Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

SEMIPARAMETRIC MIXTURES FOR BACKGROUND DENSITY ESTIMATION IN PARTICLE PHYSICS

Relatore Prof. Tommaso Dorigo
Dipartimento di Fisica e Astronomia, Università di Padova

Correlatore Prof. Igor Volobouev
Dipartimento di Fisica e Astronomia, Texas Tech University

Laureanda: Sofia Guglielmini
Matricola N2005765

Anno Accademico 2021/2022

Abstract

The work in this thesis aims to develop a method to estimate the density of events in particle physics experiments, through a semiparametric mixture of a known parametric “signal” density and an unknown nonparametric “background” density. This method relies on an assumption of local smoothness of the background around the signal. The nonparametric component is estimated with a local orthogonal polynomial expansion (LOrPE), the level of overall smoothness of which is selected through a local version of least squares cross-validation. The estimate of the background is constructed iteratively through weighting of the original signal and background sample. The mixing proportion is chosen via maximum penalized local likelihood and the penalization term is a representation of the local complexity. This term is obtained with a novel estimator of the effective degrees of freedom, that relies on rejection sampling to localize the variability of the data around the interest region. Simulation studies show how the procedure operates, in its local version and in the global one, which is also presented.

Acknowledgements

I wish to thank my supervisor Professor Igor Volobouev for his patient guidance and help, and for providing me with the great opportunity to learn and explore new ideas during the writing of this thesis; and my co-supervisor Professor Tommaso Dorigo, for giving me a glimpse into the wonderful world of particle physics and allowing me to apply my passion for statistics to challenging and interesting problems.

I would like to extend my deepest gratitude to my mother and father, whose belief in me has supported and encouraged me during my university years and throughout my entire life.

Finally, a special thanks to my friends, for always lending a listening ear and being ever-present sources of comfort and joy.

Contents

Chapter 1

Introduction

Particle physics, or high-energy physics, heavily relies on statistical methods to obtain accurate answers from the large amounts of data collected during experiments, in order to test theories. When looking for a new physics process, the data set will contain “signal” data, representing the process of interest (if present), and “background” data, which are other processes that are not the signal, but may mimic it.

The work in this thesis aims to develop a method to estimate the density of events in particle physics experiments, through a semiparametric mixture of a known parametric signal density and an unknown nonparametric background density. This chapter describes the setting of the kind of experiments being considered and the scope of this thesis, while Chapter 2 details how the semiparametric mixture is modeled.

It is assumed that the interest lies in estimating the signal fraction and the density of the background around the signal region, so a method of localization is introduced into the procedures utilized. The background density is estimated using a version of LOrPE, a smoothing technique that uses local orthogonal polynomial expansions, introduced in Chapter 3. The level of smoothness introduced by LOrPE is determined through (localized) cross-validation and the estimator is applied on weighted data. The weights are chosen iteratively in such a way that the distribution of the weighted data represents the background distribution rather than the total distribution. The method does not need a “pure background” sample to be used as training data, instead it directly utilized the “signal-plus-background” data collected in the experiment. This is quite common in high-energy physics settings, since it can be difficult to collect control samples where the signal is absent, that also present the same kind of background as the one found together with the signal. In order to correctly carry out the estimation, an assumption of smoothness of the background around the signal region is made and it is introduced through a complexity term, the effective degrees of freedom, which penalizes the likelihood. This measure is defined in Chapter 5 and a novel method of localizing it is also described. The mixing proportion represents the fraction of signal events in the sample and it is estimated through penalized maximum likelihood (Chapter 4). Chapter 6 displays a series of simulation studies which show how the method can be applied in order to estimate the background density and choose the model with the correct signal fraction, both in the global version of the algorithm for smooth data and in its localized version for data that contains a non-signal peak. The code used in this work is written in Python and C++ and relies on the software *NPStat* for nonparametric statistical modeling (Volobouev, I. [44]).

1.1 Particle physics experiments

In high-energy physics experiments, the aim is to test a certain theory or model by seeing whether empirical evidence supports it. In practice, this usually means attempting to identify the presence of a new particle (search analysis), or to measure its characteristics (measurement analysis). The processes that generate these particles are often extremely rare, so large amounts of data need to be collected in order to observe them. In the analysis, the observed particles are produced in collision reactions called *events*. The *signal* is the process of interest in the study, while the *background* is composed by all processes whose final states mimic that of the signal, but are not in fact that process. The signal events are usually a much smaller number than the background events and they differ from the background with respect to one or more discriminating variables, like the mass of the particles. The larger the difference in the distribution function of this variable between the two types of events, the easier it is to identify the signal. However, the distinction is often not immediately clear. For this reason, and because of the quantity of data to be used, statistical analysis is crucial in order to carry out inference correctly and obtain reliable answers from the experiment.

The data may come from experiments at accelerators like the Large Hadron Collider (LHC), where particles collide at high energy, and the energy from these collisions converts into new states of matter, i.e. new particles. These particles are then identified and measured by detectors, which send signals to computers that collect and analyze the data. The events registered in this way are "*signal-plus-background*", as they include the particles we already expect to see, and might potentially include new particles as well.

Figure 1.1 shows histograms representing the mass distribution of the (expected) background and signal events, for the discovery of the Higgs boson (ATLAS Collaboration [6]).

The empirical data can be compared with Monte Carlo simulations of the process as suggested by the current theory, generated by complex algorithms that replicate collisions, scattering, decays and interactions between the particles and the detector, and between the particles themselves. Since they are simulated from the known theory, these events are "*background-only*". If the experimenter has an idea of what the signal is expected to look like, signal-plus-background events may also be simulated and compared with the observed ones.

If the comparison results in a difference between the empirical evidence and the simulations, we might have a signal of new physics. In practice it is a sign that there are physics phenomena that are not predicted by the Standard Model, the theory that is currently accepted as describing elementary particles and their interactions. The Standard Model by itself is not in fact a complete theory. It does not incorporate a quantum theory of gravitation and it fails to predict the large scale structure of the universe that is heavily affected by the mysterious constituents called, according to the current state of knowledge, "dark matter"

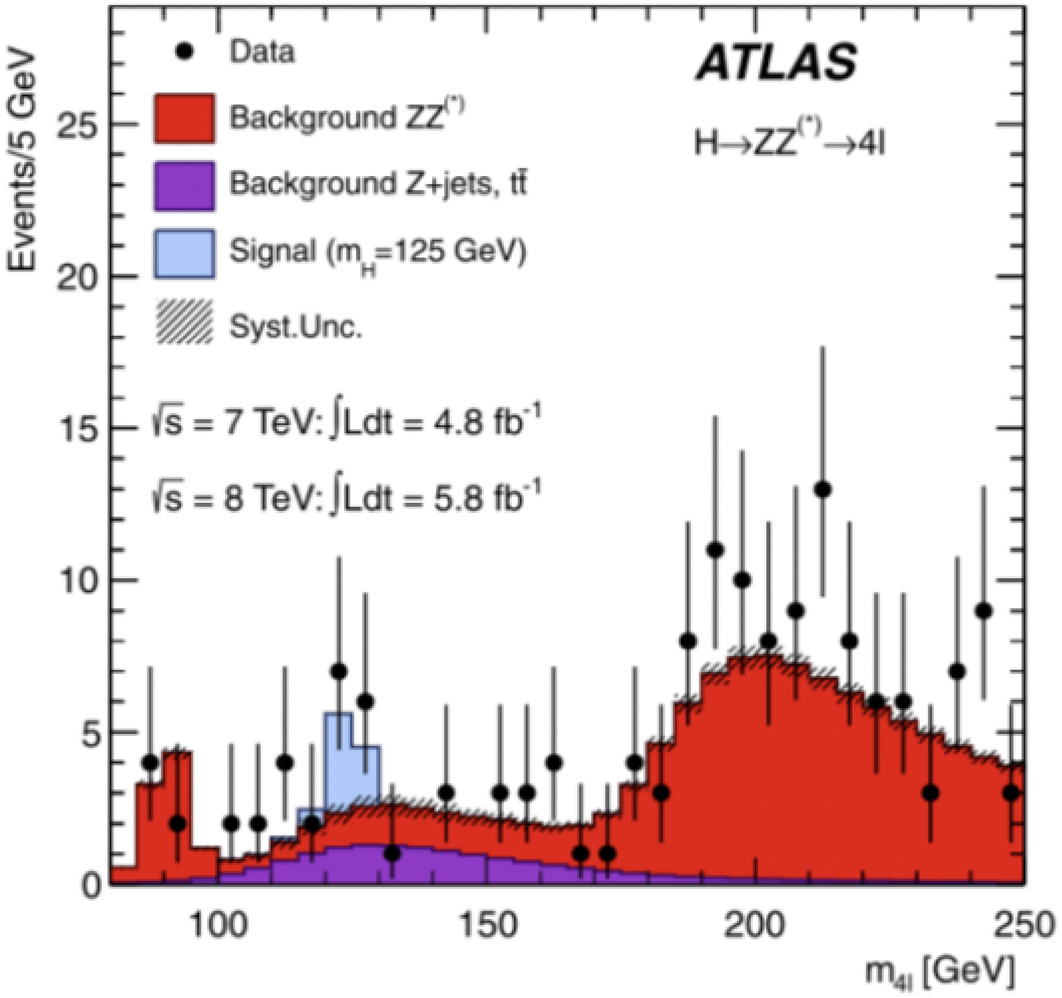

Figure 1.1: Distribution of the invariant mass for selected candidate processes, compared to the background and signal expectations.

and "dark energy".

In statistical terms, the background-only hypothesis is the null hypothesis, while the alternative is the signal-plus-background. If the null hypothesis is not rejected, this means that there is no strong evidence of new physics. However, Monte Carlo-driven models are subject to biases that can invalidate inference on the signal, particularly when looking for new processes. This is due to the fact that all background sources may not be modeled correctly, or the simulations may not completely replicate how processes are registered by the detector.

1.2 Objective

In high-energy physics analyses, the aim is, in practice, to search for a peak in the invariant mass distribution which has been reconstructed for the particles of interest. This search is based on the assumption that every selected event corresponds to a process that generates the signal. In this way, if present, the signal events will be distributed as a "Breit-Wigner" (or "Lorentz") distribution [9], which is a generalized form of the Cauchy distribution, introduced specifically to describe resonance particle production. This shape may be also convoluted with a Gaussian smearing to account for the finite experimental resolution of the mass reconstruction. The background process can take any form, but in most cases it will assume a rather smooth, featureless and monotonically decreasing shape. In our work, we do not consider the general case of the true shape that the reconstructed invariant mass distribution of a particle decay may take when the natural width of the particle be non-negligible with respect to the experimental resolution. In that general case the Lorenzian resonance form convolved with a Gaussian resolution term becomes what is known as a "Voigtian" distribution; little or nothing of the construction discussed in this work would change if we were to consider the general case.

Since the variables of interest are numbers of events, the observations may actually be modeled as realizations of a Poisson process (or a sum of two Poisson processes with different rates, one for the signal and one for the background). However, it is often the case that the overall rate of the process (i.e. the expected total number of events) is unknown and unpredictable, as it can depend not only on the physics of the phenomenon (which is governed by free parameters, such as the signal cross section, or the branching fraction of the particle decay in the studied final state, that are in fact usually the focus of the measurement), but also on the efficiency of the detector. If this happens, the problem is not one of estimation of the rates of these Poisson processes, but becomes one of density estimation: the processes are normalized and they are modelled as probability density functions, which is the case described in this work.

The analysis considered in this setting is such that there is a region of the discriminating variable (which is here taken as the invariant mass of a relevant combination of observed particles) where the signal is expected to be. The signal density is assumed, for sake of simplicity, to be a Gaussian centered in this region, with known variance. The *a priori* parametric specification of the signal shape is practically viable, as experiments are often carried out to test specific theoretical considerations. The background, on the other hand, is unknown, but we assume it contains a larger number of events at small values of the mass and a smaller number as the mass increases, and that it is smooth. This is in fact reasonable, due to the fact that the probability of collisions of given energy E between constituents of the hadrons (quarks and gluons) is in general a sharply decreasing function of E. The distribution of the data will thus look like a smooth decreasing function, potentially with a "bump" indicating a signal, as in Figure 1.2, which represents the CMS data relative to the Higgs Boson discovery [12].

The aim of this work is to develop a statistical method that evaluates the strength of the signal through the estimation of the density of the observed signal-plus-background events with a mixture of two densities. The known parametric density of the signal process is one component of the mixture, $s(x)$. The other component is the background density, $b(x)$ and it is estimated nonparametrically, since its shape can be much more complicated. For this reason, this work will refer to "semiparametric" mixtures.

The mixing proportion, α, represents the signal strength, or relative number of signal events.

$$p(x|\alpha) = \alpha s(x) + (1 - \alpha)b(x)$$

Monte Carlo simulations may contain bias and modeling errors, so often they do not provide the precision which is necessary for a two-component fit. For this reason, in high-energy physics studies, the estimation the background density is, in many cases, a process of trial and error [35], that eventually leads to a choice between many different parametric models. There is often no prior knowledge

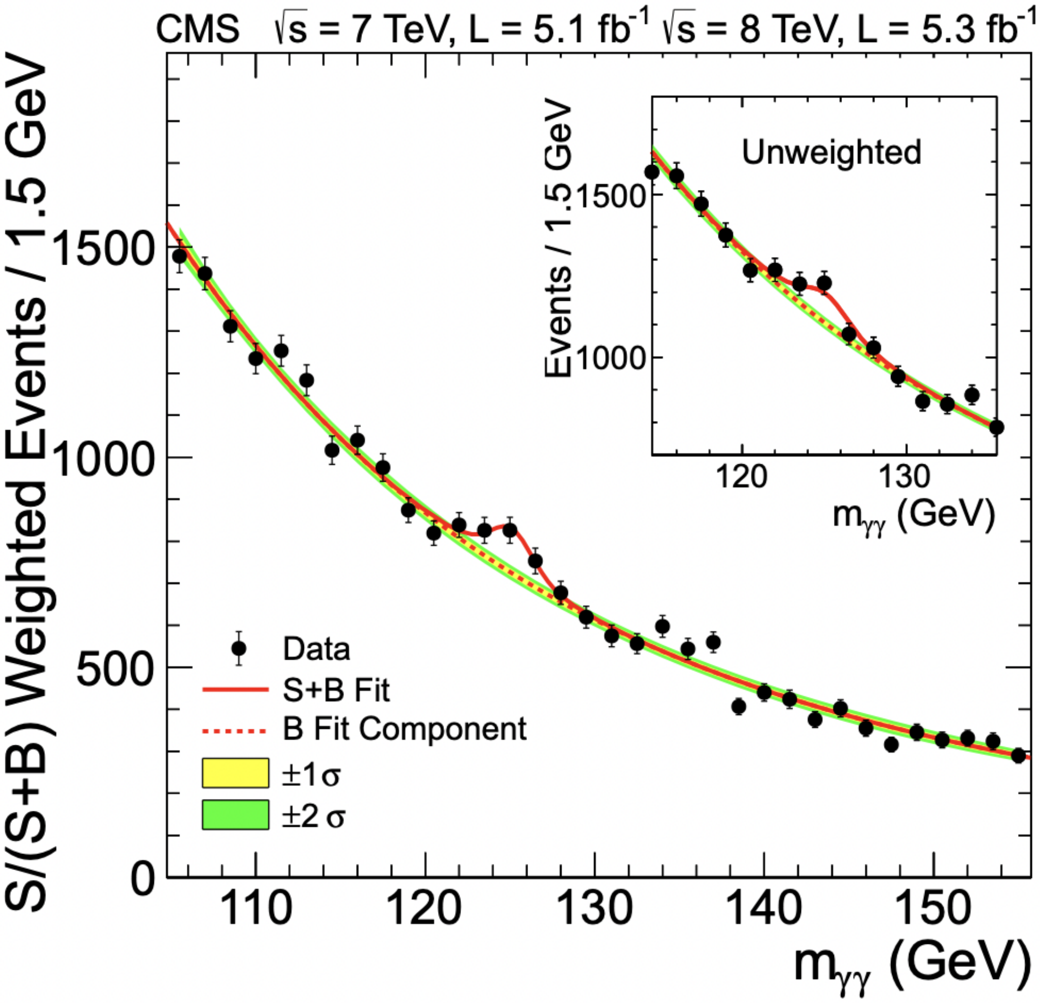

Figure 1.2: Mass distribution of the particles of interest in the Higgs Boson discovery, from CMS data. The dots represent the observed events, the lines represent the fitted background and signal, and the coloured bands represent the ±1 and ±2 standard deviation uncertainties in the background estimate.

about the specific parametric shape of the background distribution, and the model is thus chosen by the researcher. This selection procedure implicitly introduces subjective prior knowledge, which may vary from person to person and for this reason it is not optimal. By using a nonparametric estimator, the background is modelled in a very flexible way and the model is chosen on the basis of the data, excluding subjectivity as much as possible.

A typical characteristic in the setup we are considering is that the support is bounded [35], due to the researcher's choice to constrain the analysis to a certain interval of the spectrum. For univariate data, the boundary consists in the two endpoints of an interval on the discriminating variable. For this reason, the nonparametric estimator will need to account for the boundary problem with some kind of correction. Chapter 3.1 describes how LOrPE implements this correction.

In general, a semiparametric mixture with a completely arbitrary unknown component is not identifiable [32]. In practice, if further assumptions are not made, the signal fraction estimate (obtained, for example, through maximum likelihood) would be 0 and the best model would be fully nonparametric, since the nonparametric part is flexible and can also model the signal peak. This procedure leads to an estimate of the total density, but does not differentiate between signal and background densities and, crucially, does not estimate the signal fraction correctly. The following chapter contains some previous work on this kind of semiparametric mixtures from the literature and a description of the approach taken in this thesis.

Chapter 2

Semiparametric mixture estimation

A symmetry assumption on the nonparametric component has been used in the literature, like in Ma and Yao [27], and in Bordes et al. [8], who find “almost everywhere” identifiability under moment and symmetry conditions. In the case of particle physics events, this assumption is often violated, for example, as previously mentioned, the background density can be expected (even if not assumed or constrained) to be a decreasing function of the mass (Figure 1.2). Al Mohamad and Boumahdaf [3] also state that methods that use this constraint are of limited use when the proportion of the parametric part is either very high or very low: the latter is the case for the signal fraction in this kind of experiments. They thus suggest to incorporate prior linear constraints on the unknown component, but, as stated above, objective prior information on the background shape is usually not available.

Rolke and López [35] described a method for fitting this kind of semiparametric mixture specifically for high-energy physics data. Their method is only applicable assuming that a sample of pure background events is available (e.g. through a selection on the basis of a discriminating variable or via Monte Carlo simulations). Hall and Zhou [20] prove that in fact models of the type considered here are identifiable without further constraints when training data (in this context, background-only) is available. However, the work of this thesis uses data that is generated by a signal-plus-background process (if the signal is, in fact, present).

Robin et al. [34] considered an analogous mixture model and estimated the unknown density through weighted kernel density estimation, with weights which are proportional to the estimated probability that the data point was generated by that distribution. The weights are updated iteratively, for a fixed value of the mixing proportion parameter.

Xiang et al. [46] also estimated the mixing proportion by minimizing the Hellinger distance between the estimated and population densities. This is done iteratively by fixing the parameter and estimating the density, then viceversa.

Zhou and Yao [48] imposed a constraint of log-concavity of the unknown component. The estimation is carried out by semiparametric maximum likelihood through the EM algorithm.

For more work on density estimation through mixing parametric and non-parametric components, also see Olkin and Spiegelman [30], where the mixing proportion is an indicator of the goodness-of-fit of the parametric component and Schuster and Yakowitz [38], who apply semiparametric mixture estimation to flood frequency analysis.

2.1 Penalization and localization

In this thesis, the assumption imposed on the background in order to estimate it correctly is one of smoothness (at least in proximity of the signal region). The mixing proportion is thus chosen with penalized maximum likelihood, where the penalization is introduced through a measure of complexity, the effective degrees of freedom. Chapter 4 is dedicated to the selection of the value for the mixing parameter and Chapter 5 to the choice of the measure of complexity.

