# Statistical inference for inverse problems

**Nicolai Bissantz**[1] **and Hajo Holzmann**[2]

[1] Fakultät für Mathematik, Ruhr-Universität Bochum, NA 3/70, Universitätsstr, 150,
D-44780 Bochum, Germany
[2] Institut für Stochastik, Universität Karlsruhe, Germany

E-mail: nicolai.bissantz@rub.de

## Abstract

In this paper we study statistical inference for certain inverse problems. We go beyond mere estimation purposes and review and develop the construction of confidence intervals and confidence bands in some inverse problems, including deconvolution and the backward heat equation. Further, we discuss the construction of certain hypothesis tests, in particular concerning the number of local maxima of the unknown function. The methods are illustrated in a case study, where we analyze the distribution of heliocentric escape velocities of galaxies in the Centaurus galaxy cluster, and provide statistical evidence for its bimodality.

(Some figures in this article are in colour only in the electronic version)

## 1. Introduction

Inverse problems have been studied intensively in the literature on numerical analysis and mathematical physics in the last few decades. See for example [16, 31, 32] for comprehensive treatments. For linear inverse problems, the function of interest $f$ is related to another function $g$ by an operator equation $g = Kf$, where $K : \mathcal{X} \to \mathcal{Y}$ is a known linear operator between Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$, often an injective integral operator. Typically, the inverse of the operator $K$ is unbounded and the function $g$ can only be observed with errors thus, the problem of recovering $f$ is ill-posed and regularization techniques are required.

There are different ways to model the observational error in the function $g$. In the deterministic approach, one assumes that the observed function $g^\delta$ satisfies $\|g^\delta - g\| \leqslant \delta$ for some $\delta > 0$. Thus, there is a deterministic upper bound to the observational error.

In contrast, in statistics, errors and observational uncertainties are modeled as influences of random quantities. For example, an extension of the operator equation $g = Kf$ to a regression framework with an additive random error is $g = Kf + \epsilon$ for random noise $\epsilon$. In recent years, several inverse problems have been reinvestigated in such statistical frameworks, including positron emission and x-ray tomography [9, 34], Wicksell's problem [22], the heat

equation [38] or problems related to satellite gradiometry [6]. Sometimes, inverse problems also directly arise in statistical problems, such as error in variable models (cf [18, 33]) or regression models with instrumental variables [24].

Apart from an arguably more accurate model of errors and random influences, the statistical approach allows us to draw statistical inference about the unknown object $f$. For example, one can construct regions in which $f$ will be located with a certain (high) probability, or one can conclude that with a (low) error probability, $f$ will have at least two local maxima. The aim of the present paper is therefore to take a look beyond mere estimation purposes and rather to discuss statistical inference for certain inverse problems.

Since we shall investigate inverse problems from the point of view of nonparametric curve estimation, in section 2 we briefly review the concept of nonparametric (contrasted with parametric) estimation for two fundamental direct estimation problems, namely density estimation and regression estimation. Section 3 deals with statistical inference in inverse density estimation, where we focus on density deconvolution. Specifically, we discuss the construction of asymptotic confidence intervals, uniform confidence bands, inference on the number of modes and on the scale space surface for the target density $f$. Further, in section 4 we are concerned with statistical inference in some inverse regression models, namely a statistical model for the backward heat equation and for a regression deconvolution model. As a case study, in section 5 we analyze the heliocentric escape velocities of galaxies in the Centaurus galaxy cluster. Finally, we conclude with a discussion in section 6.

## 2. Nonparametric curve estimation

Before turning to curve estimation in inverse problems, in this section we briefly discuss two basic problems in direct nonparametric curve estimation, namely regression analysis and density estimation. These will be extended to indirect problems in sections 3 and 4.

### 2.1. Nonparametric regression

In a regression problem, paired observations $(Y, X)$ are available, and one is interested in determining a proper relationship between these two variables. The variable $X$ is thought of as being a predictor for the other variable $Y$, which is called dependent variable or response. The response $Y \in \mathbb{R}$ is univariate, but the predictor $X \in \mathbb{R}^d$ can be multivariate. For simplicity, we here assume that $d = 1$, which is also called a simple regression problem.

For example, one might be interested in relating budget share with log real income, or log income with age, or weight with height of a person.

In a linear regression model, one assumes that a sample $(Y_i, X_i)$, $i = 1, \ldots, n$, is available, which satisfies the relationship

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where $(\beta_0, \beta_1)$ are the unknown constant regression coefficients to be estimated, the $\epsilon_i$ are independent, normally distributed random variables with equal variance $\sigma^2$. Thus, the assumed underlying linear relationship cannot be observed precisely, but rather with a stochastic error $\epsilon_i$, the magnitude of which is controlled by the variance $\sigma^2$. The predictor variable $X_i$ can either be modeled as deterministic (fixed design) or as a random variable (random design). In the latter case, it is then usually assumed to be independent of $\epsilon_i$.

The linear regression model is an example of a parametric regression model, in which the functional relationship between $Y_i$ and $X_i$ is determined by a finite-dimensional parameter $((\beta_0, \beta_1)$ for linear regression). Other parametric regression models are polynomial regression

models $Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i + \epsilon_i$, for fixed $p \geqslant 1$, or periodic models such as $Y_i = \beta_1 \cos(\beta_2 X_i) + \epsilon_i$. Although parametric regression analysis is quite popular, there is a substantial danger of reaching incorrect conclusions from an analysis based on a parametric family which is not (at least approximately) of appropriate form.

Therefore, recent research in regression analysis has concentrated on estimating general relationships of the form

$$Y_i = \theta(X_i) + \epsilon_i, \tag{1}$$

where $\theta$ is a smooth, but otherwise unspecified function. (1) is called a nonparametric regression model, since $\theta$, except for being smooth, is not further specified. There are a variety of estimation methods for $\theta$ in (1), including local polynomial estimators [20] spline regression [17] and series estimators [17].

Apart from mere estimation purposes, a major use of nonparametric estimates $\hat{\theta}$ is for statistical inference. For example, by comparing a nonparametric estimate with a parametric estimate, one can validate a parametric form for a regression model [26]. Further, statistical inference allows us to construct confidence intervals for $\theta(x)$ based on $\hat{\theta}(x)$, i.e. intervals in which the unknown function lies with a certain (typically 95%) probability [17], or even to construct uniform confidence intervals for each $x$ in a subset $I \subset \mathbb{R}$, which are also called confidence bands [46]. These objects are useful for descriptive purposes as well as for testing statistical hypotheses on $\theta$, e.g. testing whether $\theta$ is non-negative at a certain $x$ or on an interval $I$.

## 2.2. Nonparametric density estimation

A similarly fundamental problem as the regression problem is the estimation of the (probability) density function of a univariate random sample. More specifically, assume that $X_1, \ldots, X_n$ are independent random variables with common density $f$. In the parametric approach, one assumes that $f$ belongs to a certain parametric family of densities such as the normal or the beta family, and then estimates the parameters of this family, e.g. by maximum likelihood estimation. Again, the possible shapes of a parametric family of densities are quite restricted, and therefore misspecifying the parametric family can lead to incorrect conclusions about the shape of the density $f$. Therefore, recent research has concentrated on nonparametric estimation of $f$, assuming smoothness but no parametric form.

The distribution of the $X_i$ is determined both by their probability density $f$ and by their distribution function $F$, which are related by $F(x) = \int_{-\infty}^{x} f(t)\,dt$, so that $F' = f$. Thus, the nonparametric density estimation, i.e. estimation of the derivative $f$ of $F$, may be considered as a first example of an ill-posed inverse problem arising in statistics.

