case where the true conditional survival function is the indicator function of the interval $(-\infty, \varphi(y))$, the last integral on the right hand side is strictly positive. This shows that the robust estimator asymptotically overestimates the true frontier $\varphi(y)$ if m_n does not diverge to infinity.

On the other hand, if m_n increases too fast in the sense that the condition in (2.11) does not hold, then $\hat{\varphi}^{rob}_{m_n,n}(y)$ may asymptotically underestimate the true frontier $\varphi(y)$ as one can see considering the decomposition in (2.12). Indeed, B_n tends to 0 almost surely for $n \to \infty$ as explained in the proof of the above theorem. As for A_n, the integrand converges to a non negative monotone function S with $S(\varphi(y)) < 1$, and hence the integral may tend to a limit that is smaller than the true frontier $\varphi(y)$. However, this need not be the case, and thus the condition in (2.11) is sufficient but not necessary.

Summarizing the above discussion, the sufficient condition in (2.11) implicitly defines an appropriate rate at which m_n has to diverge to infinity such that the new robust frontier estimator is consistent. This rate depends on characteristics of the true conditional survival function, and we do not know at present how to choose it in an adaptive way. Nevertheless, the simulations show that even for finite samples, large choices of m do not deteriorate the performance of the robust estimator.

The estimator is computed for each possible value of y. In practice, it is not necessary to estimate the standard deviation of the noise for each y. We can first estimate the noise level using the marginal data set of inputs only, and use the techniques developed in Section 2.2. We then use this estimated value in (2.9) even as an initial parameter of the `optim` algorithm, or as a fixed, known parameter of the noise standard deviation.

Figure 2.3 shows the estimator on the simulated data of Figure 2.1. As for the standard m-frontier, the robust m-frontier with $m = 1$ is not a satisfactory estimator. An interesting fact about the robust m-frontier is that it does not deteriorate the frontier estimation for large values of m.

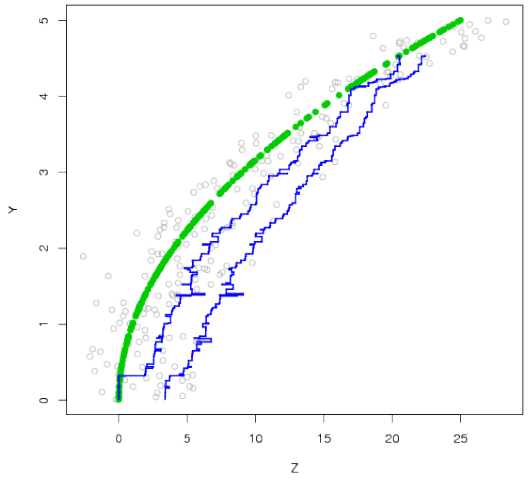

Figure 2.3: Using the same data as in Figure 2.1, the two solid lines are the robust m-frontier estimator with $m = 1$ and $m = 50$ respectively.

2.3 Auxiliary results

Lemma 2.3 *The estimator* $(S_{\hat{\delta}(n)}, \hat{\sigma}_n)$ *satisfies*

$$\gamma(S_{\hat{\delta}(n)}, \hat{\sigma}_n; \hat{S}_n^Z) \to 0 \quad \textit{as } n \to \infty.$$

Proof. By the triangle inequality, we have, for any $(S', \sigma') \in \mathcal{C} \times \mathbb{R}^+$,

$$\begin{aligned} \gamma(S_{\hat{\delta}(n)}, \hat{\sigma}_n; \hat{S}_n^Z) &= \min_{\substack{\delta \in \Delta^{(k_n, D_n)} \\ \tilde{\sigma} \in [0, D_n]}} \gamma(S_\delta, \tilde{\sigma}; \hat{S}_n^Z) \\ &\leqslant \min_{\substack{\delta \in \Delta^{(k_n, D_n)} \\ \sigma \in [0, D_n]}} \gamma(S_\delta, \sigma; S^X * \phi_\sigma) + \gamma(S^X, \phi_\sigma; \hat{S}_n^Z). \end{aligned} \tag{2.14}$$

Let $\eta > 0$ and $T > 0$ be such that $\int_T^\infty S^X(x)\,dx \leqslant \eta/2$. For n sufficiently large, we have $\sigma \leqslant D_n$ and there is $\delta \in \Delta^{(k_n, D_n)}$ with $\int_0^T |(S_\delta - S^X)(x)|\,dx \leqslant \eta/2$,

such that $\int_{\mathbb{R}} |(S_\delta - S^X)(x)|\, dx \leqslant \eta$. It follows that the first term on the right hand side of (2.14) is a vanishing sequence, because

$$\gamma(S_\delta, \sigma; S^X * \phi_\sigma) \leqslant \|(S_\delta - S^X) * \phi_\sigma\|_{L^1} \leqslant \|S_\delta - S^X\|_{L^1} \|\phi_\sigma\|_{L^1} \leqslant \eta.$$

The second term is also a vanishing sequence by virtue of Glivenko-Cantelli's and Lebesgue's Theorem. □

Lemma 2.4 *The estimator $S_{\hat{\delta}(n)}$ defined by (2.7) satisfies*

$$(P_{\hat{\delta}(n)} * \phi_{\hat{\sigma}_n}) \xrightarrow{\mathcal{D}} P^Z$$

almost surely as $n \to \infty$.

Proof. The survival function S^Z is continuous everywhere as it can be written as a convolution with some normal density. Therefore, the convergence

$$\hat{S}_n^Z(x) \xrightarrow{n\to\infty} S^Z(x) \qquad \text{a.s.}$$

holds for every $x \in \mathbb{R}$. Hence, by Lebesgue's theorem,

$$\gamma(S^X, \sigma; \hat{S}_n^Z) \xrightarrow{n\to\infty} 0 \qquad \text{a.s.}$$

The triangle inequality, together with Lemma 2.3, implies

$$\gamma(S_{\hat{\delta}(n)}, \hat{\sigma}_n; S^Z) \leqslant \gamma(S_{\hat{\delta}(n)}, \hat{\sigma}_n; \hat{S}_n^Z) + \gamma(S^X, \sigma; \hat{S}_n^Z) \xrightarrow{n\to\infty} 0 \qquad \text{a.s.}$$

A continuity argument implies

$$(S_{\hat{\delta}(n)} * \phi_{\hat{\sigma}_n})(x) \xrightarrow{n\to\infty} S^Z(x) \qquad \text{a.s.}$$

for every $x \in \mathbb{R}$, which is in fact weak convergence and concludes the proof. □

2.4 Conclusion

The contribution of this chapter is the estimation of frontiers based on observation with additive noise in the input variable. The noise is not assumed to vanish asymptotically. In this situation, the m-frontier estimator introduced by Cazals et al. (2002) is still a valuable tool, but it requires the plug-in of a consistent estimator of the conditional survival function in order to be consistent itself.

Attempting to construct such a consistent estimator, we are confronted with a deconvolution problem which we solve adapting the results of Chapter 1 to the context of the model at hand. Note that the noise level is not known, and therefore needs to be estimated from a cross section of production units.

Measurement errors are frequently encountered in empirical economic data, and the new robust estimator is designed to be consistent in this setting. The rate of convergence of the estimator is unknown, though. The study of its convergence speed might be of interest for future research in efficiency analysis.

Note that the choice of the parameter m_n is crucial for the estimation quality. In view of condition (2.11), the appropriate rate of divergence for m_n which makes the frontier estimator consistent depends on the convergence speed of the deconvolution estimator of the conditional survival function $\widehat{S}_{X|Y\geqslant y}$ in the frontier point. However, this convergence speed is not known. Thus, an adaptive choice of the parameter m_n as a function of the sample size would be of high interest. One approach to adaptivity could consist in deliberately choosing too slow a rate for m_n and trying to estimate the resulting bias. Such a proceeding has been proposed by Daouia et al. (2009). It would be interesting to investigate if this technique can be transferred to the case with noise in the input variable.

One could also be interested in the case where the measurement error is in the output rather than in the input variable. We would like to end this chapter by explaining how the above methods can be adapted to this problem. In this setting, in contrast to Section 2.2, the inputs X_i are directly observed, but only a contaminated version

$$W_i = Y_i + \eta_i, \qquad \eta_i \sim \mathcal{N}(0, \sigma^2) \tag{2.15}$$

of the true output variables Y_i is observed, with η_i independent from X_i and Y_i. Let us briefly discuss the case where both the input and the output spaces are one-dimensional, i.e. $p = q = 1$. As the frontier function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ given in (2.1) is strictly increasing, its inverse function $\varphi^{-1} : \mathbb{R}_+ \to \mathbb{R}_+$ exists. The efficiency boundary can be described by either of the functions φ and φ^{-1}. Estimating φ^{-1} is thus equivalent to estimating φ itself. We can write the inverse frontier function as

$$\varphi^{-1}(x) = \inf\{y \in \mathbb{R}_+ \mid F_{Y|X\leqslant x}(y) = 1\},$$

where $F_{Y|X\leqslant x}$ denotes the conditional distribution function of Y given $X \leqslant x$. To apply the robust m-frontier methodology, we therefore need to estimate the conditional distribution function $F_{Y|X\leqslant x}$. From the model (2.15), one can easily show that the estimation of $F_{Y|X\leqslant x}$ is again a deconvolution problem, and recalling that $F_{Y|X\leqslant x} = 1 - S_{Y|X\leqslant x}$, we can define

$$(\hat{\delta}(n), \hat{\sigma}_n) := \underset{\substack{\delta\in\Delta^{(k_n,D_n)}\\ \sigma\in[0,D_n]}}{\operatorname{argmin}} \gamma(S_\delta, \sigma; \hat{S}_{W|X\leqslant x}) \text{ and } \hat{F}_n := 1 - S_{\hat{\delta}(n)}$$

in analogy to Section 2.2.2. $\hat{F}_n$ is the deconvolving estimator of the conditional distribution function $F_{Y|X\leqslant x}$. We proceed by defining the robust m-frontier

estimator of φ^{-1} as

$$\hat{\varphi}_{m,n}^{-1}(x) := A - \int_0^A \left\{\hat{F}_n(u)\right\}^m \mathrm{d}u,$$

where $A > 0$ is some constant fixed in advance. Let m_n be a strictly divergent sequence such that $\{\hat{F}_n(\varphi(x))\}^{m_n} \to 1$ almost surely as $n \to \infty$. In analogy to Theorem 2.2, it can be shown that for such a sequence, $\hat{\varphi}_{m_n,n}^{-1}(x)$ is consistent if $A > \varphi^{-1}(x)$. Otherwise, $\hat{\varphi}_{m_n,n}^{-1}(x)$ tends to A almost surely. This suggests the following adaptive choice of A. First, one computes the estimator with some arbitrary initial value of A. If the result is close to A, recompute it repeatedly for increasing values of A until a value smaller than A is obtained.

This estimator is thus robust with respect to noise in the output variable, but note that it is not obvious how to generalize this procedure to a multidimensional setting. Moreover, it is not clear how one could cope with a situation with error in both variables. These questions could be subject to further investigation.

Chapter 3

Adaptive circular deconvolution

In this chapter, which is based on Johannes and Schwarz (2009), we deal with the estimation of circular probability densities from noisy observations. «Circular» means that the observations are points on the circle. Such models arise in numerous and various fields of application. Data with temporal structure is most naturally represented in this way – for example, times of day when events of interest happen can be represented as points on a clock face, such as requests in a computer network, financial transactions, or gun crimes (Gill and Hangartner, 2010). Replacing the clock face by a compass rose, directional data can also be treated in the circular setting. Curray (1956) considers the analysis of directional data in the context of geological research. Cochran et al. (2004) investigate migrating birds' navigation abilities using circular data. But the applications of circular data are not restricted to a spatio-temporal context: Gill and Hangartner (2010) give an overview of circular data in political science, where they can be used for example to model political preferences which are neither of temporal nor of spatial nature. For a more detailed discussion of the particularities of circular data we refer to Mardia (1972). Numerous circular data sets and examples of their statistical analysis can be found in Fisher (1993).

Let X be a circular random variable whose density f we are interested in and ε an independent additive circular error with unknown density φ. Denote by $Y = X + \varepsilon$ the contaminated observation data and by g its density. Throughout this chapter we will identify the circle with the unit interval $[0, 1)$, for notational convenience. Let $\lfloor \cdot \rfloor$ denote the floor function. Taking into account the circular nature of the data, the model can be written as $Y = X + \varepsilon - \lfloor X + \varepsilon \rfloor$.

Then, we have

$$g(y) = (f * \varphi)(y) := \int_{[0,1)} f((y-s) - \lfloor y-s \rfloor)\, \varphi(s)\, ds, \quad y \in [0,1),$$

such that $*$ denotes circular convolution. Therefore, the estimation of f is called a circular deconvolution problem. Let $L^2 := L^2([0,1))$ be the Hilbert space of square integrable complex-valued functions defined on $[0,1)$ as defined in the introduction of this thesis. We equip this space with the usual inner product $\langle f, g\rangle = \int_{[0,1)} f(x)\overline{g(x)}dx$ where $\overline{g(x)}$ denotes the complex conjugate of $g(x)$. In this chapter we suppose that f and φ, and hence g, belong to the subset $\mathcal{D}$ of all densities in L^2. It would be interesting as well to consider densities in L^1 or, more generally, in L^p spaces with $p \in [1,\infty]$. The techniques that we employ in this chapter make however use of the specific properties of L^2 spaces, for example the convergence of the Fourier series. A collection of hypotheses and techniques allowing for estimation in L^1 can be found in the monograph by Devroye and Györfi (1985).

In L^2, the densities admit representations as discrete Fourier series with respect to the exponential basis $\{e_j\}_{j\in\mathbb{Z}}$ of L^2, where $e_j(x) := \exp(-i2\pi jx)$ for $x \in [0,1)$ and $j \in \mathbb{Z}$. Given $p \in \mathcal{D}$ and $j \in \mathbb{Z}$, let $[p]_j := \langle p, e_j\rangle$ be the j-th Fourier coefficient of p. In particular, $[p]_0 = 1$. The key to the analysis of the circular deconvolution problem is the discrete convolution theorem which states that $g = f * \varphi$ if and only if $[g]_j = [f]_j[\varphi]_j$ for all $j \in \mathbb{Z}$. Therefore, as long as $[\varphi]_j \neq 0$ for all $j \in \mathbb{Z}$, which we assume from now on, we have

$$f = 1 + \sum_{|j|>0} \frac{[g]_j}{[\varphi]_j}\, e_j \quad \text{with } [g]_j = \mathbf{E}[e_j(-Y)] \text{ and } [\varphi]_j = \mathbf{E}[e_j(-\varepsilon)] \qquad \forall\, j \in \mathbb{Z}. \tag{3.1}$$

Note that representation like (3.1) also holds in the case of deconvolution on the real line when the X-density is compactly supported, but the error term ε, and hence Y, take their values in $\mathbb{R}$. In this situation, the deconvolution density still admits a discrete representation as in (3.1), but involving the characteristic functions of φ and g rather than their discrete Fourier coefficients. In fact, apart from guaranteeing the discrete representation, the circular structure of the model is only exploited in the proof of the lower bounds, the upper bounds remaining valid in the case of a compactly supported density on the real line.

In this chapter we suppose that we know neither the density $g = f * \varphi$ of the contaminated observations, nor the error density φ. But we have at our disposal two independent samples of iid. random variables

$$Y_k \sim g, \quad (k = 1,\dots,n) \qquad \text{and} \qquad \varepsilon_k \sim \varphi, \quad (k = 1,\dots,m) \tag{3.2}$$

distributed according to the densities g and φ, respectively. In practical situations, such an additional sample of the error distrubution is available for instance when calibration measurements can be performed. For example, many

digital cameras have an automatic color balance feature which necessitates taking a picture of a white sheet of paper. This can be interpreted as an observation of $Y = X + \varepsilon$ with known X, thus yielding an observation of ε itself. In the context of circular data, one may rather think of calibrating navigation devices.

Our purpose is to establish a fully data-driven estimation procedure for the deconvolution density f which attains optimal convergence rates in a minimax-sense. More precisely, given classes $\mathcal{F}_\gamma^r$ and $\mathcal{E}_\lambda^d$ (defined below) of deconvolution and of error densities, respectively, we shall measure the accuracy of an estimator $\widetilde{f}$ of f by the maximal weighted risk $\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\mathbf{E}\|\widetilde{f}-f\|_\omega^2$ defined with respect to some weighted norm

$$\|f\|_\omega^2 := \sum_{j\in\mathbb{Z}} \omega_j |[f]_j|^2, \quad f \in L^2 \tag{3.3}$$

where $\omega := (\omega_j)_{j\in\mathbb{Z}}$ is a strictly positive sequence of weights. This allows us to quantify the estimation accuracy in terms of the mean integrated square error (MISE) not only of f itself, but as well of its derivatives, for example. It is well known that even in case of a known error density the maximal risk in terms of the MISE in the circular deconvolution problem is essentially determined by the asymptotic behavior of the sequences of Fourier coefficients $([f])_{j\in\mathbb{Z}}$ and $([\varphi])_{j\in\mathbb{Z}}$ of the deconvolution density and the error density, respectively. For a fixed deconvolution density f, a faster decay of the ε-density's Fourier coefficients $([\varphi])_{j\in\mathbb{Z}}$ results in a slower optimal rate of convergence. In the standard context of an ordinary smooth deconvolution density for example, i.e. when $([f])_{j\in\mathbb{Z}}$ decays polynomially, logarithmic rates of convergence appear when the error density is super smooth, i.e. when $([\varphi])_{j\in\mathbb{Z}}$ has exponential decay. Efromovich (1997) treats exclusively this special case, for example. However, this situation and many others are covered by the density classes

$$\mathcal{F}_\gamma^r := \left\{ p \in \mathcal{D} \;\middle|\; \sum_{j\in\mathbb{Z}} \gamma_j |[p]_j|^2 =: \|p\|_\gamma^2 \leqslant r \right\} \text{ and}$$

$$\mathcal{E}_\lambda^d := \left\{ p \in \mathcal{D} \;\middle|\; d^{-1} \leqslant \frac{|[p]_j|^2}{\lambda_j} \leqslant d \quad \forall\, j \in \mathbb{Z} \right\},$$

where $r, d \geqslant 1$ and the positive weight sequences $\gamma := (\gamma_j)_{j\in\mathbb{Z}}$ and $\lambda := (\lambda_j)_{j\in\mathbb{Z}}$ specify the asymptotic behavior of the respective sequence of Fourier coefficients. In section 3.2 we show a lower bound of the maximal weighted risk which is essentially determined by the sequences γ, λ, and ω. This lower bound is composed of two main terms, each of them depending on the size of one sample, but not on the other. Let us define an orthogonal series estimator by replacing

the unknown Fourier coefficients in (3.1) by empirical counterparts, that is,

$$\widehat{f}_k := 1 + \sum_{0<|j|\leqslant k} \frac{\widehat{[g]}_j}{\widehat{[\varphi]}_j} \mathbf{1}_{[|\widehat{[\varphi]}_j|^2 \geqslant 1/m]} e_j \quad \text{with}$$

$$\widehat{[g]}_j := \frac{1}{n}\sum_{i=1}^{n} e_j(-Y_i) \quad \text{and} \quad \widehat{[\varphi]}_j := \frac{1}{m}\sum_{i=1}^{m} e_j(-\varepsilon_i). \quad (3.4)$$

For each j, we introduce a threshold for the estimated coefficient $\widehat{[\varphi]}_j$ that corresponds, as Neumann (1997) remarks, to the rate at which $[\varphi]_j$ can be estimated. Again, things work out analogously in deconvolution on the real line, where one only has to replace the empirical Fourier coefficients by the corresponding values of the empirical characteristic functions. Similar estimators have already been studied by Neumann (1997) on the real line and by Efromovich (1997) in the circular case, for example.
We show in this chapter that the estimator $\widehat{f}_k$ can attain the lower bound and is hence minimax optimal. By comparing the minimax rates in the cases of known and unknown error density, we can characterize the influence of the estimation of the error density on the quality of the estimation. In particular, depending on the Y-sample size n, we can determine the minimal ε-sample size m_n needed to attain the same upper risk bound as in the case of a known error density, up to a constant. Interestingly, the required sample size m_n is far smaller than n in a wide range of situations. For example, in the super smooth case, it is sufficient that the size of the ε-sample is a polynomial in n, that is, $m_n = n^r$ for any $r > 0$.

Of course, minimax optimality is only achieved as long as the dimension parameter k is chosen in an appropriate way. In general, this optimal choice of k depends among other things on the sequences γ and λ. However, in the special case where the error density is known to be super smooth and the deconvolution density is ordinary smooth, the optimal dimension parameter depends only on λ but not on γ. Hence, the estimator is automatically adaptive with respect to γ under the optimal choice of k. In this situation Efromovich (1997) provides an estimator which is also adaptive with respect to the super smooth error density. On the contrary, Cavalier and Hengartner (2005), deriving oracle inequalities in an indirect regression problem based on a circular convolution contaminated by Gaussian white noise, treat the ordinary smooth case only. Like in our setting, their observation scheme involves two independent samples. It is worth to note that in order to apply these estimators, one has to know in advance at least if the error density is ordinary or super smooth. We provide in this chapter a unified estimation procedure which can attain minimax rates in either of the both cases, that is, which is adaptive over a class including both ordinary and super smooth error densities. This fully adaptive method to choose the parameter k only depends on the observations and not on characteristics of

neither f nor φ. The central result of the present chapter states that for this automatic choice $\widehat{k}$, the estimator $\widehat{f}_{\widehat{k}}$ attains the lower bound up to a constant, and is thus minimax-optimal, over a wide range of sequences γ and λ, covering in particular both ordinary and super smooth error densities. A similar result has recently been derived in the context of a functional linear regression model by Comte and Johannes (2010).

As far as the two sample sizes are concerned, the assumption made by Cavalier and Hengartner (2005) on the respective noise levels can be translated to our model by stating that the ε-sample size m is at least as large as the Y-sample size n. This assumption is also used by Efromovich (1997). Note also that in the functional linear regression model, only one sample size n occurs (c.f. Comte and Johannes (2010)). However, as mentioned above, without changing the minimax rates, the ε-sample size can be reduced to m_n, which can be far smaller than n. This is a desirable property, as the observation of the additional sample from ε may be expensive in practice. Nevertheless, the minimal choice of m depends among other things on the sequences γ and λ and is hence unknown in general. In spite of the minimax rate being eventually deteriorated by choosing the sample size m smaller than n, the proposed estimator still attains this rate in many cases, that is, no price in terms of convergence rate has to be paid for adaptivity.

The adaptive choice of k is motivated by the general model selection strategy developed in Barron et al. (1999). Concretely, following Comte and Taupin (2003), who treat the case of a known error density only, we first define a contrast $\Upsilon(\cdot)$ such that the orthogonal series estimator $\widehat{f}_k$ is its argmin and $\Upsilon(\widehat{f}_k) = -\|\widehat{f}_k\|_\omega^2$. See the proof of Theorem 3.12 for the details. Then, $\widehat{k}$ is the minimizer of a penalized contrast

$$\widehat{k} := \operatorname*{argmin}_{1 \leqslant k \leqslant K} \big\{ -\|\widehat{f}_k\|_\omega^2 + \operatorname{pen}(k) \big\}.$$

Note that the norm $\|\widehat{f}_k\|_\omega^2$ can easily be computed. As in case of a known error density, it turns out that the penalty function $\operatorname{pen}(\cdot)$ as well as the upper bound K needed for the right choice of k depend on a characteristic of the error density which is now unknown. This quantity is often referred to as the degree of ill-posedness of the underlying inverse problem. Therefore, as an intermediate step, we allow the penalty function $\operatorname{pen}(\cdot)$ and the upper bound K to depend on the error density. We then show an upper risk bound for the resulting partially adaptive estimator. We prove that over a wide range of sequences γ, this choice of k yields the same upper risk bound as the optimal choice, up to a constant. Finally, we choose k fully adaptively by replacing $\operatorname{pen}(\cdot)$ and K by their empirical versions which depend only on the data. As in the case of known degree of ill-posedness, we show an upper risk bound for the now fully adaptive estimator.

This chapter is organized as follows. In the first section, we discuss the case

of direct density estimation, that is, without noise on the data. This example serves as an introduction to the tools used throughout this (and the following) chapter, such as orthogonal series estimators, minimax theory, and adaptation by model selection. In Section 3.2, we develop the minimax theory for the circular deconvolution model with respect to the weighted norms introduced above and we compute the rates which we can obtain in different configurations for the weight sequences. Section 3.3 is devoted to the construction of the adaptive estimator and an upper risk bound is shown. Again, the result is illustrated by the example configurations considered in Section 3.2. On account of legibility, some technical Lemmas are deferred to the end of the chapter.

3.1 Minimax and adaptivity – an introductory example

In order to introduce the basic concepts of this chapter – namely orthogonal series estimators, minimax theory, and adaptation – we consider in this section the case of direct density estimation without noise in the variables. This allows us to explain the method in a general way without having to cope with too many technicalities.

Orthogonal series

Let X be a real-valued random variable distributed according to some unknown density $f \in L^2[0,1]$ and suppose an iid. sample $X_1, \dots, X_n$ from X. Let $(e_j)_{j\in\mathbb{Z}}$ be the exponential basis of the Hilbert space $L^2[0,1]$ and denote by $[f]_j := \langle f, e_j\rangle = \mathbf{E}[e_j(-X)]$ the coefficients of f with respect to this basis. These coefficients can be estimated without bias by $\widehat{[f]}_j := n^{-1}\sum_{k=1}^n e_j(-X_k)$. Then, the density has the representation $f = \sum_{j\in\mathbb{Z}}[f]_j\, e_j$ and a natural estimator of f is given by the orthogonal sum $\widehat{f}_k = \sum_{|j|\leqslant k}\widehat{[f]}_j\, e_j$. The mean integrated squared error (MISE) is then easily seen to be bounded by

$$\mathbf{E}\|\widehat{f}_k - f\|^2 \leqslant kn^{-1} + \sum_{|j|>k}[f]_j^2,$$

where the first summand is the variance term and the second one the bias term. Obviously, the bias term tends to zero as k tends to infinity. Thus, $\widehat{f}_k$ is consistent if $k\to\infty$ and $k/n\to 0$ simultaneously.

Minimax

One performance measure for our estimator $\widehat{f}_k$ is its maximal risk over a given class of possible density functions. Suppose that f lies in the class $\mathcal{F}_\gamma^r$ which

we have defined in the introduction of this chapter. In this paragraph, we will only consider the special case where

$$\gamma_0 = 1 \quad \text{and} \quad \gamma_j = j^{2p} \qquad \text{for } |j| > 1$$

in order to keep things simple. The *minimax risk* is then

$$\inf_{\widetilde{f}} \sup_{f \in \mathcal{F}_\gamma^r} \mathbf{E}\|\widetilde{f} - f\|^2,$$

where the infimum is taken over all possible estimators $\widetilde{f}$ of f. Such an estimator is said to be *minimax optimal* with respect to the class $\mathcal{F}_\gamma^r$ if its maximal risk over the class $\mathcal{F}_\gamma^r$ is bounded from above by the minimax risk up to a constant. In the case of our orthogonal series estimator, with f in $\mathcal{F}_\gamma^r$, one easily sees that the maximal risk over the class $\mathcal{F}_\gamma^r$ is bounded by

$$\sup_{f \in \mathcal{F}_\gamma^r} \mathbf{E}\|\widehat{f}_k - f\|^2 \leqslant \left\{kn^{-1} + rk^{-2p}\right\}.$$

The upper bound of the maximal risk is minimal when bias and variance term are of the same order which is the case for $k = k_n^* := n^{1/(2p+1)}$ in this example. Plugging in, we obtain

$$\sup_{f \in \mathcal{F}_\gamma^r} \mathbf{E}\|\widehat{f}_{k_n^*} - f\|^2 \leqslant Cn^{-2p/(2p+1)},$$

which is known to be the minimax risk[1]. Hence, the estimator $\widehat{f}_{k_n^*}$ is minimax optimal in the class $\mathcal{F}_\gamma^r$. It is important to notice that minimax optimality is ensured only under the appropriate choice of the dimension parameter k. Its optimal value, k_n^*, however, depends on the class $\mathcal{F}_\gamma^r$ via the parameter p.

Adaptivity

If no information about the parameter p is available, the choice of the dimension parameter k has to be made based on the observations only. We call such a choice *adaptive* to the parameter p if the estimator attains the minimax optimal rate of convergence with respect to the class $\mathcal{F}_\gamma^r$ for a wide range of possible parameters p. Later on in this chapter, we are going to consider more general sequences γ.

As a method of constructing an adaptive estimator, we use the so called *model selection* which is inspired by the techniques summarized in Barron et al. (1999): First, a contrast function is defined such that the orthogonal series estimator can be written as its minimizer. Then, the adaptive choice of the

[1] This follows for example from Proposition 3.8 [o-o] with $a = s = 0$. (cf. also Tsybakov, 2004)

dimension parameter is defined as the minimizer of a penalized contrast. This construction allows us to find an upper bound of the risk like in (3.16). The technical difficulty lies in finding an appropriate decomposition of the terms appearing in this upper bound and in their statistical control.

In the context of our example model, we use the contrast function

$$\Upsilon(t) = \|t\|^2 - 2\langle t, \sum_{j\in\mathbb{Z}} \widehat{[f]}_j \, e_j \rangle.$$

For functions t in the subspace $\mathcal{S}_k := \operatorname{span}\{e_{-k}, \ldots, e_k\}$ in $L^2[0,1]$, this contrast takes the form $\Upsilon(t) = \|t - \widehat{f}_k\|^2 - \|\widehat{f}_k\|^2$ and we can obviously write the series estimator as

$$\operatorname*{argmin}_{t\in\mathcal{S}_k} \Upsilon(t) = \widehat{f}_k. \tag{3.5}$$

Noting that $\Upsilon(\widehat{f}_k) = -\|\widehat{f}_k\|^2$, we define

$$\breve{k} := \operatorname*{argmin}_{k=1,\ldots,n} \{-\|\widehat{f}_k\|^2 + 24\, k\, n^{-1}\}, \tag{3.6}$$

where $24\, k\, n^{-1}$ is the above-mentioned penalty term which has to be of the order of the estimator's variance. The adaptive estimator is then defined as $\widehat{f}_{\breve{k}}$.

Remark 3.1 The constant 24 appearing in the penalty term may seem somewhat arbitrary at first sight, but in fact this choice results from the coefficients arising in the decomposition of the risk which we perform below. The same is true for other numerical constants appearing in the subsequent results and proofs of this work. Although the focus of this work is not the optimality of the constants in the risk bounds, we have tried to give precise numerical values for the constants where ever possible. Constants in risk bounds and their optimalization in various models have been worked on by Barron et al. (1999), for example. □

Letting $\operatorname{pen}(k) = 24kn^{-1}$, we have for all $1 \leqslant k \leqslant n$ that $\Upsilon(\widehat{f}_{\breve{k}}) + \operatorname{pen}(\breve{k}) \leqslant \Upsilon(f_k) + \operatorname{pen}(k)$, using (3.5) and (3.6). Letting $f_k := \sum_{|j|\leqslant k} [f]_j \, e_j$ denote the projection of f, this implies

$$\|\widehat{f}_{\breve{k}}\|^2 - \|f_k\|^2 \leqslant 2\langle \widehat{f}_{\breve{k}} - f_k, \widehat{\Phi}_{\widehat{f}} \rangle + \operatorname{pen}(k) - \operatorname{pen}(\breve{k}),$$

and hence

$$\|\widehat{f}_{\breve{k}} - f\|^2 \leqslant \|f - f_k\|^2 + \operatorname{pen}(k) - \operatorname{pen}(\breve{k}) + 2\langle \widehat{f}_{\breve{k}} - f_k, \Phi_{\widehat{f}} - f \rangle. \tag{3.7}$$

Consider the unit ball $\mathcal{B}_k := \{f \in \mathcal{S}_k \mid \|f\| \leqslant 1\}$ and, for arbitrary $\tau > 0$ and $t \in \mathcal{S}_k$, the elementary inequality

$$2|\langle t, h\rangle| \leqslant 2\|t\| \sup_{t\in\mathcal{B}_k} |\langle t, h\rangle| \leqslant \tau\|t\|^2 + \frac{1}{\tau} \sup_{t\in\mathcal{B}_k} |\langle t, h\rangle|^2 = \tau\|t\|^2 + \frac{1}{\tau} \sum_{j=-k}^{k} |[h]_j|^2.$$

Combining this bound with (3.7) yields

$$\|\widehat{f}_{\check{k}}-f\|^2 \leqslant \|f-f_k\|^2+\tau\,\|\widehat{f}_{\check{k}}-f_k\|^2+\mathrm{pen}(k)-\mathrm{pen}(\check{k})+\frac{1}{\tau}\sup_{t\in\mathcal{B}_{k\vee\check{k}}}|\langle t,\Phi_{\widehat{f}}-f\rangle|^2.$$

Notice that $\|\widehat{f}_{\check{k}}-f_k\|^2 \leqslant 2\|\widehat{f}_{\check{k}}-f\|^2+2\|f_k-f\|^2$ and that $\|f-f_k\|^2 \leqslant r/\gamma_k$ for all $f\in\mathcal{F}_\gamma^r$ because $\gamma_k=k^{2p}$ is non decreasing. Setting $\tau:=1/4$, we obtain

$$\frac{1}{2}\|\widehat{f}_{\check{k}}-f\|^2 \leqslant \frac{3}{2}\,(r/\gamma_k)+\mathrm{pen}(k)-\mathrm{pen}(\check{k})+4\sup_{t\in\mathcal{B}_{k\vee\check{k}}}|\langle t,\Phi_{\widehat{f}}-f\rangle|^2,$$

Then, using $\mathrm{pen}(k\vee\check{k})\leqslant\mathrm{pen}(k)+\mathrm{pen}(\check{k})$,

$$\frac{1}{2}\|\widehat{f}_{\check{k}}-f\|^2 \leqslant \frac{3}{2}\,(r/\gamma_k)+4\Big(\sup_{t\in\mathcal{B}_{k\vee\check{k}}}|\langle t,\Phi_{\widehat{f}}-f\rangle|^2-6\,(k\vee\check{k})/n\Big)_+ +2\,\mathrm{pen}(k)$$

for all $k=1,\dots,n$. Finally, notice that since $k_n^*\leqslant n$ for any $p>0$, the term $\min_{k=1,\dots,n}\max(r/\gamma_k,k/n)$ is of the order of the minimax optimal risk $n^{-2p/(2p+1)}$. Thus, we obtain

$$\sup_{f\in\mathcal{F}_\gamma^r}\mathbf{E}\|\widehat{f}_{\check{k}}-f\|^2 \leqslant Cn^{-2p/(2p+1)} + C\,\mathbf{E}\Big[\sum_{k'=1}^{n}\Big(\sup_{t\in\mathcal{B}_{k'}}|\langle t,\Phi_{\widehat{f}}-f\rangle|^2-6\,(k')/n\Big)_+\Big]$$

The second term can be shown to be of order n^{-1} using an exponential inequality by Talagrand (cf. Theorem A.5), which is one of the main technical tools used in the proofs of this chapter. To end this example, we conclude from the above that the estimator $\widehat{f}_{\check{k}}$ is adaptive and minimax optimal with respect to the class $\mathcal{F}_\gamma^r$ for polynomial sequences $\gamma_j=j^{2p}$ with $p>0$. Note that the constants appearing here and in the rest of the chapter are not optimal, though.

From the next section on, we will be considering the model described in the introduction to this chapter. This involves a second unknown function, namely the density of the error distribution which is supposed to lie in some class $\mathcal{E}_\lambda^d$. The minimax risk under consideration is thus

$$\inf_{\widetilde{f}}\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\mathbf{E}\|\widetilde{f}-f\|^2.$$

Furthermore, we will allow for other sequences γ than just polynomial ones. In particular, we will show that the estimator we are going to develop is adaptive over a range of sequences including both polynomial and exponential ones. The construction of the adaptive estimator and the control of its risk follow nevertheless the outline given in this section. Though, extra terms appear in (3.7), for example, causing additional technical difficulties.

3.2 Minimax optimal estimation

In this section, we develop the minimax theory for the estimation of a circular deconvolution density under unknown error density when two independent samples from Y and ε of size n and m, respectively, are available. A lower bound depending on both sample sizes is derived and it is shown that the orthogonal series estimator $\widehat{f}_k$ defined in (3.4) attains this lower bound up to a constant if k is chosen in an appropriate way.

Here and subsequently, we will refer to any sequence $(a_n)_{n\in\mathbb{Z}}$ as a whole by omitting its index as for example in «the sequence a». Arithmetic operations on sequences are defined element-wise. Furthermore, we will denote by C universal numerical constants and by $C(\cdot)$ constants depending only on the arguments. In both cases, the values of the constants may change from line to line. Moreover, we write $a_n \lesssim b_n$ when $a_n \leqslant C\, b_n$ for all sufficiently large $n \in \mathbb{N}$ and $a_n \sim b_n$ when $a_n \lesssim b_n$ and $b_n \lesssim a_n$ simultaneously. All results in this chapter are derived under the following minimal regularity conditions.

Assumption 3.2 *Let γ, ω, and λ be strictly positive symmetric sequences of weights with $\gamma_0 = \omega_0 = \omega_1 = \lambda_0 = \lambda_1 = 1$ such that $(\omega_n/\gamma_n)_{n\in\mathbb{N}}$ and $(\lambda_n)_{n\in\mathbb{N}}$ are non increasing, respectively with $\Lambda := \sum_{j\in\mathbb{Z}} \lambda_j < \infty$.*

As the densities f and φ are real-valued, the sequences of their Fourier coefficients are symmetric. Therefore, the symmetry assumption is natural. The monotonicity of (ω/γ) ensures that the norm $\|f\|_\omega$ with respect to the weighted norm defined by ω is well defined for all $f \in \mathcal{F}^r_\gamma$. In the context of the illustration section beginning on page 61, this roughly means that we cannot estimate the $(s+1)$-th derivative of a solution f which is only s times differentiable. The monotonicity and the summability of λ are no restrictions, because the error density φ is supposed to lie in L^2 and therefore its Fourier coefficients $[\varphi]_j$ are square summable anyway.

Lower bounds

The next assertion provides a lower bound in case of a known error density, which depends on the size of the Y-sample only. Of course, this lower bound is still valid in case of an unknown error density.

Theorem 3.3 *Assume that we have a sample of n iid. copies of Y. Consider sequences ω, γ, and λ satisfying Assumption 3.2 such that $\sum_{j\in\mathbb{Z}} \gamma_j^{-1} =: \Gamma < \infty$*

and such that $\varphi \in \mathcal{E}^d_\lambda$ *for some* $d \geqslant 1$. *Define for all* $n \geqslant 1$

$$k_n^* := k_n^*(\gamma,\lambda,\omega) := \operatorname*{argmin}_{k\in\mathbb{N}} \Big\{ \max\Big(\frac{\omega_k}{\gamma_k}, \sum_{0<|j|\leqslant k} \frac{\omega_j}{n\lambda_j} \Big) \Big\} \text{ and}$$
$$\psi_n := \psi_n(\gamma,\lambda,\omega) := \max\Big(\frac{\omega_{k_n^*}}{\gamma_{k_n^*}}, \sum_{0<|j|\leqslant k_n^*} \frac{\omega_j}{n\lambda_j} \Big). \tag{3.8}$$

If in addition $\eta := \inf_{n\geqslant 1}\{\psi_n^{-1} \min(\omega_{k_n^*}\gamma_{k_n^*}^{-1}, \sum_{0<|l|\leqslant k_n^*} \omega_l (n\lambda_l)^{-1})\} > 0$, *then for all* $n \geqslant 2$

$$\inf_{\widetilde{f}} \sup_{f\in\mathcal{F}^r_\gamma} \Big\{ \mathbf{E}\|\widetilde{f} - f\|_\omega^2 \Big\} \geqslant \frac{\eta \min(r-1, 1/(8d\Gamma))}{16}\, \psi_n,$$

where the infimum is taken over all possible estimators of f.

The assumption that $\Gamma < \infty$ is used in the proof of the lower bound when constructing candidate densities for the application of Assouad's cube technique. It roughly means that f has to be continuous. The condition $\eta > 0$ ensures that the minimal risk can indeed be found by balancing bias and variance. It is satisfied for a regular behavior of the sequences γ, λ, and ω as defined for example in the illustration section of this chapter.