Furthermore, many measures and methods utilized in the following work can be localized. This means that more importance is given to a certain part of the sample space (on the mass variable), i.e. the region where the signal is expected to be. Indeed, the question of interest is about the signal process and,

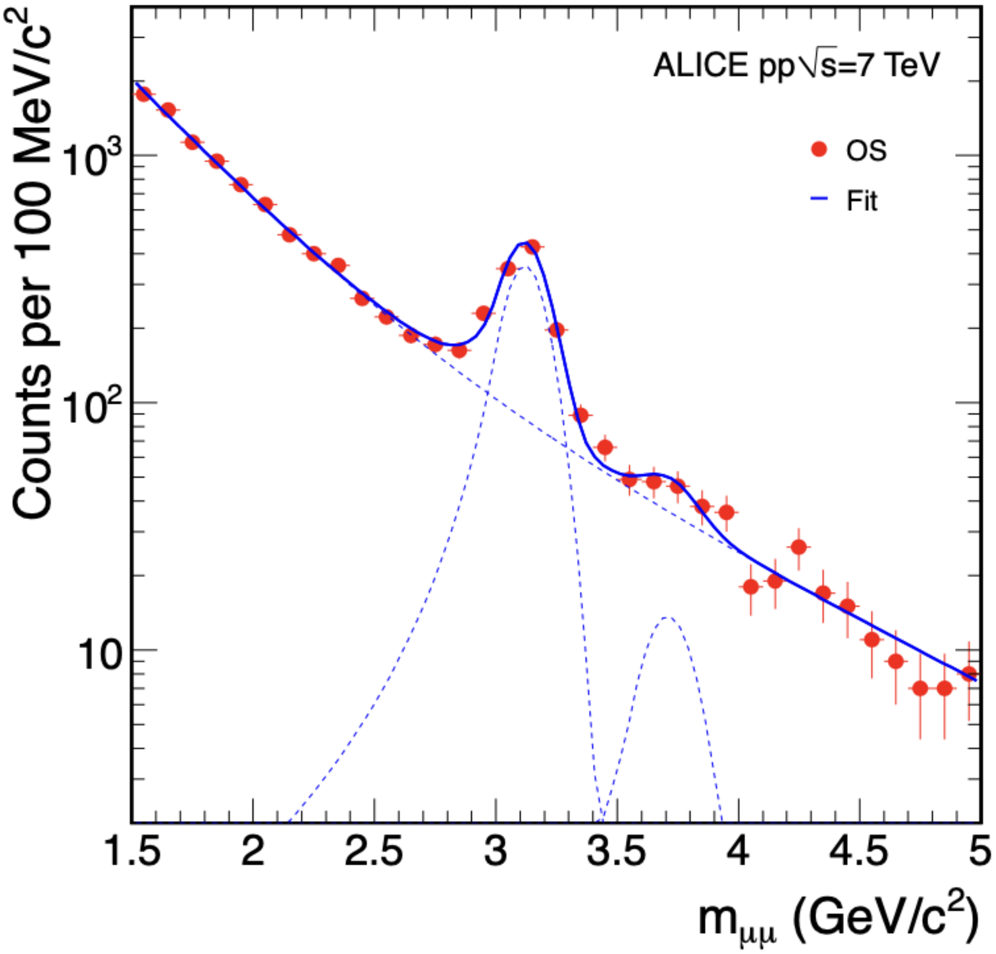

Figure 2.1: Invariant mass distribution for opposite-sign muon pairs, in the mass region between 1.5 and 5 GeV/c^2, with the result of the fit in blue.

for this reason, the distribution does not need to be estimated accurately on the whole sample space, but only around the signal region. Figure 2.1, from the ALICE experiment [4], shows the mass distribution for muon pairs, in a relatively large mass region (between 1.5 and 5 GeV/c^2). It can be seen that for large mass values (with respect to the signal region), the distribution starts to become unstable and unpredictable, due to phenomena which are usually unrelated to the process of interest. For this reason, it is a good idea to not include this data, or to give it less importance, when carrying out inference.

If we thus assume we are only interested in estimating the background around the signal, and the background is only assumed to be smooth in that region, a measure of local complexity based on the effective rank of the covariance matrix is sought (see paragraph 5.2.1). The level of overall smoothness of the nonparametric estimate of the background can also be selected in a localized way (see paragraph 3.2). Finally, in the localized case, the likelihood is also adapted in order to not consider points which are far from the signal region (see paragraph 4.1).

Chapter 3

Background density estimation

3.1 Local Orthogonal Polynomial Expansion

The estimation of the background density is carried out nonparametrically, as mentioned above. Nonparametric density estimation does not assume any parametric structure for the model, which is instead defined by a function that is not defined by a finite number of parameters. The specific estimator used in this thesis is a local orthogonal polynomial expansion of the empirical density function (or LOrPE). This method was developed by Amali Dassanayake et al. [5] and it consists in generating a truncated orthogonal polynomial series expansion for the empirical density function near each point where the density estimate is desired.

The empirical density function is defined as

$$\tilde{f}_{EDF}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i),$$

where $\delta(\cdot)$ is the Dirac delta function.

An orthogonal polynomial series is a family of polynomials such that all different polynomial in the sequence are orthogonal to each other under a certain inner product [1].

The estimate at a point x_{fit} is obtained by computing

$$\tilde{f}_{LOrPE}(x) = \sum_{k=0}^{M} c_k(x_{fit}, h) P_k\left(\frac{x - x_{fit}}{h}\right), \tag{3.1}$$

where M is the maximum degree of the polynomials, h is the bandwidth (both of these are tuning parameters that control the level of smoothness, their selection is carried out through a leave-one-out cross-validation method, as described in Chapter 3.2), P_k are polynomials, constructed such that they satisfy a normalization condition from which the coefficients c_k are then derived. The condition is that of orthogonality, under the inner product with respect to a kernel function $K(\cdot)$:

$$\frac{1}{h}\int_a^b P_j\left(\frac{x - x_{fit}}{h}\right) P_k\left(\frac{x - x_{fit}}{h}\right) K\left(\frac{x - x_{fit}}{h}\right) dx = \delta_{jk}, \tag{3.2}$$

where δ_{jk} is the Kronecker delta. In this case the kernel function is a symmetric Beta function with power parameter 4. This density is defined as:

$$f(x) = c(1 - x^2)^p,$$

where p is the power parameter and c is a normalization constant. Sufficiently far away from the support boundaries, the normalization condition (equation 3.2) generates orthonormal Gegenbauer polynomials. A Gaussian kernel may also be used instead of the symmetric Beta function and in this case Hermite polynomials would be generated [5].

Negative density estimates are dealt with either by taking $\max\{0, \tilde{f}_{LOrPE}(x_{fit})\}$ as the proposed density estimated (after renormalization), or by using a correction described by Glad et al. [16].

It is also possible to generalize equation 3.1 through the introduction of a "taper function" $t(k)$ instead of a sharp cut-off at M terms, so that higher-order polynomials are gradually suppressed instead of not being included:

$$\tilde{f}_{LOrPE}(x) = \sum_{k=0}^{\infty} t(k) c_k(x_{fit}, h) P_k \left(\frac{x - x_{fit}}{h} \right).$$

This is used in the case where the optimal M is not an integer, so the taper function is defined as:

$$t(k) = \begin{cases} 1, & \text{if } k \leq m \\ \sqrt{M - m}, & \text{if } k = m + 1 \\ 0, & \text{if } k \geq m + 2, \end{cases}$$

where m is the largest integer such that $m \leq M$ ($m = \lfloor M \rfloor$).

LOrPE may be interpreted as a linear combination of kernel density estimators, with different kernels; or as a localized version of orthogonal series density estimation, where the polynomials depend on the point of estimation (x_{fit}). The problem of the boundary bias is approached by matching local moments of the polynomials to sample values, using polynomial approximations. Note that, for maximum degree of the polynomial equal to 0, LOrPE is equivalent to kernel density estimation with a boundary kernel correction (also described by Scott [39]).

Through simulation studies, Amali Dassanayake et al. [5] observed that, at lower sample sizes, LOrPE has almost always lower Mean Square Integrated Error (MISE) than kernel density estimation; also for larger sample sizes it yields consistently minimum MISE for some distributions (it is still competitive for others). Finally, since it allows for a taper function, it is more flexible than kernel density estimation and particularly useful with sharp truncation at boundaries.

3.2 Selection of the complexity level

The optimal level of smoothing (controlled for the LOrPE estimator by the bandwidth and the maximum polynomial degree) is not attainable in practice, as it depends on the unknown density that we want to estimate. It can be chosen with rules of thumb, plug-in methods, or with automatic selection methods such as cross-validation, which is employed here. A model which is too smooth is less affected by statistical fluctuations, but risks ignoring important structures in the shape of the generating process, thus introducing a large amount of *bias*. On the other hand, a model that is too complex risks *overfitting*, or having too large variance. It will follow fluctuations closely and not perform well in prediction problems. Cross-validation is a data-based method which reutilizes the data in order to obtain a compromise between the bias and the variance of the model, ensuring that the model is not too smooth nor too complex. It consists in dividing the dataset into a certain number k of parts, or *folds*, estimating the model on $k-1$ folds, then evaluating it using the k-*th* fold. The fold which is used as validation set is rotated so that each data point acts both as estimation and validation. This method is used to compute a certain criterion for each level of complexity. The complexity that optimizes it is then selected. A global version of the Least Squares Cross-Validation criterion and a localized one are explored in this chapter. These criteria are minimized in order to obtain optimal values of the bandwidth and the maximum degree. In practice, they are computed on a grid of degree and bandwidth values. For each degree value a different range of bandwidths is explored, since a higher polynomial degree will generally require a larger bandwidth. The factors with which this range changes for each degree value are derived from theoretical considerations using the Asymptotic Mean Integrated Squared Error (AMISE) of a plug-in Gaussian distribution (Turlach et al. [43]). The actual extremes of the range of bandwidth values scanned for each maximum degree level are reported in Table 3.1. However, these are only starting points and their optimization may benefit the speed of the algorithm.

3.2.1 Least Squares Cross-Validation

The Least Squares Cross-Validation (LSCV) criterion is based on the Integrated Squared Error (ISE) [39], which we want to minimize with respect to h (a

Degree	**Minimum bandwidth**	**Maximum bandwidth**
0	0.1500	10.0000
0.25	0.1778	11.8555
0.5	0.2054	13.6930
0.75	0.2325	15.4995
1	0.2590	17.2672
1.25	0.2849	18.9917
1.5	0.3101	20.6713
1.75	0.3346	22.3052
2	0.3584	23.8941
2.25	0.3816	25.4391
2.5	0.4041	26.9417
2.75	0.4261	28.4034
3	0.4474	29.8263
3.25	0.4682	31.2119
3.5	0.4884	32.5622
3.75	0.5082	33.8789
4	0.5275	35.1636

Table 3.1: Lower and upper extremes of the bandwidth ranges scanned by the algorithm, for each maximum polynomial degree.

smoothing parameter):

$$ISE(h) = \int [\hat{f}(x) - f(x)]^2 dx = \int \hat{f}(x)^2 dx + \int f(x)^2 dx - 2 \int \hat{f}(x) f(x) dx.$$

Rudemo [37] used leave-one-out cross-validation (a number of folds equal to the number of data points) to estimate this criterion for the kernel density estimation of a histogram fit. He notes that the first term can be computed from the estimate, while the second one is not affected by the smoothing and the third one can be estimated with the aforementioned cross-validation.

$$\int \hat{f}(x) f(x) dx = E[\hat{f}(X)] \approx \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{-i}(x_i)$$

The LSCV criterion is then

$$LSCV(h) = \int \hat{f}(x)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(x_i).$$

The integral in the first term is computed through numerical integration, with the Gauss-Legendre quadrature method. This method is a way of approximating an integral with a weighted sum of values of the function of interest, evaluated at the points corresponding to the roots of the *n-th* Legendre polynomial, where n is the number of sample points (terms of the sum).

3.2.2 Weighted Least Squares Cross-Validation

The LSCV criterion may be weighted, in order to give more importance in the choice of bandwidth to the points around the region of interest.

The weight function $w(x)$ which is used in this thesis is a convolution of a standard normal and a uniform distribution (Figure 3.1). A scale and location transformation may be applied to this function, the parameters of which should be such that the signal region is included in the "flat" part of the function and coordinates corresponding to any potential non-signal structures in the tail are given a very small or approximately null weight.

A location-scale family [31] is a parametric family of distributions with scale parameter $\sigma \geq 0$ and location parameter $\mu \in \mathbb{R}$ for a univariate observation y with density

$$p(y; \mu, \sigma) = \frac{1}{\sigma} p_0 \left(\frac{y - \mu}{\sigma} \right),$$

where in this case $p_0(\cdot)$ is the original Uniform-normal convolution.

The simulation studies in Chapter 6 show how the smaller the region with non-null weights is, the smoother the overall model is, as more fluctuations are excluded in the complexity choice.

A Gaussian function was initially considered for this purpose, for analogy with the weights chosen for the localization of the effective degrees of freedom, in paragraph 5.2.1. However, simulations and logical considerations suggest that it is more appropriate to give full importance to the left-hand side part of the distribution, which is assumed to contain a high density of events and not as many unpredictable processes as the tail of the distribution.

Introducing the weight function $w(x)$ for the localization of the criterion, a

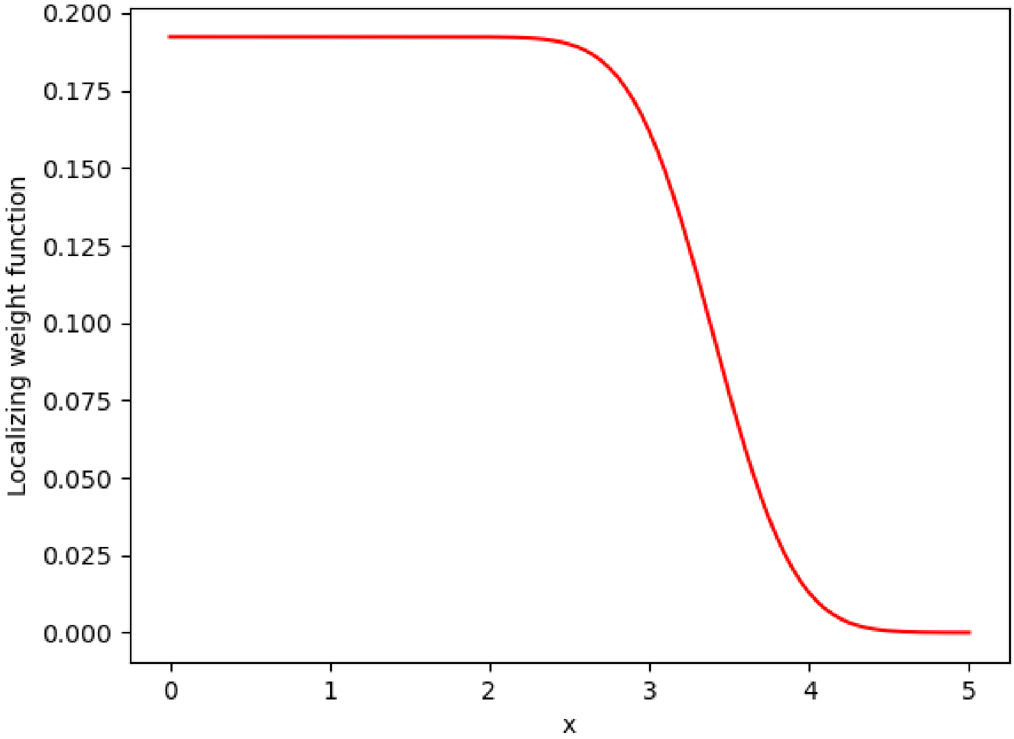

Figure 3.1: Convolution of a Uniform between -6.5 and 6.5, and a standard normal, with scale transformation parameter equal to 0.4 and location transformation with parameter 0.8.

Weighted ISE is defined:

$$\begin{aligned} WISE(h) &= \int [\hat{f}(x) - f(x)]^2 w(x)dx \\ &= \int \hat{f}(x)^2 w(x)dx + \int f(x)^2 w(x)dx - 2\int \hat{f}(x)f(x)w(x)dx \end{aligned}$$

and thus a Weighted LSCV:

$$WLSCV(h) = \int \hat{f}(x)^2 w(x)dx - \frac{2}{n}\sum_{i=1}^{n} w_i \hat{f}_{-i}(x_i),$$

where $w_i = w(x_i)$.

If any kind of weighting of the sample is to be used as well with weights w_i^s (e.g. the sampling function weights for the localization of the effective degrees of freedom or the weights introduced for the solution of the Fredholm equation in Chapter 3.3), the WLSCV criterion is computed in the following way:

$$WLSCV(h) = \int \hat{f}(x)^2 w(x)dx - \frac{2}{\sum w_i^s}\sum_{i=1}^{n} w_i^s w_i \hat{f}_{-i}(x_i),$$

where $w_i^s = w^s(x_i)$; the leave-one-out estimate $\sum_{i=1}^{n} \hat{f}_{-i}(x_i)$ and the regular estimate $\hat{f}(x)$ are computed on the weighted sample.

Loader [26] similarly introduced a local cross-validation criterion based on the application of weights to the Mean Squared Error for local regression.

3.3 Iterative algorithm for the point weights

The LOrPE smoother is used to nonparametrically estimate the background density from the data. Since the data we have is signal-plus-background and we do not have a pure background sample, we do not have the empirical density function of the background. However, the latter could be approximated by applying weights to the original sample, which are equal to the ratio of the background density and the total (mixture) density evaluated at the sample points. The empirical density function is then computed on the weighted sample:

$$EDF_b(x|\alpha) = \frac{c}{n}\sum_{i=1}^{n}\frac{b(x)}{\alpha s(x) + (1-\alpha)b(x)}\delta(x - x_i),$$

where $\delta(\cdot)$ is the Dirac delta and c is a normalization constant.

Of course, the background density is not known (nor is the total density), so we will substitute it with an estimate:

$$E\hat{D}F_b(x|\alpha) = \frac{c}{n}\sum_{i=1}^{n}\frac{\hat{b}(x)}{\alpha s(x) + (1-\alpha)\hat{b}(x)}\delta(x - x_i).$$

The mixing proportion α is also unknown. It will be substituted with a plug-in value, and the whole procedure will be repeated for a grid of possible values of α. The best model between these will be chosen with the penalized likelihood method (Chapter 4).

The LOrPE smoother, which will be used to estimate the background, can be represented as an operator applied on the empirical density function, in the general way:

$$\hat{f}_{LOrPE} = \mathcal{A}_{M,h}(EDF(x)),$$

where M and h are the smoothing parameters, respectively the maximum degree of the polynomials and the bandwidth.

We will thus apply LOrPE to the estimated empirical density function of the background, an obtain, in turn, an estimate of the background.

$$\begin{aligned}\hat{b}(x|\alpha) &= \mathcal{A}_{M,h}(E\hat{D}F_b(x|\alpha))\\ \hat{b}(x|\alpha) &= \mathcal{A}_{M,h}\left(\frac{c}{n}\sum_{i=1}^{n}\frac{\hat{b}(x)}{\alpha s(x) + (1-\alpha)\hat{b}(x)}\delta(x - x_i)\right). \qquad (3.3)\end{aligned}$$

Equation 3.3 is a nonlinear kind of Fredholm equation. Fredholm equations are generally such that an operator is applied to a function, and another function is obtained. In this case operator is quite complicated, as it includes smoothing through LOrPE, discretization and sample weights. Nevertheless, the equation can be solved iteratively, since the solution function and the function in the operator are the same (note that $\hat{b}$ is present on both sides of the equation).

The iterative procedure allows one to obtain, at convergence, a set of weights. Applying LOrPE on the sample, modulated by the weights obtained through this process, yields an estimate of the background.

The iterative algorithm consists in starting from weights chosen more or less arbitrarily. In simulations, "oracle" weights based on the true generating densities and the true α can be used as a starting point, to make convergence faster. These weights are applied to the sample, cross-validation is run on the weighted sample as described in paragraph 3.2, leading to a choice of maximum degree and bandwidth and an estimate of the background is obtained through LOrPE. This estimate is then substituted into the weight formula, and the process is repeated with the new weights. Convergence is reached at iteration k when the following condition is met:

$$\left(\frac{1}{\sum v_i} \sum_{i=1}^{n} v_i \left| 2 \frac{\omega_{i,k} - \omega_{i,k-1}}{|\omega_{i,k}| + |\omega_{i,k-1}| + 1} \right|^q \right)^{1/q} < \epsilon$$

where v_i are localizing weights (they may be, for example, the same weights used for the cross-validation localization); $q > 0$ is a parameter which is chosen *a priori* (in this work, the value 1 is used); ϵ is a tolerance parameter (10^{-6} in this work); $\omega_{i,k}$ is the weight to be solved for in the iterative process, corresponding to the *i-th* coordinate, in the *k-th* iteration:

$$\omega_{i,k} = \frac{\hat{b}^k(x_i)}{\alpha s(x_i) + (1-\alpha)\hat{b}^k(x_i)},$$

where $\hat{b}^k(x)$ is the density estimate of the background obtained in the *k-th* iteration.

The convergence criterion is also localized, as the aim is not necessarily to estimate the background accurately outside of the interest region and thus it is not expected for the algorithm to converge on the entire support. In general, the convergence of the iterations is not guaranteed. Iterations always converge for the examples used in this thesis, but theoretical proof of the convergence could be a topic of future research.

Chapter 4

Choice of the mixing proportion

Once the background is estimated for a grid of values of the mixing proportion, since the signal is assumed to be known, the total density from the semiparametric mixture $p(x|\alpha)$ can be estimated for each α, obtaining $\hat{p}(x|\alpha)$ and, from it, the empirical likelihood function

$$L(\alpha) = \prod_{i=1}^{n} \hat{p}_i(\alpha),$$

where $\hat{p}_i(\alpha) = \hat{p}(x_i|\alpha)$ is the density estimate at x_i, $i = 1, \ldots, n$.