There are several nonparametric methods for density estimation, including kernel estimates [44], estimates based on wavelets [45] or nearest-neighbor estimators [14]. Again, such nonparametric estimates $\hat{f}$ will allow statistical inference for the unknown density $f$, in the form of confidence intervals [23] or confidence bands [4], as well as testing parametric forms for $f$ [4].

Nonparametric estimates (both of densities and of regression functions) typically involve selection of an additional smoothing parameter. Although this may be seen as a disadvantage of nonparametric estimation, it also allows for additional flexibility, in particular if the estimator is investigated for distinct values of the smoothing parameter [10, 41]. In this way one can for example investigate the number of modes (i.e. local maxima) of the unknown function $f$ (or $\theta$ in the regression context).

## 3. Indirect density estimation: deconvolution

### 3.1. Deconvolution density estimation

In contrast to direct nonparametric estimation of a density $f$, in indirect density estimation one does not have observations distributed according to $f$, but rather to another density $g$ which is related to $f$ by an equation $g = Kf$, where $K$ is often an integral operator, but it can also be a nonlinear operator with an unbounded inverse.

   As an example, suppose that the density $f$ of the random variables $X_i$ is the density of interest. However, only noisy versions $Y_i = X_i + \epsilon_i$ of the $X_i$ can be observed. Here, $\epsilon_i$ are unobserved errors, distributed with a density $\psi$ and independent of $X_i$. Hence, we have for the density $g$ of the $Y_i$

$$g = f * \psi.$$

Thus, $K = K_\psi$ is the convolution operator, and recovering $f$ from observations $Y_i$ distributed according to $g$ is called the deconvolution problem. Evidently, identification and hence estimation of $f$ are only possible with some additional knowledge about $\psi$. The simplest, most common assumption is that $\psi$ is known [18]. Otherwise, additional or repeated observations [11, 40] or strong shape restrictions on $f$ and $\psi$ are required.

   Other problems which lead to indirect density estimation include Positron emission tomography [9, 34] and quantum homodyne tomography [7]. Here, in order to illustrate the main ideas and phenomena, we shall focus on density deconvolution with known error density.

   It is well known that the difficulty of recovering $f$ in the deconvolution problem depends sensitively on the smoothness of the error density $\psi$ (and also on the smoothness of $f$ itself). Roughly speaking, the error density is called ordinary smooth if its Fourier transform $|\Phi_\psi(t)|$ decays at a polynomial rate as $t \to \infty$, in which case the problem is mildly ill-posed, whereas if $|\Phi_\psi(t)|$ decays at an exponential rate as $t \to \infty$, $\psi$ is called supersmooth and the problem is severely ill-posed. In terms of optimal rates of convergence of the mean square error as well as the integrated mean square error, an ordinary-smooth error density in general leads to polynomials rates, whereas a supersmooth error density typically leads to logarithmic rates [18].

   The arguably most popular estimator in the deconvolution problem is an estimator of kernel type [18, 19, 43]. More precisely, under the assumption that $\Phi_\psi(t) \neq 0$ for all $t \in \mathbb{R}$ and that $\Phi_\mathcal{K}$, the Fourier transform of the kernel $\mathcal{K}$, has compact support, the kernel deconvolution density estimator for $f$, given by

$$\hat{f}_n(x; h) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - Y_k}{h}; h\right), \tag{2}$$

where

$$K(x; h) = \frac{1}{2\pi} \int_\mathbb{R} \exp(-\mathrm{i}tx) \frac{\Phi_\mathcal{K}(t)}{\Phi_\psi(t/h)} \, \mathrm{d}t, \tag{3}$$

is well defined. Here $h > 0$ is a smoothing parameter called bandwidth, and $K(x; h)$ is called the deconvolution kernel. The estimator in (2) can be derived by general spectral regularization methods [6, 38]. Note that it satisfies the equation

$$\hat{g}_n(x; h) = (\hat{f}_n(\cdot; h) * \psi)(x),$$

where $\hat{g}_n(x; h) = (nh)^{-1} \sum_i \mathcal{K}((x - Y_i/h))$ is the (direct) kernel estimator of $g$ with kernel $\mathcal{K}$ and bandwidth $h$. Thus, $\hat{f}_n$ can also be obtained in the particularly simple way of applying $K_\psi^{-1}$ to a sufficiently smooth estimate of $g$.

The estimation error of $\hat{f}_n$ is decomposed into a stochastic part and a deterministic part (a systematic error) as follows:

$$\hat{f}_n(x; h) - f(x) = (\hat{f}_n(x; h) - E\hat{f}_n(x; h)) + (E\hat{f}_n(x; h) - f(x)). \qquad (4)$$

Since $E\hat{f}_n(x; h) = \mathcal{K}_h * f(x)$, where $\mathcal{K}_h(x) = \mathcal{K}(x/h)/h$, the deterministic error (bias) $E\hat{f}_n(x; h) - f(x)$ does not depend on the error density $\psi$. In the following we shall discuss statistical inference for $f$ based on $\hat{f}_n$, separately for ordinary smooth and supersmooth error densities, concentrating on the stochastic part $\hat{f}_n(x; h) - E\hat{f}_n(x; h)$ in the error decomposition (4).

### 3.2. The ordinary smooth case

As mentioned above, the error density $\psi$ is called ordinary smooth if its Fourier transform decays at a polynomial rate. In order to derive the asymptotic distribution of the deconvolution kernel density estimator, we shall require the slightly more restrictive assumption,

$$\Phi_\psi(t)t^\beta \to C_\epsilon, \qquad t \to \infty, \qquad (5)$$

for some $\beta \geqslant 0$ and $C_\epsilon \in \mathbb{C} \setminus \{0\}$. Note that this implies that $\Phi_\psi(t)|t|^\beta \to \bar{C}_\epsilon, t \to -\infty$. If (5) holds, the deconvolution kernel $K(\cdot, h)$ given in (3) has a simple asymptotic form. In fact, from the dominated convergence theorem (cf [19]),

$$h^\beta K(x; h) \to K(x), \qquad h \to 0,$$

where

$$K(u) = \frac{1}{2\pi C_\epsilon} \int_0^\infty \exp(-iux)x^\beta \Phi_\mathcal{K}(x) \, dx + \frac{1}{2\pi \bar{C}_\epsilon} \int_{-\infty}^0 \exp(-iux)|x|^\beta \Phi_\mathcal{K}(x) \, dx.$$

If (5) and some additional regularity conditions hold, for $h \to 0$ and $nh \to \infty$ one can show that $\hat{f}_n(x; h) - E\hat{f}_n(x; h)$ is asymptotically normally distributed with variance of order $n^{-1/2}h^{-(\beta+1/2)}$. Specifically, if $g(x) > 0$,

$$\sqrt{n}h^{\beta+1/2}(\hat{f}_n(x; h) - E\hat{f}_n(x; h)) \to N(0, g(x)\kappa),$$
$$\kappa = \int_\mathbb{R} |t|^{2\beta}|\Phi_\mathcal{K}(t)|^2 \, dt/|C_\epsilon|^2. \qquad (6)$$

From this we can construct asymptotic level-$\alpha$ confidence intervals for the smoothed version $E\hat{f}_n(x; h) = (\mathcal{K}_h * f)(x)$ of $f$

$$C_\alpha(x) = \left[ \hat{f}_n(x; h) - \frac{q_{1-\alpha/2}(\kappa\tilde{g}(x))^{1/2}}{\sqrt{n}h^{\beta+1/2}}, \hat{f}_n(x; h) + \frac{q_{1-\alpha/2}(\kappa\tilde{g}(x))^{1/2}}{\sqrt{n}h^{\beta+1/2}} \right], \quad (7)$$

which contains the true $(\mathcal{K}_h * f)(x)$ with an asymptotic probability of 95%. Here, $\tilde{g}(x)$ is a consistent estimator for $g$, and $q_\alpha$ is the $\alpha$-quantile of the standard normal distribution.