Remark 3.4 When φ is known, it is natural to consider the orthogonal series estimator $\widetilde{f}_k := 1 + \sum_{1<|j|\leqslant k} (\widehat{[g]}_j / [\varphi]_j)\, e_j$. It is easily seen that for $|j| \leqslant k$, we have $\mathbf{E}[[\widetilde{f}]_j] = [f]_j$ and $\mathbf{V}\mathrm{ar}([\widetilde{f}]_j) \leqslant (n|[\varphi]_j|^2)^{-1}$, while $\mathbf{E}[[\widetilde{f}]_j] = 0$ and $\mathbf{V}\mathrm{ar}([\widetilde{f}]_j) = 0$ for $|j| > k$. Hence, for all $f \in \mathcal{F}^r_\gamma$ and $\varphi \in \mathcal{E}^d_\lambda$ we have

$$\begin{aligned}\mathbf{E}[\|\widetilde{f}_k - f\|_\omega^2] &\leqslant \sum_{|j|>k} \omega_j |[f]_j|^2 + \frac{1}{n} \sum_{0<|j|\leqslant k} \frac{\omega_j}{|[\varphi]_j|^2} \\ &\leqslant (r+d)\ \max\Big(\frac{\omega_k}{\gamma_k}, \sum_{0<|j|\leqslant k} \frac{\omega_j}{n\lambda_j} \Big).\end{aligned}$$

Thus, the choice k_n^* of k from (3.8) realizes the best variance-bias trade-off ψ_n. This shows that when φ is known, $\widetilde{f}_{k_n^*}$ actually attains the rate ψ_n which is hence the minimax optimal one. □

Proof of Theorem 3.3. Defining first the quantities

$$\zeta := \eta \min(r-1, 1/(8d\Gamma)) \quad \text{and} \quad \alpha_n := \psi_n \Big(\sum_{0<|j|\leqslant k_n^*} \omega_j/(\lambda_j n) \Big)^{-1},$$

we consider the function $f := 1 + (\zeta\alpha_n/n)^{1/2} \sum_{0<|j|\leqslant k_n^*} \lambda_j^{-1/2} e_j$. Lemma 3.21 shows that for any $\theta := (\theta_j) \in \{-1,1\}^{2k_n^*}$, the function

$$f_\theta := 1 + \sum_{0<|j|\leqslant k_n^*} \theta_j [f]_j e_j$$

belongs to $\mathcal{F}_\gamma^r$ and is hence a possible candidate of the deconvolution density. For each θ, the Y-density corresponding to the X-density f_θ is given by the convolution $g_\theta := f_\theta * \varphi$. We denote by g_θ^n the joint density of an iid. n-sample from g_θ and by $\mathbf{E}_\theta$ the expectation with respect to the joint density g_θ^n. Furthermore, for $0 < |j| \leqslant k_n^*$ and each θ we introduce $\theta^{(j)}$ by $\theta_l^{(j)} = \theta_l$ for $j \neq l$ and $\theta_j^{(j)} = -\theta_j$. The key argument of this proof is the following reduction scheme based on Assouad's cube technique (c.f. Korostolev and Tsybakov, 1993; Tsybakov, 2004). Let $\widetilde{f}$ denote an estimator of f. We deduce that

$$\begin{aligned}
\sup_{f\in\mathcal{F}_\gamma^r} \mathbf{E}\|\widetilde{f} - f\|_\omega^2 &\geqslant \sup_{\theta\in\{-1,1\}^{2k_n^*}} \mathbf{E}_\theta \|\widetilde{f} - f_\theta\|_\omega^2 \geqslant \frac{1}{2^{2k_n^*}} \sum_{\theta\in\{-1,1\}^{2k_n^*}} \mathbf{E}_\theta \|\widetilde{f} - f_\theta\|_\omega^2 \\
&\geqslant \frac{1}{2^{2k_n^*}} \sum_{\theta\in\{-1,1\}^{2k_n^*}} \sum_{0<|j|\leqslant k_n^*} \omega_j \mathbf{E}_\theta |[\widetilde{f} - f_\theta]_j|^2 \\
&= \frac{1}{2^{2k_n^*}} \sum_{\theta\in\{-1,1\}^{2k_n^*}} \sum_{0<|j|\leqslant k_n^*} \frac{\omega_j}{2} \Big\{ \mathbf{E}_\theta |[\widetilde{f} - f_\theta]_j|^2 + \mathbf{E}_{\theta^{(j)}} |[\widetilde{f} - f_{\theta^{(j)}}]_j|^2 \Big\}.
\end{aligned}$$

Considering the Hellinger affinity $\rho(g_\theta^n, g_{\theta^{(j)}}^n) = \int \sqrt{g_\theta^n}\sqrt{g_{\theta^{(j)}}^n}$, we obtain for any estimator $\widetilde{f}$ of f that

$$\begin{aligned}
\rho(g_\theta^n, g_{\theta^{(j)}}^n) &\leqslant \int \frac{|[\widetilde{f} - f_{\theta^{(j)}}]_j|}{|[f_\theta - f_{\theta^{(j)}}]_j|} \sqrt{g_{\theta^{(j)}}^n}\sqrt{g_\theta^n} + \int \frac{|[\widetilde{f} - f_\theta]_j|}{|[f_\theta - f_{\theta^{(j)}}]_j|} \sqrt{g_\theta^n}\sqrt{g_{\theta^{(j)}}^n} \\
&\leqslant \Big(\int \frac{|[\widetilde{f} - f_{\theta^{(j)}}]_j|^2}{|[f_\theta - f_{\theta^{(j)}}]_j|^2} g_{\theta^{(j)}}^n \Big)^{1/2} + \Big(\int \frac{|[\widetilde{f} - f_\theta]_j|^2}{|[f_\theta - f_{\theta^{(j)}}]_j|^2} g_\theta^n \Big)^{1/2}.
\end{aligned}$$

Rewriting the last estimate we obtain

$$\Big\{ \mathbf{E}_\theta |[\widetilde{f} - f_\theta]_j|^2 + \mathbf{E}_{\theta^{(j)}} |[\widetilde{f} - f_{\theta^{(j)}}]_j|^2 \Big\} \geqslant \frac{1}{2} |[f_\theta - f_{\theta^{(j)}}]_j|^2 \rho^2(g_\theta^n, g_{\theta^{(j)}}^n).$$

Bounding the Hellinger affinity from below by $1/4$ using Lemma 3.22 shows that for $n \geqslant 2$ we have

$$\Big\{ \mathbf{E}_\theta |[\widetilde{f} - f_\theta]_j|^2 + \mathbf{E}_{\theta^{(j)}} |[\widetilde{f} - f_{\theta^{(j)}}]_j|^2 \Big\} \geqslant \frac{\zeta\alpha_n}{4\lambda_j n}.$$

Combining the last lower bound and the reduction scheme yields

$$\sup_{f\in\mathcal{F}_\gamma^r} \mathbf{E}\|\widetilde{f} - f\|_\omega^2 \geqslant \frac{1}{2^{2k_n^*}} \sum_{\theta\in\{-1,1\}^{2k_n^*}} \sum_{0<|j|\leqslant k_n^*} \frac{\omega_j}{2} \frac{\zeta\,\alpha_n}{4\lambda_j n} = \frac{\zeta}{8} \alpha_n \sum_{0<|j|\leqslant k_n^*} \frac{\omega_j}{\lambda_j n}.$$

Hence, substituting the definitions of ζ and α_n we obtain the lower bound given in the theorem. □

Observe that in case $r = 1$, the lower bound is equal to zero, because in this situation the set $\mathcal{F}_\gamma^r$ reduces to a singleton containing the uniform density only. In the next theorem we state a lower bound characterizing the additional complexity due to the unknown error density, which depends only on the error sample size.

Theorem 3.5 *Assume* (3.2) *and let* ω, γ, *and* λ *be sequences satisfying Assumption 3.2. For all* $m \geqslant 2$, *let*

$$\kappa_m := \kappa_m(\gamma, \lambda, \omega) := \max_{j \in \mathbb{N}} \Big\{ \omega_j \gamma_j^{-1} \min\Big(1, \frac{1}{m\lambda_j})\Big)\Big\}. \tag{3.9}$$

If in addition there exists a density in $\mathcal{E}_\lambda^{\sqrt{d}}$ *which is bounded from below by* $1/2$, *then, for all* $m \geqslant 2$

$$\inf_{\widetilde{f}} \sup_{f \in \mathcal{F}_\gamma^r} \sup_{\varphi \in \mathcal{E}_\lambda^d} \Big\{ \mathbf{E} \|\widetilde{f} - f\|_\omega^2 \Big\} \geqslant \frac{\min(r-1, 1) \min(1/(4d), (1 - d^{-1/4})^2)}{4\sqrt{d}}\, \kappa_m,$$

where the infimum is taken over all possible estimators of f.

It is easily seen that $\mathcal{E}_\lambda^{\sqrt{d}}$ contains a density which is bounded from below by $1/2$ if $\ell := \sum_{j \in \mathbb{Z}} \lambda_j^{-1/2} < \infty$ and $\sqrt{d} \geqslant \max(4\ell^2, 1)$. It is worth to note that in case $d = 1$, the set $\mathcal{E}_\lambda^d$ of possible error densities reduces to a singleton, and hence the lower bound is equal to zero.

The proof of the last assertion is inspired by a proof given in Neumann (1997), where a similar lower bound for deconvolution on the real line is shown when both densities f and φ are ordinary smooth, i.e. when γ and λ have polynomial decay.

Proof of Theorem 3.5. For each $\theta \in \{-1, 1\}$, we construct an error density φ_θ in $\mathcal{E}_\lambda^d$ and a deconvolution density $f_\theta \in \mathcal{F}_\gamma^r$, such that $g_\theta := f_\theta * \varphi_\theta$ satisfies $g_1 = g_{-1}$. To this end, define

$$k_m^* := \operatorname*{argmax}_{|j|>0} \{\omega_j \gamma_j^{-1} \min(1, m^{-1} \lambda_j^{-1})\}$$

and $\alpha_m := \zeta \min(1, m^{-1/2} \lambda_{k_m^*}^{-1/2})$ with $\zeta := \min(1/(2\sqrt{d}), (1 - d^{-1/4}))$. Observe that

$$\begin{aligned} & 1 \geqslant (1 - \alpha_m)^2 \geqslant (1 - (1 - 1/d^{1/4}))^2 \geqslant 1/d^{1/2} \\ \text{and} \quad & 1 \leqslant (1 + \alpha_m)^2 \leqslant (1 + (1 - 1/d^{1/4}))^2 = (2 - 1/d^{1/4})^2 \leqslant d^{1/2}, \end{aligned}$$

which implies $1/d^{1/2} \leqslant (1+\theta\alpha_m)^2 \leqslant d^{1/2}$ for $\theta \in \{-1,1\}$. These inequalities will be used below without further reference. By assumption there is a density $\varphi \in \mathcal{E}_\lambda^{\sqrt{d}}$ such that $\varphi \geqslant 1/2$. For each θ, let

$$f_\theta := 1 + (1-\theta\alpha_m)\frac{\min(\sqrt{r-1},1)}{d^{1/4}}\gamma_{k_m^*}^{-1/2} e_{k_m^*} \quad \text{and} \quad \varphi_\theta := \varphi + \theta\alpha_m [\varphi]_{k_m^*} e_{k_m^*}.$$

By Lemma 3.23, we have $f_\theta \in \mathcal{F}_\gamma^r$ and $\varphi_\theta \in \mathcal{E}_\lambda^d$. Moreover, it is easily verified that

$$g_\theta = 1 + (1-\alpha_m^2)\frac{\min(\sqrt{r-1},1)}{d^{1/4}}\gamma_{k_m^*}^{-1/2}[\varphi]_{k_m^*} e_{k_m^*}$$

and hence $g_1 = g_{-1}$. We denote by g_θ^n the joint density of an iid. n-sample from g_θ and φ_θ^m the joint density of an iid. m-sample from φ_θ. Since the samples are independent from each other, $p_\theta := g_\theta^n \varphi_\theta^m$ is the joint density of all observations and we denote by $\mathbf{E}_\theta$ the expectation with respect to p_θ. Applying a reduction scheme we deduce that for each estimator $\widetilde{f}$ of f

$$\begin{aligned}\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d} \mathbf{E}\|\widetilde{f}-f\|_\omega^2 &\geqslant \max_{\theta\in\{-1,1\}} \mathbf{E}_\theta\|\widetilde{f}-f_\theta\|_\omega^2 \\ &\geqslant \frac{1}{2}\Big\{\mathbf{E}_1\|\widetilde{f}-f_1\|_\omega^2 + \mathbf{E}_{-1}\|\widetilde{f}-f_{-1}\|_\omega^2\Big\}.\end{aligned}$$

As in the proof of Theorem 3.3, employing the Hellinger affinity $\rho(p_1,p_{-1})$ we obtain for any estimator $\widetilde{f}$ of f that

$$\Big\{\mathbf{E}_1\|\widetilde{f}-f_1\|_\omega^2 + \mathbf{E}_{-1}\|\widetilde{f}-f_1\|_\omega^2\Big\} \geqslant \frac{1}{2}\|f_1-f_{-1}\|_\omega^2\rho^2(p_1,p_{-1}) \geqslant \frac{1}{8}\|f_1-f_{-1}\|_\omega^2,$$

where the last inequality follows by Lemma 3.24. Moreover, we have

$$\begin{aligned}\|f_1-f_{-1}\|^2 &= 4\alpha_m^2\omega_{k_m^*}\gamma_{k_m^*}^{-1}\frac{(r-1)\wedge 1}{d^{1/2}} \\ &= 4\frac{(r-1)\wedge 1}{d^{1/2}}\zeta^2\omega_{k_m^*}\gamma_{k_m^*}^{-1}\min\Big(1,\frac{1}{m\lambda_{k_m^*}}\Big).\end{aligned}$$

Combining the last lower bound, the reduction scheme and the definition of k_m^* implies the result of the theorem. □

Finally, by combination of both lower bounds we obtain the next corollary.

Corollary 3.6 *Under the assumptions of Theorem 3.3 and 3.5 for all $n, m \geqslant 2$*

$$\inf_{\widetilde{f}}\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\Big\{\mathbf{E}\|\widetilde{f}-f\|_\omega^2\Big\} \geqslant C(\eta,r,d,\Gamma)\ \max(\psi_n,\kappa_m).$$

Upper bound

In the next theorem and in all subsequent results, we will suppose observations according to (3.2). First, we summarize sufficient conditions to ensure the optimality of the orthogonal series estimator $\widehat{f}_k$ defined in (3.4) provided the dimension parameter k is chosen appropriately. We use the value k_n^* defined in (3.8) which, though it obviously involves the sequences ω, γ, and λ, surprisingly does not depend on the ε-sample size m. Under this choice, the estimator attains the lower bound given in Corollary 3.6 up to a constant and is hence minimax-optimal.

Theorem 3.7 *Under Assumption 3.2, we have for all $n, m \geqslant 1$ that*

$$\sup_{f \in \mathcal{F}_\gamma^r} \sup_{\varphi \in \mathcal{E}_\lambda^d} \left\{ \mathbf{E} \| \widehat{f}_{k_n^*} - f \|_\omega^2 \right\} \leqslant C \left\{ (d+r)\, \psi_n + d\, r\, \kappa_m \right\}.$$

Proof. We begin our proof with the observation that

$$\mathbf{V}\mathrm{ar}(\widehat{[g]}_j) \leqslant n^{-1} \quad \text{and} \quad \mathbf{V}\mathrm{ar}(\widehat{[\varphi]}_j) \leqslant m^{-1}$$

for all $j \in \mathbb{Z}$. Moreover, by virtue of Theorem A.3 from the appendix, there exists a constant $C > 0$ such that $\mathbf{E}|\widehat{[\varphi]}_j - [\varphi]_j|^4 \leqslant C/m^2$ for all $j \in \mathbb{Z}$ and $m \in \mathbb{N}$. These results are used below without further reference. Define now

$$\widetilde{f} := 1 + \sum_{0 < |j| \leqslant k_n^*} [f]_j \mathbf{1}\{|\widehat{[\varphi]}_j|^2 \geqslant 1/m\} e_j$$

and decompose the risk into two terms,

$$\mathbf{E} \| \widehat{f}_{k_n^*} - f \|_\omega^2 \leqslant 2\mathbf{E} \| \widehat{f}_{k_n^*} - \widetilde{f} \|_\omega^2 + 2\mathbf{E} \| \widetilde{f} - f \|_\omega^2 =: A + B, \tag{3.10}$$

which we bound separately. Consider first A which we decompose further,

$$\begin{aligned} \mathbf{E} \| \widehat{f}_{k_n^*} - \widetilde{f} \|_\omega^2 &\leqslant 2 \sum_{0 < |j| \leqslant k_n^*} \omega_j \mathbf{E} \left[\frac{|\widehat{[g]}_j - [g]_j|^2}{|\widehat{[\varphi]}_j|^2} \mathbf{1}_{|\widehat{[\varphi]}_j|^2 \geqslant 1/m} \right] \\ &\quad + 2 \sum_{0 < |j| \leqslant k_n^*} \omega_j |[f]_j|^2 \mathbf{E} \left[\frac{|\widehat{[\varphi]}_j - [\varphi]_j|^2}{|\widehat{[\varphi]}_j|^2} \mathbf{1}\{|\widehat{[\varphi]}_j|^2 \geqslant 1/m\} \right] =: A_1 + A_2. \end{aligned}$$

Using the elementary inequality $|[\varphi]_j / \widehat{[\varphi]}_j|^2 \leqslant 2|[\varphi]_j / \widehat{[\varphi]}_j - 1|^2 + 2$, the independence of $\widehat{\varphi}$ and $\widehat{g}$, and $\varphi \in \mathcal{E}_\lambda^d$ together with the definition of ψ_n given in (3.8),

we obtain

$$A_1 \leqslant 4 \sum_{0<|j|\leqslant k_n^*} \omega_j \Big\{ \frac{m\,\mathbf{V}\mathrm{ar}(\widehat{[g]}_j)\,\mathbf{V}\mathrm{ar}(\widehat{[\varphi]}_j)}{|[\varphi]_j|^2} + \frac{\mathbf{V}\mathrm{ar}(\widehat{[g]}_j)}{|[\varphi]_j|^2} \Big\}$$
$$\leqslant 8d \sum_{0<|j|\leqslant k_n^*} \frac{\omega_j}{n\lambda_j} \leqslant 8d\psi_n.$$

Moreover, we have

$$\mathbf{E}\left[\frac{|\widehat{[\varphi]}_j - [\varphi]_j|^2}{|\widehat{[\varphi]}_j|^2} \mathbf{1}\{|\widehat{[\varphi]}_j|^2 \geqslant 1/m\} \right] \leqslant \frac{2m\mathbf{E}|\widehat{[\varphi]}_j - [\varphi]_j|^4}{|[\varphi]_j|^2} + \frac{2\,\mathbf{V}\mathrm{ar}(\widehat{[\varphi]}_j)}{|[\varphi]_j|^2}$$
$$\leqslant \frac{C}{m|[\varphi]_j|^2} \leqslant \frac{C\,d}{m\lambda_j}$$

and $$\mathbf{E}\left[\frac{|\widehat{[\varphi]}_j - [\varphi]_j|^2}{|\widehat{[\varphi]}_j|^2} \mathbf{1}\{|\widehat{[\varphi]}_j|^2 \geqslant 1/m\} \right] \leqslant 1,$$

where we have used the elementary inequality and $\varphi \in \mathcal{E}_\lambda^d$ again. By combination of both bounds together with $f \in \mathcal{F}_\gamma^r$ and the definition of κ_m given in (3.9) we obtain

$$A_2 \leqslant Cd \sum_{0<|j|\leqslant k_n^*} \omega_j |[f]_j|^2 \min(1, \frac{1}{m\lambda_j}) \leqslant Cdr\ \kappa_m.$$

Consider now B which we decompose further into

$$\mathbf{E}\|\widetilde{f} - f\|_\omega^2 = \sum_{0<|j|} \omega_j |[f]_j|^2 (1 - \mathbf{1}\{0 < |j| \leqslant k_n^*\}\mathbf{1}\{|\widehat{[\varphi]}_j|^2 \geqslant 1/m\})^2$$
$$= \sum_{|j|>k_n^*} \omega_j |[f]_j|^2 + \sum_{0<|j|\leqslant k_n^*} \omega_j |[f]_j|^2 \mathbf{P}\Big(|\widehat{[\varphi]}_j|^2 < 1/m\Big) =: B_1 + B_2,$$

where $B_1 \leqslant \|f\|_\gamma^2 \omega_{k_n^*} \gamma_{k_n^*}^{-1} \leqslant r\psi_n$ because $f \in \mathcal{F}_\gamma^r$. Moreover, $B_2 \leqslant 4dr\kappa_m$ by using that

$$\mathbf{P}\Big(|\widehat{[\varphi]}_j|^2 < 1/m\Big) \leqslant 4d \min(1, \frac{1}{m\lambda_j}), \tag{3.11}$$

which we will show below. The result of the theorem follows now by combination of the decomposition (3.10) and the estimates of A_1, A_2, B_1 and B_2.

To conclude, let us prove (3.11). If $|[\varphi]_j|^2 \geqslant 4/m$, then we deduce by employing Chebychev's inequality that

$$\mathbf{P}(|\widehat{[\varphi]}_j|^2 < 1/m) \leqslant \mathbf{P}(|\widehat{[\varphi]}_j/[\varphi]_j| < 1/2) \leqslant \mathbf{P}(|\widehat{[\varphi]}_j - [\varphi]_j| > |[\varphi]_j|/2)$$
$$\leqslant 4\frac{\mathbf{V}\mathrm{ar}(\widehat{[\varphi]}_j)}{|[\varphi]_j|^2} \leqslant 4d/(m\lambda_j).$$

On the other hand, in case $|[\varphi]_j|^2 < 4/m$ the estimate $\mathbf{P}(|\widehat{[\varphi]}_j|^2 < 1/m) \leqslant 4d/(m\lambda_j)$ holds too since $1 \leqslant 4/(m|[\varphi]_j|^2) \leqslant 4d/(m\lambda_j)$. Combining the last estimates and $\mathbf{P}(|\widehat{[\varphi]}_j|^2 < 1/m) \leqslant 1$ we obtain (3.11), which completes the proof. □

Note that under slightly stronger conditions on the sequences ω, γ, and λ than Assumption 3.2, it can be shown that in case of equally large samples from Y and ε we have always the rate as in case of known error density. However, below we show that in special cases the required ε-sample size can be much smaller than the Y-sample size.

Illustration: estimation of derivatives

We will illustrate our results considering classical smoothness assumptions. As far as the deconvolution density f is concerned, recall that the class $\mathcal{F}_\gamma^r$ is a subset of the Sobolev space of p-times differentiable periodic functions if $\gamma_j \sim |j|^{2p}$ (Neubauer (1988a,b)). We call this case *ordinary smooth*. Moreover, up to a constant, for any function $h \in \mathcal{F}_\gamma^r$, the weighted norm $\|h\|_\omega$ with $\omega_j \sim j^{2s}$ equals the L^2-norm of the s-th weak derivative $h^{(s)}$ for each integer $0 \leqslant s \leqslant p$. By virtue of this relation, the results in the previous section imply also a lower as well as an upper bound of the L^2-risk for the estimation of the s-th weak derivative of f. If, on the contrary, $\gamma_j \sim \exp(|j|^{2p})$ with $p > 1$, then $\mathcal{F}_\gamma^r$ is a class of analytic functions (Kawata (1972)). We refer to this situation as *super smooth*.

As for the error densities, we consider two special cases corresponding to a regular decay of their Fourier coefficients. The error density is called *ordinary smooth* if $\lambda_j \sim |j|^{-2a}$ for some $a > 1/2$ and *super smooth* if $\lambda_j \sim \exp(-|j|^{2a})$ for some $a > 0$.

We are going to consider the following three situations: In the cases **[o-o]** and **[s-o]**, the error density is ordinary smooth, while the deconvolution density falls in the ordinary or super smooth case, respectively. **[o-s]** is the opposite case of **[s-o]**.

It is easily seen that in all these cases the minimal regularity conditions given in Assumption 3.2 and the additional conditions used in Theorems 3.3 and 3.5 translate to simple restrictions on p, a, and s which are given in the proposition below. Roughly speaking, they imply that both the deconvolution density and the error density are at least continuous. The lower bound presented in the next assertion follows now directly from Corollary 3.6.

Proposition 3.8

[o-o] *For $p > 1/2$, $a > 1$, and $0 \leqslant s \leqslant p$, we have*

$$\inf_{\widetilde{f}^{(s)}} \sup_{f\in\mathcal{F}_\gamma^r} \sup_{\varphi\in\mathcal{E}_\lambda^d} \left\{\mathbf{E}\|\widetilde{f}^{(s)} - f^{(s)}\|^2\right\} \gtrsim n^{-2(p-s)/(2p+2a+1)} + m^{-((p-s)\wedge a)/a}.$$

[s-o] *For $p > 0$, $a > 1$, and $s \geqslant 0$, we have*

$$\inf_{\widetilde{f}^{(s)}} \sup_{f\in\mathcal{F}_\gamma^r} \sup_{\varphi\in\mathcal{E}_\lambda^d} \left\{\mathbf{E}\|\widetilde{f}^{(s)} - f^{(s)}\|^2\right\} \gtrsim n^{-1}(\log n)^{(2a+2s+1)/(2p)} + m^{-1}.$$

[o-s] *For $p > 1/2$, $a > 0$, and $0 \leqslant s \leqslant p$, we have*

$$\inf_{\widetilde{f}^{(s)}} \sup_{f\in\mathcal{F}_\gamma^r} \sup_{\varphi\in\mathcal{E}_\lambda^d} \left\{\mathbf{E}\|\widetilde{f}^{(s)} - f^{(s)}\|^2\right\} \gtrsim (\log n)^{-(p-s)/a} + (\log m)^{-(p-s)/a}.$$

Proof. Since for each $0 \leqslant s \leqslant p$ we have $\mathbf{E}\|\widetilde{f}^{(s)} - f^{(s)}\|^2 \sim \mathbf{E}\|\widetilde{f} - f\|_\omega^2$ we can apply the general result given in Corollary 3.6. In both cases the additional conditions formulated in Theorem 3.3 and 3.5 are easily verified. Therefore, it is sufficient to evaluate the lower bounds ψ_n and κ_m given in (3.8) and (3.9), respectively. Note that the optimal dimension parameter $k_n^* := \operatorname{argmin}_{j\in\mathbb{N}}\{\max(\frac{\omega_j}{\gamma_j}, \sum_{0<|l|\leqslant j} \frac{\omega_l}{n\lambda_l})\}$ satisfies $n\omega_{k_n^*}/\gamma_{k_n^*} \sim \sum_{0<|l|\leqslant k_n^*} \omega_l/\lambda_l$ since both sequences (γ_j/ω_j) and $(\sum_{0<|l|\leqslant j} \frac{\omega_l}{n\lambda_l})$ are non increasing.

[o-o] The well-known approximation $\sum_{j=1}^m j^r \sim m^{r+1}$ for $r > 0$ implies $(\gamma_{k_n^*}/\omega_{k_n^*}) \sum_{0<|l|\leqslant k_n^*} \omega_l/\lambda_l \sim (k_n^*)^{2a+2p+1}$. It follows that $k_n^* \sim n^{1/(2p+2a+1)}$ and the first lower bound is $\psi_n \sim n^{-(2p-2s)/(2p+2a+1)}$. Moreover, we have $\kappa_m \sim m^{-([p-s]\wedge a)/a}$, since the minimum in $\kappa_m = \sup_{j\in\mathbb{Z}}\{|j|^{-2(p-s)} \min(1, |j|^{2a}/m)\}$ is equal to one for $|j| \geqslant m^{1/2a}$ and $|j|^{-2(p-s)}$ is non increasing.

[s-o] Approximating the sum in the same way as above, we obtain

$$(\gamma_{k_n^*}/\omega_{k_n^*}) \sum_{0<|l|\leqslant k_n^*} \omega_l/\lambda_l \sim (k_n^*)^{2a+1} \exp({k_n^*}^{2p})$$

and thus $k_n^* \sim (\log n)^{1/(2p)}$. The resulting rate is $\psi_n \sim n^{-1}(\log n)^{(2a+2s+1)/(2p)}$. Furthermore, we have $\kappa_m \sim m^{-1}$, since the supremum is taken over the expression $j^{2s}\exp(-j^{2p}) \min(1, j^{2a}/m)$ which takes its maximum at the border because of the dominating exponential term.

[o-s] Applying Laplace's Method (cf. chapter 3.7 in Olver (1974)) we have $(\gamma_{k_n^*}/\omega_{k_n^*}) \sum_{0<|l|\leqslant k_n^*} \omega_l/\lambda_l \sim (k_n^*)^{2p+((2a-1)\vee 0)} \exp(|k_n^*|^{2a})$ which implies that

$k_n^* \sim (\log n)^{1/(2a)}$ and that the first lower bound can be rewritten as $\psi_n \sim (\log n)^{-(p-s)/a}$. Furthermore, we have $\kappa_m \sim (\log m)^{-(p-s)/a}$ since the minimum in

$$\kappa_m = \sup_{j\in\mathbb{Z}}\{|j|^{-2(p-s)} \min(1, \exp(|j|^{2a})/m)\}$$

is equal to one for $|j| \geqslant (\log m)^{(1/2a)}$ and $|j|^{-2(p-s)}$ is non increasing. Consequently, the lower bounds in Proposition 3.8 follow by Corollary 3.6. □

The derivative $f^{(s)}$ can be estimated by the s-th weak derivative[2] of the estimator $\widehat{f}_k$ defined in (3.4), with k to be specified below. Given the exponential basis $\{e_j\}_{j\in\mathbb{Z}}$, we recall that for each integer $0 \leqslant s \leqslant p$ the s-th weak derivative of the estimator $\widehat{f}_k$ can be written as

$$\widehat{f}_k^{(s)} = \sum_{j\in\mathbb{Z}} (2i\pi j)^s \widehat{[f_k]}_j e_j.$$

As an immediate consequence of Theorem 3.7, the rates of the lower bound given by Proposition 3.8 are attained for $k = k_n^*$, which is summarized in the next result. We have thus proved that these rates are optimal and the proposed estimator $\widehat{f}_{k_n^*}^{(s)}$ is minimax optimal in both cases. Furthermore, it is of interest to characterize the minimal size m of the additional sample from ε needed to attain the same rate as in case of a known error density. Hence, we let the ε-sample size depend on the Y-sample size n, too.

Proposition 3.9

[o-o] *For $p > 1/2$, $a > 1$, and $0 \leqslant s \leqslant p$ with $k_n^* \sim n^{1/(2p+2a+1)}$, we have*

$$\sup_{f\in\mathcal{F}_\gamma^r} \sup_{\varphi\in\mathcal{E}_\lambda^d} \left\{\mathbf{E}\|\widehat{f}_{k_n^*}^{(s)} - f^{(s)}\|^2\right\} \lesssim n^{-2(p-s)/(2p+2a+1)} + m^{-((p-s)\wedge a)/a}$$

and for any non decreasing sequence $(m_n)_{n\geqslant 1}$ follows as $n \to \infty$

$$\sup_{f\in\mathcal{F}_\gamma^r} \sup_{\varphi\in\mathcal{E}_\lambda^d} \left\{\mathbf{E}\|\widehat{f}_{k_n^*}^{(s)} - f^{(s)}\|^2\right\} = \begin{cases} O(n^{-2(p-s)/(2p+2a+1)}) & \text{if } n^{2((p-s)\vee a)/(2p+2a+1)} = O(m_n) \\ O(m_n^{-((p-s)\wedge a)/a}) & \text{otherwise.} \end{cases}$$

[2] cf. Definition A.2 in the appendix

[s-o] *For $p > 0$, $a > 1$, and $s \geqslant 0$ with $k_n^* \sim (\log n)^{1/(2p)}$, we have*

$$\sup_{f \in \mathcal{F}_\gamma^r} \sup_{\varphi \in \mathcal{E}_\lambda^d} \left\{ \mathbf{E} \| \widehat{f}_{k_n^*}^{(s)} - f^{(s)} \|^2 \right\} \lesssim n^{-1} (\log n)^{(2a+2s+1)/(2p)} + m^{-1}$$

and for any non decreasing sequence $(m_n)_{n \geqslant 1}$ follows as $n \to \infty$

$$\sup_{f \in \mathcal{F}_\gamma^r} \sup_{\varphi \in \mathcal{E}_\lambda^d} \left\{ \mathbf{E} \| \widehat{f}_{k_n^*}^{(s)} - f^{(s)} \|^2 \right\}$$
$$= \begin{cases} O(n^{-1} (\log n)^{(2a+2s+1)/(2p)}) & \text{if } n (\log n)^{-(2a+2s+1)/(2p)} = O(m_n) \\ O(m_n^{-1}) & \text{otherwise.} \end{cases}$$

[o-s] *For $p > 1/2$, $a > 0$, and $0 \leqslant s \leqslant p$ with $k_n^* \sim (\log n)^{1/(2a)}$, we have*

$$\sup_{f \in \mathcal{F}_\gamma^r} \sup_{\varphi \in \mathcal{E}_\lambda^d} \left\{ \mathbf{E} \| \widehat{f}_{k_n^*}^{(s)} - f^{(s)} \|^2 \right\} \lesssim (\log n)^{-(p-s)/a} + (\log m)^{-(p-s)/a}$$

and for any non decreasing sequence $(m_n)_{n \geqslant 1}$ follows as $n \to \infty$

$$\sup_{f \in \mathcal{F}_\gamma^r} \sup_{\varphi \in \mathcal{E}_\lambda^d} \left\{ \mathbf{E} \| \widehat{f}_{k_n^*}^{(s)} - f^{(s)} \|^2 \right\}$$
$$= \begin{cases} O((\log n)^{-(p-s)/a}) & \text{if } \log n = O(\log m_n) \\ O((\log m_n)^{-(p-s)/a}) & \text{otherwise.} \end{cases}$$

Proof. This result follows from Theorem 3.7 and Proposition 3.8. □

In the case **[o-o]** we obtain the rate of known error density whenever the growth condition $n^{2((p-s) \vee a)/(2p+2a+1)} = O(m_n)$ is satisfied which is much less than $m_n = n$. This is even more visible in the case **[o-s]**, here the rate of known error density is attained even if $m_n = n^r$ for arbitrary small $r > 0$. Moreover, we emphasize the influence of the parameter a which characterizes the rate of the decay of the Fourier coefficients of the error density φ. Since a smaller value of a leads to faster rates of convergence, this parameter is often called *degree of ill-posedness* (cf. Natterer (1984)).

3.3 Adaptive estimation

Our objective is to construct an adaptive estimator of the deconvolution density f. Adaptation means that in spite of an unknown error density in $\mathcal{E}_\lambda^d$, the estimator should attain the optimal rate of convergence $\max(\psi_n, \kappa_m)$ over the ellipsoid $\mathcal{F}_\gamma^r$ for a wide range of different weight sequences γ and λ.

In a first step, we suppose that φ is known, but γ and r are unknown. In what follows, the orthogonal series estimator $\widehat{f}_k$ defined in (3.4) is considered and a procedure to choose the dimension parameter k based on a model selection approach via penalization is constructed. This partially adaptive choice $\widetilde{k}$ will only involve the data and the error density φ. In a second step, we replace φ by its empirical version and thus dispense with any knowledge about φ. Doing so, we obtain a fully adaptive choice $\widehat{k}$ of the dimension parameter.

The construction of the adaptive estimators follows the general model reduction scheme presented in Barron et al. (1999) which we have already discussed in the introductory example in Section 3.1. In particular, we use the same contrast as in the example. However, the choice of the penalty term and the number of models over which to minimize the penalized models present cannot be performed in a deterministic manner. This is a major technical difficulty.

Partially adaptive estimation knowing φ

First, we introduce sequences which are used below.

Definition 3.10 *For all $n, m \geqslant 1$ and $k \geqslant 0$, define*

(i) $\Delta_k := \Delta_k(\varphi) := \max\limits_{-k \leqslant j \leqslant k} \frac{\omega_j}{|[\varphi]_j|^2}$ *and* $\delta_k := \delta_k(\varphi) := 2\,k\,\Delta_k \frac{\log(\Delta_k \vee (k+2))}{\log(k+2)}$;

(ii) given $\omega_k^+ := \max_{0 \leqslant j \leqslant k} \omega_j$ *and* $N_n^\circ := \operatorname{argmax}_{1 \leqslant N \leqslant n}\{\omega_N^+ \leqslant n\}$, *let*

$$N_n := N_n(\varphi) := \operatorname*{argmin}_{1 \leqslant j \leqslant N_n^\circ} \left\{ \frac{|[\varphi]_j|^2}{j\omega_j^+} \leqslant \frac{\log(n+2)}{n} \right\} - 1,$$

defining further $b_m := (8\log(\log(m+20)))^{-1}$, *let*

$$M_m := M_m(\varphi) := \operatorname*{argmin}_{1 \leqslant j \leqslant m} \left\{ |[\varphi]_j|^2 \leqslant m^{-1+b_m} \right\} - 1;$$

with $N_n := N_n^\circ$ *and* $M_m := m$ *when the respective set in the* argmin *is empty.*

We can now define a partially adaptive choice of the dimension parameter k, namely

$$\widetilde{k} := \operatorname*{argmin}_{0 \leqslant k \leqslant (N_n \wedge M_m)} \left\{ -\|\widehat{f}_k\|_\omega^2 + 60\,\frac{\delta_k}{n} \right\}, \tag{3.12}$$

which obviously depends on the data and on the error density φ only. The fully adaptive estimator will be obtained below by introducing the empirical versions of δ, N, and M.

However, for a fixed φ, one could now derive an upper risk bound for the partially adaptive estimator $\widehat{f_{\widetilde{k}}}$, which would depend on δ, N, and M. As we wish rather to obtain a uniform upper risk bound over the class $\mathcal{E}^d_\lambda$, we now redefine the objects above referring only to the weight sequence λ and the constant d.

Definition 3.11 *Let ω^+, N°, and b as in Definition 3.10.*

(i) For all $k \geqslant 0$, define $\Delta^\lambda_k := \max_{-k\leqslant j\leqslant k} \omega_j/\lambda_j$ and

$$\delta^\lambda_k := 2\,k\,\Delta^\lambda_k\,\frac{\log(\Delta^\lambda_k \vee (k+2))}{\log(k+2)}.$$

(ii) Define two sequences N^λ and M^λ as follows,

$$N^\lambda_n := \operatorname*{argmax}_{1\leqslant j\leqslant N^\circ_n}\left\{\frac{\lambda_j}{j\omega^+_j} \geqslant \frac{4d\log(n+2)}{n}\right\},$$
$$M^\lambda_m := \operatorname*{argmax}_{1\leqslant j\leqslant m}\left\{\lambda_j \geqslant 4d\,m^{-1+b_m}\right\}.$$

If the set in the argmax *is empty, we set $N^\lambda_n := 0$ or $M^\lambda_m := 0$, respectively.*

(iii) Define two sequences N^u and M^u as follows,

$$N^u_n := N^u_n(\lambda) := \operatorname*{argmin}_{1\leqslant j\leqslant n}\left\{\frac{\lambda_j}{j\omega^+_j} < \frac{\log(n+2)}{4d\,n}\right\} - 1,$$
$$M^u_m := M^u_m(\lambda) := \operatorname*{argmin}_{1\leqslant j\leqslant m}\left\{\lambda_j < \frac{m^{-1+b_m}}{4d}\right\} - 1,$$

If the set in the argmin *is empty, we set $N^u_n := n$ or $M^u_m := m$, respectively.*

(iv) Let $\Sigma : \mathbb{R} \to \mathbb{R}$ be a non decreasing function such that for all $C > 0$

$$\sum_{k\geqslant 1} C\,\Delta^\lambda_k \exp\Big(-\frac{k\log(\Delta^\lambda_k \vee (k+2))}{3\,C\,\log(k+2)}\Big) \leqslant \Sigma(C) < \infty.$$

It is easy to see that there exists always a function Σ satisfying the defining condition. Moreover, we show in Lemma 3.25 below that the sequences defined above satisfy $N^\lambda_n \leqslant N_n \leqslant N^u_n$ and $M^\lambda_m \leqslant M_m \leqslant M^u_m$ for all $n, m \in \mathbb{N}$. In the illustration below we compute these objects explicitly.

Theorem 3.12 *Let $\zeta_d := \log(3\,d)/\log(d)$. Under Assumption 3.2, we have for all $n, m \geqslant 1$*

$$\sup_{f\in\mathcal{F}^r_\gamma}\sup_{\varphi\in\mathcal{E}^d_\lambda}\left\{\mathbf{E}\|\widehat{f}_{\widetilde{k}} - f\|^2_\omega\right\} \leqslant C\left\{(r + d\,\zeta_d)\min_{0\leqslant k\leqslant(N^\lambda_n\wedge M^\lambda_m)}\left[\max\left(\frac{\omega_k}{\gamma_k}, \frac{\delta^\lambda_k}{n}\right)\right] + r\,d\,\kappa_m\right\} + C(r, d, \Lambda, \Sigma)\left[\frac{1}{m} + \frac{1}{n}\right].$$

Before proving this theorem, we define and recall some notation. Given a function $u \in L^2[0,1]$ we denote by $[u]$ the infinite vector of Fourier coefficients $[u]_j := \langle u, e_j\rangle$. In particular we use the notations

$$\widehat{f}_k = \sum_{j=-k}^{k}\frac{\widehat{[g]}_j}{\widehat{[\varphi]}_j}\mathbf{1}\{|\widehat{[\varphi]}_j|^2 \geqslant 1/m\}e_j, \quad \widetilde{f}_k := \sum_{j=-k}^{k}\frac{\widehat{[g]}_j}{[\varphi]_j}e_j, \quad f_k := \sum_{j=-k}^{k}\frac{[g]_j}{[\varphi]_j}e_j,$$

$$\widehat{\Phi}_u := \sum_{j\in\mathbb{Z}}\frac{[u]_j}{\widehat{[\varphi]}_j}\mathbf{1}\{|\widehat{[\varphi]}_j|^2 \geqslant 1/m\}e_j, \quad \widetilde{\Phi}_u := \sum_{j\in\mathbb{Z}}\frac{[u]_j}{[\varphi]_j}e_j.$$

Furthermore, let $\widehat{g}$ be the function with Fourier coefficients $[\widehat{g}]_j := \widehat{[g]}_j$. Given $0 \leqslant k \leqslant k'$ we have then for all $t \in \mathcal{S}_k := \operatorname{span}\{e_{-k}, \ldots, e_k\}$

$$\langle t, f_{k'}\rangle_\omega = \langle t, \widetilde{\Phi}_g\rangle_\omega = \sum_{j=-k}^{k}\frac{\omega_j[t]_j[g]_j}{[\varphi]_j} = \sum_{j=-k}^{k}\omega_j[t]_j[f]_j = \langle t, f\rangle_\omega,$$

$$\langle t, \widetilde{f}_{k'}\rangle_\omega = \langle t, \widetilde{\Phi}_{\widehat{g}}\rangle_\omega = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=-k}^{k}e_j(-Y_i)\frac{\omega_j[t]_j}{[\varphi]_j} = \langle t, \widetilde{f}_k\rangle_\omega,$$

$$\langle t, \widehat{f}_{k'}\rangle_\omega = \langle t, \widehat{\Phi}_{\widehat{g}}\rangle_\omega = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=-k}^{k}e_j(-Y_i)\frac{\omega_j[t]_j}{\widehat{[\varphi]}_j}\mathbf{1}\{|\widehat{[\varphi]}_j|^2 \geqslant 1/m\} = \langle t, \widehat{f}_k\rangle_\omega.$$

Define the function $\nu = \widehat{g} - g$ with Fourier coefficients $[\nu]_j := \widehat{[g]}_j - [g]_j = \widehat{[g]}_j - \mathbf{E}\widehat{[g]}_j$. Then we have for every $t \in \mathcal{S}_k$

$$\begin{aligned}\langle t, \widehat{\Phi}_{\widehat{g}} - f\rangle_\omega &= \langle t, \widehat{\Phi}_{\widehat{g}} - \widetilde{\Phi}_g\rangle_\omega = \langle t, \widetilde{\Phi}_{\widehat{g}} - \widetilde{\Phi}_g\rangle_\omega + \langle t, \widehat{\Phi}_{\widehat{g}} - \widetilde{\Phi}_{\widehat{g}}\rangle_\omega \\ &= \langle t, \widetilde{\Phi}_\nu\rangle_\omega + \langle t, \widehat{\Phi}_{\widehat{g}} - \widetilde{\Phi}_{\widehat{g}}\rangle_\omega = \langle t, \widetilde{\Phi}_\nu\rangle_\omega + \langle t, \widehat{\Phi}_\nu - \widetilde{\Phi}_\nu\rangle_\omega + \langle t, \widehat{\Phi}_g - \widetilde{\Phi}_g\rangle_\omega.\end{aligned} \tag{3.13}$$

We are now in position to prove the result. The technical lemmas used in the proof can be found in the auxiliary results section at the end of this chapter.

