TECHNICAL REPORT

0425

POISSON INVERSE PROBLEMS

ANTONIADIS A. and BIGOT J.

IAP STATISTICS NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

POISSON INVERSE PROBLEMS

Anestis Antoniadis,
Laboratoire IMAG-LMC, University Joseph Fourier,
BP 53, 38041 Grenoble Cedex 9, France.
and
Jéremie Bigot,
Institute of Statistics, UCL,
Louvain-la-Neuve, Belgium.

September 2004

Abstract

In this paper, we focus on nonparametric estimators in inverse problems for Poisson processes involving the use of wavelet decompositions. Adopting an adaptive wavelet Galerkin discretization we find that our method combines the well know theoretical advantages of wavelet-vaguelette decompositions for inverse problems in terms of optimally adapting to the unknown smoothness of the solution, together with the remarquably simple closed form expressions of Galerkin inversion methods. Adapting the results of Barron and Sheu [2] to the context of log-intensity functions approximated by wavelet series with the use of the Kullback-Leibler distance between two point processes, we also present an asymptotic analysis of convergence rates that justify our approach. In order to shade some light on the theoretical results obtained and to examine the accuracy of our estimates in finite samples we illustrate our method by the analysis of some simulated examples. The paper ends with some concluding remarks.

Key words and phrases: Poisson Process; Integral equation; intensity function; adaptive estimation; Besov spaces; Galerkin inversion; wavelets, wavelet thresholding.
AMS 1991 subject classifications: Primary 62G07; secondary 65Dxx.

1 INTRODUCTION

In this article, the problem of estimating the intensity function of an indirectly observed, non-homogeneous Poisson process is considered. More specifically, let F be a non-observable Poisson process on a measure space (E_0, μ_0) and let $tf(x)$ be its intensity function with respect to the measure μ_0 on E_0. The function f is the function of interest and t is the "observation time" which will tend to infinity in asymptotic considerations. The observable data form another Poisson process G on, possibly another, measure space (E_1, μ_1) with an intensity function $th(y)$ with respect to a measure μ_1. The functions f and h, considered as elements of the separable Hilbert spaces $L^2(E_0, \mu_0)$ and $L^2(E_1, \mu_1)$, are related by an operator equation $h = Kf$ for some linear compact operator K mapping $L^2(E_0, \mu_0)$ into $L^2(E_1, \mu_1)$. In many applications, K is an integral operator with a kernel representing the response of a measuring device; in the special case where this linear device is translation-invariant, K reduces to a convolution operator. Observing the point process G must be understood in a measure sense. Assuming therefore that for any $v \in L^2(E_1, \mu_1)$, we observe $\int v dG$, the natural goal is to estimate f but this is a problem that is ill posed in general because K^{-1} may not be continuous. Examples range from all kinds of image debluring models, mathematical models for positron emission tomography and nuclear magnetic resonance, or unfolding problems in stereology and high energy physics, to cite only few.

Related problems of inverse estimation have been proposed in the literature, including smoothing kernel methods (Taylor [25]; Hall and Smith [12]), smoothing spline methods (Nychka and Cox [19], Nychka, Wahba, Goldfarb and Pugh [20]), Gauss-Chebyshev type quadrature methods for solving integral equations (Mase [16]), and singular value decomposition (SVD) methods (Johnstone and Silverman [11], O'Sullivan [21]; Silverman, Jones, Wilson and Nychka [23]) to cite only a few. Wavelet and multiscale analysis regularization methods for inverse problems have also recently received considerable attention in the statistics literature exploiting the fact that wavelets provide unconditional bases for a large variety of smoothness spaces. Donoho [8] introduced a wavelet-like decomposition which is specifically adapted to describe the action of K, the so-called wavelet-vaguelette decomposition (WVD), and proposed to apply a thresholding algorithm on this decomposition. In Abramovich and Silverman [1], this method was compared with the similar vaguelette-wavelet decomposition (VWD) algorithm. However, most of the techniques developed to date are based on Gaussian noise models which are not directly applicable in Poisson inverse problems. As those results are not directly applicable, an alternative approach has been proposed by Szkutnik [24] using a quasi-maximum likelihood (QML) histogram sieve estimator when restricting h to step functions. We explore in the sequel an alternative approach via wavelet based decompositions combined with thresholding strategies that address adaptivity issues. Our approach is encouraged by recent attempts to wed a multiscale analysis paradigm with Poisson estimation inverse problems introduced in Nowak and Kolaczyk [18] and also in the work of Cavalier and Koo [3] on hard threshold estimators in the tomographic data framework. To derive estimators whose constructions require as few prior knowledge assumptions on f (such as smoothness assumptions for instance) and to obtain near optimal rates of convergence for such estimators we will transfer to the Poisson case wavelet Garlekin based procedures introduced by Cohen, Hoffman and Reiss [5] in the white noise framework.

Using some concentration inequalities for suprema of integral functionals of Poisson processes which are analogous to Talagrand's inequalities for empirical processes, our method will lead to a remarkably simple analysis tool for Poisson inverse problems that is analogous to wavelet-vaguelette based methods used in Gaussian white noise inverse problems.

The Garlekin approach of Cohen, Hoffman and Reiss [5] is relatively easy to describe when the inversion problem is a white noise embedded problem. Unfortunately, as stressed above this is not the case for Poisson inverse problems. In Section 2, we propose an equivalent formulation of such problems that involves a notion of information projection similar to the one used by Barron and Sheu [2] for estimating a density. In Section 3 a linear estimator for the linear inverse problem at hand is proposed using the appropriate type of wavelets adapted to our case. In the spirit of wavelet denoising methods (Donoho [8]; Abramovich and Silverman [1]), and in order to gain in adaptivity, we then improve the estimator by applying a soft-threshold nonlinearity to the vaguelette coefficients.

The last section is devoted to the numerical implementation of our procedures. We present the results of a small Monte Carlo experiment designed to study the finite sample behavior of our estimates.

2 Preliminaries and notations

In this section we establish the notation and the general framework of the models which are adopted in this paper for the Poisson inverse problem formulated in the introduction. For this purpose let $F = \sum_{i=1}^{\tau} \delta_{X_i}$ and $G = \sum_{i=1}^{\tau} \delta_{R_i}$ be two point Poisson processes on some Borel measurable spaces $(E_0, \mathcal{B}(E_0), \mu_0)$ and $(E_1, \mathcal{B}(E_1), \mu_1)$ respectively, where δ_x denotes the Dirac measure with support $\{x\}$. Associated with these Poisson processes are the *intensity measures* defined by:

- $\lambda_F(B) = \mathbb{E}(F(B)) = \int_B tf(x)d\mu_0(x),\ B \in \mathcal{B}(E_0)$
- $\lambda_G(B') = \mathbb{E}(G(B')) = \int_{B'} th(x)d\mu_1(x),\ B' \in \mathcal{B}(E_1)$

where t is an "observation time" which will tend to infinity in our asymptotic considerations. Observing the process G, we consider the problem of estimating the rate or intensity function f, when the intensity h results from the action of a compact self-adjoint positive definite operator $K : L^2(E_0, \mu_0) \to L^2(E_1, \mu_1)$ on the intensity function of the process F, i.e. $h = Kf$. To simplify the notation, we will assume in the following without any loss of generality that the observation and unknown domains E_0 and E_1 are identical Borel subsets of $\mathbb{R}^d$ $(d \geq 1)$, say E, and that $\mu_0 = \mu_1 = \mu$ where μ denotes the Lebesgue measure. A discussion on how one can handle the case $E_0 \neq E_1$ or K not self-adjoint positive definite is deferred to the end of this paper.

In order to estimate the unknown intensity function f, we will approximate the logarithm of the intensity by a standard wavelet basis function expansion. A notable advantage of using such an exponential family intensity estimation is that it forces positivity of the resulting estimator which is not shared by other kernel or orthogonal series estimators and is also natural with an entropy based loss function. To therefore assess the quality of the estimation, we will measure the discrepancy between an estimator $\hat{f}_t$ and the true intensity function f in the sense of relative entropy (Kullback-Leibler distance) between two point processes:

$$\Delta(f; \hat{f}_t) = \int \left(f \log(\frac{f}{\hat{f}_t}) - f + \hat{f}_t \right) d\mu,$$

where logarithms above are taken with base e. One can show (see Lemma VII. 3 of [3]) that the above distance of the two intensities is also the Kullback-Leibler distance between the corresponding Poisson processes. It is well known that Δ is nonnegative and equals zero if and only if $\hat{f}_t = f$ a.e. The intensity $\hat{f}_t$ in the exponential family that is closest to f in this relative entropy sense is the so-called information projection of f (Csiszár [6]).

The ill-posed nature of the problem comes from the assumption that K is compact and therefore its inverse is not L^2 -bounded. As in Cohen *et al* [5] we will express the ill-posed condition of K by a smoothing action: K will map $L^2(E, \mu)$ into some smoothness space H^t for some $t > 0$. Following the notations in the paper cited above, we will say that K has the smoothing property of order $\nu > 0$, if K maps the Sobolev space H^s onto $H^{s+\nu}$ or the Besov space $B^s_{p,q}$ onto $B^{s+\nu}_{p,q}$. Recall that the Sobolev space $H^s(\mathbb{R})$, $s \in \mathbb{R}$, is the space of tempered distributions v such that

$$\|v\|_s^2 = \int_{\mathbb{R}} (1 + |\xi|^2)^s |\hat{v}(\xi)|^2 \, d\xi < \infty,$$

where

$$\hat{v}(\xi) = \int_{\mathbb{R}} e^{i\xi t} v(t) \, dt,$$

denotes the Fourier transform of v. The Besov spaces form another particular family of smoothness spaces. Essentially the Besov spaces $B^s_{p,q}(\mathbb{R}^d)$ consist, basically, of functions that "have s derivatives in L^p"; the parameter q provides some additional fine-tuning to the definition of these spaces. The norm $\|g\|_{B^s_{p,q}}$ in a Besov space $B^s_{p,q}$ is traditionally defined via the modulus of continuity of f in $L^p(\mathbb{R})$, of which an additional weighted L^q-norm is then taken, in which the integral is over different scales. We shall not give its details here; for our purposes it suffices that this traditional Besov norm is equivalent with a norm that can be computed from wavelet coefficients; a full study can be found in e.g. Meyer [17], Daubechies [7].

