Density estimation using multiscale local polynomial transforms

Maarten Jansen
Université libre de Bruxelles, department of Mathematics

November 2023

Abstract

The multiscale local polynomial transform (MLPT) is a combination of a kernel method for nonparametric regression or density estimation with a projection onto a basis in a multiscale framework. The MLPT is proposed for the estimation of densities with possibly one or more singular points at unknown locations. The proposed estimator reformulates the density estimation problem as a high-dimensional, sparse regression problem with asymptotically exponential response variables. The covariates in this model are the observations from the unknown density themselves. The design matrix comes from a novel extension of the MLPT for use on highly nonequidistant data.

keywordssparsity nonparametric local polynomial variable selection wavelets multiscale

1 Introduction

This paper incorporates the local polynomial smoothing method into a wavelet-like multiscale decomposition, termed the Multiscale Local Polynomial Transform (MLPT) [Jansen and Amghar, 2017]. The contribution of this paper lies in the application of the MLPT in univariate density estimation. The density estimation problem is recognised as a naturally mutiscale problem, which in this paper is transformed into a sparse, multiscale regression problem with exponentially distributed responses.

MLPT density estimation combines two classes of methods in nonparametric density estimation. The first class is based on the use of a kernel function, $K(x)$, which is mostly a continuous, symmetric, nonnegative, unimodal function [Wand and Jones, 1995, Fan and Gijbels, 1996, Simonoff, 1996] The prototype in this class of methods is the kernel density estimator

$$\widehat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

estimating the unknown density $f_X(x)$ from the iid sample $(X_1, X_2, \ldots, X_n)$. Expression (1) contains the bandwidth h as a smoothing parameter, whose value is often finetuned in an optimisation of the bias-variance trade-off. The method is fast and it requires few assumptions on $f_X(x)$ in the general, regular settings. Problems occur, however, if $f_X(x)$ has discontinuities, singularities or bounded support. For instance, if $\lim_{x\to x_0} f_X(x) = \infty$ for some finite x_0, then the estimator $\widehat{f}_X(x)$, being based on a bounded kernel function, cannot possibly reproduce such singularity, thereby leading to a bias and an inconsistency. If the position of x_0 is known, then a transformation of the random variable X and the observations X_i may induce a regular density to be estimated. Transformations in kernel density estimations have been the topic of quite some contributions [Wand et al., 1991, Park et al., 1992, Yang and Marron, 1999, Ruppert and Cline, 1994, Hossjer and Ruppert, 1995]. For densities with bounded support, such as copulas, bias occurs near the boundaries because the plain kernel density estimator does not sense the end points, and thus puts a positive value for the density outside the support. Knowing the values of the end points a and b, it is possible to transform the observations using for instance the probit function $T = \Phi^{-1}((x-a)/(b-a))$ in order to obtain an unbounded random variable. It has been reported that the estimation procedure on the transformed observations should be designed with the inverse transformation in mind. In particular, a global bandwidth for the transformed data does not perform optimally with respect to the original domain [Geenens, 2014].

The second class of methods in nonparametric density estimation is based on the projection of $f_X(x)$ onto a space generated by a set of basis functions,

$$f_j(x) = \sum_{k=1}^{n_j} s_{j,k}\varphi_{j,k}(x). \tag{2}$$

In (2), the functions $\varphi_{j,k}(x)$ are assumed to have a bounded quadratic norm and to be linearly independent, thus generating a subspace $\mathcal{V}_j \subset L_2(\mathbb{R})$. The index j refers to a possible refinement of the approximation, aiming of course at convergence for $j \to \infty$. The projection generating the coefficients is written as $s_{j,k} = \int_{-\infty}^{\infty} f_X(x)\widetilde{\varphi}_{j,k}(x)dx$, where the dual basis functions $\widetilde{\varphi}_{j,k}(x)$ satisfy the biorthogonality condition $\int_{-\infty}^{\infty} \widetilde{\varphi}_{j,k}(x)\varphi_{j,l}(x)dx = \delta_{k,l}$, with $\delta_{k,l}$ the Kronecker-delta. Biorthogonality makes the mapping of $f_X(x)$ onto $f_j(x)$ a projection, i.e., an idempotent mapping, meaning that $f_j(x)$ is projected onto itself. The dual basis and the corresponding projection is not unique. As an example, the orthogonal projection is found by $\widetilde{\varphi}_{j,k}(x) = \sum_{l=0}^{n_j-1} c_{j;k,l}\varphi_{j,l}(x)$, with $c_{j;k,l}$ the elements of the symmetric matrix C_j defined by its inverse $\left(C_j^{-1}\right)_{k,l} = \int_{-\infty}^{\infty} \varphi_{j,k}(x)\varphi_{j,l}(x)dx$. As $f_X(x)$ is the density of the observations, the projection coefficients can be estimated unbiasedly by

$$\widehat{s}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\varphi}_{j,k}(X_i). \tag{3}$$

The only source of bias in the density estimator $\widehat{f}_X(x) = \sum_{k=1}^{n_j} \widehat{s}_{j,k}\varphi_{j,k}(x)$ comes from the approximation error $f_j(x) - f_X(x)$. Points of attention for projection methods include the fact that the basis functions are defined independently from the observations. This is in contrast to kernel based methods, where the kernel functions are centered around the observations. With fixed, non data-adaptive basis functions, subtle information about the precise location of the observations is lost, once the observations have been used for evaluation in (3). Another point of attention is the positivity of the estimator $\widehat{f}_X(x)$, which needs to be enforced explicitly. An important representative of this class of methods uses wavelets as basis functions [Hall and Patil, 1995, Donoho et al., 1996]. It is well known that wavelets are of particular interest when the function to be estimated has singularities. Nonlinear wavelet methods perform well in this case, thanks to their ability to sparsely represent data with jumps. In the case of density estimation, a wavelet transform may thus be an alternative to a data transformation when the density has singular points, especially when the precise location of the singular points is unknown. A second argument for using wavelets holds also for smooth density functions. This argument focuses less on the sparsity of the wavelet representation, and more on its multiscale nature. For obvious reasons, the density of observations depends directly on the function to be estimated. As a result, the support of the analysis functions in (3) or the bandwidth in the kernel approach (1) should be optimised in a local way. In intervals of high density, the supports should be small enough to avoid unnecessary bias in the estimations. In intervals of low density, the supports should be large, in order to gather enough observations in each sum of (3). Both the support and the bandwidth represent the scale of the operation. In the estimation of a density from its observations, the scale of the problem is an intermittent parameter, making density estimation a natural multiscale problem.

Besides the loss of detailed information and the positivity issues in a projection method, the wavelet density estimation also has a specific problem, namely the evaluation of $\widetilde{\varphi}_{j,k}(X_i)$. In the course of the fast wavelet transform algorithm, used in wavelet based nonparametric regression, the basis functions $\widetilde{\varphi}_{j,k}(x)$ are not evaluated explicitly. Moreover, in most cases, no closed form is available. An explicit evaluation in an arbitrary point as in (3) leads to a suboptimal numerical computations, as the fast wavelet transform operates on dyadic (i.e., of the form $k2^{-j}$, with $k = 0, 1, \ldots, 2^j - 1$), not on arbitrary knots.

This paper is organised as follows. **Section 2** reviews the Multiscale Local Polynomial Transform. When used on covariate values from uniform densities, this data decomposition is slightly overcomplete. On highly irregular grids, with non uniformly distributed covariate values, the redundancy may get out of control. Therefore, **Subsection 2.2** introduces a new variant of the Multiscale Local Polynomial Transform that keeps the redundancy under control for application on intermittent densities of knots. **Section 3** uses the classical techniques from wavelet theory to find the scaling basis functions. Theorem 1 states that polynomials can be reconstructed in a Multiscale Local Polynomial basis, using the function values as coefficients. Thanks to this property the representation of any piecewise smooth function in a Multiscale Local Polynomial basis may use function values as coefficients. This way, the decomposition into a basis leads to expressions that are similar to those of Nadaraya-Watson or Gasser-Müller kernel smoothing. Such a property is not generally available in wavelet decompositions on irregular knots, the B-spline wavelets [Jansen, 2016] being an illustrative example.

Section 4 develops the density estimation within the MLPT framework, outined in **Subsection 4.1**. In **Subsection 4.2** first reformulates the density estimation as a nonparametric regression problem with exponentially distributed responses. Next, in **Subsection 4.3**, the regression problem is formulated in the sparse multiscale framework of the MLPT. As a fast forward and inverse map onto the sparse representation is available, a threshold procedure is proposed. In **Subsection 4.4**, the choice of the threshold through the optimisation of an information criterion is discussed.

A simulation study and a real data illustration follow in **Sections 5 and 6**. Finally, the concluding discussion in Section 7 lays the path for future analysis of the proposed method. In particular, it explains that the proposed work cannot be considered as a competitor, but rather as a complement to classical kernel density estimation, because the domains of application do not overlap. Based on this observation, the assignments for future asymptotic analysis are outlined.

2 The Multiscale Local Polynomial Transform

2.1 MLPT construction on uniformly distributed knots

In the first instance, the Multiscale Local Polynomial Transform (MLPT) is a slightly overcomplete decomposition of approximations $f_J(x)$ defined in (2), taking $j = J$, where J is the index referring to the maximal refinement of the approximation. In practice, this finest resolution corresponds to the resolution at which the observations take place. The numerical value of J is a matter of choice or convention. With n the number of observations, the finest scale is typically referred to by $J = \lceil \log_2(n) \rceil$, the smallest integer equal or larger than $\log_2(n)$. The input vector $\boldsymbol{s}_J$ contains the coefficients of the approximation in (2). The basis functions $\varphi_{J,k}(x)$ are local in the sense that they are centered around knots $x_{J,k}$, in a way that is developed further below. In the subsequent discussion, the knots will coincide with the covariates of the observations. The number of knots at the finest resolution, denoted by n_J, equals n. The vector $\boldsymbol{s}_J$ will be obtained from the response values in the knots, $f_J(x_{J,k})$, typically observed with noise, i.e., $Y_k = f_J(x_{J,k}) + \sigma Z_k$, where the precise model for the noise Z_k is not important in this discussion. The result in Theorem 1 will justify to simply take $s_{J,k} = Y_k$.

The algorithm for the actual MLPT takes the input vector $\boldsymbol{s}_J$ to be transformed into an equivalent, multiresolution decomposition, denoted by the vector $\boldsymbol{v}_L$, where L stands for the coarsest or lowest resolution level. In order to arrive at resolution level L, the MLPT iterates over the resolution levels $j = J-1, J-2, \ldots, L$. In each step of its simplest form, the covariate vector is first dyadically subsampled, meaning that the next, coarser level grid of covariate values consists of the even indexed fine level grid, i.e., $x_{j,k} = x_{j+1,2k}$, for $k = 0, 1, \ldots, n_j$. The subsampled vector then has length $n_j = \lceil n_{j+1}/2 \rceil$. More sophisticated versions may adopt other than dyadic subsampling schemes, defining the non-dyadic partition $e(j+1)$ and $o(j+1)$ of the fine scale index set $\{1, 2, \ldots, n_{j-1}\}$. Non-dyadic subsampling is particularly interesting on non-uniformly distributed covariate values. In the extreme case, the partition splits off only one element of $\boldsymbol{x}_{j+1}$, i.e., $o(j+1)$ is a singleton. Although no longer corresponding to the classical mathematical notions of even and odd, $e(j+1)$ and $o(j+1)$ will be referred to as the even and odd subsets even in non-dyadic subsampling. The corresponding subvectors are denoted as $\boldsymbol{x}_{j+1,e}$ and $\boldsymbol{x}_{j+1,o}$. The odd points will also be termed refinement points, while the even points are said to be the coarse scale grid. The resulting multiscale grid $\{\boldsymbol{x}_J, \boldsymbol{x}_{J-1}, \ldots, \boldsymbol{x}_L\}$ is a nested grid, meaning that all knots at level j belong to the set of knots at finer level $j+1$. The subsampling operation is denoted by $\boldsymbol{x}_j = \boldsymbol{x}_{j+1,e} = \widetilde{\mathbf{J}}_j \boldsymbol{x}_{j+1}$, with $\widetilde{\mathbf{J}}_j$ the $n_j \times n_{j+1}$ subsampling matrix, formed by taking all even rows of the $n_{j+1} \times n_{j+1}$ identity matrix.

The vector $\boldsymbol{s}_{j+1}$ is subsampled as well, being filtered at the occasion, using a $n_j \times n_{j+1}$ prefilter $\widetilde{\mathbf{F}}_j$ in $\widetilde{\boldsymbol{s}}_j = \widetilde{\mathbf{F}}_j \boldsymbol{s}_{j+1}$. The precise design of $\widetilde{\mathbf{F}}_j$ is closely connected to the choice of the forthcoming matrix $\mathbf{U}_j$. Along with a third matrix $\mathbf{P}_j$, these matrices fix the properties of the MLPT. The vector $\widetilde{\boldsymbol{s}}_j$ can be interpreted as a coarse scale approximation of $\boldsymbol{s}_{j+1}$. The offset between the fine scale vector and a prediction based on the coarse scale approximation is the detail vector $\boldsymbol{d}_j = \boldsymbol{s}_{j+1} - \mathbf{P}_j \widetilde{\boldsymbol{s}}_j$, which will be stored as part of $\boldsymbol{v}_L$. Finally, the coarse scale approximation is updated by $\boldsymbol{s}_j = \widetilde{\boldsymbol{s}}_j + \mathbf{U}_j \boldsymbol{d}_j$, for use as input in the next iteration step of the multiscale decomposition.

