Density estimation using multiscale local polynomial transforms

Maarten Jansen

Abstract The estimation of a density function with an unknown number of singularities or discontinuities is a typical example of a multiscale problem, with data observed at nonequispaced locations. The data are analysed through a multiscale local polynomial transform (MLPT), which can be seen as a slightly overcomplete, non-dyadic alternative for a wavelet transform, equiped with the benefits from a local polynomial smoothing procedure. In particular, the multiscale transform adopts a sequence of kernel bandwidths in the local polynomial smoothing as resolution level dependent, user controlled scales. The MLPT analysis leads to a reformulation of the problem as a variable selection in a sparse, high-dimensional regression model with exponentially distributed responses. The variable selection is realised by the optimisation of the l1-regularised maximum likelihood, where the regularisation parameter acts as a threshold. Fine-tuning of the threshold requires the optimisation of an information criterion such as AIC. This paper develops discussions on results in [9].

1 Introduction

Due to its natural intermittency, the estimation of a non-uniform density can be described as a nonequispaced multiscale problem, especially when the density contains singularities. Indeed, when the number and the locations of the singularties remain unknown, then the estimation procedure is deemed to go through all possible combinations of locations and intersingular distances. Also, since a given bandwidth in a kernel based method may be too small in a regionof low intensity and too large in a region of high intensity, a local choice of the bandwidth can be considered as an instance of multiscale processing, where the bandwidth is seen as a notion of scale.

Maarten Jansen
Université libre de Bruxelles, Belgium, e-mail: maarten.jansen@ulb.ac.be

A popular class of multiscale methods in smoothing and density estimation is based on a wavelet analysis of the data. The classical wavelet approach for density estimation [6, 3] requires an evaluation of the wavelet basis functions in the observed data or otherwise a binning of the data into fine scale intervals, defined by equispaced knots on which the wavelet transform can be constructed. The preprocessing for the equispaced (and possibly dyadic) wavelet analysis may induce some loss of details about the exact values of the observations.

This paper works with a family of multiscale transforms constructed on nonequispaced knots. With these constructions and taking the observations as knots, no information is lost at this stage of the analysis. The construction of wavelet transforms on irregular point sets is based on the lifting scheme [12, 11]. Given the transformation matrix that maps a wavelet approximation at one scale onto the approximation and offsets at the next coarser scale, the lifting scheme factorizes this matrix into a product of simpler, readily invertible operations. Based on the lifting factorization, there exist two main directions in the design of wavelets on irregular point sets. The first direction consists in the factorization of basis functions that are known to be refinable, to serve as approximation basis, termed scaling basis in a wavelet analysis. The wavelet basis for the offsets between successive scales is then constructed within the lifting factorization of the refinement equation, taking into account typical design objectives such as vanishing moments and control of variance inflation. An example of such existing refinable functions are B-spline functions defined on nested grids of knots [8]. The second approach for the construction of wavelets on irregular point sets does not factorize a scheme into lifting steps. Instead, it uses an interpolating or smoothing scheme as a basic tool in the construction of a lifting step from scratch. Using interpolating polynomials leads to the Deslauriers-Dubuc refinement scheme [2, 4]. To this refinement scheme, a wavelet transform can be associated by adding a single lifting step, designed for vanishing moments and control of variance inflation, as in the case of factorized refinement schemes. This paper follows the second approach, using local polynomial smoothing [5, Chapter 3] as a basic tool in a lifting scheme. For reasons explained in Section 2, the resulting Multiscale Local Polynomial Transform (MLPT) is no longer a wavelet transform in the strict sense, as it must be slightly overcomplete. Then, in Section 3, the density estimation problem is reformulated in a way that it can easily be handled by a MLPT. Section 4 discusses aspects of sparse selection and estimation in the MLPT domain for data from a density estimation problem. In Section 5, the sparse selection is finetuned, using information criteria and defining the degrees of freedom in this context. Finally, Section 6 presents some preliminary simulation results and further outlook.