Proof of Theorem 3.12. We define the contrast

$$\Upsilon(t) := \|t\|_\omega^2 - 2\langle t, \widehat{\Phi}_{\widehat{g}}\rangle_\omega, \quad \forall\, t \in L^2[0,1].$$

Obviously it follows for all $t \in \mathcal{S}_k$ that $\Upsilon(t) = \|t - \widehat{f}_k\|_\omega^2 - \|\widehat{f}_k\|_\omega^2$ and, hence

$$\arg\min_{t\in\mathcal{S}_k} \Upsilon(t) = \widehat{f}_k, \quad \forall\, k \geqslant 0. \tag{3.14}$$

Moreover, the adaptive choice of the dimension parameter from (3.12) can be rewritten as

$$\widetilde{k} = \operatorname*{argmin}_{0\leqslant k\leqslant (N_n\wedge M_m)} \left\{ \Upsilon(\widehat{f}_k) + 60\frac{\delta_k}{n} \right\}. \tag{3.15}$$

Let $\mathrm{pen}(k) := 60\delta_k/n$, then for all $1 \leqslant k \leqslant (N_n \wedge M_m)$ we have

$$\Upsilon(\widehat{f}_{\widetilde{k}}) + \mathrm{pen}(\widetilde{k}) \leqslant \Upsilon(\widehat{f}_k) + \mathrm{pen}(k) \leqslant \Upsilon(f_k) + \mathrm{pen}(k),$$

using first (3.15) and then (3.14). This inequality implies

$$\|\widehat{f}_{\widetilde{k}}\|_\omega^2 - \|f_k\|_\omega^2 \leqslant 2\langle \widehat{f}_{\widetilde{k}} - f_k, \widehat{\Phi}_{\widehat{g}}\rangle_\omega + \mathrm{pen}(k) - \mathrm{pen}(\widetilde{k}),$$

and hence, using (3.13), we have for all $1 \leqslant k \leqslant (N_n \wedge M_m)$

$$\begin{aligned}\|\widehat{f}_{\widetilde{k}} - f\|_\omega^2 \leqslant{}& \|f - f_k\|_\omega^2 + \mathrm{pen}(k) - \mathrm{pen}(\widetilde{k}) \\ &+ 2\langle \widehat{f}_{\widetilde{k}} - f_k, \widetilde{\Phi}_\nu\rangle_\omega + 2\langle \widehat{f}_{\widetilde{k}} - f_k, \widehat{\Phi}_\nu - \widetilde{\Phi}_\nu\rangle_\omega + 2\langle \widehat{f}_{\widetilde{k}} - f_k, \widehat{\Phi}_g - \widetilde{\Phi}_g\rangle_\omega.\end{aligned} \tag{3.16}$$

Consider the unit ball $\mathcal{B}_k := \{f \in \mathcal{S}_k \mid \|f\|_\omega \leqslant 1\}$ and, for arbitrary $\tau > 0$ and $t \in \mathcal{S}_k$, the elementary inequality

$$\begin{aligned}2|\langle t, h\rangle_\omega| &\leqslant 2\|t\|_\omega \sup_{t\in\mathcal{B}_k} |\langle t, h\rangle_\omega| \\ &\leqslant \tau\|t\|_\omega^2 + \frac{1}{\tau}\sup_{t\in\mathcal{B}_k} |\langle t, h\rangle_\omega|^2 = \tau\|t\|_\omega^2 + \frac{1}{\tau}\sum_{j=-k}^{k} \omega_j |[h]_j|^2.\end{aligned}$$

Combining the last estimate with (3.16) and $\widehat{f}_{\widetilde{k}} - f_k \in \mathcal{S}_{\widetilde{k}\vee k} \subset \mathcal{S}_{N_n\wedge M_m}$ we obtain

$$\begin{aligned}\|\widehat{f}_{\widetilde{k}} - f\|_\omega^2 \leqslant{}& \|f - f_k\|_\omega^2 + 3\tau\,\|\widehat{f}_{\widetilde{k}} - f_k\|_\omega^2 + \mathrm{pen}(k) - \mathrm{pen}(\widetilde{k}) \\ &+ \frac{1}{\tau}\sup_{t\in\mathcal{B}_{\widetilde{k}\vee k}} |\langle t, \widetilde{\Phi}_\nu\rangle_\omega|^2 + \frac{1}{\tau}\sup_{t\in\mathcal{B}_{(N_n\wedge M_m)}} |\langle t, \widehat{\Phi}_\nu - \widetilde{\Phi}_\nu\rangle_\omega|^2 \\ &+ \frac{1}{\tau}\sup_{t\in\mathcal{B}_{(N_n\wedge M_m)}} |\langle t, \widehat{\Phi}_g - \widetilde{\Phi}_g\rangle_\omega|^2.\end{aligned}$$

Notice that $\|\widehat{f}_{\widetilde{k}} - f_k\|_\omega^2 \leqslant 2\|\widehat{f}_{\widetilde{k}} - f\|_\omega^2 + 2\|f_k - f\|_\omega^2$ and that $\|f - f_k\|_\omega^2 \leqslant r\omega_k/\gamma_k$ for all $f \in \mathcal{F}_\gamma^r$ because ω/γ is non increasing. Setting $\tau := 1/8$, we obtain

$$\begin{aligned}\frac{1}{4}\|\widehat{f}_{\widetilde{k}} - f\|_\omega^2 \leqslant & \frac{7}{4}\left(r\omega_k/\gamma_k\right) + \mathrm{pen}(k) - \mathrm{pen}(\widetilde{k}) \\ & + 8 \sup_{t \in \mathcal{B}_{k \vee \widetilde{k}}} |\langle t, \widetilde{\Phi}_\nu \rangle_\omega|^2 + 8 \sup_{t \in \mathcal{B}_{(N_n \wedge M_m)}} |\langle t, \widehat{\Phi}_\nu - \widetilde{\Phi}_\nu \rangle_\omega|^2 \qquad (3.17) \\ & + 8 \sup_{t \in \mathcal{B}_{(N_n \wedge M_m)}} |\langle t, \widehat{\Phi}_g - \widetilde{\Phi}_g \rangle_\omega|^2.\end{aligned}$$

Defining the event

$$\Omega_q := \left\{ \forall\, 0 \leqslant |j| \leqslant M_m^u \;\middle|\; \Big|\frac{1}{\widehat{[\varphi]}_j} - \frac{1}{[\varphi]_j}\Big| \leqslant \frac{1}{2|[\varphi]_j|} \;\wedge\; |\widehat{[\varphi]}_j|^2 \geqslant 1/m \right\}, \quad (3.18)$$

consider the following decomposition of the risk:

$$\mathbf{E}\|\widehat{f}_{\widetilde{k}} - f\|_\omega^2 = \mathbf{E}\|\widehat{f}_{\widetilde{k}} - f\|_\omega^2 \mathbf{1}_{\Omega_q} + \mathbf{E}\|\widehat{f}_{\widetilde{k}} - f\|_\omega^2 \mathbf{1}_{\Omega_q^c}. \quad (3.19)$$

We bound these two terms separately. Consider the first term. By Lemma 3.25 below and $\mathbf{1}_{[|\widehat{[\varphi]}_j|^2 \geqslant 1/m]} \mathbf{1}_{\Omega_q} = \mathbf{1}_{\Omega_q}$, it follows that for all $1 \leqslant |j| \leqslant (N_n \wedge M_m)$,

$$\left(\frac{[\varphi]_j}{\widehat{[\varphi]}_j} \mathbf{1}_{[|\widehat{[\varphi]}_j|^2 \geqslant 1/m]} - 1 \right)^2 \mathbf{1}_{\Omega_q} = |[\varphi]_j|^2 \, \mathbf{1}_{\Omega_q} \left| \frac{1}{\widehat{[\varphi]}_j} - \frac{1}{[\varphi]_j} \right|^2 \leqslant \frac{1}{4}.$$

Hence, $\sup_{t \in \mathcal{B}_k} |\langle t, \widehat{\Phi}_\nu - \widetilde{\Phi}_\nu \rangle_\omega|^2 \mathbf{1}_{\Omega_q} \leqslant \frac{1}{4} \sup_{t \in \mathcal{B}_k} |\langle t, \widetilde{\Phi}_\nu \rangle_\omega|^2$ for all $0 \leqslant k \leqslant (N_n \wedge M_m)$, and (3.17) implies

$$\begin{aligned}\frac{1}{4}\|\widehat{f}_{\widetilde{k}} - f\|_\omega^2 \mathbf{1}_{\Omega_q} \leqslant & \frac{7}{4}\left(r\omega_k/\gamma_k\right) + 10\Big(\sup_{t \in \mathcal{B}_{k \vee \widetilde{k}}} |\langle t, \widetilde{\Phi}_\nu \rangle_\omega|^2 - (6\,\delta_{k \vee \widetilde{k}})/n \Big)_+ \\ & + \Big(60\,\delta_{k \vee \widetilde{k}}\Big)/n + \mathrm{pen}(k) - \mathrm{pen}(\widetilde{k}) + 8 \sup_{t \in \mathcal{B}_{(N_n \wedge M_m)}} |\langle t, \widehat{\Phi}_g - \widetilde{\Phi}_g \rangle_\omega|^2. \quad (3.20)\end{aligned}$$

Moreover, we have that $60\,\delta_{k \vee \widetilde{k}}/n = \mathrm{pen}(k \vee \widetilde{k}) \leqslant \mathrm{pen}(k) + \mathrm{pen}(\widetilde{k})$. Notice further that

$$\Delta_k \leqslant d\,\Delta_k^\lambda, \qquad \delta_k \leqslant d\,\zeta_d\,\delta_k^\lambda, \quad \text{and} \quad \delta_k/\Delta_k \geqslant 2\,k\,\zeta_d^{-1}\,\frac{\log(\Delta_k^\lambda \vee (k+2))}{\log(k+2)} \quad (3.21)$$

with $\zeta_d = \log(3d)/\log d$. From Lemma 3.25 it follows that

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\mathbf{E}\|\widehat{f}_{\widetilde{k}}-f\|_\omega^2\mathbf{1}_{\Omega_q}\leqslant 480\,(r+d\,\zeta_d)\min_{0\leqslant k\leqslant N_n^\lambda\wedge M_m^\lambda}[\max(\omega_k/\gamma_k,\delta_k^\lambda/n)]$$
$$+40\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\sum_{0\leqslant k'\leqslant(N_n^u\wedge M_m^u)}\mathbf{E}\Big(\sup_{t\in\mathcal{B}_{k'}}|\langle t,\widetilde{\Phi}_\nu\rangle_\omega|^2-(6\,\delta_{k'})/n\Big)_+$$
$$+32\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\mathbf{E}\Big[\sup_{t\in\mathcal{B}_{(N_n^u\wedge M_m^u)}}|\langle t,\widehat{\Phi}_g-\widetilde{\Phi}_g\rangle_\omega|^2\Big].$$

In order to bound the second term, we apply Lemma 3.27 setting $\delta_k^*=\delta_k$ and $\Delta_k^*=\Delta_k$. By virtue of (3.21), we have for all $k\geqslant 0$

$$\mathbf{E}\Big(\sup_{t\in\mathcal{B}_k}|\langle t,\widetilde{\Phi}_\nu\rangle_\omega|^2-6\,\frac{\delta_k}{n}\Big)_+\leqslant C\left\{\frac{1}{n^2}\,\exp\big(-K_2\sqrt{n}\big)\,d\,\zeta_d\,\delta_k^\lambda\right.$$
$$\left.+\,\frac{\|\varphi\|^2\,\|f\|^2}{n}\,d\,\Delta_k^\lambda\exp\left(-\frac{k}{3\,\|\varphi\|^2\,\|f\|^2\,\zeta_d}\,\frac{\log(\Delta_k^\lambda\vee(k+2))}{\log(k+2)}\right)\right\}.$$

Due to Lemmas 3.25 and 3.26 (i) and to the properties of the function Σ from Definition 3.11, we have

$$\sum_{k=0}^{N_n^u}\mathbf{E}\Big(\sup_{t\in\mathcal{B}_k}|\langle t,\widetilde{\Phi}_\nu\rangle_\omega|^2-6\,\frac{\delta_k}{n}\Big)_+\leqslant\frac{C}{n}\,d\,\Sigma(\|\varphi\|^2\,\|f\|^2\,\zeta_d).$$

It is readily verified that $\|\varphi\|^2\leqslant d\Lambda$ for all $\varphi\in\mathcal{E}_\lambda^d$ and $\|f\|^2\leqslant r$ for all $f\in\mathcal{F}_\gamma^r$. The remaining term can be controlled by virtue of Lemma 3.28, which shows

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\mathbf{E}\|\widehat{f}_{\widetilde{k}}-f\|_\omega^2\mathbf{1}_{\Omega_q}\leqslant C\Big\{(r+d\,\zeta_d)\min_{0\leqslant k\leqslant(N_n^\lambda\wedge M_m^\lambda)}[\max(\omega_k/\gamma_k,\delta_k^\lambda/n)]$$
$$+\,r\,d\,\kappa_m+d\,\Sigma(r\,d\,\Lambda\,\zeta_d)\,n^{-1}\Big\}.\quad(3.22)$$

Consider the second term from (3.19). Let

$$\breve{f}_k:=1+\sum_{0<|j|\leqslant k}[f]_j\mathbf{1}\{|\widehat{[\varphi]}_j|^2\geqslant 1/m\}e_j.$$

It is easy to see that $\|\widehat{f}_k-\breve{f}_k\|^2\leqslant\|\widehat{f}_{k'}-\breve{f}_{k'}\|^2$ for all $k\leqslant k'$ and $\|\breve{f}_k-f\|^2\leqslant\|f\|^2$ for all $k\geqslant 0$. Thus, using that $0\leqslant\widetilde{k}\leqslant(N_n^\circ\wedge m)$, we can write

$$\mathbf{E}\|\widehat{f}_{\widetilde{k}}-f\|_\omega^2\mathbf{1}_{\Omega_q^c}\leqslant 2\{\mathbf{E}\|\widehat{f}_{\widetilde{k}}-\breve{f}_{\widetilde{k}}\|_\omega^2\mathbf{1}_{\Omega_q^c}+\mathbf{E}\|\breve{f}_{\widetilde{k}}-f\|_\omega^2\mathbf{1}_{\Omega_q^c}\}$$
$$\leqslant 2\Big\{\mathbf{E}\|\widehat{f}_{(N_n^\circ\wedge m)}-\breve{f}_{(N_n^\circ\wedge m)}\|_\omega^2\mathbf{1}_{\Omega_q^c}+\|f\|_\omega^2\,\mathbf{P}[\Omega_q^c]\Big\}.$$

Moreover, applying Theorem A.3 from the appendix,

$$
\begin{aligned}
&\mathbf{E}\|\widehat{f}_{(N_n^\circ\wedge m)}-\check{f}_{(N_n^\circ\wedge m)}\|_\omega^2\mathbf{1}_{\Omega_q^c}\\
&\quad\leqslant 2m\sum_{0<|j|\leqslant(N_n^\circ\wedge m)}\omega_j\Big\{\mathbf{E}(\widehat{[g]}_j-[\varphi]_j[f]_j)^2\mathbf{1}_{\Omega_q^c}+\mathbf{E}([\varphi]_j[f]_j-\widehat{[\varphi]}_j[f]_j)^2\mathbf{1}_{\Omega_q^c}\Big\}\\
&\quad\leqslant 2m\Big\{\sum_{0<|j|\leqslant(N_n^\circ\wedge m)}\omega_j\Big[\mathbf{E}\left(\widehat{[g]}_j-[g]_j\right)^4\Big]^{1/2}\mathbf{P}[\Omega_q^c]^{1/2}\\
&\qquad\qquad+\sum_{0<|j|\leqslant(N_n^\circ\wedge m)}\omega_j|[f]_j|^2[\mathbf{E}(\widehat{[\varphi]}_j-[\varphi]_j)^4]^{1/2}\mathbf{P}[\Omega_q^c]^{1/2}\Big\}\\
&\quad\leqslant 2m\Big\{2m(\max_{1\leqslant j\leqslant N_n^\circ}\omega_j)(Cn^{-1})+(Cm^{-1})\|f\|_\omega^2\Big\}\,\mathbf{P}[\Omega_q^c]^{1/2},
\end{aligned}
$$

which implies, using Definition 3.10 (ii),

$$
\begin{aligned}
\mathbf{E}\|\widehat{f}_{\widetilde{k}}-f\|_\omega^2\mathbf{1}_{\Omega_q^c}&\leqslant 4C\Big(m^2+\|f\|_\omega^2\Big)\,\mathbf{P}[\Omega_q^c]^{1/2}+2\|f\|_\omega^2\,\mathbf{P}[\Omega_q^c]\\
&\leqslant 6Cm^2(1+\|f\|_\omega^2)\mathbf{P}[\Omega_q^c]^{1/2}.
\end{aligned}
\tag{3.23}
$$

It follows by Lemma 3.29 that for all $m\in\mathbb{N}$

$$
\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\mathbf{E}\|\widehat{f}_{\widehat{k}}-f\|_\omega^2\mathbf{1}_{\Omega_p^c}\leqslant C(d)(1+r)\,m^{-1}. \tag{3.24}
$$

The result of the theorem follows by combination of the last estimate with the bound from (3.22). □

A comparison with the lower bound from Corollary 3.6 shows that this upper bound ensures minimax optimality of the estimator $\widehat{f}_{\widetilde{k}}$ only if

$$
\psi_{n,m}^\diamond:=\min_{0\leqslant k\leqslant(N_n^\lambda\wedge M_m^\lambda)}\left[\max\left(\frac{\omega_k}{\gamma_k},\frac{\delta_k^\lambda}{n}\right)\right]
$$

is in the same order as $\psi_n=\min_{k\in\mathbb{N}}\Big\{\max\Big(\frac{\omega_k}{\gamma_k},\sum_{0<|j|\leqslant k}\frac{\omega_j}{n\lambda_j}\Big)\Big\}$. Note that, by construction, $\delta_k^\lambda\geqslant\sum_{0<|j|\leqslant k}\omega_j\lambda_j^{-1}$ for all $k\geqslant 1$. Also, δ^λ is direcly related to the penalty function. The next assertion is a immediate consequence of Theorem 3.12 and we omit its proof.

Corollary 3.13 *Under Assumption 3.2 and if additionally*

$$
\eta^\diamond:=\sup_{n,m\geqslant 1}\{\psi_{n,m}^\diamond/\psi_n\}<\infty,
$$

we have for all $n,m\geqslant 1$

$$
\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\Big\{\mathbf{E}\|\widehat{f}_{\widetilde{k}}-f\|_\omega^2\Big\}\leqslant C(\eta^\diamond,\Sigma,r,d,\Lambda)\ \max(\psi_n,\kappa_m).
$$

In Theorem 3.7, we have shown the minimax optimality of the orthogonal series estimator under the optimal choice k_n^* of the dimension parameter. Comparing Corollary 3.13 to this theorem, it is noteworthy that the only additional assumption needed to ensure minimax optimality of the partially adaptive estimator is $\eta^\diamond < \infty$.

Remark 3.14 The partially adaptive choice $\widetilde{k}$ still depends on $\varphi \in \mathcal{E}_\lambda^d$. However, we can already define a procedure depending only on the sequence λ and the constant d, namely

$$\widetilde{k}^\lambda := \operatorname*{argmin}_{1 \leqslant k \leqslant (N_n^\lambda \wedge M_m^\lambda)} \left\{ -\|\widehat{f}_k\|_\omega^2 + 60\, \frac{d\, \delta_k^\lambda}{n} \right\}.$$

Roughly speaking, this choice requires knowledge of the degree of ill-posedness of the underlying inverse problem only. It is straightforward to derive an upper risk bound for $\widehat{f}_{\widetilde{k}^\lambda}$, which is, up to minor changes in the constants, the same as the one in Theorem 3.12. Its proof follows the lines of the proof of Theorem 3.12, using the new penalty term $\operatorname{pen}(k) = 60\, d\, \delta_k^\lambda$. The only change occurs when applying Lemma 3.27, where one uses $\delta_k^* = d\, \delta_k^\lambda$ and $\Delta_k^* = d\, \Delta_k^\lambda$ rather than $\delta_k^* = \delta_k$ and $\Delta_k^* = \Delta_k$. □

Fully adaptive estimation

We begin by defining empirical versions of the sequences from Definition 3.10.

Definition 3.15 *For all $n, m \geqslant 1$ and $k \geqslant 0$, define*

(i) $\widehat{\Delta}_k := \max\limits_{-k \leqslant j \leqslant k} \frac{\omega_j}{|\widehat{[\varphi]}_j|^2} \mathbf{1}_{[|\widehat{[\varphi]}_j|^2 \geqslant 1/m]}$ *and* $\widehat{\delta}_k := k \widehat{\Delta}_k \frac{\log(\widehat{\Delta}_k \vee (k+2))}{\log(k+2)}$*;*

(ii) given N_n°, ω^+, and b from Definition 3.10,

$$\widehat{N}_n := \operatorname*{argmin}_{1 \leqslant j \leqslant N_n^\circ} \left\{ \frac{\min(|\widehat{[\varphi]}_j|^2, |\widehat{[\varphi]}_{-j}|^2)}{j \omega_j^+} < \frac{\log(n+2)}{n} \right\} - 1,$$

$$\widehat{M}_m := \operatorname*{argmin}_{1 \leqslant j \leqslant m} \left\{ \min(|\widehat{[\varphi]}_j|^2, |\widehat{[\varphi]}_{-j}|^2) < m^{-1+b_m} \right\} - 1,$$

with $\widehat{N}_n := N_n^\circ$ and $\widehat{M}_m := m$ if the respective sets in the argmin *are empty.*

We can now define a data-driven choice of k which, in contrast to $\widetilde{k}$, does not depend on the sequences δ, N, or M, but only on $\widehat{\delta}$, $\widehat{N}$, and $\widehat{M}$:

$$\widehat{k} := \operatorname*{argmin}_{0 \leqslant k \leqslant (\widehat{N}_n \wedge \widehat{M}_m)} \left\{ -\|\widehat{f}_k\|_\omega^2 + 600\, \frac{\widehat{\delta}_k}{n} \right\}. \tag{3.25}$$

The constant 600 arising in the definition of $\widehat{k}$, though convenient for deriving the theory, may be far too large in practice and instead be determined by means of a simulation study as in Comte et al. (2007), for example.

In the proof of Theorem 3.12, we have used the inequalities

$$(N_n^\lambda \wedge M_m^\lambda) \leqslant (N_n \wedge M_m) \leqslant (N_n^u \wedge M_m^u)$$

which hold by Lemma 3.25. In the proof of the next theorem, we consider the event $\{(N_n^\lambda \wedge M_m^\lambda) \leqslant (\widehat{N}_n \wedge \widehat{M}_m) \leqslant (N_n^u \wedge M_m^u)\}$ on which we can imitate the proof of Theorem 3.12. In order to control the risk on the complement of this event, we need to bound its probability. This necessitates the following assumption.

Assumption 3.16 *Suppose* $m^6 \exp\left(-m\lambda_{M_m^u+1}/(72\,d)\right) \leqslant C(\lambda, d)$ *for* $m \geqslant 1$.

Theorem 3.17 *Under Assumptions 3.2 and 3.16 we have for all* $n, m \geqslant 1$

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\left\{\mathbf{E}\|\widehat{f}_{\widehat{k}} - f\|_\omega^2\right\} \leqslant C\left\{(r + d\,\zeta_d)\min_{0\leqslant k\leqslant(N_n^\lambda\wedge M_m^\lambda)}\left[\max\left(\frac{\omega_k}{\gamma_k}, \frac{\delta_k^\lambda}{n}\right)\right] + r\,d\,\kappa_m\right\} + C(r, d, \lambda, \Sigma)\left[\frac{1}{m} + \frac{1}{n}\right].$$

Remark 3.18 Up to a change in the constant of the negligible terms, we obtain the same bound as for the partially adaptive estimator in Theorem 3.12. In comparison to the latter theorem, the only additional condition is Assumption 3.16. Note that the number 6 appearing in this assumption is just the exponent needed in order to show that the remainder term in the proof of the upper bound is in fact of negligible order: see (3.30) and Lemmas 3.29 and 3.30. Note that in Lemma 3.26 (ii) we show that

$$m^6 \exp\left(-m\lambda_{M_m^u}/(72\,d)\right) \leqslant C(d)$$

for all $m \geqslant 1$ using only Assumption 3.2. It is however not obvious to us that Assumption 3.2 implies $m^6 \exp\left(-m\lambda_{M_m^u+1}/(72\,d)\right) \leqslant C(d)$ for sufficiently large m. Though, in the illustrations below we show that Assumption 3.16 is satisfied. □

Proof of Theorem 3.17. We begin the proof by defining the event $\Omega_{qp} := \Omega_q \cap \Omega_p$ where Ω_q is given in (3.18) and

$$\Omega_p := \left\{(N_n^\lambda \wedge M_m^\lambda) \leqslant (\widehat{N}_n \wedge \widehat{M}_m) \leqslant (N_n^u \wedge M_m^u)\right\}. \tag{3.26}$$

Observe that on Ω_q we have $(1/2)\Delta_k \leqslant \widehat{\Delta}_k \leqslant (3/2)\Delta_k$ for all $0 \leqslant k \leqslant M_m^u$ and hence $(1/2)[\Delta_k \vee (k+2)] \leqslant [\widehat{\Delta}_k \vee (k+2)] \leqslant (3/2)[\Delta_k \vee (k+2)]$, which implies

$$(1/2)k\Delta_k\Big(\frac{\log[\Delta_k \vee (k+2)]}{\log(k+2)}\Big)\Big(1 - \frac{\log 2}{\log(k+2)}\frac{\log(k+2)}{\log(\Delta_k \vee [k+2])}\Big)$$
$$\leqslant \widehat{\delta}_k \leqslant (3/2)k\Delta_k\Big(\frac{\log(\Delta_k \vee [k+2])}{\log(k+2)}\Big)\Big(1 + \frac{\log 3/2}{\log(k+2)}\frac{\log(k+2)}{\log(\Delta_k \vee [k+2])}\Big).$$

Using $\log(\Delta_k \vee (k+2))/\log(k+2) \geqslant 1$, we conclude from the last estimate that

$$\begin{aligned}\delta_k/10 \leqslant (\log 3/2)/(2\log 3)\delta_k &\leqslant (1/2)\delta_k[1 - (\log 2)/\log(k+2)] \leqslant \widehat{\delta}_k \\ &\leqslant (3/2)\delta_k[1 + (\log 3/2)/\log(k+2)] \leqslant 3\delta_k.\end{aligned}$$

Letting $\mathrm{pen}(k) := 60\,\delta_k n^{-1}$ and $\widehat{\mathrm{pen}}(k) := 600\,\widehat{\delta}_k n^{-1}$, it follows that on Ω_q

$$\mathrm{pen}(k) \leqslant \widehat{\mathrm{pen}}(k) \leqslant 30\,\mathrm{pen}(k) \qquad \forall\, 0 \leqslant k \leqslant M_m^u.$$

On $\Omega_{qp} = \Omega_q \cap \Omega_p$, we have $\widehat{k} \leqslant M_m^u$. Thus,

$$\begin{aligned}\Big(\mathrm{pen}(k \vee \widehat{k}) + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k})\Big)\mathbf{1}_{\Omega_{qp}} &\\ \leqslant \Big(\mathrm{pen}(k) + \mathrm{pen}(\widehat{k}) + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k})\Big)\mathbf{1}_{\Omega_{qp}} &\\ \leqslant 31\,\mathrm{pen}(k) \qquad \forall\, 0 \leqslant k \leqslant M_m^u.& \quad (3.27)\end{aligned}$$

Now consider the decomposition

$$\mathbf{E}\|\widehat{f}_{\widehat{k}} - f\|_\omega^2 = \mathbf{E}\|\widehat{f}_{\widehat{k}} - f\|_\omega^2 \mathbf{1}_{\Omega_{qp}} + \mathbf{E}\|\widehat{f}_{\widehat{k}} - f\|_\omega^2 \mathbf{1}_{\Omega_{qp}^c}. \tag{3.28}$$

We bound the two terms separately. Consider the first term. Following the proof of (3.20) line by line, one sees that we have for all for $0 \leqslant k \leqslant (N_n^\lambda \wedge M_m^\lambda)$

$$\begin{aligned}(1/4)\|\widehat{f}_{\widehat{k}} - f\|_\omega^2 \mathbf{1}_{\Omega_{qp}} &\leqslant (7/4)(r\omega_k/\gamma_k) + 10\sum_{j=0}^{N_n^u}\Big(\sup_{t\in\mathcal{B}_j}|\langle t, \widetilde{\Phi}_\nu\rangle_\omega|^2 - 6\frac{\delta_j}{n}\Big)_+ \\ &\quad + 8\sup_{t\in\mathcal{B}_{N_n^u \wedge M_m^u}}|\langle t, \widehat{\Phi}_g - \widetilde{\Phi}_g\rangle_\omega|^2 + \Big(\mathrm{pen}(k \vee \widehat{k}) + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k})\Big)\mathbf{1}_{\Omega_{qp}} \\ &\leqslant (7/4)(r\omega_k/\gamma_k) + 10\sum_{j=0}^{N_n^u}\Big(\sup_{t\in\mathcal{B}_j}|\langle t, \widetilde{\Phi}_\nu\rangle_\omega|^2 - 6\frac{\delta_j}{n}\Big)_+ \\ &\quad + 8\sup_{t\in\mathcal{B}_{N_n^u \wedge M_m^u}}|\langle t, \widehat{\Phi}_g - \widetilde{\Phi}_g\rangle_\omega|^2 + 31\,\mathrm{pen}(k),\end{aligned}$$

where the last inequality follows from (3.27). The second and the third term are controlled by Lemmas 3.27 and 3.28, respectively (cf. proof of (3.22)). Consequently,

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\mathbf{E}\|\widehat{f}_{\widehat{k}}-f\|_\omega^2\mathbf{1}_{\Omega_{qp}}\leqslant C\Big\{(r+d\,\zeta_d)\min_{0\leqslant k\leqslant(N_n^\lambda\wedge M_m^\lambda)}[\max(\omega_k/\gamma_k,\delta_k^\lambda/n)] + r\,d\,\kappa_m+d\,\Sigma(r\,d\,\Lambda\,\zeta_d)\,n^{-1}\Big\}. \quad (3.29)$$

Consider the second term from (3.28). Following the proof of (3.23), and replacing Ω_q^c by Ω_{qp}^c therein, we obtain

$$\mathbf{E}\|\widehat{f}_{\widehat{k}}-f\|_\omega^2\mathbf{1}_{\Omega_{qp}^c}\leqslant C\,m^2(1+\|f\|_\omega^2)\mathbf{P}[\Omega_{qp}^c]^{1/2}. \quad (3.30)$$

It follows by Lemmas 3.29 and 3.30 that for all $m\geqslant 1$,

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\mathbf{E}\|\widehat{f}_{\widehat{k}}-f\|_\omega^2\mathbf{1}_{\Omega_{qp}^c}\leqslant C(\lambda,d)(1+r)\,m^{-1}.$$

The theorem follows combining the last estimate with (3.24) and (3.29). □

A comparison of Theorem 3.17 with the lower bound from Corollary 3.6 shows that this upper bound does not necessarily ensure minimax optimality of the estimator $\widehat{f}_{\widehat{k}}$. However, as in the partially adaptive case (cf. Corollary 3.13), under the additional assumption $\eta^\diamond<\infty$, the next assertion establishes its optimality.

Corollary 3.19 *Under Assumptions 3.2 and 3.16 and the additional condition* $\eta^\diamond:=\sup_{n,m\geqslant 1}\{\psi_{n,m}^\diamond/\psi_n\}<\infty$, *we have for all* $n,m\geqslant 1$

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\Big\{\mathbf{E}\|\widehat{f}_{\widehat{k}}-f\|_\omega^2\Big\}\leqslant C(\eta^\diamond,\Sigma,r,d,\Lambda)\ \max(\psi_n,\kappa_m).$$

Illustration: estimation of derivatives

Let us continue with the example from Section 3.2. The following result shows that without any prior knowledge on the error density φ, the adaptive penalized estimator automatically attains the optimal rate in the cases **[o-s]** and **[s-o]** and in the case **[o-o]** if $p-s\geqslant a$. Recall that the computation of the dimension parameter $\widehat{k}$ given in (3.25) involves the sequence N°, which in our illustration satisfies $N_n^\circ\sim n^{1/(2s)}$.

Proposition 3.20 *Let $(m_n)_{n\geqslant 1}$ a non decreasing sequence of integers.*

[o-o] *We have that*

$$\Delta_k^\lambda \sim k^{2a+2s}, \quad \delta_k^\lambda \sim k^{2a+2s+1}, \quad \psi_{n,m_n}^\diamond \sim (k_n^* \wedge M_{m_n}^\lambda)^{-2(p-s)}$$
$$N_n^\lambda \sim (n/\log n)^{1/(2a+2s+1)}, \quad M_{m_n}^\lambda \sim m_n^{(1-b_m)/(2a)}.$$

In the case $p-s > a$ the adaptive estimator $\widehat{f}_{\widehat{k}}^{(s)}$ attains the optimal rates. In the case $p-s \leqslant a$, if $n^{2a/(2p+2a+1)} = O(m_n)$, we have

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\left\{\mathbf{E}\|\widehat{f}_{\widehat{k}}^{(s)} - f^{(s)}\|^2\right\} = \begin{cases} O(n^{-2(p-s)/(2p+2a+1)}) & \text{if } n^{2a/(2p+2a+1)} = O(m_n^{1-b_{m_n}}) \\ O(m_n^{-(p-s)/a}\, m_n^{b_{m_n}}) & \text{otherwise,} \end{cases}$$

while if $m_n = o(n^{2a/(2p+2a+1)})$ we have

$$\sup_{f\in\mathcal{F}_\gamma^r}\sup_{\varphi\in\mathcal{E}_\lambda^d}\left\{\mathbf{E}\|\widehat{f}_{\widehat{k}}^{(s)} - f^{(s)}\|^2\right\} = O(m_n^{-(p-s)/a} m_n^{b_{m_n}}).$$

[s-o] *The sequences Δ^λ, δ^λ, N^λ, and M^λ are the same as above. We have that $\psi_{n,m_n}^\diamond \sim (k_n^* \wedge M_{m_n}^\lambda)^{2s}\exp(-(k_n^* \wedge M_{m_n}^\lambda)^{2p})$, and $\widehat{f}_{\widehat{k}}^{(s)}$ attains the optimal rates.*

[o-s] *We have that*

$$\Delta_k^\lambda = k^{2s}\exp(k^{2a}), \quad \delta_k^\lambda \sim k^{2a+2s+1}\exp(k^{2a})(\log k)^{-1},$$
$$\psi_{n,m}^\diamond \sim (k_n^* \wedge M_{m_n}^\lambda)^{-2(p-s)},$$
$$N_n^\lambda \sim (\log(n/(\log n)^{(2a+2s+1)/(2a)}))^{1/(2a)}, \quad M_{m_n}^\lambda \sim ((1-b_m)\log m_n)^{1/(2a)},$$

and the adaptive estimator $\widehat{f}_{\widehat{k}}^{(s)}$ attains the optimal rates.

Proof. In view of Proposition 3.8 we apply Theorem 3.17, where we only have to check the additional Assumption 3.16. The result follows then by an evaluation of the upper bound.

[o-o] It is easily seen that $m\,(\lambda_{M_m^u+1}\log m)^{-1} = o(1)$ as $m \to \infty$. Hence, Assumption 3.16 is satisfied in this case. Since $k_n^* \sim n^{1/(2a+2p+1)}$, we have $k_n^* \lesssim N_n^\lambda$. Thus, the upper bound is

$$(k_n^* \wedge M_{m_n}^\lambda)^{-2(p-s)} + m_n^{-(1\wedge((p-s)/a))}. \tag{3.31}$$

We consider two cases. First, let $p - s > a$. Suppose that $n^{2(p-s)/(2p+2a+1)} = O(m_n)$. Then,

$$\frac{k_n^*}{M_{m_n}^\lambda} \sim \frac{n^{1/(2a+2p+1)}}{\left(m_n^{1-b_{m_n}}\right)^{(1/2a)}} = \frac{n^{1/(2a+2p+1)}}{m_n^{1/2(p-s)}} \left(m_n^{-a+(p-s)(1-b_{m_n})}\right)^{\frac{1}{2(p-s)a}} = o(1).$$

This means that $k_n^* \lesssim M_{m_n}^\lambda$, so the resulting upper bound is $(k_n^*)^{-2(p-s)} + m_n^{-1} \lesssim (k_n^*)^{-2(p-s)}$. Suppose now that $m_n = o(n^{2(p-s)/(2p+2a+1)})$. If in addition $k_n^* = O(M_{m_n}^\lambda)$, then the first summand in (3.31) reduces to $(k_n^*)^{-2(p-s)}$ and hence the upper bound is m_n^{-1}. On the other hand, if $M_{m_n}^\lambda / k_n^* = o(1)$, then the first term is $(M_{m_n}^\lambda)^{-2(p-s)} \sim (m_n^{-a+(p-s)(1-b_{m_n})})^{1/a} m_n^{-1} \lesssim m_n^{-1}$, since $p - s > a$. Combining both cases, we obtain the result in case $p - s > a$.

Now assume $p - s \leqslant a$. First, suppose that $k_n^* = O(M_{m_n}^\lambda)$. Then, the first summand in (3.31) reduces to $(k_n^*)^{-2(p-s)}$ and moreover it follows that $n^{2a/(2p+2a+1)} = O(m_n)$. Therefore, the upper bound is $(k_n^*)^{-2(p-s)}$. Consider now $M_{m_n}^\lambda = o(k_n^*)$. Then (3.31) can be rewritten as $(m_n^{1-b_{m_n}})^{-(p-s)/a} + m_n^{-(p-s)/a}$ which results in the rate $(m_n^{1-b_m})^{-(p-s)/a}$. Combining both cases gives the result. More precisely, $m_n = o(n^{2a/(2p+2a+1)})$ implies $M_{m_n}^\lambda = o(k_n^*)$. On the other hand, in case $n^{2a/(2p+2a+1)} = O(m_n)$, if $k_n^* / M_{m_n}^\lambda = O(1)$, then the rate is $(k_n^*)^{-2p}$, while if $M_{m_n}^\lambda / k_n^* = o(1)$, we have the rate $(m_n^{1-b_m})^{-(p-s)/a}$.

[s-o] As in case [o-o], Assumption 3.16 holds. Recall that $k_n^* \sim (\log n)^{1/(2p)}$. If $n(\log n)^{-(2a+2s+1)/(2p)} = O(m_n)$, then $k_n^* \lesssim M_{m_n}^\lambda$ and

$$m_n^{-1} \lesssim \psi_{n,m_n}^\diamond \sim n^{-1} (\log n)^{(2a+2s+1)/(2p)}.$$

In the opposite case, we have $\psi_{n,m_n}^\diamond \lesssim m_n^{-1}$, which proves the result.

[o-s] To see that Assumption 3.16 is satisfied in this setting, one can proceed as follows. Define the sequence $\widetilde{M}^u$ exactly as M^u but replacing b_m by $a_m = b_m^{2^k}$. Then, $\widetilde{M}^u$ satisfies assertion Lemma 3.26 (ii), the proof being similar to the one for M^u. On the other hand, one can show that $\widetilde{M}_m^u - M_m^u \to \infty$ as $m \to \infty$, which amounts to showing Assumption 3.16.