For a self-adjoint positive definite operator K, the smoothing property can be expressed by the ellipticity property:

$$\langle Kf, f \rangle \sim \|f\|^2_{H^{-\nu/2}}, \tag{2.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $L^2(E, \mu)$ and $H^{-\nu}$ is the dual space of H^ν appended with appropriate boundary conditions depending on the problem (homogeneous, periodic, etc. . .).

As already explained in the introduction a key ingredient for solving the Poisson inverse problem is the use of standard wavelet bases of $L^2(E, \mu)$ which allow the characterization of the function spaces that describe both the smoothness of the solution and the smoothing action of the operator K since wavelet bases provide also an unconditional bases for a variety of other useful Banach spaces of functions, such as Hölder spaces, Sobolev spaces and, more generally, Besov spaces. Assume that we have a scaling function ϕ and a wavelet function ψ. Scaling and wavelet functions at scale j (i.e. resolution level 2^j) will be denoted by ϕ_λ and ψ_λ, where the index λ summarizes both the usual scale and space parameters j and k (e.g. for one-dimensional wavelets, $\lambda = (j, k)$ and $\psi_{j,k} = 2^{j/2}\psi(2^j \cdot -k)$). If $d \geq 2$, the notation ψ_λ stands for the adaptation of scaling and wavelet functions to multi-dimensional domains. The notation $|\lambda| = j$ will be used to denote a wavelet at scale j, while $|\lambda| < j$ denotes some wavelet at scale j', with $0 \leq j' < j$ (we shall assume, merely for notational convenience, that the usual coarse level of approximation j_0 is equal to 0). With these notations, we assume that:

- The scaling functions $(\phi_\lambda)_{|\lambda|=j}$ span a finite dimensional space V_j within a multiresolution hierarchy $V_0 \subset V_1 \subset \ldots \subset L^2(E, \mu)$, such that $dim(V_j) = 2^{jd}$ (periodic wavelets for notational convenience).

- We assume that we are in the orthonormal case i.e. the scaling functions $(\phi_\lambda)_{|\lambda|=j}$ are an orthonormal basis of V_j, and the wavelets $(\psi_\lambda)_{|\lambda|=j}$ form an orthonormal basis of W_j which is the orthogonal complement of V_j into V_{j+1}.

- For any $g \in L^2(E, \mu)$, its wavelet decomposition can be written as:
$$g = \sum_{|\lambda|=0} \alpha_\lambda \phi_\lambda + \sum_{j \geq 0}^{\infty} \sum_{|\lambda|=j} \beta_\lambda \psi_\lambda,$$
where $\alpha_\lambda = \langle g, \phi_\lambda \rangle$ and $\beta_\lambda = \ g, \psi_\lambda \rangle$.

- To simplify the notation we shall use the convenient slight abuse of notation that sweeps up the coarsest-j scaling functions into the ψ_λ as well, i.e, we will sometimes write $(\psi_\lambda)_{|\lambda|=-1}$ for $(\phi_\lambda)_{|\lambda|=0}$. We thus denote the complete d-dimensional, inhomogeneous wavelet basis by $\{\psi_\lambda; \lambda \in \Lambda\}$. By truncating the wavelet decomposition at level j, we obtain the orthogonal projection onto V_j:
$$P_j g = \sum_{|\lambda|<j} \beta_\lambda \psi_\lambda,$$

- We also assume that $\|\psi_\lambda\|_\infty = \|\psi\|_\infty 2^{|\lambda|d/2}$

Wavelets provide unconditional bases for the Besov spaces, and one can express whether or not a function g on E belongs to a Besov space by a fairly simple and completely explicit requirement on the absolute value of the wavelet coefficients of g. We shall not give its details here; for our purposes it suffices that the traditional Besov norm is equivalent with a norm that can be computed from wavelet coefficients. More precisely, let us assume that the original 1-dimensional ϕ and ψ are in $C^L(\mathbb{R})$, with $L > s$, that $\sigma = s + d(1/2 - 1/p) \geq 0$, and define the norm $\| \cdot \|_{s,p,q}$ by

$$\|g\|_{s,p,q} = \left(\sum_{j=0}^{\infty} \left(2^{j\sigma p} \sum_{\lambda \in \Lambda, |\lambda| = j} |\langle g, \psi_\lambda \rangle|^p \right)^{q/p} \right)^{1/q}.$$

Then this norm is equivalent to the traditional Besov norm, that is, there exist strictly positive constants A and B such that

$$A\|g\|_{s,p,q} \leq \|g\|_{B^s_{p,q}} \leq B\|g\|_{s,p,q}.$$

The condition that $\sigma \geq 0$ is imposed to ensure that $B^s_{p,q}(\mathbb{R}^d)$ is a subspace of $L^2(\mathbb{R}^d)$; we shall restrict ourselves to this case in this paper. When $0 < p, q < 1$, the Besov spaces can still be defined as complete metric spaces, although they are no longer Banach spaces. This allows for more local variability in local smoothness than is typical for functions in the usual Hölder or Sobolev spaces. For instance, a real function g on $\mathbb{R}$ that is piecewise continuous, but for which each piece is locally in C^s, can be an element of $B^s_{p,p}(\mathbb{R})$, despite the possibility of discontinuities at the transition from one piece to the next, provided $p > 0$ is sufficiently small, and some technical conditions are met on the number and size of the discontinuities, and on the decay at ∞ of g.

To end this section and since our estimation procedures will be based on a wavelet Galerkin projection method, we recall here some useful results on linear Galerkin projection methods for solving linear problems with additive white noise of the form $g = Kf + \sigma dW$. For a more detailed description the reader is referred to the paper by Cohen *et . al* [5]. Let $g \in L^2(E, \mu)$, then the function $g_j \in V_j$ is said to be the Galerkin approximation of g if for all $v \in V_j$:

$$\langle Kg_j, v \rangle = \langle Kg, v \rangle.$$

Let $G_j \in \mathbb{R}^{2^{jd}}$ be the vector of wavelet coefficients of $g_j \in V_j$, then the Galerkin projection method for approximating g amounts to solve the following linear system:

$$G_j = K_j^{-1} G^K,$$

where $K_j = (\langle K\psi_\lambda, \psi_\mu \rangle)_{|\lambda|<j,|\mu|<j}$ is a symmetric positive definite matrix, and $G^K = (\langle Kg, \psi_\mu \rangle)_{|\mu|<j}$ is a "data" vector. Now, define the Galerkin wavelets $u_\lambda^j \in V_j$ as:

$$\langle K u_\lambda^j, v \rangle = \langle \psi_\lambda, v \rangle, \text{ for all } v \in V_j.$$

Let U_λ^j be the vector of wavelet coefficients of $u_\lambda^j \in V_j$, then:

$$U_\lambda^j = K_j^{-1} \Upsilon_\lambda,$$

where $\Upsilon_\lambda = (\langle \psi_\lambda, \psi_\mu \rangle)_{|\mu|<j}$ is a vector with zero entries except the λ-th component which is equal to 1. Note that,

$$\begin{aligned} \langle u_\lambda^j, Kg \rangle &= (U_\lambda^j)^T (\langle \psi_\mu, Kg \rangle)_{|\mu|<j} \\ &= (U_\lambda^j)^T G^K \\ &= \Upsilon_\lambda^T K_j^{-1} G^K = \Upsilon_\lambda^T G_j = G_{j,\lambda}, \end{aligned}$$

where $G_{j,\lambda} = \langle g_j, \psi_\lambda \rangle$ denotes the λ-th component of G_j. Hence, if we define $g_j \in V_j$ by $\langle g_j, \psi_\lambda \rangle = \langle Kg, u_\lambda^j \rangle$, then g_j is the Galerkin approximation of g.

3 Information projection based estimation

One of the most important problems confronting an investigator in statistical modeling is the choice of an appropriate model to characterize the underlying data. This determination can often be facilitated by the use of a model selection criterion, which judges the propriety of a fitted model by assessing whether it offers an optimal balance between "goodness of fit" and parsimony. Kullback's directed divergence between the true model and a fitted approximating model is such a criterion. The directed divergence, also known as the Kullback-Leibler information, the I-divergence, or the relative entropy, assesses the dissimilarity between two statistical models.

Information projection for the estimation of density functions has been studied by Barron and Sheu [2]. They obtained various existence results and asymptotic bounds for the distance $\int p \log(p/q)$ between two probability density functions p and q. Their estimation procedure is based on sequences of exponential families spanned by orthogonal functions such as polynomials, splines and trigonometric series. Estimation of density functions by approximation of log-densities with wavelets has been considered by Koo and Kim [13].

In this section, we adapt the results of Barron and Sheu [2] to the context of log-intensity functions approximated by wavelet series with the use of the Kullback-Leibler distance between two point processes. More precisely, let $j \geq 0$. If θ denotes a vector in $\mathbb{R}^{2^{jd}}$, then θ_λ denotes its λ-th component. The wavelet based exponential family $\mathcal{E}_j$ at scale j will be defined as the set of functions:

$$\mathcal{E}_j = \left\{ f_{j,\theta}(.) = \exp(\sum_{|\lambda|<j} \theta_\lambda \psi_\lambda(.)),\ \theta = (\theta_\lambda)_{|\lambda|<j} \in \mathbb{R}^{2^{jd}} \right\}.$$

Following Csiszár [6], the intensity $f_{j,\theta}$ in the exponential family $\mathcal{E}_j$ that is closest to the true intensity f in the relative entropy sense is characterized as the unique intensity function in the family for which $\langle f_{j,\hat{\theta}_t}, \psi_\lambda \rangle = \langle f, \psi_\lambda \rangle$. It seems therefore natural to estimate the unknown intensity function f, by searching for some $\hat{\theta}_t \in \mathbb{R}^{2^{jd}}$ such that:

$$\langle f_{j,\hat{\theta}_t}, \psi_\lambda \rangle = \frac{1}{t} \int u_\lambda^j dG = \hat{\alpha}_\lambda^t, \text{ for all } |\lambda| < j.$$

If there exists a solution to this problem, then $f_{j,\hat{\theta}_t}$ will be called *the Galerkin information projection* estimate of f at scale j.