The log-likelihood function is thus

$$l(\alpha) = \sum_{i=1}^{n} \log \hat{p}_i(\alpha).$$

As mentioned in Chapter 1.2, the estimation of α cannot be carried out correctly through the maximization of the likelihood, as a model of this kind is not identifiable, and a possible solution is to introduce a penalty term into the likelihood, which will account for the complexity of the model.

As stated by Cole et al. [13], the penalized log-likelihood is the log-likelihood with a penalty subtracted from it that will pull or shrink the final estimates away from the maximum likelihood estimates, toward values that have some grounding in information outside of the likelihood as good guesses for the parameters. The idea behind penalizing the complexity here is based on an assumption of smoothness of the background in the proximity of the signal region, and on the fact that, if the value of α is close to the true signal fraction, the estimate of the background will be smooth, because it does not need to compensate a wrong value of α with a peak (if the assumed α is smaller than the true one; Figure 6.3, left) or with a dip (if the assumed α is larger than the true one; Figure 6.3, right).

The penalized log-likelihood, to be maximized in order to get the final estimate of the signal fraction (and thus the density estimate) is

$$l_d(\alpha) = \sum_{i=1}^{n} \log \hat{p}_i(\alpha) - \gamma \cdot d,$$

where d is a measure of complexity based on the effective rank of the covariance matrix of the bootstrap density, which will be shown in Chapter 5, and γ is a parameter that defines how much the complexity should influence the final estimate. In this thesis, the penalization parameter γ is chosen empirically and the simulation studies show that it is possible to find a value that returns a correct estimate of α.

Further research may include the search of a value for γ based on theoretical considerations.

Good and Gaskins [18] introduced the idea of penalizing the log-likelihood for roughness in nonparametric density estimation. Their approach is proposed with a Bayesian interpretation, which introduces information on how much better one density estimate is than another one. They also describe some conditions under which the estimate is consistent. Akaike [2] formulated the Akaike Information Criterion, which uses a complexity-based penalization of the likelihood for model selection.

The choice of α here is treated as a model selection problem, however Silverman [41] describes maximum penalized likelihood estimation in a formal way, Nong et al. [29] carry out hypothesis tests with a penalized likelihood approach, while Good and Gaskins [19] specifically use likelihood penalization for “bump-hunting”. Chen [11] applied the penalized likelihood method to finite mixture models and was able to obtain formal likelihood ratio tests.

4.1 Local likelihood

If localized cross-validation is used, the fit will not be required to be accurate on the entire sample space, thus likelihood also needs to be localized. Localizing weights may thus applied to the likelihood function in order to obtain a kind of composite likelihood function [25]:

$$cL(\alpha) = \prod_{i=1}^{n} (\hat{p}_i(\alpha))^{w_i},$$

the weights w_i are, again, coordinate-based and they may be the same weights used for the localization of the cross-validation criterion, or the convergence of the Fredholm equation.

The log-likelihood function is

$$cl(\alpha) = \sum_{i=1}^{n} w_i \log \hat{p}_i(\alpha).$$

The penalization is then applied in the same way as the global version:

$$cl_d(\alpha) = \sum_{i=1}^{n} w_i \log \hat{p}_i(\alpha) - \gamma \cdot d,$$

where in this case d is a localized measure of complexity, as described in paragraph 5.2.1.

Loader [26] and Tibshirani and Hastie [42] provided detailed descriptions of local likelihood, also for density estimation. Wang et al. [45] proved consistency, asymptotic normality and other important properties of the maximum weighted likelihood estimators.

Chapter 5

Effective degrees of freedom

In parametric statistics, the complexity of a model is represented by the number of estimated parameters. In a nonparametric or semiparametric setting, this concept is not as obvious, so the number of parameters (which is not defined, of course) is usually replaced by that of "effective degrees of freedom". Effective degrees of freedom for a linear smoother S_h, such that $\hat{y} = S_h y$ (where h is a smoothing parameter) are defined by Buja et al. [10] as $tr(S_h)$, i.e. the trace of the smoothing matrix (or *hat* matrix). Other definitions from the same authors are $tr(S_h S_h^T)$ and $n - tr(2S_h - S_h S_h^T)$. These formulations are all motivated by analogies with the linear regression model and can be extended to nonlinear smoothers, although they may, in this case, depend on the distribution of the data [21]. In general, the number of effective degrees of freedom is a quantity that is related to the complexity or dimensionality of the system at hand, so in this way also to the degree of smoothing given by the operator.

This concept has been studied extensively for nonparametric regression, not as much for nonparametric density estimation.

McCloud and Parmeter [28] state that the trace of the usual smoothing matrix is not ideal to determine effective degrees of freedom in density estimation and they suggest transforming the density estimator (in their case, a kernel density estimator) to resemble a regression estimator in order to obtain a new kind of hat matrix, the trace of which is used to estimate the effective degrees of freedom.

Barron and Cover [7] used the concept of "description length" in order to carry out consistent density estimation.

Gao and Jojic [15] computed effective degrees of freedom for neural networks through a Monte Carlo method and used them in model selection problems.

Ye [47] introduced "generalized degrees of freedom", which are applicable to complex modeling procedures, like the ones utilized in data mining, and can be used as measures of complexity. The definition of this quantity is based on the sum of the average sensitivities of the fitted values to perturbation in the observed value. Hauenstein and Dormann [22] reviewed the use of generalized degrees of freedom, tested it and compared it with cross-validation.

During the development of this thesis, several methods of estimating the level of smoothing for the density estimation of the background were considered and tested.

5.1 Binned data

In an early stage of the studies reported here, data binned into a histogram was considered. Since smoothing tends to lower variance, an estimator of complexity was first developed as

$$df_{ratio} = \sum_{i=1}^{K} \frac{\sigma_{a,i}^2}{\sigma_{b,i}^2},$$

where K is the number of bins, $\sigma_{a,i}^2$ is the variance in bin i after smoothing, $\sigma_{b,i}^2$ is the variance in bin i before smoothing. The reasoning behind this estimator is that the more smoothing there is, the lower the complexity is and the lower the ratio of the variance after and before smoothing is, so this quantity should diminish as the amount of smoothing increases.

Since the sample size of the simulated data was itself generated from a Poisson distribution, the number of observations in each bin is also distributed as a Poisson density, so the variance before smoothing corresponds to the number of entries in that bin. The variance after smoothing is calculated in the following way:

$$\hat{y} = S_h y$$
$$Var(\hat{y}) = S_h Var(y) S_h^T.$$

In the case of simulations, the number of entries in the bins may be the expected (according to the distribution function of the data-generating process) or observed one: the quantities obtained in these two ways were both considered and analyzed. They are analogous to, respectively, the Pearson and Neyman Chi-squared measures for goodness of fit analyses, often used in high-energy physics [24]. For this reason, we will refer to these measures with df_P and df_N, using the initials of Pearson and Neyman; and define them as follows:

$$df_P = \sum_{i=1}^{K} \frac{[S_h N S_h^T]_{ii}}{N_{ii}},$$

$$df_N = \sum_{i=1}^{K} \frac{[S_h \hat{N} S_h^T]_{ii}}{\hat{N}_{ii}},$$

where S_h is the smoothing matrix, N is a diagonal matrix, with the expected number of entries in each bin as diagonal elements, $\hat{N}$ is an analogous matrix with observed values instead of expected ones. Note that, for large sample sizes, these two measures tend to the same values.

Given the previous considerations on the decreasing shape of the background density, the data used to study the behaviour of the estimators of complexity is generated from a truncated exponential distribution (unless specified otherwise), with density function

$$f_Y(y|\lambda, b) = \frac{e^{-y/\lambda}/\lambda}{1 - e^{-b/\lambda}},$$

where b is the upper limit of the distribution, the lower limit is 0 and λ is a parameter which corresponds to the mean in the non-truncated version of the distribution.

For data generated from an exponential with $\lambda = 3$, truncated between 0 and 5, the dependence of these measures on the bandwidth of the LOrPE smoother is

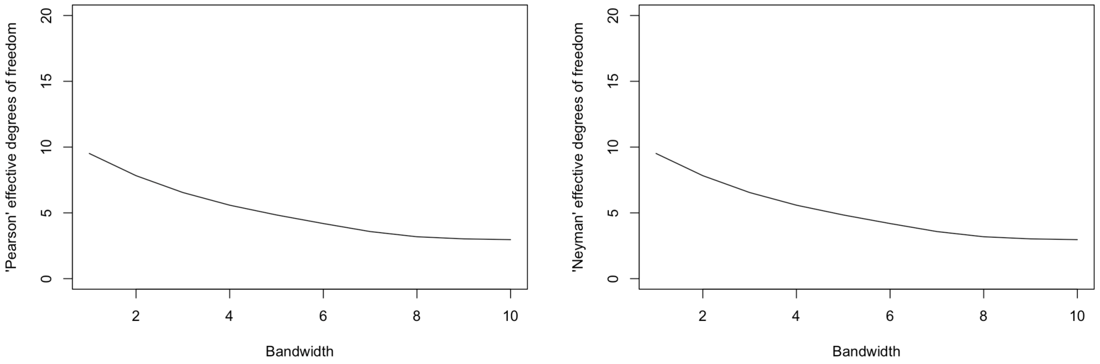

Figure 5.1: Dependence of df_P (left) and df_N (right) on the bandwidth, for an exponential with $\lambda = 3$, truncated between 0 and 5. The degree is fixed at 3.

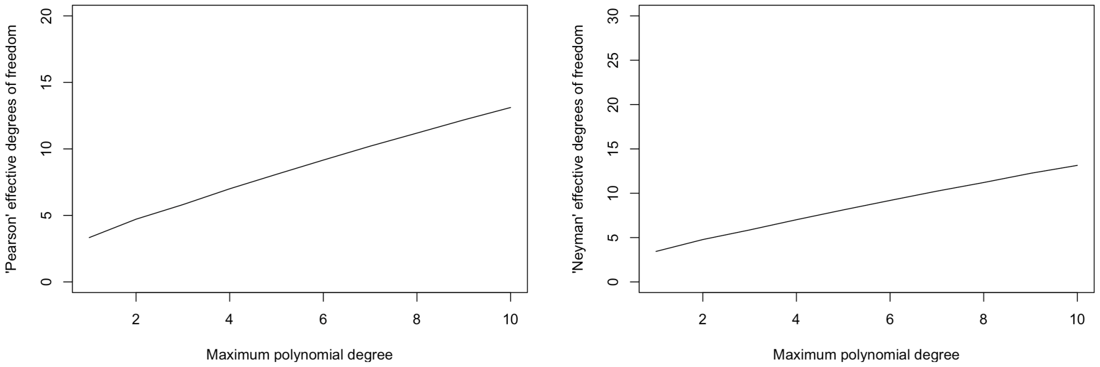

Figure 5.2: Dependence of df_P (left) and df_N (right) on the degree, for an exponential with $\lambda = 3$, truncated between 0 and 5. The bandwidth is fixed at 3.

decreasing (Figure 5.1); and it is increasing with the degree of the polynomials used in the smoother (Figure 5.2), which is in line with what is expected from the estimators.

However, if λ is smaller, these measures tend to actually be larger for very high levels of smoothing (large bandwidths).

This phenomenon is interpreted as being due to the fact that if λ is small, then the density decays quickly towards 0 and so the bins in the tail have a very small number of entries compared to the bins near 0. With higher levels of smoothing, this curve becomes more flattened and the number of entries in the tail bins actually increases, and so does the variance. If the difference between the bins is large, this effect dominates (Figure 5.3).

Since the level of smoothing is related to the covariance between the value of the estimated density at different points, another way of estimating the density

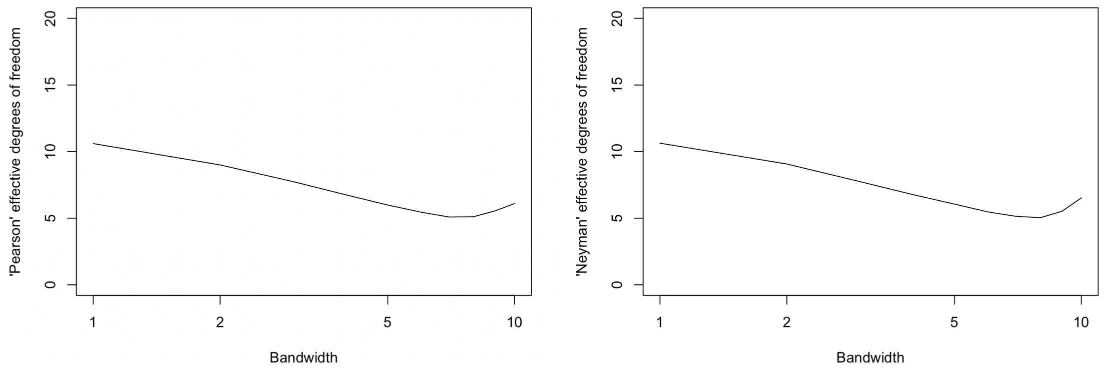

Figure 5.3: Dependence of df_P (left) and df_N (right) on the bandwidth, for an exponential with $\lambda = 1$, truncated between 0 and 5. The degree is fixed at 3. The bandwidth axis is in the logarithmic scale in order to visually highlight the non-monotonicity of the plots.

is through the covariance matrix of the estimated density. In the binned case, this matrix is $S_h N S_h^T$.

This is analogous to estimating the "effective dimensionality" of a set of variables. The structure of a set of variables, as described by the covariance (or correlation) matrix, can in fact be summarized by an equivalent number of orthogonal dimensions, which quantifies the effective dimensionality of the original variables. This is closely tied to the concept of information entropy. In information theory, *entropy* may be defined intuitively as the information content of a probability distribution [14].

Thus, in order to obtain a single value that estimates the overall level of complexity, we considered the use of the entropy-based effective rank (Roy and Vetterli [36]), which is defined as follows.

Assume the covariance matrix has size K (i.e. the number of bins in the discretized density) and let $\sigma = (\sigma_1, \ldots, \sigma_K)$ be its singular values. Take

$$p_i = \sigma_i / ||\sigma||_1, \quad i = 1, \ldots, K;$$

where $||\sigma||_1 = \sum_{i=1}^{K} |\sigma_i|$ is the L_1 norm. Then

$$H(p_1, \ldots, p_K) = -\sum_{i=1}^{K} p_i \log p_i$$

is the Shannon entropy [40].

An indicator of effective complexity is thus the entropy-based effective rank of the covariance matrix,

$$df_{er} = erank(C) = e^{H(p_1,\ldots,p_K)}.$$

Good [17] also used entropy as a roughness penalty, in discrete probability function estimation.

Another measure that was considered is a trace-based effective rank:

$$df_{tr} = \frac{\sum_{i=1}^{K} \lambda_i}{\lambda_1}, \tag{5.1}$$

where λ_i is the *i-th* eigenvalue of the covariance matrix and λ_1 is the largest one.

Del Giudice [14] reviews other measures of effective complexity, which have been also considered here in an exploratory analysis, the results of which are reported in the Appendix, in Section A.

5.2 Unbinned data

The final objective of this study is to work on a non-discretized density function, as the mass is in fact a continuous variable. In this case, it is not possible to obtain the covariance matrix analytically, as there are no bins the distribution of which may be known.

The covariance is thus estimated through the bootstrap method, which is implemented in the following way.

1. Resample the original data (real or simulated);

2. apply the smoother on this data and obtain a vector of values of the estimated density corresponding to a grid of coordinates of length K;

3. repeat this a large number of times, always on the same grid;

4. compute the $K \times K$ variance and covariance matrix for the estimated densities.

Once the covariance matrix is obtained, a measure of the effective degrees of freedom can be obtained through the estimators introduced above.

It can be seen that the complexity computed from the bootstrap estimate of the covariance matrix are equivalent to the ones computed from the analytical covariance matrix: Figure 5.4 shows, for different values of the bandwidth, the entropy-based and the trace-based effective ranks for the two methods of computing the covariance, both on binned data, and the bootstrap estimate for the unbinned case. For large sample sizes and a very fine grid of coordinates or finely binned data, these measures are expected to be a representation of the true complexity of the model and to converge to the same values in the binned and unbinned cases. In fact, the figure represents these measures with 50 bins and they are the same as their unbinned versions. For an increasing number of bins, the number of effective degrees of freedom from the bootstrap covariance on binned data converges to the unbinned measure (Figure 5.5). For the comparison, in the unbinned case the grid of coordinates is chosen so that the points at which the estimated density is evaluated correspond to the centers of the bins of the binned case.

5.2.1 Local effective degrees of freedom

The effective degrees of freedom are localized through the use of a “Gaussian dip” sampling function $d(x)$ (between 0 and 1) (Figure 5.6). It is such that the center of its “dip” corresponds to the center of the region of interest and defined as:

$$d(x) = \left(1 + a \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right)^{-1},$$

where μ is the location of the dip, $\sigma > 0$ and $a > 0$ represents the “amplitude” of the dip.

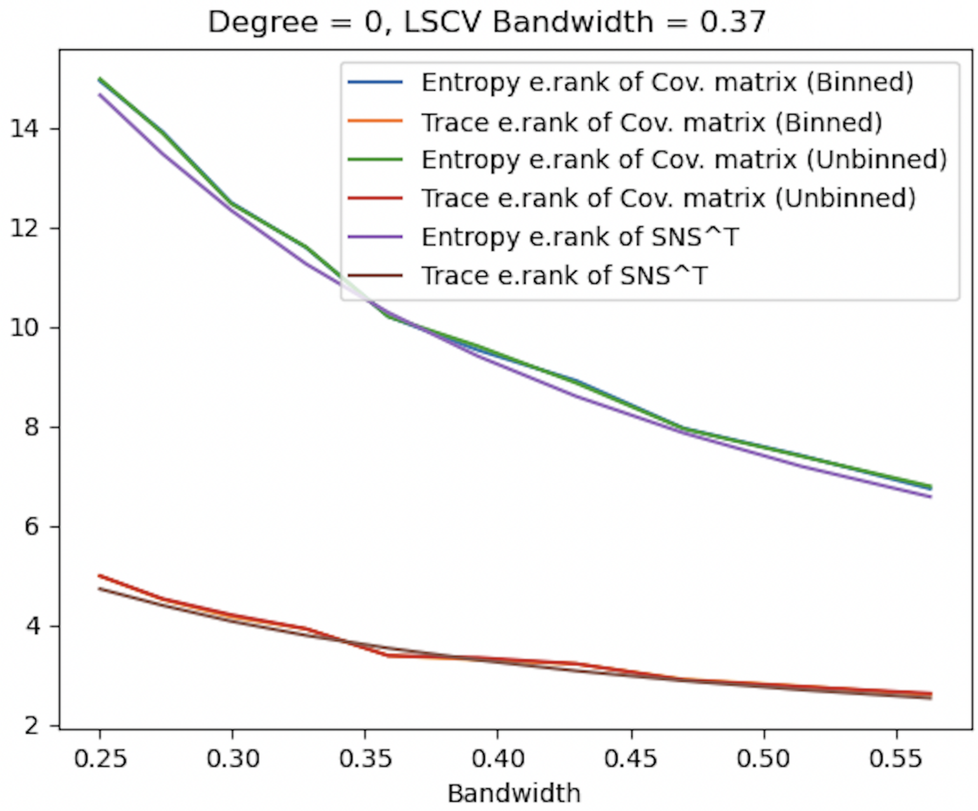

Figure 5.4: Dependence of df_{er}, df_{tr} on the bandwidth, as computed from the analytical covariance matrix on binned data, with 50 bins (purple, brown); from the bootstrap estimate on binned data (blue, orange); from the bootstrap estimate on unbinned data (green, red). The degree parameter is 0, the bandwidth varies around the one chosen with Cross-Validation. Note that the bootstrap measures for binned and unbinned data overlap.