In order to obtain a confidence interval for $f(x)$, one has to deal with the bias $|E\hat{f}_n(x; h) - f(x)|$. Roughly speaking, for $j$-times differentiable densities $f$, this bias decays as $h^j$. More precisely, if $f$ satisfies a Sobolev condition of type

$$\int_\mathbb{R} |\Phi_f(t)||t|^j \, dt < \infty,$$

and if, as is assumed above, a flat-top kernel is used (cf [37]), then

$$|E\hat{f}_n(x) - f(x)| = \frac{1}{2\pi} \left| \int_\mathbb{R} \exp(-itx)(1 - \Phi_\mathcal{K}(ht))\Phi_f(t) \, dt \right| = o(h^j), \quad (8)$$

5

uniformly in $x$. Therefore, in order to guarantee that (6) continues to hold if $E\hat{f}_n(x; h)$ is replaced by $f(x)$, one needs that $h^j = O(1/(\sqrt{n}h^{\beta+1/2}))$, or equivalently,

$$h = O(n^{-\frac{1}{2\beta+2j+1}}). \tag{9}$$

Note that due to the use of a flat-top kernel, the estimate for the bias in (8) is $o(h^j)$, and hence even for optimal estimation, the variance and the squared bias are not exactly balanced, the variance dominates the squared bias. Hence, explicit undersmoothing (i.e. choosing the bandwidth too small as compared to optimal estimation) is not required, and the bandwidth may be chosen at the order of $\sim n^{-\frac{1}{2\beta+2j+1}}$. In this case, (7) is also a valid asymptotic confidence interval for $f(x)$. We shall come back to the issue of bandwidth selection in section 3.4.

The confidence intervals $C_\alpha(x)$ in (7) are only valid asymptotically, i.e. for large sample sizes $n$. In order to improve the finite sample performance, statisticians use a resampling method called the bootstrap. The basic idea is to replace the unknown underlying population by the sample, and resample from this given sample [3].

Thus, resampling $n$-times from the observations $Y_1, \ldots, Y_n$ we obtain an i.i.d. sample $Y_1^*, \ldots, Y_n^*$ with distribution $G_n$, the empirical distribution function of $Y_1, \ldots, Y_n$. Denote by $E^*$ the conditional expectation given $Y_1, \ldots, Y_n$. The bootstrap estimator of $f$ is given by

$$\hat{f}_n^*(x; h) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - Y_k^*}{h}; h\right).$$

Since $E^*\hat{f}_n^*(x; h) = \hat{f}_n(x; h)$, we simulate the distribution of $|Z_n^*(x)|$, where

$$Z_n^*(x) = \frac{n^{1/2}h^{\beta+1/2}}{\tilde{g}(x)^{1/2}}(\hat{f}_n^*(x; h) - \hat{f}_n(x; h)). \tag{10}$$

Let $q_{1-\alpha}^*$ denote the $(1 - \alpha)$-quantile of this simulated distribution. Then

$$C_\alpha^*(x) = \left[\hat{f}_n(x; h) - \frac{q_{1-\alpha}^* \tilde{g}^{1/2}(x)}{\sqrt{n}h^{\beta+1/2}}, \hat{f}_n(x; h) + \frac{q_{1-\alpha}^* \tilde{g}^{1/2}(x)}{\sqrt{n}h^{\beta+1/2}}\right] \tag{11}$$

is a level-$\alpha$ bootstrap confidence interval for $\mathcal{K}_h * f(x)$ for general $h$, and for a bandwidth $h$ chosen according to (9) also for $f(x)$ itself.

The confidence interval $C_\alpha(x)$ in (7) or the bootstrap confidence interval $C_\alpha^*(x)$ in (11) can be constructed at each $x$ in some interval $I$, and in this way one can obtain a pointwise confidence band. However, this picture is misleading since the curve does not lie uniformly within this band (with a certain error probability $\alpha$), but only at each point $x$. In other words, the probability that the true $f$ will not lie in the band at some $x \in I$ may be much higher than $\alpha$. Therefore, instead of pointwise confidence bands one is rather interested in uniform confidence bands, which are designed to satisfy this uniformity property.

In fact, under some additional regularity conditions, in [5] it is shown that for an undersmoothing bandwidth $h$ (here, explicit undersmoothing is still required), if $g > 0$ uniformly on $I$, one has that

$$P(\hat{f}_n(x; h) - b_n(x, t)$$
$$\leqslant f(x) \leqslant \hat{f}_n(x; h) + b_n(x, t) \text{ for all } x \in [0, 1]) \rightarrow \exp(-2\exp(-t)), \tag{12}$$

where

$$b_n(x, t) = \left(\frac{\tilde{g}_n(x)\kappa}{nh^{2\beta+1}}\right)^{1/2} \left(\frac{t}{(2\log(1/h))^{1/2}} + d_n\right),$$

$$d_n = (2\log(1/h))^{1/2} + \frac{\log\left(\frac{1}{2\pi}C_{K,2}^{1/2}\right)}{(2\log(1/h))^{1/2}}, \qquad C_{K,2} = \frac{\int_{\mathbb{R}} x^{2\beta+1}\Phi_{\mathcal{K}}^2(x)\,\mathrm{d}x}{\int_{\mathbb{R}} x^{2\beta}\Phi_{\mathcal{K}}^2(x)\,\mathrm{d}x}.$$

Thus, choosing $t_\alpha$ such that $\exp(-2\exp(-t_\alpha)) = 1 - \alpha$, one obtains the following level-$\alpha$ confidence band on $I = [0, 1]$,

$$C_\alpha^{\mathrm{uni}}(I) = \{[\hat{f}_n(x; h) - b_n(x, t_\alpha), \hat{f}_n(x; h) + b_n(x, t_\alpha)], x \in I\}.$$

Let us make some remarks concerning the uniform band $C_\alpha^{\mathrm{uni}}(I)$. First, (12) and hence the construction of $C_\alpha^{\mathrm{uni}}(I)$ can easily be transfered to any compact interval $I$ on which $g > 0$. Second, the uniform confidence band is wider than the pointwise band by a factor of $(\log(1/h))^{1/2}$. Thus, the additional cost of uniformity is quite low at a first glance. However, the convergence in (12) is rather slow (much slower than in (6)), and bootstrapping is highly recommended. For constructing bootstrap confidence bands, one bootstraps the distribution of $\sup_{x \in I} |Y_n^{*}(x)|$, where $Y_n^{*}(x)$ is defined in (10). If $q_{1-\alpha}^{*,\mathrm{unif}}$ denotes its $(1 - \alpha)$-quantile, then a bootstrap confidence band is given by

$$C_\alpha^{*,\mathrm{unif}}(x) = \left\{ \left[ \hat{f}_n(x; h) - \frac{q_{1-\alpha}^{*,\mathrm{unif}} \tilde{g}^{1/2}(x)}{\sqrt{n} h^{\beta+1/2}}, \hat{f}_n(x; h) + \frac{q_{1-\alpha}^{*,\mathrm{unif}} \tilde{g}^{1/2}(x)}{\sqrt{n} h^{\beta+1/2}} \right], x \in I \right\}. \tag{13}$$