We already pointed the advantage of such an approach since one can guarantee that the intensity function estimates are positive. The following lemma states some of the I-projection properties onto $\mathcal{E}_j$ (see also Csiszár [6]).

Lemma 3.1 *Let $\alpha \in \mathbb{R}^{2^{jd}}$. Assume that there exists some $\theta(\alpha) \in \mathbb{R}^{2^{jd}}$ such that for all $|\lambda| < j$:*

$$\langle f_{j,\theta(\alpha)}, \psi_\lambda \rangle = \alpha_\lambda.$$

Then, for any intensity function $f \in L^2(E, \mu)$ such that $\langle f, \psi_\lambda \rangle = \alpha_\lambda$ and for all $\theta \in \mathbb{R}^{2^{jd}}$, the following Pythagorian-like identity holds:

$$\Delta(f; f_{j,\theta}) = \Delta(f; f_{j,\theta(\alpha)}) + \Delta(f_{j,\theta(\alpha)}, f_{j,\theta}).$$

A consequence of the above lemma, and since $\Delta(f;h) > 0$ unless $f = h$ almost everywhere, is that $\theta(\alpha)$ (if it exists) uniquely minimizes $\Delta(f; f_{j,\theta})$ for $\theta \in \mathbb{R}^{2^{jd}}$.
From now on assume that there exists some constant $A_j < \infty$ such that for all $v \in V_j$:

$$\|v\|_\infty \leq A_j \|v\|_{L^2}.$$

A key lemma relating distances between the intensities in the parametric family to distance between the corresponding wavelet coefficients is then the following:

Lemma 3.2 *Let $\theta_0 \in \mathbb{R}^{2^{jd}}$, $\alpha_{0,\lambda} = \langle f_{j,\theta_0}, \psi_\lambda \rangle$ and $\alpha \in \mathbb{R}^{2^{jd}}$ a given vector. Let $b = \exp(\|\log(f_{j,\theta_0})\|_\infty)$ and $e = \exp(1)$. If $\|\alpha - \alpha_0\|_2 \leq \frac{1}{2ebA_j}$, then the solution $\theta(\alpha)$ to*

$$\langle f_{j,\theta(\alpha)}, \psi_\lambda \rangle = \alpha_\lambda \textit{ for all } |\lambda| < j$$

exists and satisfies:

$$\begin{aligned} \|\theta(\alpha) - \theta_0\|_2 &\leq 2eb\|\alpha - \alpha_0\|_2 & (3.1) \\ \|\log(\frac{f_{j,\theta(\alpha_0)}}{f_{j,\theta(\alpha)}})\|_\infty &\leq 2ebA_j\|\alpha - \alpha_0\|_2 & (3.2) \\ \Delta(f_{j,\theta(\alpha_0)}; f_{j,\theta(\alpha)}) &\leq 2eb\|\alpha - \alpha_0\|_2^2. & (3.3) \end{aligned}$$

The proof of this lemma relies upon a series of lemmas on bounds within exponential families for the Kullback-Leibler distance and is given in the Appendix A.

4 Linear estimation

Let M be some fixed constant and let $F^s_{p,q}(M)$ denote the set of scaled intensity functions such that

$$F^s_{p,q}(M) = \left\{ f = \exp(g),\ \|g\|_{B^s_{p,q}} \leq M \right\}.$$

Note that assuming that $f \in F^s_{p,q}(M)$ implies that f is strictly positive.

For $f \in F^s_{p,q}(M)$ we define

$$\begin{aligned} D_j &= \|g - P_j g\|_{L^2} \\ \gamma_j &= \|g - P_j g\|_\infty \end{aligned}$$

Basic to our analysis is a decomposition of the relative entropy between the true and the estimated intensity into the sum of two terms which correspond to approximation error and estimation error (bias and variance in a familiar mean squared error analysis). The proof of this result, which relies upon some concentration inequalities for Poisson processes, is postponed in the Appendix A .

Theorem 4.1 *Assume that ψ is compactly supported and that $f \in F^s_{p,q}(M)$ (with $s > d/p \geq d/2$). Let $M_1 > 1$ be a constant such that $M_1^{-1} \leq f \leq M_1$ (see Lemma 7.4), and let $\epsilon_j = 2M_1^2 e^{2\gamma_j+1} D_j A_j$. If $\epsilon_j \leq 1$, the information projection exists, i.e. there exists $\theta_j^* \in \mathbb{R}^{2^{jd}}$ such that:*

$$\langle f_{j,\theta_j^*}, \psi_\lambda \rangle = \langle f, \psi_\lambda \rangle \text{ for all } |\lambda| < j$$

and the approximation error satisfies

$$\Delta(f; f_{j,\theta_j^*}) \leq C e^{\gamma_j} D_j^2$$

Moreover, suppose that ψ is in $H^{s+d/2-\nu}$ (with $s > \nu - d/2$) and has r vanishing moments with $r > s + d/2$. Let $\delta_j^t = 4M_1^2 e^{2\epsilon_j + 2\gamma_j + 2} A_j^2 \rho_{j,t}$, where $\rho_{j,t} = (2^{j(\nu+d/2)}/\sqrt{t} + 2^{j(\nu+\frac{3}{2}d)}/t)^2 + 2^{-2js}$. If $\delta_j^t \leq 1$, then for every $\eta^2 \leq \frac{1}{\delta_j^t}$ there is a set of probability less than $\exp(-\eta)$, such that outside this set there exists some $\hat{\theta}_t \in \mathbb{R}^{2^{jd}}$ which satisfies:

$$\langle f_{j,\hat{\theta}_t}, \psi_\lambda \rangle = \frac{1}{t} \int u_\lambda^j dG \text{ for all } |\lambda| < j$$

and the estimation error satisfies

$$\Delta(f_{j,\theta_j^*}; f_{j,\hat{\theta}_t}) \leq C\eta^2 e^{1+\gamma_j+\epsilon_j} \rho_{j,t}.$$

Note that by using the above theorem, explicit bounds are obtained which are applicable for each finite value of j and t, subject to ϵ_j and $\delta_j^t \leq 1$. We can now state the general result on the nonadaptive Galerkin information projection estimator of the unknown intensity function.

Theorem 4.2 *Assume that ψ is compactly supported and that $f \in F^s_{2,2}(M)$ (with $s > d/2$). Moreover, suppose that ψ is in $H^{s+d/2-\nu}$ (with $s > \nu - d/2$) and has r vanishing moments with $r > s + d/2$. Let $j(t)$ be such that $2^{-j(t)} = (\frac{1}{t})^{1/(2s+2\nu+d)}$. Then, with probability tending to 1 as $t \to \infty$, the Galerkin information projection exists and satisfies*

$$\Delta(f; f_{j(t),\hat{\theta}_t}) \leq O\left((\frac{1}{t})^{2s/(2s+2\nu+d)} \right)$$

The above estimator therefore converges almost surely with the optimal rate for intensities in $F^s_{2,2}(M)$. However, the main defect of the estimator defined in Theorem 4.2 is that it is suited for smooth functions and does not attain the optimal rates when for example $g = \log(f)$ has singularities. We therefore propose in the next section another estimator derived by applying an appropriate nonlinear thresholding procedure.

5 Nonlinear estimation

It is well known that linear estimators do not achieve the optimal rates of convergence when the functions to be recovered belong to Besov spaces $B^s_{p,p}$ with index $1 \leq p < 2$ (the case of functions which are not very smooth). In order to attain such a rate we need therefore some kind of nonlinear procedure and this is our aim in this section.

Our estimation procedure simply consists of applying a soft thresholding algorithm on the data to which we apply the Galerkin information projection inversion which was described previously, exploiting the fact that the model with Poisson intensity is not too different than the usual Gaussian white-noise model.

Let us first recall that the coefficients defining *the Galerkin information projection* estimate of f at scale j, as derived in the previous section, are given by:

$$\hat{\alpha}_{t,\lambda} = \frac{1}{t}\int u^j_\lambda dG = (U^j_\lambda)^T \frac{1}{t}\left(\int \psi_\mu dG\right)_{|\mu|<j},$$

where $U^j_\lambda = K^{-1}_j \Psi_\lambda$.
For some $j \geq 0$ (to be fixed further), we define the following thresholded coefficients:

$$\delta_t(\hat{\alpha}_{t,\lambda}) = (U^j_\lambda)^T \left(T_{\epsilon(t)}(\frac{1}{t}\int \psi_\mu dG)\right)_{|\mu|<j} \quad \text{for all } |\lambda| < j,$$

where $T_{\epsilon(t)}(x) = \operatorname{sign}(x)(x - \epsilon(t))_+$ for $x \in \mathbb{R}$ denotes the usual soft thresholding operator.

In order to build an optimal solution for the Poisson inverse problem we will use a level dependent wavelet thresholding procedure by setting $\epsilon(t) = 2^{\nu|\lambda|}t^{-1/2}\sqrt{|\log t|}$. The role of $2^{\nu|\lambda|}$ is to take into account the amplification of the noise by the inversion process. The following theorem shows that the resulting estimator behaves in an optimal way provided that the cutoff resolution level $j(t)$ is chosen such as $2^{-j(t)} \leq t^{-1/(2\nu)}$ where ν is the degree of ill-posedness of the estimator.

Theorem 5.1 *Assume that ψ is compactly supported and that $f \in F^s_{p,p}(M)$ with $s > 0$ and $1/p = 1/2 + s(2\nu + d)$. Moreover, suppose that ψ is in $H^{s+d/2-\nu}$ (with $s > \nu - d/2$) and has r vanishing moments with $r > s + d/2$. Also assume that K is an isomorphism between L^2 and H^ν and that it has the smoothing property of order ν with respect to the space $B^s_{p,p}$. Then, the above described Galerkin information projection estimator, say $f_{j(t),\hat{\theta}_t}$, satisfies the minimax rate*

$$\mathbb{E}\left(\Delta(f; f_{j(t),\hat{\theta}_t})\right) \leq O\left((\frac{1}{t}\sqrt{|\log t|})^{2s/(2s+2\nu+d)} \right)$$

provided that $j(t)$ is such that $2^{-j(t)} \leq t^{-1/(2\nu)}$.