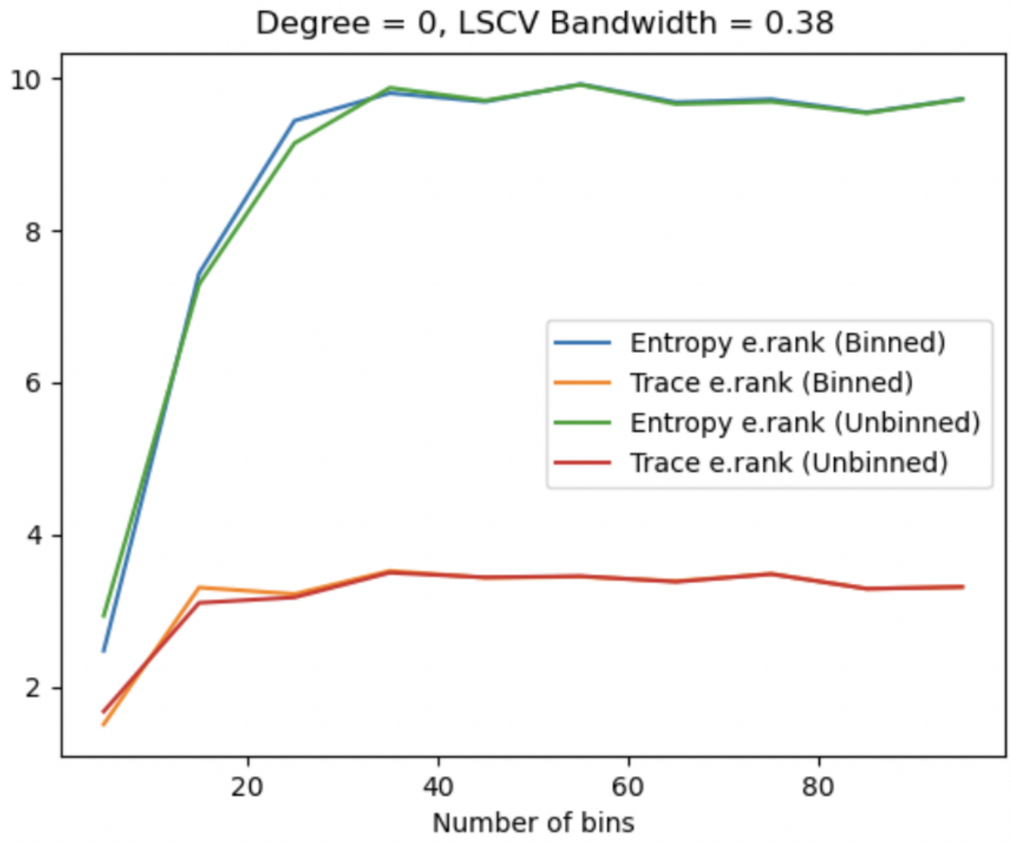

Figure 5.5: Dependence of df_{er}, df_{tr} from the bootstrap covariance matrix on the number of bins, compared to the unbinned version. The degree parameter is 0, the bandwidth is chosen via LSCV.

This function can be used to generate weighted random samples from some density $f(x)$ as follows.

1. Generate a random number x distributed with density $f(x)$;
2. calculate $d(x)$;
3. generate a random number r distributed uniformly in $[0,1]$;
4. accept x if $r \leq d(x)$ and reject otherwise. Assign the weight $d(x)^{-1}$ to the accepted point;
5. repeat previous steps until some predefined number of points, n, is accepted.

This procedure is equivalent to generating n random points x_i i.i.d. with density $\rho(x) = \frac{d(x)f(x)}{\int_a^b d(y)f(y)dy}$, and assigning weights $d(x_i)^{-1}$ to these points. These weights need to be assigned as less points will be accepted in the dip region and so they need to weigh more in order to maintain the original shape of the target function.

The method described above can not only be used to sample from any continuous function $f(x)$, but from the empirical density function of the data as well: sampling from it through this sampling function generates a weighted resampled set of the original data. By taking the effective rank of the covariance matrix of the smoothed densities computed on the weighted sample, an estimate of a kind of localized effective degrees of freedom may be obtained. The intuition behind this method relies on the fact that this weighted resampling causes less data points to be accepted in the region of interest so, when repeating the resampling many times (as it is done when computing the bootstrap variance and covariance matrix), the resampled data sets will have more variability in that region. More variability translates in higher complexity, viceversa in regions far from the coordinates of interest. For this reason the method is explored as a way of localizing the measure of model complexity.

Section C of the Appendix contains a series of plots, showing the dependence of the global and local effective degrees of freedom, obtained as described above,

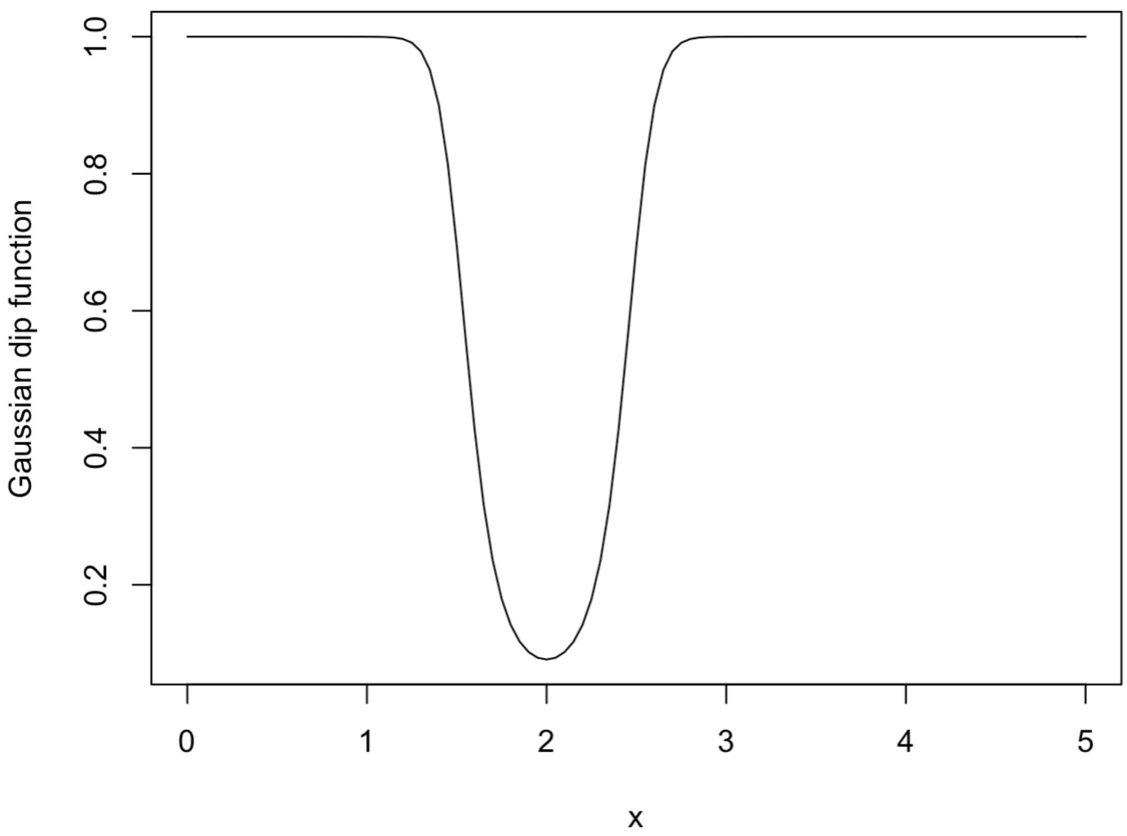

Figure 5.6: Gaussian dip sampling function with $\lambda = 2$, standard deviation = 0.2 and amplitude = 10.

on the bandwidth and maximum degree of the LOrPE operator. The plots also represent the behaviour of these measures when the underlying distribution of the simulated data and the parameters of the localizing Gaussian dip function vary.

The exploration of other methods of localization of the effective degrees of freedom is reported in the appendix, in Section B.

Chapter 6

Simulation studies

This chapter presents some simulation studies, which encompass different cases to which the procedure for the semiparametric density estimation may be applied.

6.1 Global procedure

The first case considered is one in which the background density is an exponential with λ equal to 2, truncated between 0 and 5.

The signal is assumed to have a Gaussian shape with mean equal to 2 and standard deviation equal to 0.1. The total number of simulated signal-plus-background observations is a realization of a Poisson distribution with mean 5000. The signal fraction is 0.05.

Let us first apply the global procedure, scanning a grid of values for α between 0 and 0.1, with a 0.005 step. Small value of the signal fraction are scanned as this is in line with the usual rarity of signal processes in particle physics experiments. The model chosen by the algorithm, using the log-likelihood penalized with the entropy-based effective rank of the covariance matrix, is in fact the one where the mixing proportion is 0.05, which is the true signal fraction (Figure 6.1, left). Simulations show that this measure of complexity is more reliable than the trace-based one, so it will be used in the following examples. Plotting the estimated and true densities over the histogram also shows how the density of the background (and thus the total density, since the signal is known) is correctly estimated (Figure 6.2). Since this is the global version of the procedure, and the background is overall smooth, the density is estimated well on the entire support. On the other hand, Figure 6.1 (right) shows the likelihood of each α, and it illustrates how the semiparametric mixture is not identifiable without considering the smoothness assumption: the maximum-likelihood value of α is, in fact, 0. The algorithms scans maximum degree values between 0 and 4, with a 0.25 step. Non-integer values of the degree represent the use of a taper function, mentioned in paragraph 3.1. The bandwidth values which are scanned by the cross-validation algorithm are reported in Table 3.1. The bandwidth and degree chosen by the localized cross-validation are such that the overall complexity is lower when the signal fraction is correct, since the background does not need to compensate for the incorrectly modeled signal peak (see Figure 6.1 for the density estimate with assumed $\alpha = 0.05$). If the assumed α is smaller than 0.05, the complexity is larger, to allow for the background to model a peak to be added to the signal one. For α larger than 0.05, the bandwidth also allows for larger complexity, and the background models a dip to compensate for the assumed signal. The latter is not in practice a problem in the setting considered here, as we are interested in signals which are very small. Figure 6.3 displays this effect, with assumed $\alpha = 0.005$ (left) and 0.1 (right). Figure 6.4 shows how the effective degrees of freedom are in fact smaller around the true signal fraction. 500 resampling iterations are used for the computation of the bootstrap covariance matrix of the estimated densities, and they are computed on a grid of 500 coordinates. The value for γ, the penalization parameter which decides the weight given to the complexity, is here chosen empirically through simulations and it is 2.

If the signal fraction is instead 0.03, the correct value is again chosen and the density estimate is very accurate (Figure 6.5).

The algorithm also identifies the absence of a signal, and selects $\alpha = 0$ in this case (Figure 6.6).

Consider now a $80 - 20$ mixture of a Beta distribution with parameters 1 and

3 and a Uniform between 0 and 1 as background distribution, and a Gaussian signal with 0.3 and standard deviation 0.02. The signal fraction is $\alpha = 0.03$. The chosen α is 0.035 (Figure 6.7, left) which is still quite close to the correct one. See Figure 6.7 (right) for the fit to the histogram.

Let us now introduce a nuisance structure in the tail of the distribution, by using, as a background density, a $95-5$ mixture of an exponential with $\lambda = 2$, truncated between 0 and 5, and a Gaussian with mean 4.5 and standard deviation 0.05, truncated between 0 and 5. This results in a high, narrow peak at the tail of the distribution. This corresponds to a structure that has a different scale than the ones associated with the background near the signal, so it may not be effectively estimated by a model with the same smoothness level. Note that the nuisance may also consist in smaller fluctuations. The signal is here again a Gaussian with mean equal to 2 and standard deviation equal to 0.1. The signal fraction is 0.05.

This extreme case of severe non-smoothness illustrates precisely why localization may be necessary, as a potential element of nuisance is the reason why the localization method developed in this thesis is useful. Indeed, the non-localized procedure fails in this case, since the estimation of complexity also takes into consideration the peak at the tail, which is not of interest. Indeed, if the cross-validation takes into consideration the nuisance in the tail, since the bandwidth is fixed on the entire support, it chooses an optimal complexity level which is too high for the smooth background near the signal, so the model chosen is not smooth. Furthermore, the maximum penalized likelihood signal fraction is 0.035, not the correct 0.05 (Figure 6.8). The complexity, estimated by the effective degrees of freedom, does not only take into account the peak or dip in the signal region, but also the non-signal peak, so it is not an accurate penalization for the choice of α (Figure 6.9).

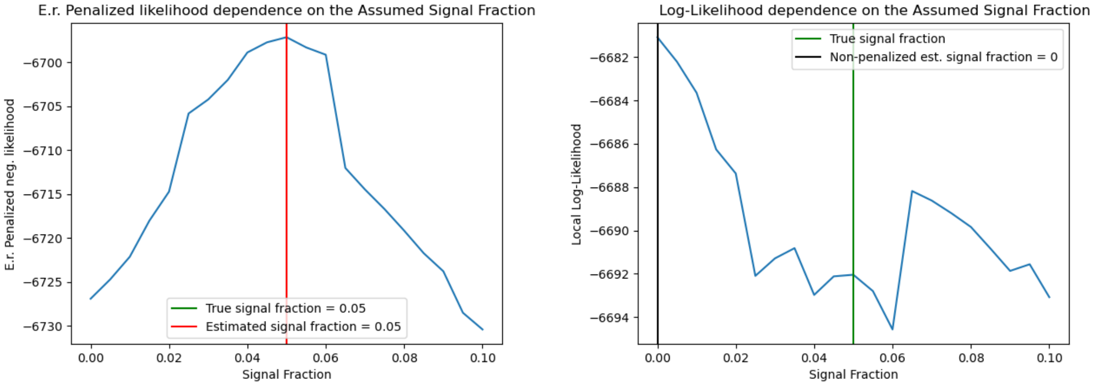

Figure 6.1: Dependence of the penalized likelihood (left), and non-penalized likelihood (right) on the assumed signal fraction, for the smooth truncated exponential background, when the true signal fraction is 0.05.

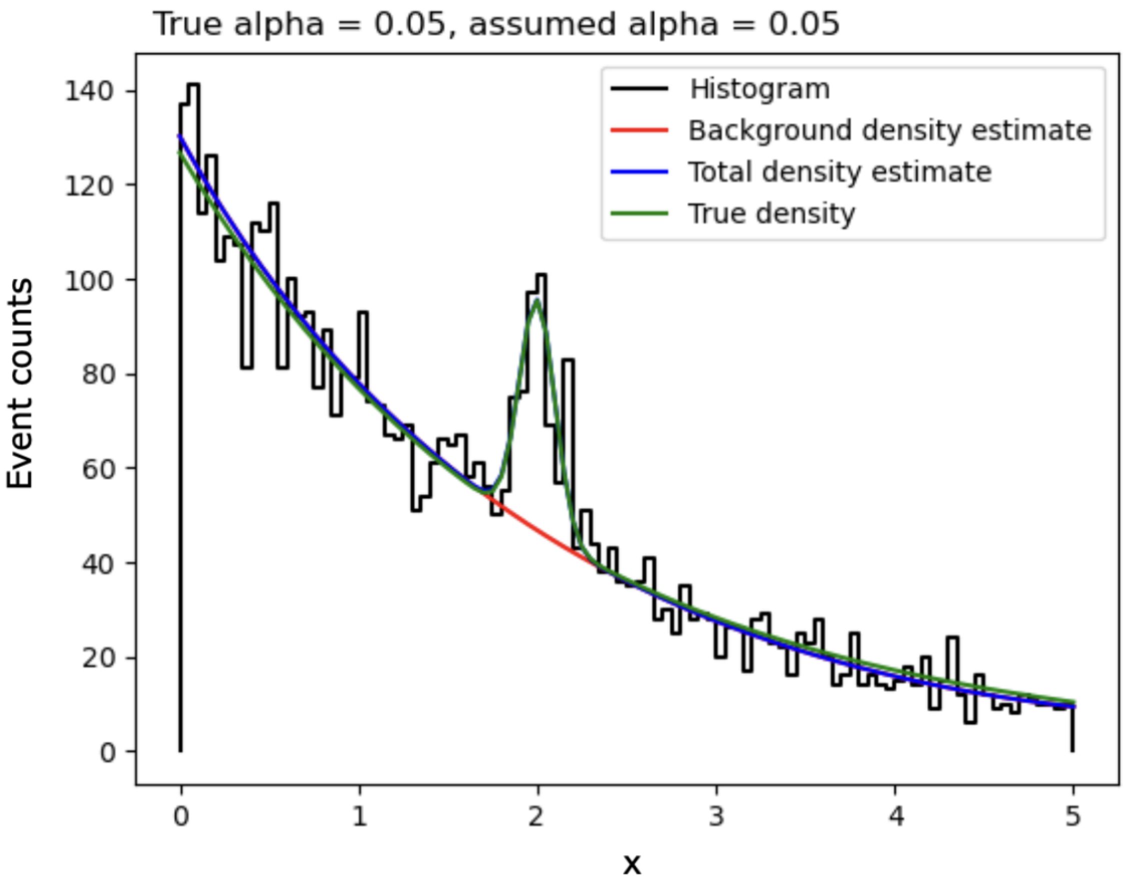

Figure 6.2: Histogram of the data (black), true total density (green) and estimated densities using the maximum penalized likelihood value of α (background density in red, total density in blue), for the smooth truncated exponential background, when the true signal fraction is 0.05.

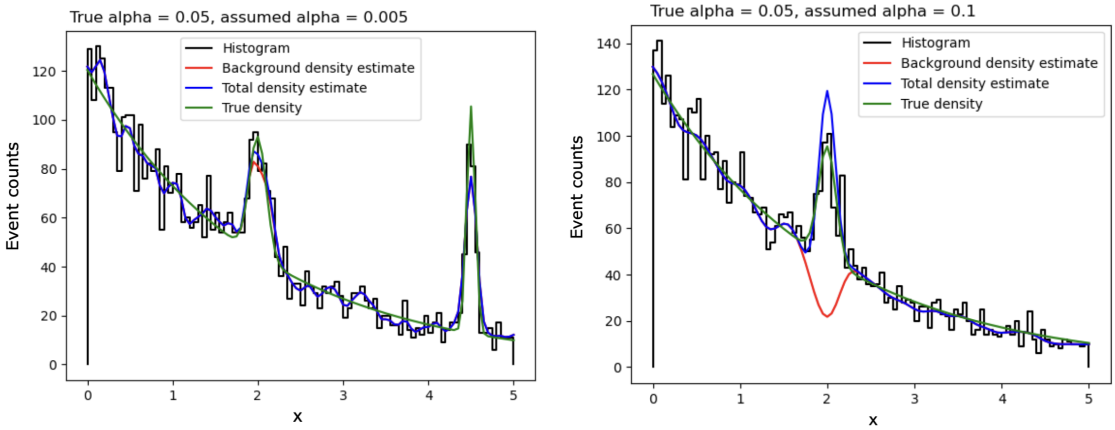

Figure 6.3: Histogram of the data (black), true total density (green) and estimated densities using $\alpha = 0.005$ (left) and $\alpha = 0.1$ (left), for the smooth truncated exponential background, when the true signal fraction is 0.05.

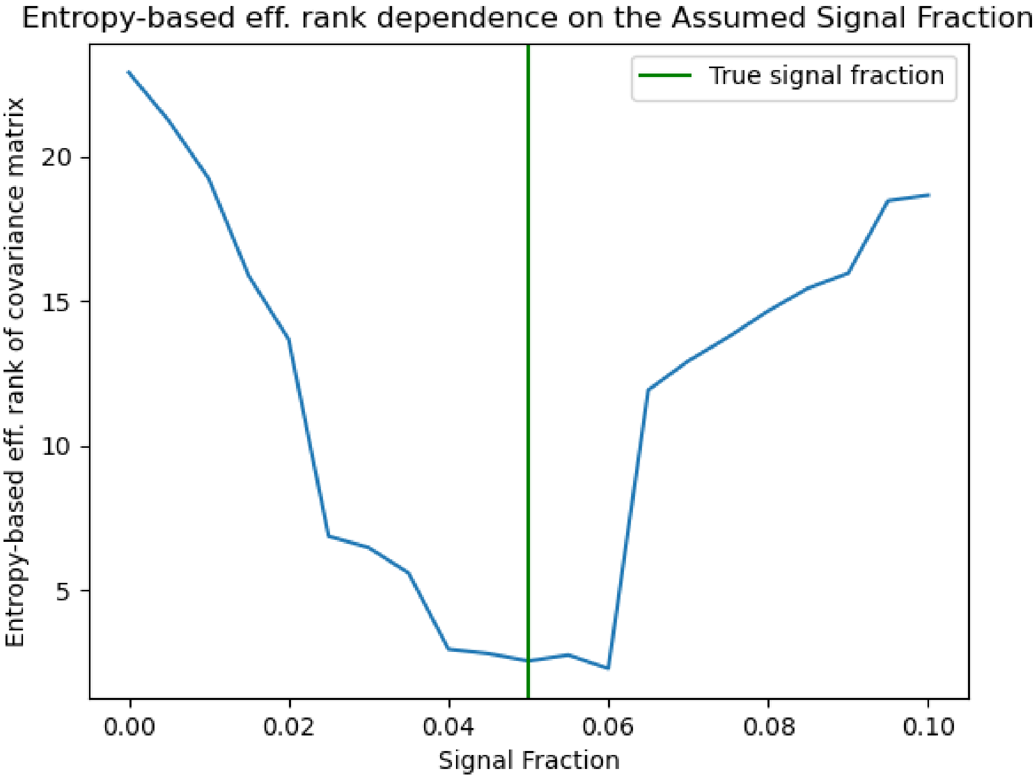

Figure 6.4: Dependence of the effective degrees of freedom on the assumed signal fraction, for the smooth truncated exponential background, when the true signal fraction is 0.05.