2 The Multiscale Local Polynomial Transform (MLPT)

Let $\boldsymbol{Y}$ be a sample vector from the additive model $Y_i = f(x_i) + \sigma_i Z_i$, where the covariates x_i may be non-equidistant and the noise Z_i may be correlated. The underlying function, $f(x)$, is assumed to be approximated at resolution level J by a linear combination of basis functions $\varphi_{J,k}(x)$, in

$$f_J(x) = \sum_{k=0}^{n_J-1} \varphi_{J,k}(x) s_{J,k} = \Phi_J(x) \boldsymbol{s}_J,$$

where $\Phi_J(x)$ is a row vector containing the basis functions. The choice of coefficients $\boldsymbol{s}_J$ is postponed to the moment when the basis functions are specified. At this moment, one could think of a least squares projection as one of the possibilities.

The Multiscale Local Polynomial Transform (MLPT) [7] finds the sparse coefficient vector $\boldsymbol{v}$ in $\boldsymbol{s}_J = \mathbf{X}\boldsymbol{v}$, using a linear operation $\boldsymbol{v} = \widetilde{\mathbf{X}}\boldsymbol{s}_J$. Just like in wavelet decomposition, the coefficient vector of several subvectors $\boldsymbol{v} = [\, \boldsymbol{s}_L^T \; \boldsymbol{d}_L^T \; \boldsymbol{d}_{L+1}^T \; \ldots \; \boldsymbol{d}_{J-1}^T \,]^T$, leading to the following basis transformation

$$\Phi_J(x)\boldsymbol{s}_J = \Phi_J(x)\mathbf{X}\boldsymbol{v} = \Phi_L(x)\boldsymbol{s}_L + \sum_{j=L}^{J-1} \Psi_j(x)\boldsymbol{d}_j,$$

where we inytroduced $\Phi_L(x)$ and $\Psi_j(x)$ for the submatrices of the transformed basis $\Phi_J(x)\mathbf{X}$, corresponding to the subvectors of the coefficient vector $\boldsymbol{v}$. The detail vectors $\boldsymbol{d}_j$ are associated to successive resolution levels through the decomposition algorithm, corresponding to the analysis matrix $\widetilde{\mathbf{X}}$,
for $j = J-1, J-2, \ldots, L$

- **Subsamplings**, i.e., keep a subset of the current approximation vector, $\boldsymbol{s}_{j+e,e} = \mathbf{J}_j \boldsymbol{s}_{j+1}$, with $\mathbf{J}_j$ a $n_j \times n_{j+1}$ submatrix of the identity matrix.
- **Prediction**, i.e., compute the detail coefficients at scale j as offsets from a prediction based on the subsample.
  $\boldsymbol{d}_j = \boldsymbol{s}_{j+1} - \mathbf{P}_j \boldsymbol{s}_{j+1,e}$
- **Update**, the remaining approximation coefficients. The idea is that $\boldsymbol{s}_j$ can be interpreted as smoothed, filtered, or averaged values of $\boldsymbol{s}_{j+1}$.
  $\boldsymbol{s}_j = \boldsymbol{s}_{j+1,e} + \mathbf{U}_j \boldsymbol{d}_j$

Before elaborating on the different steps of this decomposition, we develop the inverse transform $\mathbf{X}$ by straightforwardly undoing the two lifting steps in reverse order.
for $j = L, L+1, \ldots, J-1$

- **Undo update**, using $\boldsymbol{s}_{j+1,e} = \boldsymbol{s}_j - \mathbf{U}_j \boldsymbol{d}_j$.
- **Undo prediction**, using $\boldsymbol{s}_{j+1} = \boldsymbol{d}_j + \mathbf{P}_j \boldsymbol{s}_{j+1,e}$.

2.1 Local polynomial smoothing as prediction

The transform in this paper adopts a smoothing operation as prediction, thus incorporating the covariate values as parameters of the analysis. As an example, the Nadaraya-Watson kernel prediction leads to

$$P_{j;k,\ell} = \frac{K\left(\frac{x_{j+1,k}-x_{j,\ell}}{h_{j+1}}\right)}{\sum_{l=1}^{n_j} K\left(\frac{x_{j+1,k}-x_{j,l}}{h_{j+1}}\right)}.$$

In this expression, $K(u)$ denotes a kernel function, i.e., a positive function with integral 1. The parameter h_{j+1} is the bandwidth. While in (uniscale) kernel smoothing this is a smoothing parameter, aiming at optimal balance between bias and variance in the estimation, it acts as a user controlled scale parameter in a Multiscale Kernel Transform (MKT). This is in contrast to a discrete wavelet transform, where the scale is inherently fixed to be dyadic, i.e., the scale at level j is twice the scale at level $j+1$. In a MKT, an also in the forthcoming MLPT, the scale can be chosen in a data adaptive way, taking the irrgeularity of the covariate grid into account. For instance, when the covariates can be considered as ordered indepenent realisations from a uniform density, it is recommended that the scale is taken to be $h_j = h_0 \log(n_j)/n_j$ [10]. The scales at fine resolution levels are then a bit larger, allowing them cope with the non-equidistancy of the covariates.