The design of the $n_j \times n_{j+1}$ update matrix $\mathbf{U}_j$ follows in Section 3.1. In data smoothing [Jansen and Amghar, 2017], the update can be omitted because the prefilter $\widetilde{\mathbf{F}}_j$ offers nearly the same benefits. As explained in Section 3.1, however, in density estimation, the update is necessary to ensure that the estimated density integrates to one. As a result, no prefilter is needed in this context, so we set $\widetilde{\mathbf{F}}_j = \widetilde{\mathbf{J}}_j$ for the remainder of this article. The $n_{j+1} \times n_j$ prediction matrix $\mathbf{P}_j$ performs a local polynomial estimation in $\boldsymbol{x}_{j+1}$ based on the values in $\widetilde{\boldsymbol{s}}_j$ and the covariates in $\boldsymbol{x}_j$. This matrix is defined as follows.

Definition 1 *The local polynomial prediction matrix* $\mathbf{P}_j$ *at level* j *in a Multiscale Local Polynomial*

Transform of order $\widetilde{p}$ has entries given by $P_{j;k,l} = P_{j,l}(x_{j+1,k}; \boldsymbol{x}_j)$ where

$$P_{j,l}(x; \boldsymbol{x}_j) = \mathrm{X}^{(\widetilde{p})}(x) \left(\mathbf{X}_j^{(\widetilde{p})^T} \mathbf{W}_j(x) \mathbf{X}_j^{(\widetilde{p})} \right)^{-1} \left(\mathbf{X}_j^{(\widetilde{p})^T} \mathrm{W}_{j;l,l}(x) \right). \tag{4}$$

In (4), $\mathrm{X}^{(\widetilde{p})}(x)$ is a row matrix of power functions, $\mathrm{X}^{(\widetilde{p})}(x) = [1\, x\, \ldots\, x^{\widetilde{p}-1}]$. The $n_j \times \widetilde{p}$ matrix $\mathbf{X}_j^{(\widetilde{p})}$ has elements $\mathbf{X}_{j;k,r}^{(\widetilde{p})} = x_{j,k}^{r-1}$. The $n_j \times n_j$ weight matrix $\mathbf{W}_j(x)$ has a diagonal structure with elements $\mathrm{W}_{j;l,l}(x) = K\left(\frac{x-x_{j,l}}{h_j}\right)$. Here, the function $K(x)$ is the kernel function and h_j is the bandwidth at resolution level j.
The order $\widetilde{p}$ is also termed the number of dual vanishing moments.

The output is composed as $\boldsymbol{v}_L = \left[\begin{array}{ccccc} \boldsymbol{s}_L & \boldsymbol{d}_L & \boldsymbol{d}_{L+1} & \ldots & \boldsymbol{d}_{J-1} \end{array} \right]$. The output has length $n_L + \sum_{j=L+1}^{J} n_j = \mathcal{O}(2n_J)$. The transform thus expands a vector of size n_J into a vector of twice that size, following a scheme known in other applications and with other predictions as a Laplacian pyramid [Burt and Adelson, 1983]. As the transform is overcomplete, the inverse transform is not unique. A straightforward reconstruction first undoes the update $\widetilde{\boldsymbol{s}}_j = \boldsymbol{s}_j - \mathbf{U}_j \boldsymbol{d}_j$ and then the prediction $\boldsymbol{s}_{j+1} = \mathbf{P}_j \widetilde{\boldsymbol{s}}_j + \boldsymbol{d}_j$. The two steps can be assembled into the reconstruction formula

$$\boldsymbol{s}_{j+1} = (\mathbf{I}_{j+1} - \mathbf{P}_j \mathbf{U}_j)\boldsymbol{d}_j + \mathbf{P}_j \boldsymbol{s}_j. \tag{5}$$

The reconstruction in (5) does not depend on the prefilter $\widetilde{\mathbf{F}}_j$. As a result, the design of the prefilter does not need to take any effect from the reconstruction into account. In particular, a prefilter can be designed without bothering about variance propagation in the reconstruction. This would be one of the benefits of a prefilter above an update, were it not for the disadvantages in density estimation, developed in Section 3.1.

2.2 An alternative reconstruction

This paper proposes a more advanced reconstruction, referred to as weighted reconstruction,

$$\boldsymbol{s}_{j+1} = \mathbf{Q}_{j+1} (\mathbf{P}_j \widetilde{\boldsymbol{s}}_j + \boldsymbol{d}_j) + (\mathbf{I}_{j+1} - \mathbf{Q}_{j+1}) \widetilde{\mathbf{J}}_j^T \widetilde{\boldsymbol{s}}_j. \tag{6}$$

The elements of $n_{j+1} \times n_{j+1}$ matrix $\mathbf{Q}_{j+1}$ are taken to be between 0 and 1, thus making (6) a weighted average between two reconstructions. The first reconstruction, $\mathbf{P}_j \widetilde{\boldsymbol{s}}_j + \boldsymbol{d}_j$, is the one used in (5). The second one is a simple upsampling $\widetilde{\mathbf{J}}_j^T \widetilde{\boldsymbol{s}}_j$, i.e., starting from the vector $\widetilde{\boldsymbol{s}}_j$, it inserts zeros at the locations of the odds in $\boldsymbol{s}_{j+1}$. Starting from $\boldsymbol{s}_{j+1}$, and with $\widetilde{\boldsymbol{s}}_j = \widetilde{\mathbf{F}}_j \boldsymbol{s}_{j+1}$, we have perfect reconstruction by (6) if

$$\mathbf{Q}_{j+1} + (\mathbf{I}_{j+1} - \mathbf{Q}_{j+1}) \widetilde{\mathbf{J}}_j^T \widetilde{\mathbf{F}}_j = \mathbf{I}_{j+1}.$$

Taking $\mathbf{Q}_{j+1} = \mathbf{I}_{j+1}$ reduces (6) to (5), for which perfect reconstruction is guaranteed. In some situations, however, it is interesting not to recover the even components $\boldsymbol{s}_{j+1,e}$ using (5), as it involves a prediction and a detail coefficient. Instead, reconstruction is possible straight from $\widetilde{\boldsymbol{s}}_j$, at least if $\widetilde{\boldsymbol{s}}_j = \boldsymbol{s}_{j+1,e}$, so the prefilter needs to be trivial, $\widetilde{\mathbf{F}}_j = \widetilde{\mathbf{J}}_j$. Although not strictly necessary, it is a natural choice to take $\mathbf{Q}_{j+1}$ a diagonal matrix. Since there is only one way to reconstruct the odds in $\boldsymbol{s}_{j+1}$, the submatrix $\mathbf{Q}_{j+1,o,o}$ must be the $(n_{j+1} - n_j) \times (n_{j+1} - n_j)$ identity matrix. In particular we propose to take $Q_{j+1,k,k} = q(x_{j+1,k}; \boldsymbol{x}_{j+1,o}, h_j)$, where $q(x; \boldsymbol{x}_{j+1,o}, h_j)$ is a continuous weight function, which equals one in $x_{j+1,\ell}$ if $\ell \in o(j+1)$. On the other hand, we propose to set $q(x; \boldsymbol{x}_{j+1,o}, h_j) = 0$ in points x far from any knot in $\boldsymbol{x}_{j+1,o}$. More precisely, we suggest a zero weight if $\min_{\ell \in o(j+1)} |x - x_{j+1,\ell}| > h_j$ and a smooth transition between zero and one for points closer to any of the $\boldsymbol{x}_{j+1,o}$ than the bandwidth h_j.

The benefit from this weighted reconstruction lies in the values $\boldsymbol{s}_{j+1,e}$ far from the refinement points in $\boldsymbol{x}_{j+1,o}$. If these values are kept at coarse scale ($\widetilde{\mathbf{F}}_j = \widetilde{\mathbf{J}}_j$), then there is no need to keep detail coefficients coding for the offset between the value and a smoothing prediction. The reconstruction (6) thus allows us to keep the redundancy under control. The alternative reconstruction is particularly interesting when the partitioning in $e(j+1)$ and $o(j+1)$ is far from dyadic, i.e., far from the alternating even-odd split. Non-dyadic splits are useful in the decomposition on a highly heterogeneous set of covariates, because the bandwidth h_j may be too small in regions with few covariate values. The local polynomial prediction in (4) is then possibly unbounded, leading to coefficients with uncontrolled variances. This situation occurs when the covariates come from random, highly nonuniform, densities, which is the very framework of this paper. Non-dyadic subdivision equipped with weighted reconstruction is a way to control the variance propagation.

In an extreme approach, $o(j+1)$ is just a singleton, thus incorporating a lifting scheme with one coefficient at-a-time [Nunes et al., 2006] into a Multiscale Local Polynomial Transform.

In a scheme with a general non-dyadic split, a proper construction of a local polynomial smoothing with degree $\widetilde{p} - 1$ requires at least $\widetilde{p}$ knots. Therefore we define the active set of knots, i.e., the set of splittable or predictable knots by

$$A_j = \{x_{j+1,k}, k = 1, 2, \ldots, n_{j+1}, |x_{j+1,k} - x_{j+1,k+2l+1}| < h_j \text{ for at least } \widetilde{p}_j \text{ values of } l\}. \qquad (7)$$

In this definition, the parameter $\widetilde{p}_j$ controls the smoothness of the reconstruction from the refinement across the resolution levels. The value can be taken level dependent, in order to combine sharp reconstructions at fine scales, using small values of $\widetilde{p}_j$, with smooth reconstructions at coarse scales, using larger values of $\widetilde{p}_j$. Obviously, we need that $\widetilde{p}_j \geq \widetilde{p}$ at all levels.

The set of points $o(j+1)$ that are actually taken out at level j is then given by a recursion. If $x_{j+1,i} = \min(A_j)$, then $i \in o(j+1)$. Furthermore, if $k - 1 \notin o(j+1)$ and $x_{j+1,k} \in A_j$, then $k \in o(j+1)$.

3 Working in an MLPT basis

3.1 The construction of the basis

The Multiscale Local Polynomial Transform can also be described in terms of the basis functions $\varphi_{J,k}(x)$ of the expansion (2). Indeed, taking $j = J$ in (2), we can write the finest scale approximation as $f_J(x) = \Phi_J(x)\boldsymbol{s}_J$, where $\Phi_J(x)$ is a row vector of basis functions $\varphi_{J,k}(x)$. The transform rewrites this approximation as

$$f_J(x) = \Phi_L(x)\boldsymbol{s}_L + \sum_{j=L}^{J-1} \Psi_j(x)\boldsymbol{d}_j, \tag{8}$$

where the rows of functions $\Phi_L(x)$ and $\Psi_j(x)$ can be found through an adjoint transform. Using the reconstruction of (5), the adjoint transform is

$$\begin{aligned} \Phi_j(x) &= \Phi_{j+1}(x)\mathbf{P}_j, \text{ and,} & (9)\\ \Psi_j(x) &= \Phi_{j+1}(x)(\mathbf{I}_{j+1} - \mathbf{P}_j\mathbf{U}_j) = \Phi_{j+1}(x) - \Phi_j(x)\mathbf{U}_j. & (10) \end{aligned}$$

This can be seen, step by step, by imposing the equality

$$\Phi_{j+1}(x)\boldsymbol{s}_{j+1} = \Phi_j(x)\boldsymbol{s}_j + \Psi_j(x)\boldsymbol{d}_j, \tag{11}$$

in which (5) is substituted. The weighted reconstruction in (6) leads to a generalised version of the two scale equation (9),

$$\Phi_j(x) = \Phi_{j+1}(x)\left[\mathbf{Q}_{j+1}\mathbf{P}_j + (\mathbf{I}_{j+1} - \mathbf{Q}_{j+1})\widetilde{\mathbf{J}}_j^T\right]. \tag{12}$$

Unless otherwise stated, further discussions work with the simple two scale equation (9), in order to keep the expressions as simple as possible. All further conclusions apply to the weighted reconstruction if (9) is replaced by (12).

Expression (8) is not a decomposition into a basis, because the collection of functions in $\Phi_L(x)$ and $\Psi_j(x)$ is overcomplete. The collections $\Phi_L(x)$ and $\Psi_j(x)$ for each j separately are, however, linearly independent.

Expression (10) is used in the design of the update matrix $\mathbf{U}_j$. We impose that $\int_{-\infty}^{\infty} \Psi_j(x)^T dx = \mathbf{0}_j$, so that any processing of the detail coefficients $\boldsymbol{d}_j$ preserves the integral of the function. More precisely, defining $\widehat{f}_J(x) = \Phi_L(x)\boldsymbol{s}_L + \sum_{j=L}^{J-1} \Psi_j(x)\widehat{\boldsymbol{d}}_j$, for any values of $\widehat{\boldsymbol{d}}_j$, it holds that $\int_{-\infty}^{\infty} \widehat{f}_J(x)dx = \int_{-\infty}^{\infty} f_J(x)dx$, with $f_J(x)$ defined in (8). Defining the moments $M_j^q = \int_{-\infty}^{\infty} \Phi_j(x)^T x^q dx$, the integration of (9) becomes $M_j^q = \mathbf{P}_j^T M_{j+1}^q$, while the integration of (10) leads to the following condition on $\mathbf{U}_j$,

$$M_{j+1}^q = \mathbf{U}_j^T M_j^q, \tag{13}$$

which is imposed at least for $q = 0$, so that the zero integral condition (11) is satisfied. The matrix $\mathbf{U}_j$ is taken to have a sparse band structure, close to that of $\widetilde{\mathbf{J}}_j$, while satisfying (13) for $q = 0, \ldots, p-1$, where

p denotes the number of primal vanishing moments. As condition (13) may lead to uncontrolled variances of $\boldsymbol{s}_j$, additional variance control conditions can be applied [Jansen, 2016], leading to a slightly higher number of nonzero elements in $\widetilde{\mathbf{J}}_j$. These variance control conditions have been developed for B-spline wavelet transforms. In the framework of the overcomplete MLPT, the straightforward implementation of these conditions turns out to lead to ill conditioned linear systems. Further research, beyond the scope of this paper, is necessary to find fast and stable methods for variance control in MLPT.