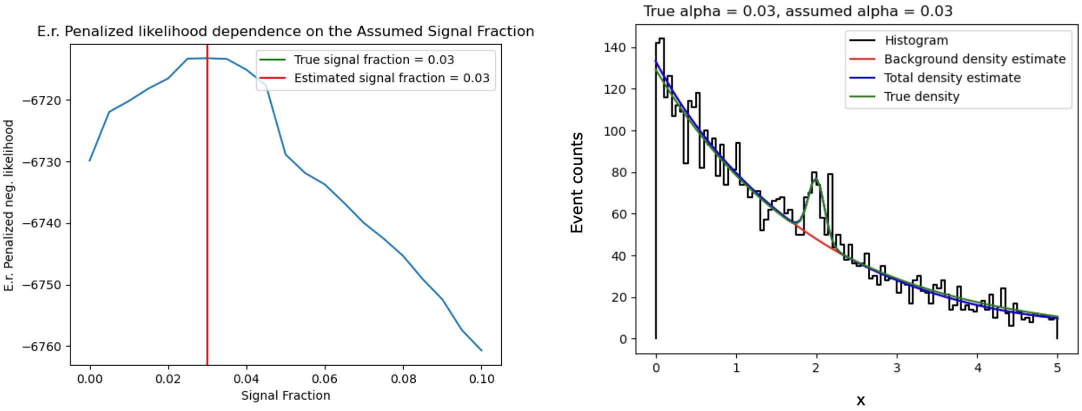

Figure 6.5: Dependence of the penalized likelihood on the assumed signal fraction, for the smooth truncated exponential background, when the true signal fraction is 0.03 (left); true and estimated densities with the maximum penalized likelihood value of α (right).

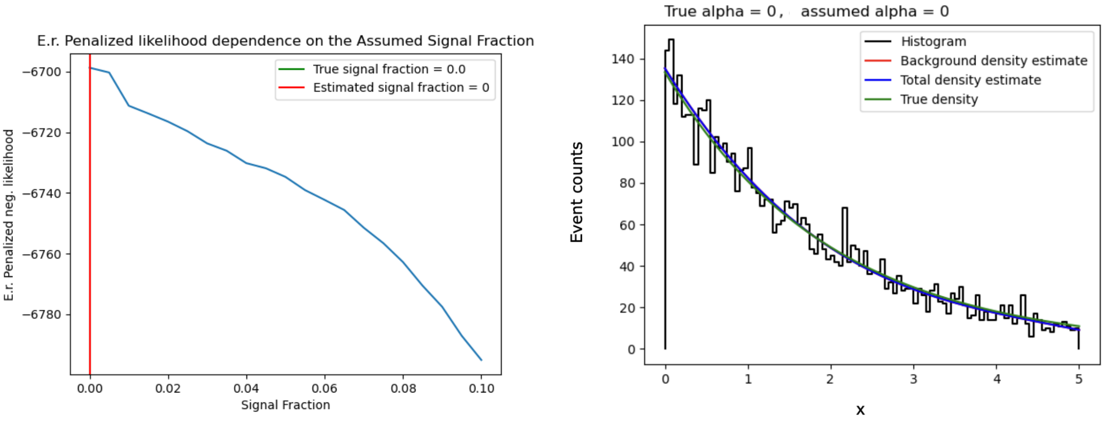

Figure 6.6: Dependence of the penalized likelihood on the assumed signal fraction, for the smooth truncated exponential background, when the true signal fraction is 0 (left); true and estimated densities with the maximum penalized likelihood value of α (right).

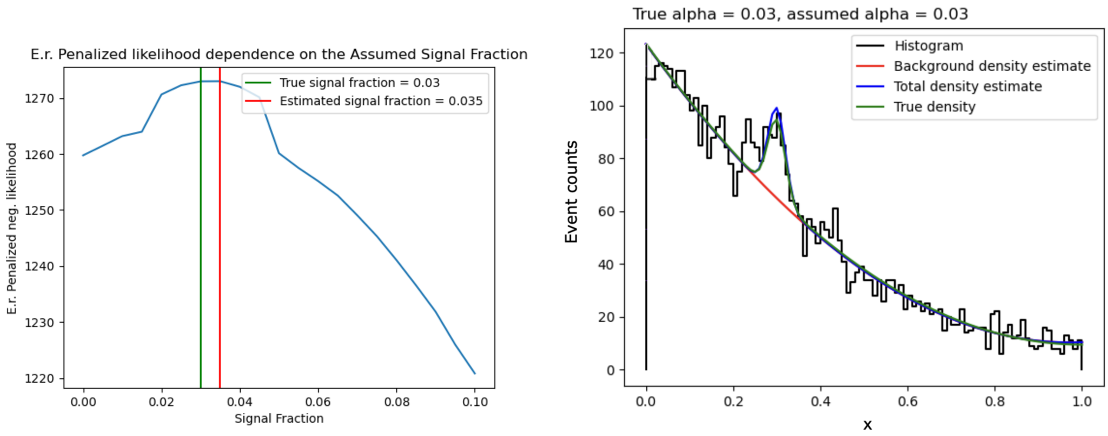

Figure 6.7: Dependence of the penalized likelihood on the assumed signal fraction, for the Beta-Uniform mixture background, when the true signal fraction is 0.03 (left); true and estimated densities with the maximum penalized likelihood value of α (right).

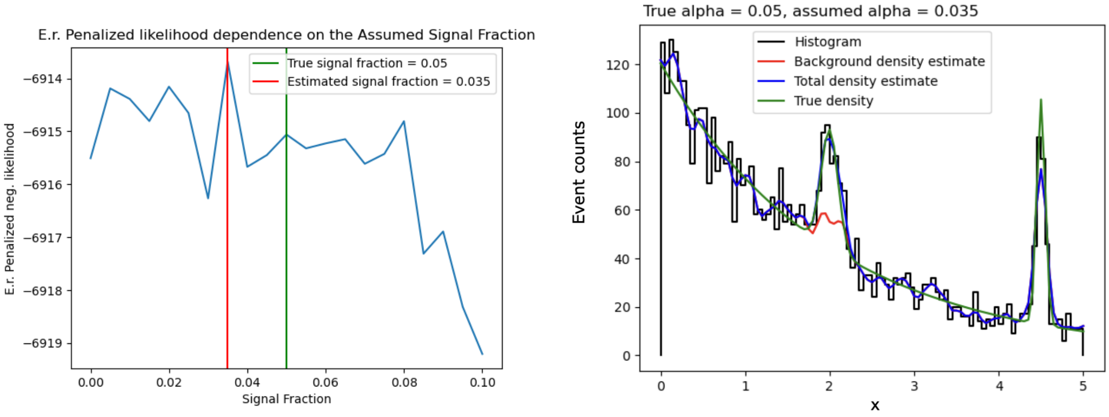

Figure 6.8: Dependence of the penalized global likelihood on the assumed signal fraction, for the truncated exponential background with nuisance, when the true signal fraction is 0.05 (left); true and estimated densities with the maximum penalized likelihood value of α (right).

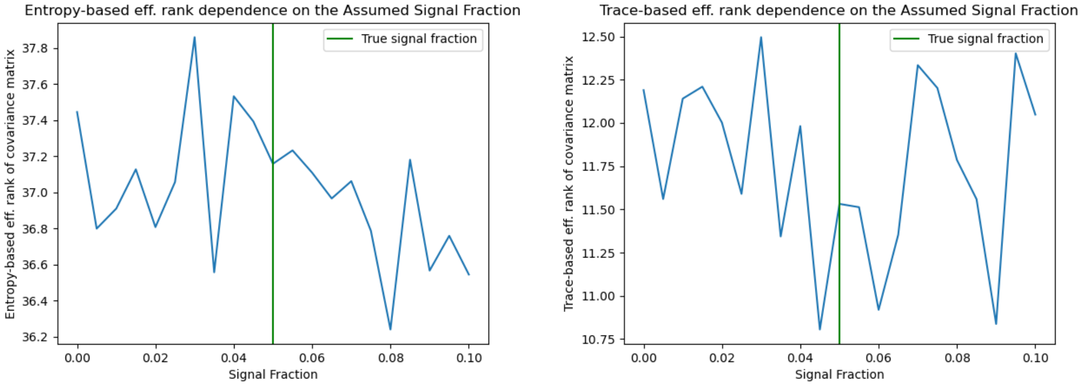

Figure 6.9: Dependence of the effective degrees of freedom (entropy-based, left and trace-based, right) on the assumed signal fraction, for the truncated exponential background with nuisance, when the true signal fraction is 0.05.

6.2 Local procedure

The simulations in this paragraph showcase how the introduction of localization allows the method to work correctly and accurately also on data that contains structures with different scales from the signal. In this case, since the algorithm can be slower when localized, a fixed maximum polynomial degree equal to 0 was used. This means using kernel density estimation with a boundary correction, as described in paragraph 3.1. The localization of the bandwidth choice makes the goodness-of-fit of the model worse far from the signal region, but, as stated before, this is not a problem of interest. The penalization using the localized effective degrees of freedom also only penalizes the model which contains a peak or a dip in the background density in the signal region, and ignores irregularities outside of it. The dependence of the effective degrees of freedom once again has a minimum at the true signal fraction (Figure 6.11, left). Since the degree is fixed, this shape is also reflected in the bandwidth choice for each assumed α (Figure 6.11, right). This behaviour allows the density to be estimated correctly around the signal (Figure 6.12, left) and the mixing proportion to be chosen appropriately, 0.05 in this case (Figure 6.12, right). Note that, without penalization, the maximum local likelihood estimate of α is 0 (Figure 6.13). In the case just described, the sampling function for the computation of the localized effective degrees of freedom is a Gaussian dip centered on 2, with $\sigma = 0.2$ and amplitude equal to 10. A location transformation with parameter 0.8 is applied to the localizing function (Figure 6.10, left). A transformation with parameter 1.5 was also explored (Figure 6.10, right), and the algorithm still chose the correct signal fraction (Figure 6.14, left), however the function seemed to give too much importance to a region with fluctuations and thus the bandwidth choice was small, and the estimate not very smooth (Figure 6.14, right).

If the true signal fraction is instead 0, so there is no signal fraction, the model also chooses the correct background density. For the truncated exponential background with a nuisance element at the tail, the smoothest model estimate of α is, correctly, 0 (Figure 6.15, left), and the density estimate is accurate (Figure 6.15, right).

The exemplifying cases considered can, of course, be treated by simple truncation, so that the non-signal structure is not included in the support. However, in practice, contamination might not be this obvious and it may not be possible to find a suitable margin for truncation.

These simulations have been run using "oracle" weights for faster convergence (as mentioned in paragraph 3.3), but the procedure yields the same results if starting from an arbitrary density as a first approximation of the background. For example, by starting the iterations with a uniform approximation of the background density instead of the correct truncated exponential with nuisance, the model estimate is equally as accurate (Figure 6.16, right) and the chosen α is correct (0.05 in this case, see Figure 6.16, left).

6.3 Comparison with the parametric fit

Let us now consider the first case of the truncated exponential background with $\lambda = 2$ and signal fraction equal to 0.05. If we assume the shape of the background density to be known, a fully parametric fit (obtained by maximizing the likelihood with the optimization tool Minuit [23]) yields a correct estimate

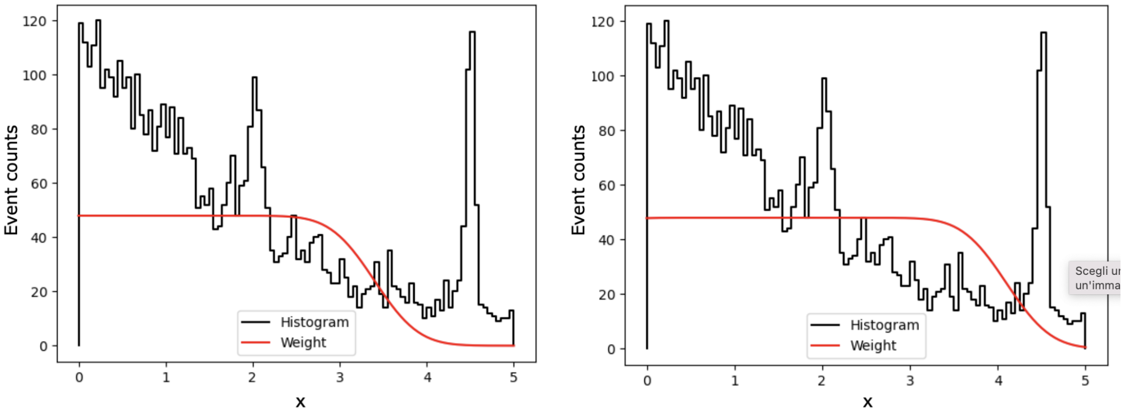

Figure 6.10: Convolution of a Uniform between −6.5 and 6.5, and a standard normal, with scale transformation parameter equal to 0.4 and location transformation with parameter 0.8 (left) and 1.5 (right), plotted in red over the histogram of the truncated exponential with nuisance simulated data, in black. In the plots, the weight function is multiplied by a factor that depends on the scale of the histogram, for sake of visual clarity.

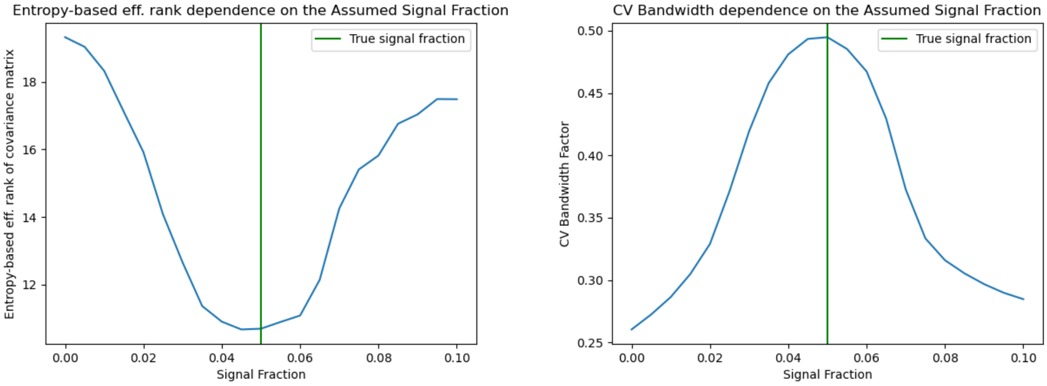

Figure 6.11: Dependence of the localized entropy-based effective degrees of freedom (left) and localized cross-validation bandwidth (right) on the assumed signal fraction, for the truncated exponential with nuisance background, when the true signal fraction is 0.05.

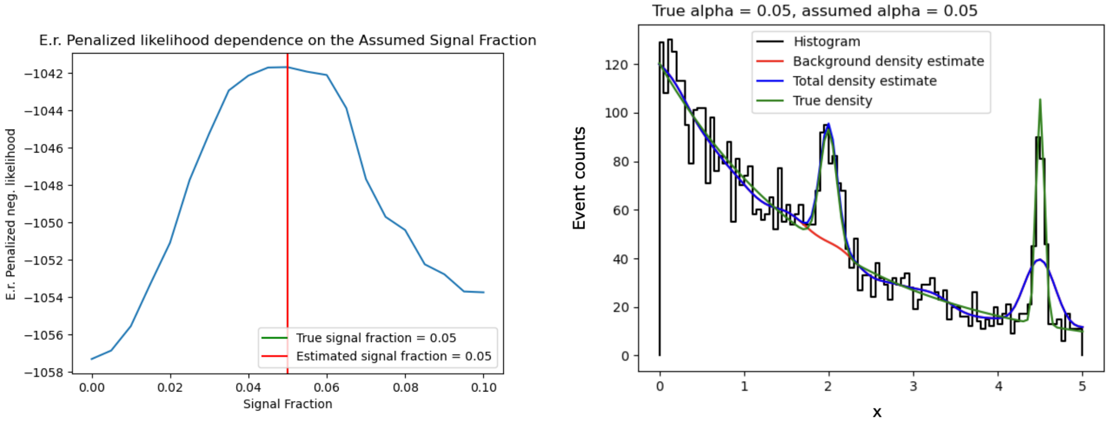

Figure 6.12: Dependence of the penalized local likelihood on the assumed signal fraction, for the truncated exponential with nuisance, when the true signal fraction is 0.05 (left); true and estimated densities with the maximum penalized local likelihood value of α (right).

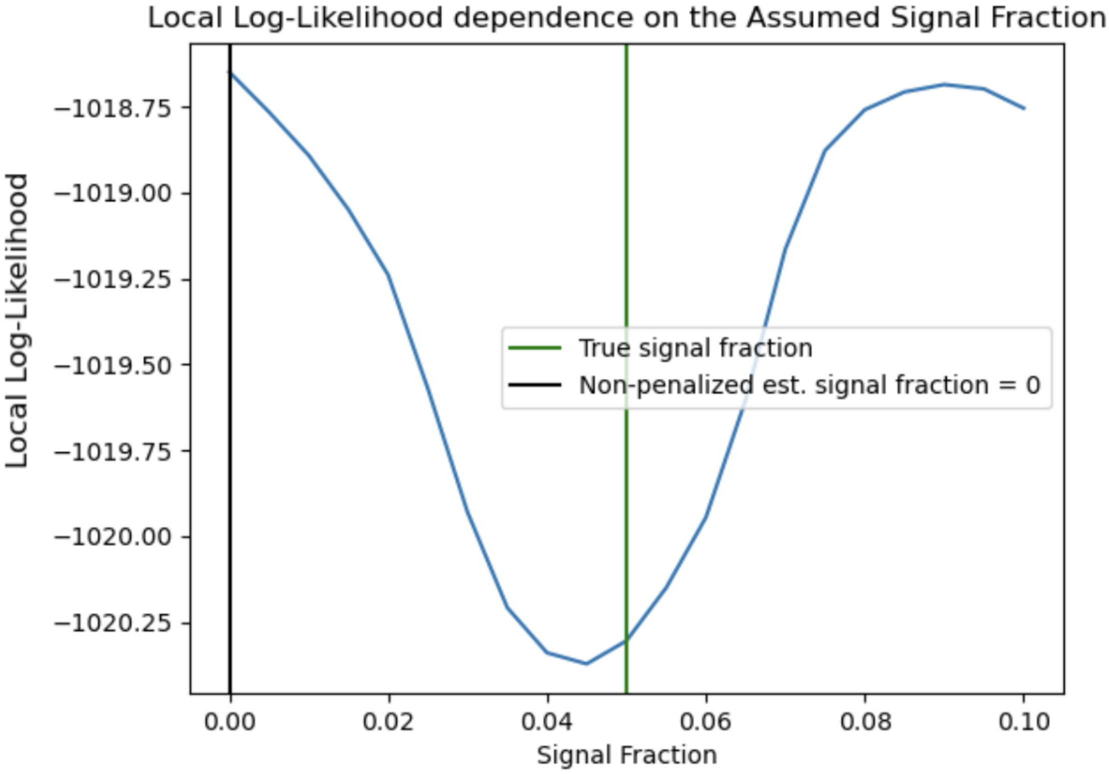

Figure 6.13: Dependence of the non-penalized local likelihood on the assumed signal fraction, for the truncated exponential with nuisance, when the true signal fraction is 0.05.

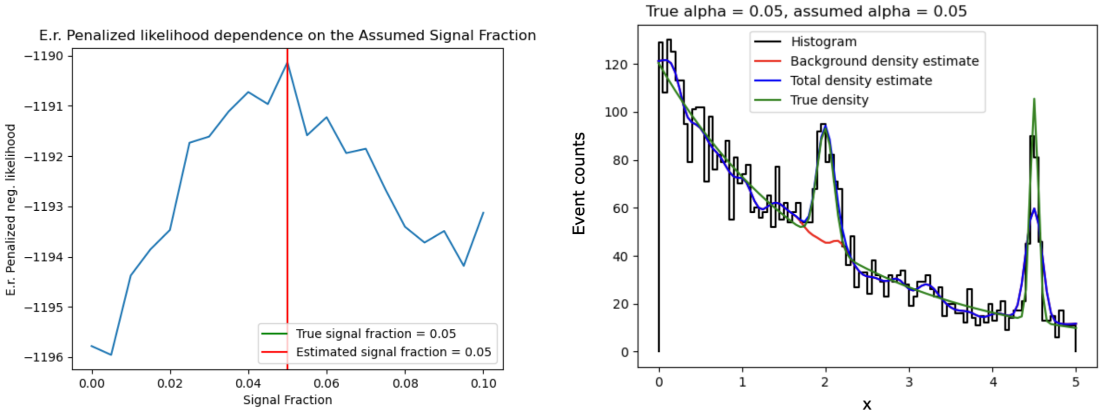

Figure 6.14: Dependence of the penalized local likelihood on the assumed signal fraction, for the truncated exponential with nuisance, when the true signal fraction is 0.05 (left); true and estimated densities with the maximum penalized local likelihood value of α (right), with 1.5 location parameter for the localizing function.