### 3.3. Normally distributed errors

In the supersmooth case, the error density has a Fourier transform with exponentially decreasing tails. Apart from logarithmic rates of convergence, supersmooth errors can also lead to a distinct asymptotic behavior [28, 43]. Here we discuss the results for centered normally distributed errors, which have characteristic function $\Phi_\psi(t) = \mathrm{e}^{-|t|^2 \sigma^2}$, $\sigma^2 > 0$. When using the sinc kernel with characteristic function $\Phi_{\mathcal{K}}(t) = 1_{[-1,1]}(t)$, in [43] the following asymptotic behavior for $h \to 0$ and $nh \to \infty$ is obtained:

$$\frac{\sqrt{n}}{h\,\mathrm{e}^{\sigma^2/h^2}} (\hat{f}_n(x; h) - E\hat{f}_n(x; h)) \to N\left(0, \frac{1}{8\pi^2\sigma^4}\right). \tag{14}$$

Thus, the form of the asymptotic variance changes, in contrast to (6) it does no longer depend on $g$ or $x$ at all. The asymptotics in (14) could be used for constructing asymptotic confidence intervals similarly as in (7), but with proper changes according to the form of the asymptotic variance, in particular, the estimate of $g(x)$ is not needed. However, this is not recommended since convergence in (14) is quite slow, one should instead use bootstrap confidence intervals. Another sensitive problem is the choice of the bandwidth in the supersmooth case, since this falls into the so-called bias-dominating case [8]. This means that for optimal estimation, the bandwidth is chosen in such a way that the squared bias dominates the variance. However, for the construction of confidence intervals one requires bandwidth which leads to variance domination, but still to consistent estimation. This is still possible theoretically, but the construction of an adequate data-driven bandwidth choice is a hard, yet unsolved problem. Concerning uniform confidence bands, no asymptotic results are available for supersmooth deconvolution. Of course, one can simply bootstrap a supremum-type statistic, but no theoretical justification is available here. Thus, there are still several open problems for statistical inference for supersmooth deconvolution.

### 3.4. Bandwidth selection and scale space theory

Use of the estimator $\hat{f}_n$ in (2) requires choosing the smoothing parameter $h$. There are several data-driven procedures for this purpose, see [12] for an overview. However, most of these methods tend to smooth the estimator too much (oversmoothing). In contrast, for the statistical inference in sections 3.2 and 3.3, we need a bandwidth $h$ for which the variance dominates the squared bias (i.e. the variance decreases at a slower rate than the bias). Actual undersmoothing

in practice (i.e. choosing the bandwidth too small, so that the variance strongly dominates the squared bias) results in poor performance of $\hat{f}_n$; thus, one should tend not to *oversmooth* in applications.

As indicated in section 2.2, the need to choose a bandwidth is often seen as a disadvantage when applying nonparametric estimators. However, it also allows for additional flexibility, and varying the bandwidth can also be used for statistical inference, e.g. for testing for the number of modes of $\hat{f}_n$. In fact, for most error densities the estimator (2) is also well defined if the normal kernel is used for $\|$ instead of a kernel with compactly supported Fourier transform (exceptions are normal errors). In this case, the number of modes of $\hat{f}_n$ is a monotonically decreasing, right-continuous function of $h$ [2, 41]. Therefore, there are certain minimal bandwidths (called critical bandwidths) for which the estimator just has $j$ and not yet $j + 1$ modes. This observation can be used to construct a statistical test for the hypothesis that $f$ has at most $j$ modes for some fixed $j$ [2]. Since the test is an extension of the classical Silverman test [41] to deconvolution problems, we call it 'Silverman test for deconvolution'. Investigating the estimator for different critical bandwidths is also a useful descriptive tool.

In [10], direct nonparametric kernel density estimation is discussed from the point of view of scale space theory from computer vision, and these ideas can also be extended to deconvolution density estimation. For (compact) intervals $I \subset \mathbb{R}$ and $H \subset (0, \infty)$, call

$$\{\hat{f}_n(x; h), x \in I, h \in H\},$$

the empirical scale space surface and $\{E \hat{f}_n(x; h), x \in I, h \in H\}$ the theoretical scale space surface. The monotonicity of the number of modes of $\hat{f}_n(x; h)$ is called causality in the scale space literature [35]. Intuitively, causality means that at coarser scales (i.e. for larger $h$), no additional features (i.e. modes) occur. For the deconvolution kernel density estimator, the scale space surface (if $I = \mathbb{R}$) also satisfies the causality property. Further, following [10] one can also show convergence in distribution of the scale space surface if the bandwidth is bounded away from zero.

## 4. Inverse regression models

### 4.1. Inverse regression: estimators and example

In an indirect regression model, independent observations $(z_k, Y_k), k = 1, \ldots, n$, satisfying

$$Y_k = (K\theta)(z_k) + \epsilon_k \tag{15}$$

are available. Here the $z_k$ are design points, the $\epsilon_k$ s are i.i.d. errors with $E\epsilon_k = 0$, $E\epsilon_k^2 = \sigma^2 < \infty$, and $K$ is a compact injective operator between $L_2$-spaces $L_2(\mu_1)$ and $L_2(\mu_2)$. The aim is not to estimate the regression function $m(z) = K\theta(z)$ of $Y_i$ given $z_i$, but rather the function $\theta$, therefore, we speak of indirect regression. We shall assume that the $z_i$ are fixed (i.e. non-random) design points. Since $K$ is assumed to be compact, we can consider its singular value decomposition: there exist orthonormal bases $(\phi_k)$ of $L_2(\mu_1)$ and $(\psi_k)$ of $L_2(\mu_2)$, and singular values $\lambda_k > 0$, such that $K\phi_k = \lambda_k \psi_k$ and $K^*\psi_k = \lambda_k \phi_k$. Here $K^*$ denotes the adjoint operator of $K$.

**Example 1** (The heat equation). Suppose that the state $u(x, t)$ of a system at a spatial point $x$ at time $t$ is governed by the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial^2 x}, \qquad 0 \leqslant x \leqslant 1, \qquad 0 \leqslant t \leqslant T,$$

with boundary and initial conditions

$$u(0, t) = u(1, t) = 0, \qquad u(x, 0) = \theta(x), \qquad 0 \leqslant t \leqslant T, \tag{16}$$

and suppose that we have noisy observations of the form $Y_k = u(z_k, T_0) + \epsilon_k$. The aim is to recover the initial condition $\theta$, which is related to $u(z, T_0)$ by

$$u(x, T_0) = \left(K_{T_0}\theta\right)(x) = \sum_{k=1}^{\infty} \exp(-k^2\pi^2 T_0)\langle\theta, \phi_k\rangle\phi_k(x),$$

where $\phi_k(x) = \sqrt{2}\sin(k\pi x), k \geqslant 1$, and $\langle\theta, \phi_k\rangle = \int_0^1 \theta(x)\phi_k(x)\,\mathrm{d}x$ (cf [32], p 267). Evidently, the operator $K_{T_0}$ is self-adjoint with eigenfunctions $\phi_k$ and eigenvalues $\lambda_k = \exp(-k^2\pi^2 T_0)$.

In the following we discuss construction of estimators for real-valued basis functions $\phi_j$ and $\psi_j$. We first construct estimators for the Fourier coefficients $b_k = \langle K\theta, \psi_k\rangle$, assuming that the design points are approximately uniform w.r.t. the measure $\mu_2$, as follows:

$$\hat{b}_{j,n} = \frac{1}{n}\sum_{k=1}^{n}\psi_j(z_k)Y_k.$$

A truncated Fourier series estimator of $\theta$ is then given by

$$\hat{\theta}_n(x; M) = \sum_{j=1}^{M}\frac{\hat{b}_{j,n}}{\lambda_j}\phi_j(x) = \sum_{k=1}^{n}Y_k w_{k,n}(x), \qquad w_{k,n}(x) = \frac{1}{n}\sum_{j=1}^{M}\frac{\psi_j(z_k)\phi_j(x)}{\lambda_j}, \qquad (17)$$

where $M = M(n)$ is a regularization parameter, namely the truncation parameter in the series estimator. The estimator $\hat{\theta}_n(x; M)$ is the spectral cut-off estimator as suggested, e.g., in [38]. Note that it makes explicit use of the basis functions $\phi_k$ and $\psi_k$. Other estimators, e.g. based on iterative regularization methods, which do not require knowledge of these functions, are discussed in [6].