Repeated refinement allows us to write all functions in terms of the initial basis $\Phi_J(x)$,

$$\Phi_j(x) = \Phi_J(x)\mathbf{P}_{J-1}\mathbf{P}_{J-2}\ldots\mathbf{P}_j. \tag{14}$$

The finest scale functions in $\Phi_J(x)$, are in principle free to choose. In the first instance we take $\varphi_{J,k}(x) = \chi_{J,k}(x)$, with $\chi_{J,k}(x)$ the characteristic (or identicator) function of the interval $I_{J,k}$. The intervals $I_{J,k}$ are chosen to form a partition of $[0,1]$ so that $x_{J,k} \in I_{J,k}$. Now, the scaling functions at scale j are piecewise constant functions, determined by the partition at scale J. The dependence on the fine scale J is made explicit in the notation $\Phi_j^{[J]}(x)$.

3.2 Superresolution

The piecewise constant basis $\Phi_j^{[J]}(x)$ can be replaced by an alternative basis of continuous functions $\overline{\Phi}_j^{[J]}(x)$, defined by iterated local polynomial smoothing. More precisely, the iterations start off with $\overline{\Phi}_j^{[j]}(x)$, whose columns contain functions $\overline{\varphi}_{j,k}^{[j]}(x)$ satisfying $\overline{\varphi}_{j,k}^{[j]}(x_{j,l}) = \delta_{k,l}$. Then, for $i = j, j+1, \ldots, J-1$, we define

$$\overline{\Phi}_j^{[i+1]}(x) = \sum_{l=0}^{n_i-1} \overline{\Phi}_j^{[i]}(x_{i,l})P_{i,l}(x;\boldsymbol{x}_i). \tag{15}$$

It is straightforward to show that the alternative $\overline{\Phi}_j^{[J]}(x)$ satisfies the iterative refinement (14) for fixed superscript $[J]$, and interpolates the piecewise constant basis $\overline{\Phi}_j^{[J]}(x)$ in all fine scale knots.

Proposition 1 *The set of continuous, linearly independent functions $\overline{\Phi}_j^{[J]}$, defined recursively by (15), interpolates the set of piecewise constant, linearly independent functions $\Phi_j^{[J]}(x)$, i.e., $\overline{\Phi}_j^{[J]}(\boldsymbol{x}_J) = \Phi_j^{[J]}(\boldsymbol{x}_J) = \mathbf{P}_{J-1}\mathbf{P}_{J-2}\ldots\mathbf{P}_j$. In this expression, $\Phi_j^{[J]}(\boldsymbol{x}_J)$ stands for the $n_J \times n_j$ matrix with elements $\varphi_{j,k}^{[J]}(x_{J,l})$ on row k, column l, while $\overline{\Phi}_j^{[J]}(\boldsymbol{x}_J)$ is of course the same matrix, now using the functions in $\overline{\Phi}_j^{[J]}(x)$.*

The approximative construction (15) can be refined beyond the resolution of the observations by inserting knots without observations between the elements of $\boldsymbol{x}_J$. This leads to the approximation $\overline{\Phi}_j^{[J^*]}(x)$ at superresolution J^*, where J^* can be arbitrarily fine. Likewise, the actual scaling functions can be defined through a refinement as in (14) up to the superresolution J^*.

Superresolution refinement is also possible in the context of wavelet decompositions. It is a useful numerical tool for finding the scaling functions when no closed form is available.

3.3 Finest scale coefficients

The approximation (15) has the interesting property to reproduce polynomials of degree $\widetilde{p}-1$ from their function values in the knots.

Lemma 1 *For any $J^* > j$, and with $\boldsymbol{x}_j^q$ the vector of observations $x_{j,k}^q$ from a power functions x^q, we find $\overline{\Phi}_j^{[J^*]}(x)\boldsymbol{x}_j^q = x^q$.*

By fixing $\Phi_j^{[J^*]}(x)$ through refinement up to superresolution, Lemma 1 leads to $\Phi_j^{[J^*]}(x)\boldsymbol{x}_j^q = x^q + \mathcal{O}\left(\Delta_{J^*}\right)$, where $\Delta_j = \max_{k=1,\ldots,n_j-1}(x_{j,k} - x_{j,k-1})$ and J^* is an arbitrarily fine superresolution.

The reconstruction of polynomials from function values as in Lemma 1 is shared with interpolating wavelet schemes, such as the Deslauriers-Dubuc lifting scheme [Deslauriers and Dubuc, 1989, Donoho and Yu, 1999]. The downside of the Deslauriers-Dubuc refinement scheme is that on nonequispaced settings, it may produce unbounded oscillations in the prediction $P_{j,l}(x;\boldsymbol{x}_j)$ and hence in the basis functions $\overline{\Phi}_j^{[J^*]}(x)$. Nonequispaced wavelet decompositions with bounded scaling functions do exist, B-splines [Jansen, 2016] being an important example. These scaling functions, however, have the problem that they do not reproduce polynomials from function values.

A second property is the compact support of the basis functions.

Lemma 2 *If at each level j, and for each knot $x_{j+1,k}$, there are at least $\widetilde{p}$ knots in $\boldsymbol{x}_j$ within distance h_j from $x_{j+1,k}$, then $\overline{\varphi}_{j,k}(x)$ has a support comprised in $[x_{j,k} - h_j^*, x_{j,k} + h_j^*]$, where $h_j^* = \sum_{i=j}^{J-1} h_i$.*

As a conclusion, the Multiscale Local Polynomial Transform combines three features: first, using non-dyadic refinement and weighted reconstruction from the overcomplete decomposition, it is possible to have bounded refinement in $\mathbf{P}_j$ and hence bounded scaling functions in $\overline{\Phi}_j^{[J^*]}(x)$. Second, the support of the basis functions is bounded. Third, the basis functions reproduce polynomials from their function values in the knots. These three ingredients are needed for the following result.

Theorem 1 *Let $f(x) \in C^{\widetilde{p}+\alpha}(a_j, b_j)$ with positive α, and $(a_j, b_j) \subset$ Dom(f), the domain of f. Furthermore, let $\Phi_j(x)$ be a scaling basis defined on the knots $\boldsymbol{x}_j$. Assuming that*

(A1) $\Phi_j(x)\boldsymbol{x}_j^q = x^q$ *for* $q = 0, 1, \ldots, \widetilde{p}-1$,

(A2) all functions in $\Phi_j(x)$ are bounded by $|\varphi_{j,k}(x)| \leq M$ for some positive M independent from k,

(A3) all functions in $\Phi_j(x)$ have bounded support, meaning that $\varphi_{j,k}(x_1)\varphi_{j,k}(x_2) \neq 0$ implies $|x_1 - x_2| < 2h_j^*$,

(A4) the number of knots within distance h_j^* of x is bounded from above, independently from j,

then the approximation $f_j(x) = \Phi_j(x) f(\boldsymbol{x}_j)$, *where* $f(\boldsymbol{x}_j)$ *is the vector of function values* $f(x_{j,k})$ *in the knots* $x_{j,k}$, *has an approximation error uniformly bounded by* $|f_j(x) - f(x)| \leq K h_j^{*\widetilde{p}}$.

Proof. See AppendixA. □

The result in Theorem 1 states that the approximation $f_j(x)$ of $f(x)$ using its function values in the knots achieves the same convergence rate as the approximation (2) with coefficients from a least squares or other projection onto the basis. The approximation with function values thus combines a basis decomposition with elements from a kernel approximation. As an example, the Nadaraya-Watson and Gasser-Müller local estimators also take function values as coefficients, $\widehat{f}_{\text{local},j}(x) = \sum_{k=0}^{n_j-1} f(x_{j,k}) K_{j,k}(x)$, where $K_{j,k}(x) = K\left(\frac{x-x_{j,k}}{h}\right) \Big/ \sum_{l=0}^{n_j-1} K\left(\frac{x-x_{j,l}}{h}\right)$, in the Nadaraya-Watson case, while for Gasser-Müller,

$$K_{j,k}(x) = \frac{1}{h} \int_{(x_{j,k-1}+x_{j,k})/2}^{(x_{j,k}+x_{j,k+1})/2} K\left(\frac{u-x}{h}\right) du.$$

In the framework of MLPT density estimation, the importance of Theorem 1 lies in the definition of the finest scaling coefficient vector $\boldsymbol{S}_J$ as unbiased or nearly unbiased fine scale estimators of the density. In other words, thanks to Theorem 1, it is possible to take as fine scale coefficients a vector of variables $S_{J,k}$ so that $E(S_{J,k}) \approx f_X(x_{J,k})$. Such a fine scale nearly unbiased estimator can be obtained by a sort of adaptive, fine scale histogram, further developed in Section 4.3, Expression (19). The fine scale values $\boldsymbol{S}_J$ act as a pilot estimator of $f_X(x)$, which is asymptotically unbiased but noisy. The subsequent MLPT regression then finds a trade-off between bias and variance.

4 Multiscale Local Polynomial density estimation

4.1 The proposed density estimation procedure

Let $\boldsymbol{X}_n$ represent a vector of n i.i.d. observations from an unknown density $f_X(x)$, which may have one or more, isolated, singular points at unknown locations. These singular points are points of discontinuity or infinite values. This paper proposes the following procedure for the estimation of $f_X(x)$ from $\boldsymbol{X}_n$.

1. Define the following covariates and responses in a nonparametric additive model as in Section 4.2:
   (a) Let $\Delta X_{n;i} = X_{(n;i)} - X_{(n;i-1)}$ be the differences between the ordered values of $\boldsymbol{X}_n$.

   (b) Define the intermediate values

   $$\xi_{n;i} = \left[X_{(n;i-1)} + X_{(n;i)}\right]/2,$$

   which will be used as covariates.

   (c) Define the fine scale response variables $S_{J,i}^{[0]} = 1/\Delta X_{n;i}$. As in Section 2.1, the subscript J refers to the highest, i.e., finest scale in the subsequent multiscale analysis.

(d) Reduce the variance in the response variables $\boldsymbol{S}_J^{[0]}$, at the price of a small scale bias by a prefilter, defining $\boldsymbol{S}_J$, as in (19).

Then, based on the result in Lemma 4, the observations $\boldsymbol{S}_J$ are modelled as approximately normally distributed variables with expected values $\boldsymbol{\theta}_j$, where $\theta_{J,i} = f_X(\xi_{n;i})$.

2. Write the nonparametric model as a high-dimensional, sparse linear regression model $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$, where the design matrix is the inverse MLPT matrix.

   Using the forward transform matrix, the coefficients $\boldsymbol{\beta}$ are estimated by a soft-thresholding scheme, from which the estimator of the density can be found by reconstruction, as in (18).

4.2 Density estimation as a generalised linear model

Let $X_{(n;i)}$, with $i = 1, 2, \ldots, n$, be the order statistics of a random sample of size n from the distribution with probability density function $f_X(x)$. Then, for the differences $\Delta X_{n;i} = X_{(n;i)} - X_{(n;i-1)}$, also termed spacings [Pyke, 1965] in the literature, there exists a mean value $\overline{\xi}_{n;i} \in [X_{(n;i-1)}, X_{(n;i)}]$, so that

$$f_X(\overline{\xi}_{n;i})\Delta X_{n;i} = \Delta U_{n;i}, \tag{16}$$

where $\Delta U_{n;i} = U_{(n;i)} - U_{(n;i-1)} = F_X(X_{(n;i)}) - F_X(X_{(n;i-1)})$ are uniform spacings, i.e., differences between successive order statistics from a sample of a uniform random variable.

A straightforward calculation leads to the following well known result [Pyke, 1965]:

Lemma 3 *For a given n, and for $i = 2, 3, \ldots, n$, all values of $\Delta U_{n;i}$ are identically distributed with cumulative distribution $F_{\Delta U_{n;i}}(u) = F_{U_{(n;1)}}(u) = 1 - (1-u)^n$.*

From here, it follows immediately that $(n+1)\Delta U_{n;i} \overset{\mathrm{d}}{\to} \exp(1)$. Although the uniform spacings are not independent, they are so up to a random normalisation. Indeed, we have the following result [Devroye, 1986, Chapter 5, Theorem 2.2].

Theorem 2 *The joint distribution of the spacings $\Delta U_{n;i}$ is given by the joint distribution of iid exponential random variables, normalised by their sum, i.e., by the joint distribution of the values $\Lambda_i / \sum_{j=1}^n \Lambda_j$, where $\Lambda_j \sim \exp(\lambda)$ for a common value of λ, while all Λ_j are independent.*

This seems to justify the interpretation of $\Delta U_{n;i}$ as a noise factor in (16). Nevertheless, since we would like to use the form (16) as a basis for nonparametric regression, we propose the following result for conditional convergence in distribution, thereby refining the marginal convergence result following from Lemma 3.

Proposition 2 *Let D_α be a subset of $[0,1] \times [0,1]$ so that $(0, v) \in D_\alpha$ for all $v \in [0,1]$. Let $\alpha(t, v)$ be a twice differentiable bivariate function defined on D_α, satisfying the following*

(P1) For a given t, there exists a positive, Riemann integrable function $A_0(t)$, so that $\left|\frac{\partial\alpha}{\partial v}(t,v)\right| \leq A_0(t)$.