We have $k_n^* \sim (\log n)^{1/2a}$. The upper bound becomes $(k_n^* \wedge M_{m_n}^\lambda)^{-2(p-s)} + (\log m_n)^{-(p-s)/a} \sim (k_n^* \wedge M_{m_n}^\lambda)^{-2(p-s)}$. Distinguishing $k_n^* \lesssim M_m^\lambda$ and the opposite case shows the result. □

The adaptive estimator always attains the minimal rates if $n \lesssim m_n$. We underline that these rates are still attained when $m_n \lesssim n$ except in the case **[os]**

when the error density is smoother than the s-th derivative of the deconvolution density ($p - s \leqslant a$) and when at the same time m_n grows far more slowly than n. The estimation of φ is negligible as soon as $m_n^{1-b_{m_n}}$ grows at least as fast as $n^{2a/(2p+2a+1)}$ in this situation, while in the non adaptive case, only m_n has to satisfy this condition. In the lossy case, the convergence rate differs from the optimal one by a factor $m_n^{b_{m_n}}$ only. The exponent b_{m_n} however tends to zero as n tends to infinity.

If one were considering the **[os]** case only, one could replace the bound m^{-1+b_m} by $m^{-1}\log m$ in the definition of M^u (Definition 3.11). Using this definition, Assumption 3.16 would still hold, and applying Theorem 3.17, the adaptive estimator misses the optimal rates by a logarithmic factor in the lossy case only. However, Assumption 3.16 is violated in the super smooth case under this definition of M^u.

3.4 Auxiliary results

Lemma 3.21 *In the context of Theorem 3.3, $f_\theta \in \mathcal{F}_\gamma^r$ for all $\theta \in \{-1,1\}^{2k_n^*}$.*

Proof. The assertion is easily verified if $f \in \mathcal{F}_\gamma^r$. In order to show that f belongs indeed to $\mathcal{F}_\gamma^r$, we first notice that f integrates to one. Moreover, f is non negative because $|\sum_{0<|j|\leqslant k_n^*}[f]_j e_j| \leqslant 1$, and $\|f\|_\gamma^2 \leqslant r$, which can be seen as follows. By employing the condition $\sum_{j\in\mathbb{Z}} \gamma_j^{-1} = \Gamma < \infty$ we have

$$
\begin{aligned}
|\sum_{0<|j|\leqslant k_n^*}[f]_j e_j| \leqslant \sum_{0<|j|\leqslant k_n^*}|[f]_j| &= \Big(\frac{\zeta\alpha_n}{n}\Big)^{1/2}\sum_{0<|j|\leqslant k_n^*}\lambda_j^{-1/2} \\
&\leqslant \Big(\zeta\alpha_n\Big)^{1/2}\Big(\sum_{0<|j|\leqslant k_n^*}\gamma_j^{-1}\Big)^{1/2}\Big(\sum_{0<|j|\leqslant k_n^*}\frac{\gamma_j}{n\lambda_j}\Big)^{1/2} \\
&\leqslant \Big(\zeta\alpha_n\Gamma\Big)^{1/2}\Big(\sum_{0<|j|\leqslant k_n^*}\frac{\gamma_j}{n\lambda_j}\Big)^{1/2}.
\end{aligned}
$$

Since ω/γ is non increasing the definition of ζ, α_n and η implies

$$
|\sum_{0<|j|\leqslant k_n^*}[f]_j e_j| \leqslant \Big(\zeta\Gamma\Big)^{1/2}\Big(\frac{\gamma_{k_n^*}}{\omega_{k_n^*}}\alpha_n\sum_{0<|j|\leqslant k_n^*}\frac{\omega_j}{\lambda_j n}\Big)^{1/2} \leqslant \Big(\frac{\zeta\Gamma}{\eta}\Big)^{1/2} \leqslant 1 \quad (3.32)
$$

as well as $\|f\|_\gamma^2 \leqslant 1 + \zeta\frac{\gamma_{k_n^*}}{\omega_{k_n^*}}\alpha_n\Big(\sum_{0<|j|\leqslant k_n^*}\frac{\omega_j}{n\lambda_j}\Big) \leqslant 1 + \zeta/\eta \leqslant r$. □

Lemma 3.22 *In the context of Theorem 3.3, we have* $\rho(g_\theta^n, g_{\theta^{(j)}}^n) \geqslant 1/4$.

Proof. We consider first the Hellinger distance

$$H^2(g_\theta, g_{\theta^{(j)}}) := \int \Big(\sqrt{g}_\theta - \sqrt{g}_{\theta^{(j)}}\Big)^2 = \int \frac{\Big|g_\theta - g_{\theta^{(j)}}\Big|^2}{\Big(\sqrt{g}_\theta + \sqrt{g}_{\theta^{(j)}}\Big)^2} \leqslant 4\|g_\theta - g_{\theta^{(j)}}\|^2 = 16|[f]_j|^2|[\varphi]_j|^2 \leqslant \frac{16\zeta d}{\eta\, n},$$

where we have used that $\alpha_n \leqslant 1/\eta$, $\varphi \in \mathcal{E}_\lambda^d$ and $g_\theta \geqslant 1/2$ because the expression $|\sum_{0<|j|\leqslant k_n^*}[g_\theta]_j e_j|$ is bounded by $1/2$, which can be seen as follows. Using the condition $\sum_{j\in\mathbb{Z}}\gamma_j^{-1} = \Gamma < \infty$ and $\varphi \in \mathcal{E}_\lambda^d$ we obtain in analogy to the proof of (3.32) that

$$\Big|\sum_{0<|j|\leqslant k_n^*}[g_\theta]_j e_j\Big| \leqslant \sum_{0<|j|\leqslant k_n^*}|[f]_j||[\varphi]_j| \leqslant \Big(\frac{\zeta\alpha_n d}{n}\Big)^{1/2}\sum_{0<|j|\leqslant k_n^*}\lambda_j^{-1/2} \leqslant \Big(\frac{\zeta d\Gamma}{\eta}\Big)^{1/2} \leqslant 1/2.$$

Therefore, the definition of ζ implies $H^2(g_\theta, g_{\theta^{(j)}}) \leqslant 2/n$. By using the independence, i.e. $\rho(g_\theta^n, g_{\theta^{(j)}}^n) = \rho(g_\theta, g_{\theta^{(j)}})^n$, together with the identity $\rho(g_\theta, g_{\theta^{(j)}}) = 1 - \frac{1}{2}H^2(g_\theta, g_{\theta^{(j)}})$ it follows $\rho(g_\theta^n, g_{\theta^{(j)}}^n) \geqslant (1-n^{-1})^n \geqslant 1/4$ for all $n \geqslant 2$. □

Lemma 3.23 *In the context of Theorem 3.5, we have* $f_\theta \in \mathcal{F}_\gamma^r$ *and* $\varphi_\theta \in \mathcal{E}_\lambda^d$ *for* $\theta \in \{-1, 1\}$.

In order to show $f_\theta \in \mathcal{F}_\gamma^r$, we first observe that f_θ integrates to one. Moreover, f_θ is non negative because $|(1-\theta\alpha_m)\frac{1\wedge\sqrt{r-1}}{d^{1/4}}\gamma_{k_m^*}^{-1/2}| \leqslant \gamma_{k_m^*}^{-1/2} \leqslant 1$ and

$$\|f_\theta\|_\gamma^2 = 1 + \gamma_{k_m^*}|[f_\theta]_{k_m^*}|^2 \leqslant 1 + \gamma_{k_n^*}|(1-\theta\alpha_m)\frac{1\wedge\sqrt{r-1}}{d^{1/4}}\gamma_{k_m^*}^{-1/2}|^2 \leqslant r.$$

Consider now φ_θ. Obviously, it integrates to one. Furthermore, as $\varphi \geqslant 1/2$, the function $\varphi_\theta = \varphi + \theta\alpha_m[\varphi]_{k_m^*}e_{k_m^*}$ is non negative since

$$|\theta\alpha_m[\varphi]_{k_m^*}e_{k_m^*}| \leqslant \alpha_m\lambda_{k_m^*}^{1/2}d^{1/2} \leqslant \zeta m^{-1/2}\sqrt{d} \leqslant 1/2$$

using the definition of α_m and ζ. To check that $\varphi_\theta \in \mathcal{E}_\lambda^d$, it remains to show that $1/d \leqslant [\varphi_\theta]_j^2/\lambda_j \leqslant d$ for all $|j| > 0$. Since $\varphi \in \mathcal{E}_\lambda^{\sqrt{d}}$, it follows from the

definition of φ_θ that these inequalities are satisfied for all $j \neq k_m^*$ and moreover that

$$1/d \leqslant \frac{|[\varphi]_{k_m^*}|^2}{\sqrt{d}\lambda_{k_m^*}} \leqslant \frac{(1+\theta\alpha_m)^2|[\varphi]_{k_m^*}|^2}{\lambda_{k_m^*}} \leqslant \frac{\sqrt{d}|[\varphi]_{k_m^*}|^2}{\lambda_{k_m^*}} \leqslant d,$$

which completes the proof. □

Lemma 3.24 *In the context of Theorem 3.5, $\rho(p_1, p_{-1}) \geqslant 1/4$ for all $m \geqslant 2$.*

From the independence and the fact that $g_1 = g_{-1}$, it is easily seen that Hellinger affinity satisfies

$$\rho(p_1, p_{-1}) = \rho(g_1, g_{-1})^n \rho(\varphi_1, \varphi_{-1})^m = \rho(\varphi_1, \varphi_{-1})^m = \Big(1 - \frac{1}{2}H^2(\varphi_1, \varphi_{-1})\Big)^m.$$

Hence, we conclude $\rho(p_1, p_{-1}) \geqslant (1-1/m)^m \geqslant 1/4$, for all $m \geqslant 2$, since

$$\begin{aligned} H^2(\varphi_1, \varphi_{-1}) &\leqslant \int \frac{\Big|\varphi_1 - \varphi_{-1}\Big|^2}{\varphi_1 + \varphi_{-1}} = \int \frac{\Big|\varphi_1 - \varphi_{-1}\Big|^2}{\varphi} \leqslant 2\int |\varphi_1 - \varphi_{-1}|^2 \\ &\leqslant 2\int 4\alpha_m^2 |[\varphi]_{k_m^*}|^2 e_{k_m^*}^2 \leqslant 8d\alpha_m^2 \lambda_{k_m^*} = 8d\zeta^2 m^{-1} \leqslant 2m^{-1} \end{aligned}$$

where we have used that $\varphi \geqslant 1/2$ and the definition of α_m and ζ. □

Lemma 3.25 *Under Assumption 3.2, we have for all $n, m \in \mathbb{N}$*

$$N_n^\lambda \leqslant N_n \leqslant N_n^u \qquad \text{and} \qquad M_m^\lambda \leqslant M_m \leqslant M_m^u.$$

Proof. First, we prove that $N_n^\lambda \leqslant N_n$. If $N_n^\lambda = 0$ or $N_n = N_n^\circ$, there is nothing to show. Noting that

$$N_n^\lambda = 0 \iff \max_{1 \leqslant j \leqslant N_n^\circ} \frac{\lambda_j}{j\omega_j^+} < \frac{4d\log(n+2)}{n}$$

$$\text{and} \quad N_n = 0 \iff \max_{1 \leqslant j \leqslant N_n^\circ} \frac{\lambda_j}{j\omega_j^+} < \frac{d\log(n+2)}{n},$$

we deduce that in case $N_n = 0$, we also have $N_n^\lambda = 0$. It remains the case where $N_n^\lambda > 0$ and $N_n^\circ > N_n > 0$, which implies

$$\min_{1 \leqslant j \leqslant N_n^\lambda} \frac{\lambda_j}{j\omega_j^+} \geqslant \frac{4d\log(n+2)}{n} \quad \text{and} \quad \frac{\log(n+2)}{n} > \frac{|[\varphi]_{N_n+1}|^2}{N_n\,\omega_{N_n+1}} \geqslant \frac{\lambda_{N_n+1}}{dN_n\omega_{N_n+1}^+}$$

and therefore $N_n + 1 > N_n^\lambda$, which proves the claim.

Let us now prove $N_n \leqslant N_n^u$. If $N_n = 0$ or $N_n^u = n$, this is trivial. On the other hand, if $n > N_n^u \geqslant 0$ and $N_n^\circ \geqslant N_n > 0$, it follows from the definitions that

$$\min_{1\leqslant j\leqslant N_n} \frac{d\lambda_j}{j\omega_j^+} \geqslant \min_{1\leqslant j\leqslant N_n} \frac{|[\varphi]_j|^2}{j\omega_j^+} \geqslant \frac{\log(n+2)}{n}$$

$$\text{and} \quad \frac{\lambda_{N_n^\circ+1}}{(N_n^\circ+1)\omega^+_{N_n^\circ+1}} < \frac{\log(n+2)}{4dn},$$

which implies $N_n^\circ + 1 > N_n$ and hence the claim. Similar arguments show the corresponding estimates in m. □

Lemma 3.26 *Under Assumption 3.2, we have for all $n, m \geqslant 3$ that*

(i) $\delta_{N_n^u}/n \leqslant 32\, d^2$

(ii) $m^7 \exp\left(-\frac{m\lambda_{M_m^u}}{72\, d}\right) \leqslant C(d)$

and for $m \geqslant \exp(512 \log(3d)^2)$ that

(iii) $\min_{1\leqslant j\leqslant M_m^u} |[\varphi]_j|^2 \geqslant \frac{2}{m}$.

Proof. (i) For $N_n^u = 0$, we have $\delta_{N_n^u} = 0$ and there is nothing to show. If $0 < N_n^u \leqslant n$, one can show that $\omega^+_{N_n^u}/\lambda_{N_n^u} \leqslant 4dn/(N_n^u \log(n+2))$, which we use in the following computation:

$$\begin{aligned}
\delta_{N_n^u} &= N_n^u \, \frac{\omega^+_{N_n^u}}{\lambda_{N_n^u}} \, \frac{\log((\omega^+_{N_n^u}/\lambda_{N_n^u}) \vee (N_n^u+2))}{\log(N_n^u+2)} \\
&\leqslant \frac{4dn}{\log(n+2)} \, \frac{\log\left(\frac{4dn}{N_n^u \log(n+2)} \vee (N_n^u+2)\right)}{\log(N_n^u+2)} \\
&\leqslant n \begin{cases} 4d & (\log(n+2) \geqslant 4d) \\ 4d(4d+\log(4d))/(\log(n+2)) & (\text{otherwise}), \end{cases}
\end{aligned}$$

which implies $\delta_{N_n^u}/n \leqslant 4d(4d+\log(4d)) \leqslant 32d^2$ for all $n \geqslant 1$.

(ii) For $0 < M_m^u \leqslant m$, we have $\lambda_{M_m^u} \geqslant m^{-1+b_m}\,(4d)^{-1}$. Hence,

$$m^7 \exp\left(-\frac{m\lambda_{M_m^u}}{72d}\right) \leqslant \exp\left(-\frac{m^{b_m}}{288d^2} + 7\log m\right).$$

This proves the claim, because $\log m \lesssim m^{b_m}$. Note that $M_m^u = 0$ cannot occur as we suppose $\lambda_1 = 1$.

(iii) We have that

$$\min_{1\leqslant j\leqslant M_m^u} |[\varphi]_j|^2 \geqslant \min_{1\leqslant j\leqslant M_m^u} \frac{\lambda_j}{d} \geqslant \frac{m^{b_m}}{4d^2\, m} \geqslant \frac{2}{m},$$

where the last step holds for $m \geqslant \exp(512 \log(3d)^2)$ as some algebra shows. □

Lemma 3.27 *Let δ^* and Δ^* be sequences such that for all $k \geqslant 1$*

$$\delta_k^* \geqslant \sum_{-k\leqslant j\leqslant k} \frac{\omega_j}{|[\varphi]_j|^2} \qquad \text{and} \qquad \Delta_k^* \geqslant \max_{0\leqslant |j|\leqslant k} \frac{\omega_j}{|[\varphi]_j|^2}$$

and let $K_2 := (\sqrt{2}-1)/(21\sqrt{2})$. Then, for all $k \geqslant 1$,

$$\begin{aligned}&\mathbf{E}\Big[\Big(\sup_{t\in\mathcal{B}_k} |\langle t, \widetilde{\Phi}_\nu\rangle_\omega|^2 - \frac{6\,\delta_k^*}{n}\Big)_+\Big] \\ &\leqslant C\left\{\frac{\|\varphi\|^2\, \|f\|^2}{n}\,\Delta_k^* \exp\left(-\frac{1}{6\,\|\varphi\|^2\, \|f\|^2}(\delta_k^*/\Delta_k^*)\right) + \frac{1}{n^2}\,\exp\left(-K_2\,\sqrt{n}\right)\delta_k^*\right\}.\end{aligned}$$

Proof. For $t \in \mathcal{S}_k$ define the function $r_t := \sum_{k\leqslant j\leqslant k} \omega_j [t]_j \overline{[\varphi]}_j^{-1} e_j$, then it is readily seen that $\langle t, \widetilde{\Phi}_\nu\rangle_\omega = \frac{1}{n}\sum_{k=1}^n r_t(Y_k) - \mathbf{E}[r_t(Y_k)]$. Next, we compute constants H_1, H_2, and v verifying the three inequalities required in Talagrand's inequality (Theorem A.5), which then implies the result. Consider H_1 first:

$$\sup_{t\in\mathcal{B}_k} \|r_t\|_\infty^2 = \sup_{y\in\mathbb{R}} \sum_{-k\leqslant j\leqslant k} \omega_j |[\varphi]_j|^{-2}\, |e_j(y)|^2 = \sum_{-k\leqslant j\leqslant k} \omega_j |[\varphi]_j|^{-2} \leqslant \delta_k^* =: H_1^2.$$

Next, find H_2. Notice that

$$\mathbf{E}[\sup_{t\in\mathcal{B}_k} |\langle t, \widetilde{\Phi}_\nu\rangle_\omega|^2] = \frac{1}{n} \sum_{-k\leqslant j\leqslant k} \omega_j |[\varphi]_j|^{-2}\, \mathbf{V}\mathrm{ar}(e_j(Y_1)).$$

As $\mathbf{V}\mathrm{ar}(e_j(Y_1)) \leqslant \mathbf{E}[|\, e_j(Y_1)\,|^2] = 1$, we have $\mathbf{E}[\sup_{t\in\mathcal{B}_k} |\langle t, \widetilde{\Phi}_\nu\rangle|^2] \leqslant \delta_k^*/n$ and we set $H_2 := \delta_k^*/n$.
Finally, consider v. Given $t \in \mathcal{B}_k$, let $\underline{[t]} := ([t]_{-k}, \ldots, [t]_k)^t$ and for a sequence $(z_j)_{j\in\mathbb{Z}}$ denote by $D_k(z) := \mathrm{diag}[z_{-k}, \ldots, z_k]$ the corresponding diagonal matrix. Define the Hermitian matrix

$$A_k := \left(\overline{[\varphi]}_j^{-1} [\varphi]_{j'}^{-1} [\varphi]_{j-j'}\, [f]_{j-j'}\right)_{j,j'=-k,\ldots,k}.$$

Straightforward algebra shows

$$\sup_{t\in\mathcal{B}_k} \mathbf{V}\mathrm{ar}(r_t(Y_1)) \leqslant \sup_{t\in\mathcal{B}_k} \langle A_k D_k(\omega)\, \underline{[t]}, D_k(\omega)\underline{[t]}\rangle_{\mathbb{C}^{2k+1}}$$

and hence, by the Cauchy-Schwarz inequality,

$$\sup_{t\in\mathcal{B}_k}\frac{1}{n}\sum_{k=1}^{n}\mathbf{V}\mathrm{ar}(r_t(Y_k)) \leqslant \|D_k(\sqrt{\omega})A_kD_k(\sqrt{\omega})\|_{\mathbb{C}^{2k+1}}.$$

We have $A_k = D_k([\varphi]^{-1})\ B_k\ D_k(\overline{[\varphi]}^{-1})$, where $B_k := \big([\varphi]_{j-k}\,[f]_{j-k}\big)_{j,k=-k,\dots,k}$. Consequently,

$$\sup_{t\in\mathcal{B}_k}\frac{1}{n}\sum_{k=1}^{n}\mathbf{V}\mathrm{ar}(r_t(Y_k)) \leqslant \|D_k(\sqrt{\omega}\ [\varphi]^{-1})\|^2_{\mathbb{C}^{2k+1}}\ \|B_k\|_{\mathbb{C}^{2k+1}}.$$

We have that $\|D_k(\sqrt{\omega}\ [\varphi]^{-1})\|^2_{\mathbb{C}^{2k+1}} = \max_{0\leqslant|j|\leqslant k}\omega_j|[\varphi]_j|^{-2} \leqslant \Delta_k^*$. It remains to show the boundedness of $\|B_k\|_{\mathbb{C}^{2k+1}}$. Let ℓ^2 be the space of square-summable sequences in $\mathbb{C}$ and define the operator $B:\ell^2\to\ell^2$ by $(Bz)_k := \sum_{j\in\mathbb{Z}}[\varphi]_{j-k}\,[f]_{j-k}z_j$, $k\in\mathbb{Z}$. Then it is easily verified that for any $z\in\ell^2$ with $\|z\|_{\ell^2}=1$, the Cauchy-Schwarz inequality yields $\|Bz\|^2_{\ell^2}\leqslant\|\varphi\|^2\ \|f\|^2$, and hence $\|B\|^2_{\ell^2}\leqslant\|\varphi\|^2\ \|f\|^2$. Given the orthogonal projection Π_k in ℓ^2 onto $\mathcal{S}_k$ the operator $\Pi_k\,B\,\Pi_k:\mathcal{S}_k\to\mathcal{S}_k$ has the matrix representation B_k via the isomorphism $\mathcal{S}_k\cong\mathbb{C}^{2k+1}$ and hence $\|\Pi_k\,B\,\Pi_k\|_{\ell^2}=\|B_k\|_{\mathbb{C}^{2k+1}}$. Orthogonal projections having a norm bounded by 1, we conclude that $\|B_k\|_{\mathbb{C}^{2k+1}}\leqslant\|B\|_{\ell^2}$ for all $k\in\mathbb{N}$, which implies

$$\sup_{t\in\mathcal{B}_k}\frac{1}{n}\sum_{k=1}^{n}\mathbf{V}\mathrm{ar}(r_t(Y_k)) \leqslant \|\varphi\|^2\ \|f\|^2\,\Delta_k^* =: v$$

and thus completes the proof. □

Lemma 3.28 *For every $m\geqslant 1$ and $k\geqslant 0$ we have*

$$\sup_{f\in\mathcal{F}^r_\gamma}\mathbf{E}\bigg[\sup_{t\in\mathcal{B}_k}|\langle t,\widehat{\Phi}_g-\widetilde{\Phi}_g\rangle_\omega|^2\bigg] \leqslant C\,r\ \max_{j\in\mathbb{N}}\Big\{\frac{\omega_j}{\gamma_j}\min\Big(1,\frac{1}{m[\varphi]_j^2})\Big)\Big\} \leqslant C\,d\,r\,\kappa_m(\gamma,\lambda,\omega).$$

Proof. Firstly, as $f\in\mathcal{F}^r_\gamma$, it is easily seen that

$$\mathbf{E}\bigg[\sup_{t\in\mathcal{B}_k}|\langle t,\widehat{\Phi}_g-\widetilde{\Phi}_g\rangle_\omega|^2\bigg] \leqslant r\ \sup_{-k\leqslant j\leqslant k}\frac{\omega_j}{\gamma_j}\,\mathbf{E}[|R_j|^2],$$

where R_j is defined by

$$R_j := \left(\frac{[\varphi]_j}{\widehat{[\varphi]}_j}\mathbf{1}_{[|\widehat{[\varphi]}_j|^2\geqslant 1/m]}-1\right).$$

The result then follows from $\mathbf{E}[|R_j|^2] \leqslant C \min\Big\{1, \frac{1}{m|[\varphi]_j|^2}\Big\}$, which can be shown as follows. Consider the identity

$$\mathbf{E}|R_j|^2 = \mathbf{E}\Big[\,\Big|\frac{[\varphi]_j}{\widehat{[\varphi]}_j} - 1\Big|^2 \mathbf{1}_{[|\widehat{[\varphi]}_j|^2 \geqslant 1/m]}\,\Big] + \mathbf{P}[|\widehat{[\varphi]}_j|^2 < 1/m] =: R_j^I + R_j^{II}.$$

Trivially, $R_j^{II} \leqslant 1$. If $1 \leqslant 4/(m\,|[\varphi]_j|^2)$, then $R_j^{II} \leqslant 4 \min\Big\{1, \frac{1}{m|[\varphi]_j|^2}\Big\}$. Otherwise, we have $1/m < |[\varphi]_j|^2/4$ and hence, using Chebychev's inequality,

$$R_j^{II} \leqslant \mathbf{P}[|\widehat{[\varphi]}_j - [\varphi]_j| > |[\varphi]_j|\,/2\,] \leqslant \frac{4\ \mathbf{V}\mathrm{ar}(\widehat{[\varphi]}_j)}{|[\varphi]_j|^2} \leqslant 4 \min\Big\{1, \frac{1}{m|[\varphi]_j|^2}\Big\},$$

where we have used that $\mathbf{V}\mathrm{ar}(\widehat{[\varphi]}_j) \leqslant m^{-1}$ for all j. Now consider R_j^I. We find that

$$R_j^I = \mathbf{E}\Big[\frac{|\widehat{[\varphi]}_j - [\varphi]_j|^2}{|\widehat{[\varphi]}_j|^2}\ \mathbf{1}_{[|\widehat{[\varphi]}_j|^2 \geqslant 1/m]}\Big] \leqslant m\,\mathbf{V}\mathrm{ar}(|\widehat{[\varphi]}_j) \leqslant 1. \tag{3.33}$$

On the other hand, using that $\mathbf{E}[|\widehat{[\varphi]}_j - [\varphi]_j|^4] \leqslant C/m^2$ (cf. Theorem A.3 in the appendix), we obtain

$$\begin{aligned} R_j^I &\leqslant \mathbf{E}\Big[\frac{|\widehat{[\varphi]}_j - [\varphi]_j|^2}{|\widehat{[\varphi]}_j|^2}\ \mathbf{1}_{[|\widehat{[\varphi]}_j|^2 \geqslant 1/m]}\ 2\Big\{\frac{|\widehat{[\varphi]}_j - [\varphi]_j|^2}{|[\varphi]_j|^2} + \frac{|\widehat{[\varphi]}_j|^2}{|[\varphi]_j|^2}\Big\}\Big] \\ &\leqslant \frac{2\,m\,\mathbf{E}[|\widehat{[\varphi]}_j - [\varphi]_j|^4]}{|[\varphi]_j|^2} + \frac{2\ \mathbf{V}\mathrm{ar}(\widehat{[\varphi]}_j)}{|[\varphi]_j|^2} \leqslant \frac{2\,C}{m\,|[\varphi]_j|^2} + \frac{2}{m\,|[\varphi]_j|^2}. \end{aligned}$$

Combining with (3.33) yields $R_j^I \leqslant 2(C+1) \min\Big\{1, \frac{1}{m|[\varphi]_j|^2}\Big\}$, which completes the proof. □

Lemma 3.29 *Under Assumption 3.2,* $\mathbf{P}[\Omega_q^c] \leqslant C(d)\,m^{-6}$ *for all* $m \geqslant 1$.

Proof. The estimate is obvious for $m < \exp(512\log(3d)^2) =: m_0$. Consider the complement of Ω_q given by

$$\Omega_q^c = \Big\{\exists\ 0 < |j| \leqslant M_m^u \ \Big|\ \Big|\frac{[\varphi]_j}{\widehat{[\varphi]}_j} - 1\Big| > \frac{1}{2}\ \vee\ |\widehat{[\varphi]}_j|^2 < 1/m\Big\}.$$

Due to Lemma 3.26 (iii), we have $|[\varphi]_j|^2 \geqslant 2/m$ for all $m \geqslant m_0$ and for all $0 < |j| \leqslant M_m^u$. This yields

$$\Omega_q^c \subseteq \Big\{\exists\ 0 < |j| \leqslant M_m^u \ \Big|\ \Big|\frac{\widehat{[\varphi]}_j}{[\varphi]_j} - 1\Big| > \frac{1}{3}\Big\}.$$

By Hoeffding's inequality (Theorem A.4), for all $0 < |j| \leqslant M_m^u$

$$\mathbf{P}[|\widehat{[\varphi]}_j/[\varphi]_j - 1| > 1/3] \leqslant 2\,\exp\left(-\frac{m\,|[\varphi]_j|^2}{72}\right) \leqslant 2\exp\left(-\frac{m\,\lambda_{M_m^u}}{72d}\right) \quad (3.34)$$

which implies the result by Lemma 3.26 (ii). □

Lemma 3.30 *Under Assumptions 3.2 and 3.16, the event* Ω_p *defined in* (3.26) *satisfies*

$$\mathbf{P}(\Omega_p^c) \leqslant C(\lambda, d)\, m^{-6} \qquad \forall\, n, m \geqslant 1.$$

Proof. Let $\Omega_I := \{(N_n^\lambda \wedge M_m^\lambda) > (\widehat{N}_n \wedge \widehat{M}_m)\}$ and $\Omega_{II} := \{(\widehat{N}_n \wedge \widehat{M}_m) > (N_n^u \wedge M_m^u)\}$. Then we have $\Omega_p^c = \Omega_I \cup \Omega_{II}$. Consider

$$\Omega_I = \{\widehat{N}_n < (N_n^\lambda \wedge M_m^\lambda)\} \cup \{\widehat{M}_m < (N_n^\lambda \wedge M_m^\lambda)\}$$

first. By definition of N_n^λ, we have that $\min_{1\leqslant|j|\leqslant N_n^\lambda} \frac{|[\varphi]_j|^2}{|j|\,\omega_j^+} \geqslant \frac{4(\log(n+2))}{n}$, which implies

$$\begin{aligned}\{\widehat{N}_n < (N_n^\lambda \wedge M_m^\lambda)\} &\subset \left\{\exists 1 \leqslant |j| \leqslant (N_n^\lambda \wedge M_m^\lambda) \;\middle|\; \frac{|\widehat{[\varphi]}_j|^2}{|j|\,\omega_j^+} < \frac{\log(n+2)}{n}\right\} \\ &\subset \bigcup_{1\leqslant|j|\leqslant N_n^\lambda \wedge M_m^\lambda} \left\{\frac{|\widehat{[\varphi]}_j|}{|[\varphi]_j|} \leqslant 1/2\right\} \subset \bigcup_{1\leqslant|j|\leqslant N_n^\lambda \wedge M_m^\lambda} \left\{\left|\frac{\widehat{[\varphi]}_j}{[\varphi]_j} - 1\right| \geqslant 1/2\right\}.\end{aligned}$$

One can see that from $\min_{1\leqslant|j|\leqslant M_m^\lambda} |[\varphi]_j|^2 \geqslant 4m^{-1+b_m}$ it follows in the same way that

$$\left\{\widehat{M}_m < (N_n^\lambda \wedge M_m^\lambda)\right\} \subset \bigcup_{1\leqslant|j|\leqslant N_n^\lambda \wedge M_m^\lambda} \left\{\left|\frac{\widehat{[\varphi]}_j}{[\varphi]_j} - 1\right| \geqslant 1/2\right\}.$$

Therefore, $\Omega_I \subset \bigcup_{1\leqslant|j|\leqslant M_m^u}\left\{|\widehat{[\varphi]}_j/[\varphi]_j - 1| \geqslant 1/2\right\}$, since $M_m^\lambda \leqslant M_m^u$. Hence, applying Hoeffding's inequality and Lemma 3.26 (ii) as in (3.34) yields

$$\mathbf{P}[\Omega_I] \leqslant \sum_{1\leqslant|j|\leqslant M_m^u} 2\,\exp\left(-\frac{m\,|[\varphi]_j|^2}{72}\right) \leqslant C(d)\, m^{-6}. \quad (3.35)$$

Consider $\Omega_{II} = \{\widehat{N}_n > (N_n^u \wedge M_m^u)\} \cap \{\widehat{M}_m > (N_n^u \wedge M_m^u)\}$. In case $(N_n^u \wedge M_m^u) = N_n^u$, use $\frac{\log(n+2)}{4n} \geqslant \max_{|j| \geqslant N_n^u+1} \frac{|[\varphi]_j|^2}{|j|\,\omega_j^+}$, such that

$$\Omega_{II} \subset \{\widehat{N}_n > N_n^u\} \subset \left\{ \forall 1 \leqslant |j| \leqslant N_n^u + 1 \;\middle|\; \frac{|\widehat{[\varphi]}_j|^2}{|j|\,\omega_j^+} \geqslant \frac{\log(n+2)}{n} \right\}$$
$$\subset \left\{ \frac{|\widehat{[\varphi]}_{N_n^u+1}|}{|[\varphi]_{N_n^u+1}|} \geqslant 2 \right\} \subset \left\{ |\widehat{[\varphi]}_{N_n^u+1}/[\varphi]_{N_n^u+1} - 1| \geqslant 1 \right\}.$$

In case $(N_n^u \wedge M_m^u) = M_m^u$, it follows from $m^{-1+b_m} \geqslant 4 \max_{|j| \geqslant M_m^u+1} |[\varphi]_j|^2$ that

$$\Omega_{II} \subset \{\widehat{M}_m > M_m^u\} \subset \left\{ |\widehat{[\varphi]}_{M_m^u+1}/[\varphi]_{M_m^u+1} - 1| \geqslant 1 \right\}.$$

Therefore, we have $\Omega_{II} \subset \left\{ |\widehat{[\varphi]}_{(N_n^u \wedge M_m^u)+1}/[\varphi]_{(N_n^u \wedge M_m^u)+1} - 1| \geqslant 1 \right\}$. Applying Hoeffding's inequality as in (3.34) and using Assumption 3.16, we obtain for all $m \geqslant 1$

$$\mathbf{P}[\Omega_{II}] \leqslant 2 \exp\left(- \frac{m\,|[\varphi]_{M_m^u+1}|^2}{72} \right) \leqslant C(\lambda, d)\, m^{-6}. \tag{3.36}$$

Combining (3.35) and (3.36) implies the result. □

3.5 Conclusion

In this chapter, we have developed a minimax theory for the circular deconvolution problem with two independent samples. In particular, we have shown lower risk bounds in each of the two sample sizes. We have defined an orthogonal series estimator that can attain the lower risk bounds and we have shown its minimax optimality under an appropriate choice of the regularization parameter. Finally, we have defined a data-driven choice of this parameter and we have proved that the resulting adaptive estimator is still minimax optimal for a wide range of deconvolution and error density classes.

Minimax optimality means that the lower bound for the maximal risk is attained up to a numerical constant. While our results are exact and not asymptotic, this constant is however fairly large. Although some constant-tuning might be possible in order to optimize the results, there is probably not very much room for substantial improvement. It seems that the adaptation over a wide range of density classes, including both ordinary and super smooth functions, has to be paid for in terms of constants. It could be interesting, however, to consider the results in the ordinary smooth case only and to compare the resulting constants in the results with those in Cavalier and Hengartner (2005).

As far as the size m_n of the ε-sample is concerned, a natural question is how to choose it in advance. The illustrations in this chapter show that in some cases, m_n has to grow sufficiently fast in n in order to obtain the same optimal rates as if the error density were known. However, the necessary rate of divergence for m_n depends on the classes $\mathcal{F}_\gamma^r$ and $\mathcal{E}_\lambda^d$. If no information at all about the classes is available, the choice $m_n = n$ will yield the desired result. However, one might want to reduce the size of the error sample by devising a procedure to estimate the required sample size. For instance, one could start with a small m and perform additional calibration measurements successively until the resulting estimator does not change much anymore by adding further measurements. It is however not obvious what stopping criterion could yield satisfactory results.

Another interesting question is how identification could be preserved if the variable X is not supported on the circle, but on an unknown compact interval of the real line. Due to the compactness of the support, the density f would still have a series representation, but the basis obviously depends on the support.

Moreover, as in the case of Chapter 1, an interesting problem would be the renunciation of the independence between X and ε and an investigation of identifiability conditions in this case. Finally, one could ask if the methods presented in this chapter still work when the two samples are dependent or when there is a time series structure in the observations. The following proofs would be affected: In the proof of the lower bound in the case of a known error density in Theorem 3.3, we have used the iid. structure of the observations when controlling the Hellinger affinity of the candidate models in Lemma 3.21. The same remark holds for the proof of the second lower bound in Theorem 3.5 where we need the independence of the two samples in addition. We would therefore need new technical tools in order to control the Hellinger affinity of dependent variables. In the proof of the upper bounds, we have used the independence of the two samples when applying Petrov's and Talagrand's inequalities. Again, new technical results would be needed in order to obtain similar results. Additional assumptions such as mixing conditions would probably be required.

Finally, let us consider the case where some Fourier coefficients of the error density are zero, that is $[\varphi]_j = 0$ for $j \in \mathcal{J}_0$ for some unknown set $\mathcal{J}_0 \subset \mathbb{Z}$. Note that $-j \in \mathcal{J}_0$ if and only if $j \in \mathcal{J}_0$. Suppose that the other coefficients of φ still follow the decay imposed by the class $\mathcal{E}_\lambda^d$, that is

$$d^{-1} \leqslant \frac{|[\varphi]_j|^2}{\lambda_j} \leqslant d \quad \forall\, j \in \mathbb{Z} \setminus \mathcal{J}_0.$$

The convolution theorem states that $[g]_j = [\varphi]_j\,[f]_j$ for all $j \in \mathbb{Z}$. As a consequence, the solution's coefficients $[f]_j$ with $j \in \mathcal{J}_0$ are not identifiable. We can

thus only identify the function

$$f^+ = 1 + \sum_{\substack{0<|j| \\ j \notin \mathcal{J}_0}} [f]_j \, e_j.$$

Consequently, we cannot expect the risk $\mathbf{E}\|\widehat{f}_k - f\|_\omega^2$ to tend to zero. Let us therefore rather consider the risk corresponding to the identified part of the solution, that is $\mathbf{E}\|\widehat{f}_k - f^+\|_\omega^2$. In order to compute this risk, decompose the estimator $\widehat{f}_k$ according to the set $\mathcal{J}_0$, that is

$$\widehat{f}_k^+ := 1 + \sum_{\substack{0<|j|\leqslant k \\ j \notin \mathcal{J}_0}} \frac{\widehat{[g]}_j}{\widehat{[\varphi]}_j} \mathbf{1}_{[|\widehat{[\varphi]}_j|^2 \geqslant 1/m]} e_j \qquad \widehat{f}_k^\circ := \sum_{\substack{0<|j|\leqslant k \\ j \in \mathcal{J}_0}} \frac{\widehat{[g]}_j}{\widehat{[\varphi]}_j} \mathbf{1}_{[|\widehat{[\varphi]}_j|^2 \geqslant 1/m]} e_j.$$

Obviously, we have $\widehat{f}_k = \widehat{f}_k^+ + \widehat{f}_k^\circ$, but note that we cannot compute the parts of the estimator in practice as we do not know the set $\mathcal{J}_0$. The risk can be written as

$$\mathbf{E}\|\widehat{f}_k - f^+\|_\omega^2 = \mathbf{E}\|\widehat{f}_k^+ - f^+\|_\omega^2 + \mathbf{E}\|\widehat{f}_k^\circ\|_\omega^2.$$

The control of the first term follows immediately from Theorem 3.7. It remains to show that the second term is of negligible order. It can be written as

$$\mathbf{E}\|\widehat{f}_k^\circ\|_\omega^2 = \sum_{\substack{0<|j|\leqslant k \\ j \in \mathcal{J}_0}} \omega_j \mathbf{E}[|\widehat{[g]}_j|^2] \, \mathbf{E}[|\widehat{[\varphi]}_j|^{-2} \mathbf{1}_{[|\widehat{[\varphi]}_j|^2 \geqslant 1/m]}].$$

Noting that $\mathbf{E}[|\widehat{[g]}_j|^2] \leqslant n^{-1}$ and, by Hoeffding's inequality (cf. Theorem A.4), $\mathbf{E}[|\widehat{[\varphi]}_j|^{-2} \mathbf{1}_{[|\widehat{[\varphi]}_j|^2 \geqslant 1/m]}] \leqslant C m^{-1}$, we conclude that the remainder term $\mathbf{E}\|\widehat{f}_k^\circ\|_\omega^2$ is of negligible order. Thus, even if some coefficients of the error density are zero, the results of this chapter still hold for the identifiable part of the solution.

Chapter 4

Non parametric instrumental regression

Non parametric instrumental regression models have attracted increasing attention in the econometrics and statistics literature (e.g. Florens, 2003; Darolles et al., 2001; Newey and Powell, 2003; Hall and Horowitz, 2007; Blundell et al., 2007). In instrumental regression, the dependence of a response Y to the variation of an *endogenous* vector Z of explanatory variables is characterized by

$$Y = \varphi(Z) + U \tag{4.1a}$$

for some error term U. Endogenous means that that Z and U are not stochastically independent ($Z \not\perp U$). Additionally, a vector of exogenous instruments W (meaning that $W \perp U$) such that

$$\mathbf{E}[U|W] = 0 \tag{4.1b}$$

is supposed to be observed. The non parametric relationship is hence modeled by the regression function φ, which is also called structural function in this context. Typical examples of such settings are error-in-variable models, simultaneous equations or treatment models with endogenous selection. It is worth noting that in the presence of instrumental variables, the model equations (4.1a–4.1b) are the natural generalization of a standard parametric model (eg. Amemiya, 1974) to the non parametric situation. This extension has first been introduced by Florens (2003) and Newey and Powell (2003), while its identification has been studied e.g. in Carrasco et al. (2007), Darolles et al. (2001) and Florens et al. (2011). Recent applications and extensions of this approach include non parametric tests of exogeneity (Blundell and Horowitz,

2007), quantile regression models (Horowitz and Lee, 2007), or semi-parametric modeling (Florens et al., 2009), for example.
There is a vast literature on the non parametric estimation of the structural function φ based on a sample of (Y, Z, W). For example, Ai and Chen (2003), Blundell et al. (2007) or Newey and Powell (2003) consider sieve minimum distance estimators, while Darolles et al. (2001), Gagliardini and Scaillet (2006) or Florens et al. (2011) consider penalized least squares estimators. The optimal estimation in a minimax sense has been worked on by Hall and Horowitz (2005) and Chen and Reiss (2011). The authors prove a lower bound for the mean integrated squared error (MISE) and propose an estimator which can attain optimal rates. In the present chapter, we extend this result by considering not only the MISE of the estimation of φ but, more generally, a risk defined with respect to the weighted norm $\|\cdot\|_\omega$ that we have already defined in (3.3) in the previous chapter. This allows us for example to consider the estimation of the derivatives of φ, too.