A slightly more complex prediction, adopted in this paper, is based on local polynomial smoothing. It fills the kth row of $\mathbf{P}_j$ with $\boldsymbol{P}(x_{j+1,k})$, where the row vector $\boldsymbol{P}_j(x)$ is given by

$$\boldsymbol{P}_j(x) = \boldsymbol{X}^{(\tilde{p})}(x)\left(\mathbf{X}_j^{(\tilde{p})^T}\mathbf{W}_j(x)\mathbf{X}_j^{(\tilde{p})}\right)^{-1},$$

with the row vector of power functions, $\boldsymbol{X}^{(\tilde{p})}(x) = [1 \quad x \quad \ldots \quad x^{\tilde{p}-1}]$ and the corresponding Vandermonde matrix at resolution level j, $\mathbf{X}_j^{(\tilde{p})} = [\mathbf{1} \;\; \boldsymbol{x}_j \;\; \ldots \;\; \boldsymbol{x}_j^{\tilde{p}-1}]$. The diagonal matrix of weight functions is given by $(\mathbf{W}_j)_{\ell\ell}(x) = K\left(\frac{x-x_{j,\ell}}{h_j}\right)$.

The prediction matrix has dimension $n_{j+1} \times n_j$. This expansive or redundant prediction is in contrast to lifting schemes for critically downsampled wavelet transform, such as the Deslauriers-Dubuc or B-spline refinement schemes. In these schemes, the prediction step takes the form $\boldsymbol{d}_j = \boldsymbol{s}_{j+1,o} - \mathbf{P}_j \boldsymbol{s}_{j+1,e}$, where $\boldsymbol{s}_{j+1,o} = \mathbf{J}_j^c \boldsymbol{s}_{j+1}$, with $\mathbf{J}_j^c$ the $(n_{j+1}-n_j) \times n_{j+1}$ subsampling operation, complementary to $\mathbf{J}_j$. In the MLPT, a critical downsampling with $\mathbf{J}_j$ and $\mathbf{J}_j^c$ would lead to fractal like basis functions [7]. Omission of the complementary subsampling leads to slight redudancy, where n data points are transformed into roughly $2n$ MLPT coefficients, at least if n_j is approximately half of n_{j+1} at each scale. The corresponding scheme is known in signal processing literature as the Laplacian pyramid [1]. With an output size of $2n$, the MLPT is less redundant than the non-decimated wavelet transform (cycle spinning, stationary wavelet transform) which produces outputs of size $n\log(n)$. Nevertheless, the inverse MLPT shares with the non-decimated wavelet transform an additional smoothing occuring in the reconstruction after processing. This is because processed coefficients are unlikely to be exact decompositions of an existing data vector. The reconstruction thus involves some sort of projection.

2.2 The update lifting step

The second lifting step, the update $\mathbf{U}_j$, serves multiple goals, leading to a combination of design conditions [8]. An important objective, especially in the context of density estimation, is to make sure that all functions in $\Psi_j(x)$ have zero integral. When $f_j(x) = \Phi_L(x)\boldsymbol{s}_L + \sum_{j=L}^{J-1} \Psi_j(x)\boldsymbol{d}_j$, then any processing that modifies the detail coefficients $\boldsymbol{d}_j$, e.g., using thresholding, preserves the integral of $f_j(x)$, which is interesting if we want to impose that $\int_{-\infty}^{\infty} f_j(x)dx = 1$ for an estimation or approximation of a density function. Another important goal of the update, leading to additional design conditions, is to control the variance propagation throughout the transformation. This prevents the noise on a single observation from proceeding unchanged all the way to coarse scales.

2.3 The MLPT frame

Examples of MLPT functions are depicted in Figure 1. It should be noted that these functions are defined on an irregular grid of knots. Nothing of the grid irregularity is reflected in the approximation and detail functions $\Phi_L(x)$ and $\Psi_j(x)$. Also, as the detail functions form an overcomplete set, they are not basis functions in the strict sense. Instead, the set of $\Phi_L(x)$ and $\Psi_j(x)$ for $j = L, L+1, \ldots, J-1$ is called a frame.