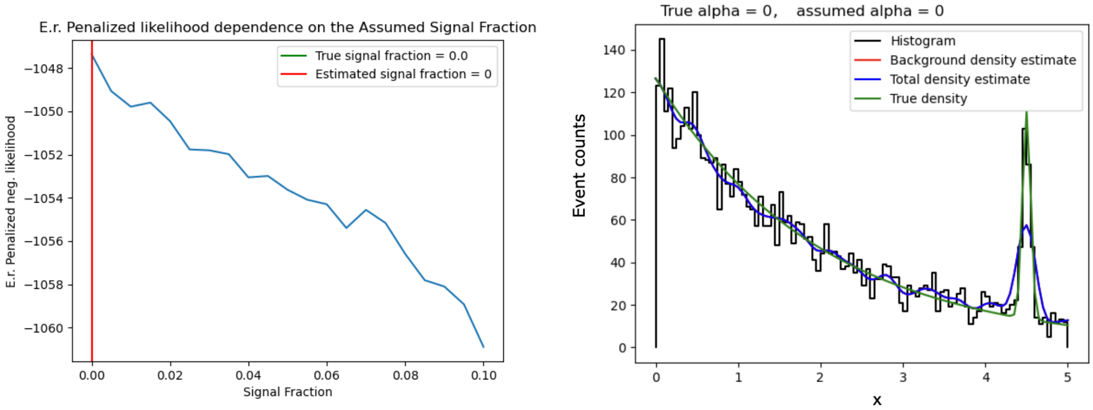

Figure 6.15: Dependence of the penalized local likelihood on the assumed signal fraction, for the truncated exponential with nuisance, when the true signal fraction is 0 (left); true and estimated densities with the maximum penalized local likelihood value of α (right).

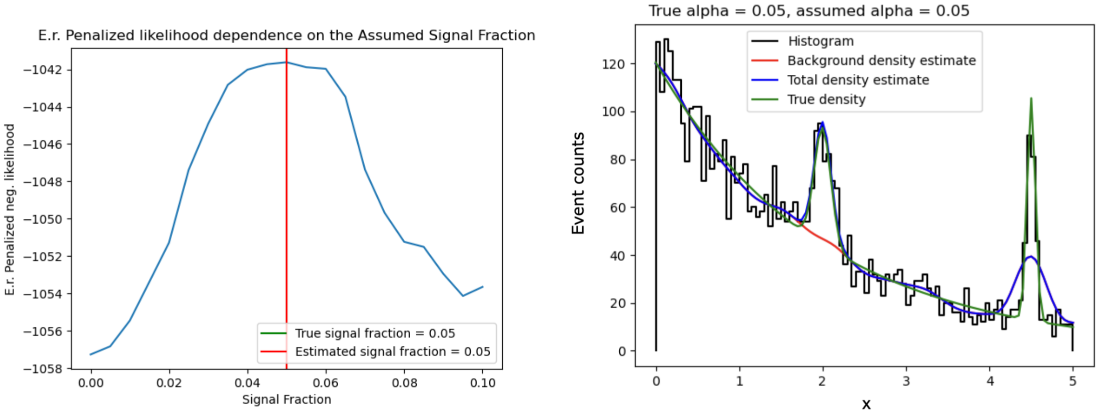

Figure 6.16: Dependence of the penalized local likelihood on the assumed signal fraction, for the truncated exponential with nuisance, when the true signal fraction is 0.05 (left); true and estimated densities with the maximum penalized local likelihood value of α (right). The weights are obtained by starting from a Uniform approximation of the background.

of the density (Figure 6.17), with an estimated λ equal to 1.928 ± 0.044. The estimate of α is 0.058 ± 0.005. This method allows one to obtain the level of uncertainty of the estimates and thus to carry out statistical tests and compute confidence intervals on α. This is not yet implemented in the semiparametric version, where the choice of α is treated as a model selection problem, however it may be introduced during future research. Some previous work on the formalization of penalized likelihood and local likelihood estimation is cited in Chapter 4.

Furthermore, the shape of the background is not assumed to be known, as is often the case in practice. If we thus assume an incorrect background density (while still considering the signal to be known), the performance of the parametric method is worse than the semiparametric one. If we use a Gaussian with mean 0 and unknown standard deviation, the estimated signal fraction is 0.039 ± 0.005, which is quite far from the true one, 0.05, and the fit is not accurate (Figure 6.18). The fully parametric fit is thus correct if the model chosen is correct, but it is not robust with respect to model misspecification.

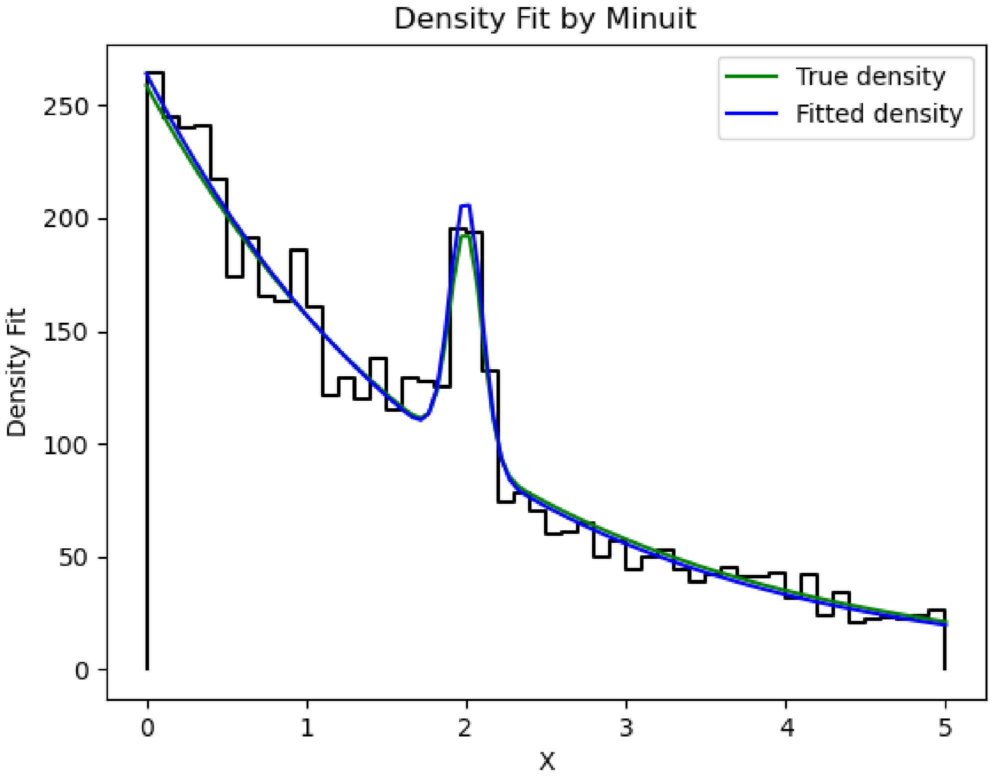

Figure 6.17: True (green) and estimated (blue) total densities with the parametric method, by assuming the correct parametric specification, in the smooth truncated exponential background case.

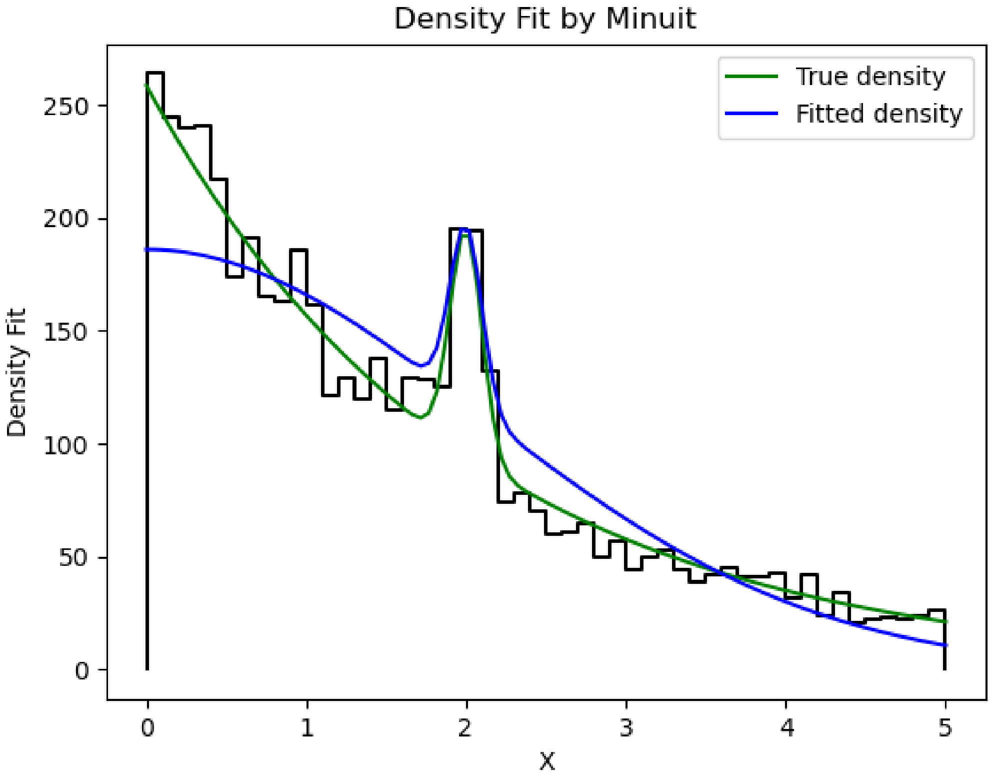

Figure 6.18: True (green) and estimated (blue) total densities with the parametric method, by assuming the incorrect parametric specification (half normal), in the smooth truncated exponential background case.

Chapter 7

Conclusions

The method developed in this thesis allows the density of the background events to be estimated effectively around a region of interest, through an iterative method that uses signal-plus-background data to nonparametrically estimate the sole background density. The estimate is produced using Local Orthogonal Polynomial Expansions and the complexity of this estimator is chosen using a (potentially localized) cross-validation criterion. The total density is estimated with a mixture of a known parametric signal and the nonparametric estimate of the background, its mixing proportion parameter is chosen via maximum penalized (local) likelihood. The penalization term accounts for the assumption of background smoothness and it is based on a estimator of the entropy of the covariance matrix, which is computed on resampled data. The data may be resampled using an algorithm that localizes the complexity around the signal region. The method is empirically shown to work in its global version on smooth-background simulated data, and in the localized version in the presence of a non-signal peak in the tail of the distribution.

Chapter 8

Further research and discussion

As mentioned before, the penalization term in the likelihood is defined up to a multiplicative factor, so the penalization parameter is not uniquely determined: simulations show that the penalized likelihood method works if enough weight is given to the smoothness term, however further studies to obtain a value that is supported by theory would improve the procedure. If the local complexity were computed in a sharply defined interval instead of a smooth one, the value of the smoothness term might be chosen such that, by summing the local degrees of freedom computed on each of the intervals that make up the support, the global degrees of freedom be obtained. A starting point for the theory behind the choice of the penalization parameter might thus be the search for a way of obtaining global effective degrees of freedom through combining the local effective degrees of freedom.

The effect that was described in Chapter 5 (Figure 5.3), can be further explored in a separate discussion, including the identification of "admissible" smoothers, which would be ones such that high levels of smoothing do not increase the variability with respect to that of the original data.

Theoretical studies are also to be carried out on the distribution of the maximum penalized likelihood estimator of the signal fraction parameter α, in order to get an appropriate measure of uncertainty on it. This is necessary for hypothesis testing and building confidence intervals.

The signal density is here assumed to be completely known, however a point of future research is the extension of the algorithm to include the estimation of the signal location, scale, or both of these parameters. The entire procedure may be adapted to the multivariate setting, in order to use more than one discriminating variable. Finally, further optimization of the algorithm to make it faster and more efficient may be needed to work with very large amounts of data, which is crucial when analyzing real high-energy physics data.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. US Government printing office, 1964.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*. Akademiai Kiado, 1973.

[3] D. Al Mohamad and A. Boumahdaf. Semiparametric two-component mixture models when one component is defined through linear constraints. *IEEE Transactions on Information Theory*, 64(2):795–830, 2018.

[4] ALICE Collaboration. Rapidity and transverse momentum dependence of inclusive J/ψ production in pp collisions at $s = 7$ TeV. *Physics Letters B*, 704:442–455, 2011.

[5] D.P. Amali Dassanayake, I. Volobouev, and A. A. Trindade. Local orthogonal polynomial expansion for density estimation. *Journal of Nonparametric Statistics*, 29(4):806–830, 2017.

[6] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1):1–29, 2012.

[7] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.

[8] L. Bordes, C. Delmas, and P. Vandekerkhove. Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian Journal of Statistics*, 33(4):733–752, 2006.

[9] Gregory Breit and Eugene Wigner. Capture of slow neutrons. *Physical Review*, 49(7):519, 1936.

[10] A. Buja, T. J. Hastie, and R. J. Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510, 1989.

[11] J. Chen. Penalized likelihood-ratio test for finite mixture models with multinomial observations. *Canadian Journal of Statistics*, 26(4):583–599, 1998.

[12] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30–61, 2012.

[13] S. R. Cole, H. Chu, and S. Greenland. Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *American Journal of Epidemiology*, 179(2):252–260, 2014.

[14] M. Del Giudice. Effective dimensionality: A tutorial. *Multivariate Behavioral Research*, 56(3):527–542, 2021.

[15] T. Gao and V. Jojic. Degrees of freedom in deep neural networks. *arXiv:1603.09260 [cs.LG]*, 2016.

[16] I. K. Glad, N. L. Hjort, and N. G. Ushakov. Correction of density estimators that are not densities. *Scandinavian Journal of Statistics*, 30(2):415–427, 2003.

[17] I. J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, 34(3):911–934, 1963.

[18] I. J. Good and R. A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.

[19] I. J. Good and R. A. Gaskins. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, 75(369):42–56, 1980.

[20] P. Hall and X. Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of Statistics*, 31(1):201–224, 2003.

[21] T. J. Hastie and R. J. Tibshirani. *Generalized additive models.* CRC Press, 1990.

[22] Wood S. N. Hauenstein, S. and C. F. Dormann. Computing AIC for black-box models using generalized degrees of freedom: A comparison with cross-validation. *Communications in Statistics-Simulation and Computation*, 47 (5):1382–1396, 2018.

[23] F. James and M. Roos. Minuit - A system for function minimization and analysis of the parameter errors and correlations. *Computer Physics Communications*, 10(6):343–367, 1975.

[24] X. Ji, W. Gu, X. Qian, H. Wei, and C. Zhang. Combined Neyman–Pearson chi-square: An improved approximation to the Poisson-likelihood chi-square. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 961:163677, 2020.

[25] B. G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–239, 1988.

[26] C. Loader. *Local regression and likelihood.* Springer Science & Business Media, 2006.

[27] Y. Ma and W. Yao. Flexible estimation of a semiparametric two-component mixture model with one parametric component. *Electronic Journal of Statistics*, 9(1):444–474, 2015.

[28] N. McCloud and C. F. Parmeter. Determining the number of effective parameters in kernel density estimation. *Computational Statistics & Data Analysis*, 143:106843, 2020.

[29] Q. V. Nong, C. T. Ng, W. Lee, and Y. Lee. Hypothesis testing via a penalized-likelihood approach. *Journal of the Korean Statistical Society*, 48(2):265–277, 2019.

[30] I. Olkin and C. H. Spiegelman. A semiparametric approach to density estimation. *Journal of the American Statistical Association*, 82(399):858–865, 1987.

[31] L. Pace and A. Salvan. *Principles of statistical inference from a Neo-Fisherian perspective.* World Scientific, 1997.

[32] R. K. Patra and B. Sen. Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):869–893, 2016.

[33] A. Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961.

[34] S. Robin, A. Bar-Hen, J. Daudin, and L. Pierre. A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics & Data Analysis*, 51(12):5483–5493, 2007.

[35] W. A. Rolke and A. M. López. Estimating a signal in the presence of an unknown background. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 685:16–21, 2012.

[36] O. Roy and M. Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, 2007.

[37] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):65–78, 1982.

[38] E. Schuster and S. Yakowitz. Parametric/nonparametric mixture density estimation with application to flood-frequency analysis 1. *Journal of the American Water Resources Association*, 21(5):797–804, 1985.

[39] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization.* John Wiley & Sons, 2015.

[40] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[41] B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, 10 (3):795–810, 1982.

[42] R. J. Tibshirani and T. J. Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567, 1987.

[43] B. A. Turlach et al. Bandwidth selection in kernel density estimation: A rewiew. Technical report, Humboldt Universitaet Berlin.

[44] Volobouev, I. NPStat — Non-parametric Statistical Modeling and Analysis. `https://npstat.hepforge.org`.

[45] X. Wang, C. van Eeden, and J. V. Zidek. Asymptotic properties of maximum weighted likelihood estimators. *Journal of Statistical Planning and Inference*, 119(1):37–54, 2004.

[46] S. Xiang, W. Yao, and J. Wu. Minimum profile Hellinger distance estimation for a semiparametric mixture model. *Canadian Journal of Statistics*, 42 (2):246–267, 2014.

[47] J. Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998.

[48] Y. Zhou and W. Yao. Maximum likelihood estimation of a semiparametric two-component mixture model using log-concave approximation. *arXiv:1903.11200v1 [stat.ME]*, 2019.

Appendix A

Other eigenvalue-based estimators

The estimators reviewed here are based on the eigenvalue decomposition of the covariance matrix, and are derived from the general form of the Rényi entropy [33]:

$$H_q = \frac{1}{1-q} \log \left(\sum_{i=1}^{K} p_i^q \right),$$

for different values of q, which defines what specific measure of entropy is to be used. p_i is the normalized *i-th* eigenvalue:

$$p_i = \frac{\lambda_i}{\sum_{j=1}^{K} \lambda_j},$$

where λ_i is the *i-th* eigenvalue of the covariance matrix.

The first three estimators are obtained from different measures of entropy, while n_C is based on the variance of the eigenvalues.

- n_1 is based on the Shannon entropy ($q = 1$):

$$n_1 = \prod_{i=1}^{K} \left(\frac{\lambda_i}{\sum_{j=1}^{K} \lambda_j} \right)^{-\frac{\lambda_i}{\sum_{j=1}^{K} \lambda_j}}.$$

- n_2 is based on the Rényi entropy with $q = 2$, it is more conservative than

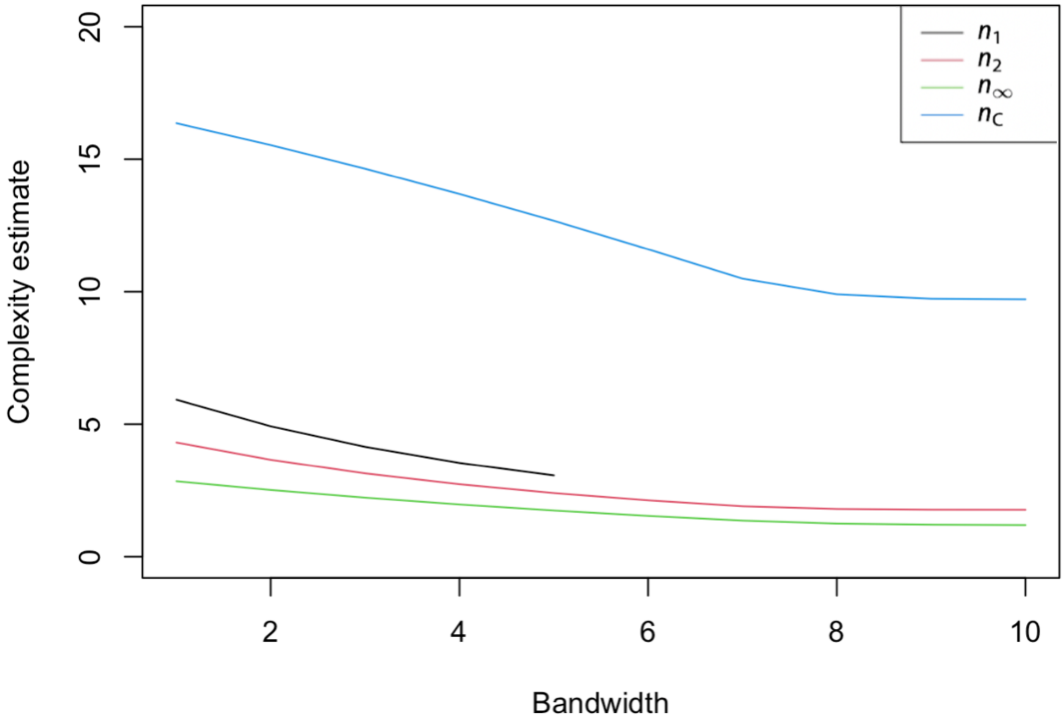

Figure A.1: Dependence of n_1 (black), n_2 (red), n_∞ (green), n_C (blue) on the bandwidth. Computed from $S_h N S_h^T$ matrix on binned data, generated from a truncated exponential with $\lambda = 1$ between 0 and 5. Degree is fixed at 3.

n_1. Its computation may fail (Figure A.1) if one or more of the eigenvalues is 0 or close to 0, which can be the case for the covariance matrix (Figure A.3).

$$n_2 = \frac{(\sum_{j=1}^K \lambda_j)^2}{\sum_{j=1}^K \lambda_j^2}.$$

- n_∞ is a conservative estimator (note that it corresponds to the trace-based effective rank):

$$n_\infty = \frac{\sum_{j=1}^K \lambda_j}{\max_j \lambda_j};$$

- n_C systematically overestimate the effective dimensionality and is thus not recommended:

$$n_C = K - \frac{K^2}{(\sum_{j=1}^K \lambda_j)^2} Var(\lambda).$$

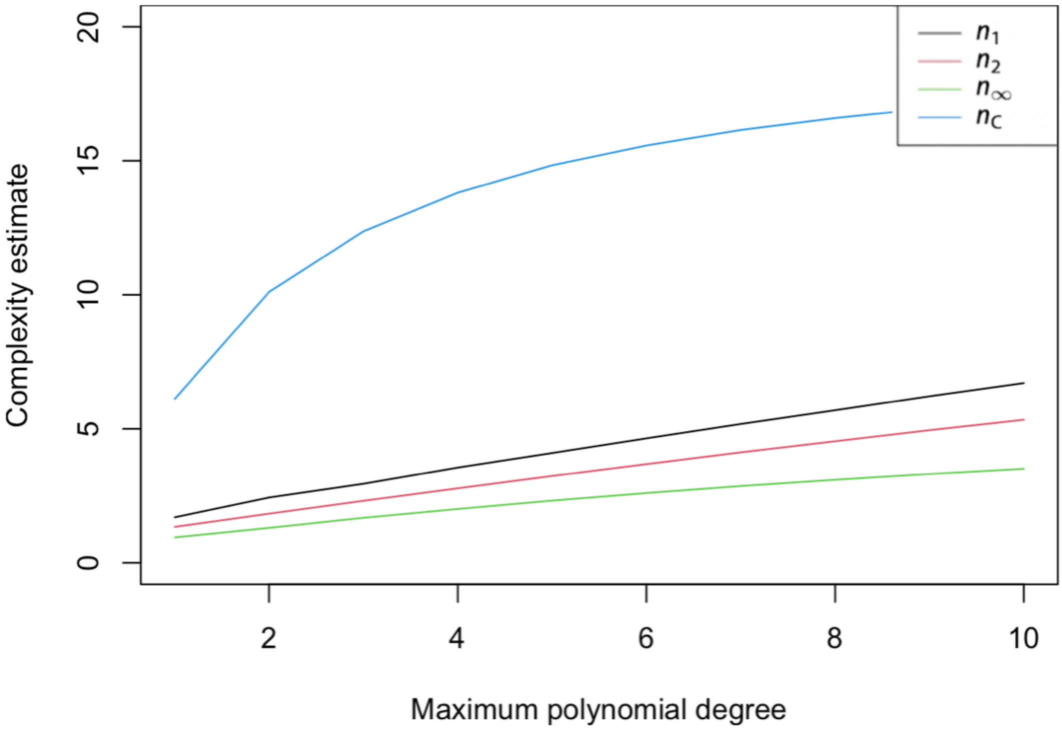

Figure A.2: Dependence of n_1 (black), n_2 (red), n_∞ (green), n_C (blue) on the degree. Computed from $S_h N S_h^T$ matrix on binned data, generated from a truncated exponential with $\lambda = 1$ between 0 and 5. Bandwidth is fixed at 3.