In contrast to the kernel deconvolution density estimator, for the estimator (17) in indirect regression, no results on asymptotic normality and the construction of confidence intervals are available, since these are harder to derive for series estimators (such as (17)) than for kernel estimators (such as (2)).

Here we present some new results on asymptotic normality of $\hat{\theta}_n(x; M)$ in (17), specifically for the examples introduced above. To this end, we consider the decomposition of the estimation error into a stochastic and a deterministic part

$$\hat{\theta}_n(x; M) - \theta(x) = (\hat{\theta}_n(x; M) - E\hat{\theta}_n(x; M)) + (E\hat{\theta}_n(x; M) - \theta(x)).$$

In the following section we will derive confidence intervals for $E\hat{\theta}_n(x; M)$. These can be used to construct confidence intervals for the regression function $\theta$ itself after additional estimation of the bias.

### 4.2. Variance and asymptotic normality

In order to check asymptotic normality of the estimator (17), we use the following theorem which is lemma 3.1 in [17].

**Theorem 1.** *Suppose that in model (15), the weights $w_{k,n}(x)$ of the estimator $\hat{\theta}_n(x; M)$ in (17) satisfy*

$$\frac{\max_{1\leqslant k\leqslant n}|w_{k,n}(x)|}{\left(\sum_{j=1}^{n}w_{j,n}^2(x)\right)^{1/2}} \to 0. \qquad (18)$$

*Then $\left(\sigma^2\sum_{j=1}^{n}w_{j,n}^2(x)\right)^{-1/2}(\hat{\theta}_n(x; M) - E\hat{\theta}_n(x; M)) \to N(0, 1)$.*

The denominator in (18) is proportional to the variance of $\hat{\theta}_n(x; M)$. In order to derive lower bounds on the variance (required to check (18)), we shall use the following assumption.

**Assumption 1** (Orthogonal design)**.** Suppose that

$$\frac{1}{n}\sum_{k=1}^{n}\psi_i(z_k)\psi_j(z_k) = \delta_{i,j}, \qquad i,j = 1,\ldots,n. \tag{19}$$

Let us mention that approximate orthogonality of the design is sufficient for the following results. In the case of an orthogonal design, the variance can be easily evaluated as follows. Under assumption 1 we have for the variance that

$$\text{Var}\,\hat{\theta}_n(x;M) = \frac{\sigma^2}{n}\sum_{j=1}^{M}\frac{\phi_j^2(x)}{\lambda_j^2}.$$

**Example 1** (continued)**.** In example 1, an orthogonal design is given by the uniform design $z_k = (k-1/2)/n, k = 1,\ldots,n$. In this case we get for the variance that

$$\text{Var}\,\hat{\theta}_n(x;M) = \frac{2\sigma^2}{n}\sum_{j=1}^{M}\exp(2j^2\pi^2 T_0)\sin^2(j\pi x).$$

Note that at the boundary $x = 0$ or $x = 1$, the estimator (17) satisfies the boundary condition (16): $\hat{\theta}(x) = 0$ for $x = 0, 1$. Therefore, in the following we restrict ourselves to $x \in (0, 1)$.

To bound the order of the variance from below, we have to control the trigonometric function $\sin^2(x)$. To this end we prove that there exists $\kappa = \kappa(x)$ such that, if $\sin^2(j\pi x) < \kappa(x)$ for some index $j$, then $\sin^2((j+1)\pi x) \geqslant \kappa(x)$. In fact, choose $l = l(x) \in \mathbb{N}$ with $0 < 1/l < x < 1 - 1/l < 1$ and set $\kappa(x) := \sin^2(1/(2l)) > 0$. Suppose that $j$ is such that $(jx) \in \mathcal{B}_{1/(2l)}$, where

$$\mathcal{B}_{1/(2l)} := \left\{x : (x \bmod 1) \in \left[0, \frac{1}{2l}\right) \cup \left(1 - \frac{1}{2l}, 1\right]\right\},$$

so that $\sin^2(j\pi x) < \kappa(x)$. Then, since $1/l < x < 1 - 1/l$, we have that $(j+1)x \notin \mathcal{B}_{1/(2l)}$, and $\sin^2((j+1)\pi x) \geqslant \kappa(x)$, which proves our claim. Using this we estimate

$$\text{Var}\,\hat{\theta}_n(x;M) = \frac{2\sigma^2}{n}\sum_{j=1}^{M}\exp(2j^2\pi^2 T_0)\sin^2(j\pi x) \geqslant \frac{\sigma^2\kappa(x)}{n}\sum_{j=1}^{M-1}\exp(2j^2\pi^2 T_0),$$

since $\exp(2j^2\pi^2 T_0)$ is monotonically increasing in $j$. The order of the variance is now bounded from below as follows:

$$\sum_{j=1}^{M}\exp(2j^2\pi^2 T_0) \geqslant M\int_0^1 \exp(2M^2u^2\pi^2 T_0)\,\mathrm{d}u \sim \frac{M}{4M^2\pi^2 T_0}\exp(2M^2\pi^2 T_0),$$

where in the last step we used lemma 5 in [43]. Therefore, for $x \in (0, 1)$ and some $c(x) > 0$,

$$\text{Var}\,\hat{\theta}_n(x;M) \geqslant c(x)\frac{\exp(2(M-1)^2\pi^2 T_0)}{n(M-1)}.$$

On the other hand, the weights $w_{k,n}(x)$ can be bounded as follows:

$$|w_{k,n}(x)| \leqslant C\frac{\exp((M+1)^2\pi^2 T_0)}{n(M+1)},$$

by arguing as above. Therefore,

$$\frac{\max_{1\leqslant k\leqslant n}|w_{k,n}(x)|}{\left(\sum_{j=1}^{n}w_{j,n}^2(x)\right)^{1/2}} = O\left(\frac{\exp(4M\pi^2 T_0)}{(nM)^{1/2}}\right),$$

which tends to zero if we choose $M \to \infty$ such that $M \leqslant (\log n)/(8\pi^2 T_0)$. Thus, under these conditions, theorem 1 applies.

An asymptotic level-$\alpha$ confidence interval for $E\hat{\theta}_n(x; M)$ is now given as follows:

$$C_\alpha(x) = [\hat{\theta}_n(x; M) - \hat{\sigma} S_n q_{1-\alpha/2}, \hat{\theta}_n(x; M) + \hat{\sigma} S_n q_{1-\alpha/2}], \qquad x \in (0, 1), \tag{20}$$

where $\hat{\sigma}^2$ is an estimator of the error variance $\sigma^2$ (e.g. of difference type as in [13]), $S_n^2 = 2/n \sum_{j=1}^{M} \exp(2j^2\pi^2 T_0) \sin^2(j\pi x)$, and $q_{1-\alpha/2}$ the $(1 - \alpha/2)$-quantile of the standard normal distribution.