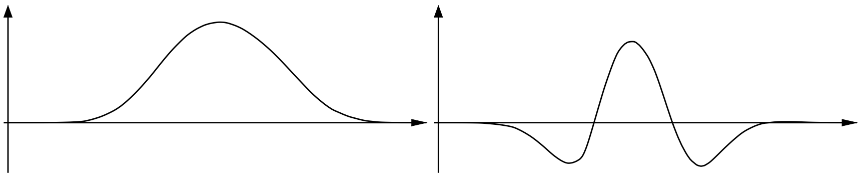


Fig. 1 Left panel: approximation function, i.e., one element of $\Phi_L(x)$. Right panel: detail or offset function, i.e., one element of $\Psi_j(x)$. It holds that $\int_{-\infty}^{\infty} \Psi_j(x)dx = \mathbf{0}_j^T$.

Unlike in a B-spline wavelet decomposition, observation in the knots are valid fine scale approximation coefficients [9]. More precisely, the approximation

$$f_J(x) = \sum_{i=1}^{n} f(x_i)\varphi_{J,i}(x),$$

has a convergence rate equal to that of least squares projection. This property is important when incorporating a MLPT model into the regression formulation of the problem of the density estimation problem in Section 3.

2.4 The MLPT on highly irregular grids

The regression formulation of the density estimation problem in Section 3 will lead to regression on highly irregular grids, that is, grids that are far more irregular than ordered observations from a random variable. On these grids, it is not possible to operate at fine scales, even if these scales are a bit wider than in the equidistant case, as discussed in Section 2.1. In order to cope with the irregularity, the fine scales would be so wide that fine details are lost, and no asymptotic result would be possible. An alternative solution, adopted here, is to work with dyadic scales, but only processing coefficients that have sufficient nearby neighbours within the current scale. Coefficients in sparsely sampled neighbourhoods are forwarded to coarser scales. The implementation of such a scheme requires the introduction of a smooth transition between active and non-active areas at each scale [9]. More precisely, the reconstruction from the local polynomial prediction $\boldsymbol{s}_{j+1} = \boldsymbol{d}_j + \mathbf{P}_j \boldsymbol{s}_{j+1,e}$, is replaced by a weighted form

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

The diagonal matrix $\mathbf{Q}_{j+1}$ has values between 0 and 1. The value is 0 when a coefficient is not surrounded by enough neighbours to apply a regular local polynomial prediction $\mathbf{P}_j$, and it gradually (not suddenly, that is) tends to one in areas with sufficiently dense observations to apply proper polynomial prediction.

3 A regression model for density estimation

Let $\boldsymbol{X}$ be a sample of independent realisation from an unknown density $f_X(x)$ on a bounded interval, which we take, without loss of generality, to be $[0, 1]$. The density function has an unknown number of singularities, i.e., points $x_0 \in [0, 1]$ where $\lim_{x \to x_0} f_X(x) = \infty$, as well as other discontinuities.

As in [9], we consider the spacings $\Delta X_{n;i} = X_{(n;i)} - X_{(n;i-1)}$, i.e., the differences between the successive ordered observations $X_{(n;i)}$. Then, by the mean value theorem, we have that there exists a value $\overline{\xi}_{n;i} \in [X_{(n;i-1)}, X_{(n;i)}]$ for which $f_X(\overline{\xi}_{n;i})\Delta X_{n;i} = \Delta U_{n;i}$, where $\Delta U_{n;i} = U_{(n;i)} - U_{(n;i-1)} = F_X(X_{(n;i)}) - F_X(X_{(n;i-1)})$. Unfortunately, the value of $\overline{\xi}_{n;i}$ cannot be used as such in the subsequent asymptotic result, due to technical issues in the proof. Neverthless, for a fairly free choice of $\xi_{n;i} \in [X_{(n;i-1)}, X_{(n;i)}]$, close to $\overline{\xi}_{n;i}$, the theorem provides nonparametric regression of $\Delta X_{n;i}$ on $\xi_{n;i}$. For details on the proof, we refer to [9].