(P2) $\lim_{t\to 0} t\frac{\partial\alpha}{\partial t}(t,v) = 0$.

Let $U_{(n;i)}$, with $i = 1, 2, \ldots, n$, be the order statistics of a random sample of size n from a uniform distribution on $[0,1]$. Define $\Delta U_{n;i} = U_{(n;i)} - U_{(n;i-1)}$ and consider the implicit definition $V_{n;i} = [1 - \widehat{\alpha}_{n;i}]U_{(n;i-1)} + \widehat{\alpha}_{n;i}U_{n;i}$, where $\widehat{\alpha}_{n;i} = \alpha(\Delta U_{n;i}, V_{n;i})$.

Then, with $i/n \to \rho$ for $n \to \infty$, it follows that $(n+1)\Delta U_{n;i}|V_{n;i} = v \overset{\mathrm{d}}{\to} \Delta_v \sim \exp(1/\mu(v))$, where $\mu(v) = 1/E(\Delta_v)$ is given by $\frac{1}{\mu(v)} = \frac{1-\rho}{1-v}\,[1-\alpha(0,v)] + \frac{\rho}{v}\,\alpha(0,v)$. Moreover, the convergence of the distribution function is uniform in v.

Remark 1 *Whereas $V_{n;i}$ is presented in Proposition 2 as an implicitly defined random variable, its definition becomes explicit when α is given as a function of $U_{(n;i-1)}$ and $U_{(n;i)}$ instead of a function of $V_{n;i}$ and $\Delta U_{n;i}$. A situation where α is a constant holds as a special case.*

Proof. See AppendixB. □

The result in Proposition 2 can be simplified a bit further. Indeed, it is well known that the empirical quantile function of a sample from a uniform random variable converges with probability one uniformly to the population quantile function $Q_U(p) = p$. As a result, if $i/n \to \rho$, then $P(U_{(n;i-1)} \to \rho) = 1 = P(U_{(n;i)} \to \rho)$. By the sandwich theorem, $P(V_{n;i} \to \rho) = 1$, and so, $(n+1)\Delta U_{n;i}|V_{n;i} \overset{\mathrm{d}}{\to} D \sim \exp(1)$ almost surely.

The statement of Proposition 2 can be applied to $V_{n;i} = F_X(\overline{\xi}_{n;i})$, with $\overline{\xi}_{n;i}$ defined by (16), leading to the following asymptotic nonparametric regression model.

Corollary 1 *Let $f_X(x)$ be a continuously differentiable density function with ordered observations $X_{(n;i)}$, $i = 1, 2, \ldots, n$. Define the spacings $\Delta X_{n;i} = X_{(n;i)} - X_{(n;i-1)}$ and the middle values $\overline{\xi}_{n;i}$ so that $f_X(\overline{\xi}_{n;i})\Delta X_{n;i} = F_X(X_{(n;i)}) - F_X(X_{(n;i)})$. Then, almost surely,*

$$f_X(\overline{\xi}_{n;i})(n+1)\Delta X_{n;i}|\overline{\xi}_{n;i} \overset{\mathrm{d}}{\to} D \sim \exp(1). \tag{17}$$

Proof. See AppendixD. □

As the covariate values in $\overline{\xi}_{n;i}$ depend on the unknown density function through the definition in (16), the subsequent analysis redefines the covariates as $\xi_{n;i} = \left[X_{(n;i-1)} + X_{(n;i)}\right]/2$, for which Proposition 2 still holds, meaning that $(n+1)\Delta U_{n;i}|V_{n;i} = F_X(\xi_{n;i}) \overset{\mathrm{d}}{\to} D \sim \exp(1)$, almost surely. The nonparametric regression model in (16) is replaced by $f_X(\xi_{n;i})\Delta X_{n;i} = \Delta U_{n;i} - \frac{f'_X(\zeta_{n;i,0})+f'_X(\zeta_{n;i,1})}{2} \cdot \frac{[\Delta X_{n;i}]^2}{4}$, with $\zeta_{n;i,0} \in [X_{(n;i-1)}, \xi_{n;i}]$ and $\zeta_{n;i,1} \in [\xi_{n;i}, X_{n;i}]$. Conditioned on $\xi_{n;i}$, the second term converges almost surely to zero in probability.

Taking the covariate values in the midpoints $\xi_{n;i}$ of the observations and taking as responses the spacings between the observations, the nonparametric regression model (17) stores the information from the original sample in both covariates and responses. The duplication of the information allows us, further on, to transform the response variables without losing information.

4.3 Estimation in a multiscale transform of exponential observations

The nonparametric regression model in (17) can be studied within the framework of a generalised linear model (GLM), where we observe a sample $\boldsymbol{Y}$ from the density $f_{Y_i}(y;\boldsymbol{\theta},\boldsymbol{\phi}) = \exp\left[\frac{-y\theta_i - b(\theta_i)}{d(\phi_i)}\right] h(y_i,\phi_i)$. The parameter vector $\boldsymbol{\theta}$ is identified with the unknown density values, $\theta_i = f_X(\xi_{n;i})$. For the sake of simplicity in subsequent expressions, we switch signs of $\boldsymbol{\theta}$ in the classical definition of a exponential family in the literature. In the light of Theorem 1 the density values can operate as finest scaling coefficients in a MLPT decomposition. To this end, the parameter vector $\boldsymbol{\theta}$ is further developed by sparse regression $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ are the coefficients of the MLPT decomposition. The design matrix, $\mathbf{X}$, is identified as the reconstruction matrix from a MLPT, while the responses are given by the rescaled spacings, $Y_i = (n+1)\Delta X_{n;i}$. The expected responses are $\mu_i = E(Y_i) = 1/\theta_i$. In our case, $d(\phi_i) = 1 = h(y_i,\phi_i)$, so the dispersion parameter $\boldsymbol{\phi}$ has no effect.

In order to impose sparsity, the estimation of $\boldsymbol{\beta}$ should include a variable selection. Variable selection is typically achieved by a regularisation of the maximum likelihood estimator, adding a term that controls a sparsity norm of the proposed solution. Using the ℓ_1 norm to express sparsity, as in the literature on the lasso [Chen and Donoho, 1995, Tibshirani, 1996, van de Geer, 2008, Wang et al., 2015] leads to the following ℓ_1 regularised maximum log-likelihood problem $\max_{\boldsymbol{\beta}} \log L(\boldsymbol{\theta};\boldsymbol{Y}) - \lambda\|\boldsymbol{\beta}\|_1$.

The working independence score with respect to $\boldsymbol{\beta}$ for a sample from the GLM with regression $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$ and with $d(\phi_i) = 1 = h(y_i,\phi_i)$, is given by $\nabla \log L(\boldsymbol{\theta};\boldsymbol{Y}) = \mathbf{X}^T(\boldsymbol{Y}-\boldsymbol{\mu})$. The Karush-Kuhn-Tucker (KKT) conditions for the ℓ_1 regularised maximum log-likelihood problem are given by $\mathbf{X}_j^T(\boldsymbol{Y}-\boldsymbol{\mu}) = \lambda\mathrm{sign}(\beta_j)$ if $\beta_j \neq 0$, and $\left|\mathbf{X}_j^T(\boldsymbol{Y}-\boldsymbol{\mu})\right| < \lambda$ if $\beta_j = 0$, both for $j = 1,2,\ldots,m$. Even without the partitioning of the parameter vector into subsets of zeros and non-zeros, these expressions are highly nonlinear, as $\mu_i = 1/\theta_i$ while $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$, leading to iterative solvers with uncertain convergence.

In the specific context where the design $\mathbf{X}$ represents the reconstruction of a sparse data transformation, we wish to make use of the availability of a fast forward transformation $\widetilde{\mathbf{X}}$, in a soft-thresholding scheme

$$\widehat{\boldsymbol{\theta}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\cdot\mathrm{ST}(\widetilde{\mathbf{X}}\boldsymbol{S}_J,\lambda). \tag{18}$$

In this scheme, the finest scaling vector $\boldsymbol{S}_J$ is derived from the (asymptotically) exponential observations $\boldsymbol{Y}$ in a way so that two conditions are fulfilled.

1. The finest scaling vector acts as a pilot estimator of the parameter vector $\boldsymbol{\theta}$. Hence, we impose that at least asymptotically, $E(\boldsymbol{S}_J) = \boldsymbol{\theta}$.

2. The distribution of the MLPT transformed data should not show heavy tails, so that $\widetilde{\mathbf{X}}(\boldsymbol{S}_J - \boldsymbol{\theta})$

does suffer from outliers.

In combination with the latter condition, the assumption of sparsity in the MLPT decomposition $\boldsymbol{\beta} = \widetilde{\mathbf{X}}\boldsymbol{\theta}$ ensures that large values in $\widetilde{\mathbf{X}}\boldsymbol{S}_J$ can be attributed to the $\boldsymbol{\theta}$, thus motivating the use of a threshold.

As a choice for $\boldsymbol{S}_J$ we propose to take $S_{J,k} = 1/\overline{Y}_k$, where $\overline{Y}_k$ is defined by adaptive local averaging,

$$\overline{Y}_k = \frac{1}{|\partial k|} \sum_{l \in \partial k} Y_l, \tag{19}$$

with ∂k the smallest set of the form $\{k, k-1, k+1, k-2, k+2, \ldots\}$ so that either $|\partial k| = r$ for a user defined integer r or $\xi_l - \xi_m > h_{J,0}$ for at least one pair $l, m \in \partial k$ and with $h_{J,0}$ a user defined minimum working scale at the finest resolution level J. Although the local averaging induces a certain loss in the response data, the duplication of the input into both grid $\boldsymbol{\xi}$ and response $\boldsymbol{Y}$ ensures that no information is really lost. The local average operation is denoted by $\overline{\boldsymbol{Y}} = \widetilde{\mathbf{Q}}\boldsymbol{Y}$, where the matrix $\widetilde{\mathbf{Q}}$ is viewed as being parametrically dependent on the grid vector $\boldsymbol{\xi}$. This way, the adaptive local averaging can be modelled as a linear, and thus continuous operation on $\boldsymbol{Y}$.

The process of local averaging is an extreme case of k-nearest estimator [Loftsgaarden and Queensberry, 1965] or it can be seen as a special case of variable kernel estimator [Breiman et al., 1977, Terrell and Scott, 1992].

The asymptotic distribution of $\boldsymbol{S}_J$ follows from the subsequent analysis.

Lemma 4 *Let $\boldsymbol{Y}$ be a vector of iid exponential random variables, with parameter θ, i.e., $E(Y_i) = 1/\theta$ and with cumulative sums $X_k = \sum_{i=1}^{k} Y_i$. For a given positive value of h, define $N_h = \min\{k | X_k > h\}$. Then for $S_h = N_h/X_{N_h}$, we have*

$$\begin{aligned} E(S_h) &= \theta(h\theta + 1)\exp(h\theta)\left[-\mathrm{Ei}(-h\theta)\right], & (20) \\ \mathrm{var}(S_h) &= \theta^2 \exp(2h\theta)\left[\mathrm{Ei}(-h\theta)\right]^2 h\theta + & (21) \\ &\quad \theta^2\left[(h\theta)^2 + 3h\theta + 1\right] \cdot \left\{\frac{1}{h\theta} - \exp(h\theta)\left[-\mathrm{Ei}(-h\theta)\right] - \exp(2h\theta)\left[\mathrm{Ei}(-h\theta)\right]^2\right\}. \end{aligned}$$

These expressions use the notation for the exponential integral $\mathrm{Ei}(\tau) = -\int_{-\tau}^{\infty}\left[\exp(-t)/t\right] dt$. Moreover, for $h \to \infty$ the variables S_h are asymptotically normally distributed, $S_h \sim \mathrm{AN}(\theta, \theta/h)$. Finally, let $S_{h,r} = S_h$ if $N_h \leq r$ and $S_{h,r} = (r-1)/X_r$ otherwise, then

$$E(S_{h,r}) = \theta\left\{P(N_h \geq r) + P(N_h \leq r)\left[1 + \frac{h\theta P(N_h \leq r-1)}{P(N_h \leq r)}\right]\exp(h\theta)\left[-\mathrm{Ei}(-h\theta)\right]\right\}. \tag{22}$$

Moreover, for $h \to \infty$ and $r \to \infty$ the variables $S_{h,r}$ are asymptotically normally distributed.