### 4.3. Estimation of the bias

In order to construct confidence intervals for $\theta$ itself, we have to deal with the bias, which for fixed design consists of an additional discretization bias. In fact, we can decompose it into a discretization bias and a truncation bias as follows:

$$E\hat{\theta}_n(x; M) - \theta(x) = \sum_{j=1}^{M} \frac{n^{-1} \sum_{i=1}^{n} \psi_j(z_i)(K\theta)(z_i) - b_j}{\lambda_j} \phi_j(x) + \sum_{j=M+1}^{\infty} a_j \phi_j(x),$$

where $b_j = \langle K\theta, \psi_j \rangle$ and $a_j = \langle \theta, \phi_j \rangle$. For estimating the truncation bias, one needs smoothness assumptions on $\theta$ (or equivalently a source condition on $\theta$). Typically, one assumes that the function $\theta$ is contained in a smoothness class of type

$$\Theta_\alpha = \left\{ f \in L_2(\mu_1) : f(x) = \sum_{j \geqslant 1} c_j \phi_j(x), c_j \leqslant C_1 j^{-(1+\alpha)} \right\}$$

for some $\alpha > 0$ (cf [38]). If in addition, the functions $\phi_j$ are uniformly bounded by some $C_2 > 0$, we can estimate the truncation bias by

$$\left| \sum_{j=M+1}^{\infty} a_j \phi_j(x) \right| \leqslant \frac{C_1 C_2 M^{-\alpha}}{\alpha}. \tag{21}$$

The discretization bias is more difficult to handle in general, and requires assumptions on the design, in particular that averaging w.r.t. the design points is close to integration w.r.t. $\mu_2$. In our examples one can establish a uniform estimate of the form

$$\left| n^{-1} \sum_{k=1}^{n} \psi_j(z_k)(K\theta)(z_k) - b_j \right| \leqslant An^{-1} \tag{22}$$

for some constant $A > 0$, independent of $j$. From (21) and (22) it follows that

$$|E\hat{\theta}_n(x; M) - \theta(x)| \leqslant \frac{A}{n} \sum_{j=1}^{M} \lambda_j^{-1} + \frac{C}{\alpha} M^{-\alpha}.$$

**Example 1** (continued). Arguing as in [17], pp 106–107, (22) is satisfied if $\theta \in \Theta_\alpha$ for $\alpha > 1/2$. In this case,

$$|E\hat{\theta}_n(x; M) - \theta(x)| \leqslant C \left( \frac{\exp((M + 1)^2\pi^2 T_0)}{n(M + 1)} + M^{-\alpha} \right).$$

Under the above assumption on $M$, the first term is negligible as compared to the standard deviation. However, for the second term $M^{-\alpha}$, a more specific choice of the smoothing parameter is required to achieve domination of the standard deviation. In fact, if we let $M - 1 = (\sqrt{2T_0}\pi)^{-1}(\log n)^{1/2}$, then we get for the bias a rate of $(\log n)^{-\alpha/2}$, whereas for the variance we get a rate of $(\log n)^{-1/4}$, which is slower since we assume that $\alpha > 1/2$. For such an $M$, the confidence interval (20) is also a valid asymptotic confidence interval for $\theta$.

*4.4. Deconvolution*

Suppose that $\theta, \varphi \in L^2[0, 1]$ are periodic functions, and consider the convolution model

$$Y_i = (K_\varphi \theta)(z_k) + \epsilon_i, \qquad (K_\varphi \theta)(z) = \varphi * \theta(z).$$

On the space $L^2_{\mathbb{C}}[0, 1]$ of complex-valued square-integrable functions, the operator $K$ is a normal operator for real-valued $\varphi \in L^2[0, 1]$ with eigenfunctions $\phi_j(x) = \exp(2\pi i j x)$, $j \in \mathbb{Z}$, and eigenvalues

$$\lambda_j = \int_0^1 \varphi(t) \, e^{-2\pi i j t} \, dt, \qquad j \in \mathbb{Z}.$$

Here, the basis functions are evidently complex-valued, and the estimator needs some minor modifications as follows:

$$\hat{\theta}_n(x; M) = \sum_{j=-M}^{M} \frac{\hat{b}_{j,n}}{\lambda_j} \phi_j(x), \qquad \text{where} \qquad \hat{b}_{j,n} = \frac{1}{n} \sum_{k=1}^{n} Y_k \overline{\phi_j}(z_k),$$

and $^-$ denotes complex conjugation. Since $\hat{b}_{j,n} = \overline{\hat{b}}_{-j,n}$ and $\lambda_j = \overline{\lambda}_{-j}$, the estimator $\hat{\theta}_n$ is real-valued. Its properties are developed along the lines for real-valued basis functions. For example, the orthogonality property (19) now reads

$$\frac{1}{n} \sum_{k=1}^{n} \phi_i(z_k) \overline{\phi}_j(z_k) = \delta_{i,j}, \qquad i, j = 1, \ldots, n,$$

and is satisfied for the uniform design. In this case the variance of $\hat{\theta}(x; M)$ is given by

$$\text{Var} \, \hat{\theta}_n(x; M) = \frac{\sigma^2}{n} \sum_{j=-M}^{M} \frac{|\phi_j(x)|^2}{|\lambda_j|^2} = \frac{\sigma^2}{n} \left( \frac{1}{|\lambda_0|^2} + 2 \sum_{j=1}^{M} \frac{1}{|\lambda_j|^2} \right),$$

since $|\phi_j(x)|^2 = 1$. Its order depends on the decay of the coefficients $\lambda_j$. Both for polynomial decay and for exponential decay one can prove asymptotic normality as in (18), and use this to construct confidence intervals for $E\hat{\theta}_n(x; M)$. After estimating the bias $E\hat{\theta}_n(x; M) - \theta(x)$, these yield confidence intervals for the regression function $\theta$ itself. The bias is dealt with similarly to as in example 1, and we hence omit the explicit derivation.

## 5. Case study—analyzing the deep structure of the Centaurus galaxy cluster

*5.1. Observational data*

In this section we show how to apply some of our methods in a case study. The problem we consider is estimation of the density of heliocentric escape velocities of galaxies in the Centaurus galaxy cluster. To this end we (re-)analyze a dataset of heliocentric velocities in or close to the Centaurus cluster [15], which is an ensemble of many individual galaxies. In particular, we aim to provide statistically significant evidence for the multimodality of the escape velocity distribution in the cluster. Based on the original dataset (we shall use a slightly extended version provided by R J Lucey) bimodality of the distribution of heliocentric velocities in the Centaurus galaxy cluster was discussed. From the velocity distribution one can estimate the deep (distance) structure of galaxies from a distribution of escape velocities. Such results are of fundamental importance for verifying cosmic evolution models by testing their predictions on the structure of the universe on different scales.

At our disposal are 274 measurements of heliocentric velocities in the Centaurus cluster ([15], where our data are a slightly revised version of table 11 provided by J R Lucey,

personal communication). The distance of the galaxies included in the data from the cluster center at $12^h 46^m 56^{deg}0.3, -40^{deg}56'49''$ (1950.0) is $\leqslant 6.9^{deg}$. In the following sections we show evidence for a bimodality of the distribution in the 'region of interest' given by $\approx 2000\,\text{km s}^{-1} \leqslant d \leqslant 5500\,\text{km s}^{-1}$. This selection of the 'region of interest' is suggested by the known escape velocity (and hence distance) of the galaxies in the Centaurus cluster. We will comment further on the selection of the 'region of interest' from the methodological point of view below. In section 5.2 we discuss statistical modeling of the data, which includes choosing the distribution of the noise. This is a difficult and important step in any data analysis, as statistical significance in general depends somewhat on the distribution of the noise, in particular on its approximate variance. Then, in section 5.3, we estimate the density and provide associated bootstrap confidence bands to assess the error level of the estimate. Finally, in section 5.4, we use the Silverman test for deconvolution to assess the modality structure of the density of escape velocities.