Theorem 1. *Let $f_X(x)$ be an almost everywhere twice continuously differentiable density function on $x \in [0, 1]$. Define $A_{M,\delta} \subset [0, 1]$ as the set where $f_X(x) \geq \delta$ and $f_X'(x) \leq M$, with δ, M arbitrary, strictly positive real numbers. Then there exist values $\xi_{n;i} \in [X_{(n;i-1)}, X_{(n;i)}]$, so that with probability one, for all intervals $[X_{(n;i-1)}, X_{(n;i)}] \subset A_{M,\delta}$, the value of $f_X(\xi_{n;i})(n + 1)\Delta X_{n;i}$, given the covariate $\xi_{n;i}$,* *converges in distribution to an exponential random variable, i.e.*

$$f_X(\xi_{n;i})(n+1)\Delta X_{n;i}|\xi_{n;i} \xrightarrow{\mathrm{d}} D \sim \exp(1), a.s.$$

We thus consider a model with exponentially distributed response variable $Y_i = (n+1)\Delta X_{n;i}|\xi_{n;i}$ and the vector of parameters $\theta_i = f_X(\xi_{n;i}) = 1/\mu_i$ with $\mu_i = E(Y_i)$, for which we propose a sparse MLPT model $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$, with $\mathbf{X}$ the inverse MLPT matrix defined on the knots in $\boldsymbol{\xi}$.

The formulation of the density estimation problem as a sparse regression model induces no binning or any other loss of information. On the contrary, the information on the values of X_i is duplicated: a first, approximative copy can be found in the covariate values $\xi_{n;i}$. A second copy defines the design matrix. The duplication prevents loss of information when in subsequent steps some sort of binning is performed on the response variables.

4 Sparse variable selection and estimation in the exponential regression model

For the i.i.d. exponential responses $\boldsymbol{Y} \sim \exp(|\boldsymbol{\theta}|)$ with $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$, and $\mu_i = 1/\theta_i$, the score is given by $\nabla \log L(\boldsymbol{\theta};\boldsymbol{Y}) = \mathbf{X}^T(\boldsymbol{Y}-\boldsymbol{\mu})$, so that the maximum ℓ_1 regularised log-likelihood estimator $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \log L(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_1$ can be found by solving the Karush-Kuhn-Tucker (KKT) conditions

$$\begin{array}{ll} \mathbf{X}_j^T(\boldsymbol{Y}-\boldsymbol{\mu}) = \lambda\mathrm{sign}(\beta_j) & \text{if } \beta_j \neq 0, \\ \left|\mathbf{X}_j^T(\boldsymbol{Y}-\boldsymbol{\mu})\right| < \lambda & \text{if } \beta_j = 0. \end{array}$$

Even if we knew which components of $\boldsymbol{\beta}$ were nonzero, the KKT would still be highly nonlinear. This is in contrast to the additive normal model, where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. The estimator *given the selection* then follows from a shrunk least squares solution. Indeed, with $\mathcal{I}$ the set of selected components, we have $\widehat{\boldsymbol{\beta}}_{\mathcal{I}} = \left(\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}\right)^{-1}\mathrm{ST}_\lambda\left(\mathbf{X}_{\mathcal{I}}^T\boldsymbol{Y}\right)$, where $\mathrm{ST}_\lambda(x)$ is the soft-threshold function. In the case of orthogonal design, i.e., when $\mathbf{X}_{\mathcal{I}}^T\mathbf{X}_{\mathcal{I}}$ is the identity matrix, this reduces to straightforward soft-thresholding in the transformed domain. In the case of non-orthogonal, but still Riesz-stable, design, straightforward thresholding is still a good approximation and a common practice, for instance in B-spline wavelet thresholding. For the model with exponential response, the objective is to find appropriate values of $\boldsymbol{S}_J$, so that $\widehat{\boldsymbol{\beta}} = \mathbf{X}\cdot\mathrm{ST}_\lambda(\widetilde{\mathbf{X}}\boldsymbol{S}_J)$. can be used as estimator. For this we need at least that

(C1) the expected value of $\boldsymbol{S}_J$ is close to $\boldsymbol{\theta}$, so that $E(\widetilde{\mathbf{X}}\boldsymbol{S}_J) \approx \widetilde{\mathbf{X}}\boldsymbol{\theta} = \boldsymbol{\beta}$,
(C2) the MLPT decomposition $\boldsymbol{\beta} = \widetilde{\mathbf{X}}\boldsymbol{\theta}$ is sparse,
(C3) the MLPT decomposition of the errors, $\widetilde{\mathbf{X}}(\boldsymbol{S}_J - \boldsymbol{\theta})$ has no outliers, i.e., no heayvy tailed distributions.