Note that the lower bound on $j(t)$ does not depend on the unknown smoothness of f and therefore Theorem 5.1 allows to build an adaptive solution to our Poisson inverse problem. The assumption that K^{-1} maps H^ν into L^2 in the above theorem is also implicit in the vaguelette-wavelet method of Donoho [8] for white noise inverse problems.

6 Implementation and some numerical results

The purpose of this section is to describe the implementation of our approach and to briefly explore the performance of our method from a numerical point of view. As in Cohen et al. [5] we will focus on a simple example of a logarithmic potential kernel in dimension one. We will consider its action on two typical test intensity functions, which, together with their folded versions by the action of K are displayed in Figure 6.1.

The logarithmic potential operator K that we will consider is defined by

$$Kf(x) = \int_0^1 k(x,y)f(y)\,dy,$$

where

$$k(x,y) = -\log(\frac{1}{2}|\sin\frac{y-x}{2}|), \quad x,y \in [0,1].$$

Such a kernel is singular on the diagonal $x = y$ but integrable. The corresponding operator is known to be an elliptic operator of order -1, which maps $H^{-1/2}$ into $H^{1/2}$ and therefore satisfies the assumptions made in this paper with $\nu = 1$. The first test function we will consider is

$$f(x) = \max\{1 - |30(x - 0.5)|, 0.1\}, \quad x \in [0,1],$$

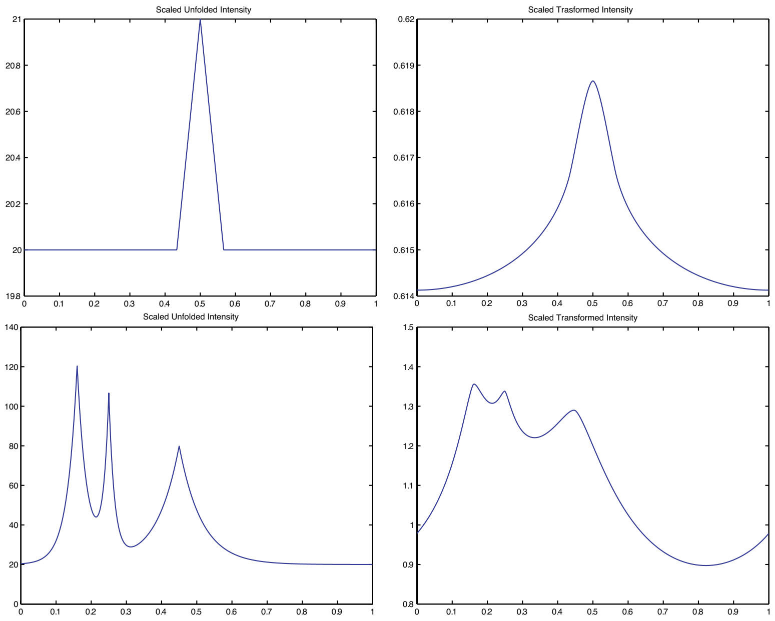

Figure 6.1: Artificial intensity functions (left) and their folded versions (right) by the action of the logarithmic potential kernel. Top-left : the intense peak ; Bottom-left: the burst-like intensity.

which presents an intense peak and is badly approximated by the singular functions of K but has a very sparse representation in a wavelet basis. We will also consider a fast rise-exponential model often used by astrophysicists to model burst phenomena which is of the form $f(x) = f_0 + \sum_{i=1}^{3} f_i(x)$ where f_0 models the relatively constant background level of gamma-ray photon arrivals and f_i models the ith peak in the burst of a form

$$f_i(x) = \begin{cases} A_i \exp\left(-|x - m_i| / \sigma_{r,i}^{\nu_i}\right), & \text{if } x \leq m_i \\ A_i \exp\left(-|x - m_i| / \sigma_{r,i}^{\nu_i}\right), & \text{if } x \leq m_i. \end{cases}$$

In the expression above m_i denotes the location of the ith peak, and the factors A_i, $\sigma_{r,i}$, $\sigma_{d,i}$ and ν_i control respectively the amplitude, the rise, the decay and the peakedness.

Most often data can only be observed in binned form because of the discrete nature of the measurement apparatus or because binning may be enforced by data handling and computing efficiency. We therefore have chosen to discretize K for a maximal resolution level $J = 11$ by computing the stiffness matrix K_J with entries

$$(K_J)_{\ell,k=0,\ldots,2^J-1} = \left(\langle K_J \phi_{J,\ell}, \phi_{J,k} \rangle\right)_{\ell,k=0,\ldots,2^J-1}$$

where the $\phi_{J,k} = 2^{J/2} I_{[k2^{-J},(k+1)2^{-J})}$ are the Haar scaling functions. Each integral

$$\langle K_J \phi_{J,\ell}, \phi_{J,k} \rangle = \int_0^1 \int_0^1 k(x,y)\phi_{J,\ell}(x)\phi_{J,k}(y)dxdy$$

was computed by Riemann approximation at scale 2^{-16}. Note that the kernel k is such that $k(x,y) = h(x-y)$ where $h(\cdot)$ is a 1-periodic function. The discretization K_J of K is therefore a Toeplitz cyclic matrix and the Fast Fourier Transform makes the computation of the action of K on functions approximated in the Haar basis numerically fast and easy. But this is not the only reason we have used the Haar based discretization for both f and K_J. More specifically, the Haar scaling basis at resolution J, induces a partition of the interval $[0,1)$ into 2^J disjoints and measurable bins $B_k = [k2^{-J},(k+1)2^{-J})$. Integrating the function $tK_J f$ with respect to the Poisson counting measure G simply leads to observed data consisting of counts observed in the bins B_k. By the Poisson nature of G, these are independent Poisson counts with expected values within each bin $t\int_{B_k} h$, $k = 0,\ldots,2^J-1$. Moreover, taking a high resolution J permits the approximation of $\int_{B_k} h$ by $2^{-J}h(k/2^J)$ and this is what we have done for creating the simulated data in the examples.

For the examples treated in this paper, the estimation was implemented using Symmlets with 6 vanishing moments. Since the set $\{x_1,\ldots,x_n\}$, of the $n = 2^J$ points at which the data is sampled is dyadic any scalar product involving a wavelet at a lower resolution is computed via the discrete wavelet transform. The Information projection estimator was obtained by solving the system of equations given in Theorem 4.1 at a maximal resolution $J_{\max}$.

To find the estimate $\hat{\boldsymbol{\theta}}_t$ we have used, inspired by a similar approach in Koo and Kim [14], a modified version of the Newton-Raphson method. Let $S(\boldsymbol{\theta})$ denote the $J_{\max}$-dimensional vector of elements

$$S(\boldsymbol{\theta})_\lambda = \left(\delta_t(\hat{\alpha}_{t,\lambda}) - \langle f_{t,\boldsymbol{\theta}}, \psi_\lambda \rangle\right)$$

and $H(\boldsymbol{\theta})$, the $J_{\max} \times J_{\max}$ Hessian matrix whose (λ,λ') entry is given by

$$H(\boldsymbol{\theta})_{\lambda,\lambda'} = \int f_{t,\boldsymbol{\theta}}(x)\psi_\lambda(x)\psi_{\lambda'}(x)dx.$$

The method to compute $\hat{\boldsymbol{\theta}}_t$ is to start with an initial guess $\boldsymbol{\theta}^0$ and iteratively determine $\boldsymbol{\theta}^{m+1}$ according to

$$\boldsymbol{\theta}^{m+1} = \boldsymbol{\theta}^m + H^{-1}(\boldsymbol{\theta}^m)S(\boldsymbol{\theta}^m)$$

with a standard criterion for stopping the iterations.

One difficulty in implementing the above algorithm is that Symmlets have no close functional expression and the integration involved in the computation of the Hessian H can be very time consuming if ones has to compute a table of the appropriate values of the function ψ_λ. We have used instead an efficient filter-bank algorithm for computing such integrals similar to the one used by Vannucci and Corradi [26] or Kovac and Silverman [15] to compute the diagonal elements of the covariance structure of wavelet coefficients, which amounts in computing the fast 2-dimensional wavelet transform of the diagonal matrix whose diagonal is the vector $f_{t,\theta}(x_i)$, $i = 1, \ldots, J_{\max}$ and to retain only the diagonal blocks of the transform.

For our first example, we consider the peaky function and chose a maximal resolution $J_{\max} = 10$ and an "observation time" $t = 10^8$ (corresponding to the noise level used by Cohen et al. [5] for a similar white noise model). A typical sample from the simulated model is shown in Figure 6.2.

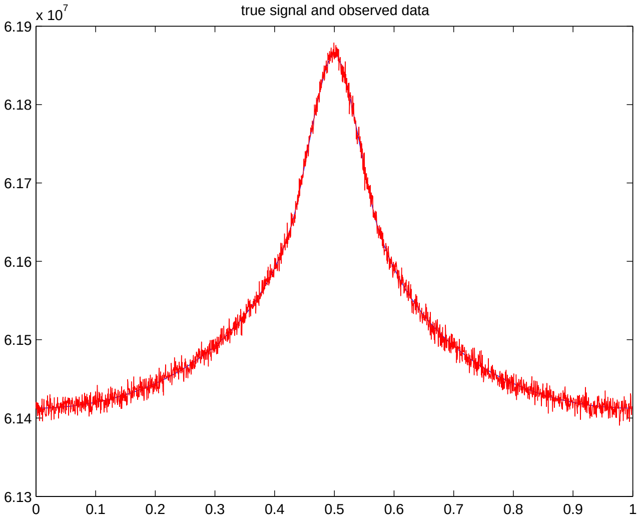

Figure 6.2: Simulated Poisson data obtained by folding the peaky function with the logarithmic potential.

Let us recall that our non-linear information projection estimator depends on the cutoff level $j(t)$ given by $2^{-j(t)} \leq (\frac{1}{t})^{1/(2\nu)}$ and the level dependent thresholds $\epsilon(t) = 2^{\nu|\lambda|}t^{-1/2}\sqrt{|\log t|}$. We therefore have used these expression with $\nu = 1$ to estimate the unfolded intensity function. Figure 6.3 displays the non-linear Galerkin estimator for the folded Poisson data displayed in Figure 6.2. We observe, and this is true also for the second example, that the peak is very well estimated. However, some oscillations are observed on the right side of the central peak. A possible remedy to this defect could be to use a translation invariant procedure but such an approach is beyond the scope of this paper.