It has been noticed by Newey and Powell (2003) and Florens (2003) that the non parametric estimation of the structural function φ leads to an ill-posed inverse problem in general. Consider the model equations (4.1a–4.1b). Taking the conditional expectation with respect to the instruments W on both sides in equation (4.1a) yields the conditional moment equation

$$\mathbf{E}[Y|W] = \mathbf{E}[\varphi(Z)|W]. \tag{4.2}$$

Therefore, the estimation of the structural function φ is linked to the inversion of equation (4.2), which is not stable in general and hence an ill-posed inverse problem (for a comprehensive review of inverse problems in econometrics, see Carrasco et al. (2007)). This instability is generally accounted for by the application of regularization techniques which however involve the choice of a smoothing parameter. It is well known that the resulting estimation procedure can attain optimal rates only if this parameter is chosen in an appropriate way. In general, this choice requires knowledge of characteristics of the structural function, such as the number of its derivatives, which are not known in practice. Thus, an essential problem in this theoretical framework is the data driven choice of smoothing parameters. In this chapter, an adaptive method is proposed which indeed does not depend on any properties of φ. However, it still necessitates that some characteristics of the underlying operator be known.

One objective in this chapter is the minimax optimal non parametric estimation of the structural function φ based on an iid. sample of (Y, Z, W) satisfying the model equations (4.1a–4.1b). For the moment being, suppose that the structural function can be represented as $\varphi = \sum_{j=1}^{k} [\varphi]_j e_j$ using only k pre-specified basis functions $e_1, \ldots, e_k$, and that only the coefficients $[\varphi]_j$ with respect to that base are unknown. In this situation, the conditional moment equation (4.2) reduces to a multivariate linear conditional moment equation, that is, $\mathbf{E}[Y|W] = \sum_{j=1}^{k} [\varphi]_j \mathbf{E}[e_j(Z)|W]$. Solving this equation is a classical

textbook problem in econometrics (cf. Pagan and Ullah, 1999). A popular approach consists in replacing the conditional moment equation by an unconditional one: given k functions $f_1, \ldots, f_k$, one can consider k unconditional moment equations instead of the multivariate conditional moment equation, that is, $\mathbf{E}[Y f_l(W)] = \sum_{j=1}^k [\varphi]_j \mathbf{E}[e_j(Z) f_l(W)]$, $l = 1, \ldots, k$. Notice that once the functions $\{f_l\}_{l=1}^k$ are chosen, all the unknown quantities in the unconditional moment equations can be estimated by simply substituting empirical versions for the theoretical expectation. Moreover, a least squares solution of the estimated equation leads to a consistent and asymptoticly normal estimator of the parameter vector $([\varphi]_j)_{j=1}^k$ under mild assumptions. The choice of the functions $\{f_l\}_{l=1}^k$ directly influences the asymptotic variance of the estimator and thus the question of optimal instruments arises (cf. Newey, 1990). One advantage of this approach is that estimator is easily computable. However, in many situations an infinite number of functions $\{e_j\}_{j \geqslant 1}$ and associated coefficients $([\varphi]_j)_{j \geqslant 1}$ is needed to represent the structural function φ. Considering an infinite number of functions $\{f_l\}_{l \geqslant 1}$ then for each $k \geqslant 1$ we could still consider the finite dimensional least squares estimator described above. The choice of the basis functions $\{e_j\}_{j \geqslant 1}$ reflects a priori information about the structural function φ, such as smoothness.

Notice that the dimension k plays the role of a smoothing parameter and one might expect that the estimator of the structural function φ is consistent as k tends to infinity at a suitable rate. Unfortunately, this is not true in general. Let $\varphi_k := \sum_{j=1}^k [\varphi_k]_j e_j$ denote a least squares solution of the reduced unconditional moment equations. This means that the vector of coefficients $([\varphi_k]_j)_{j=1}^k$ minimizes the quantity $\sum_{l=1}^k \{\mathbf{E}[Y f_l(W)] - \sum_{j=1}^k \beta_j \mathbf{E}[e_j(Z) f_l(W)]\}^2$ over all vectors $(\beta_j)_{j=1}^k$. Then, φ_k converges to the true structural function as k tends to infinity only under an additional assumption (the «extended link condition» introduced below) on the basis $\{f_j\}_{j \geqslant 1}$. We are going to develop a least squares estimator $\widehat{\varphi}_k$ of φ based on dimension reduction and thresholding, and we show that it can attain optimal rates of convergence in terms of a weighted risk – provided the choice of the dimension parameter k is made in the optimal way. It is worth to note that all the results in this chapter are obtained without any additional smoothness assumption on the joint density of (Y, Z, W). In fact, such a density need not even exist.

Our main contribution is the development of a method to choose the dimension parameter k in a fully data driven way, that is, not depending on characteristics of φ, and assuming only that the underlying conditional expectation operator is «smoothing» in a sense to be made precise below. The central result of the present chapter states that for this automatic choice $\widehat{k}$, the least squares estimator $\widehat{\varphi}_{\widehat{k}}$ can attain the lower bound up to a constant, and is thus minimax-optimal. The adaptive choice of k is made following the same general model selection methodology which we have used in Chapter 3 and which has been developed in Barron et al. (1999). More specifically, $\widehat{k}$ is again the mini-

mizer of a penalized contrast. We illustrate all of our results by considering the estimation of derivatives of the structural function under a smoothing conditional expectation operator. Typically, one distinguishes finitely and infinitely smoothing such operators. Loubes and Marteau (2009) propose an adaptive estimator for the case where the operator is known to be finitely smoothing. They derive oracle inequalities and obtain convergence rates which differ from the optimal ones by a logarithmic factor. In contrast to this, we provide a unified estimation procedure which can attain minimax-optimal rates in either of the both cases. In other words, our estimation procedure attains optimal rates without knowing in advance if the operator is finitely or infinitely smoothing.

This chapter, which is based on Johannes and Schwarz (2010), is organized as follows. As in the last chapter, we begin by discussing a basic example in the first section, namely a non parametric regression model with exogenous regressors, and thus without the need for instrumental variables. This example allows us to discuss in a simplified framework how the adaptation techniques we have elaborated on in Chapter 3 can be applied in the context of a regression model. In Section 4.2, we develop the minimax theory for the non parametric instrumental regression model with respect to the weighted risk. We derive, as an illustration, the optimal convergence rates for the estimation of derivatives in the finitely and in the infinitely smoothing case. Finally, in Section 4.3, we construct the adaptive estimator. An upper risk bound is shown and convergence rates for the finitely and infinitely smoothing case are found to coincide with minimax optimal ones. Some auxiliary results are deferred to the end of the chapter.

4.1 An introductory example

As in Chapter 3, in order to give an account of the techniques to be used further on, we begin with a basic example. Consider the non parametric regression model

$$Y = \varphi(X) + U,$$

where $\varphi \in L^2[0,1]$ is the real-valued regression function, $X \sim U[0,1]$ the uniformly distributed exogenous regressor, and $U \sim \mathcal{N}(0,1)$ a normally distributed error. Suppose that we observe an iid. sample $(Y_j, X_j)_{j=1,\dots,n}$. As in Section 3.1, let $(e_j)_{j\in\mathbb{N}}$ denote the exponential basis of the Hilbert space $L^2[0,1]$. For every $j \in \mathbb{N}$, a natural and unbiased estimator of the coefficient $[\varphi]_j$ is given by $\widehat{[\varphi]}_j := n^{-1}\sum_{k=1}^n Y_k e_j(-X_k)$. An orthogonal series estimator of φ can then be defined by $\widehat{\varphi}_k = \sum_{j=1}^k \widehat{[\varphi]}_j\, e_j$.

As we want to develop minimax theory again, consider the case where the regression function φ lies in the class $\mathcal{F}_\gamma^\rho$ defined below in (4.4). In this introductory section, we only consider the special case where $\gamma_0 = 1$ and $\gamma_j = j^{2p}$

for $|j| > 1$. As in the case of the deconvolution model considered in Section 3.1, the maximal risk over this class is bounded by

$$\sup_{\varphi \in \mathcal{F}_\gamma^\rho} \mathbf{E}\|\widehat{\varphi}_k - \varphi\|^2 \leqslant k n^{-1} + \rho k^{-2p}.$$

The optimal choice for k is $k = k_n^* = n^{1/(2p+1)}$, and the resulting upper bound

$$\sup_{\varphi \in \mathcal{F}_\gamma^\rho} \mathbf{E}\|\widehat{\varphi}_{k_n^*} - \varphi\|^2 \leqslant C\, n^{-2p/(2p+1)},$$

which is known to be the optimal minimax risk (c.f. Tsybakov, 2004). The optimal choice k_n^* still depends on the parameter p. Therefore, as in Section 3.1 of the previous chapter, we define the data driven choice

$$\check{k} := \operatorname*{argmin}_{k=1,\dots,n} \{-\|\widehat{\varphi}_k\|^2 + c\, k\, n^{-1}\}$$

for some constant c. Letting $\Phi_{\widehat{\varphi}} := \sum_{j\in\mathbb{N}} \widehat{[\varphi]}_j\, e_j$ and following exactly the lines of the argument explained in Section 3.1, one shows that

$$\|\widehat{\varphi}_{\check{k}} - \varphi\|^2 \leqslant \|\varphi - \varphi_k\|^2 + \operatorname{pen}(k) - \operatorname{pen}(\check{k}) + 2\langle \widehat{\varphi}_{\check{k}} - \varphi_k, \Phi_{\widehat{\varphi}} - \varphi\rangle$$

and then

$$\begin{aligned}\sup_{\varphi \in \mathcal{F}_\gamma^\rho} \mathbf{E}\|\widehat{\varphi}_{\check{k}} - \varphi\|^2 \leqslant & C\, n^{-2p/(2p+1)} \\ & + C\, \mathbf{E}\Bigg[\sum_{k'=1}^{n} \Bigg(\sup_{t\in\mathcal{B}_{k'}} |\langle t, \Phi_{\widehat{\varphi}} - \varphi\rangle|^2 - c\,(k')/n\Bigg)_+\Bigg].\end{aligned} \tag{4.3}$$

The remainder term can be controlled by a concentration inequality for the normal distribution. Let us roughly outline the next step in the development of the model. Unlike in this example, we consider an endogenous regressor Z, but we assume that instrumental variables W with $\mathbf{E}[U|W] = 0$ are available. The set of observations consists of n iid. copies of (Y, Z, W). Furthermore, we will not suppose the error U to be normally distributed; instead, we will define general classes of error distributions. Thus, the maximal risk not only depends on a class of solutions, but also on classes of error distributions and of conditional expectation operators.

In this more general setting, the remainder term in (4.3) cannot be controlled by standard concentration inequalities anymore. We need rather to apply Talagrand's inequality, which unlike in Section 3.1 is not possible directly, because the $\widehat{[\varphi]}_j$ are not bounded almost surely – a finer decomposition of the term is required. The result however remains the same in the sense that the remainder term is of order n^{-1} and hence negligible with respect to the minimax rate. The optimality of the adaptive estimator is thus established.

4.2 Minimax optimal estimation

In this section, we develop a minimax theory for the estimation of the structural function and its derivatives in nonparametric instrumental regression models.

4.2.1 Basic model assumptions

It is convenient to rewrite the moment equation (4.2) in terms of an operator between Hilbert spaces. Therefore, let us first introduce the Hilbert spaces

$$L_Z^2 = \big\{\varphi : \mathbb{R}^p \to \mathbb{R} \;\big|\; \|\varphi\|_Z^2 := \mathbf{E}[\varphi^2(Z)] < \infty\big\},$$
$$L_W^2 = \big\{\psi : \mathbb{R}^q \to \mathbb{R} \;\big|\; \|\psi\|_W^2 := \mathbf{E}[\psi^2(W)] < \infty\big\},$$

endowed with the inner products $\langle\varphi,\tilde\varphi\rangle_Z = \mathbf{E}[\varphi(Z)\tilde\varphi(Z)]$, $\varphi,\tilde\varphi \in L_Z^2$, and $\langle\psi,\tilde\psi\rangle_W = \mathbf{E}[\psi(W)\tilde\psi(W)]$, $\psi,\tilde\psi \in L_W^2$, respectively. Then the conditional expectation of Z given W defines a linear operator $T\varphi := \mathbf{E}[\varphi(Z)|W]$, $\varphi \in L_Z^2$, which maps L_Z^2 to L_W^2. The moment equation (4.2) can be rewritten as

$$g := \mathbf{E}[Y|W] = \mathbf{E}[\varphi(Z)|W] =: T\varphi,$$

where the function g belongs to L_W^2. The estimation of the structural function φ is thus linked to the inversion of the conditional expectation operator T. Moreover, we suppose throughout this chapter that the operator T is compact, which is the case under fairly mild assumptions. For example, if the triple (Y,Z,W) has a joint density, it is sufficient to demand that it be square integrable – or continuous, if its support is compact – in order for T to be compact (c.f. Carrasco et al., 2007). Consequently, unlike in a multivariate linear instrumental regression model, a continuous generalized inverse of T does not exist as long as the range of the operator T is an infinite dimensional subspace of L_W^2. This corresponds to the setup of statistical ill-posed inverse problems with unknown operator outlined in the introduction to this thesis. For a detailed discussion in the context of inverse problems see Chapter 2.1 in Engl et al. (1996), while in the special case of a nonparametric instrumental regression we refer to Carrasco et al. (2007). In what follows, we always assume that the joint distribution of (Y,Z,W) is such that $g = \mathbf{E}[Y|W]$ lies in the range of T and that T is injective. Thus, the first two Hadamard conditions are satisfied. Note that this assumption does not imply that every $g \in L_W^2$ has a preimage under T.

4.2.2 Complexity of the problem: a lower bound

In this section we show that the obtainable accuracy of any estimator of the structural function φ is essentially determined by additional regularity conditions imposed on φ and the conditional expectation operator T. In this chapter,

these conditions are characterized through different weighted norms in L^2_Z with respect to a pre-specified orthonormal basis $\{e_j\}_{j\geqslant 1}$ of L^2_Z. We formalize these conditions as follows.

Minimal regularity conditions

As in (3.3) in the previous chapter, given a strictly positive sequence of weights $\omega := (\omega_j)_{j\geqslant 1}$, we denote by $\|\cdot\|_\omega$ the weighted norm given by

$$\|f\|_\omega := \sum_{j=1}^{\infty} \omega_j |\langle f, e_j\rangle_Z|^2, \qquad \forall f \in L^2_Z.$$

We shall measure the accuracy of any estimator $\widehat{\varphi}$ of the unknown structural function in terms of a weighted risk, that is $\mathbf{E}\|\widehat{\varphi} - \varphi\|^2_\omega$, for a pre-specified sequence of weights $\omega := (\omega_j)_{j\geqslant 1}$. This general approach allows as to consider not only the estimation of the structural function itself but also of its derivatives, as we have already discussed in the illustration section of the previous chapter (cf. p. 61). Moreover, given a sequence of weights $\gamma := (\gamma_j)_{j\geqslant 1}$ we suppose, here and subsequently, that for some constant $\rho > 0$ the structural function φ belongs to the ellipsoid

$$\mathcal{F}^\rho_\gamma := \Big\{ f \in L^2_Z \;\Big|\; \|f\|^2_\gamma \leqslant \rho \Big\}, \tag{4.4}$$

which captures all the prior information (such as smoothness) about the unknown structural function φ. Furthermore, as usual in the context of ill-posed inverse problems, we specify the mapping properties of the conditional expectation operator T. Therefore, consider the sequence $(\|Te_j\|_W)_{j\geqslant 1}$, which converges to zero since T is compact. In what follows, we impose restrictions on the decay of this sequence. Denote by $\mathcal{T}$ the set of all injective compact operator mapping L^2_Z to L^2_W. Given a strictly positive sequence of weights $\lambda := (\lambda_j)_{j\geqslant 1}$ and a constant $d \geqslant 1$, we define the subset $\mathcal{T}^d_\lambda$ of $\mathcal{T}$ by

$$\mathcal{T}^d_\lambda := \Big\{ T \in \mathcal{T} \;\Big|\; \|f\|^2_\lambda / d \leqslant \|Tf\|^2_W \leqslant d\,\|f\|^2_\lambda, \quad \forall f \in L^2_Z \Big\}. \tag{4.5}$$

Notice that for all $T \in \mathcal{T}^d_\lambda$ it follows that $d^{-1} \leqslant \|Te_j\|^2_W/\lambda_j \leqslant d$. Furthermore, let us denote by $T^* : L^2_W \to L^2_Z$ the adjoint of T which satisfies $T^*\psi = \mathbf{E}[\psi(W)|Z]$ for all $\psi \in L^2_W$. Let $T \in \mathcal{T}$. Considering the case where $\{e_j\}_{j\geqslant 1}$ are the eigenfunctions of T^*T, one sees immediately that the sequence λ specifies the decay of the eigenvalues of T^*T. All results of this work are derived under regularity conditions on the structural function φ and the conditional expectation operator T described by the sequences γ and λ, respectively. However, below we provide illustrations of these conditions by assuming a «regular decay» of these sequences. The next assumption summarizes our minimal regularity conditions on these sequences.

Assumption 4.1 *Let $\gamma := (\gamma_j)_{j\in\mathbb{N}}$, $\omega := (\omega_j)_{j\in\mathbb{N}}$ and $\lambda := (\lambda_j)_{j\in\mathbb{N}}$ be strictly positive sequences of weights with $\gamma_0 = \omega_0 = \lambda_0 = 1$ and $\Gamma := \sum_{j\in\mathbb{N}} \gamma_j^{-1} < \infty$, such that (ω/γ), (λ/ω), and λ are non-increasing, respectively.*

As in the case of the analogue Assumption 3.2 in Chapter 3, it is worth noting that the monotonicity assumption on (ω/γ) only ensures that $\|\varphi\|_\omega$ is finite for all $\varphi \in \mathcal{F}_\gamma^\rho$, and hence the weighted risk is a well-defined measure of accuracy for estimators of φ. Heuristically, this reflects the fact that we cannot estimate the $(s+1)$-th derivative if the structural function has only s derivatives. In the illustration in Section 4.2.3, the additional assumption $\Gamma := \sum_{j\in\mathbb{N}} \gamma_j^{-1} < \infty$ can be interpreted as a continuity assumption on φ.

The lower bound

The next assertion provides a lower bound for the risk with respect to the weighted norm. Thus, we extend the result of Chen and Reiss (2011), who show a lower bound for the mean integrated squared error.

Theorem 4.2 *Suppose that the iid. (Y, Z, W)-sample of size n obeys the model (4.1a–4.1b), that the distribution of the error term U belongs to the class*

$$\mathcal{U}_\sigma := \{P_U \mid \mathbf{E}[U|W] = 0 \text{ and } \mathbf{E}[U^4|W] \leqslant \sigma^4\}$$

with $\sigma > 0$ and that $\sup_{j\geqslant 1} \mathbf{E}[e_j^4(Z)|W] \leqslant \eta$, $\eta \geqslant 1$. Consider sequences γ, ω and λ satisfying Assumption 4.1 such that the conditional expectation operator T associated to (Z, W) belongs to $\mathcal{T}_\lambda^d$, $d \geqslant 1$. Define for all $n \geqslant 1$

$$k_n^* := k_n^*(\gamma, \lambda, \omega) := \operatorname*{argmin}_{k\in\mathbb{N}} \Big\{ \max\Big(\frac{\omega_k}{\gamma_k}, \sum_{j=1}^{k} \frac{\omega_j}{n\lambda_j} \Big) \Big\} \text{ and}$$

$$R_n^* := R_n^*(\gamma, \lambda, \omega) := \max\Big(\frac{\omega_{k_n^*}}{\gamma_{k_n^*}}, \sum_{j=1}^{k_n^*} \frac{\omega_j}{n\lambda_j} \Big). \quad (4.6)$$

If in addition $\kappa := \inf_{n\geqslant 1}\{(R_n^*)^{-1} \min(\omega_{k_n^*}\gamma_{k_n^*}^{-1}, \sum_{l=1}^{k_n^*} \omega_l (n\lambda_l)^{-1})\} > 0$ and $\sigma^4 \geqslant 8(3 + 2\rho^2\Gamma^2)$, then for all $n \geqslant 1$ and for any estimator $\widetilde{\varphi}$ of φ, we have

$$\sup_{P_U\in\mathcal{U}_\sigma} \sup_{\varphi\in\mathcal{F}_\gamma^\rho} \mathbf{E}\|\widetilde{\varphi} - \varphi\|_\omega^2 \geqslant \frac{\kappa}{4} \min\left(\rho, \frac{1}{2d}\right) R_n^*.$$

Remark 4.3 As the proofs of the lower bounds in the previous chapter, the proof of the last assertion is based on Assouad's cube technique (c.f. Korostolev and Tsybakov, 1993; Tsybakov, 2004), which consists in constructing $2^{k_n^*}$ candidates of structural functions which have the largest possible $\|\cdot\|_\omega$-distance

but are still statistically non distinguishable. In the last theorem, the additional moment condition $\sup_{j\geqslant 1}\mathbf{E}[e_j^4(Z)|W]\leqslant\eta$ is obviously satisfied if the basis functions $\{e_j\}$ are uniformly bounded (e.g. the trigonometric basis considered in Section 4.2.3). However, if V denotes a Gaussian random variable with mean zero and variance one, which is moreover independent of (Z,W), then the additional condition $\sigma^4\geqslant 8(1+2\rho^2\Gamma^2\eta)$ ensures that for all structural functions $\varphi\in\mathcal{F}_\gamma^r$, the distribution of the error term $U:=V-\varphi(Z)+[T\varphi](W)$ belongs to $\mathcal{U}_\sigma$. This specific case is only needed to simplify the calculation of the distance between distributions corresponding to different structural functions. A similar assumption has been used by Chen and Reiss (2011).

On the other hand, below we derive an upper bound assuming that the distribution of error term U belongs to $\mathcal{U}_\sigma$ and that the joint distribution of (Z,W) satisfies additional moment conditions. In this situation, Theorem 4.2 provides a lower bound for any estimator as long as σ is sufficiently large. Note further that this lower bound tends only to zero if ω/γ is a vanishing sequence. In other words, in case $\gamma\equiv 1$, uniform consistency over all φ with $\|\varphi\|_Z^2\leqslant\rho$ can only be achieved with respect to a weighted norm weaker than the L_Z^2-norm, that is, if ω is a sequence tending to zero. This reflects the fact that the ill-posedness of the underlying inverse problem could be redeemed by changing the topological structure of the spaces as we have discussed in the introduction of this thesis. Finally, it is important to note that the regularity conditions imposed on the structural function φ and the conditional expectation operator T involve only the basis $\{e_j\}_{j\geqslant 1}$ in L_Z^2. Therefore, the lower bound derived in Theorem 4.2 does not capture the influence of the basis $\{f_l\}_{l\geqslant 1}$ in L_W^2 used to construct the estimator. In other words, the proposed estimator of φ can only attain this lower bound if $\{f_l\}_{l\geqslant 1}$ is appropriately chosen. □

Proof of Theorem 4.2. Consider a pair (Z,W) with associated conditional expectation operator $T\in\mathcal{T}_d^\lambda$. Let

$$\zeta:=\kappa\min(\rho,1/(2d))\quad\text{and}\quad\alpha_n:=R_n^*\Big(\sum_{j=1}^{k_n^*}\omega_j/(\lambda_j n)\Big)^{-1}.$$

Then, the function $\varphi:=(\zeta\alpha_n/n)^{1/2}\sum_{j=1}^{k_n^*}\lambda_j^{-1/2}e_j$ belongs to the class $\mathcal{F}_\gamma^\rho$, because the monotonicity of (γ/ω) implies $\|\varphi\|_\gamma^2\leqslant\rho\kappa(\gamma_{k_n^*}/\omega_{k_n^*})R_n^*\leqslant\rho$, using successively the definitions of α_n and κ. Based on φ, the candidates for the structural function are defined as

$$\varphi_\theta:=\sum_{j=1}^{k_n^*}\theta_j[\varphi]_j e_j$$

for every $\theta:=(\theta_j)\in\{-1,1\}^{k_n^*}$. These functions obviously belong to $\mathcal{F}_\gamma^\rho$, too. Let $V\sim\mathcal{N}(0,1)$ be a random variable independent of (Z,W). For every

$\theta := (\theta_j) \in \{-1,1\}^{k_n^*}$, the distribution of the random variable

$$U_\theta := [T\varphi_\theta](W) - \varphi_\theta(Z) + V$$

then belongs to $\mathcal{U}_\sigma$ for all $\sigma^4 \geqslant 8(3+2\rho^2\Gamma^2\eta)$: Firstly, $\mathbf{E}[U_\theta|W=0]$. Secondly, we have

$$|\mathbf{E}[f(Z)|W]|^4 \leqslant \rho^2\Gamma \sum_{j\in\mathbb{N}} \gamma_j^{-1}\mathbf{E}[e_j^4(Z)|W] \leqslant \rho^2\Gamma^2\eta$$

for all all $f \in \mathcal{F}_\gamma^\rho$, which follows from the condition $\Gamma = \sum_{j\in\mathbb{N}}\gamma_j^{-1} < \infty$ together with $\sup_j \mathbf{E}[e_j^4(Z)|W] \leqslant \eta$, applying the Cauchy-Schwarz inequality twice. From this estimate we conclude $\mathbf{E}[\varphi_\theta^4(Z)|W] \leqslant \eta\rho^2\Gamma^2$ and $|[T\varphi_\theta](W)|^4 \leqslant \mathbf{E}[\varphi_\theta^4(Z)|W] \leqslant \eta\rho^2\Gamma^2$. By combination of the last two bounds we obtain $\mathbf{E}[U_\theta^4|W] \leqslant 8\{2\eta\rho^2\Gamma^2+3\}$.

Consequently, for any θ, the tuple (Y,Z,W) defined by $Y := \varphi_\theta(Z)+U_\theta$ obeys the model (4.1a–4.1b). Let $(Y_i,Z_i,W_i)_{i=1,\dots,n}$ be n iid. copies of (Y,Z,W) and denote their joint distribution by P_θ.

Under the law P_θ, the conditional distribution of Y_i given W_i is then Gaussian with mean $[T\varphi_\theta](W_i)$ and variance 1. Furthermore, for $j=1,\dots,k_n^*$ and for each θ we introduce $\theta^{(j)}$ by $\theta_l^{(j)} = \theta_l$ for $j \neq l$ and $\theta_j^{(j)} = -\theta_j$. Then, it is easily seen that the log-likelihood of P_θ with respect to $P_{\theta^{(j)}}$ is given by

$$\log\Big(\frac{dP_\theta}{dP_{\theta^{(j)}}}\Big) = \sum_{i=1}^n 2(Y_i - [T\varphi_\theta](W_i))\theta_j[\varphi]_j[Te_j](W_i) + 2[\varphi]_j^2\sum_{i=1}^n |[Te_j](W_i)|^2.$$

Its expectation with respect to P_θ satisfies

$$\mathbf{E}_{P_\theta}[\log(dP_\theta/dP_{\theta^{(j)}})] = 2n[\varphi]_j^2\|Te_j\|_W^2 \leqslant 2nd[\varphi]_j^2\lambda_j,$$

because $T \in \mathcal{T}_d^\lambda$. In terms of the Kullback-Leibler divergence, this means

$$KL(P_\theta, P_{\theta^{(j)}}) \leqslant 2\,d\,n\,[\varphi]_j^2\lambda_j.$$

Since the Hellinger distance satisfies $H^2(P_\theta,P_{\theta^{(j)}}) \leqslant KL(P_\theta,P_{\theta^{(j)}})$, we can use the definition of φ, the property $\alpha_n \leqslant \kappa^{-1}$, and the definition of ζ successively and obtain that

$$H^2(P_\theta,P_{\theta^{(j)}}) \leqslant 2\,d\,n\,[\varphi]_j^2\lambda_j \leqslant 2d\,\zeta\,\alpha_n \leqslant 1. \tag{4.7}$$

Considering the Hellinger affinity $\rho(P_\theta,P_{\theta^{(j)}}) = \int\sqrt{dP_\theta dP_{\theta^{(j)}}}$, we can write for any estimator $\widetilde{\varphi}$ of φ that

$$\begin{aligned}\rho(P_\theta,P_{\theta^{(j)}}) &\leqslant \int \frac{|[\widetilde{\varphi}-\varphi_{\theta^{(j)}}]_j|}{|[\varphi_\theta-\varphi_{\theta^{(j)}}]_j|}\sqrt{dP_\theta dP_{\theta^{(j)}}} + \int\frac{|[\widetilde{\varphi}-\varphi_\theta]_j|}{|[\varphi_\theta-\varphi_{\theta^{(j)}}]_j|}\sqrt{dP_\theta dP_{\theta^{(j)}}}\\ &\leqslant \Big(\int\frac{|[\widetilde{\varphi}-\varphi_{\theta^{(j)}}]_j|^2}{|[\varphi_\theta-\varphi_{\theta^{(j)}}]_j|^2}dP_{\theta^{(j)}}\Big)^{1/2} + \Big(\int\frac{|[\widetilde{\varphi}-\varphi_\theta]_j|^2}{|[\varphi_\theta-\varphi_{\theta^{(j)}}]_j|^2}dP_\theta\Big)^{1/2}.\end{aligned}$$

Rewriting the last estimate using the identity $\rho(P_\theta, P_{\theta^{(j)}}) = 1 - \frac{1}{2}H^2(P_\theta, P_{\theta^{(j)}})$ and (4.7), we obtain

$$\left\{\mathbf{E}_\theta|[\widetilde{\varphi} - \varphi_\theta]_j|^2 + \mathbf{E}_{\theta^{(j)}}|[\widetilde{\varphi} - \varphi_{\theta^{(j)}}]_j|^2\right\} \geqslant \frac{1}{8}|[\varphi_\theta - \varphi_{\theta^{(j)}}]_j|^2 = \frac{1}{2}[\varphi]_j^2.$$

We combine the last estimate with the following reduction scheme, which is the key argument of this proof:

$$\begin{aligned}
\sup_{P_U \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{F}_\gamma^\rho} \mathbf{E}_{P_\theta}\|\widetilde{\varphi} - \varphi\|_\omega^2 &\geqslant \sup_{\theta \in \{-1,1\}^{k_n^*}} \mathbf{E}_{P_\theta}\|\widetilde{\varphi} - \varphi_\theta\|_\omega^2 \\
&\geqslant \frac{1}{2^{k_n^*}} \sum_{\theta \in \{-1,1\}^{k_n^*}} \sum_{j=1}^{k_n^*} \omega_j \mathbf{E}_{P_\theta}|[\widetilde{\varphi} - \varphi_\theta]_j|^2 \\
&= \frac{1}{2^{k_n^*}} \sum_{\theta \in \{-1,1\}^{k_n^*}} \sum_{j=1}^{k_n^*} \frac{\omega_j}{2}\left\{\mathbf{E}_{P_\theta}|[\widetilde{\varphi} - \varphi_\theta]_j|^2 + \mathbf{E}_{P_{\theta^{(j)}}}|[\widetilde{\varphi} - \varphi_{\theta^{(j)}}]_j|^2\right\} \\
&\geqslant \frac{1}{2^{k_n^*}} \sum_{\theta \in \{-1,1\}^{k_n^*}} \sum_{j=1}^{k_n^*} \frac{\omega_j}{4}[\varphi]_j^2 = \frac{\zeta\alpha_n}{4} \sum_{j=1}^{k_n^*} \frac{\omega_j}{n\lambda_j}.
\end{aligned}$$

Hence, from the definition of ζ and α_n we obtain the lower bound given in the theorem. □

4.2.3 Minimax-optimal Estimation by dimension reduction and thresholding

In addition to the basis $\{e_j\}_{j\geqslant 1}$ of L_Z^2 considered in the last section, we introduce now a basis $\{f_l\}_{l\geqslant 1}$ in L_W^2. In this section we derive the asymptotic properties of the least squares estimator under minimal assumptions on these two bases. More precisely, we suppose that the structural function φ belongs to some ellipsoid $\mathcal{F}_\gamma^r$ and that the conditional expectation satisfies a link condition, i.e., $T \in \mathcal{T}_d^\lambda$. Furthermore, we introduce an additional condition linked to the basis $\{f_l\}_{l\geqslant 1}$. Then, under slightly stronger moment conditions, we show that the proposed estimator attains the lower bound derived in the last section. All these results are illustrated under classical smoothness assumptions at the end of this section.

Matrix and operator notations

Given $k \geqslant 1$, let $\mathcal{E}_k$ and $\mathcal{F}_k$ denote the subspace of L_Z^2 and L_W^2 spanned by the functions $\{e_j\}_{j=1}^k$ and $\{f_l\}_{l=1}^k$, respectively. E_k and $E_k^\perp$ (resp. F_k and $F_k^\perp$) denote the orthogonal projection mappings on $\mathcal{E}_k$ (resp. $\mathcal{F}_k$) and its orthogonal

complement $\mathcal{E}_k^\perp$ (resp. $\mathcal{F}_k^\perp$), respectively. Given a matrix K, its inverse is denoted by K^{-1} and its transposed matrix by K^t. Let $[\varphi]$, $[\psi]$ and $[K]$ denote the (infinite) vector and matrix of the function $\varphi \in L_Z^2$, $\psi \in L_W^2$ and the operator $K : L_Z^2 \to L_W^2$ with the entries $[\varphi]_j = \langle \varphi, e_j \rangle$, $[\psi]_l = \langle \psi, f_l \rangle$ and $[K]_{lj} = \langle Ke_j, f_l \rangle$, respectively. The upper k-sub-vector and $(k \times k)$-sub-matrix of $[\varphi]$, $[\psi]$ and $[K]$ are denoted by $[\varphi]_{\underline{k}}$, $[\psi]_{\underline{k}}$ and $[K]_{\underline{k}}$, respectively. Note that $[K^*]_{\underline{k}} = [K]_{\underline{k}}^t$. The diagonal matrix with entries v is denoted by $\mathrm{diag}(v)$ and the identity matrix is denoted by I. Clearly, $[E_k\varphi]_{\underline{k}} = [\varphi]_{\underline{k}}$ and if we restrict F_kKE_k to an operator from $\mathcal{E}_k$ into $\mathcal{F}_k$, then it has the matrix $[K]_{\underline{k}}$. Moreover, if $v \in \mathbb{R}^k$ then $\|v\|$ denotes the Euclidean norm of v, and given a $(k \times k)$-matrix M, let $\|M\| := \sup_{\|v\| \leqslant 1} \|Mv\|$ denote its spectral-norm and $\mathrm{tr}(M)$ its trace.

Consider the conditional expectation operator T associated to the regressor Z and the instrument W. If $[e(Z)]$ and $[f(W)]$ denote the infinite random vector with entries $e_j(Z)$ and $f_j(W)$ respectively, then $[T]_{\underline{k}} = \mathbf{E}[f(W)]_{\underline{k}}[e(Z)]_{\underline{k}}^t$ which is, throughout the chapter, assumed to be non singular for all $k \geqslant 1$ (or, at least for sufficiently large k), such that $[T]_{\underline{k}}^{-1}$ always exists. Note that it is a nontrivial problem to determine in under what precise conditions such an assumption holds (see e.g. Efromovich and Koltchinskii (2001) and references therein).

Definition of the estimator

Let $(Y_1, Z_1, W_1), \ldots, (Y_n, Z_n, W_n)$ be an iid. sample of (Y, Z, W). Since $[T]_{\underline{k}} = \mathbf{E}[f(W)]_{\underline{k}}[e(Z)]_{\underline{k}}^t$ and $[g]_{\underline{k}} = \mathbf{E}Y[f(W)]_{\underline{k}}$ can be written as expectations, we can construct estimators by using their empirical counterparts, that is,

$$\widehat{[T]}_{\underline{k}} := (1/n) \sum_{i=1}^n [f(W_i)]_{\underline{k}}[e(Z_i)]_{\underline{k}}^t \quad \text{and} \quad \widehat{[g]}_{\underline{k}} := (1/n) \sum_{i=1}^n Y_i [f(W_i)]_{\underline{k}}.$$

Then the estimator of the structural function φ is defined by

$$\widehat{\varphi}_k := \sum_{j=1}^k [\widehat{\varphi}_k]_j e_j \text{ with } [\widehat{\varphi}_k]_{\underline{k}} := \begin{cases} \widehat{[T]}_{\underline{k}}^{-1} \widehat{[g]}_{\underline{k}}, & \text{if } \widehat{[T]}_{\underline{k}} \text{ is nonsingular} \\ & \text{and } \|\widehat{[T]}_{\underline{k}}^{-1}\| \leqslant \sqrt{n}, \\ 0, & \text{otherwise,} \end{cases} \tag{4.8}$$

where the dimension parameter $k = k(n)$ has to tend to infinity as the sample size n increases. This estimator $\widehat{\varphi}_k$ takes its inspiration from the linear Galerkin approach (c.f. Efromovich and Koltchinskii, 2001; Hoffmann and Reiss, 2008). One could also consider other regularization techniques in this context, for example the Tikhonov regularization which has been briefly discussed in the introduction. But the Tikhonov method has a great disadvantage, namely the so called *saturation effect* which is an absolute upper bound on the convergence

speed of the regularized solution. This limitation prevents the estimator from attaining optimal convergence rates for a wide range of functional classes $\mathcal{F}_\gamma^\rho$. For a more detailed discussion of this phenomenon see for example Engl et al. (1996).

Note that the estimator $[\widehat{T}]_{\underline{k}}$ only depends on the observations of the couple (Z, W) and the estimator $[\widehat{g}]_{\underline{k}}$ only on (Y, W). Instead of supposing iid. observations of the triple (Y, Z, W), one could therefore work as well with two separated samples of the couples just mentioned. These two samples could even be of different size. This case may possibly be handled taking into account the theoretical framework of the previous chapter. In this chapter, however, we base our estimation on iid. observations of (Y, Z, W), which corresponds to the case where we have two samples of equal size.

Extended link condition

Consistency of this estimator is only possible if the least squares solution $\varphi_k = \sum_{j=1}^k [\varphi_k]_j e_j$ with $[\varphi_k]_{\underline{k}} = [T]_{\underline{k}}^{-1}[g]_{\underline{k}}$ converges to φ as $k \to \infty$, which is not true in general. However, the condition $\sup_{k\in\mathbb{N}} \|[T]_{\underline{k}}^{-1}[TE_k^\perp]_{\underline{k}}\| < \infty$ is known to be necessary to ensure convergence of φ_k. Notice that this condition involves also the basis $\{f_l\}_{l\geqslant 1}$ in L_W^2. In what follows, we introduce an alternative but stronger condition to guarantee the convergence, which extends the link condition (4.5), that is, $T \in \mathcal{T}_d^\lambda$. We denote by $\mathcal{T}_{d,D}^\lambda$ for some $D \geqslant d$ the subset of $\mathcal{T}_d^\lambda$ given by

$$\mathcal{T}_{d,D}^\lambda := \Big\{ T \in \mathcal{T}_d^\lambda \;\Big|\; \sup_{k\in\mathbb{N}} \|[\mathrm{diag}(\lambda)]_{\underline{k}}^{1/2}[T]_{\underline{k}}^{-1}\|^2 \leqslant D \Big\}. \tag{4.9}$$

Remark 4.4 The link condition (4.5) implies the extended link condition (4.9) for a suitable $D > 0$ if $\{e_j\}$ and $\{f_j\}$ are the singular functions of T and if $[T]$ is only a small perturbation of $\mathrm{diag}(\lambda^{1/2})$, or if T is strictly positive (for a detailed discussion we refer to Efromovich and Koltchinskii (2001) and Cardot and Johannes (2010)). We underline that once both bases $\{e_j\}_{j\geqslant 1}$ and $\{f_l\}_{l\geqslant 1}$ are specified, the extended link condition (4.9) restricts the class of joint distributions of (Z, W) to those for which the least squares solution φ_k is L^2-consistent. Moreover, we show below that under the extended link condition the least squares estimator of φ given in (4.8) can attain minimax-optimal rates of convergence. In this sense, given a joint distribution of (Z, W), a basis $\{f_l\}_{l\geqslant 1}$ satisfying the extended link condition can be interpreted as a set of optimal instruments. Furthermore, for each pre-specified basis $\{e_j\}_{j\geqslant 1}$, we can theoretically construct a basis $\{f_l\}_{l\geqslant 1}$ of optimal instruments such that the extended link condition is not a stronger restriction than the link condition (4.5) (see Johannes and Breunig (2009) for more details). □

The upper bound

The following theorem provides an upper bound under the extended link condition (4.9) and an additional moment condition on the bases or, more precisely, on the random vectors $[e(Z)]$ and $[f(W)]$. We begin this section by formalizing this additional condition.

Assumption 4.5 *There exists $\eta \geqslant 1$ such that the joint distribution of (Z, W) satisfies*

(i) $\sup_{j\in\mathbb{N}} \mathbf{E}[e_j^2(Z)|W] \leqslant \eta^2$ *and* $\sup_{l\in\mathbb{N}} \mathbf{E}[f_l^4(W)] \leqslant \eta^4$*;*

(ii) $\sup_{j,l\in\mathbb{N}} \mathbf{V}\mathrm{ar}(e_j(Z)f_l(W)) \leqslant \eta^2$ *and*
$\sup_{j,l\in\mathbb{N}} \mathbf{E}|e_j(Z)f_l(W) - \mathbf{E}[e_j(Z)f_l(W)]|^8 \leqslant 8!\eta^6\, \mathbf{V}\mathrm{ar}(e_j(Z)f_l(W))$.