As $\theta_i = 1/\mu_i = 1/E(Y_i)$ it may be interesting to start the search for appropriate fine scale coefficients $S_{J,i}$ from $S_{J,i}^{[0]} = 1/Y_i$. Unfortunately, $S_{J,i}^{[0]}$ is heavy tailed. Experiments show that the heavy tails cannot be dealt properly by truncation of $1/Y_i$ in $S_{J,i}^{[1]} = \min(1/Y_i, s_{\max})$ without loss of substantial information about the position and nature of the singular points in the density function.

Therefore, a prefilter with a binning effect is proposed, however keeping track of the original values of $\boldsymbol{Y}$ through the covariate values in the design $\mathbf{X}$. More precisely, let

$$\boldsymbol{S}_J = \boldsymbol{\Pi}\mathbf{D}_{h_{J,0}}\widetilde{\boldsymbol{\Pi}}\boldsymbol{S}_J^{[0]}. \tag{2}$$

The matrices $\widetilde{\boldsymbol{\Pi}}$ and $\boldsymbol{\Pi}$ represent a forward and inverse, one coefficient at-a-time, Unbalanced Haar transform defined on the data adaptive knots $t_{J,i} = \sum_{k=0}^{i-1} Y_k$ and $t_{J,0} = 0$. An Unbalanced Haar transform on the vector of knots $\boldsymbol{t}_J$ is defined by

$$\begin{aligned} s_{j,k} &= \frac{\Delta_{j+1,2k}s_{j+1,2k} + \Delta_{j+1,2k+1}s_{j+1,2k+1}}{\Delta_{j,k}} = \frac{\Delta_{j+1,2k}s_{j+1,2k} + \Delta_{j+1,2k+1}s_{j+1,2k+1}}{\Delta_{j+1,2k} + \Delta_{j+1,2k+1}}, \\ d_{j,k} &= s_{j+1,2k+1} - s_{j,k}, \end{aligned}$$

where $\Delta_{J,k} = t_{J,k} - t_{J,k} = Y_k$ and $\Delta_{j,k} = \Delta_{j+1,2k} + \Delta_{j+1,2k+1}$. In the coefficient at-a-time version, the binning operation $\Delta_{j+1,2k} + \Delta_{j+1,2k+1}$ takes place on a single pair $\Delta_{j+1,k}$ and $\Delta_{j+1,k+1}$, chosen so that $\Delta_{j,k} = \Delta_{j+1,k} + \Delta_{j+1,k+1}$ is as small as possible. Finally, the diagonal matrix $\mathbf{D}_{h_{J,0}}$ in (2), replaces all details $d_{j,k}$ by zero for which the scale $\Delta_{j,k}$ is smaller than a minimum width h_J. The overall effect of (2) is that small values in $\boldsymbol{Y}$ are recursively added to their neighbours until all binned values are larger than $h_{J,0}$. For values of $h_{J,0}$ sufficiently large, it can be analysed that the coefficients of $\boldsymbol{S}_J$ are close to being normally distributed with expected value asymptotically equal to $\boldsymbol{\theta}$ and variance asymptotically equal to $\boldsymbol{\theta}/h_{J,0}$ [9]. Unfortunately, a large value of $h_{J,0}$ also introduces binning bias. In order to reduce this bias and to let $h_{J,0}$ be sufficiently large, the choice of $h_{J,0}$ is combined with a limit on the number of observations in one bin [9].

5 Finetuning the selection threshold

The estimator $\widehat{\boldsymbol{\beta}} = \mathbf{X} \cdot \mathrm{ST}_\lambda(\widetilde{\mathbf{X}}\boldsymbol{S}_J)$. applies a threshold on the MLPT of $\boldsymbol{S}_J$. The input $\boldsymbol{S}_J$ is correlated and heteroscedastic, while the transform is not orthogonal. For all these reasons, the errors on $\widetilde{\mathbf{X}}\boldsymbol{S}_J$ are correlated and heteroscadastic. In an additive normal model where variance and mean are two seperate parameters, the threshold would be taken proporional to the standard deviation. In the context of the exponential model with approximate variance function $\mathrm{var}(S_{J,i}) = E(S_{J,i})/h_{J,0}$, coefficients with large variances tend to carry more information, i.e., they have a larger expected value as well. As a result, there is no argument for a threshold linearly depending on the local standard deviation. This paper adopts a single threshold for all coefficients to begin with. Current research also investigates the use of block thresholding methods.