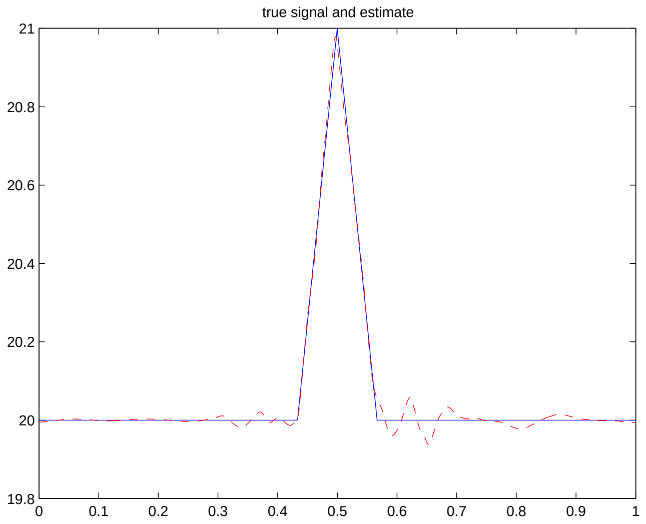

Figure 6.3: The unknow intensity (solid) and the non-inear Garlekin estimate (dashed).

Our second example concerns the burst-like intensity function. The data displayed in Figure 6.4 is simulated as above using for an unfolded intensity a burst signal with a constant intensity of 20 assigned to the background and three peaks.

Since we are using the same logarithmic potential kernel to fold the intensity, we have also used here a value of ν equal to 1. Maximal resolution, smooth cutoff level and thresholds were chosen exactly as in the previous example and the estimation procedure provides us the fit in Figure 6.5 for the unfolded burst-like intensity, confirming the good behavior of our procedure even for complicated intensity structures.

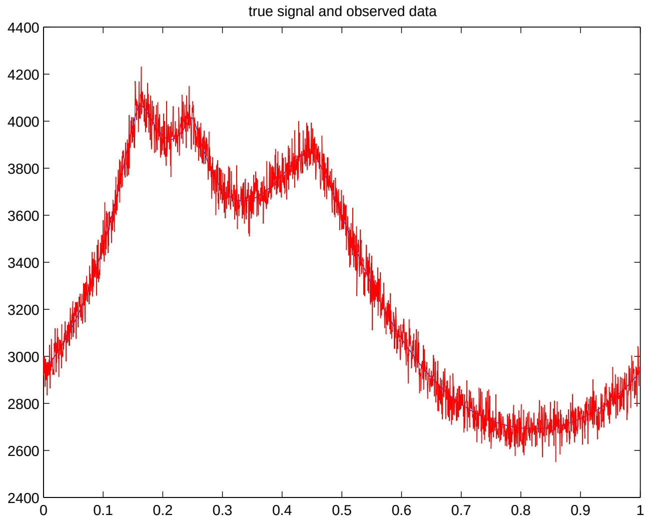

Figure 6.4: Simulated Poisson data obtained by folding the burst-like intensity with the logarithmic potential.

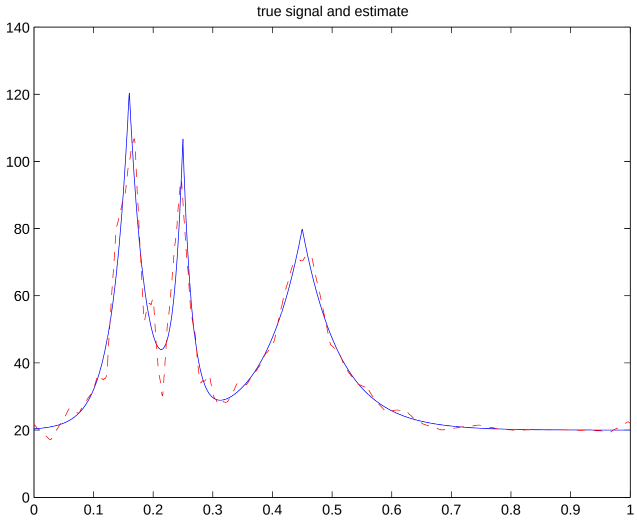

Figure 6.5: The unknow intensity (solid) and the non-inear Garlekin estimate (dashed).

7 Conclusions

The methodology of this paper was motivated by wavelet-vaguelette (WVD) methods that have been developed in the literature for solving inverse problems with Gaussian white-noise perturbations. Such methods have been proved very efficient for recovering signals folded by homogeneous operators, but are not designed to handle more general operators or Poisson inverse problems. The method developed here combines the numerical simplicity of Galerkin projection methods with the optimality of adaptive wavelet thresholding methods, which are easier to handle than WVD approaches. The underlying intuition is again that if the operator involved in the folding process lends to a fairly sparse representation in wavelets, and if the intensity to be unfolded is mostly smooth with some singularities, then the inversion of the truncated operator will not be too onerous, and the approximate representation of the intensity will do a good job of capturing the singularities. Moreover, the adopted thresholding makes the method adaptive.

Our attention was restricted to inverse problems, in which E_0 and E_1 are identical Borel subsets of $\mathbb{R}^d$ and the operator K is self-adjoint positive definite. In the case where $E_0 \neq E_1$ and K is not self-adjoint but just injective one may choose, as it is done in Cohen *et al.* [5], $E_1 = K(E_0)$ and solve instead the inverse problem $K^*h = K^*Kf$ where K^* denotes the adjoint of K. In such a way we are led back to the wavelet Galerkin information projection method developed in this paper.

Another approach, which is worth investigating in the future, would be to derive a Landweber-type iterative algorithm that involves a denoising procedure at each iteration step and provides a sequence of approximations converging in norm to the maximum penalized likelihood minimizer as it is done by Figueiredo and Nowak [10], who derive such an iterative algorithm for inverting a convolution operator acting on objects that are sparse in the wavelet domain.

ACKNOWLEDGEMENTS. We thank Marc Hoffmann for interesting and stimulating discussions and also for providing us the Matlab numerical procedures used in the paper by Cohen *et al.* [5] . This research was supported by funds from the IAP research network nr P5/24 of the Belgian Government (Federal Office for Scientific, Technical and Cultural Affairs) and the EC-HPRN-CT-2002-00286 Breaking Complexity network. Both authors would like to thank the Institute of Statistics of the Université Catholique de Louvain for the hospitality.

Appendix: Proofs of the main results.

We first derive the proof of Lemma 3.1.
PROOF OF LEMMA 3.1. Note that:

$$\Delta(f; f_{j,\theta}) = D(f, f_{j,\theta}) + \int (-f + f_{j,\theta(\alpha)}) d\mu + \int (-f_{j,\theta(\alpha)} + f_{j,\theta}) d\mu,$$

where

$$D(f, f_{j,\theta}) = \int f \log(\frac{f}{f_{j,\theta}}) d\mu = D(f, f_{j,\theta(\alpha)}) + D(f_{j,\theta(\alpha)}, f_{j,\theta}),$$

by Lemma 4 of Barron and Sheu [2], which completes the proof. □

To prove Lemma 3.2 we need some preliminary lemmas on bounds within exponential families for the Kullback-Leibler distance; their proofs can be easily derived from the proofs of Lemma 1 and Lemma 4 of Barron and Sheu [2] and thus they are omitted.

Lemma 7.1 *Let f and h be two intensity measures in $L^2(E, \mu)$. Assume that $\log(\frac{f}{h})$ is bounded, then*

$$\begin{aligned} \Delta(f; h) &\geq \frac{1}{2} e^{-\| \log(\frac{f}{h}) \|_\infty} \int f (\log(\frac{f}{h}))^2 d\mu \\ \Delta(f; h) &\leq \frac{1}{2} e^{\| \log(\frac{f}{h}) \|_\infty} \int f (\log(\frac{f}{h}))^2 d\mu \end{aligned}$$

Lemma 7.2 *For $j \geq 0$, let $\theta_0, \theta \in \mathbb{R}^{2dj}$ and $b = \exp(\|log(f_{j,\theta_0})\|_\infty)$, then*

$$\begin{aligned} \| \log(\frac{f_{j,\theta_0}}{f_{j,\theta}}) \|_\infty &\leq A_j \|\theta_0 - \theta\|_2 \\ \Delta(f_{j,\theta_0}; f_{j,\theta}) &\geq \frac{1}{2b} e^{-A_j \|\theta_0 - \theta\|_\infty} \|\theta_0 - \theta\|_2^2 \\ \Delta(f_{j,\theta_0}; f_{j,\theta}) &\leq \frac{b}{2} e^{A_j \|\theta_0 - \theta\|_\infty} \|\theta_0 - \theta\|_2^2 \end{aligned}$$

We can proceed to the proof of Lemma 3.2.
PROOF OF LEMMA 3.2. This proof is inspired by the proof of Lemma 5 of Barron and Sheu [2]. Let

$$F(\theta) = \theta \cdot \alpha - H(\theta),$$

where $H(\theta) = \int f_{j,\theta}(x)d\mu(x)$. If $\alpha = \alpha_0$, then the result is trivial. Now, if $\alpha \neq \alpha_0$, note that for any $\theta \in \mathbb{R}^{2dj}$

$$\Delta(f_{j,\theta_0}; f_{j,\theta}) = \alpha_0 \cdot (\theta_0 - \theta) + H(\theta) - H(\theta_0).$$

Hence,

$$F(\theta_0) - F(\theta) = \Delta(f_{j,\theta_0}; f_{j,\theta}) - (\alpha_0 - \alpha) \cdot (\theta_0 - \theta).$$

So, by Lemma 7.2 and the Cauchy-Schwartz inequality, we have that

$$F(\theta_0) - F(\theta) \geq \frac{1}{2b} e^{-A_j \|\theta_0 - \theta\|_\infty} \|\theta_0 - \theta\|_2^2 - \|\alpha_0 - \alpha\|_2 \|\theta_0 - \theta\|_2.$$