This assumption restricts the set of possible joint distribution of (Z, W). It is however noticeable that it is satisfied for any joint distribution and for sufficiently large η if the bases $\{e_j\}_{j\geqslant 1}$ and $\{f_l\}_{l\geqslant 1}$ are uniformly bounded. As the bases are not pre-specified, the assumption is not too restrictive. Recall the notation $a_n \lesssim b_n$ and $a_n \sim b_n$ from Chapter 3 (p. 54).

Theorem 4.6 *Suppose that the iid. (Y, Z, W)-sample of size n obeys the model (4.1a–4.1b) and that the joint distribution of (Z, W) fulfils Assumption 4.5 for some $\eta \geqslant 1$. Consider sequences γ, ω and λ satisfying Assumption 4.1 such that the conditional expectation operator T associated to (Z, W) belongs to $\mathcal{T}^{\lambda}_{d,D}$, where $d, D \geqslant 1$. Let k_n^*, R_n^*, and κ be as given in Theorem* (4.2)*. If in addition $\sup_{k\in\mathbb{N}} k^3/\gamma_k =: \zeta < \infty$, then we have for all $n \in \mathbb{N}$ with $(k_n^*)^3 \geqslant 4D\zeta/\kappa$ that*

$$\sup_{P_U\in\mathcal{U}_\sigma}\sup_{\varphi\in\mathcal{F}^{\rho}_{\gamma}} \mathbf{E}\|\widehat{\varphi}_{k_n^*} - \varphi\|_\omega^2 \lesssim D\,\eta^4\,(\sigma^2 + 4\Gamma\, D\, d\,\rho)\, R_n^*$$
$$\cdot\Big\{4D\zeta/\kappa + \max\left(1, \frac{\lambda_{k_n^*}}{\omega_{k_n^*}} \max_{1\leqslant j\leqslant k_n^*} \frac{\omega_j}{\lambda_j}\right) + (k_n^*)^3 \left|P\Big(\|[\widehat{T}]_{\underline{k_n^*}} - [T]_{\underline{k_n^*}}\|^2 > \frac{\lambda_{k_n^*}}{4D}\Big)\right|^{1/4}\Big\}$$
$$+ \rho P\Big(\|[\widehat{T}]_{\underline{k_n^*}} - [T]_{\underline{k_n^*}}\|^2 > \frac{\lambda_{k_n^*}}{4D}\Big).$$

The proof of this theorem is given in a separate paragraph (cf. p. 104).

Remark 4.7 We emphasize that the bound in the last theorem is not asymptotic. Also, note that the term $\max\left(1, \lambda_{k_n^*}/\omega_{k_n^*} \max_{1\leqslant j\leqslant k_n^*} \omega_j/\lambda_j\right)$ is uniformly bounded by a constant since ω/λ is non decreasing. □

A comparison with the lower bound from Theorem 4.2 shows that the last assertion does not establish the minimax-optimality of the estimator. However, the upper bound in Theorem 4.6 can be improved by imposing a moment

condition stronger than Assumption 4.5. To be more precise, consider the centered random variable $e_j(Z)f_l(W) - \mathbf{E}[e_j(Z)f_l(W)]$. Then Assumption 4.5 (ii) imposes that its 8-th moment be uniformly bounded over $j, l \in \mathbb{N}$. In the next Assumption we suppose that these random variables satisfy Cramer's condition uniformly, which is known to be sufficient to obtain an exponential bound for their large deviations (c.f. Bosq, 1998).

Assumption 4.8 *There exists $\eta \geqslant 1$ such that the joint distribution of (Z, W) satisfies Assumption 4.5 (i)-(ii) and in addition*

(iii) $\sup_{j,l\in\mathbb{N}} \mathbf{E}|e_j(Z)f_l(W) - \mathbf{E}[e_j(Z)f_l(W)]|^k \leqslant \eta^{k-2} k! \, \mathbf{V}\mathrm{ar}(e_j(Z)f_l(W))$, *for all $k \geqslant 3$.*

Obviously this assumption is stronger than Assumption 4.5 (ii). But as in case of the latter assumption, whenever the bases $\{e_j\}_{j\geqslant 1}$ and $\{f_l\}_{l\geqslant 1}$ are uniformly bounded, it follows that any joint distribution of (Z, W) satisfies Assumption 4.8 for sufficiently large η. This is a consequence of the well-known fact that Cramer's condition is satisfied in particular if the random variable $e_j(Z)f_l(W) - \mathbf{E}[e_j(Z)f_l(W)]$ is bounded.

On the other hand, we show that under this additional condition the deviation probability tends to zero faster than R_n^*. Hence, the rate R_n^* is optimal and $\widehat{\varphi}_{k_n^*}$ is minimax-optimal, which is summarized in the next assertion.

Theorem 4.9 *Suppose that the assumptions of Theorem 4.6 are satisfied. In addition, assume that the joint distribution of (Z, W) fulfils Assumption 4.8 and that the sequence (ω/λ) is non-decreasing. For all $n \in \mathbb{N}$ with $(\log k_n^*)/k_n^* \leqslant \kappa/(280 D\eta^2\zeta)$ and $(\log R_n^*)/k_n^* \geqslant -\kappa/(40 D\eta^2\zeta)$ we have*

$$\sup_{P_U\in\mathcal{U}_\sigma} \sup_{\varphi\in\mathcal{F}_\gamma^\rho} \mathbf{E}\|\widehat{\varphi}_{k_n^*} - \varphi\|_\omega^2 \lesssim D^2 \, \eta^4 \, \zeta \, \kappa^{-1} (\sigma^2 + \Gamma \, D \, d \, \rho) \; R_n^*.$$

The proof of this theorem is given in a separate paragraph below.

Remark 4.10 From Theorems 4.2 and 4.9 it follows that the estimator $\widehat{\varphi}_{k_n^*}$ attains the optimal rate R_n^* for all sequences γ, ω and λ satisfying the minimal regularity conditions from Assumption 4.1. Let us briefly discuss the role of the sequences γ, ω and λ. Theorem 4.2 and 4.9 show that the faster the sequence λ decreases, the slower the obtainable optimal rate of convergence becomes. On the other hand, a faster increase of γ or decrease of ω leads to a faster optimal rate. In other words, as expected, a structural function satisfying a stronger regularity condition can be estimated faster, and measuring the accuracy with respect to a weaker norm leads to faster rates, too. □

Proof of the upper bounds

We begin by defining and recalling notations to be used in the proofs of this section. Given $k > 0$, denote $\varphi_k := \sum_{j=1}^{k} [\varphi_k]_j e_j$ with $[\varphi_k]_{\underline{k}} = [T]_{\underline{k}}^{-1}[g]_{\underline{k}}$ which is well-defined since $[T]_{\underline{k}}$ is non singular. Then, the identities $[T(\varphi - \varphi_k)]_{\underline{k}} = 0$ and $[\varphi_k - E_k\varphi]_{\underline{k}} = [T]_{\underline{k}}^{-1}[TE_k^{\perp}\varphi]_{\underline{k}}$ hold true. Furthermore, let $[\Xi]_{\underline{k}} := \widehat{[T]}_{\underline{k}} - [T]_{\underline{k}}$ and define vectors $[B]_{\underline{k}}$ and $[S]_{\underline{k}}$ by

$$[B]_j := \frac{1}{n}\sum_{i=1}^{n} U_i f_j(W_i)$$

$$[S]_j := \frac{1}{n}\sum_{i=1}^{n} f_j(W_i)\{\varphi(Z_i) - [\varphi_k]_{\underline{k}}^t [e(Z_i)]_{\underline{k}}\},\ 1 \leqslant j \leqslant k,$$

such that $\widehat{[g]}_{\underline{k}} - \widehat{[T]}_{\underline{k}}[\varphi_k]_{\underline{k}} = [B]_{\underline{k}} + [S]_{\underline{k}}$. Note that $\mathbf{E}[B]_{\underline{k}} = 0$ due to the mean independence, i.e., $\mathbf{E}[U|W] = 0$, and that $\mathbf{E}[S]_{\underline{k}} = [T\varphi]_{\underline{k}} - [T\varphi_k]_{\underline{k}} = 0$. Moreover, let us introduce the events

$$\Omega := \{\|\widehat{[T]}_{\underline{k}}^{-1}\| \leqslant \sqrt{n}\} \quad \text{and} \quad \Omega_{1/2} := \{\|[\Xi]_{\underline{k}}\|\,\|[T]_{\underline{k}}^{-1}\| \leqslant 1/2\}.$$

Observe that $\Omega_{1/2} \subset \Omega$ in case $\sqrt{n} \geqslant 2\|[T]_{\underline{k}}^{-1}\|$. Indeed, if $\|[\Xi]_{\underline{k}}\|\|[T]_{\underline{k}}^{-1}\| \leqslant 1/2$ then the identity $\widehat{[T]}_{\underline{k}} = [T]_{\underline{k}}\{I + [T]_{\underline{k}}^{-1}[\Xi]_{\underline{k}}\}$ implies $\|\widehat{[T]}_{\underline{k}}^{-1}\| \leqslant 2\|[T]_{\underline{k}}^{-1}\|$ by the usual Neumann series argument. Moreover, in case T satisfies the extended link condition (4.9), that is $T \in \mathcal{T}_{d,D}^{\lambda}$, then

$$2\|[T]_{\underline{k}}^{-1}\| \leqslant 2\|[\mathrm{diag}(\lambda)]_{\underline{k}}^{-1/2}\|\|[\mathrm{diag}(\lambda)]_{\underline{k}}^{1/2}[T]_{\underline{k}}^{-1}\| \leqslant 2\sqrt{D/\lambda_k}$$

since λ is non increasing. Finally, given k_n^*, R_n^* and κ defined in Theorem 4.2 we have $\kappa^{-1}\omega_{k_n^*}\gamma_{k_n^*}^{-1} \geqslant R_n^* \geqslant \sum_{j=1}^{k_n^*} \omega_j (n\lambda_j)^{-1}$ by using successively the definition of κ and R_n^*. By combination of the last estimate and the condition $\sup_{k\in\mathbb{N}} k^3\gamma_k^{-1} \leqslant \zeta$ it follows that $(k_n^*)^3(n\lambda_{k_n^*})^{-1} \leqslant \kappa^{-1}(k_n^*)^3\gamma_{k_n^*}^{-1} \leqslant \kappa^{-1}\zeta$. Thus, for all $n \in \mathbb{N}$ with $(k_n^*)^3 \geqslant 4D\zeta\kappa^{-1}$ we have $4\|[T]_{\underline{k_n^*}}^{-1}\|^2 \leqslant 4D\lambda_{k_n^*}^{-1} \leqslant n4D\zeta\kappa^{-1}(k_n^*)^{-3} \leqslant n$, and hence $\Omega_{1/2} \subset \Omega$. These notations and results will be used below without further reference.

We shall prove in the end of this section three technical lemmas (4.21 – 4.23) which are used in the following proofs.

Proof of Theorem 4.6. Define $\widetilde{\varphi}_{k_n^*} := \varphi_{k_n^*}\mathbf{1}_{\Omega}$ and decompose the risk into two terms,

$$\mathbf{E}\|\widehat{\varphi}_{k_n^*} - \varphi\|_{\omega}^2 \leqslant 2\{\mathbf{E}\|\widehat{\varphi}_{k_n^*} - \widetilde{\varphi}_{k_n^*}\|_{\omega}^2 + \mathbf{E}\|\widetilde{\varphi}_{k_n^*} - \varphi\|_{\omega}^2\} =: 2\{A_1 + A_2\}, \quad (4.10)$$

which we bound separately. Consider first A_2. By combination of $\Omega^c \subset \Omega^c_{1/2}$ and the identity $\|\widetilde{\varphi}_{k_n^*} - \varphi\|^2_\omega = \|\varphi_{k_n^*} - \varphi\|^2_\omega \mathbf{1}_\Omega + \|\varphi\|^2_\omega \mathbf{1}_{\Omega^c}$ we deduce

$$\mathbf{E}\|\widetilde{\varphi}_{k_n^*} - \varphi\|^2_\omega \leqslant \|\varphi_{k_n^*} - \varphi\|^2_\omega + \|\varphi\|^2_\omega P(\Omega^c_{1/2}).$$

Since (ω/γ) is monotonically decreasing, the last estimate together with (4.36) in Lemma 4.22 implies for all $\varphi \in \mathcal{F}^\rho_\gamma$

$$\mathbf{E}\|\widetilde{\varphi}_{k_n^*} - \varphi\|^2_\omega \leqslant 4\, D\, d\, \rho\, R_n^* \max\left(1, \frac{\lambda_{k_n^*}}{\omega_{k_n^*}} \max_{1 \leqslant j \leqslant k_n^*} \frac{\omega_j}{\lambda_j}\right) + \rho P(\Omega^c_{1/2}) \tag{4.11}$$

by employing the definition of R_n^*.
Consider A_1. From the identity $\widehat{[g]}_{\underline{k_n^*}} - \widehat{[T]}_{\underline{k_n^*}}[\varphi_m]_{\underline{k_n^*}} = [B]_{\underline{k_n^*}} + [S]_{\underline{k_n^*}}$ follows

$$\begin{aligned}
[\widehat{\varphi}_{k_n^*} - \widetilde{\varphi}_{k_n^*}]_{\underline{k_n^*}} &= \{[T]^{-1}_{\underline{k_n^*}} + [T]^{-1}_{\underline{k_n^*}}([T]_{\underline{k_n^*}} - \widehat{[T]}_{\underline{k_n^*}})\widehat{[T]}^{-1}_{\underline{k_n^*}}\}\{[B]_{\underline{k_n^*}} + [S]_{\underline{k_n^*}}\}\mathbf{1}_\Omega \\
&= [T]^{-1}_{\underline{k_n^*}}\{[B]_{\underline{k_n^*}} + [S]_{\underline{k_n^*}}\}\mathbf{1}_\Omega - [T]^{-1}_{\underline{k_n^*}}[\Xi]_{\underline{k_n^*}}\widehat{[T]}^{-1}_{\underline{k_n^*}}\{[B]_{\underline{k_n^*}} + [S]_{\underline{k_n^*}}\}\mathbf{1}_\Omega.
\end{aligned}$$

By making use of this identity we decompose A_1 further into two terms

$$\begin{aligned}
&\mathbf{E}\|\widehat{\varphi}_{k_n^*} - \widetilde{\varphi}_{k_n^*}\|^2_\omega \leqslant 2\mathbf{E}[\|[\mathrm{diag}(\omega)]^{1/2}_{k_n^*}[T]^{-1}_{k_n^*}\{[B]_{k_n^*} + [S]_{k_n^*}\}\|^2\mathbf{1}_\Omega] \\
&+ 2\mathbf{E}[\|[\mathrm{diag}(\omega)]^{1/2}_{k_n^*}[T]^{-1}_{k_n^*}[\Xi]_{k_n^*}\widehat{[T]}^{-1}_{k_n^*}\{[B]_{k_n^*} + [S]_{k_n^*}\}\|^2\mathbf{1}_\Omega] =: 2\{A_{11} + A_{12}\}
\end{aligned} \tag{4.12}$$

which we bound separately. In case of A_{11} we employ successively (4.35) from Lemma 4.21 with $M := [\mathrm{diag}(\omega)]^{1/2}_{k_n^*}[T]^{-1}_{k_n^*}$, the elementary inequality $\mathrm{tr}(A^tB^tBA) \leqslant \|A\|^2\mathrm{tr}(B^tB)$ valid for all $(k \times k)$ matrices A and B and the extended link condition (4.9), that is, $\|[\mathrm{diag}(\lambda)]^{1/2}_{k_n^*}[T]^{-1}_{k_n^*}\|^2 \leqslant D$. Thereby, we obtain

$$\begin{aligned}
&\mathbf{E}[\|[\mathrm{diag}(\omega)]^{1/2}_{\underline{k_n^*}}[T]^{-1}_{\underline{k_n^*}}\{[B]_{\underline{k_n^*}} + [S]_{\underline{k_n^*}}\}\|^2\mathbf{1}_\Omega] \\
&\leqslant (2/n)\, D\, \mathrm{tr}\left([\mathrm{diag}(\lambda)]^{-1/2}_{\underline{k_n^*}}[\mathrm{diag}(\omega)]_{\underline{k_n^*}}[\mathrm{diag}(\lambda)]^{-1/2}_{\underline{k_n^*}}\right)\{\sigma^2 + \eta^2\,\Gamma\,\|\varphi - \varphi_{k_n^*}\|^2_\gamma\} \\
&= 2D\{\sigma^2 + \eta^2\,\Gamma\,\|\varphi - \varphi_{k_n^*}\|^2_\gamma\}\sum_{j=1}^{k_n^*}\frac{\omega_j}{n\lambda_j}.
\end{aligned} \tag{4.13}$$

Consider now A_{12}. Observe that $\|[\mathrm{diag}(\omega)]^{1/2}_{\underline{k_n^*}}[T]^{-1}_{\underline{k_n^*}}\|^2 \leqslant D \max_{1 \leqslant j \leqslant k_n^*} \omega_j/\lambda_j$ for all $T \in \mathcal{T}^\lambda_{d,D}$. Applying the last inequality together with

$$\|\widehat{[T]}^{-1}_{\underline{k_n^*}}\|^2\mathbf{1}_{\Omega_{1/2}} \leqslant 4D/\lambda_{k_n^*} \quad \text{and} \quad \|\widehat{[T]}^{-1}_{\underline{k_n^*}}\|^2\mathbf{1}_\Omega \leqslant n,$$

we see that there exists a numerical constant $C > 0$ such that

$$\begin{aligned}
\mathbf{E}[\|[\mathrm{diag}(\omega)]_{\underline{k_n^*}}^{1/2}[T]_{\underline{k_n^*}}^{-1}[\Xi]_{\underline{k_n^*}}\widehat{[T]}_{\underline{k_n^*}}^{-1}\{[B]_{\underline{k_n^*}}+[S]_{\underline{k_n^*}}\}\|^2\mathbf{1}_\Omega] \\
&\leqslant D\max_{1\leqslant j\leqslant k_n^*}\frac{\omega_j}{\lambda_j}\Big\{4D\lambda_{k_n^*}^{-1}\mathbf{E}\|[\Xi]_{\underline{k_n^*}}\|^2\|[B]_{\underline{k_n^*}}+[S]_{\underline{k_n^*}}\|^2\mathbf{1}_{\Omega_{1/2}} \\
&\qquad + n\mathbf{E}\|[\Xi]_{\underline{k_n^*}}\|^2\|[B]_{\underline{k_n^*}}+[S]_{\underline{k_n^*}}\|^2\mathbf{1}_{\Omega_{1/2}^c}\Big\} \\
&\leqslant D\max_{1\leqslant j\leqslant k_n^*}\frac{\omega_j}{\lambda_j}\Big\{4D\lambda_{k_n^*}^{-1}\big(\mathbf{E}\|[\Xi]_{\underline{k_n^*}}\|^4\big)^{1/2} \\
&\qquad + n\big(\mathbf{E}\|[\Xi]_{\underline{k_n^*}}\|^8\big)^{1/4}P(\Omega_{1/2}^c)^{1/4}\Big\}\big(\mathbf{E}\|[B]_{\underline{k_n^*}}+[S]_{\underline{k_n^*}}\|^4\big)^{1/2} \\
&\leqslant C\max_{1\leqslant j\leqslant k_n^*}\frac{\omega_j}{n\lambda_j}D\,\eta^4\,(\sigma^2+\Gamma\,\|\varphi-\varphi_{k_n^*}\|_\gamma^2) \\
&\qquad \Big\{4D\frac{(k_n^*)^3}{\lambda_{k_n^*}n}+(k_n^*)^3|P(\Omega_{1/2}^c)|^{1/4}\Big\}
\end{aligned}$$

where the last bound follows from (4.32), (4.33) and (4.34) in Lemma 4.21. By combination of the last bound and (4.13) via the decomposition (4.12) there exists a numerical constant $C > 0$ such that

$$\begin{aligned}
\mathbf{E}\|\widehat{\varphi}_{k_n^*}-\widetilde{\varphi}_{k_n^*}\|_\omega^2 &\leqslant C\,D\,\eta^4\,(\sigma^2+\Gamma\,\|\varphi-\varphi_{k_n^*}\|_\gamma^2) \\
&\qquad \Big\{4D\zeta/\kappa+(k_n^*)^3|P(\Omega_{1/2}^c)|^{1/4}\Big\}\sum_{j=1}^{k_n^*}\frac{\omega_j}{n\lambda_j}.
\end{aligned}$$

Furthermore, taking into account the estimate (4.36) from Lemma 4.22 with $\omega = \gamma$ and the definition of R_n^*, the last inequality implies

$$\mathbf{E}\|\widehat{\varphi}_{k_n^*}-\widetilde{\varphi}_{k_n^*}\|_\omega^2 \leqslant C\,D\,\eta^4\,(\sigma^2+4\Gamma\,Dd\rho)\Big\{4D\zeta/\kappa+(k_n^*)^3|P(\Omega_{1/2}^c)|^{1/4}\Big\}R_n^*.$$

Finally, using the decomposition (4.10), the result of the theorem follows from the last estimate and (4.11), since $\Omega_{1/2}^c \subset \{\|\widehat{[T]}_{\underline{k_n^*}}-[T]_{\underline{k_n^*}}\|^2 > \lambda_{k_n^*}/(4D)\}$. □

Proof of Theorem 4.9. We start our proof with the observation that under Assumption 4.8

$$\begin{aligned}
P\Big(\|\widehat{[T]}_{\underline{k_n^*}}-[T]_{\underline{k_n^*}}\|^2 > \frac{\lambda_{k_n^*}}{4D}\Big) &\leqslant 2\exp\{-\frac{n\lambda_{k_n^*}}{20D\eta^2(k_n^*)^2}+2\log k_n^*\} \\
&\leqslant 2\exp\{-\frac{\kappa}{20D\eta^2\zeta}k_n^*+2\log k_n^*\}
\end{aligned}$$

using first (4.38) and the estimate $(k_n^*)^3(n\lambda_{k_n^*})^{-1} \leqslant \kappa^{-1}(k_n^*)^3\gamma_{k_n^*}^{-1} \leqslant \kappa^{-1}\zeta$. From this estimate we conclude for all $n \in \mathbb{N}$ with

$$(\log k_n^*)/k_n^* \leqslant \kappa/(280D\eta^2\zeta) \quad \text{and} \quad (\log R_n^*)/k_n^* \geqslant -\kappa/(40D\eta^2\zeta)$$

that

$$(k_n^*)^{12} P\Big(\|[\widehat{T}]_{\underline{k_n^*}} - [T]_{\underline{k_n^*}}\|^2 > \frac{\lambda_{k_n^*}}{4D}\Big) \leqslant 2,$$
$$(R_n^*)^{-1} P\Big(\|[\widehat{T}]_{\underline{k_n^*}} - [T]_{\underline{k_n^*}}\|^2 > \frac{\lambda_{k_n^*}}{4D}\Big) \leqslant 2.$$

By employing these estimates the assertion follows now from Theorem 4.6. □

Illustration: estimation of derivatives

To illustrate the previous results, we will describe in this section the prior information about the unknown structural function φ by its degree of smoothness. In order to simplify the presentation, we follow Hall and Horowitz (2005) and suppose that the marginal distribution of the scalar regressor Z and the scalar instrument W are uniformly distributed on the interval $[0, 1]$. It is worth noting that all the results below can be extended to the multivariate case in a straightforward way. In the univariate case, it follows that both Hilbert spaces L_Z^2 and L_W^2 are isomorphic to $L^2[0, 1]$, endowed with the usual norm $\|\cdot\|$ and inner product $\langle\cdot, \cdot\rangle$.

In the last sections, we have seen that the choice of the basis $\{e_j\}_{j\geqslant 1}$ is directly linked to the a priori assumptions we are willing to impose on the structural function. In case of classical smoothness assumptions, it is natural to consider the Sobolev space of periodic functions. Therefore, we introduce the trigonometric basis

$$\psi_1 :\equiv 1,\ \psi_{2j}(s) := \sqrt{2}\cos(2\pi j s),\ \psi_{2j+1}(s) := \sqrt{2}\sin(2\pi j s), s \in [0, 1],\ j \in \mathbb{N}.$$

and choose $\{e_j = \psi_j\}$. It is well-known that for a weight sequence γ with $\gamma_1 = 1$ and $\gamma_j = j^{2p}$ for $j \geqslant 2$, the ellipsoid $\mathcal{F}_\gamma^\rho$ is a subset of the Sobolev space of p-times differentiable periodic functions. In the rest of this section we will suppose that the prior information about the unknown structural function φ is characterized by such a Sobolev ellipsoid, i.e. that φ is $p \geqslant 0$ times differentiable. In this illustration, we consider the estimation of derivatives of the structural function φ. We therefore recall that, up to a constant, for any function $h \in \mathcal{F}_\gamma^\rho$ the weighted norm $\|h\|_\omega$ with $\omega_0 = 1$ and $\omega_j = j^{2s}$, $j \geqslant 2$, equals the L^2-norm of the s-th weak derivative $h^{(s)}$ for each integer $0 \leqslant s \leqslant p$. By virtue of this relation, the results in the previous section imply also a lower as well as an upper bound of the L^2-risk for the estimation of the s-th weak derivative of φ. Finally, we restrict our attention to conditional expectation operator $T \in \mathcal{T}_d^\lambda$ with either

[p-λ] a polynomially decreasing sequence λ, i.e., $\lambda_0 = 1$ and $\lambda_j = j^{-2a}$, $j \geqslant 2$, for some $a > 0$, or

[e-λ] an exponentially decreasing sequence λ, i.e., $\lambda_0 = 1$ and $\lambda_j = \exp(-j^{2a})$, $j \geqslant 2$, for some $a > 0$.

It is easily seen that the minimal regularity conditions given in Assumption 4.1 are satisfied if $p > 1/2$. Roughly speaking, this means that the structural function is at least continuous. The lower bound presented in the next assertion follows now directly from Theorem 4.2. Note that the additional condition, $\sup_{j\geqslant 1} \mathbf{E}[e_j^4(Z)|W] \leqslant \eta$, $\eta \geqslant 8$, is satisfied since the trigonometric basis is bounded uniformly by two.

Proposition 4.11 *Suppose an iid. sample of size n from the model (4.1a–4.1b). If $\gamma_j = j^{2p}$ with $p > 1/2$, then we have for any estimator $\widetilde{\varphi}^{(s)}$ of $\varphi^{(s)}$, $0 \leqslant s < p$,*

[p-λ] $$\sup_{P_U\in\mathcal{U}_\sigma} \sup_{\varphi\in\mathcal{F}_\gamma^p} \left\{\mathbf{E}\|\widetilde{\varphi}^{(s)} - \varphi^{(s)}\|^2\right\} \gtrsim n^{-2(p-s)/(2p+2a+1)},$$

[e-λ] $$\sup_{P_U\in\mathcal{U}_\sigma} \sup_{\varphi\in\mathcal{F}_\gamma^p} \left\{\mathbf{E}\|\widetilde{\varphi}^{(s)} - \varphi^{(s)}\|^2\right\} \gtrsim (\log n)^{-(p-s)/a}.$$

Proof. Since for each $0 \leqslant s \leqslant p$ we have $\mathbf{E}\|\widetilde{f}^{(s)} - f^{(s)}\|^2 \sim \mathbf{E}\|\widetilde{f} - f\|_\omega^2$ we apply the general result given Theorem 4.2. In both cases, the additional conditions formulated in Theorem 4.2 are easily verified. Therefore, it is sufficient to evaluate the lower bound R_n^* given in (4.6). Note that the optimal dimension parameter k_n^* satisfies $R_n^* \sim \omega_{k_n^*}/\gamma_{k_n^*} \sim \sum_{l=1}^{k_n^*} \omega_l/(n\lambda_l)$ since both sequences (γ_j/ω_j) and $(\sum_{0<|l|\leqslant j} \frac{\omega_l}{n\lambda_l})$ are non-increasing.

[p-λ] The well-known approximation $\sum_{j=1}^k j^r \sim k^{r+1}$ for $r > 0$ implies $n \sim (\gamma_{k_n^*}/\omega_{k_n^*}) \sum_{l=1}^{k_n^*} \omega_l/\lambda_l \sim (k_n^*)^{2a+2p+1}$. It follows that $k_n^* \sim n^{1/(2p+2a+1)}$ and the lower bound writes $R_n^* \sim n^{-(2p-2s)/(2p+2a+1)}$.

[e-λ] Applying Laplace's Method (c.f. Chapter 3.7 in Olver (1974)) we have $n \sim (\gamma_{k_n^*}/\omega_{k_n^*}) \sum_{l=1}^{k_n^*} \omega_l/\lambda_l \sim (k_n^*)^{2p} \exp(|k_n^*|^{2a})$ which implies that $k_n^* \sim \{\log(n/(\log n)^{p/a})\}^{1/(2a)} = (\log n)^{1/(2a)}(1 + o(1))$ and that the lower bound can be rewritten as $R_n^* \sim (\log n)^{-(p-s)/a}$. □

In this section, the basis of L^2_W is given by the trigonometric basis $\{f_l = \psi_l\}_{l\geqslant 1}$. The additional moment conditions formalized in Assumption 4.8 are thus automatically fulfilled since the bases $\{e_j\}_{j\geqslant 1}$ and $\{f_l\}_{l\geqslant 1}$ are both uniformly bounded. We suppose that the associated conditional expectation operator T satisfies the extended link condition (4.9), that is, $T \in \mathcal{T}_{d,D}^{\lambda}$. By this means, we restrict the set of possible joint distributions of (Z, W) to those having the trigonometric basis as optimal instruments. As an estimator of $\varphi^{(s)}$, we shall consider the s-th weak derivative of the estimator $\widehat{\varphi}_k$ defined in (4.8). Recall that for each integer $0 \leqslant s \leqslant p$, the s-th weak derivative of the estimator $\widehat{\varphi}_k$ is

$$\widehat{\varphi}_k^{(s)}(t) = \sum_{j\in\mathbb{Z}} (2i\pi j)^s \int_0^1 \widehat{\varphi}_k(u) \exp(-2i\pi ju)du \exp(-2i\pi jt).$$

Applying Theorem 4.6, the rates of the lower bound given in the last assertion are seen to coincide, up to a constant, with an upper bound of the L^2-risk of the estimator $\widehat{\varphi}_k^{(s)}$, which is the statement of the next proposition. This proves that these rates are optimal and the estimator $\widehat{\varphi}_k^{(s)}$ is minimax optimal in both cases.

Proposition 4.12 *Suppose that the iid. (Y, Z, W)-sample of size n obeys the model (4.1a–4.1b). Let $\gamma_j = j^{2p}$ for $p \geqslant 3/2$. For $0 \leqslant s < p$ consider the estimator $\widehat{\varphi}_{k_n^*}$ given in* (4.8).

[**p-λ**] *In the polynomial decreasing case with $k_n^* \sim n^{1/(2p+2a+1)}$,*

$$\sup_{P_U \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{F}_\gamma^\rho} \left\{ \mathbf{E} \|\widehat{\varphi}_{k_n^*}^{(s)} - \varphi^{(s)}\|^2 \right\} \lesssim n^{-2(p-s)/(2p+2a+1)}.$$

[**e-λ**] *In the exponentially decreasing case with $k_n^* \sim (\log n)^{1/(2a)}$,*

$$\sup_{P_U \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{F}_\gamma^\rho} \left\{ \mathbf{E} \|\widehat{\varphi}_{k_n^*}^{(s)} - \varphi^{(s)}\|^2 \right\} \lesssim (\log n)^{-(p-s)/a}.$$

Proof. Since in both cases the dimension parameter is chosen in the optimal way (see the proof of Proposition 4.11), the result follows from Theorem 4.9.□

Remark 4.13 We emphasize the interesting role of the parameters p and a characterizing the regularity conditions imposed on φ and T respectively: As we see from Propositions 4.11 and 4.12, if the value of a increases, the obtainable optimal rate of convergence decreases. Therefore, the parameter a is often called *degree of ill-posedness* (c.f. Natterer, 1984). On the other hand, an increase of the quantity p leads to a faster optimal rate. In other words, as expected, a smoother structural function can be estimated faster. Finally, as opposed to the polynomial case, in the exponential case the smoothing parameter k_n^* does not depend on the value of p. It follows that the proposed estimator is automatically adaptive, i.e. it does not depend on an a-priori knowledge of the degree of smoothness of the structural function φ. However, the choice of the smoothing parameter does depend on the properties of T, more precisely, the value of a. □

4.3 Adaptive estimation under smoothness assumptions

In this section, our objective is to construct a fully adaptive estimator of the structural function φ. Adaptation means that in spite of the conditional expectation operator T being unknown, the estimator should attain the optimal rate of convergence over the ellipsoid $\mathcal{F}_\gamma^\rho$ for a wide range of different weight sequences γ. However, we will suppose that the operator T is diagonal with respect to the trigonometric basis $\{\psi_j\}$. In this situation, for example, an operator with polynomially decreasing λ having a degree of ill-posedness a behaves like a-times integrating, and hence it is also called *finitely smoothing*. On the other hand, when the sequence λ is exponentially decreasing with degree of ill-posedness a, the operator behaves like integrating infinitely many times, and hence it is also called *infinitely smoothing*. Thus, this additional condition imposes in fact a smoothing condition on the unknown conditional expectation operator T. Even though we assume that the operator is smoothing, we do not impose any a-priori knowledge about the specific decay of λ. Our starting point is the estimator given in (4.8), which in this situation takes the form

$$\widehat{\varphi}_k = \sum_{j=1}^{k} \frac{\widehat{[g]}_j}{\widehat{[T]}_{jj}} \mathbf{1}_{[\inf_{1\leqslant j\leqslant k} \widehat{[T]}_{jj}^2 \geqslant 1/n]}\, \psi_j, \tag{4.14}$$

with $\widehat{[g]}_j$ and $\widehat{[T]}_{jj}$ defined in (4.8). In the last section, we have shown that this estimator is minimax-optimal provided the dimension parameter k is chosen in the optimal way. In what follows, the dimension parameter k is chosen using a model selection approach via penalization. This choice will only involve the data and none of the sequences γ and λ describing the underlying smoothness. First, we introduce some sequences which are used below.

Definition 4.14

(i) For all $k \geqslant 1$, define $\Delta_k := \max_{1\leqslant j\leqslant k} \omega_j/\lambda_j$, $\tau_k := \max_{1\leqslant j\leqslant k}(\omega_j)_{\vee 1}/\lambda_j$ with $(q)_{\vee 1} := \max(q,1)$ and

$$\delta_k := k\Delta_k \frac{\log(\tau_k \vee (k+2))}{\log(k+2)}.$$

Let further Σ be a non-decreasing function such that for all $C > 0$

$$\sum_{k\geqslant 1} C\,\tau_k \exp\Big(-\frac{k\log(\tau_k \vee (k+2))}{6C\log(k+2)}\Big) \leqslant \Sigma(C) < \infty \tag{4.15}$$

and $\sup_{n\in\mathbb{N}} \exp\big(-K_2\ C^{-1}\ n^{1/6} + \frac{5}{3}\log n\big) \leqslant \Sigma(C)$ with the constant $K_2 = (\sqrt{2}-1)/(21\sqrt{2})$.

(ii) Define a sequence N as follows,

$$N_n := N_n(\lambda, d) := \max \Big\{ 1 \leqslant N \leqslant n \;\Big|\; n^7 \exp\Big(- \frac{n\,\lambda_N}{288 d}\Big) \leqslant \Big(\frac{2016\,d}{\lambda_1}\Big)^7 \quad \text{and} \quad \delta_N / n \leqslant 1 \Big\}.$$

It is easy to see that there exists always a function Σ satisfying condition (4.15). Consider the estimator $\widehat{\varphi}_{\widetilde{k}}$ defined by choosing the dimension parameter $\widetilde{k}$ such that

$$\widetilde{k} := \operatorname*{argmin}_{1 \leqslant k \leqslant N_n} \Big\{ -\|\widehat{\varphi}_k\|_\omega^2 + c\,\frac{\delta_k}{n} \Big\}$$

for some constant $c > 0$. It is shown in the previous chapter and in Comte and Johannes (2010) that such an estimator can attain minimax-optimal rates in the context of a circular deconvolution problem and a functional linear model, respectively. However, this estimator is only partially adaptive, since the dimension parameter is chosen using a criterion function that involves the sequences N and δ which depend on λ and d. We solve this problem by defining empirical versions of these sequences. The fully adaptive estimator is then defined analogously to the one above, but uses the estimated rather than the original sequences.

Definition 4.15 *Let $\widehat{\delta} := (\widehat{\delta}_k)_{k \geqslant 1}$, $\widehat{N} := (\widehat{N}_n)_{n \geqslant 1}$, be as follows.*

(i) Given $\widehat{\Delta}_k := \max_{1 \leqslant j \leqslant k} \omega_j \widehat{[T]}_{jj}^{-2} \mathbf{1}_{[\inf_{1 \leqslant j \leqslant k} \widehat{[T]}_{jj}^2 \geqslant 1/n]}$ and $\widehat{\tau}_k := \max_{0 \leqslant j \leqslant k} (\omega_j)_{\vee 1} \widehat{[T]}_{jj}^{-2} \mathbf{1}_{[\inf_{1 \leqslant j \leqslant k} \widehat{[T]}_{jj}^2 \geqslant 1/n]}$ let

$$\widehat{\delta}_k := k \widehat{\Delta}_k \, \frac{\log(\widehat{\tau}_k \vee (k+2))}{\log(k+2)}.$$

(ii) Given $N_n^u := \operatorname{argmax}_{1 \leqslant N \leqslant n} \big\{ \max_{1 \leqslant j \leqslant N} \omega_j / n \leqslant 1 \big\}$, let

$$\widehat{N}_n := \quad \operatorname*{argmin}_{1 \leqslant j \leqslant N_n^u} \Big\{ \frac{|\widehat{[T]}_j|^2}{|j| (\omega_j)_{\vee 1}} < \frac{\log n}{n} \Big\}.$$

It worth to stress that all these sequences do not involve any a-priori knowledge about neither the target function φ nor the operator T. Now, we choose the dimension parameter as

$$\widehat{k} := \operatorname*{argmin}_{1 \leqslant k \leqslant \widehat{N}_n} \Big\{ - \|\widehat{f}_k\|_\omega^2 + 540\, \mathbf{E}[Y^2]\, \frac{\widehat{\delta}_k}{n} \Big\}. \tag{4.16}$$

Throughout this chapter we do not address the issue that the value $\mathbf{E}[Y^2]$ is not known in practice. Anyway, it can easily be estimated by its empirical counterpart. Moreover, the constant 540, though suitable for the theory, may probably be chosen much smaller in practice by a simulation study (cf. Comte et al. (2006) in the context of a deconvolution problem).

Our main result below requires the following Assumption.

Assumption 4.16 *The sequence N from Definition 4.14 (ii) satisfies the conditions*

$$\max_{j \geqslant N_n} \frac{\lambda_j}{j(\omega_j)_{\vee 1}} \leqslant \frac{\log n}{4dn} \qquad \textit{and} \qquad d^{-1} \min_{1 \leqslant j \leqslant N_n} \lambda_j \geqslant 2/n.$$

Remark 4.17 Assumption 4.16 satisfied for sufficiently large n by construction. Let us illustrate briefly this assumption in the setting of the examples introduced in Section 4.2.3. Recall the distinction between finitely and infinitely smoothing conditional expectation operators. The sequences from Definition 4.14 take the following forms in the two respective cases.

[**fs**] In the finitely smoothing case, we have

$$\Delta_k = k^{2a+2s}, \quad \delta_k \sim k^{2a+2s+1}, \quad N_n \sim n^{1/(2a+2s+1)}.$$

[**is**] In the infinitely smoothing case, we have

$$\Delta_k = k^{2s} \exp(k^{2a}), \quad \delta_k \sim k^{2a+2s+1} \exp(k^{2a})(\log k)^{-1},$$
$$N_n \sim \left(\log \frac{n \ \log\log n}{(\log n)^{(2a+2s+1)/(2a)}} \right)^{1/(2a)}.$$

The sequence N satisfies Assumption 4.16 in either case. □

We are now able to state the main result of this chapter providing an upper risk bound for the fully adaptive estimator in the case where the eigenfunctions of the operator T^*T are known.

Theorem 4.18 *Assume that we have a sample of size n of (Y, Z, W). Consider sequences ω, γ, and λ satisfying Assumption 4.1 such that the conditional expectation operator T associated to (Z, W) belongs to $T \in \mathcal{T}^{\lambda}_{d,D}$, $d, D \geqslant 1$ and is diagonal with respect to the trigonometric basis $\{\psi_j\}$. Let the sequences δ and N be as in Definition 4.14 and suppose that Assumption 4.16 holds. Define*

further $N_n^l := \operatorname{argmax}_{1\leqslant j\leqslant N_n}\big\{\frac{\lambda_j}{j(\omega_j)_{\vee 1}} \geqslant \frac{4d\log n}{n}\big\}$. *Consider the estimator* $\widehat{\varphi}_{\widehat{k}}$ *defined in* (4.14) *with* $\widehat{k}$ *given by* (4.16). *Then for all* $n \geqslant 1$

$$\sup_{P_U\in\mathcal{U}_\sigma}\sup_{\varphi\in\mathcal{F}_\gamma^\rho}\Big\{\mathbf{E}\|\widehat{\varphi}_{\widehat{k}}-\varphi\|_\omega^2\Big\} \lesssim (2\rho\Gamma+\sigma^2+1)^4 d\,\zeta_d\bigg[\min_{1\leqslant k\leqslant N_n^l}\bigg\{\max\bigg(\frac{\omega_k}{\gamma_k},\frac{\delta_k}{n}\bigg)\bigg\} + \rho\max_{j\geqslant 1}\bigg\{\frac{\omega_j}{\gamma_j}\min\bigg(1,\frac{1}{n\lambda_j}\bigg)\bigg\} + \frac{1}{n}\bigg\{\Sigma\bigg(\frac{(2\rho\Gamma+\sigma^2)\zeta_d+V_{U|Z}}{V_{U|Z}^2}\bigg)+1\bigg\}\bigg],$$

where $V_{U|Z} := \mathbf{E}[\mathbf{V}\mathrm{ar}(U|Z)]$ *and* $\zeta_d := (\log 3d)/\log 3$.