## 5.2. Statistical modeling of the data

The first important step in the application of statistical methodology in general is the formulation of a statistical model for the data. This consists of a model equation which links the quantity of interest, here $f$, to the observations, and in the definition of a model for the distribution of the random noise. Here, we choose the density deconvolution model, i.e. we assume that the observations $Y_i$ are given by $Y_i = X_i + \varepsilon_i$ (cf section 3), where $X_i$ are the true heliocentric escape velocities of the observed galaxies, which are random realizations from the density of interest $f$, and $\varepsilon_i$ are i.i.d. noise terms with density $\psi$. In practical applications, it is in general a difficult task to determine the noise density $\psi$ with high precision, since the noise is a complex combination of many sources of error, which includes calibration errors in the evaluation of the observed data and random measurement errors. If the density of the noise is unimodal, popular models for the density of the noise are Gaussian and double-exponential (Laplace) densities. Many theoretical results, such as our asymptotic confidence bands for deconvolution data, do not hold for supersmooth noise densities, as e.g. the Gaussian case. However, in [39] it is shown that misspecifying a supersmooth noise density as ordinary smooth is much less critical from the point of view of the risk of the estimator than vice versa. Hence, in practical applications where the density of the noise can only be approximately estimated but is expected to have a simple unimodal structure we suggest using a Laplace density with an estimated variance for the deconvolution methodology rather than a Gaussian density. Here we use a Laplace distribution with standard deviation $100\,\text{km s}^{-1}$ [36].

## 5.3. Estimating the density of heliocentric velocities and its associated confidence bands

The second step of our analysis consists in assessing the shape of the unknown density $f$. To this end, we compute an estimate and its associated confidence bands according to equations (2) and (13), respectively. Here, the bandwidth for the estimator $\hat{f}_n(x; h)$ is determined by a simple data-driven bandwidth selector which aims at minimizing the $L_\infty$-distance between the estimates $\hat{f}$ and $f$. This bandwidth estimator was shown to perform well in simulation studies w.r.t. the coverage probabilities and area of the resulting confidence bands [5]. The 'region of interest' for which the confidence bands hold is $(2000, 5500)\,\text{km s}^{-1}$. For some details on the necessity of the selection of such a region we refer to section 5.4. In practical applications with sample sizes of order several hundred, Bissantz *et al* [5] have shown that bootstrap confidence bands are significantly more reliable than their asymptotic counterparts. Hence, we determine bootstrap confidence bands, where $\approx 100$ bootstrap simulations turned

13
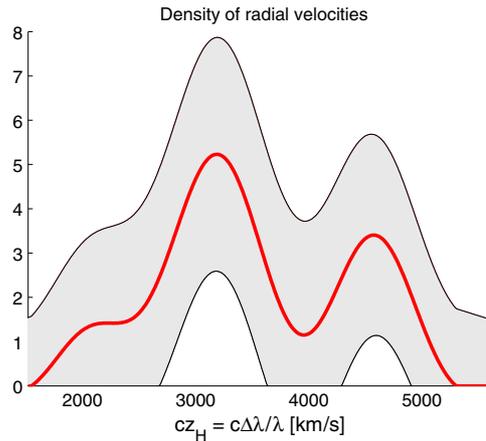
**Figure 1.** Estimator (thick line) and bootstrap confidence bands with a nominal coverage probability of 90% for the estimation of the density of heliocentric radial velocities $cz_H = c\Delta\lambda/\lambda$ (km s$^{-1}$) in the Centaurus cluster of galaxies from 274 observations.

out to be sufficient in simulation studies. Figure 1 shows the resulting estimator and confidence bands with a nominal coverage of 90%. From the additional information on the (uniform) precision of the estimator it is now possible to derive further properties of the galaxy cluster, such as its total mass and other properties of the dynamical structure. However, whereas this information is of large importance on its own (and in a number of cases the major interest in the study of a galaxy cluster), from the figure we can only conclude that there is some indication for bimodality, but unimodality cannot be excluded due to the large width of the confidence bands. A more sophisticated analysis of the modality structure of the data is thus required and will be provided by the Silverman test for deconvolution in the following section.

### 5.4. Statistical evidence for multimodality in the escape velocity distribution: applying Silverman's test for deconvolution

The graphical assessment of the density of escape velocities discussed in the preceding section provided some indications for multimodality, but did not allow us to exclude a unimodal density. Therefore, we use Silverman's test for deconvolution (cf section 3.4) in this final part of our case study to provide statistical significance of multimodality of the density. Indeed, Silverman's test allows us to test and, given the *p*-value is small enough, to reject the hypothesis of unimodality of $f$.

The 'region of interest' for Silverman's test was defined to be $v_h \in (2000, 5500)$ km s$^{-1}$, similarly as for the confidence bands. Such a restriction is not only suggested by (crude) previous knowledge of the escape velocities in the Centaurus cluster, but also necessary for proper performance of the statistical methods for deconvolution. Neither the confidence bands nor the Silverman test perform reasonably well if zero (or very close to zero) regions of the density are included in the region of interest. Similar problems exist for direct density estimation (cf [25]).

We now turn to the results of applying the Silverman test for deconvolution to the escape velocities. Figure 2 shows the critical densities for one, two and three modes. The first and second modes appear at approximately the same locations where the confidence bands indicated the existence of modes. This provides additional evidence for the existence of these modes (but still does not show their existence in a statistically significant way). In more detail,
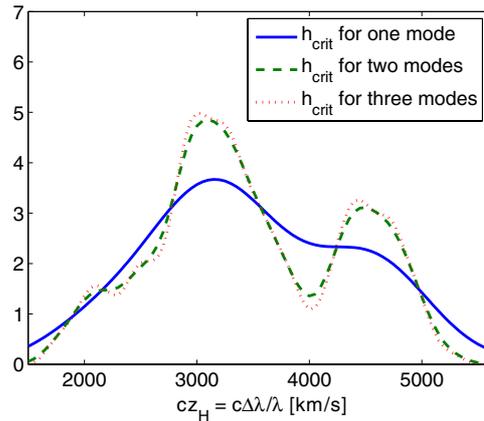
**Figure 2.** Estimates of the density of heliocentric escape velocities $cz_H = c\Delta\lambda/\lambda$ (km s$^{-1}$) in the Centaurus cluster of galaxies corresponding to the critical bandwidth for one mode (solid line), two modes (dashed line) and three modes (dotted lines). See the text for details.

the first and second modes, which are the major ones, are located at $\approx 3.1 \times 10^3$ km s$^{-1}$ and $4.5 \times 10^3$ km s$^{-1}$. This compares well with the results of Lucey *et al* [36], who obtained 3041 km s$^{-1}$ and 4570 km s$^{-1}$, respectively, from a maximum likelihood fit of a mixture of two normal densities to the data.

In order to evaluate statistical significance for the existence of more than one mode we performed the Silverman test for deconvolution to test the unimodality null hypothesis against a multimodal alternative, where we used the calibrated version of the test with a level of 5% (cf [2]). The calibration constant $\lambda_{5\%}$ was simulated based on a normal test density. It turns out that the null hypothesis is rejected, and we conclude that the distribution of velocities in the Centaurus cluster is (at least) bimodal. We remark that Lucey *et al* [36] already used several tests for normality (the Lilliefors test, Shapiro–Wilk test, and tests for skewness and kurtosis) to show non-normality of the distribution, which however, is not a suitable method to reject unimodality of the data.

In recent years large, in part automated observation systems have provided additional, more precise observations of escape velocities toward the Centaurus clusters. These datasets support the conclusions derived from the data used in our application (cf e.g. [42]). In our case study, we nevertheless used the dataset of Dickens *et al* [15] since it resembles well the data quality in many practical applications, for example our recent analysis of the young massive cluster luminosity function in the Antennae galaxies [1].