The threshold or any other selection parameter can be finetuned by optimatisation of the estimated distance between the data generating process and the model under consideration. Estimation of that distance leads to an information criterion. This paper works with an Akaike's Information Criterion for the estimation of the Kullback-Leibler distance. As data generating process, we consider the (asymptotic) independent exponential model for the spacings, and not the asymptotic additive, heteroscedastic normal model for $\boldsymbol{S}_J$. This choice is motivated by the fact that a model for $\boldsymbol{S}_J$ is complicated as it should account for the correlation structure, while the spacings are nearly independent. Moreover, finetuning w.r.t. the spacings is not affected by the loss of information in the computation of $\boldsymbol{S}_J$.

The resulting information criterion is given by the sum of two terms, $\mathrm{AIC}(\widehat{\boldsymbol{\theta}}) = \widehat{\ell}(\widehat{\boldsymbol{\theta}}) - \widehat{\nu}(\widehat{\boldsymbol{\theta}})$. The first term is the empirical log-likelihood

$$\widehat{\ell}(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \left[\log(\widehat{\theta}_i) - \widehat{\theta}_i Y_i\right],$$

while the second term is an estimator of the degrees of freedom

$$\nu(\widehat{\boldsymbol{\theta}}) = E\left[\widehat{\boldsymbol{\theta}}^T(\boldsymbol{\mu} - \boldsymbol{Y})\right].$$

The degrees of freedom are also the bias of $\widehat{\ell}(\widehat{\boldsymbol{\theta}})$ as estimator of the expected log-likelihood, taken over the unknown data generating process. The expected loglilelihood in its turn is the part of the Kullback-Leibler distance that depends on the estimated parameter vector.

An estimator of the degrees of freedom is developed in [9], leading to the expression

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

where $\widehat{\boldsymbol{\Theta}}^{-1}$ is a diagonal matrix with slightly shifted versions of the observed values, i.e., $\widehat{\Theta}_{ii}^{-1} = Y_{i-1}$. The matrix $\boldsymbol{\Upsilon}$ is a diagonal matrix with the observations, i.e., $\Upsilon_{ii} = Y_i$. The diagonal matrix $\mathbf{D}_\lambda$ has zeros and ones on the diagonal. The ones correspond to nonzero coefficients in the thresholded MLPT decomposition.

6 Illustration and concluding discussion

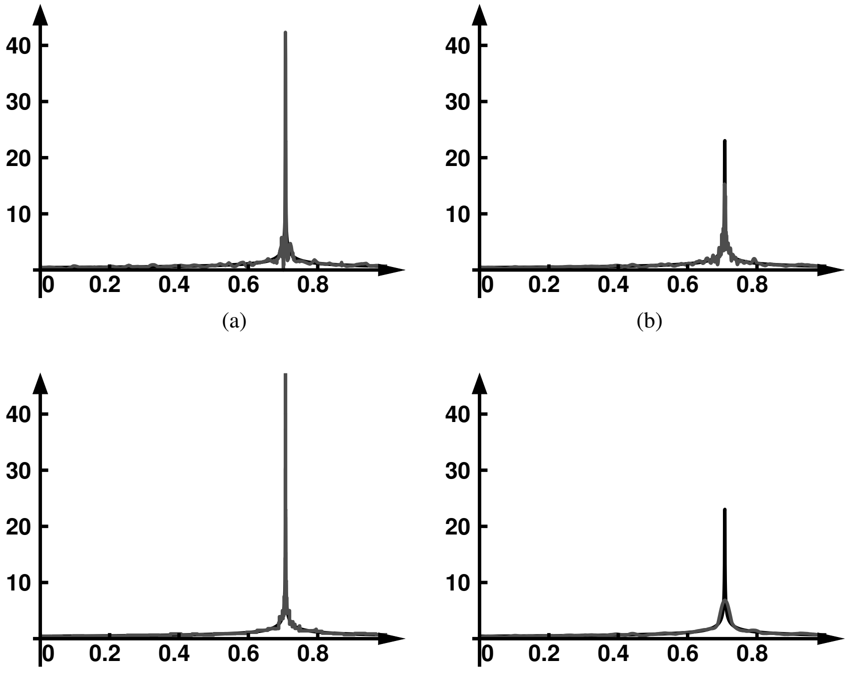
Ongoing research concentrates on motivated choices for the tuning parameters in the proposed data transformation and processing. In particular, the data transformation depends on the choice of the finest resolution bandwidth h_J, the degree of the local polynomial in the prediction step, the precise design of the update step. Also the Unbalanced Haar prefilter is parametrised by a fine scale $h_{J,0}$. Processing parameters include the threshold value, which is selected using the AIC approach of Section 5, and the sizes of the blocks in the block threshold procedure.