Proof. See AppendixE. □

4.4 Degrees of freedom

For the selection of an optimal threshold in (18), we adopt the Kullback-Leibler distance as the objective function. More precisely, suppose that the data generating process is an independently exponentially distributed random vector with parameter $\boldsymbol{\theta}$ and let $\widetilde{\boldsymbol{\theta}}$ be the value of that parameter in the model under consideration. The Kullback-Leibler (KL) distance, defined as

$$\mathrm{KL}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} E\left[\log f_{Y_i}(Y_i;\theta_i)\right] - E\left[\log f_{Y_i}(Y_i;\widetilde{\theta}_i)\right],$$

becomes

$$\mathrm{KL}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\left[\log(\theta_i) - 1\right] + \frac{1}{n}\sum_{i=1}^{n}\left[\widetilde{\theta}_i\mu_i - \log(\widetilde{\theta}_i)\right].$$

The first sum depends on the unobserved $\boldsymbol{\theta}$, but it has no effect, as it is a constant for all models $\widetilde{\boldsymbol{\theta}}$. The second term equals $-\ell(\widetilde{\boldsymbol{\theta}})$, with

$$\ell(\widetilde{\boldsymbol{\theta}}) = \sum_{i=1}^{n}\left[\log(\widetilde{\theta}_i) - E(\widetilde{\theta}_i)\mu_i\right]$$

the expected log-likelihood of the model under consideration, $\widetilde{\boldsymbol{\theta}}$. The expectation is taken over the unknown data generating process. In practice, the expected log-likelihood is estimated by its empirical counterpart, evaluated in a sample dependent estimator, $\widehat{\ell}(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n}\left[\log(\widehat{\theta}_i) - \widehat{\theta}_i Y_i\right]$. The bias of the estimator is given by $\nu(\widehat{\boldsymbol{\theta}}) = E\left[\widehat{\ell}(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}})\right] = E\left[\widehat{\boldsymbol{\theta}}^T(\boldsymbol{\mu} - \boldsymbol{Y})\right]$. In the context of a linear regression model, the quantity $\nu = E\left[\widehat{\boldsymbol{\mu}}^T(\boldsymbol{Y} - \boldsymbol{\mu})\right]/\sigma^2$ is termed the number of (generalised) degrees of freedom, for which various expressions have been developed under a variety of selection and estimation procedures [Efron, 1986, Ye, 1998, Tibshirani and Taylor, 2012, Zou et al., 2007, Gao and Fang, 2011, Zhang et al., 2012, Tibshirani, 2015, You et al., 2016, Vaiter et al., 2017]. In our case of soft-thresholding MLPT applied to inverse exponential random variables, further development of this bias or degrees of freedom is based on the following generalisation of Stein's Lemma.

Lemma 5 *Let $f_Y(y)$ be a density defined on $[a,b]$, where a and b may be infinite. Suppose that $E(Y) = \mu$ and $\mathrm{var}(Y) = \sigma^2$ are finite. Define $f_{Y'}(y) = \frac{1}{\sigma^2}\int_a^y f_Y(u)(\mu - u)du$, then $f_{Y'}(y)$ is a density and $E\left[(Y-\mu)g(Y)\right] = \sigma^2 E\left[g'(Y')\right]$, at least for any function $g(u)$ defined on $[a,b]$ for which a function $g'(y)$ exists, satisfying*

$$\int_a^y g'(u)du = g(y).$$

Proof. It is straightforward to verify that $f_{Y'}(y)$ is a density. The rest of the lemma follows from integration by parts. □

In the case where $Y \sim \exp(\theta)$, we find $Y' \sim \Gamma(\theta, 2)$ and $E\left[g'(Y')\right] = \theta E\left[g'(Y)Y\right]$. It then holds that $E\left[(Y-\mu)g(Y)\right] = \sigma^2\theta E\left[g'(Y)Y\right] = E\left[g'(Y)Y\right]/\theta$. In a bivariate case (X,Y), a conditional version of (5) reads $E\left[(Y-\mu_{Y|X})g(X,Y)\middle|X\right] = \sigma^2_{Y|X}E\left[\frac{\partial g(X,Y')}{\partial y}\middle|X\right]$. For independent variables, it follows that $E\left[(Y-\mu_Y)g(X,Y)\right] = \sigma^2_Y E\left[\frac{\partial g(X,Y')}{\partial y}\right]$.

We now introduce the notation $\boldsymbol{Y}_i'$ for the random vector that is equal to $\boldsymbol{Y}$, except for the ith component, which is Y_i'. Then, for independent, exponential observations Y_i, we have

$$\nu(\widehat{\boldsymbol{\theta}}) = -\sum_{i=1}^{n}\sigma^2_{Y_i}E\left[\frac{\partial\widehat{\theta}_i(\boldsymbol{Y}_i')}{\partial Y_i}\right] = -\sum_{i=1}^{n}\sigma^2_{Y_i}\theta_i E\left[Y_i\frac{\partial\widehat{\theta}_i(\boldsymbol{Y})}{\partial Y_i}\right] = -\sum_{i=1}^{n}\theta_i^{-1}E\left[Y_i\frac{\partial\widehat{\theta}_i(\boldsymbol{Y})}{\partial Y_i}\right].$$

The subsequent development also introduces the diagonal matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Upsilon}$ with elements θ_j and Y_i on the diagonals. The expression above becomes $\nu(\widehat{\boldsymbol{\theta}}) = -E\left\{\mathrm{Tr}\left[\boldsymbol{\Theta}^{-1}\boldsymbol{\Upsilon}\mathbf{J}_{\boldsymbol{Y}}(\widehat{\boldsymbol{\theta}})\right]\right\}$, where $\mathbf{J}_{\boldsymbol{Y}}(\widehat{\boldsymbol{\theta}})$ is the Jacobian matrix of $\widehat{\boldsymbol{\theta}}$ w.r.t. $\boldsymbol{Y}$.

The Jacobian matrix is given by

$$\mathbf{J}_{\boldsymbol{Y}}(\widehat{\boldsymbol{\theta}}) = \mathbf{X}\mathbf{D}_\lambda\widetilde{\mathbf{X}}\mathbf{J}_{\boldsymbol{Y}}(\boldsymbol{S}_J) = -\mathbf{X}\mathbf{D}_\lambda\widetilde{\mathbf{X}}\overline{\boldsymbol{\Upsilon}}^{-2}\widetilde{\mathbf{Q}},$$

leading to

$$\nu(\widehat{\boldsymbol{\theta}}) = E\left\{\mathrm{Tr}\left[\mathbf{X}\mathbf{D}_\lambda\widetilde{\mathbf{X}}\overline{\boldsymbol{\Upsilon}}^{-2}\widetilde{\mathbf{Q}}\boldsymbol{\Upsilon}\boldsymbol{\Theta}^{-1}\right]\right\} = E\left\{\mathrm{Tr}\left[\mathbf{D}_\lambda\widetilde{\mathbf{X}}\overline{\boldsymbol{\Upsilon}}^{-2}\widetilde{\mathbf{Q}}\boldsymbol{\Upsilon}\boldsymbol{\Theta}^{-1}\mathbf{X}\right]\right\}.$$

The value of $\nu(\widehat{\boldsymbol{\theta}})$ can be estimated by

$$\widehat{\nu}(\widehat{\boldsymbol{\theta}}) = \mathrm{Tr}\left[\mathbf{D}_\lambda\widetilde{\mathbf{X}}\overline{\boldsymbol{\Upsilon}}^{-2}\widetilde{\mathbf{Q}}\boldsymbol{\Upsilon}\widehat{\boldsymbol{\Theta}}^{-1}\mathbf{X}\right],$$

where the estimator $\widehat{\boldsymbol{\Theta}}^{-1}$ can be taken to be a diagonal matrix with slightly shifted versions of the observed values, i.e., $\widehat{\theta}_{ii}^{-1} = Y_{i-1}$, so that the diagonal elements of $\widehat{\boldsymbol{\Theta}}^{-1}$ and those of $\boldsymbol{\Upsilon}$ are pairwise independent. As for computational complexity, a fast evaluation of $\widehat{\nu}(\widehat{\boldsymbol{\theta}})$ for a sequence of values of λ follows from the expression $\widehat{\nu}(\widehat{\boldsymbol{\theta}}) = \sum_{i\in\mathcal{I}_\lambda}W_{ii}$, with $\mathcal{I}_\lambda = \{i|D_{\lambda,ii} = 1\}$ and where $\mathbf{W} = \widetilde{\mathbf{X}}\overline{\boldsymbol{\Upsilon}}^{-2}\widetilde{\mathbf{Q}}\boldsymbol{\Upsilon}\widehat{\boldsymbol{\Theta}}^{-1}\mathbf{X}$ is independent from λ.

5 Simulation study

The proposed MLPT density estimator is applied to the power law on $[0,1]$, $f_X(x) = K|x-x_0|^k$, for $x\in[0,1]$, and with $K = (k+1)/[x_0^{k+1} + (1-x_0)^{k+1}]$. Throughout the simulations we set $x_0 = 0.345$ and $k = -1/2$.

Figure 1 has a visual comparison of the MLPT approach with a straightforward kernel density estimation and a kernel density estimation on probit transformed data. The simple kernel approach obviously

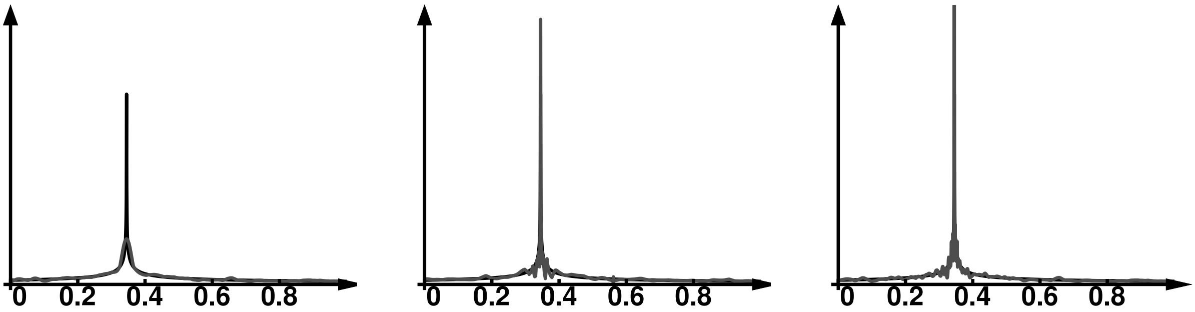

Figure 1: Left panel, in grey line: the result of kernel density estimation. The true density function in black line. Middle: density estimation using the proposed MLPT approach. Right panel: estimation using a probit transformation of the observations, assuming knowledge of the exact position of the singularity.

oversmooths near the singularity. Unlike the MLPT approach, the probit transform, $Y = \Phi^{-1}(X - x_0)$ for $X > x_0$ and $Y = \Phi^{-1}(X - x_0 + 1)$ for $X < x_0$, relies on exact knowledge of the singularity. Yet, the MLPT approach succeeds in locating the singularity and in reconstructing the density with less fluctuations than the kernel estimator on the probit transformed data.

The sample size in Figure 1 has been $n = 2000$. The MLPT approach involves the choice of quite some parameters. First, Definition (19) of the finest scaling coefficients from the observations $\boldsymbol{X}$ contains the parameters r and $h_{J,0}$, the latter being the scale at which the coefficient is defined. In the example of Figure 1, this scale is taken to be half the finest scale of the actual Multiscale Local Polynomial Transform, i.e., $h_{J,0} = h_J/2$. This choice is motivated by the idea that the fine scaling coefficients should carry information concentrated within a fine resolution. Higher values of $h_{J,0}$ may be recommended as well, since they promote normality of the fine scaling coefficients, albeit at the price of some bias. Thanks to the duplication of information in the model setup, discussed in Sections 4.2 and 4.3, there is always at least a theoretical possibility to remedy this bias. The other parameter in the definition of the finest scaling coefficients (19) is fixed throughout the simulations at $r = 10$. This value prevents the distribution of $S_{J,k}$ from showing heavy tails, as investigated in Lemma 4.

A second class of parameters comes with the actual MLPT. The number of dual vanishing moments $\widetilde{p}$ is set to two throughout the simulations, meaning that all results use Multiscale Local Linear Transforms. The number of primal vanishing moments is set to two as well, meaning that all detail basis functions have a zero integral and zero first moment, i.e., $\int_{-\infty}^{\infty} \psi_{j,k}(x)xdx = 0$. Obviously, the kernel function used in the transform is another parameter to choose. This text has worked with the cosine kernel, leaving further analysis on the impact of this choice to future research. An important issue is the choice of the bandwidths h_j at each level of the transform. The bandwidths act as user controlled scales at the corresponding levels. On a regular grid of covariates with regular subdivision, the scale is taken to be inversely proportional to the number of covariates in the grid at that resolution level, i.e., $h_j = \mathcal{O}(1/n_j)$. In particular, dyadic subdivision would call for dyadic bandwidths. For covariates at intermittent locations that can be modelled as ordered independent realisations from a uniform random variable, the irregularity can be dealt with by taking the bandwidths at each level slightly larger, more precisely $h_j = \mathcal{O}(\log(n_J)/n_J)$ [Jansen and Amghar, 2017]. Observations from other than uniform densities in a MLPT density estimation are too far from equidistant to be treated by a single bandwidth at a time. Therefore, we return to dyadic bandwidths, but, as proposed in Section 2.2, only subsample in regions where each prediction is based on a sufficient number of left over covariates within the distance of the bandwidth. As in (7), let $\widetilde{p}_j$ denote the level dependent minimum number of left over neighbours imposed for a subsampling and a prediction to take place, then in Figure 1 we take $\widetilde{p}_j = \widetilde{p} + 2$ at all levels. In a scheme with dyadic bandwidths, the bandwidth at the finest level, h_J, remains a parameter to be finetuned. As this parameter plays a crucial role in the asymptotics of the method (see Theorem 1), its assessment should depend, in an explicit or implicit way, on the sample size. The simulation of Figure 1 works with $h_J = 0.6/n$. The exact choice of h_J remains a topic of further investigation, especially because the analysis is quite sensitive to the value of this parameter.

Third, the selection of coefficients in the Multiscale Local Polynomial decomposition may be steered by a smoothing parameter for which Section 4.4 has proposed the optimisation of the information criterion $\widehat{\ell}(\widehat{\boldsymbol{\theta}}) - \widehat{\nu}(\widehat{\boldsymbol{\theta}})$. In its simplest form, the selection takes the form of a threshold procedure on the coefficients with normalised basis functions. As this operation does not guarantee positivity of the estimated density function, the procedure adopted in the simulation study adds to the selection another set of coefficients that makes the estimator positive on the entire domain. More sophisticated selection procedures include block thresholding and tree structured selection. The result in Figure 1 adopted level dependent thresholds, for a more adaptive reconstruction.