## 6. Discussion

Inverse problems are omnipresent in scientific data analysis. In this paper we discussed inverse problems from a statistical point of view, which apart from estimation purposes allows us to draw statistical inference, e.g. the construction of confidence intervals or hypothesis tests, about the unknown objects. Here we took the point of view of nonparametric curve estimation, but other inference techniques such as Bayesian methods can be used in this context as well [30]. While for nonparametric estimation, also in the context of inverse problems, the concept of adaptivity has been thoroughly investigated, for statistical inference such as confidence bands it has been shown recently that adaptivity sometimes cannot be achieved [21].

Our main examples were deconvolution problems. However, many other operators $K$ exist, which are of similar practical importance in applications from physics, image analysis, etc. These are technically more difficult to handle and an extension of our results is the aim of some of our current research. Moreover, sometimes, apart from the measurements error in $g$ there is an observational error in the operator $K$ as well. Estimation in such situations has been studied very recently [27], but further statistical inference methods still have to be developed.

## Acknowledgment

## References

[1] Anders P, Bissantz N, Boysen L, de Grijs R, Fritze V and Alvensleben U 2007 The young star cluster system in the Antennae: evidence for a turnover in the luminosity function *Mon. Not. R. Astron. Soc.* **377** 91–106

[2] Balabdaoui F, Bissantz K, Bissantz N and Holzmann H 2007 Testing the number of modes of a density of noisy observations *Working Paper* University of Göttingen

[3] Bickel P J and Freedman D A 1981 Some asymptotic theory for the bootstrap *Ann. Stat.* **9** 1196–217

[4] Bickel P J and Rosenblatt M 1973 On some global measures of the deviations of density function estimates *Ann. Stat.* **1** 1071–95

[5] Bissantz N, Dümbgen L, Holzmann H and Munk A 2007 Nonparametric confidence bands in deconvolution density estimation *J. R. Stat. Soc. Ser. B.* **69** 483–506

[6] Bissantz N, Hohage T, Munk A and Ruymgaart F 2007 Convergence rates of general regularization methods for statistical inverse problems *SIAM J. Numer. Anal.* **45** 2610–36

[7] Butucea C, Guta M and Artiles L 2007 Minimax and adaptive estimation of the Wigner function in quantum homodyne tomography with noisy data *Ann. Stat.* at press

[8] Butucea C and Tsybakov A B 2007 Sharp optimality for density deconvolution with dominating bias *Theory of Probability and its Applications* at press

[9] Cavalier L 2000 Efficient estimation of a density in a problem of tomography *Ann. Stat.* **28** 630–47

[10] Chaudhuri P and Marron J S 2000 Scale space view of curve estimation *Ann. Stat.* **28** 408–28

[11] Delaigle A, Hall P and Meister A 2007 On deconvolution with repeated measurements *Ann. Stat.* at press

[12] Delaigle A and Gijbels I 2004 Practical bandwidth selection in deconvolution kernel density estimation *Comput. Stat. Data Anal.* **45** 249–67

[13] Dette H, Munk A and Wagner T 1998 Estimating the variance in nonparametric regression—what is a reasonable choice? *J. R. Stat. Soc. Ser. B* **60** 751–64

[14] Devroye L and Lugosi G 2001 *Combinatorial Methods in Density Estimation* (Berlin: Springer)

[15] Dickens R J, Currie M J and Lucey R J 1986 The Centaurus cluster of galaxies: I. The data *Mon. Not. R. Astron. Soc.* **220** 679–711

[16] Engl H W, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Dordrecht: Kluwer)

[17] Eubank R L 1999 *Nonparametric Regression and Spline Smoothing* 2nd edn (New York: Dekker)

[18] Fan J 1991 On the optimal rates of convergence for nonparametric deconvolution problems *Ann. Stat.* **19** 1257–72

[19] Fan J 1991 Asymptotic normality for deconvolution kernel density estimators *Sankhya Ser.* A **53** 97–110

[20] Fan J and Gijbels I 1996 *Local Polynomial Modelling and Its Applications* (London: Chapman and Hall)

[21] Genovese C R and Wasserman L 2007 Adaptive confidence bands *Ann. Stat.* at press

[22] Groeneboom P and Jongbloed G 1995 Isotonic estimation and rates of convergence in Wicksell's problem *Ann. Stat.* **23** 1518–42

[23] Hall P 1992 Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density *Ann. Stat..* **20** 675–94

[24] Hall P and Horowitz J L 2005 Nonparametric methods for inference in the presence of instrumental variables *Ann. Stat.* **33** 2904–29

[25] Hall P and York M 2001 On the calibration of Silverman's test for multimodality *Stat. Sin.* **11** 515–36

[26] Härdle W and Mammen E 1993 Comparing nonparametric versus parametric regression fits *Ann. Stat.* **21** 1926–47
[27] Hoffman M and Reiss M 2008 Nonlinear estimation for linear inverse problems with error in the operator *Ann. Stat.* **36** 310–36
[28] Holzmann H and Boysen L 2006 Integrated square error asymptotics for supersmooth deconvolution *Scand. J. Stat.* **33** 849–60
[29] Hubble E 1929 A relation between distance and radial velocity among extra-galactic nebulae *Proc. Natl Acad. Sci.* **15** 168–73
[30] Kaipio J and Somersalo E 2005 *Statistical and Computational Inverse Problems* (Berlin: Springer)
[31] Kirsch A 1996 *An Introduction to the Mathematical Theory of Inverse Problems* (Berlin: Springer)
[32] Kress R 1999 *Linear Integral Equations* 2nd edn *Applied Mathematical Sciences* vol 82 (New York: Springer)
[33] Johnstone I M, Kerkyacharian G, Picard D and Raimondo M 2004 Wavelet deconvolution in a periodic setting *J. R. Stat. Soc. Ser.* B **66** 547–73
[34] Johnstone I M and Silverman B W 1990 Speed of estimation in positron emission tomography and related inverse problems *Ann. Stat.* **18** 251–80
[35] Lindeberg T 1993 *Scale-Space Theory in Computer Vision* vol 256 (Dordrecht: Kluwer)
[36] Lucey J R, Currie M J and Dickens R J 1986 The Centaurus cluster of galaxies: II. The bimodal velocity structure *Mon. Not. R. Astron. Soc.* **221** 453–72
[37] Politis D N and Romano J P 1999 Multivariate density estimation with general flat-top kernels of infinite order *J. Multivariate Anal.* **68** 1–25
[38] Mair B A and Ruymgaart F H 1996 Statistical inverse estimation in Hilbert scales *SIAM J. Appl. Math.* **56** 1424–44
[39] Meister A 2004 On the effect of misspecifying the error density in a deconvolution problem *Can. J. Stat.* **32** 439–49
[40] Neumann M H 1997 On the effect of estimating the error density in nonparametric deconvolution *J. Nonparametr. Stat.* **7** 307–30
[41] Silverman B W 1981 Using kernel density estimates to investigate multimodality *J. R. Stat. Soc. Ser.* B **43** 97–9
[42] Stein P, Jerjen H and Federspiel M 1997 Velocity structure of the dwarf galaxy population in the Centaurus cluster *Astron. Astrophys.* **327** 952–62
[43] van Es A J and Uh H-W 2004 Asymptotic normality for kernel type deconvolution density estimators: crossing the Cauchy boundary *J. Nonparametr. Stat.* **16** 261–77
[44] Wand M P and Jones M C 1995 *Kernel Smoothing* (London: Chapman and Hall)
[45] Wasserman L 2006 *All of Nonparametric Statistics* (Berlin: Springer)
[46] Xia Y 1998 Bias-corrected confidence bands in nonparametric regression *J. R. Stat. Soc. Ser.* B **60** 797–811