This inequality is strict if $\theta \neq \theta_0$. For all θ such that $\|\theta_0 - \theta\|_2 = 2eb\|\alpha_0 - \alpha\|_2$,

$$F(\theta_0) - F(\theta) > 2eb\|\alpha_0 - \alpha\|_2^2 \left(e^{1 - 2A_j eb\|\alpha_0 - \alpha\|_2} - 1\right).$$

The right side is positive whenever $2A_j eb\|\alpha_0 - \alpha\|_2 \leq 1$. Hence, $F(\theta_0) > F(\theta)$ for all θ such that $\|\theta_0 - \theta\|_2 = 2eb\|\alpha_0 - \alpha\|_2$. Consequently, F has an extreme point θ^* such that $\|\theta_0 - \theta^*\|_2 < 2eb\|\alpha_0 - \alpha\|_2$. The gradient of F at θ^* must satisfy:

$$< f_{j,\theta^*}, \psi_\lambda >= \alpha_\lambda \text{ for all } |\lambda| < j$$

and so $\theta(\alpha) = \theta^*$. Hence, inequality (3.1) immediately follows. Inequality (3.2) follows from Lemma 7.2. Since $F(\theta(\alpha)) \geq F(\theta_0)$, we have that:

$$\begin{array}{rcl} \Delta(f_{j,\theta(\alpha_0)}; f_{j,\theta(\alpha)}) & \leq & (\alpha_0 - \alpha) \cdot (\theta_0 - \theta(\alpha)) \\ & \leq & \|\alpha_0 - \alpha\|_2 \|\theta_0 - \theta\|_2 \\ & \leq & 2eb\|\alpha_0 - \alpha\|_2^2, \end{array}$$

which completes the proof □.

To prove the main results of this paper we shall need a series of technical lemmas, stated and proved below. Throughout this section, C will denote a constant whose value may change from line to line.

Lemma 7.3 *If ψ is compactly supported, then*

$$\| \sum_{|\lambda|=j} \beta_\lambda \psi_\lambda \|_\infty \leq C 2^{jd/2} \|\beta_j\|_2$$

PROOF: This lemma immediately follows from the proof of Lemma 1 of Koo and Kim [13]. □.

The following lemma is similar to Lemma 2 of Koo and Kim [13]:

Lemma 7.4 *Assume that ψ is compactly supported and that $f \in F^s_{p,q}(M)$. If $s > d/p \geq d/2$, then there exists a constant M_1 such that*

$$0 < \frac{1}{M_1} \leq f \leq M_1 < \infty.$$

PROOF: Let $g = \log(f) = \sum_{j=-1}^{\infty} \sum_{|\lambda|=j} \beta_\lambda \psi_\lambda$. Since $\|g\|_{B^s_{p,q}} \leq M$, we have

$$\|\beta_j\|_p = \left(\sum_{|\lambda|=j} |\beta_\lambda|^p \right)^{1/p} \leq M 2^{-js'},$$

where $s' = s + d(1/2 - 1/p)$. If $s > d/p \geq d/2$, then:

$$\|\beta_j\|_2 \leq \|\beta_j\|_p \leq C 2^{-js'}. \tag{7.1}$$

Then, by Lemma 7.3

$$\|g\|_\infty \leq \sum_{j=-1}^{\infty} \| \sum_{|\lambda|=j} \beta_\lambda \psi_\lambda \|_\infty \leq \sum_{j=0}^{\infty} C 2^{jd/2} \|\beta_j\|_2 \leq C \sum_{j=0}^{\infty} 2^{j(d/2 - s')} \leq C \sum_{j=0}^{\infty} 2^{-j(s-d/p)}.$$

Since $s > d/p$, $\sum_{j=0}^{\infty} 2^{-j(s-d/p)} < \infty$ and therefore there exists some constant $M_1 > 1$ such that $\|g\|_\infty = \| \log f \|_\infty \leq \log M_1$. □

Now, we give bounds for A_j, D_j and γ_j.

Lemma 7.5 *Assume that ψ is compactly supported, then:*

$$A_j = C 2^{jd/2}$$

Moreover suppose that $f \in F^s_{p,q}(M)$. If $s > d/p \geq d/2$, then

$$\begin{aligned} D_j &\leq C 2^{-j(s + d(1/2 - 1/p))} \\ \gamma_j &\leq C 2^{-j(s - d/p)} \end{aligned}$$

Proof: The result for A_j immediately follows from Lemma 7.3. Note that from equation (7.1),

$$D_j^2 = \sum_{j' \geq j} \|\beta_{j'}\|_2^2 \leq C \sum_{j' \geq j} 2^{-2j'(s+d(1/2-1/p))} = O(2^{-2j(s+d(1/2-1/p))}).$$

By definition,

$$\gamma_j = \|g - P_j g\|_\infty \leq A_j D_j \leq C 2^{-j(s-d/p)},$$

which completes the proof. □

We may now proceed to the proofs of our main results.

PROOF OF THEOREM 4.1.

Approximation error term: Let $g = \log(f) = \sum_{j=-1}^{\infty} \sum_{|\lambda|=j} \beta_\lambda \psi_\lambda$, and for all $|\lambda| < j$, let $\alpha_{j,\lambda} = \langle \exp(P_j g), \psi_\lambda \rangle$ and $\alpha_\lambda = \langle f, \psi_\lambda \rangle$. Observe that the coefficients $(\alpha_{j,\lambda} - \alpha_\lambda)$, $|\lambda| < j$, are the coefficients of the orthogonal projection of $f - \exp(P_j g)$ onto V_j. Hence by Bessel's inequality

$$\|\alpha_j - \alpha\|_2^2 \leq \|f - \exp(P_j g)\|_{L^2}^2.$$

Given our assumptions on ψ and f, Lemma 7.4 implies that

$$\|\alpha_j - \alpha\|_2^2 \leq M_1 \int \frac{(f - \exp(P_j g))^2}{f} d\mu,$$

and so by Lemma 2 of Barron and Sheu [2],

$$\|\alpha_j - \alpha\|_2^2 \leq M_1 e^{2\|g - P_j g\|_\infty} \int f (g - P_j g)^2 d\mu \leq M_1^2 e^{2\gamma_j} D_j^2.$$

Now, apply Lemma 3.2 with $\theta_{0,\lambda} = \beta_\lambda$, $\alpha_\lambda = \langle f, \psi_\lambda \rangle$ for all $|\lambda| < j$ and $b = e^{\|\log(\exp(P_j g))\|_\infty}$. Since $\|\log(f/\exp(P_j g))\|_\infty = \gamma_j$, we have that $\|\log(\exp(P_j g))\|_\infty \leq \log M_1 + \gamma_j$ and therefore $b \leq M_1 e^{\gamma_j}$. From Lemma 3.2, we have that if $M_1 e^{\gamma_j} D_j \leq \frac{1}{2ebA_j}$, i.e if $\epsilon_j \leq 1$, then $\theta_j^* = \theta(\alpha)$ exists. By Lemma 3.1 (Pythagorean-like relationship), we obtain that

$$\Delta(f; f_{j,\theta_j^*}) \leq \Delta(f; \exp(P_j g)).$$

Thence, by Lemma 7.1,

$$\begin{aligned} \Delta(f; f_{j,\theta_j^*}) &\leq \frac{1}{2} e^{\|g - P_j g\|_\infty} \int f (g - P_j g)^2 d\mu. \\ &\leq \frac{1}{2} M_1 e^{\gamma_j} D_j^2, \end{aligned}$$

which completes the proof of the first assertion of the theorem.

Estimation error term: Using the above notation and proof, and since, by assumption $\epsilon_j \leq 1$, let $\theta_j^* \in \mathbb{R}^{2jd}$ be the parameter vector achieving the minimum of the relative entropy for intensities in the exponential family. For all $|\lambda| < j$, define $\alpha_{0,\lambda} = \langle f, \psi_\lambda \rangle = \langle f_{j,\theta_j^*}, \psi_\lambda \rangle$ and let $\hat{\alpha}_{t,\lambda} = \frac{1}{t} \int u_\lambda^j dG$. It is easy to see that $\mathbb{E}(\hat{\alpha}_{t,\lambda}) = \langle u_\lambda^j, Kf \rangle$. We now have

$$\|\hat{\alpha}_t - \alpha_0\|_2^2 = \sum_{|\lambda|<j} (\hat{\alpha}_{t,\lambda} - \alpha_{0,\lambda})^2 \leq 2 \sum_{|\lambda|<j} \left(\hat{\alpha}_{t,\lambda} - \langle u_\lambda^j, Kf \rangle \right)^2 + 2 \left(\langle u_\lambda^j, Kf \rangle - \alpha_{0,\lambda} \right)^2. \tag{7.2}$$

Concerning the second member of the right hand side of inequality (7.2), and since $\alpha_{0,\lambda} = \langle f, \psi_\lambda \rangle$ note that

$$\begin{aligned} (\langle u_\lambda^j, Kf \rangle - \langle f, \psi_\lambda \rangle)^2 &= (\langle f, Ku_\lambda^j - \psi_\lambda \rangle)^2 \leq \|f\|_{L^2}^2 \|Ku_\lambda^j - \psi_\lambda\|_{L^2}^2 \\ &= \|f\|_{L^2}^2 \left(\|Ku_\lambda^j\|_{L^2}^2 + 1 - 2\langle Ku_\lambda^j, \psi_\lambda \rangle \right) \end{aligned}$$

By definition we have that $\langle Ku_\lambda^j, \psi_\lambda \rangle = \langle \psi_\lambda, \psi_\lambda \rangle = 1$. It follows that

$$\begin{aligned} \|Ku_\lambda^j\|_{L^2}^2 &= \sum_{|\mu|<j} \langle Ku_\lambda^j, \psi_\mu \rangle^2 + \sum_{|\mu|\geq j} \langle Ku_\lambda^j, \psi_\mu \rangle^2 \\ &= 1 + \sum_{|\mu|\geq j} \langle Ku_\lambda^j, \psi_\mu \rangle^2. \end{aligned}$$