The simulation of Figure 1 is repeated $n_s = 200$ times for two samplesizes, $n = 100$ and $n = 1000$. The results are summarised in Table 1. The table compares bias and variance for the MLPT approach with both a simple kernel density estimation, using the same kernel function as in the MLPT approach, and the aforementioned probit transform. The root mean integrated squared biases (RMISB) and variances (RMIV) in the table are defined by $\mathrm{RMISB} = \left(\int_0^1 \left[E\widehat{f}_X(x) - f_X(x)\right]^2 dx\right)^{1/2}$, and $\mathrm{RMIV} = \left(\int_0^1 E\left[\widehat{f}_X(x) - E\widehat{f}_X(x)\right]^2 dx\right)^{1/2}$, where the expected values can be estimated by the mean over the simulation runs, as in $E\widehat{f}_X(x) \approx \frac{1}{n_s}\sum_{s=1}^{n_s} \widehat{f}_{X,s}(x)$. In this expression, $\widehat{f}_{X,s}(x)$ is the estimate of simulation s, evaluated in x. The evaluation of $\widehat{f}_{X,s}(x)$ can be realised by including the point of interest x, into the superresolution level of Proposition 1. The integrals can then be approximated numerically on a fine grid x. The tabled values seem to suggest that, at least for sample size $n = 1000$, the bias in a simple kernel approach is compensated by a much smaller variance than the competitors. It should be kept in mind, however, that the integrated biases and variances report on the global quality, whereas the visual comparison in Figure 1 draws the attention towards the local problems near the singularity. The probit transform and MLPT approaches perform relatively well near that singularity, yet even their local variances near the singularity are much larger than that of the simple kernel approach.

	RMISB		RMIV	
	$n = 100$	$n = 1000$	$n = 100$	$n = 1000$
Kernel	0.54890	0.54666	0.54118	0.16998
Probit+Kernel	0.164690	0.056735	2.16094	0.67754
MLPT	0.35631	0.10695	0.66695	0.43758

Table 1: Comparative study for classical kernel density estimations, probit transformed kernel estimation and MLPT for power law with singularity, based on 200 simulation runs.

On the other hand, the variance of the MLPT estimator remains relatively large, also for $n = 1000$, compared to that of the probit transform method. The reason is that, in contrast to the case of B-spline wavelet transforms [Jansen, 2016], there is no variance control available in the current implementation of the MLPT, as explained in Section 3.1.

The bias of the MLPT estimator decreases as the sample size increases. Although the MLPT does not know the precise location of the singularity, its bias comes close to that of the probit transform kernel estimator, which hinges on the precise location of the singularity.

6 Application to call center data

The MLPT density estimation method can be applied to the call center data used in Desmet et al. [2010]. This dataset contains the precise moments (in seconds) of 39553 phone calls during one month. The objective is to estimate the distribution of the phone calls over the day, i.e., the density of the phone calls on a domain of 24 hours, i.e., 86400 seconds. The performance is compared to the that of a simple kernel density estimator with Sheather-Jones bandwidth, as in Desmet et al. [2010], given by $h_K = 950.4$ (seconds, i.e., 0.264 hours).

The example illustrates how some of the MLPT parameters can be chosen as a function of the single scale kernel bandwidth h_K, whereas other parameters reflect the resolution of the observations. First, the scale $h_{J,0}$ in the definition of the fine scaling coefficients $\boldsymbol{S}_J$ is set to $h_{J,0} = h_K/16$. This value is large enough to model $\boldsymbol{S}_J$ as normally distributed data. The value is also much larger than the finest scale h_J of the actual multiscale analysis. The latter is not based on the kernel bandwidth, but is set to 10 seconds, thereby keeping track of the resolution of 1 second in the observations. The number of resolution levels in the MLPT is chosen such that the coarsest scale h_L is larger than the unique bandwidth in the kernel approach h_K. In this illustration we take dyadic bandwidths and $h_{L+2} \leq h_K \leq h_{L+1} = 2h_{L+2}$. This choice provides us with two levels beyond the kernel estimation bandwidth, leaving us the flexibility to combine coarser and finer bandwidths at-a-time. Since the fine scale coefficients are nearly homoscedastic normal random variables, the MLPT coefficients are approximately homoscedastic within each resolution level. Level dependent thresholds for coefficients with homoscedastic normal errors can be found using classical methods. The results in the analysis of Figures 2 and 3 were obtained with minimum Generalised Cross Validation thresholds. The MLPT estimator clearly preserves more

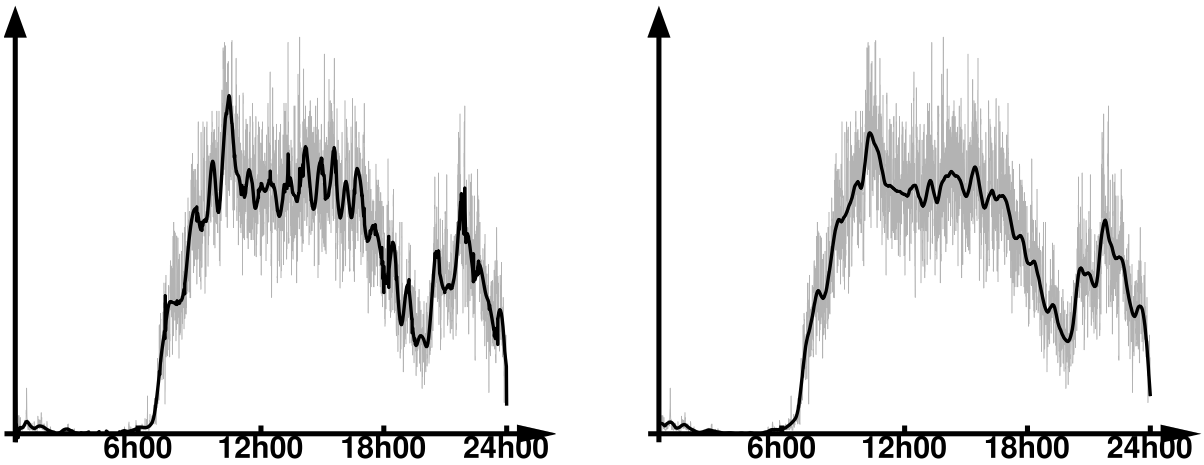

Figure 2: Left panel: MLPT threshold density estimator, further developed in Figure 3. Right panel: single bandwidth kernel density estimator. The grey curve in both panels is the fine scale pilot estimator $\boldsymbol{S}_J$.

details and sharp transitions than the kernel density estimator.

7 Concluding discussions and outlook

For application in density estimation, this article has equipped the MLPT with a novel refinement scheme for a multiscale analysis of highly non-equidistant samples. Thanks to degree of the local polynomial, the adopted kernel function, the finest scale bandwidth and the number of primal vanishing moments in the detail basis functions, the MLPT offers a great variety of analysis, basis functions and reconstructions, in a way quite similar to the multitude of wavelet transforms and basis functions.

With the interpolating Deslauriers-Dubuc wavelet transforms, the MLPT shares the advantage that function values are valid fine scaling coefficients. In the context of density estimation , this motivates the use of a pilot or prototype estimator which is nearly unbiased. Its variance is allowed to be large, although heavy tails should be avoided. The evaluation of the pilot estimator can then be used as fine scaling coefficients $\boldsymbol{S}_J$.

From there, the density estimation problem can be rewritten as a high-dimensional, sparse, generalised linear model with exponential response, for which this article provides a MLPT analysis, including a finetuning of the sparse variable selection and estimation using an information criterion based on a generalised notion of degrees of freedom.

Further research may focus on the choice of the MLPT parameters on the smoothness and numerical properties of the transform. In particular, the control of variance propagation throughout the analysis, especially near the boundaries of a finite interval, requires further attention.

The MLPT approach is mostly complementary to classical kernel density estimation for smooth den-

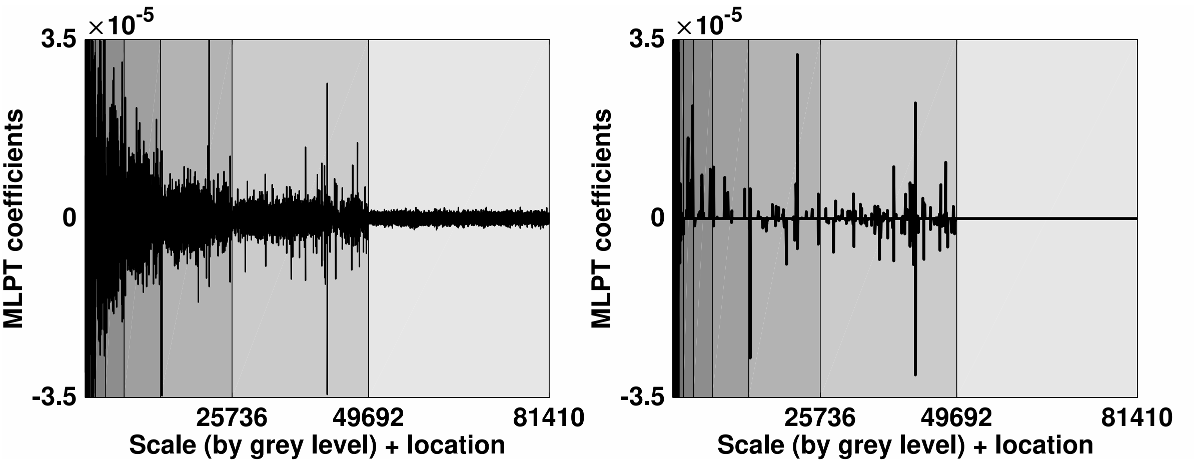

Figure 3: Left panel: MLPT coefficients from the fine scaling coefficients S_J in Figure 2. Right panel: Thresholded MLPT coefficients using minimum Generalised Cross Validation thresholds.

sity functions. This is because the MLPT of a smooth function shows no sparsity in the detail coefficients. Indeed, the detaul coefficients of a smooth function are all (near-)zero, thus making any coefficient selection by thresholding pointless. Any MLPT based smooth density estimation should then operate on the coarse approximation only, which is complementary to the approach proposed in this paper. As a result, the performance of the method in this paper cannot be compared with kernel density estimation.

Instead, future analysis of the performance of a MLPT based kernel density estimation should first consider coarse scale approaches for smooth densities, selecting the coarsest scale such that its performance competes with the well known asymptotic results for kernel density estimation. The introduction of fine scale detail analysis has to proceed without undoing the coarse scale performance of smooth densities. On top of that, the fine scale details can be used to reconstruct singularities at unknown locations. Asymptotic study of the preformance, which is beyond the scope of the current paper, is a topic of onging and future research.

References

L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19:135–144, 1977.

P. J. Burt and E. H. Adelson. Laplacian pyramid as a compact image code. *IEEE Trans. Commun.*, 31 (4):532–540, 1983.

S. Chen and D. L. Donoho. Atomic decomposition by basis pursuit. Technical Report 479, Department of Statistics, Stanford University, May 1995.

G. Deslauriers and S. Dubuc. Symmetric iterative interpolation processes. *Constructive Approximation*, 5:49–68, 1989.

L. Desmet, I. Gijbels, and A. Lambert. Estimation of irregular probability densities. *Institute of Mathematical Statistics Collections*, 7:46–61, 2010.

L. Devroye. *Non-Uniform Random Variate Generation*. Springer, 1986.

D. L. Donoho and T. P. Y. Yu. Deslauriers-Dubuc: Ten years after. In S. Dubuc and G. Deslauriers, editors, *Spline Functions and the Theory of Wavelets*, CRM Proceedings and Lecture Notes. American Mathematical Society, 1999.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2):508–539, 1996.

B. Efron. How biased is the apparent error rate of a prediction rule? *J. American Statistical Association*, 81(394):461–470, 1986.

J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications*. Chapman and Hall, London, 1996.

X. Gao and Y. Fang. A note on the generalized degrees of freedom under the l_1 loss function. *Journal of Statistical Planning and Inference*, 141:677–686, 2011.

G. Geenens. Probit transformation for kernel density estimation on the unit interval. *J. American Statistical Association*, 109:346–358, 2014.

P. Hall and P. Patil. Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *The Annals of Statistics*, 23(3):905–928, 1995.

O. Hossjer and D. Ruppert. Asymptotics for the transformation kernel density estimator. *The Annals of Statistics*, 23:1198–1222, 1995.

M. Jansen. Non-equispaced B-spline wavelets. *International Journal of Wavelets, Multiresolution and Information Processing*, 14(6), 2016. doi: 10.1142/S0219691316500569.

M. Jansen and M. Amghar. Multiscale local polynomial decompositions using bandwidths as scales. *Statistics and Computing*, 27(5):1383–1399, 2017. doi: 10.1007/s11222-016-9692-8.

D. O. Loftsgaarden and C. P. Queensberry. A nonparametric estimate of a multivariate density function. *Ann. Math. Stat.*, 36:1049–1051, 1965.

M. Nunes, M. Knight, and G. P. Nason. Adaptive lifting for nonparametric regression. *Statistics and Computing*, 16(2):143–159, 2006.

B. U. Park, S. S. Chung, and K. H. Seog. An empirical investigation of the shifted power transformation method in density estimation. *Computational Statistics and Data Analysis*, 14:183–191, 1992.

R. Pyke. Spacings (with discussion). *Journal of the Royal Statistical Society, Series B*, 27:395–449, 1965.

D. Ruppert and D. B. H. Cline. Bias reduction in kernel density estimation by smoothed empirical transformations. *The Annals of Statistics*, 22:185–210, 1994.