Compare the last assertion with the lower bound given in Theorem 4.2. It is easily seen that if (ω/λ) is non-decreasing, the second term in the upper bound of Theorem 4.18 is always smaller than the first one. Thus, in this situation the fully adaptive estimator attains the lower bound up to a constant as long as $\sup_{k\geqslant 1}\{\delta_k/(\sum_{1\leqslant j\leqslant k}\omega_j/\lambda_j)\} < \infty$ and if the optimal dimension parameter k_n^* given in Theorem 4.2 is smaller than N_n^l. This is summarized in the next assertion.

Corollary 4.19 *Let the assumptions of Theorem 4.18 be satisfied. If in addition* (ω/λ) *is non-decreasing and we have* $\sup_{k\geqslant 1}\{\delta_k/(\sum_{1\leqslant j\leqslant k}\omega_j/\lambda_j)\} < \infty$ *and* $\sup_{n\in\mathbb{N}}(k_n^*/N_n^l) \leqslant 1$, *then*

$$\sup_{P_U\in\mathcal{U}_\sigma}\sup_{\varphi\in\mathcal{F}_\gamma^\rho}\Big\{\mathbf{E}\|\widehat{\varphi}_{\widehat{k}}-\varphi\|_\omega^2\Big\} = O(R_n^*), \qquad \textit{as} \quad n\to\infty,$$

where k_n^* *and* R_n^* *are given in* (4.6).

It is worth to note that the additional assumptions in the last assertion are sufficient, but not necessary, to establish the order optimality of the estimator, as follows from the example **[is]** below.

Before proving Theorem 4.18, we define some notation to be used in the proof. Given $u \in L^2[0,1]$ we denote by $[u]$ the infinite vector of Fourier coefficients $[u]_j := \langle u, \psi_j\rangle$. In particular we use the notations

$$\widehat{\varphi}_k = \sum_{j=1}^k \frac{\widehat{[g]}_j}{\widehat{[T]}_{jj}}\mathbf{1}\{\inf_{1\leqslant j\leqslant k}\widehat{[T]}_{jj}^2 \geqslant 1/n\}\psi_j,\ \widetilde{\varphi}_k := \sum_{j=1}^k \frac{\widehat{[g]}_j}{[T]_{jj}}e_j,\ \varphi_k := \sum_{j=1}^k \frac{[g]_j}{[T]_{jj}}\psi_j,$$
$$\widehat{\Phi}_u := \sum_{j\in\mathbb{N}} \frac{[u]_j}{\widehat{[T]}_{jj}}\mathbf{1}\{\inf_{1\leqslant j\leqslant k}\widehat{[T]}_{jj}^2 \geqslant 1/n\}\psi_j, \quad \widetilde{\Phi}_u := \sum_{j\in\mathbb{N}}\frac{[u]_j}{[T]_{jj}}\psi_j.$$

Furthermore, let $\widehat{g}$ be the function with Fourier coefficients $[\widehat{g}]_j := \widehat{[g]}_j$ and observe that $\mathbf{E}\widehat{g} = g$. Given $1 \leqslant k \leqslant k'$ we have then for all $t \in \mathcal{S}_k := \mathrm{span}\{\psi_1, \ldots, \psi_k\}$

$$\begin{aligned}
\langle t, \widehat{\varphi}_{k'}\rangle_\omega = \langle t, \widehat{\Phi}_{\widehat{g}}\rangle_\omega &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^k Y_i\psi_j(W_i)\frac{\omega_j[t]_j}{\widehat{[T]}_{jj}}\mathbf{1}_{[\inf_{1\leqslant j\leqslant k}\widehat{[T]}^2_{jj}\geqslant 1/n]} = \langle t, \widehat{\varphi}_k\rangle_\omega,\\
\langle t, \widetilde{\varphi}_{k'}\rangle_\omega = \langle t, \widetilde{\Phi}_{\widehat{g}}\rangle_\omega &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^k Y_i\psi_j(W_i)\frac{\omega_j[t]_j}{[T]_{jj}} = \langle t, \widetilde{\varphi}_k\rangle_\omega, \qquad (4.17)\\
\langle t, \varphi_{k'}\rangle_\omega = \langle t, \widetilde{\Phi}_g\rangle_\omega &= \sum_{j=1}^k \frac{\omega_j[t]_j[g]_j}{[T]_{jj}} = \sum_{j=1}^k \omega_j[t]_j[\varphi]_j = \langle t, \varphi\rangle_\omega.
\end{aligned}$$

Consider the contrast $\Upsilon(t) := \|t\|^2_\omega - 2\langle t, \widehat{\Phi}_{\widehat{g}}\rangle_\omega$, for all $t \in L^2[0,1]$. Obviously it follows for all $t \in \mathcal{S}_k$ that $\Upsilon(t) = \|t - \widehat{\varphi}_k\|^2_\omega - \|\widehat{\varphi}_k\|^2_\omega$ and, hence

$$\arg\min_{t\in\mathcal{S}_k}\Upsilon(t) = \widehat{\varphi}_k, \quad \forall\, k \geqslant 1. \qquad (4.18)$$

Then, the adaptive choice $\widehat{k}$ of the dimension parameter can be rewritten as

$$\widehat{k} = \operatorname*{argmin}_{1\leqslant k\leqslant N_n}\{\Upsilon(\widehat{\varphi}_k) + \widehat{\mathrm{pen}}(k)\} \qquad \text{with} \qquad \widehat{\mathrm{pen}}(k) := 540\,\mathbf{E}[Y^2]\frac{\widehat{\delta}_k}{n}. \quad (4.19)$$

Then for all $1 \leqslant k \leqslant N_n$, we have that $\Upsilon(\widehat{\varphi}_{\widehat{k}}) + \widehat{\mathrm{pen}}(\widehat{k}) \leqslant \Upsilon(\widehat{\varphi}_k) + \widehat{\mathrm{pen}}(k) \leqslant \Upsilon(\varphi_k) + \widehat{\mathrm{pen}}(k)$, using first (4.19) and then (4.18). This inequality implies

$$\|\widehat{\varphi}_{\widehat{k}}\|^2_\omega - \|\varphi_k\|^2_\omega \leqslant 2\langle\widehat{\varphi}_{\widehat{k}} - \varphi_k, \widehat{\varphi}_{\widehat{k}}\rangle_\omega + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k}),$$

which together with the identities given in (4.17) for all $1 \leqslant k \leqslant N_n$ implies

$$\begin{aligned}
\|\widehat{\varphi}_{\widehat{k}} - \varphi\|^2_\omega &= \|\varphi - \varphi_k\|^2_\omega + \|\widehat{\varphi}_{\widehat{k}}\|^2_\omega - \|\varphi_k\|^2_\omega - 2\langle\widehat{\varphi}_{\widehat{k}} - \varphi_k, \varphi\rangle_\omega\\
&\leqslant \|\varphi - \varphi_k\|^2_\omega + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k}) + 2\langle\widehat{\varphi}_{\widehat{k}} - \varphi_k, \widehat{\Phi}_{\widehat{g}} - \widetilde{\Phi}_g\rangle_\omega \quad (4.20)
\end{aligned}$$

Consider the unit ball $\mathcal{B}_k := \{f \in \mathcal{S}_k \mid \|f\|_\omega \leqslant 1\}$ and, for arbitrary $\tau > 0$ and $t \in \mathcal{S}_k$, the elementary inequality

$$\begin{aligned}
2|\langle t, h\rangle_\omega| &\leqslant 2\|t\|_\omega \sup_{t\in\mathcal{B}_k}|\langle t, h\rangle_\omega|\\
&\leqslant \tau\|t\|^2_\omega + \frac{1}{\tau}\sup_{t\in\mathcal{B}_k}|\langle t, h\rangle_\omega|^2 = \tau\|t\|^2_\omega + \frac{1}{\tau}\sum_{j=1}^k \omega_j|[h]_j|^2.
\end{aligned}$$

Combining the last estimate with (4.20) and $\widehat{\varphi}_{\widehat{k}} - \varphi_k \in \mathcal{S}_{\widehat{k}\vee k}$ we obtain

$$\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \leqslant \|\varphi - \varphi_k\|_\omega^2 + \tau\,\|\widehat{\varphi}_{\widehat{k}} - \varphi_k\|_\omega^2 + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k}) + \frac{1}{\tau}\sup_{t\in\mathcal{B}_{\widehat{k}\vee k}} |\langle t, \widehat{\Phi}_{\widehat{g}} - \widetilde{\Phi}_g\rangle_\omega|^2.$$

Letting $\tau := 1/3$ it follows from $\|\widehat{\varphi}_{\widehat{k}} - \varphi_k\|_\omega^2 \leqslant 2\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 + 2\|\varphi_k - \varphi\|_\omega^2$ that

$$\frac{1}{3}\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \leqslant \frac{5}{3}\|\varphi - \varphi_k\|_\omega^2 + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k}) + 3\sup_{t\in\mathcal{B}_{\widehat{k}\vee k}} |\langle t, \widehat{\Phi}_{\widehat{g}} - \widetilde{\Phi}_g\rangle_\omega|^2.$$

Consider the functions $\widehat{\nu}$ and $\widehat{\mu}$ with $[\widehat{\nu}]_j = \frac{1}{n}\sum_{i=1}^n Y_i \mathbf{1}_{[|Y_i|\leqslant n^{1/3}]}\psi_j(W_i)$ and $[\widehat{\mu}]_j = \frac{1}{n}\sum_{i=1}^n Y_i \mathbf{1}_{[|Y_i|>n^{1/3}]}\psi_j(W_i)$ respectively, as well as their centered versions $\nu = \widehat{\nu} - \mathbf{E}[\widehat{\nu}]$ and $\mu = \widehat{\mu} - \mathbf{E}[\widehat{\mu}]$, then we have $\widehat{g} - g = \nu + \mu$ and

$$\frac{1}{3}\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \leqslant \frac{5}{3}\|\varphi - \varphi_k\|_\omega^2 + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k}) + 6\sup_{t\in\mathcal{B}_{\widehat{k}\vee k}} |\langle t, \widetilde{\Phi}_\nu\rangle_\omega|^2 + 12\sup_{t\in\mathcal{B}_{\widehat{k}\vee k}} |\langle t, \widehat{\Phi}_\nu - \widetilde{\Phi}_\nu\rangle_\omega|^2 + 12\sup_{t\in\mathcal{B}_{\widehat{k}\vee k}} |\langle t, \widehat{\Phi}_\mu + \widehat{\Phi}_g - \widetilde{\Phi}_g\rangle_\omega|^2$$

Decompose $|\langle t, \widehat{\Phi}_\nu - \widetilde{\Phi}_\nu\rangle_\omega|^2 = |\langle t, \widehat{\Phi}_\nu - \widetilde{\Phi}_\nu\rangle_\omega|^2 \mathbf{1}_{\Omega_q} + |\langle t, \widehat{\Phi}_\nu - \widetilde{\Phi}_\nu\rangle_\omega|^2 \mathbf{1}_{\Omega_q^c}$ further using

$$\Omega_q := \left\{\forall\, 1 \leqslant j \leqslant N_n \;\middle|\; \left|\widehat{[T]}_{jj}^{-1} - [T]_{jj}^{-1}\right| \leqslant \frac{1}{2|[T]_{jj}|} \,\wedge\, \widehat{[T]}_{jj}^{\,2} \geqslant 1/n\right\}. \tag{4.21}$$

Since $\mathbf{1}_{\widehat{[[T]}_{jj}^2 \geqslant 1/n]}\mathbf{1}_{\Omega_q} = \mathbf{1}_{\Omega_q}$, it follows that for all $1 \leqslant j \leqslant N_n$ we have

$$\left(\frac{[T]_{jj}}{\widehat{[T]}_{jj}}\mathbf{1}_{\widehat{[[T]}_{jj}^2 \geqslant 1/n]} - 1\right)^2 \mathbf{1}_{\Omega_q} = |[T]_{jj}|^2\, \mathbf{1}_{\Omega_q}\left|\widehat{[T]}_{jj}^{-1} - [T]_{jj}^{-1}\right|^2 \leqslant \frac{1}{4}.$$

Hence, $\sup_{t\in\mathcal{B}_k} |\langle t, \widehat{\Phi}_\nu - \widetilde{\Phi}_\nu\rangle_\omega|^2\, \mathbf{1}_{\Omega_q} \leqslant \frac{1}{4}\sup_{t\in\mathcal{B}_k} |\langle t, \widetilde{\Phi}_\nu\rangle_\omega|^2$ for all $1 \leqslant k \leqslant N_n$ and

$$\frac{1}{3}\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \leqslant \frac{5}{3}\|\varphi - \varphi_k\|_\omega^2 + 9\sup_{t\in\mathcal{B}_{\widehat{k}\vee k}} |\langle t, \widetilde{\Phi}_\nu\rangle_\omega|^2 + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k}) + 12\sup_{t\in\mathcal{B}_{\widehat{k}\vee k}} |\langle t, \widehat{\Phi}_\nu - \widetilde{\Phi}_\nu\rangle_\omega|^2\, \mathbf{1}_{\Omega_q^c} + 12\sup_{t\in\mathcal{B}_{\widehat{k}\vee k}} |\langle t, \widehat{\Phi}_\mu + \widehat{\Phi}_g - \widetilde{\Phi}_g\rangle_\omega|^2. \tag{4.22}$$

Define $\Delta_k^T := \max_{1\leqslant j\leqslant k} \omega_j/|[T]_{jj}|^2$, $\tau_k^T := \max_{1\leqslant j\leqslant k}(\omega_j)_{\vee 1}/|[T]_{jj}|^2$, and $\delta_k^T := k\Delta_k^T \left\{\log(\tau_k^T \vee (k+2)) \,/\, \log(k+2)\right\}$. Then, it is easily seen that

$$\delta_k^T \leqslant \delta_k\, d\, \frac{\log(3d)}{\log 3} = \delta_k\, d\, \zeta_d \qquad \forall\, k \geqslant 1. \tag{4.23}$$

with $\zeta_d = (\log 3d)/(\log 3)$. Moreover, define the event $\Omega_{qp} := \Omega_q \cap \Omega_p$ where Ω_q is given in (4.20) and

$$\Omega_p := \Big\{ N_n^l \leqslant \widehat{N}_n \leqslant N_n \Big\}.$$

Observe that on Ω_q we have $(1/2)\Delta_k^T \leqslant \widehat{\Delta}_k \leqslant (3/2)\Delta_k^T$ for all $1 \leqslant k \leqslant N_n$ and hence $(1/2)[\Delta_k^T \vee (k+2)] \leqslant [\widehat{\Delta}_k \vee (k+2)] \leqslant (3/2)[\Delta_k^T \vee (k+2)]$, which implies

$$\begin{aligned}(1/2)k\Delta_k^T \Big(\frac{\log[\Delta_k^T \vee (k+2)]}{\log(k+2)}\Big)\Big(1 - \frac{\log 2}{\log(k+2)}\frac{\log(k+2)}{\log(\Delta_k^T \vee [k+2])}\Big)\\ \leqslant \widehat{\delta}_k \leqslant (3/2)k\Delta_k^T \Big(\frac{\log(\Delta_k^T \vee [k+2])}{\log(k+2)}\Big)\Big(1 + \frac{\log 3/2}{\log(k+2)}\frac{\log(k+2)}{\log(\Delta_k^T \vee [k+2])}\Big).\end{aligned}$$

Using $\log(\Delta_k^T \vee (k+2))/\log(k+2) \geqslant 1$, we conclude from the last estimate that

$$\begin{aligned}\delta_k^T/10 \leqslant (\log 3/2)/(2\log 3)\delta_k^T \leqslant (1/2)\delta_k^T[1 - (\log 2)/\log(k+2)] \leqslant \widehat{\delta}_k\\ \leqslant (3/2)\delta_k^T[1 + (\log 3/2)/\log(k+2)] \leqslant 3\delta_k^T.\end{aligned}$$

Recalling that $\widehat{\mathrm{pen}}(k) = 540\,\mathbf{E}[Y^2]\,\widehat{\delta}_k n^{-1}$, we define

$$\mathrm{pen}(k) := 54\,\mathbf{E}[Y^2]\,\delta_k^T n^{-1}, \tag{4.24}$$

then it follows that on Ω_q we have

$$\mathrm{pen}(k) \leqslant \widehat{\mathrm{pen}}(k) \leqslant 30\,\mathrm{pen}(k) \qquad \forall\, 1 \leqslant k \leqslant N_n.$$

On $\Omega_{qp} = \Omega_q \cap \Omega_p$, we have $\widehat{k} \leqslant N_n$. Thus,

$$\begin{aligned}\Big(\mathrm{pen}(k \vee \widehat{k}) + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k})\Big)\mathbf{1}_{\Omega_{qp}}\\ \leqslant \Big(\mathrm{pen}(k) + \mathrm{pen}(\widehat{k}) + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k})\Big)\mathbf{1}_{\Omega_{qp}}\\ \leqslant 31\,\mathrm{pen}(k) \qquad \forall 1 \leqslant k \leqslant N_n.\end{aligned} \tag{4.25}$$

Furthermore, we obviously have $\widehat{\Delta}_k \leqslant n\Delta_k^T$ for every $1 \leqslant k \leqslant N_n$, which implies $\widehat{\delta}_k \leqslant n\,(1 + \log n)\,\delta_k^T$. Consequently, $\widehat{\mathrm{pen}}(k) \leqslant 540\,\mathbf{E}[Y^2]\,n\,(1 + \log n)$, because $\delta_k^T/n \leqslant d\zeta_d\delta_k/n \leqslant d\zeta_d$ for all $1 \leqslant k \leqslant N_n$ by (4.23) and the definition of N_n. On $\Omega_q^c \cap \Omega_p$, we have $\widehat{k} \leqslant N_n$ and hence $\mathrm{pen}(k \vee \widehat{k}) \leqslant \mathrm{pen}(N_n) \leqslant 54\,\mathbf{E}[Y^2]$, which implies

$$(\mathrm{pen}(k \vee \widehat{k}) + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k}))\mathbf{1}_{\Omega_q^c \cap \Omega_p} \leqslant 594\,\mathbf{E}[Y^2]\,n\,(1 + \log n)\mathbf{1}_{\Omega_q^c \cap \Omega_p}. \tag{4.26}$$

We note further that for all $\varphi \in \mathcal{F}_\gamma^\rho$ with $\sum_{j\in\mathbb{N}} \gamma_j^{-1} = \Gamma < \infty$ and for all $z \in [0,1]$ we have $|\varphi(z)|^2 \leqslant \rho \sum_{j\in\mathbb{N}} \gamma_j^{-1}\psi_j^2(z) \leqslant 2\rho\Gamma$ using the Cauchy-Schwarz inequality. Thereby, given $m \geqslant 1$ such that $\mathbf{E}[U^{2m}|W] \leqslant \sigma^{2m}$, it follows that

$$\mathbf{E}[Y^{2m}|W] \leqslant 2^{2m}(2\rho\Gamma + \sigma^2)^m \text{ and, hence } \mathbf{E}[Y^{2m}] \leqslant 2^{2m}(2\rho\Gamma + \sigma^2)^m. \tag{4.27}$$

At the end of this chapter we will prove the technical Lemmata which are used in the following proof.

Proof of Theorem 4.18. The proof is based on the decomposition

$$\mathbf{E}\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 = \mathbf{E}\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \mathbf{1}_{\Omega_{qp}} + \mathbf{E}\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \mathbf{1}_{\Omega_{qp}^c}.$$

In Lemma 4.24 below we show that for all $n \geqslant 1$ and all $1 \leqslant k \leqslant N_n^l$ we have

$$\begin{aligned}\mathbf{E}\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \mathbf{1}_{\Omega_{qp}} \leqslant C\Bigg\{ \|\varphi - \varphi_k\|_\omega^2 + \mathrm{pen}(k) + d\rho \max_{j \geqslant 1} \left[\frac{\omega_j}{\gamma_j} \min\left(1, \frac{1}{n\lambda_j}\right)\right] \\ + \frac{(2\rho\Gamma + \sigma^2)^4}{n} + \frac{(2\rho\Gamma + \sigma^2 + 1) d\, \zeta_d}{n}\ \Sigma\left(\frac{(2\rho\Gamma + \sigma^2)\zeta_d + V_{U|Z}}{V_{U|Z}^2}\right)\Bigg\},\end{aligned} \tag{4.28}$$

$$\mathbf{E}\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \mathbf{1}_{\Omega_{qp}^c} \leqslant \frac{C}{n}\,(2\rho\Gamma + \sigma^2). \tag{4.29}$$

The result follows using (4.27), that is, $\mathrm{pen}(k) \leqslant 54\,(2\rho\Gamma + \sigma^2)\, d\, \zeta_d\, \delta_k n^{-1}$, and by employing the monotonicity of ω/γ, that is $\|\varphi - \varphi_k\|_\omega^2 \leqslant \rho\omega_k/\gamma_k$. □

Illustration: estimation of derivatives (continued)

The following result shows that even without any prior knowledge on the structural function φ and for all smoothing operators T, the fully adaptive penalized estimator automatically attains the optimal rate in the finitely and in the infinitely smoothing case. Recall that the computation of the dimension parameter $\widehat{k}$ given in (4.16) involves the sequence N^u, which in our illustration satisfies $N_n^u \sim n^{1/(2s)}$ since $\omega_j = j^{2s}$, $j \geqslant 1$.

Proposition 4.20 *Suppose that the i.i.d. (Y, Z, W)-sample of size n obeys the model (4.1a–4.1b) and that $P_U \in \mathcal{U}_\sigma$, $\sigma > 0$. Consider the estimator $\widehat{\varphi}_{\widehat{k}}$ given in* (4.8) *with $\widehat{k}$ defined by* (4.16).

[fs] *In the finitely smoothing case, we obtain*

$$\sup_{P_U \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{W}_p^\rho} \left\{\mathbf{E}\|\widehat{\varphi}_{\widehat{k}}^{(s)} - \varphi^{(s)}\|^2\right\} = O(n^{-2(p-s)/(2p+2a+1)}).$$

[is] *In the infinitely smoothing case, we have*

$$\sup_{P_U \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{W}_p^\rho} \left\{\mathbf{E}\|\widehat{\varphi}_{\widehat{k}}^{(s)} - \varphi^{(s)}\|^2\right\} = O((\log n)^{-(p-s)/a}).$$

Proof. In the light of the proof of Proposition 4.11 we apply Theorem 4.18, where in both cases the additional conditions are easily verified (Remark 4.17)

and the result follows by an evaluation of the upper bound. Note further that (ω/λ) is in both cases non decreasing, and hence the second term in the upper bound of Theorem 4.18 is always smaller than the first one.
In case **[fs]** we have $N_n^l \sim (n/(\log n))^{1/(2a+2s+1)}$ and $k_n^* := n^{1/(2a+2p+1)}$. Note that $k_n^* \lesssim N_n^l$. Thus, the upper bound is of order $O((k_n^*)^{-2(p-s)} + n^{-1})$, which equals $O(n^{-2(p-s)/(2a+2p+1)})$.
In case **[is]** we have

$$N_n^l \sim \{\log(n/(\log n)^{(2p+2a+1)/(2a)})\}^{1/(2a)} = (\log n)^{1/(2a)}(1+o(1)) \sim k_n^*.$$

Thereby, the upper bound is of order $O((k_n^*)^{-2(p-s)} + n^{-1})$, which equals $O((\log n)^{-(p-s)/a})$. □

4.4 Auxiliary results

Lemma 4.21 *Suppose that the distribution P_U of U belongs to $\mathcal{U}_\sigma$, $\sigma > 0$ and that the joint distribution of (Z, W) satisfies Assumption 4.5. If in addition $\varphi \in \mathcal{F}_\gamma^r$ with $\Gamma = \sum_{j=1}^\infty \gamma_j^{-1} < \infty$, then there exists a constant $C > 0$ such that for all $k \in \mathbb{N}$ and for all $z \in \mathbb{R}^k$*

$$\mathbf{E}|z^t\,[B]_{\underline{k}}|^2 \leqslant (1/n)\,\|z\|^2\,\sigma^2, \tag{4.30}$$

$$\mathbf{E}|z^t\,[S]_{\underline{k}}|^2 \leqslant (1/n)\,\|z\|^2\,\eta^2\,\Gamma\,\|\varphi-\varphi_k\|_\gamma^2 \tag{4.31}$$

$$\mathbf{E}\|[B]_{\underline{k}}\|^4 \leqslant C\cdot\Big((k/n)\cdot\sigma^2\cdot\eta^2\Big)^2, \tag{4.32}$$

$$\mathbf{E}\|[S]_{\underline{k}}\|^4 \leqslant C\cdot\Big((k/n)\cdot\eta^2\cdot\Gamma\cdot\|\varphi-\varphi_k\|_\gamma^2\Big)^2, \tag{4.33}$$

$$\mathbf{E}\|[\Xi]_{\underline{k}}\|^8 \leqslant C\cdot\Big((k^2/n)\cdot\eta^2\Big)^4. \tag{4.34}$$

Moreover, given a $(k\times k)$ matrix M, we have

$$\mathbf{E}\|M\{[B]_{\underline{k}}+[S]_{\underline{k}}\}\|^2 \leqslant (2/n)\,\mathrm{tr}(M^tM)\{\sigma^2+\eta^2\,\Gamma\,\|\varphi-\varphi_k\|_\gamma^2\}. \tag{4.35}$$

Proof. The proof of (4.30) - (4.34) can be found in Johannes and Breunig (2009) and we omit the details. The estimate (4.35) follows by applying (4.30) and (4.31) to the identity $\|M\{[B]_{\underline{k}}+[S]_{\underline{k}}\}\|^2 = \sum_{j=1}^k \|M_j^t\{[B]_{\underline{k}}+[S]_{\underline{k}}\}\|^2$, where M_j denotes the j-th column of M^t, which completes the proof. □

Lemma 4.22 *Let $g = T\varphi$ and for each $k \in \mathbb{N}$ denote $\varphi_k := [T]_{\underline{k}}^{-1}[g]_{\underline{k}}$. Given sequences λ and γ satisfying Assumption 4.1 let $T \in \mathcal{T}_{d,D}^\lambda$ and $\varphi \in \mathcal{F}_\gamma^r$. For*

each strictly positive sequence $\omega := (\omega_j)_{j\in\mathbb{N}}$ *such that* ω/γ *is non increasing we obtain for all* $k \in \mathbb{N}$

$$\|\varphi - \varphi_k\|_\omega^2 \leqslant 4\, D\, d\, \rho\, \frac{\omega_k}{\gamma_k} \max\left(1, \frac{\lambda_k}{\omega_k} \max_{1\leqslant j\leqslant k} \frac{\omega_j}{\lambda_j}\right) \tag{4.36}$$

Proof. The condition $T \in \mathcal{T}_{d,D}^{\lambda}$, that is, $\sup_{k\in\mathbb{N}} \|[\mathrm{diag}(\lambda)]_{\underline{k}}^{1/2} [T]_{\underline{k}}^{-1}\|^2 \leqslant D$ and $\|Tf\|^2 \leqslant d\|f\|_\lambda^2$ for all $f \in L_Z^2$, together with the identity

$$[E_k\varphi - \varphi_k]_{\underline{k}} = -[T]_{\underline{k}}^{-1} [TE_k^\perp \varphi]_{\underline{k}}$$

imply

$$\|E_k\varphi - \varphi_k\|_\lambda^2 \leqslant D\|TE_k^\perp\varphi\|^2 \leqslant Dd\|E_k^\perp\varphi\|_\lambda^2 \leqslant Dd\gamma_k^{-1}\lambda_k\rho$$

for all $\varphi \in \mathcal{F}_\gamma^\rho$ because (λ/γ) is monotonically non increasing. From this estimate we conclude

$$\begin{aligned}\|E_k\varphi - \varphi_k\|_w^2 &= \|[\mathrm{diag}(w)]_{\underline{k}}^{1/2} [E_k\varphi - \varphi_k]_{\underline{k}}\|^2 \\ &\leqslant \|[\mathrm{diag}(w)]_{\underline{k}}^{1/2} [\mathrm{diag}(\lambda)]_{\underline{k}}^{-1/2}\|^2 \|E_k\varphi - \varphi_k\|_\lambda^2 \leqslant Dd\rho \frac{\lambda_k}{\gamma_k} \max_{1\leqslant j\leqslant k} \frac{\omega_j}{\lambda_j}. \end{aligned} \tag{4.37}$$

Furthermore, since (ω/γ) is non increasing, we have $\|E_k\varphi - \varphi\|_w^2 \leqslant \rho\omega_k/\gamma_k$ for all $f \in \mathcal{F}_\gamma^\rho$. The assertion follows now by combination of the last estimate and (4.37) via a decomposition based on an elementary triangular inequality.□

Lemma 4.23 *Suppose that the joint distribution of* (Z, W) *satisfies Assumption 4.8. If in addition the sequence* λ *fulfills Assumption 4.1, then for all* $k \in \mathbb{N}$ *we have*

$$P(\|[\Xi]_{\underline{k}}\|^2 > \frac{\lambda_k}{4D}) \leqslant 2\exp\{-\frac{n\lambda_k}{k^2(20D\eta^2)} + 2\log k\}. \tag{4.38}$$

Proof. The proof of the assertion can be found in Johannes and Breunig (2009) and we omit the details. □

Lemma 4.24 *The inequalities* (4.28) *and* (4.29) *hold.*

Proof. Consider first (4.28). Defining $\mathrm{pen}(k) := 54\,\mathbf{E}[Y^2]\,\delta_k^T n^{-1}$ and using the estimate (4.22), we have

$$\begin{aligned}\frac{1}{3}\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \leqslant \frac{5}{3}\|\varphi - \varphi_k\|_\omega^2 + 9\Bigg(\sup_{t\in\mathcal{B}_{k\vee\widehat{k}}} |\langle t, \Phi_\nu\rangle_\omega|^2 - 6\,\frac{\mathbf{E}[Y^2]\,\delta_{k\vee\widehat{k}}^T}{n}\Bigg)_+ \\ + \mathrm{pen}(k\vee\widehat{k}) + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k}) \\ + 12\sup_{t\in\mathcal{B}_{k\vee\widehat{k}}} |\langle t, \widehat{\Phi}_\nu - \Phi_\nu\rangle_\omega|^2\, \mathbf{1}_{\Omega_q^c} + 12\sup_{t\in\mathcal{B}_{k\vee\widehat{k}}} |\langle t, \widehat{\Phi}_\mu + \widehat{\Phi}_g - \Phi_g\rangle_\omega|^2\end{aligned}$$

and, hence using that $\widehat{k} \leqslant N_n$ on Ω_p we obtain for all $1 \leqslant k \leqslant N_n^l$

$$\begin{aligned}\frac{1}{3}\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \mathbf{1}_{\Omega_{qp}} &\leqslant \frac{5}{3}\|\varphi - \varphi_k\|_\omega^2 + 9\sum_{k=1}^{N_n}\Big(\sup_{t\in\mathcal{B}_k}|\langle t, \Phi_\nu\rangle_\omega|^2 - 6\frac{\mathbf{E}[Y^2]\delta_k^T}{n}\Big)_+ \\ &\quad + 12\sup_{t\in\mathcal{B}_{N_n}}|\langle t, \widehat{\Phi}_\mu + \widehat{\Phi}_g - \Phi_g\rangle_\omega|^2 + \Big(\mathrm{pen}(k\vee\widehat{k}) + \widehat{\mathrm{pen}}(k) - \widehat{\mathrm{pen}}(\widehat{k})\Big)\mathbf{1}_{\Omega_{qp}} \\ &\leqslant \frac{5}{3}\|\varphi - \varphi_k\|_\omega^2 + 9\sum_{k=1}^{N_n}\Big(\sup_{t\in\mathcal{B}_k}|\langle t, \Phi_\nu\rangle_\omega|^2 - 6\frac{\mathbf{E}[Y^2]\delta_k^T}{n}\Big)_+ \\ &\qquad + 12\sup_{t\in\mathcal{B}_{N_n}}|\langle t, \widehat{\Phi}_\mu + \widehat{\Phi}_g - \Phi_g\rangle_\omega|^2 + 31\,\mathrm{pen}(k),\end{aligned}$$

where the last inequality follows from (4.25). The second term is bounded by employing Lemma 4.25. In order to control the third term, apply Lemmata 4.26 and 4.27. Consequently, combining these estimates proves inequality (4.28).

Consider now (4.29). Let $\breve{\varphi}_k := \sum_{j=1}^k [\varphi]_j \mathbf{1}\{\widehat{[T]}_{jj}^2 \geqslant 1/n\}\psi_j$. It is easy to see that $\|\widehat{\varphi}_k - \breve{\varphi}_k\|^2 \leqslant \|\widehat{\varphi}_{k'} - \breve{\varphi}_{k'}\|^2$ for all $k' \leqslant k$ and $\|\breve{\varphi}_k - \varphi\|^2 \leqslant \|\varphi\|^2$ for all $k \geqslant 1$. Thus, using that $1 \leqslant \widehat{k} \leqslant N_n^u$, we can write

$$\begin{aligned}\mathbf{E}\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \mathbf{1}_{\Omega_{qp}^c} &\leqslant 2\{\mathbf{E}\|\widehat{\varphi}_{\widehat{k}} - \breve{\varphi}_{\widehat{k}}\|_\omega^2 \mathbf{1}_{\Omega_{qp}^c} + \mathbf{E}\|\breve{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \mathbf{1}_{\Omega_{qp}^c}\} \\ &\leqslant 2\Big\{\mathbf{E}\|\widehat{\varphi}_{N_n^u} - \breve{\varphi}_{N_n^u}\|_\omega^2 \mathbf{1}_{\Omega_{qp}^c} + \|\varphi\|_\omega^2\,\mathbf{P}[\Omega_{qp}^c]\Big\}.\end{aligned}$$

Moreover, since $\sup_{j\geqslant 1}\mathbf{E}[Y^4\psi_j^4(W)] \leqslant 64(2\rho\Gamma + \sigma^2)^2$ and $\mathbf{E}\psi_j^4(W)\psi_j^4(Z) \leqslant 16$ due to (4.27), it follows from Theorem A.3 in the appendix that

$$\begin{aligned}&\mathbf{E}\|\widehat{\varphi}_{N_n^u} - \breve{\varphi}_{N_n^u}\|_\omega^2 \mathbf{1}_{\Omega_{qp}^c} \\ &\quad\leqslant 2n\sum_{j=1}^{N_n^u}\omega_j\Big\{\mathbf{E}(\widehat{[g]}_j - [T]_{jj}[\varphi]_j)^2\mathbf{1}_{\Omega_{qp}^c} + \mathbf{E}([T]_{jj}[\varphi]_j - \widehat{[T]}_{jj}[\varphi]_j)^2\mathbf{1}_{\Omega_{qp}^c}\Big\} \\ &\quad\leqslant 2n\Big\{\sum_{j=1}^{N_n^u}\omega_j\Big[\mathbf{E}\left(\widehat{[g]}_j - [g]_j\right)^4\Big]^{1/2}\mathbf{P}[\Omega_{qp}^c]^{1/2} \\ &\qquad\qquad + \sum_{j=1}^{N_n^u}\omega_j|[\varphi]_j|^2[\mathbf{E}(\widehat{[T]}_{jj} - [T]_{jj})^4]^{1/2}\mathbf{P}[\Omega_{qp}^c]^{1/2}\Big\} \\ &\quad\leqslant Cn\Big\{n\,(2\rho\Gamma + \sigma^2) + (n^{-1}\|\varphi\|_\omega^2)\Big\}\,\mathbf{P}[\Omega_{qp}^c]^{1/2},\end{aligned}$$

where we have used that $\sum_{j=1}^{N_n^u}\omega_j \leqslant n(\max_{1\leqslant j\leqslant N_n^u}\omega_j) \leqslant n^2$ due to Definition 4.15 (ii). Since (ω/γ) is non-increasing, (4.29) follows from Lemmas 4.29 and 4.30, which completes the proof. □

Lemma 4.25 *There exists a numerical constant $C > 0$ such that*

$$\sum_{k=1}^{N_n} \mathbf{E}\Big[\Big(\sup_{t\in\mathcal{B}_k} |\langle t, \Phi_\nu\rangle_\omega|^2 - \frac{6\,\mathbf{E}[Y^2]\,\delta_k^T}{n}\Big)_+\Big] \leqslant \frac{C}{n}\left\{(2\rho\Gamma + \sigma^2 + 1)d\,\zeta_d\;\Sigma\left(\frac{(2\rho\Gamma+\sigma^2)\zeta_d + V_{U|Z}}{V_{U|Z}^2}\right)\right\}.$$

where $\Sigma(\cdot)$ is the function from Definition 4.14

Proof. For $t \in \mathcal{S}_k$, define $r_t(y,w) := \sum_{j=1}^k \omega_j y \mathbf{1}_{[|y|\leqslant n^{1/3}]}\psi_j(w)[t]_j[T]_{jj}^{-1}$. Then it is readily seen that $\langle t, \Phi_\nu\rangle_\omega = \frac{1}{n}\sum_{k=1}^n r_t(Y_k, W_k) - \mathbf{E}[r_t(Y_k, W_k)]$.