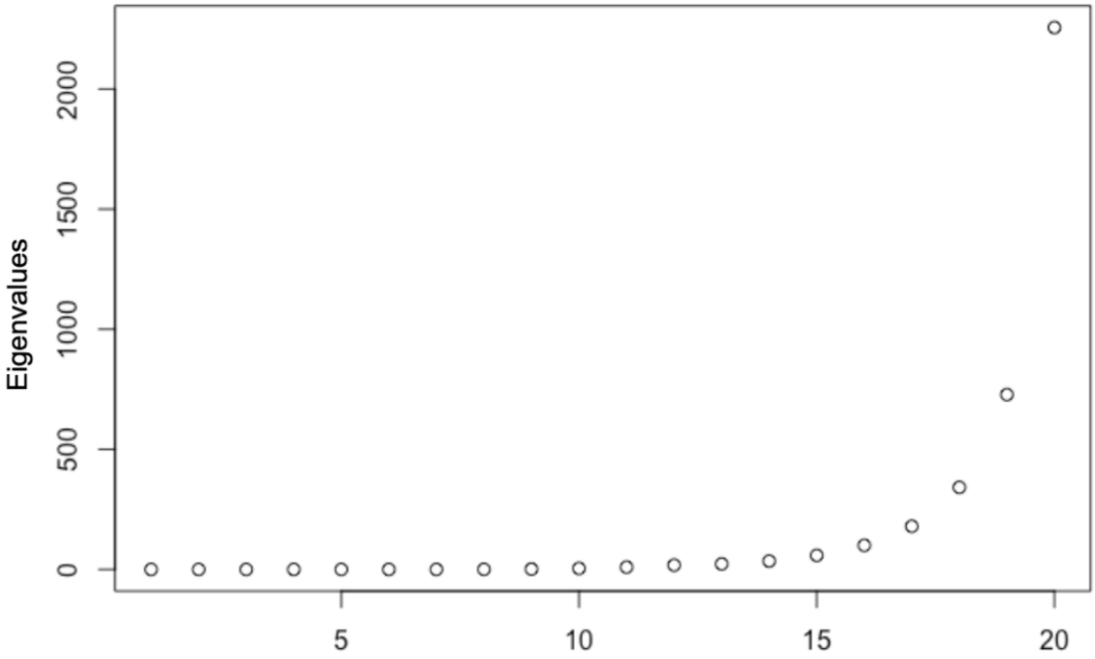

Figure A.3: Eigenvalues of the $S_h N S_h^T$ covariance matrix of the smoothed densities, in the binned case (20 bins), from smallest to largest. It can be seen that many of these values are very close to 0.

Appendix B

Effective degrees of freedom localization

Different definitions of effective degrees of freedom were considered, and so were different ways of introducing localization.

For df_{ratio} (see Chapter 5), a the following weight function was proposed: The weight w_i, corresponding to the $i-th$ bin is a Gaussian function of the distance between the the bin and the region of interest:

$$w_i = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{d_i^2}{2\sigma^2}\right\},$$

where $\sigma > 0$ is the standard deviation of the Gaussian and d_i is the euclidean distance from the center of the bin to the center of the region of interest.

The localized version of the estimator is thus

$$ldf_{ratio} = \sum_i^K w_i \frac{\sigma_{a,i}^2}{\sigma_{b,i}^2},$$

where $\sigma_{a,i}^2$ and $\sigma_{b,i}^2$ are, respectively, the variances in the *i-th* bin after and before smoothing.

With the expected and observed variances:

$$ldf_P = \sum_i^K w_i \frac{[S_h N S_h^T]_{ii}}{N_{ii}};$$

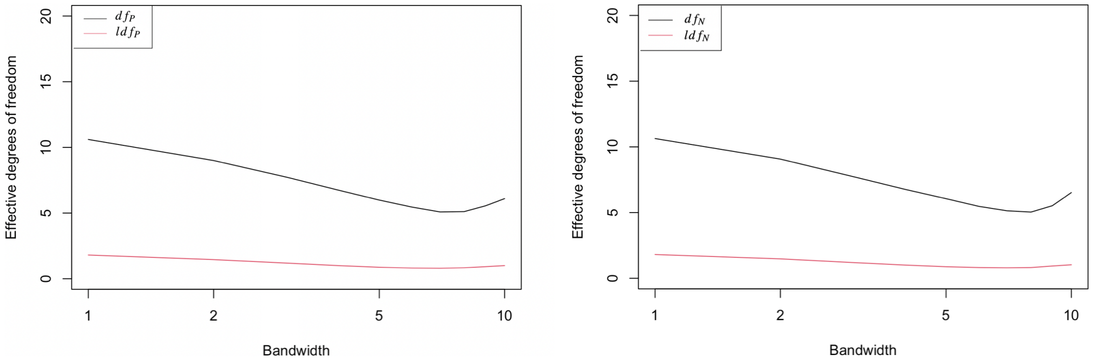

Figure B.1: Dependence of df_P, ldf_P (left), df_N and ldf_N (right) on the bandwidth, for an exponential with $\lambda = 1$, truncated between 0 and 5. The degree is fixed at 3.

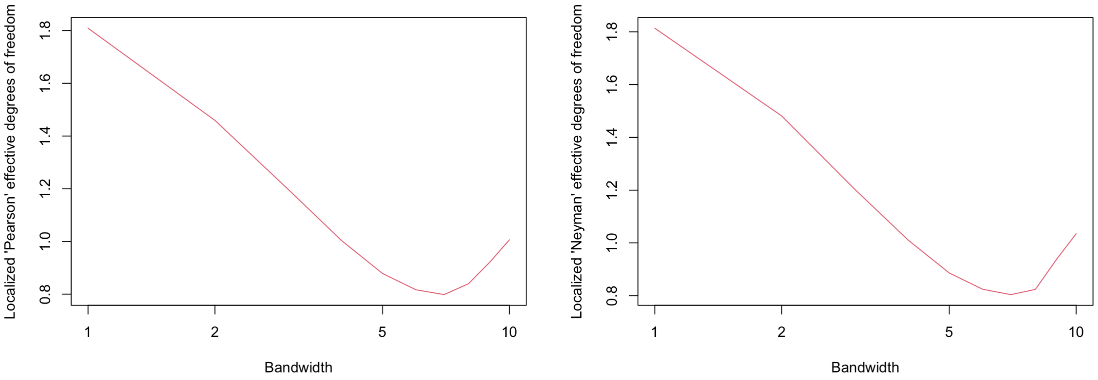

Figure B.2: Dependence of ldf_P (left) and ldf_N (right) on the bandwidth, for an exponential with $\lambda = 1$, truncated between 0 and 5. The degree is fixed at 3.

$$ldf_N = \sum_{i}^{K} w_i \frac{[S_h \hat{N} S_h^T]_{ii}}{\hat{N}_{ii}},$$

The estimator, for both the "Neyman" and "Pearson" versions, meets the expectation for the localized degrees of freedom to be less than the global ones (Figure B.1). However, it also presents the effect described in Chapter 5, Figure B.2.

Another option for localization was creating a diagonal matrix W of weights, where each diagonal element W_{ii} is w_i. The weights are applied to the covariance matrix C through the transformation $C_W = W^{1/2} C W^{1/2}$. Figure B.3 offers a visualization of the C (left) and the C_W (right) matrices. The C matrix used for the figures is the analytical form, for binned data. The effective rank measures of effective degrees of freedom obtained from this matrix are, for large numbers,

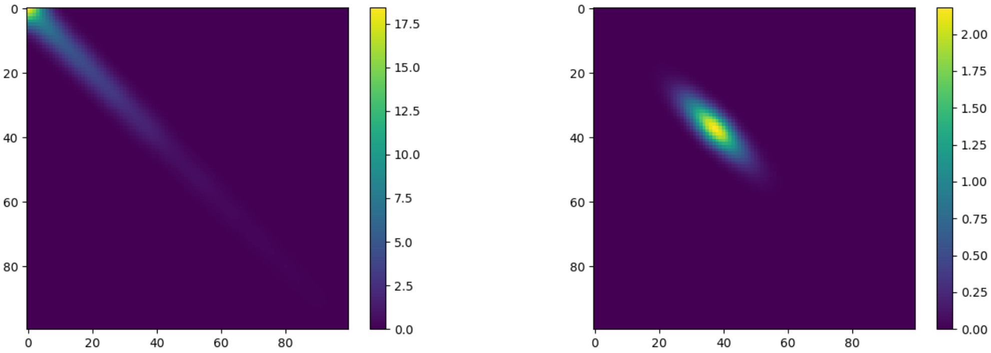

Figure B.3: Covariance matrix before (left) and after applying localizing weights (right), computed from binned data (100 bins). Horizontal axis indicates columns, vertical axis indicates rows.

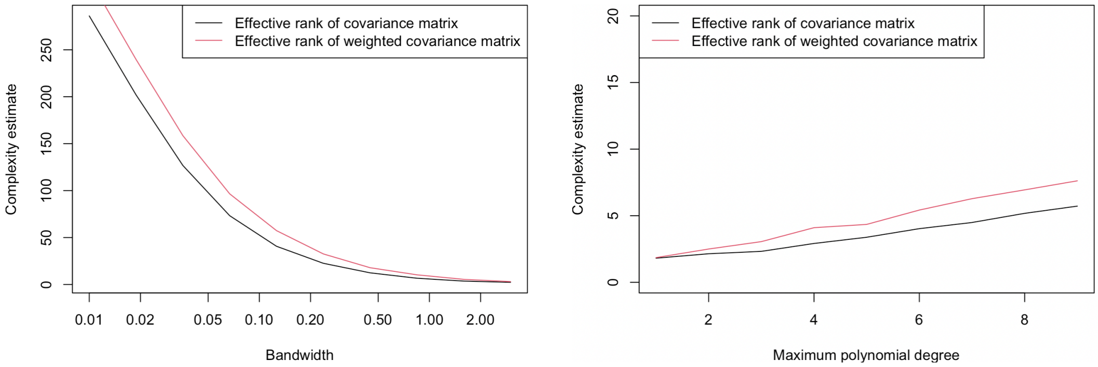

Figure B.4: Dependence of the entropy-based effective rank of the covariance matrix (black) and the same measure computed on the weighted covariance matrix (red) on the bandwidth (left) and degree (right), for an exponential with $\lambda = 1$, truncated between 0 and 5.

equivalent to the ones obtained from the bootstrap estimate for unbinned (or binned) data. Note that these measures also do not depend on the normalization of the weights, as it can be written as a multiplicative constant, which does not affect the computation of the entropy-based effective rank [36], or the trace-based one (all eigenvalues are also modified by a multiplicative factor which simplifies in the computation in Equation 5.1).

However, this method of localizing the covariance matrix does not result in a measure that represents local complexity. As shown in Figure B.4, this local measure is not smaller than the global one.

Appendix C

Plots of the effective degrees of freedom

Figures C.1 - C.7 show the trace-based and entropy-based effective rank of the bootstrap covariance matrix, in their global (non weighted) version and local (weighted) version, and the distributions used to generate the data for each plot. Although in practice the bandwidth is chosen via (localized) cross-validation, in each of these plots it varies in the same interval for each plot, in order to see the dependence of these measures as not only the bandwidth varies, but also the polynomial degree.

A sample size of 500 has been used for each of the simulated samples and 500 bootstrap iterations for the computation of the covariance matrix of the estimated density. It has been verified that a higher number of sample points or bootstrap iterations and the use of resampling rather than toy samples (generated from the original density), do not notably improve or otherwise change the estimates.

In all these situations, it can be seen that the local degrees of freedom are, as expected, a smaller number than the global ones.

Note that, since the bandwidth ranges here are not chosen on the basis of cross-validation, it is not useful to compare the effective degrees of freedom for the different distributions. For example, if uniform data were used in practice, the bandwidth chosen would be much larger, as the distribution is completely

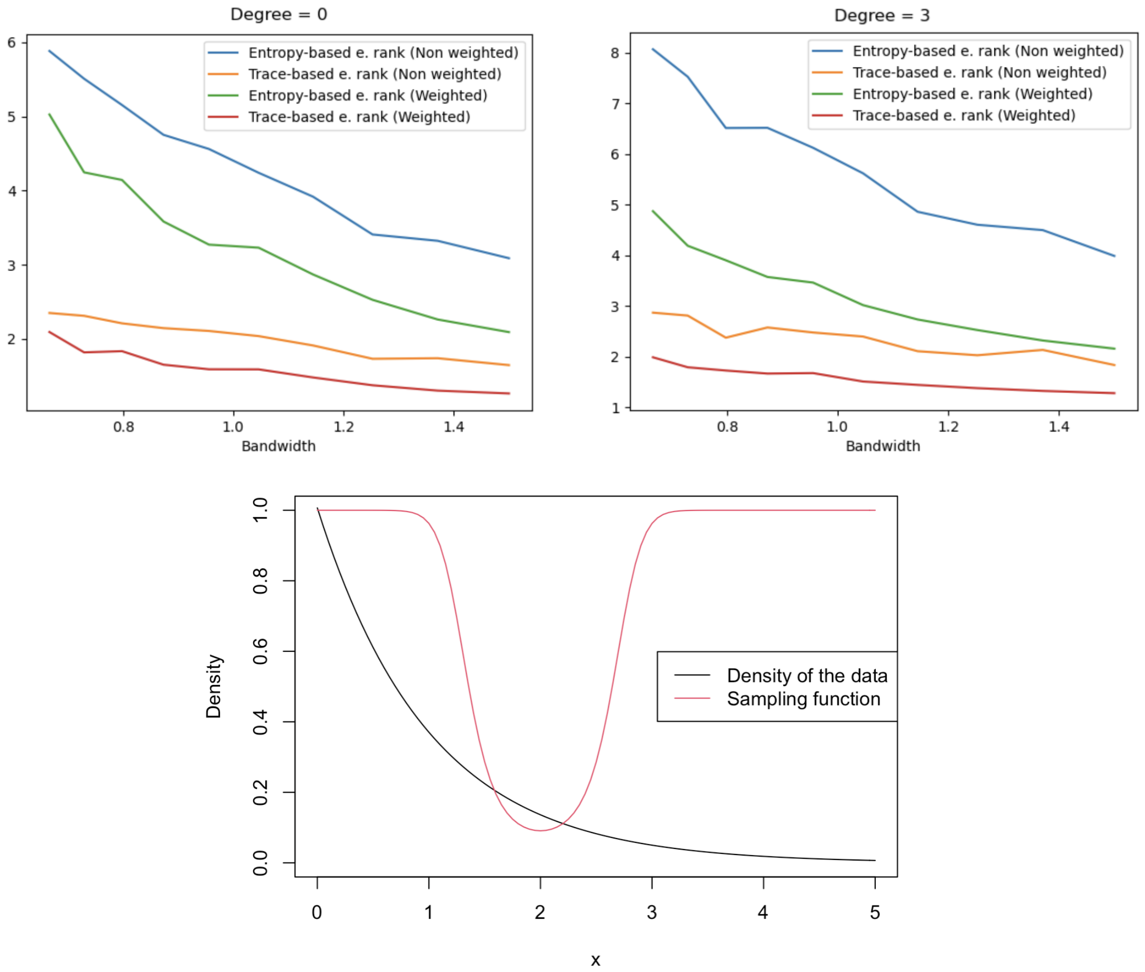

Figure C.1: Dependence of the entropy-based (global version in blue, local version in green) and trace-based (global version in orange, local version in red) effective rank of the bootstrap covariance matrix on the bandwidth, with maximum degree of the polynomials equal to 0 (top left) and 3 (top right). Data generated from a truncated exponential with $\lambda = 1$, localization through a Gaussian dip function with mean 2 and standard deviation 0.3 (bottom).

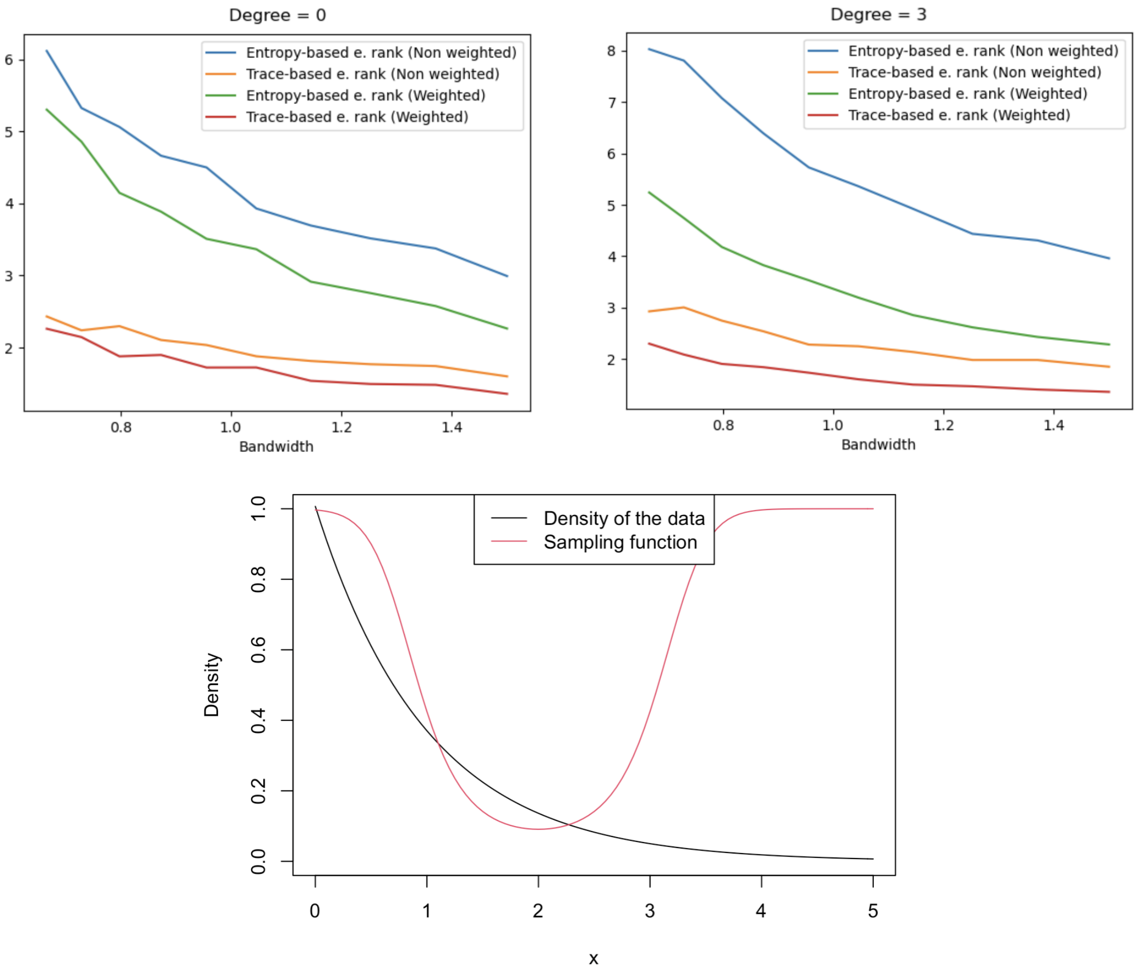

Figure C.2: Dependence of the entropy-based (global version in blue, local version in green) and trace-based (global version in orange, local version in red) effective rank of the bootstrap covariance matrix on the bandwidth, with maximum degree of the polynomials equal to 0 (top left) and 3 (top right). Data generated from a truncated exponential with $\lambda = 1$, localization through a Gaussian dip function with mean 2 and standard deviation 0.5 (bottom).