Given the assumptions on the wavelet ψ and the operator K, u_λ^j belongs to $H^{s+d/2-\nu}$ and hence Ku_λ^j belongs to $H^{s+d/2}$. Since $r > s + d/2$, it follows from approximation theory that

$$\sum_{|\mu|\geq j} \langle Ku_\lambda^j, \psi_\mu \rangle^2 \leq 2^{-2j(s+d/2)},$$

and therefore

$$(\langle u_\lambda^j, Kf \rangle - \langle f, \psi_\lambda \rangle)^2 \leq \|f\|_{L^2}^2 2^{-2js-jd}.$$

Thence we obtain for the second member of the right hand side of inequality (7.2)

$$\sum_{|\lambda|<j} (\langle u_\lambda^j, Kf \rangle - \langle f, \psi_\lambda \rangle)^2 \leq \|f\|_{L^2}^2 2^{-2js}. \tag{7.3}$$

To control the first term of the right hand side of inequality (7.2), let $S_j = \operatorname{span}\{u_\lambda^j; |\lambda| < j\}$ and set

$$\chi^2(S_j) = \sum_{|\lambda|<j} \left(\hat{\alpha}_{t,\lambda} - \langle u_\lambda^j, Kf \rangle \right)^2 = \sum_{|\lambda|<j} \left(\frac{1}{t} \int u_\lambda^j dG - \langle u_\lambda^j, Kf \rangle \right)^2.$$

Noticing that

$$\chi(S_j) = \sup_{\{a\in\mathbb{R}^{jd};\|(a_\lambda)_{|\lambda|<j}\|_2<1\}} \int \frac{\sum_{|\lambda|<j} a_\lambda u_\lambda^j}{t} \left(dG - tKf d\mu\right),$$

we can use Corollary 1 of Reynaud-Bouret [22] about concentration inequalities for Poisson processes to get, for any $w > 0$ and $\epsilon > 0$,

$$\mathbb{P}\left\{\chi(S_j) \geq (1+\epsilon)\mathbb{E}(\chi(S_j)) + \sqrt{12v_0 w} + \kappa(\epsilon) b_0 w\right\} \leq \exp(-w), \tag{7.4}$$

where $\kappa(\epsilon) = 1.25 + 32/\epsilon$,

$$v_0 = \sup_{\{a\in\mathbb{R}^{jd};\|a\|_2<1\}} \int \frac{\sum_{|\lambda|<j}(a_\lambda u_\lambda^j)^2}{t^2} tKf d\mu,$$

and

$$b_0 = \sup_{\{a\in\mathbb{R}^{jd};\|a\|_2<1\}} \frac{\|\sum_{|\lambda|<j} a_\lambda u_\lambda^j\|_\infty}{t}.$$

In what follows we provide a precise control of the constants v_0 and b_0 involved in inequality (7.4). It is easy to show that

$$v_0 \leq \frac{\|Kf\|_\infty}{t} \sup_{\{a\in\mathbb{R}^{jd};\|a\|_2<1\}} \int \left(\sum_{|\lambda|<j} a_\lambda u_\lambda^j\right)^2 d\mu.$$

For any vector $a \in \mathbb{R}^{jd}$ we have

$$\begin{aligned} \int \left(\sum_{|\lambda|<j} a_\lambda u_\lambda^j\right)^2 d\mu &= \sum_{|\lambda|<j,|\lambda'|<j} a_\lambda a_{\lambda'} \int u_\lambda^j u_{\lambda'}^j d\mu \qquad (7.5) \\ &\leq \sum_{|\lambda|<j} a_\lambda a_{\lambda'} \|u_\lambda^j\|_{L^2} \|u_{\lambda'}^j\|_{L^2}. \qquad (7.6) \end{aligned}$$

As argued in Cohen *et al.* [5], the ellipticity property (2.1) yields

$$\begin{aligned} \|u_\lambda^j\|_{H^{-\nu/2}}^2 &\leq\ < Ku_\lambda^j, u_\lambda^j > = < \psi_\lambda, u_\lambda^j > \\ &\leq \|\psi_\lambda\|_{L^2} \|u_\lambda^j\|_{L^2} = \|u_\lambda^j\|_{L^2}. \end{aligned}$$

and the inverse inequality (see Cohen *et al* [5]) states that $\|u_\lambda^j\|_{L^2} \leq 2^{\nu j/2} \|u_\lambda^j\|_{H^{-\nu/2}}$ which implies that (dividing by $\|u_\lambda^j\|_{L^2}$):

$$\|u_\lambda^j\|_{L^2} \leq 2^{\nu j}. \tag{7.7}$$

Using the above bound in the last inequality (7.6) we finally obtain:

$$\int \left(\sum_{|\lambda|<j} a_\lambda u_\lambda^j \right)^2 d\mu \leq 2^{2\nu j} \left(\sum_{|\lambda|<j} a_\lambda \right)^2 \leq 2^{j(2\nu+d)} \|a\|_2^2.$$

It follows that

$$v_0 \leq \|Kf\|_\infty \frac{2^{j(2\nu+d)}}{t}.$$

Now, by definition of the constants A_j and since $\sum_{|\lambda|<j} a_\lambda u_\lambda^j \in V_j$, we have

$$\| \sum_{|\lambda|<j} a_\lambda u_\lambda^j \|_\infty \leq A_j \| \sum_{|\lambda|<j} a_\lambda u_\lambda^j \|_{L^2},$$

and it follows that, for $\|a\|_2 \leq 1$,

$$\| \sum_{|\lambda|<j} a_\lambda u_\lambda^j \|_{L^2} \leq 2^{j(\nu+d)},$$

which combined with the statements of Lemma 7.5 gives

$$b_0 = O\left(\frac{2^{j(\nu+\frac{3}{2}d)}}{t} \right).$$

Now, recall that $\text{Var}(\frac{1}{t} \int u_\lambda^j dG) = \frac{1}{t} \int (u_\lambda^j)^2 Kf d\mu$. Hence, using again the bound in inequality (7.7) we have

$$\sum_{|\lambda|<j} \text{Var}(\frac{1}{t} \int u_\lambda^j dG) \leq \frac{1}{t} \|Kf\|_\infty 2^{j(d+2\nu)}, \tag{7.8}$$

which implies that

$$\mathbb{E}(\chi^2(S_j)) \leq \frac{1}{t} \|Kf\|_\infty 2^{j(d+2\nu)}.$$

Combining equations (7.8) and (7.3) yields finally

$$\mathbb{E}\left(\sum_{|\lambda|<j} (\frac{1}{t} \int u_\lambda^j dG - \langle f, \psi_\lambda \rangle)^2 \right) \quad \leq \quad 2\left(\frac{1}{t} \|Kf\|_\infty 2^{j(d+2\nu)} + \|f\|_{L^2}^2 2^{-2js} \right)$$

From Cauchy-Schwartz inequality and expression (7.4) with $\epsilon = w$ it follows that there exists a constant $C > 0$, such that, for any $w > 0$:

$$\mathbb{P}\left\{ \chi^2(S_j) \geq C(1+w)^2 (2^{j(\nu+d/2)}/\sqrt{t} + 2^{j(\nu+\frac{3}{2}d)}/t)^2 \right\} \leq \exp(-w).$$

Combining the above inequalities, and using the fact that f is bounded in L^2 we get finally:

$$\mathbb{P}\left\{\|\hat{\alpha}_t - \alpha_0\|_2^2 \geq C(1+w)^2\rho_{j,t}\right\} \leq \exp(-w).$$

It remains to set $\eta = (1+w)$ and recall that $\rho_{j,t} = (2^{j(\nu+d/2)}/\sqrt{t} + 2^{j(\nu+\frac{3}{2}d)}/t)^2 + 2^{-2js}$ to get

$$\mathbb{P}\left\{\|\hat{\alpha}_t - \alpha_0\|_2^2 \geq C_1\eta^2\rho_{j,t}\right\} \leq \exp(-\eta).$$

Now, applying Lemma 3.2 with $\theta_0 = \theta_j^*$, $\alpha = \hat{\alpha}_t$ and $b = e^{\|\log(f_{j,\theta_j^*})\|_\infty}$, we have that $\|\log(f_{j,\theta_j^*}/\exp(P_j g))\|_\infty \leq \epsilon_j$, and so $b \leq M_1 e^{\epsilon_j+\gamma_j}$. Hence, if $\eta^2\rho_j^t \leq \frac{1}{4e^2b^2A_j^2}$, that is if $\delta_j^t \leq 1/\eta^2$, then except in the set above, Lemma 3.2 implies that $\hat{\theta}_t = \theta(\hat{\alpha}_t) \in \mathbb{R}^{2^{jd}}$ exists and satisfies

$$\begin{aligned}\Delta(f_{j,\theta_j^*}; f_{j,\hat{\theta}_t}) &\leq 2eb\|\alpha - \alpha_0\|_2^2 \\ &\leq 2M_1 e^{1+\epsilon_j+\gamma_j}\eta^2\rho_{j,t},\end{aligned}$$

which completes the proof. □

Proof of Theorem 4.2. From the bounds for A_j, Δ_j and γ_j given by Lemma 7.5 and since $s > d/2$, we obtain that $\gamma_{j(t)} \to 0$ as $t \to \infty$ and so $\epsilon_{j(t)} = 0(A_j\Delta_j) = 0(2^{-j(t)(s-d/2)})$. Hence, $\epsilon_{j(t)} \to 0$ as $t \to \infty$ which implies that $\delta_{j(t)}^t = 0(2^{-j(t)(2s-d)})$. Since $\epsilon_{j(t)} \to 0$ and $\delta_{j(t)}^t \to 0$ as $t \to \infty$, Theorem 4.1 implies that

$$\Delta(f; f_{j(t),\theta_{j(t)}^*}) \leq O\left(2^{-2j(t)s}\right),$$

while for the estimation error, we have that as $t \to \infty$, then with probability tending to 1, $f_{j(t),\hat{\theta}_t}$ exists and by Borel-Cantelli's Lemma satisfies

$$\Delta(f_{j(t),\theta_{j(t)}^*}; f_{j(t),\hat{\theta}_t}) \leq O\left(2^{-2j(t)s}\right).$$