J. S. Simonoff. *Smoothing Methods in Statistics*. Springer series in Statistics. Springer, 1996.

G. R. Terrell and D. W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992.

R. J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

R. J. Tibshirani. Degrees of freedom and model search. *Statistica Sinica*, 25(3):1265–1296, 2015.

R. J. Tibshirani and J. E. Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2): 1198–1232, 2012.

S. Vaiter, C. Deledalle, J. Fadili, G. Peyré, and C. Dossal. The degrees of freedom of partly smooth regularizers. *Annals of the Institute of Statistical Mathematics*, 69(4):791–832, 2017.

S. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36 (2):614–645, 2008.

M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995.

M. P. Wand, J. S. Marron, and D. Ruppert. Transformations in density estimation. *J. American Statistical Association*, 86:343–353, 1991.

L. Wang, Y. Yoy, and H. Lian. Convergence and sparsity of lasso and group lasso in high-dimensional generalized linear models. *Stat. Papers*, 56(3):819–828, 2015.

L. Yang and J. S. Marron. Iterated transformation-kernel density estimation. *J. American Statistical Association*, 94:580–589, 1999.

J. Ye. On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.*, 93:120–131, 1998.

C. You, S. Müller, and J. T. Ormerod. On generalized degrees of freedom with application in linear mixed models selection. *Statistics and Computing*, 26(1-2):199–210, 2016.

B. Zhang, X. Shen, and S. L. Mumford. Generalized degrees of freedom and adaptive model selection in linear mixed-effects models. *Computational Statistics and Data Analysis*, 56(3):574–586, 2012.

H. Zou, T. J. Hastie, and R. J. Tibshirani. On the “degrees of freedom” of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.

Appendices

A Proof of Theorem 1

Let $\partial_j(x) \subset \{0, 1, \ldots, n_j - 1\}$ denote the x dependent set of indices for which $|x - x_{j,k}| \leq h_j^*$. The cardinality $|\partial_j(x)|$ is bounded from above by (A4). Let C be the smallest integer, independent from j, so that $|\partial_j(x)| \leq C$. For the sake of (A1), it also holds that $\widetilde{p} \leq |\partial_j(x)|$.

Then $f_j(x) = \sum_{k \in \partial_j(x)} f(x_{j,k})\varphi_{j,k}(x)$. As $f(x) \in C^{\widetilde{p}+\alpha}[a, b]$, there exists a polynomial $p_x(u)$ of degree $\widetilde{p}-1$ so that $|f(u)-p_x(u)| \leq L|u-x|^{\widetilde{p}}$. The polynomial $p_x(u)$ satisfies $p_x(x) = f(x)$, and being of degree $\widetilde{p}-1$, it can be decomposed as (A1), i.e., $p_x(u) = \sum_{k=0}^{n_j-1} p_x(x_{j,k})\varphi_{j,k}(u)$, from which we find throught (A2,A3) that $|f_j(x) - f(x)| = |f_j(x) - p_x(x)| \leq \sum_{k \in \partial_j(x)} |f(x_{j,k}) - p_x(x_{j,k})| \cdot |\varphi_{j,k}(x)| \leq C \cdot h_j^{*\widetilde{p}} \cdot M$.

B Proof of Proposition 2

The joint density function of $U_{(n;i-1)}$ and $U_{(n;i)}$ is given by $f_{\boldsymbol{U}_{(n;i-1,i)}}(u_1, u_2) = (i-1)i\binom{n}{i}u_1^{i-2}(1-u_2)^{n-i}$. The Jacobian of the transformation $U_1 = V - \alpha\Delta$ and $U_2 = V + (1-\alpha)\Delta$ is given by $J = 1 - \frac{\partial\alpha}{\partial v}(\Delta, V)\Delta$, so we find, on a region defined by $\alpha(t, v)t \leq v \leq 1 - [1 - \alpha(t, v)]t$,

$$f_{\Delta U_{n;i}, V_{n;i}}(t, v) = (i-1)i\binom{n}{i}(v - \alpha(t, v)t)^{i-2}(1 - v - [1 - \alpha(t, v)]t)^{n-i}\left(1 - \frac{\partial\alpha}{\partial v}(t, v)t\right).$$

From this, we find $F_{\Delta U_{n;i}|V_{n;i}}(t|v) = \int_0^t \exp\left[L_{\Delta V_n}(r, v)\right] dr \Big/ \int_0^{m(v,t)} \exp\left[L_{\Delta V_n}(r, v)\right] dr$, where $m(v, t) = \min\left(v/\alpha(t, v), (1-v)/(1-\alpha(t, v))\right)$ and

$$\begin{aligned} L_{\Delta V_n}(t, v) &= \log\left[(v - \alpha t)^{i-2}(1 - v - (1-\alpha)t)^{n-i}\left(1 - \frac{\partial\alpha}{\partial v}t\right)\right] \\ &= (i-2)\log(v - \alpha t) + (n-i)\log(1 - v - (1-\alpha)t) + \log\left(1 - \frac{\partial\alpha}{\partial v}t\right) \end{aligned}$$

The asymptotic distribution of $(n+1)\Delta U_{n;i}|V_{n;i}$ is given by $F_\infty(t|v) = \lim_{n\to\infty} F_{\Delta U_{n;i}|V_{n;i}}(t/(n+1)|v)$. When n and i grow large, the first two terms of $L_{\Delta V_n}(t, v)$ tend to dominate, with an increasingly sharp peak at the origin, followed by a decay. We further assume that n is large enough so that $\frac{\partial L_{\Delta V_n}}{\partial t}(0, v)$ is negative. In the spirit of Laplace's method for the approximation of integrals, we write

$I_n(t,v) = \int_0^t \exp\left[L_{\Delta V_n}(r,v)\right] dr \approx \widetilde{I}_n(t,v)$, where

$$\begin{aligned} \widetilde{I}_n(t,v) &= \int_0^t \exp\left[L_{\Delta V_n}(0,v) + \frac{\partial L_{\Delta V_n}}{\partial t}(0,v) \cdot r\right] dr \\ &= \exp\left[L_{\Delta V_n}(0,v)\right] \int_0^t \exp\left[\frac{\partial L_{\Delta V_n}}{\partial t}(0,v) \cdot r\right] dr \\ &= \exp\left[L_{\Delta V_n}(0,v)\right] \left|\frac{\partial L_{\Delta V_n}}{\partial r}(0,v)\right|^{-1} \left\{1 - \exp\left[\frac{\partial L_{\Delta V_n}}{\partial t}(0,v) \cdot t\right]\right\}. \end{aligned}$$

More precisely, using Assumption *(P1)* in Proposition 2, Appendix C shows that $I_n(t,v)/\widetilde{I}_n(t,v) \to 1$, uniformly in v and t. As a result, the asymptotic distribution can be expressed as

$$F_\infty(t|v) = \lim_{n\to\infty} \frac{1 - \exp\left[\frac{\partial L_{\Delta V_n}}{\partial t}(0,v) \cdot t/(n+1)\right]}{1 - \exp\left[\frac{\partial L_{\Delta V_n}}{\partial t}(0,v) \cdot m(v,t)\right]}, \tag{23}$$

provided that the limit on the right hand side exists. The convergence holds uniformly for v. We have

$$\begin{aligned} \frac{\partial L_{\Delta V_n}}{\partial t}(t,v) &= \frac{i-2}{v-\alpha t}\left(-\alpha - t \cdot \frac{\partial \alpha}{\partial t}\right) + \frac{n-i}{1-v-(1-\alpha)t}\left(-(1-\alpha) + t \cdot \frac{\partial \alpha}{\partial t}\right) \\ &\quad + \left(1 - \frac{\partial \alpha}{\partial v} t\right)^{-1} \left(-\frac{\partial \alpha}{\partial v} - t \cdot \frac{\partial^2 \alpha}{\partial t \partial v}\right). \end{aligned}$$

Defining the function $g_t(v) = t\frac{\partial \alpha}{\partial t}(t,v)$, Assumption *(P2)* states that $g_0(v) = 0$ and from there we also have $\lim_{t\to 0} t\frac{\partial^2 \alpha}{\partial t \partial v}(t,v) = g_0'(v) = 0$. From Assumption *(P1)*, it follows that $\lim_{t\to 0} t\frac{\partial \alpha}{\partial v}(t,v) = 0$. All together, this leads to

$$\frac{\partial L_{\Delta V_n}}{\partial t}(0,v) = -\frac{i-2}{v}\alpha(0,v) - \frac{n-i}{1-v}[1-\alpha(0,v)] - \frac{\partial \alpha}{\partial v}(0,v).$$

As a result, for any $M > 0$, there exists an integer n^*, so that for all $n = n^*, n^*+1, \ldots$ and for all values of v and t, it holds that $\frac{\partial L_{\Delta V_n}}{\partial t}(0,v)m(v,t) < -M$. The denominator in (23) tends to $\lim_{n\to\infty} 1 - \exp\left[\frac{\partial L_{\Delta V_n}}{\partial t}(0,v) \cdot m(v,t)\right] = 1 - 0 = 1$. For the numerator, we find

$$\begin{aligned} &\lim_{n\to\infty} \exp\left[\frac{\partial L_{\Delta V_n}}{\partial t}(0,v) \cdot \frac{t}{n+1}\right] \\ &= \exp\left[-t \lim_{n\to\infty}\left(\frac{i-2}{v} \cdot \frac{\alpha(0,v)}{n+1} + \frac{n-i}{1-v} \cdot \frac{1-\alpha(0,v)}{n+1} + \frac{\partial \alpha}{\partial v}(0,v)\frac{1}{n+1}\right)\right] \\ &= \exp\left[-t\left(\frac{\rho}{v}\alpha(0,v) + \frac{1-\rho}{1-v}[1-\alpha(0,v)]\right)\right], \end{aligned}$$

and so, $F_\infty(t|v) = 1 - \exp(-t/\mu(v))$. with $\mu(v)$ as in the statement of Proposition 2. This limit holds uniformly in v and t. Indeed, we have the inequality $|\exp(-\gamma_n/u) - \exp(-\gamma/u)| \leq |\gamma/\gamma_n - 1|$ where we can substitute $\gamma_n = t(i-2)\alpha(0,v)/(n+2)$, $\gamma = t\rho\alpha(0,v)$ and $u = v$ to find that $\exp\left[-t\left(\frac{i-2}{v}\cdot\frac{\alpha(0,v)}{n+1}\right)\right]$ converges uniformly to $\exp\left[-t\left(\frac{\rho}{v}\alpha(0,v)\right)\right]$. The other terms in the exponential function follow a similar analysis, resulting in a product sequence of three uniformly convergent and bounded sequences.

C A variant of Laplace's method for the approximation of an integral

The uniform convergence of the ratio $I_n(v,s)/\widetilde{I}_n(v,s) \to 1$, defined and used in the proof of Proposition 2, follows from the identification $g_n(t,s) = L_{\triangle V_n}(t,s)$ in the lemma below. The function $L_{\triangle V_n}(t,s)$ can be verified to satisfy all the stated assumptions.

Lemma 6 *Let $g_n(x,y)$ be a sequence of functions defined on $[0,1]\times[0,1]$, so that $\lim_{n\to\infty}\frac{1}{n}g_n(x,y) = g(x,y)$ exists. Assume that the sequence has the following properties.*

(A1) The $g_n(x,y)$ are uniformly bounded from above.

(A2) For any convergent sequence a_n, the functions $\int_0^{a_n}\exp\left[g_n(x,y)/n\right]dx$ converge uniformly to $\int_0^{a_n}\exp\left[g(x,y)\right]dx$.

(A3) The functions $g_n(x,y)$ are continuously differentiable, with $\frac{\partial g_n}{\partial x}(x,y) < 0$. Moreover, there exists a positive constant κ, so that $\frac{\partial g_n}{\partial x}(0,y) < -\kappa$ for any y,

(A4) $\frac{\partial g_n}{\partial x}(0,y) = -\infty$ only when $g_n(0,y) = -\infty$,

(A5) For fixed x, the sequence $\frac{1}{n}\frac{\partial g_n}{\partial x}(x,y)$ converges to the function $\frac{\partial g}{\partial x}(x,y)$, that is continuous w.r.t. y. The convergence is uniform on every compact subset of $A_x = \left\{y\in[0,1],\ \frac{\partial g}{\partial x}(x,y) \text{ is finite}\right\}$.

Then, with $I_n(y) = \int_0^{a_n} e^{g_n(x,y)}dx$, and $\widetilde{I}_n(y) = \frac{\exp\left[g_n(0,y)\right]}{\left|\frac{\partial g_n}{\partial x}(0,y)\right|}\left[1 - \exp\left(-a_n\left|\frac{\partial g_n}{\partial x}(0,y)\right|\right)\right]$, we have a uniform convergence $I_n(y)/\widetilde{I}_n(y) \to 1$ for $y\in[0,1]$ and $n\to\infty$.