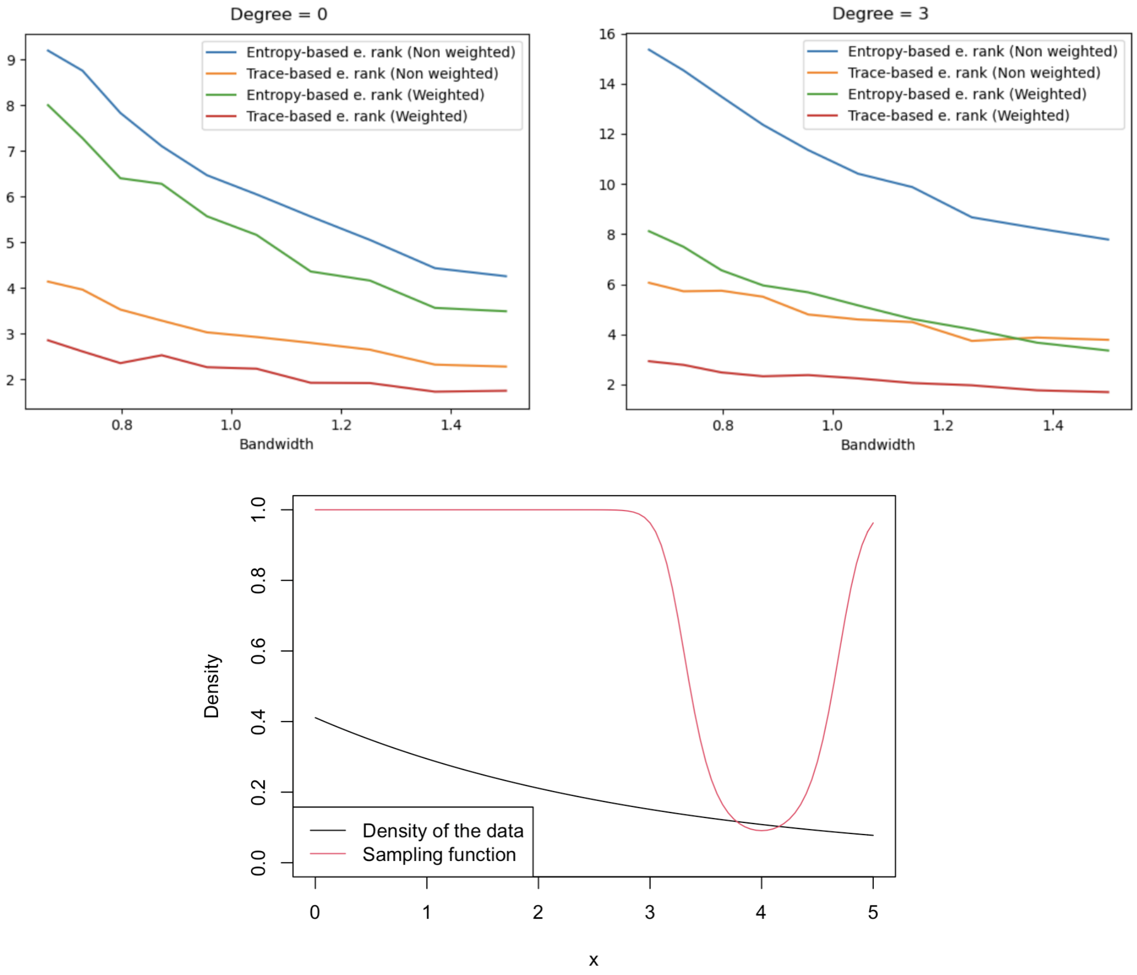

Figure C.3: Dependence of the entropy-based (global version in blue, local version in green) and trace-based (global version in orange, local version in red) effective rank of the bootstrap covariance matrix on the bandwidth, with maximum degree of the polynomials equal to 0 (top left) and 3 (top right). Data generated from a truncated exponential with $\lambda = 3$, localization through a Gaussian dip function with mean 4 and standard deviation 0.3 (bottom).

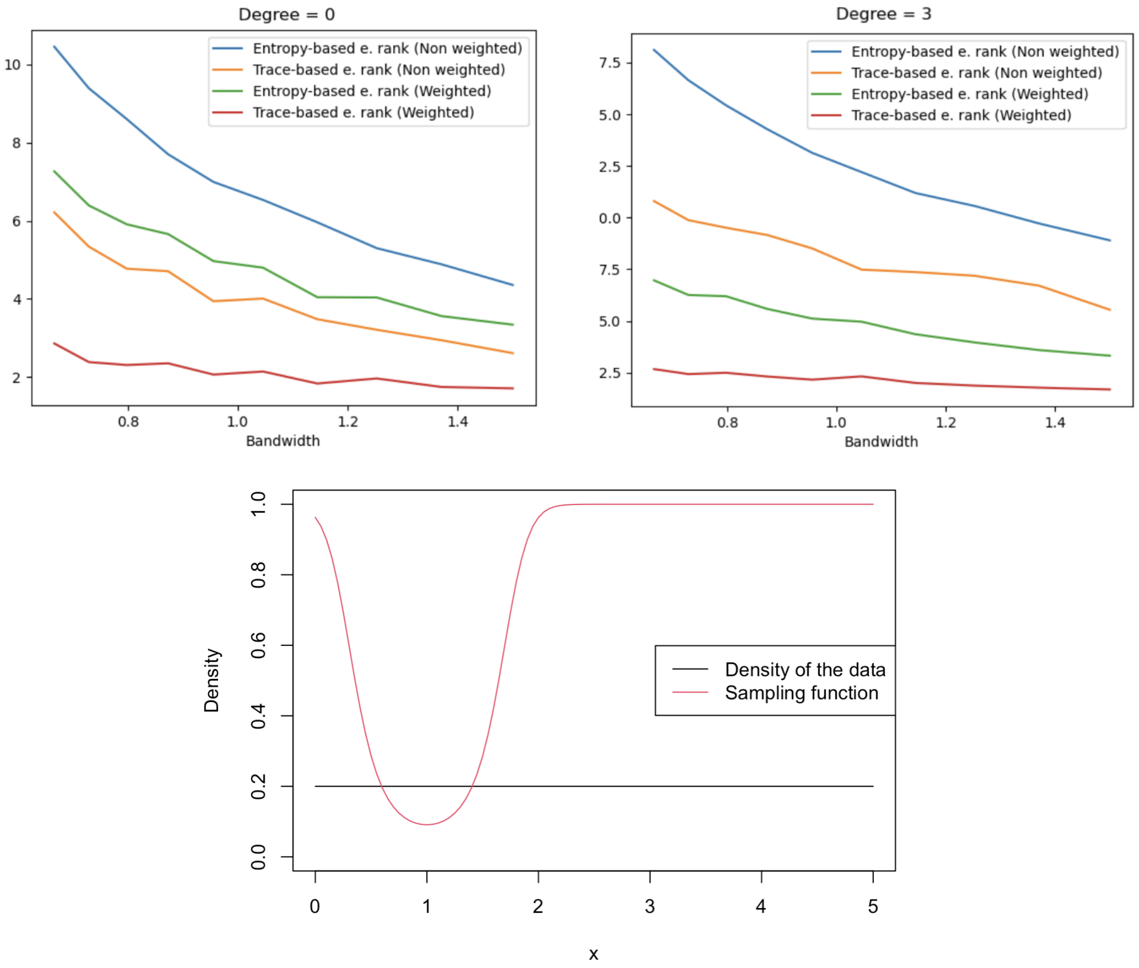

Figure C.4: Dependence of the entropy-based (global version in blue, local version in green) and trace-based (global version in orange, local version in red) effective rank of the bootstrap covariance matrix on the bandwidth, with maximum degree of the polynomials equal to 0 (top left) and 3 (top right). Data generated from a uniform between 0 and 5, localization through a Gaussian dip function with mean 1 and standard deviation 0.3 (bottom).

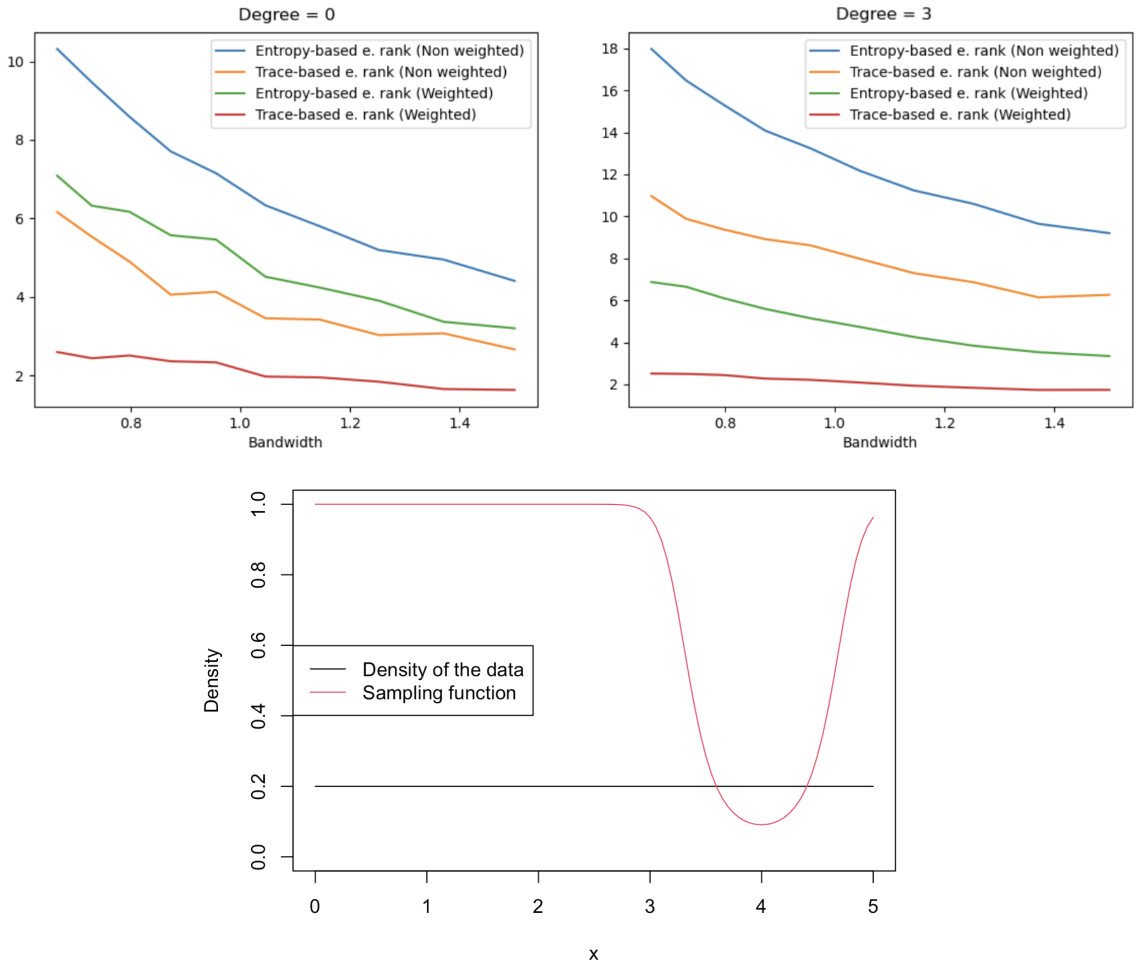

Figure C.5: Dependence of the entropy-based (global version in blue, local version in green) and trace-based (global version in orange, local version in red) effective rank of the bootstrap covariance matrix on the bandwidth, with maximum degree of the polynomials equal to 0 (top left) and 3 (top right). Data generated from a uniform between 0 and 5, localization through a Gaussian dip function with mean 4 and standard deviation 0.3 (bottom).

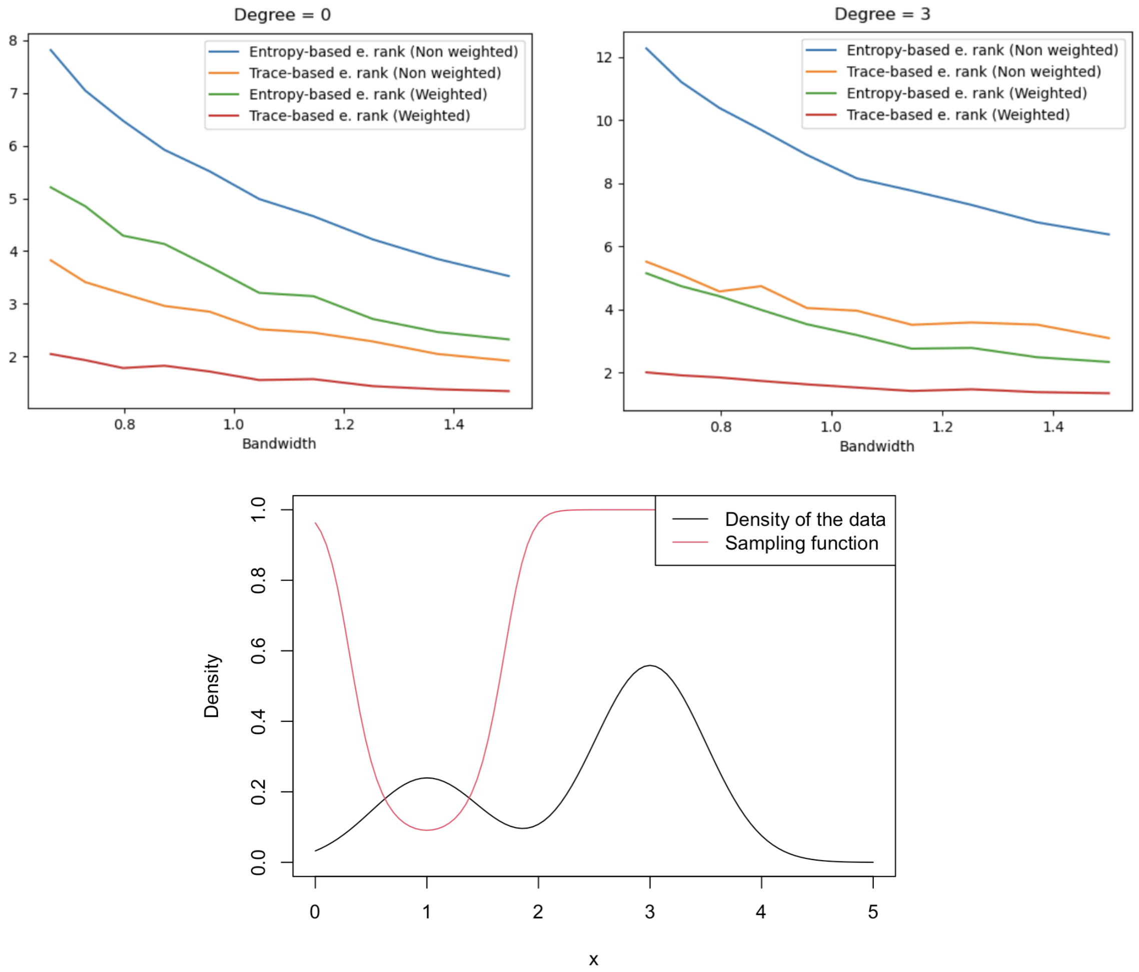

Figure C.6: Dependence of the entropy-based (global version in blue, local version in green) and trace-based (global version in orange, local version in red) effective rank of the bootstrap covariance matrix on the bandwidth, with maximum degree of the polynomials equal to 0 (top left) and 3 (top right). Data generated from a Gaussian mixture with means 1 and 3, standard deviations 0.5, mixing proportions 0.3 and 0.7, localization through a Gaussian dip function with mean 1 and standard deviation 0.3 (bottom).

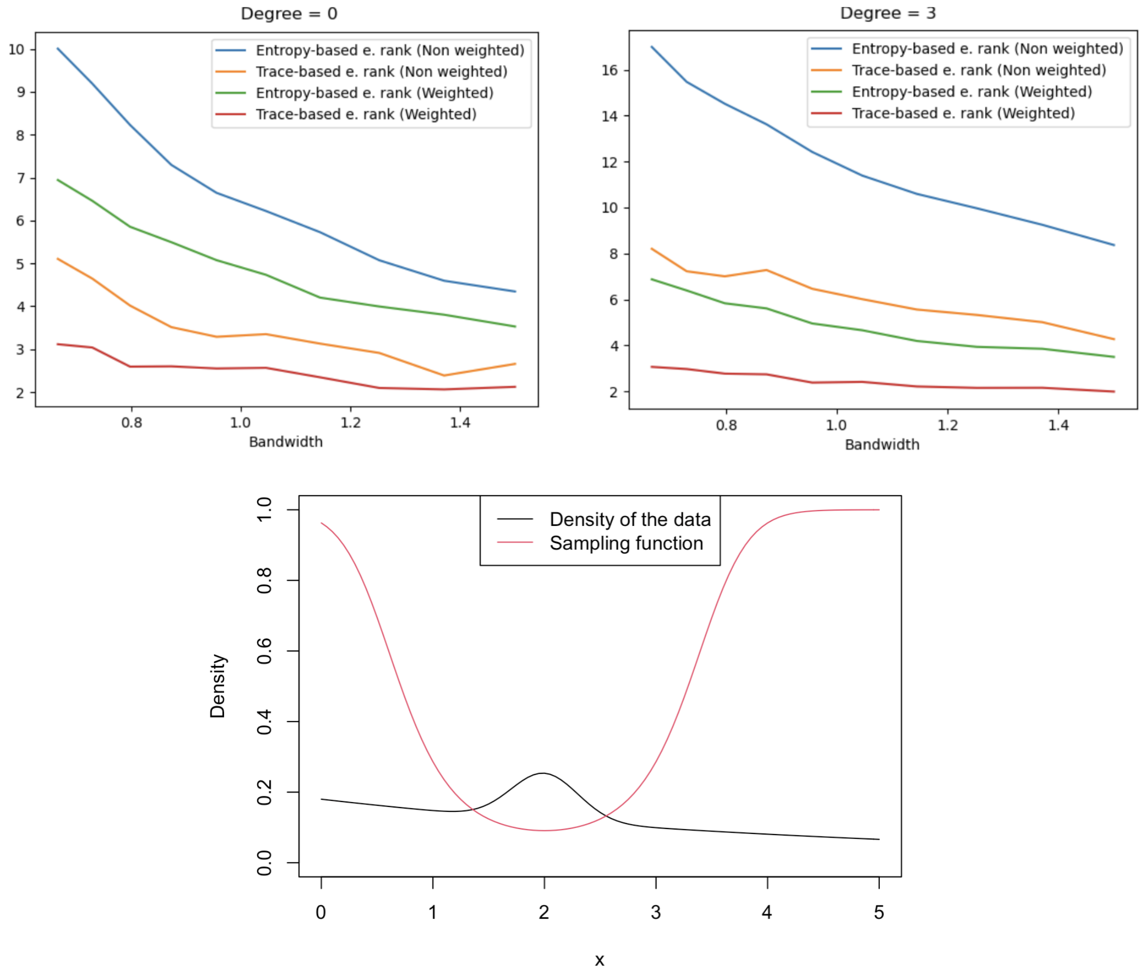

Figure C.7: Dependence of the entropy-based (global version in blue, local version in green) and trace-based (global version in orange, local version in red) effective rank of the bootstrap covariance matrix on the bandwidth, with maximum degree of the polynomials equal to 0 (top left) and 3 (top right). Data generated from a mixture of a truncated exponential with $\lambda = 5$ and a normal with with mean 2 and standard deviation 0.3, mixing proportions 0.9 and 0.1, localization through a Gaussian dip function with mean 2 and standard deviation 0.6 (bottom).

smooth, and thus the effective degrees of freedom would be less. However, fixing the bandwidth range allows to compare, for each distribution, the different polynomial degrees and verify that with a higher degree (higher complexity), the effective degrees of freedom are in fact a larger number, the bandwidth being equal.