For the result in Figure 2, the MLPT adopted a local linear prediction step. In the wavelet literature, the transform is said to have two dual vanishing moments, i.e., $\tilde{p} = 2$, meaning that all detail coefficients of a linear function are zero. The MLPT for the figure also includes an update step designed for two primal vanishing moments, meaning that $\int_{-\infty}^{\infty} \Psi_j(x) x^r \, dx = 0$ for $r = 0$ and $r = 1$. Block sizes were set to one, i.e., classical thresholding was used.

The density function in the simulation study is the power law $f_X(x) = K|x - x_0|^k$ on the finite interval $[0, 1]$, with a singular point $x_0 = 1/2$ in this simulation study and $k = -1/2$. The sample size is $n = 2000$. The MLPT approach, unaware of the presence and location of x_0, is compared with a kernel density estimation applied to a probit transform of the observations, $Y = \Phi^{-1}(X - x_0)$ for $X > x_0$ and $Y = \Phi^{-1}(X - x_0 + 1)$ for $X < x_0$. This transform uses the information on the singularity's location, in order to create a random variable whose density has no end points of a finite interval, nor any singular points inside. In this experiment, the MLPT outperforms the Probit transformed kernel estimation, both in the reconstruction of the singular peak and in the reduction of the oscillations next to the peak. With the current procedure, this is not always the case. Further reserch concentrates on the design making the MLPT analyses as close as possible to orthogonal projections, using appropriate update steps. With an analysis close to orthogonal projection, the variance propagation throughout the analysis, processing and reconstruction can be more easily controlled, thereby reducing spurious effects in the reconstruction. Both MLPT and Probit transformation outperform a straightforward uniscale kernel density estimation. This estimation, illustrated the Figure 2(d), oversmooths the sharp peaks of the true density.

References

1. P. J. Burt and E. H. Adelson. Laplacian pyramid as a compact image code. *IEEE Trans. Commun.*, 31(4):532–540, 1983.
2. G. Deslauriers and S. Dubuc. Symmetric iterative interpolation processes. *Constructive Approximation*, 5:49–68, 1989.
3. 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.
4. 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.
5. J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications*. Chapman and Hall, London, 1996.
6. 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.
7. M. Jansen. Multiscale local polynomial smoothing in a lifted pyramid for non-equispaced data. *IEEE Transactions on Signal Processing*, 61(3):545–555, 2013.
8. M. Jansen. Non-equispaced B-spline wavelets. *International Journal of Wavelets, Multiresolution and Information Processing*, 14(6), 2016.



(a)

(b)

(c)

(d)

Fig. 2 Pannel (a): power law and its estimation from $n = 2000$ observations using the MLPT procedure of this paper. Pannel (b): estimation from the same observations using a probit transform centered around the location of the singularity, thus hinging on the knowledge of this location. Pannel (c): estimation using the finest possible Haar wavelet transform. This transform involves full processing of many resolution levels. Pannel (d): estimation using straightforward uniscale kernel density estimation.

9. M. Jansen. Density estimation using multiscale local polynomial transforms. Technical report, Department of Mathematics, ULB, 2019.
10. M. Jansen and M. Amghar. Multiscale local polynomial decompositions using bandwidths as scales. *Statistics and Computing*, 27(5):1383–1399, 2017.
11. M. Jansen, G. Nason, and B. Silverman. Multiscale methods for data on graphs and irregular multidimensional situations. *Journal of the Royal Statistical Society, Series B*, 71(1):97–125, 2009.
12. W. Sweldens. The lifting scheme: a construction of second generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546, 1998.