Next, we compute constants H_1, H_2, and v verifying the three inequalities required in Talagrand's inequality (Theorem A.5). Consider H_1 first:

$$\sup_{t\in\mathcal{B}_k}\|r_t\|_\infty^2 = \sup_{y,w}\sum_{j=1}^k \omega_j\left(y\mathbf{1}_{[|y|\leqslant n^{1/3}]}[T]_{jj}^{-1}\psi_j(w)\right)^2 \leqslant 2n^{2/3}\delta_k^T =: H_1^2$$

Next, find H_2. Notice that

$$\begin{aligned}\mathbf{E}[\sup_{t\in\mathcal{B}_k}|\langle t,\Phi_\nu\rangle_\omega|^2] &= \frac{1}{n}\sum_{j=1}^k \omega_j |[T]_{jj}|^{-2}\;\mathbf{V}\mathrm{ar}(Y\mathbf{1}_{[|Y|\leqslant n^{1/3}]}\psi_j(W)) \\ &\leqslant \frac{1}{n}\sum_{j=1}^k \omega_j|[T]_{jj}|^{-2}\;\mathbf{E}[\mathbf{E}[Y^2|W]\,\psi_j(W)^2] \leqslant 2\mathbf{E}[Y^2]\frac{\delta_k^T}{n} =: H_2^2\end{aligned}$$

As for v, we note that due to (4.27) for all $\varphi \in \mathcal{F}_\gamma^\rho$ the condition $P_U \in \mathcal{U}_\sigma$, i.e., $\mathbf{E}[U^2|W] \leqslant \sigma^2$, implies $\mathbf{E}[Y^2|W] \leqslant 2(2\rho\Gamma + \sigma^2)$, and hence

$$\begin{aligned}\sup_{t\in\mathcal{B}_k}\mathbf{V}\mathrm{ar}(r_t(Y,W)) &\leqslant \sup_{t\in\mathcal{B}_k}\mathbf{E}\left[\left(Y\sum_{j=1}^k \frac{\omega_j[t]_j}{[T]_{jj}}\psi_j(W)\right)^2\right] \\ &= \sup_{t\in\mathcal{B}_k}\mathbf{E}\left[\mathbf{E}[Y^2|W]\,\left(\sum_{j=1}^k \frac{\omega_j[t]_j}{[T]_{jj}}\psi_j(W)\right)^2\right] \\ &\leqslant 2(2\rho\Gamma+\sigma^2)\sup_{t\in\mathcal{B}_k}\sum_{j,j'=1}^k \frac{\omega_j\omega_{j'}[t]_j[t]_{j'}}{[T]_{jj}[T]_{j'j'}}\mathbf{E}[\psi_j(W)\psi_{j'}(W)] \\ &\leqslant 2(2\rho\Gamma+\sigma^2)\max_{1\leqslant j\leqslant k}\frac{\omega_j}{[T]_{jj}^2}\sup_{t\in\mathcal{B}_k}\sum_{j=1}^k \omega_j[t]_j^2 \leqslant 2(2\rho\Gamma+\sigma^2)\Delta_k^T =: v,\end{aligned}$$

Employing Theorem A.5 we conclude

$$\sum_{k=1}^{N_n} \mathbf{E}\Big[\Big(\sup_{t\in\mathcal{B}_k} |\langle t, \Phi_\nu\rangle_\omega|^2 - \frac{6\,\mathbf{E}[Y^2]\,\delta_k^T}{n}\Big)_+\Big]$$
$$\leqslant C\left\{\frac{\mathbf{E}[Y^2]}{n}\sum_{k=1}^{N_n}\frac{(2\rho\Gamma+\sigma^2)}{\mathbf{E}[Y^2]}\,\Delta_k^T \exp\Big(-\frac{\mathbf{E}[Y^2]}{6(2\rho\Gamma+\sigma^2)}(\delta_k^T/\Delta_k^T)\Big)\right.$$
$$\left. + n^{2/3}\exp\Big(-K_2\sqrt{\mathbf{E}[Y^2]}n^{1/6}\Big)\sum_{k=1}^{N_n}\frac{\delta_k^T}{n^2}\right\}.$$

The definition of N_n together with (4.23) implies $\sum_{k=1}^{N_n}\delta_k^T/n^2 \leqslant \zeta_d$. Thereby, using (4.23), $\Delta_k^T \leqslant d\tau_k$ and the function Σ given in Definition 4.14, there exists a numerical constant $C>0$ such that

$$\sum_{k=1}^{N_n} \mathbf{E}\Big[\Big(\sup_{t\in\mathcal{B}_k} |\langle t, \Phi_\nu\rangle_\omega|^2 - \frac{6\,\mathbf{E}[Y^2]\,\delta_k^T}{n}\Big)_+\Big]$$
$$\leqslant \frac{C}{n}\left\{\mathbf{E}[Y^2]d\,\Sigma\Big(\frac{(2\rho\Gamma+\sigma^2)\zeta_d}{\mathbf{E}[Y^2]}\Big) + \zeta_d\Sigma\Big(\frac{1}{\sqrt{\mathbf{E}[Y^2]}}\Big)\right\}.$$

Moreover, we have $\mathbf{E}[Y^2] \leqslant 2(2\rho\Gamma+\sigma^2)$ and

$$\inf_{\varphi\in\mathcal{F}_\gamma^\rho}\mathbf{E}[Y^2] \geqslant \inf_{\varphi\in L_Z^2}\mathbf{E}[(\varphi(Z)+U)^2] \geqslant \mathbf{E}[(U-\mathbf{E}[U|Z])^2] = \mathbf{E}[\mathbf{V}\mathrm{ar}(U|Z)] = V_{U|Z}^2,$$

which implies the result. □

Lemma 4.26 *For every $n\in\mathbb{N}$ we have*

$$\mathbf{E}\Big[\sup_{t\in\mathcal{B}_{N_n}}|\langle t, \widehat{\Phi}_\mu\rangle_\omega|^2\Big] \leqslant 2^9(2\rho\Gamma+\sigma^2)^4 n^{-1}.$$

Proof. Since $[\mu]_j = [\widehat{\mu}]_j - \mathbf{E}[\widehat{\mu}]_j$ and $\mathbf{V}\mathrm{ar}[\widehat{\mu}]_j \leqslant n^{-1}\mathbf{E}[Y^2\mathbf{1}_{[|Y|>n^{1/3}]}\psi_j^2(W)]$, it is easily seen that

$$\mathbf{E}\Big[\sup_{t\in\mathcal{B}_{N_n}}|\langle t, \widehat{\Phi}_\mu\rangle_\omega|^2\Big] \leqslant n\sum_{j=1}^{N_n}\omega_j\,\mathbf{V}\mathrm{ar}[\widehat{\mu}]_j$$
$$\leqslant \sum_{j=1}^{N_n}\mathbf{E}\left[\Big(\mathbf{E}[Y^4|W]\mathbf{E}[\mathbf{1}_{[|Y|>n^{1/3}]}|W]\Big)^{1/2}\psi_j^2(W)\right].$$

Moreover, we have $\mathbf{E}[Y^{12}|W] \leqslant 2^{12}(2\rho\Gamma+\sigma^2)^6$for all $\varphi \in \mathcal{F}_\gamma^\rho$ and $U \in \mathcal{U}_\sigma$ due to (4.27) with $m=6$, and hence by Markov's inequality

$$\mathbf{E}[\mathbf{1}_{[|Y|>n^{1/3}]}|W] \leqslant 2^{12}(2\rho\Gamma+\sigma^2)^6 n^{-4}.$$

Combining these estimates, we obtain

$$\mathbf{E}\Big[\sup_{t\in\mathcal{B}_{N_n}}|\langle t,\widehat{\Phi}_\mu\rangle_\omega|^2\Big] \leqslant \sum_{j=1}^{N_n}\mathbf{E}\Big[2^8(2\rho\Gamma+\sigma^2)^4 n^{-2}\psi_j^2(W)\Big] \leqslant 2^9 N_n(2\rho\Gamma+\sigma^2)^4 n^{-2}.$$

The result follows now from $N_n \leqslant n$. □

Lemma 4.27 *There is a numerical constant $C>0$ such that for all $\varphi \in \mathcal{F}_\gamma^\rho$ and every $k,n\in\mathbb{N}$*

$$\mathbf{E}\Big[\sup_{t\in\mathcal{B}_k}|\langle t,\widehat{\Phi}_g-\Phi_g\rangle_\omega|^2\Big] \leqslant Cd\rho\max_{j\geqslant 1}\Big\{\frac{\omega_j}{\gamma_j}\min\big(1,\frac{1}{n\lambda_j}\big)\Big\}.$$

Proof. Firstly, as $\varphi\in\mathcal{F}_\gamma^\rho$, it is easily seen that

$$\mathbf{E}\Big[\sup_{t\in\mathcal{B}_k}|\langle t,\widehat{\Phi}_g-\Phi_g\rangle_\omega|^2\Big] \leqslant \sum_{j=1}^{k}[\varphi]_j^2\omega_j\mathbf{E}[R_j^2] \leqslant \rho\max_{j\geqslant 1}\Big\{\frac{\omega_j}{\gamma_j}\mathbf{E}[R_j^2]\Big\}$$

where R_j is defined by

$$R_j := \left(\frac{[T]_{jj}}{\widehat{[T]}_{jj}}\mathbf{1}_{\widehat{[T]}_{jj}^2\geqslant 1/n]}-1\right). \tag{4.39}$$

The result follows from $\mathbf{E}R_j^2 \leqslant Cd\min\big(1,\frac{1}{n\lambda_j}\big)$, which can be shown as follows. Consider the identity

$$\mathbf{E}|R_j|^2 = \mathbf{E}\Big[\Big|\frac{[T]_{jj}}{\widehat{[T]}_{jj}}-1\Big|^2\mathbf{1}_{\widehat{[T]}_{jj}^2\geqslant 1/n]}\Big]+\mathbf{P}[\widehat{[T]}_{jj}^2<1/n] =: R_j^I+R_j^{II}. \tag{4.40}$$

Trivially, $R_j^{II}\leqslant 1$. If $1\leqslant 4/(n\,[T]_{jj}^2)$, then obviously $R_j^{II}\leqslant 4/(n[T]_{jj}^2)\leqslant 4d/(n\lambda_j)$. Otherwise, we have $1/n<[T]_{jj}^2/4$ and hence, using Chebychev's inequality,

$$R_j^{II}\leqslant \mathbf{P}[|\widehat{[T]}_{jj}-[T]_{jj}|>|[T]_{jj}|/2] \leqslant \frac{4\,\mathbf{V}\mathrm{ar}(\widehat{[T]}_{jj})}{[T]_{jj}^2}\leqslant\frac{16}{n[T]_{jj}^2}\leqslant\frac{16d}{n\lambda_j},$$

where we have used that $\mathbf{V}\mathrm{ar}(\widehat{[T]}_{jj}) \leqslant 4/n$ for all j. Combining both estimates we have $R_j^I \leqslant 16d \min\left(1, \frac{1}{n\lambda_j}\right)$. Now consider R_j^I. We find that

$$R_j^I = \mathbf{E}\left[\frac{|\widehat{[T]}_{jj} - [T]_{jj}|^2}{\widehat{[T]}_{jj}^2} \mathbf{1}_{\widehat{[[T]}_{jj}^2 \geqslant 1/n]}\right] \leqslant n\,\mathbf{V}\mathrm{ar}(\widehat{[T]}_{jj}) \leqslant 4. \tag{4.41}$$

Using that $\mathbf{E}[|\widehat{[T]}_{jj} - [T]_{jj}|^4] \leqslant c/n^2$ for some numerical constant $c > 0$ (cf. Theorem A.3 in the appendix), there exists a numerical constant $c > 0$ such that

$$\begin{aligned} R_j^I &\leqslant \mathbf{E}\left[\frac{|\widehat{[T]}_{jj} - [T]_{jj}|^2}{\widehat{[T]}_{jj}^2} \mathbf{1}_{\widehat{[[T]}_{jj}^2 \geqslant 1/n]} 2\left\{\frac{|\widehat{[T]}_{jj} - [T]_{jj}|^2}{[T]_{jj}^2} + \frac{\widehat{[T]}_{jj}^2}{[T]_{jj}^2}\right\}\right] \\ &\leqslant \frac{2\,n\,\mathbf{E}[|\widehat{[T]}_{jj} - [T]_{jj}|^4]}{[T]_{jj}^2} + \frac{2\ \mathbf{V}\mathrm{ar}(\widehat{[T]}_{jj})}{[T]_{jj}^2} \leqslant \frac{c}{n\,[T]_{jj}^2} \leqslant \frac{cd}{n\lambda_j}. \end{aligned}$$

Combining with (4.41) gives $R_j^I \leqslant Cd \min\left\{1, \frac{1}{n\lambda_j}\right\}$ for some numerical constant $C > 0$, which completes the proof. □

Lemma 4.28 *There is a numerical constant $C > 0$ such that*

$$\mathbf{E}\left[\sup_{t \in \mathcal{B}_{N_n}} |\langle t, \widehat{\Phi}_\nu - \Phi_\nu\rangle_\omega \mathbf{1}_{\Omega_q^c}|^2\right] \leqslant Cd(\mathbf{P}[\Omega_q^c])^{(1/2)}.$$

Proof. Given R_j from (4.39) we begin our proof observing that

$$\begin{aligned} \mathbf{E}\left[\sup_{t \in \mathcal{B}_{M_m}} |\langle t, \widehat{\Phi}_\nu - \Phi_\nu\rangle_\omega \mathbf{1}_{\Omega_q^c}|^2\right] &\leqslant \sum_{j=1}^{N_n} \frac{\omega_j}{[T]_{jj}^2}\, \mathbf{E}[[\nu]_j^2\, R_j^2\, \mathbf{1}_{\Omega_q^c}] \\ &\leqslant \sum_{j=1}^{N_n} \frac{\omega_j}{[T]_{jj}^2}\, \left(\mathbf{E}[[\nu]_j^8]\mathbf{E}[R_j^8]\right)^{1/4} \mathbf{P}[\Omega_q^c]^{1/2}, \end{aligned}$$

where we have applied Cauchy-Schwarz twice. By Petrov's inequality, there exists a numerical constant $c > 0$ such that $E[[\nu]_j^8] \leqslant cn^{-4/3}$ and hence, because $d\delta_k \geqslant \sum_{j=1}^k \frac{\omega_j}{[T]_{jj}^2}$,

$$\mathbf{E}\left[\sup_{t \in \mathcal{B}_{M_m}} |\langle t, \widehat{\Phi}_\nu - \Phi_\nu\rangle_\omega \mathbf{1}_{\Omega_q^c}|^2\right] \leqslant \mathbf{P}[\Omega_q^c]^{1/2} d\delta_k \max_{1 \leqslant j \leqslant N_n} (\mathbf{E}[R_j^8])^{1/4}$$

In analogy to (4.40), we decompose the moment of R_j into two terms

$$\mathbf{E}[R_j^8] = \mathbf{E}\left[\left|\frac{[T]_{jj} - \widehat{[T]}_{jj}}{\widehat{[T]}_{jj}}\right|^8 \mathbf{1}_{\widehat{[[T]}_{jj}^2 \geqslant 1/n]}\right] + \mathbf{P}[\widehat{[T]}_{jj}^2 < 1/n],$$

which we bound by a constant using Petrov's inequality. □

Lemma 4.29 *We have* $\mathbf{P}[\Omega_q^c] \leqslant 2(2016d/\lambda_1)^7\, n^{-6}$, *where* Ω_q *is the event defined in* (4.21).

Proof. Consider the complement of Ω_q given by

$$\Omega_q^c = \Big\{ \exists\, 1 \leqslant j \leqslant N_n \;\Big|\; \Big| \frac{[T]_{jj}}{\widehat{[T]}_{jj}} - 1 \Big| > \frac{1}{2} \;\vee\; \widehat{[T]}_{jj}^{\,2} < 1/n \Big\}.$$

It follows from Assumption 4.16 (i) that $[T]_{jj}^2 \geqslant 2/n$ for all $1 \leqslant j \leqslant N_n$. This yields

$$\mathbf{P}(\Omega_q^c) \leqslant \sum_{j=1}^{N_n} \mathbf{P}\Big[\Big| \frac{\widehat{[T]}_{jj}}{[T]_{jj}} - 1 \Big| > \frac{1}{3} \Big].$$

From Hoeffding's inequality follows

$$\mathbf{P}[|\widehat{[T]}_{jj}/[T]_{jj} - 1| > 1/3] \leqslant 2 \exp\Big(- \frac{n[T]_{jj}^2}{288} \Big),$$

which implies the result by definition of N_n. □

Lemma 4.30 *Consider the event* Ω_p *defined in* (4.21). *Then we have*

$$\mathbf{P}(\Omega_p^c) \leqslant 4 \left(\frac{2016\, d}{\lambda_1} \right)^7 n^{-6}, \qquad \forall\, n \geqslant 1.$$

Proof. Let $\Omega_I := \{N_n^l > \widehat{N}_n\}$ and $\Omega_{II} := \{\widehat{N}_n > N_n\}$. Then we have $\Omega_p^c = \Omega_I \cup \Omega_{II}$. Consider Ω_I first. By definition of N_n^l, we have that $\min_{1 \leqslant j \leqslant N_n^l} \frac{|[T]_j|^2}{|j|(\omega_j)_{\vee 1}} \geqslant \frac{4(\log n)}{n}$, which implies

$$\begin{aligned} \{\widehat{N}_n < N_n^l\} &\subset \Big\{ \exists 1 \leqslant j \leqslant N_n^l \;\Big|\; \frac{\widehat{[T]}_{jj}^{\,2}}{|j|(\omega_j)_{\vee 1}} < \frac{\log n}{n} \Big\} \\ &\subset \bigcup_{1 \leqslant j \leqslant N_n^l} \Big\{ \frac{|\widehat{[T]}_{jj}|}{|[T]_{jj}|} \leqslant 1/2 \Big\} \subset \bigcup_{1 \leqslant j \leqslant N_n^l} \Big\{ |\widehat{[T]}_{jj}/[T]_{jj} - 1| \geqslant 1/2 \Big\}. \end{aligned}$$

Therefore, $\Omega_I \subset \bigcup_{1 \leqslant |j| \leqslant N_n} \Big\{ |\widehat{[\varphi]}_j/[\varphi]_j - 1| \geqslant 1/2 \Big\}$, since $N_n^l \leqslant N_n$. Hence, as in (4.23) applying Hoeffding's inequality together with the definition of N_n gives

$$\mathbf{P}[\Omega_I] \leqslant \sum_{j=1}^{N_n} 2 \exp\Big(- \frac{n\,[T]_{jj}^2}{288} \Big) \leqslant 2 \left(\frac{2016\, d}{\lambda_1} \right)^7 n^{-6}. \tag{4.42}$$

Consider Ω_{II}. Recall that $\frac{\log n}{4n} \geqslant \max_{|j|\geqslant N_n} \frac{[T]_{jj}^2}{|j|(\omega_j)_{\vee 1}}$ due to Assumption 4.16, and hence

$$\{\widehat{N}_n > N_n\} \subset \Big\{\forall 1 \leqslant j \leqslant N_n \;\Big|\; \frac{\widehat{[T]}_{jj}^2}{|j|(\omega_j)_{\vee 1}} \geqslant \frac{\log n}{n}\Big\}$$
$$\subset \Big\{\frac{|\widehat{[T]}_{N_n}|}{|[T]_{N_n}|} \geqslant 2\Big\} \subset \Big\{|\widehat{[T]}_{N_n}/[T]_{N_n} - 1| \geqslant 1\Big\}.$$

Hoeffding's inequality and the definition of N yield $\mathbf{P}[\Omega_{II}] \leqslant 2(2016d/\lambda_1)^7 n^{-6}$, which by combining with (4.42) implies the result. □

4.5 Conclusion

In this chapter, we have developed a minimax theory for the estimation of the structural function in a nonparametric regression model with instrumental variables. We have defined an estimator based on the Galerkin solution which can attain the minimax optimal rate when the dimension parameter is chosen in an appropriate way. This choice, however, depends on characteristics of the conditional expectation operator which are not known.

In order to solve this problem, we have proposed a data-driven estimator which attains the minimax optimal rate over a wide range of classes. Unfortunately, we still need the additional assumption that the eigenfunctions of the conditional expectation operator are known, in which case the proposed estimator takes the form of an orthogonal series. Furthermore, we have shown in (4.17) that if $\mathcal{S}_k$ is the subspace generated by the first k eigenfunctions, then we have for all $k \leqslant k'$ and $t \in \mathcal{S}_k$ that $\langle t, \widehat{\varphi}_{k'}\rangle_\omega = \langle t, \widehat{\varphi}_k\rangle_\omega$. If, however, $\mathcal{S}_k$ is generated by an arbitrary set of linearly independent functions, this is not true in general. In particular, the estimate (4.20) on which the proof is essentially based, does not hold anymore.

Let us briefly outline a promising approach which might allow to drop the restrictive assumption of known eigenfunctions. Recall that in the proof of the upper risk bound for the adaptive estimator, we have first represented the orthogonal series estimator as a minimum contrast estimator in (4.18) by defining a suitable contrast function Υ. The data-driven choice of the dimension parameter was then the minimizer of this contrast subject to a stochastic penalty term that was the empirical version of the function pen from (4.24). The question is if similar proof techniques could work for unknown eigenfunctions if we use a different contrast function. We define a promising candidate of such a new contrast by

$$\Psi(k) := \max_{k\leqslant j\leqslant N_n} \{\|\widehat{\varphi}_j - \widehat{\varphi}_k\|_\omega^2 - \mathrm{pen}(j)\},$$

where N_n is the sequence defined in Definition 4.14. A partially adaptive choice of the dimension parameter k is then defined as

$$\widehat{k} := \operatorname*{argmin}_{1 \leqslant j \leqslant N_n} \{\Psi(j) + \operatorname{pen}(j)\}.$$

Using the monotonicity of pen, one can show that

$$\Psi(k) \leqslant 6 \sup_{k \leqslant j \leqslant N_n} \left[\|\widehat{\varphi}_j - \varphi_j\|_\omega^2 - (1/6)\operatorname{pen}(j)\right]_+ + 3 \sup_{k \leqslant j \leqslant N_n} \|\varphi_j - \varphi_j\|_\omega^2.$$

By the definition of $\widehat{k}$, it follows that for all $1 \leqslant k \leqslant N_n$

$$\begin{aligned}
\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 &\leqslant \left[\|\widehat{\varphi}_{\widehat{k}} - \widehat{\varphi}_{\widehat{k}\wedge k}\|_\omega^2 + \|\widehat{\varphi}_{\widehat{k}\wedge k} - \widehat{\varphi}_k\|_\omega^2 + \|\widehat{\varphi}_k - \varphi\|_\omega^2\right] \\
&\leqslant 3\left[\Psi(k) + \operatorname{pen}(\widehat{k}) + \Psi(\widehat{k}) + \operatorname{pen}(k) + \|\widehat{\varphi}_k - \varphi\|_\omega^2\right] \\
&\leqslant 6\left[\Psi(k) + \operatorname{pen}(k)\right] + 3\|\widehat{\varphi}_k - \varphi\|_\omega^2 \\
&\leqslant 42 \sup_{k \leqslant j \leqslant N_n} \left(\|\widehat{\varphi}_j - \varphi_j\|_\omega^2 - (1/6)\operatorname{pen}(j)\right)_+ + 18 \sup_{k \leqslant j \leqslant N_n} \|\varphi_j - \varphi_k\|_\omega^2 \\
&\qquad + 7\operatorname{pen}(k) + 6\|\varphi_k - \varphi\|_\omega^2
\end{aligned}$$

Moreover, we have

$$\sup_{k \leqslant j \leqslant N_n} \|\varphi_j - \varphi_k\|_\omega^2 \leqslant 4 \sup_{k \leqslant j \leqslant N_n} \|\varphi_j - \varphi\|_\omega^2.$$

And hence, for all $1 \leqslant k \leqslant N_n$,

$$\begin{aligned}
\|\widehat{\varphi}_{\widehat{k}} - \varphi\|_\omega^2 \leqslant 42 \sup_{k \leqslant j \leqslant N_n} &\left(\|\widehat{\varphi}_j - \varphi_j\|_\omega^2 - (1/6)\operatorname{pen}(j)\right)_+ \\
&+ 78 \sup_{k \leqslant j \leqslant N_n} \|\varphi_j - \varphi\|_\omega^2 + 7\operatorname{pen}(k).
\end{aligned}$$

Using $\sup_{k \leqslant j \leqslant N_n} \|\varphi_j - \varphi\|_\omega^2 \lesssim \|\varphi_k - \varphi\|_\omega^2$, we may expect a similar estimate as (4.28) to hold. However, it remains to control the term

$$\mathbf{E}\left[\sup_{k \leqslant j \leqslant N_n} \left(\|\widehat{\varphi}_j - \varphi_j\|_\omega^2 - (1/6)\operatorname{pen}(j)\right)_+\right]. \tag{4.43}$$

The analogous term appearing in the proof of Theorem 4.18 is controlled by Lemmas 4.24 and 4.25. The control of (4.43) is technically demanding and still work in progress.

Another interesting research question concerns the choice of the basis $\{f_j\}$ in the image space of the operator, because the estimator is only minimax optimal if this basis is appropriately chosen. The construction of the optimal

basis, though possible in theory (cf. Johannes and Breunig, 2009), requires a priori knowledge about the operator. It is natural to ask how a basis could be constructed without prior knowledge about the operator. A possible approach could be a criterion allowing to choose a basis from a finite set («library») of possible bases and to investigate to what extend the performance of the estimator could be improved depending on the available bases in the library.

Conclusion and future research

Before mentioning some yet unanswered questions which have arisen during the work on this thesis and which in my view are promising starting points for interesting future research projects, let me briefly explain which results of this thesis were the most challenging to prove and which were the most surprising to obtain. In the deconvolution model on the real line, the hardest problem consisted in obtaining a fully identified model without imposing any condition on the Fourier transform of the target density. The vast majority of identification conditions from the literature is based on conditions in the frequency domain and so I was surprised to see that in fact an easy to interpret condition is sufficient to guarantee identification in the deconvolution model under a normally distributed error with unknown variance. Another difficulty was the understanding of why the identification condition is not sufficient to ensure consistency as well, which is illustrated by a counterexample. While the proof of the identification result is finally surprisingly short and elementary, the consistency theorem demanded a greater technical effort.

In the last two chapters, the greatest challenge consisted in the construction of the adaptive estimator and in the control of its risk. While the general model selection procedure has already been used by other authors, its application in the particular models necessitated the solution of many technical difficulties. Surprising results in this context were the fact that adaptation is possible over a range of density classes including both polynomially and exponentially smooth densities, and that it is not even very costly in terms of convergence rates. In fact, the optimal rates are attained for a vast variety of classes.

Let us now turn to the still open questions. In Chapter 1 and 3, we have considered density deconvolution problems on the real line and on the circle. In both cases, the error term was assumed to be independent from the uncon-

taminated data. A natural modification of the model would be a relaxation of the independence condition. How could the dependence be modeled (possibly using copulas?) and which restrictions would be necessary in order to preserve the identifiability of the solution by the contaminated sample? Would similar estimation methods still yield comparable results? In Chapter 3, developing the minimax theory, we have considered density classes defined in terms of characteristic functions. This corresponds well to classical smoothness assumptions, but in case of densities with pronounced local features such as discontinuities or sharp peaks, it would be of interest as well to consider wavelet based classes.

Chapter 2 was devoted to robust frontier estimation in the presence of noise in the data. We have considered the case where the noise was in the input variable and argued that a slight modification of the same technique would still work when the noise was in the output variable instead. However, it is not obvious how to treat the case of error in both the input and the output variable, which seems however a realistic assumption in real life applications. The two errors could be independent of each other or correlated in some way. In either case, identification problems arise. What modeling assumptions would allow for identification and for consistent estimation in this case? Moreover, the estimation of productivity which is measured by the distance of an individual production unit to the frontier is another interesting problem. In the context of the model discussed in Chapter 2, the horizontal or vertical distance suggest themselves, but one could also be interested in directional distances, which presents us with the problem of formulating the model such that an underlying deconvolution problem becomes manifest.

In Chapter 4, we have developed an adaptive estimator in the context of a nonparametric regression model with instrumental variables. We needed the assumption that the eigenfunctions of the conditional expectation operator were known. Note that we have used an analogous implicit assumption in the circular deconvolution model in Chapter 3. In that case, this was no restriction because the eigenfunctions of the convolution operator are indeed known. In the regression framework, however, this assumption is restrictive and the most natural question is how to estimate the regression function in the same model when the eigenfunctions are not known. We have outlined an approach to this question in the concluding section of Chapter 4, but much technical work remains to be done. One could further investigate the problem of missing data, that is the case where some replications of the instrument or of the regressor have not been observed.

Finally, one could drop the iid. hypothesis and examine the effect of a time series structure in the data in any of the models we have treated in this thesis.

Auxiliary definitions and results

Definition A.1 (Hellinger, Kullback-Leibler) *Consider probability measures P and Q on some measurable space $(\Omega, \mathcal{A})$. Suppose that P and Q have densities p and q, respectively, with respect to a dominating measure λ. The* Hellinger distance *between P and Q is defined as*

$$H^2(P,Q) := \int_\Omega (\sqrt{p} - \sqrt{q})^2 \,\mathrm{d}\lambda = 2\,(1 - \int_\Omega \sqrt{pq}\,\mathrm{d}\lambda).$$

This distance does not depend on the choice of the dominating measure. The quantity $\rho(P,Q) := \int_\Omega \sqrt{pq}\,\mathrm{d}\lambda$ is called the Hellinger affinity *of P and Q. The Kullback-Leibler divergence between P and Q is given by*

$$KL(P,Q) := \begin{cases} \int_\Omega \log \frac{dP}{dQ}\, dP & (P \ll Q) \\ 0 & (otherwise). \end{cases}$$

For any two probability measrues P and Q, the Hellinger distance and the Kullback-Leibler divergence satisfy

$$H^2(P,Q) \leqslant KL(P,Q).$$

A more detailed discussion of distances between probability measures can be found in Tsybakov (2004), for example.

Definition A.2 (Weak derivative) *Let $u \in L^2(I)$ for some bounded interval $I \subset \mathbb{R}$. A function $v \in L^2(I)$ is called weak derivative of u if*

$$\int_I v(t)\varphi(t)\,\mathrm{d}t = -\int_I u(t)\varphi'(t)\,\mathrm{d}t$$

for every infinitely differentiable function φ with compact support. If a weak derivative exists, it is unique in $L^2(I)$.

Theorem A.3 (Petrov (1995)) *Let $X_1, \ldots, X_n$ be independent random variables with zero means, and let $p \geqslant 2$. Then*

$$\mathbf{E}\left[\Big|\sum_{k=1}^{n} X_k\Big|^p\right] \leqslant C(p)\, n^{p/2-1} \sum_{k=1}^{n} \mathbf{E}[|X_k|^p],$$

where $C(p)$ is a positive constant depending only on p.

Theorem A.4 (Hoeffding (1963)) *Let $X_1, \ldots, X_n$ be independent real-valued random variables. Assume that there are intervals $[a_i, b_i] \subset \mathbb{R}$ such that $\mathbf{P}(X_i \in [a_i, b_i]) = 1$ for all $i = 1, \ldots, n$. Letting $S_n := \sum_{i=1}^n X_i$, we have, for all $t > 0$,*

$$\mathbf{P}(S_n - \mathbf{E}[S_n] \geqslant t) \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

$$\mathbf{P}(|S_n - \mathbf{E}[S_n]| \geqslant t) \leqslant 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Theorem A.5 (Talagrand (1996)) *Let $T_1, \ldots, T_n$ be independent random variables and $\nu_n^*(r) = (1/n)\sum_{i=1}^n \big[r(T_i) - \mathbf{E}[r(T_i)]\big]$, for r belonging to a countable class $\mathcal{R}$ of measurable functions. Then,*

$$\mathbf{E}[\sup_{r\in\mathcal{R}} |\nu_n^*(r)|^2 - 6H_2^2]_+ \leqslant C\left(\frac{v}{n}\exp(-(nH_2^2/6v)) + \frac{H_1^2}{n^2}\exp(-K_2(nH_2/H_1))\right)$$

with numerical constants $K_2 = (\sqrt{2}-1)/(21\sqrt{2})$ and $C > 0$ and where

$$\sup_{r\in\mathcal{R}} \|r\|_\infty \leqslant H_1, \quad \mathbf{E}\left[\sup_{r\in\mathcal{R}} |\nu_n^*(r)|\right] \leqslant H_2, \quad \sup_{r\in\mathcal{R}} \frac{1}{n}\sum_{i=1}^n \mathbf{V}\mathrm{ar}(r(T_i)) \leqslant v.$$

List of symbols

ψ^X	Characteristic function of the random variable X
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathcal{N}_\sigma$	Abbreviation for $\mathcal{N}(0, \sigma^2)$
$\mathcal{B}(E)$	Borel σ-algebra of the topological space E
$\mathbb{N}$	The set of positive integers
$\mathbb{Z}$	The set of integers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_+$	The set of positive real numbers
$\mathbb{C}$	The set of complex numbers
$\lvert A\rvert$	The Lebesgue measure of a Borel set $A \subset \mathbb{R}$
$\mathbf{1}_A$,$\mathbf{1}_{[A]}$,$\mathbf{1}_{\{A\}}$	Indicator function of the event A
$(X_{(k)})_{k=1,\dots,n}$	Ordered version of the (random) vector $(X_k)_{k=1,\dots,n} \in \mathbb{R}^n$
$\xrightarrow{\mathcal{D}}$	Weak convergence
$\xrightarrow{\mathcal{V}}$	Vague convergence (p.22)
$[f]_j$	j-th coefficient of f with respect to a given basis (p.3)
$f * g$	Convolution of the functions f and g (p.14)
$a_n \lesssim b_n$	«There is $C > 0$ such that $a_k < Cb_k$ all $k \geqslant 1$»
$a_n \sim b_n$	$a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously
$\mathbf{P}[A]$	Probability of the event A
$\mathbf{E}[X]$	Expected value of the random variable X
iid.	«independent and identically distributed»
a.s.	«almost surely»
$a_n = O(b_n)$	$\limsup_{n\to\infty} \lvert a_n/b_n\rvert < \infty$
$a_n = o(b_n)$	$\lim_{n\to\infty} \lvert a_n/b_n\rvert = 0$
cdf	«cumulative distribution function»
MISE	«mean integrated squared error»
$KL(P, Q)$	Kullback-Leibler divergence between the measures P and Q (p.98)
$H^2(P, Q)$	Hellinger distance between the measures P and Q (p. 131)

$\rho^2(P,Q)$	Hellinger affinity between the measures P and Q (p. 131)
$\perp$	stochastic independence
$\text{span}\{v_1,\ldots,v_n\}$	the linear subspace generated by the vectors $v_1,\ldots,v_n$
$P \ll Q$	absolute continuity of the measure P with respect to Q

Bibliography

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71:1795–1843.

Amemiya, T. (1974). The nonlinear two-stage least square estimator. *Journal of Econometrics*, 2:105–110.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413.

Bigot, J. and Van Bellegem, S. (2009). Log-density deconvolution by wavelet thresholding. *Scandinavian Journal of Statistics*, 36:749–763.

Blundell, R., Chen, X., and Kristensen, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica*, 75:1613–1669.

Blundell, R. and Horowitz, J. L. (2007). A non-parametric test of exogeneity. *The Review of Economic Studies*, 74:1035–1058.

Bosq, D. (1998). *Nonparametric statistics for stochastic processes*. Springer, New York.

Butucea, C. and Matias, C. (2005). Minimax estimation of the noise level and of the deconvolution density in a semiparametric deconvolution model. *Bernoulli*, 11:309–340.

Butucea, C., Matias, C., and Pouet, C. (2008). Adaptivity in convolution models with partially known noise distribution. *Electronic Journal of Statistics*, 2:897–915.

Cardot, H. and Johannes, J. (2010). Thresholding projection estimators in functional linear models. *Journal of Multivariate Analysis*, 101:395–408.

Carrasco, M., Florens, J.-P., and Renault, E. (2007). *Linear Inverse Problems and Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization*, volume 6B of *Handbook of Econometrics*. J. Heckman and E. Leamer.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83:1184–1186.

Cavalier, L. and Hengartner, N. W. (2005). Adaptive estimation for inverse problems with noisy operators. *Inverse Problems*, 21:1345–1361.

Cazals, C., Florens, J.-P., and Simar, L. (2002). Nonparametric frontier estimation: A robust approach. *Journal of Econometrics*, 106:1–25.

Chen, X. and Reiss, M. (2011). On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*. In press.

Chung, K. (1968). *A course in probability theory*. Harcourt, Brace and World.

Cochran, W. W., Mouritsen, H., and Wikelski, M. (2004). Migrating songbirds recalibrate their magnetic compass daily from twylight cues. *Science*, 304:405–408.

Comte, F. and Johannes, J. (2010). Adaptive estimation in circular functional linear models. *Mathematical Methods of Statistics*, 19:42–63.

Comte, F., Rozenholc, Y., and Taupin, M.-L. (2006). Penalized contrast estimator for adaptive density deconvolution. *Canadian Journal of Statistics*, 34:431–452.

Comte, F., Rozenholc, Y., and Taupin, M.-L. (2007). Finite sample penalization in adaptive density deconvolution. *Journal of Statistical Computation and Simulation*, 77:977–1000.

Comte, F. and Taupin, M.-L. (2003). Adaptive density deconvolution for circular data. Discussion Paper MAP5 2003-10, Université Paris 5.

Curray, J. R. (1956). The analysis of two-dimensional orientation data. *The Journal of Geology*, 64:117–131.

Daouia, A., Florens, J.-P., and Simar, L. (2009). Regularization in nonparametric frontier estimators. Discussion paper, Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Belgium.

Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. (2001). Nonparametric instrumental regression. *Econometrica*. To appear.

Daskovska, A., Simar, L., and Van Bellegem, S. (2010). Forecasting the Malmquist productivity index. *Journal of Productivity Analysis*, 33:97–107.

De Borger, B., Kerstens, K., Moesen, W., and Vanneste, J. (1994). A non-parametric free disposal hull (FDH) approach to technical efficiency: an illustration of radial and graph efficiency measures and some sensitivity results. *Swiss Journal of Economics and Statistics*, 130:647–667.

Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *The Annals of Statistics*, 36:665–685.

Deprins, D., Simar, L., and Tulkens, H. (1984). Measuring labor inefficiency in post offices. In Marchand, M., Pestieau, P., and Tulkens, H., editors, *The Performance of Public Enterprises: Concepts and Measurements*, pages 243–267, Amsterdam. North-Holland.

Devroye, L. and Györfi, L. (1985). *Nonparametric density estimation. The L_1 view.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Efromovich, S. (1997). Density estimation for the case of supersmooth measurement error. *Journal of the American Statistical Association*, 92:526–535.

Efromovich, S. and Koltchinskii, V. (2001). On inverse problems with unknown operators. *IEEE Transactions on Information Theory*, 47:2876–2894.

Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of inverse problems.* Mathematics and its Applications. Kluwer Academic Publishers.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19:1257–1272.

Färe, R., Grosskopf, S., and Knox Lovell, C. (1985). *The Measurements of Efficiency of Production*, volume 6 of *Studies in Productivity Analysis*. Springer, New York.

Fisher, N. (1993). *Statistical analysis of circular data.* Cambridge University Press.

Florens, J.-P. (2003). Inverse problems and structural econometrics: The example of instrumental variables. In Dewatripont, M., Hansen, L. P., and Turnovsky, S. J., editors, *Advances in Economics and Econometrics: Theory and Applications – Eight World Congress*, volume 36 of *Econometric Society Monographs*. Cambridge University Press.

Florens, J.-P., Johannes, J., and Van Bellegem, S. (2009). Instrumental regression in partially linear models. Discussion Paper 0537, Institut de statistique, biostatistique et scieces actuarielles, Université catholique de Louvain (first version 2005, revised).

Florens, J.-P., Johannes, J., and Van Bellegem, S. (2011). Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory*. To appear.

Gagliardini, P. and Scaillet, O. (2006). Tikhonov regularization for functional minimum distance estimators. Swiss Finance Institute Research Paper No. 06-30.

Gill, J. and Hangartner, D. (2010). Circular data in political science and how to handle it. *Political Analysis*, 18:316–336.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52.

Hall, P. and Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33:2904–2929.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35:70–91.

Hall, P. and Simar, L. (2002). Estimating a changepoint, boundary, or frontier in the presence of observation error. *Journal of the American Statistical Association*, 97:523–534.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:pp. 13–30.

Hoffmann, M. and Reiss, M. (2008). Nonlinear estimation for linear inverse problems with error in the operator. *The Annals of Statistics*, 36:310–336.

Horowitz, J. L. and Lee, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, 75:1191–1208.

Horrace, W. C. and Parmeter, C. F. (2011). Semiparametric deconvolution with unknown error variance. *Journal of Productivity Analysis*, 35:129–141.

Jain, N. C. and Orey, S. (1979). Vague convergence of sums of independent random variables. *Israel Journal of Mathematics*, 33:317–348.

Johannes, J. and Breunig, C. (2009). On rate optimal local estimation in nonparametric instrumental regression. Technical report, University Heidelberg (submitted.). `arxiv:0902.2103`.

Johannes, J. and Schwarz, M. (2009). Adaptive circular deconvolution by model selection under unknown error distribution. Discussion Paper 0931, Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain.

Johannes, J. and Schwarz, M. (2010). Adaptive nonparametric instrumental regression by model selection. Discussion Paper 1026, Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain. Submitted.

Johannes, J., Van Bellegem, S., and Vanhems, A. (2011). Convergence rates for ill-posed inverse problems with an unknown operator. *Econometric Theory*. To appear.

Johnstone, I., Kerkyacharian, G., Picard, D., and Raimondo, M. (2004). Wavelet deconvolution in a periodic setting. *Journal of the Royal Statistical Society: Series B*, 66:547–573.

Kawata, T. (1972). *Fourier analysis in probability theory.* Academic Press, New York.

Kneip, A., Park, B., and Simar, L. (1998). A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory*, 14:783–793.

Kneip, A., Simar, L., and Van Keilegom, I. (2010). Boundary estimation in the presence of measurement error with unknown variance. Discussion Paper, Institut de statistique, biostatistique et sciences actuarielles, Univertité catholique de Louvain.

Korostolev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory for Image Reconstruction.*, volume 82 of *Lecture Notes in Statistics*. Springer.

Leleu, H. (2006). A linear programming framework for free disposal hull technologies and cost functions: Primal and dual models. *European Journal of Operational Research*, 168:340–344.

Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis*, 65:139–165.

Loubes, J.-M. and Marteau, C. (2009). Oracle inequality for instrumental variable regression. `arXiv:0901.4321v1`.

Mardia, K. (1972). *Statistics of directional data.* Probability and Mathematical Statistics. A Series of Monographs and Textbooks. Vol. 13. Academic Press.

Matias, C. (2002). Semiparametric deconvolution with unknown noise variance. *European Series in Applied and Industrial Mathematics: Probability and Statistics*, 6:271–292.

Meister, A. (2006). Density estimation with normal measurement error with unknown variance. *Statistica Sinica*, 16:195–211.

Meister, A. (2007). Deconvolving compactly supported densities. *Mathematical Methods in Statistics*, 16:63–76.

Meister, A. (2009). *Deconvolution problems in nonparametric statistics.* Lecture Notes in Statistics 193, Springer.

Meister, A., Stadtmüller, U., and Wagner, C. (2010). Density deconvolution in a two-level heteroscedastic model with unknown error density. *Electronic journal of Statistics*, 4:36–57.

Natterer, F. (1984). Error bounds for Tikhonov regularization in Hilbert scales. *Applicable Analysis*, 18:29–37.

Neubauer, A. (1988a). An a posteriori parameter choice for Tikhonov regularization in Hilbert scales leading to optimal convergence rates. *SIAM Journal on Numerical Analysis*, 25:1313–1326.

Neubauer, A. (1988b). When do Sobolev spaces form a Hilbert scale? *Proceedings of the American Mathematical Society*, 103:557–562.

Neumann, M. H. (1997). On the effect of estimating the error density in nonparametric deconvolution. *Journal of Nonparametric Statistics*, 7:307–330.

Neumann, M. H. (2007). Deconvolution from panel data with unknown error distribution. *Journal of Multivariate Analysis*, 98:1955–1968.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58:809–837.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71:1565–1578.

Olver, F. (1974). *Asymptotics and special functions.* Academic Press.

Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics.* Cambridge University Press.

Park, B. U., Sickles, R. C., and Simar, L. (2003). Semiparametric efficient estimation of AR(1) panel data models. *Journal of Econometrics*, 117:279–311.

Park, B. U., Simar, L., and Weiner, C. (2000). The FDH estimator for productivity efficiency scores: asymptotic properties. *Econometric Theory*, 16:855–877.

Pensky, M. and Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *The Annals of Statistics*, 27:2033–2053.

Petrov, V. V. (1995). *Limit theorems of probability theory. Sequences of independent random variables.* Oxford Studies in Probability. Clarendon Press.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837.

Schwarz, M. and Van Bellegem, S. (2010). Consistent density deconvolution under partially known error distribution. *Statistics and Probability Letters*, 80:236–241.

Schwarz, M., Van Bellegem, S., and Florens, J.-P. (2011). Nonparametric frontier estimation from noisy data. In *Festschrift in honour of Léopold Simar*. Springer (to appear).

Seiford, L. and Thrall, R. (1990). Recent developments in DEA: The mathematical programming approach to frontier analysis. *Journal of Econometrics*, 46:7–38.

Shephard, R. W. (1970). *Theory of cost and production functions.* Princeton University Press.

Simar, L. (2007). How to improve the performances of DEA/FDH estimators in the presence of noise? *Journal of Productivity Analysis*, 28:183–201.

Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, 21:169–184.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563.

Tsybakov, A. B. (2004). *Introduction to nonparametric estimation. (Introduction à l'estimation non-paramétrique.).* Mathématiques & Applications 41, Springer.

Walker, J. S. (1988). *Fourier analysis.* Oxford University Press.