Proof. Define $A_{M'}(\rho) = \liminf_{n\to\infty}\left\{y\in[0,1], \sup_{x\in[0,\rho)}\frac{1}{n}\left|\frac{\partial g_n}{\partial x}(x,y)\right| < M'\right\}$, for arbitrary $M' > 0$. The set $A_{M'}(\rho)$ is not empty, as it covers the set $\left\{y\in[0,1], \sup_{x\in[0,\rho)}\left|\frac{\partial g}{\partial x}(x,y)\right| < M' - \epsilon\right\}$ for any $\epsilon > 0$. Then, for arbitrary $\varepsilon > 0$, there exists a $\delta > 0$, so that for any $\xi\in[0,\delta)$, for any $y\in A_{M'}(\delta)$, and for n sufficiently large, we have $\frac{1}{n}\left|\frac{\partial g_n}{\partial x}(\xi,y) - \frac{\partial g_n}{\partial x}(0,y)\right| \leq \left|\frac{1}{n}\frac{\partial g_n}{\partial x}(\xi,y) - \frac{\partial g}{\partial x}(\xi,y)\right| + \left|\frac{\partial g}{\partial x}(\xi,y) - \frac{\partial g}{\partial x}(0,y)\right| + \left|\frac{\partial g}{\partial x}(0,y) - \frac{\partial g_n}{\partial x}(0,y)\right| < \varepsilon$. On the other hand, for $x\in[0,\delta)$, we have $g_n(x,y) - g_n(0,y) = \frac{\partial g_n}{\partial x}(\xi,y)x$, where $\xi\in[0,x]$ depends on x and y. As a result, for any $\varepsilon > 0$, there exists a set $A_{M'}(\delta)\times[0,\delta)$ in which, for n sufficiently large, $g_n(0,y) + \left(\frac{\partial g_n}{\partial x}(0,y) - n\varepsilon\right)x \leq g_n(x,y) \leq g_n(0,y) + \left(\frac{\partial g_n}{\partial x}(0,y) + n\varepsilon\right)x$. Defining $\delta_n = \min(\delta, a_n)$, and for $y \in A_{M'}(\delta)$, this leads to the upper bound

$$\begin{aligned}
\int_0^{a_n} e^{g_n(x,y)}dx &= \int_0^{\delta_n} e^{g_n(x,y)}dx + \int_{\delta_n}^{a_n} e^{g_n(x,y)}dx \\
&\le \int_0^{\delta_n} e^{g_n(0,y)} \cdot \exp\left[\left(\frac{1}{n}\frac{\partial g_n}{\partial x}(0,y)+\varepsilon\right)nx\right]dx + \int_{\delta_n}^{a_n} e^{g_n(x,y)}dx \\
&= \frac{e^{g_n(0,y)}}{\left|\frac{\partial g_n}{\partial x}(0,y)\right| - n\varepsilon}\left\{1-\exp\left[-\delta_n n\left(\frac{1}{n}\left|\frac{\partial g_n}{\partial x}(0,y)\right|-\varepsilon\right)\right]\right\} \\
&\quad + \int_{\delta_n}^{a_n} e^{g_n(x,y)}dx \qquad (24)
\end{aligned}$$

For $x > \delta$, we have $g_n(x,y) \le g_n(\delta,y) \le g_n(0,y) - (\kappa + n\varepsilon)\delta$. The last term in (24) is then further bounded by

$$\begin{aligned}
\int_{\delta_n}^{a_n} e^{g_n(x,y)}dx &= \int_{\delta_n}^{a_n} e^{g_n(x,y)/n}e^{g_n(x,y)(n-1)/n}dx \\
&\le \int_{\delta_n}^{a_n} e^{g_n(x,y)/n}\exp\left[(n-1)g_n(0,y)/n-(n-1)/n(\kappa+n\varepsilon)\delta\right]dx \\
&\sim \frac{\exp\left[(n-1)g_n(0,y)/n\right]\int_{\delta_n}^{a_n} e^{g(x,y)}dx}{\exp\left[(n-1)\kappa\delta/n\right]\exp\left[(n-1)\varepsilon\delta\right]},
\end{aligned}$$

where the asymptotic equivalence holds uniformly in y.

From here, we find $\int_{\delta_n}^{a_n} e^{g_n(x,y)}dx/\widetilde{I}_n(y) \to 0$, uniformly for any choice of ε. Let $\overline{I}_{n,1}(y) = \frac{e^{g_n(0,y)}}{\left|\frac{\partial g_n}{\partial x}(0,y)\right|-n\varepsilon}\left\{1-\exp\left[-\delta_n n\left(\left|\frac{1}{n}\frac{\partial g_n}{\partial x}(0,y)\right|+\varepsilon\right)\right]\right\}$ be the first term in (24), then

$$\frac{\overline{I}_{n,1}(y)}{\widetilde{I}_n(y)} = \frac{\frac{1}{n}\left|\frac{\partial g_n}{\partial x}(0,y)\right|}{\left\{\frac{1}{n}\left|\frac{\partial g_n}{\partial x}(0,y)\right|-\varepsilon\right\}}\frac{\left\{1-\exp\left[-\delta_n n\left(\left|\frac{1}{n}\frac{\partial g_n}{\partial x}(0,y)\right|+\varepsilon\right)\right]\right\}}{\left\{1-\exp\left[-a_n n\left|\frac{1}{n}\frac{\partial g_n}{\partial x}(0,y)\right|\right]\right\}},$$

which converges uniformly to $\frac{\left|\frac{\partial g}{\partial x}(0,y)\right|}{\left|\frac{\partial g}{\partial x}(0,y)\right|+\varepsilon}$. In a similar, though slightly simpler way, a lower bound with arbitrary ε is obtained

$$\frac{\left|\frac{\partial g}{\partial x}(0,y)\right|}{\left|\frac{\partial g}{\partial x}(0,y)\right|-\varepsilon} \le \lim_{n\to\infty}\frac{I_n(y)}{\widetilde{I}_n(y)}.$$

The proof is completed by letting $M' \to \infty$. □

D Proof of Corollary 1

Define the function $\alpha(t,v) = (v-u_0)/t$ with u_0 solving the equation $t = f_X(F_X^{-1}(v))[F_X^{-1}(u_0+t) - F_X^{-1}(u_0)]$. This is, let x_0 the value for which $u_0 = F_X(x_0)$ and x_1 the value for which $u_0+t = F_X(x_1)$, then x_0 and x_1 satisfy $F_X(x_1) - F_X(x_0) = f_X(\overline{\xi})[x_1 - x_0]$, and $F_X(\overline{\xi}) = v = u_0 + \alpha(t,v)t$.

Using de l'Hôpital's rule we find $\alpha(0,v) = \lim_{t\to 0}(v-u_0)/t = -\lim_{t\to 0}\frac{\partial u_0}{\partial t}$. As $\frac{\partial \alpha}{\partial t} = (1/t)\left[-\frac{\partial u_0}{\partial t} - \alpha(t,v)\right]$, it follows straightforwardly that $\lim_{t\to 0} t\frac{\partial \alpha}{\partial t}(t,v) = 0$, which is Asssumption *(P2)* in Proposition 2. As for Asssumption *(P1)*, we have that

$$\frac{\partial \alpha}{\partial v}(t,v) = \frac{1}{t}\left\{1 - \frac{f_X'(\overline{\xi})}{\left[f_X(\overline{\xi})\right]^3}\,\frac{f_X(x_1)f_X(x_0)}{f_X(x_1) - f_X(x_0)}\,t\right\}.$$

There are several ways to obtain this result. For instance, by defintion, we have

$$\frac{\partial \alpha}{\partial v} = \frac{1 - \frac{\partial u_0}{\partial v}}{t},$$

in which we substitute $\frac{\partial u_0}{\partial v}$ using the expression

$$\frac{\partial u_0}{\partial \overline{\xi}} = \frac{dv}{d\overline{\xi}}\frac{\partial u_0}{\partial v} = f_X(\overline{\xi})\frac{\partial u_0}{\partial v}.$$

With $t = u_1 - u_0 = (x_1 - x_0)f_X(\overline{\xi})$ a constant throughout these calculations, we have

$$\frac{\partial(x_1 - x_0)}{\partial \overline{\xi}} = \frac{-t}{\left[f_X(\overline{\xi})\right]^2}\,f_X'(\overline{\xi}),$$

but also

$$\frac{\partial(x_1 - x_0)}{\partial \overline{\xi}} = \frac{\partial u_1}{\partial \overline{\xi}}\frac{dx_1}{du_1} - \frac{\partial u_0}{\partial \overline{\xi}}\frac{dx_0}{du_0} = \frac{\partial u_0}{\partial \overline{\xi}}\left(\frac{1}{f_X(x_1)} - \frac{1}{f_X(x_0)}\right).$$

Putting together these expressions leads to the aforementioned result, which can be further developed into

$$\frac{\partial \alpha}{\partial v}(t,v) = \frac{1}{t}\left\{1 - \frac{f_X'(\overline{\xi})f_X(x_1)f_X(x_0)}{\left[f_X(\overline{\xi})\right]^2 f_X'(\overline{\zeta})}\right\},$$

with $\overline{\zeta} \in (x_0, x_1)$.

Define $D_\alpha = \{(t,v) \in [0,1]\times[0,1]$ s.t.$\overline{\xi}$ exists$\}$, then it is straightforward that $(0,v) \in D_\alpha$ for any $v \in [0,1]$. With r an arbitrary positive real number and
$D_r = \{(t,v) \in D_\alpha; \max(|f'(\overline{\xi})|, 1/|f'(\overline{\xi})|, |f(\overline{\xi})|, 1/|f(\overline{\xi})|) < r\}$,
the corollary follows from letting $r \to \infty$.

E Proof of Lemma 4

Given the memorylessness of the exponential distribution, the Poisson count $N_h - 1$ is independent from the exponentially distributed value in $U_h = X_{N_h} - h$. As a result

$$E(S_h) = E\left(\frac{N_h}{X_{N_h}}\right) = E(N_h) \cdot E\left(\frac{1}{h+U_h}\right) = (h\theta + 1)\int_0^\infty \frac{1}{h+u}\theta \exp(-u\theta)du,$$

leading to (20). The expression (21) follows in a similar way, using the fact that for independent variables X and Y, it holds that $\mathrm{var}(XY) = [E(X)]^2 \,\mathrm{var}(Y) + [E(Y)]^2 \,\mathrm{var}(X) + \mathrm{var}(X)\mathrm{var}(Y)$. For the asymptotic analysis, we have

$$\sigma_{N_h}\frac{N_h - E(N_h)}{X_{N_h}} = \frac{[N_h - E(N_h)]/\sigma_{N_h}}{X_{N_h}/E(X_{N_h})} \cdot \frac{\mathrm{var}(N_h)}{E(X_{N_h})} \overset{\mathrm{d}}{\to} Z\theta \sim N(0, \theta^2).$$

Indeed, in the first factor, the numerator converges in distribution to a normal random variable, while the denominator converges in distribution to a constant. The second factor is $(h\theta)/(h+1/\theta) \to \theta$. We thus find that

$$\frac{\sigma_{N_h}}{\theta}\frac{N_h - E(N_h)}{X_{N_h}} = \frac{S_h - \theta}{\sqrt{\theta/h}} + \frac{\theta - E(N_h)/X_{N_h}}{\sqrt{\theta/h}} \overset{\mathrm{d}}{\to} Z \sim N(0,1). \qquad (25)$$

Using the identity $E(N_h) = \theta E(X_{N_h})$, the second term becomes $\frac{\theta - E(N_h)/X_{N_h}}{\sqrt{\theta/h}} = \sqrt{h\theta}\left(1 - E(X_{N_h})/X_{N_h}\right)$. With $\eta = \varepsilon/\sqrt{h\theta}$, and applying the inequality $P\left(|1 - 1/X| > \eta/(1-\eta)\right) \le P\left(|1-X| > \eta\right)$, followed by Chebyshev's inequality, we find for the second term in (25) that

$$\begin{aligned} P\left(\left|\sqrt{h\theta}\left(1 - \frac{E(X_{N_h})}{X_{N_h}}\right)\right| > \varepsilon\right) &\le P\left(\left|\frac{X_{N_h}}{E(X_{N_h})} - 1\right| > \frac{\varepsilon/\sqrt{h\theta}}{\varepsilon/\sqrt{h\theta}+1}\right) \\ &\le \frac{\mathrm{var}(X_{N_h}/E(X_{N_h}))}{\left[\frac{\varepsilon/\sqrt{h\theta}}{\varepsilon/\sqrt{h\theta}+1}\right]^2} \\ &= \frac{h\theta}{\varepsilon^2}\frac{1/\theta^2}{(h+1/\theta^2)^2}\left(\frac{\varepsilon}{\sqrt{h\theta}} + 1\right)^2 \to 0 \end{aligned}$$

as $h \to \infty$. From this it follows that the first term in (25) is asymptotically standard normal.

For (22), we have

$$\begin{aligned} E(S_{h,r}) &= E(S_{h,r}|N_h > r)P(N_h > r) + E(S_{h,r}|N_h \le r)P(N_h \le r) \\ &= E\left(\frac{r-1}{X_r}\middle| N_h > r\right)P(N_h > r) + E\left(\frac{1}{h+U_h}\right) \cdot E(N_h|N_h \le r)P(N_h \le r), \end{aligned}$$

using the fact that the density of the exponential excess value U_h does not change by the information that $N_h \leq r$. Given that $N_h - 1$ is a Poisson count, it holds that $E(N_h|N_h \leq r)P(N_h \leq r) = \sum_{k=1}^{r} kP(N_h = k) = P(N_h \leq r) + h\theta P(N_h \leq r-1)$. Furthermore, we have that $X_r|N_h > r$ has the distribution of a truncated gamma, $T_{\theta,r}|T_{\theta,r} \leq h$, where $T_{\theta,r} \sim \Gamma(\theta, r)$, from which it follows that

$$\begin{aligned} E\left(\frac{r-1}{X_r}\middle| N_h > r\right) P(N_h > r) &= E\left(\frac{r-1}{T_{\theta,r}}\middle| T_{\theta,r} \leq h\right) P(T_{\theta,r} \leq h) \\ &= \theta P(T_{\theta,r-1} \leq h) = \theta P(N_h > r-1) = \theta P(N_h \geq r). \end{aligned}$$

Assembling the expressions concludes the proof of (22). □