The result now follows from the Pythagorean-like relationship (Lemma 3.1)

$$\Delta(f; f_{j(t),\hat{\theta}_t}) = \Delta(f; f_{j(t),\theta_{j(t)}^*}) + \Delta(f_{j(t),\theta_{j(t)}^*}; f_{j(t),\hat{\theta}_t}).$$

□

PROOF OF THEOREM 5.1. Note that

$$\begin{aligned}
\|\delta_t(\hat{\alpha}_t) - \alpha_0\|_2^2 &= \sum_{|\lambda|<j} (\delta_t(\hat{\alpha}_{t,\lambda}) - \langle f, \psi_\lambda\rangle)^2 \\
&\leq 2\left(\sum_{|\lambda|<j} (\delta_t(\hat{\alpha}_{t,\lambda}) - \langle u^j_\lambda, Kf\rangle)^2 + (\langle u^j_\lambda, Kf\rangle - \langle f, \psi_\lambda\rangle)^2\right)
\end{aligned}$$

From the proof of Theorem 4.1 (see eq. (7.3)), we have (with the assumed conditions on ψ and K)

$$\sum_{|\lambda|<j} (\langle u^j_\lambda, Kf\rangle - \langle f, \psi_\lambda\rangle)^2 \leq \|f\|_{L^2}^2 2^{-2js}.$$

Remark that the space $B^s_{p,p}$ is continuously embedded in H^ζ whenever $\zeta \leq s + d/2 - d/p = 2\nu s/(2\nu + d)$. Moreover, since $2\nu/(2\nu + d) < 1$ and f is uniformly bounded we therefore obtain the estimate

$$\sum_{|\lambda|<j} (\langle u^j_\lambda, Kf\rangle - \langle f, \psi_\lambda\rangle)^2 \leq O\left(2^{-4js\frac{\nu}{2\nu+d}}\right).$$

This gives the optimal order $\left(\frac{1}{t}\right)^{2s/(2s+2\nu+d)}$ provided $j(t)$ is large enough so that $2^{-j(t)} \leq (\frac{1}{t})^{\frac{\nu+d/2}{\nu(2s+2\nu+d)}}$. For $t > 1$, we have $(\frac{1}{t})^{1/(2\nu)} \leq (\frac{1}{t})^{\frac{\nu+d/2}{\nu(2s+2\nu+d)}}$, since $s \geq 0$, with equality if $s = 0$. Therefore, if $2^{-j(t)} \leq (\frac{1}{t})^{1/(2\nu)}$, we obtain

$$\sum_{|\lambda|<j} (\langle u^j_\lambda, Kf\rangle - \langle f, \psi_\lambda\rangle)^2 \leq O\left(\left(\frac{1}{t}\right)^{2s/(2s+2\nu+d)}\right). \tag{7.9}$$

Note also that

$$\begin{aligned}
\sum_{|\lambda|<j} (\delta_t(\hat{\alpha}_{t,\lambda}) - \langle u^j_\lambda, Kf\rangle)^2 &= \sum_{|\lambda|<j} \left[\Psi^T_\lambda K_j^{-1}\left(T_{\epsilon(t)}(\frac{1}{t}\int \psi_\mu dG) - \langle h, \psi_\mu\rangle\right)_{|\mu|<j}\right]^2 \\
&= \|K_j^{-1}\left(T_{\epsilon(t)}(\frac{1}{t}\int \psi_\mu dG) - \langle h, \psi_\mu\rangle\right)_{|\mu|<j}\|_2^2,
\end{aligned}$$

where $h = Kf$. Since K is an isomorphism between L^2 and H^ν, using the proof of Theorem 1 of Cohen *et al.* [5] we have, that for any $U = (u_\lambda)_{|\lambda|<j} \in \mathbb{R}^{2jd}$,

$$\|K_j^{-1}U\|_2^2 \leq C\|U\|_{H^\nu} = \sum_{|\lambda|<j} 2^{2\nu|\lambda|}|u_\lambda|^2. \tag{7.10}$$

Hence, it follows that

$$\sum_{|\lambda|<j} (\delta_t(\hat{\alpha}_{t,\lambda}) - \langle u^j_\lambda, Kf\rangle)^2 \leq \sum_{|\lambda|<j} 2^{2\nu|\lambda|} \left(T_{\epsilon(t)}(\frac{1}{t}\int \psi_\lambda dG) - \langle h, \psi_\lambda\rangle \right)^2$$

We remark that $T_{\epsilon(t)}(\frac{1}{t}\int \psi_\lambda dG) - \langle h, \psi_\lambda\rangle$ is exactly the error when estimating $\langle h, \psi_\lambda\rangle$ by the thresholding procedure on the "data" $\frac{1}{t}\int \psi_\lambda dG$. We are thus left with finding a threshold $\epsilon(t)$ and some appropriate bounds for the estimation of h based on the thresholded coefficients $T_{\epsilon(t)}(\frac{1}{t}\int \psi_\lambda dG)$, $|\lambda| < j$, which would yield a bound for $\mathbb{E}\|\delta_t(\hat{\alpha}_t) - \alpha_0\|_2^2$.

To simplify the notation set $\hat{\beta}_{\lambda,t} = \frac{1}{t}\int \psi_\lambda dG$ and let $\beta_{\lambda,0} = \int \psi_\lambda h d\mu$. Since G is a Poisson process with intensity th it is easy to see that $\mathbb{E}(\hat{\beta}_{\lambda,t}) = \beta_{\lambda,0}$ and that $\text{Var}(\hat{\beta}_{\lambda,t}) = \frac{1}{t}\int \psi_\lambda^2 h d\mu = \frac{1}{t}\sigma_\lambda^2$. In order to bound the intensity estimation risk by a corresponding white noise model risk, we will apply Lemma V of Cavalier and Koo [3] to construct an approximation $\hat{\eta}_{\lambda,t}$ having an *exact* Gaussian distribution with the same mean $\beta_{\lambda,0}$ and the same variance $\text{Var}(\hat{\beta}_{\lambda,t})$. To this end, let $g_\lambda = \frac{1}{\sigma_\lambda}\psi_\lambda$ and note that $\int g_\lambda^2 h d\mu = 1$ and $\|g_\lambda\|_\infty = \frac{1}{\sigma_\lambda}\|\psi_\lambda\|_\infty = \frac{2^{|\lambda|d/2}}{\sigma_\lambda} := H_\lambda$ say. We construct $\hat{\eta}_{\lambda,t} = \beta_{\lambda,0} + t^{-1/2}\sigma_\lambda Z_\lambda$ by the following recipe.

Firstly, if $\sigma_\lambda \geq C2^{|\lambda|d}\frac{\log^3 t}{t}$, then use Lemma V of Cavalier and Koo [?] to construct Z_λ and note that

$$V_{\lambda,t} = \mathbb{E}\left(\hat{\beta}_{\lambda,t} - \hat{\eta}_{\lambda,t}\right)^2 = \frac{\sigma_\lambda^2}{t}\mathbb{E}(t^{-1/2}S_{\lambda,t} - Z_\lambda)^2 \leq C2^{|\lambda|d}t^{-2},$$

where $S_{\lambda,t} = \int \psi_\lambda(dG - thd\mu)$.

Secondly, if $\sigma_\lambda < C2^{|\lambda|d}\frac{\log^3 t}{t}$, choose an independent $Z_\lambda \sim N(0,1)$ and simply use the inequality

$$V_{\lambda,t} \leq 2\text{Var}(\hat{\beta}_{\lambda,t}) + 2t^{-1}\sigma_\lambda^2 \leq C2^{|\lambda|d}\frac{\log^3 t}{t^2}.$$

In either case, we have therefore for all $|\lambda| < j$ and all $t > 0$,

$$V_{\lambda,t} \leq C2^{|\lambda|d}\frac{\log^3 t}{t^2}.$$

To apply the Gaussian approximation to $T_{\epsilon(t)}(\hat{\beta}_{\lambda,t})$ note that

$$\mathbb{E}\left(T_{\epsilon(t)}(\hat{\beta}_{\lambda,t}) - \beta_{\lambda,0}\right)^2 \leq 2\mathbb{E}\left(T_{\epsilon(t)}(\hat{\beta}_{\lambda,t}) - T_{\epsilon(t)}(\hat{\eta}_{\lambda,t})\right)^2 + 2\mathbb{E}\left(T_{\epsilon(t)}(\hat{\eta}_{\lambda,t}) - \beta_{\lambda,0}\right)^2$$

Since the mapping $y \to T(y, \epsilon)$ is a contraction (see Donoho *et al.* [9]) regardless of the value of ϵ, it follows that

$$\mathbb{E}\left(T_{\epsilon(t)}(\hat{\beta}_{\lambda,t}) - \beta_{\lambda,0}\right)^2 \leq 2V_{\lambda,t} + 2r(\epsilon(t); t^{-1/2}\sigma_\lambda; \beta_{\lambda,0}),$$

where $r(\epsilon(t); \sigma; \beta)$ is the Gaussian mean squared error $\mathbb{E}(T_\epsilon(\beta + \sigma Z) - \beta)^2$ for estimation of β from a single Gaussian observation with mean β and variance σ^2. Since all intensities $h \in F^s_{p,p}(M)$ are uniformly bounded we have $\sigma_\lambda \leq \|h\|_\infty$ and therefore

$$\mathbb{E}\left(T_{\epsilon(t)}(\hat{\beta}_{\lambda,t}) - \beta_{\lambda,0}\right)^2 \leq 2V_{\lambda,t} + 2r(\epsilon(t); t^{-1/2}\|h\|_\infty; \beta_{\lambda,0}). \tag{7.11}$$

Using the level dependent threshold $\epsilon(t) = 2^{\nu|\lambda|}t^{-1/2}\sqrt{|\log t|}$, the upper bound in inequality (7.11), the fact that h belongs to a Besov ball $B^{s+\nu}_{p,p}(\tilde{M})$ for some finite constant $\tilde{M}$ and the stability property (7.10), we obtain the rate

$$\mathbb{E}\left(\sum_{|\lambda|<j}(\delta_t(\hat{\alpha}_{t,\lambda}) - \langle u^j_\lambda, Kf\rangle)^2\right) = O\left((\frac{1}{t}\sqrt{|\log t|})^{2s/(2s+2\nu+d)}\right),$$

as a particular case of classical results on soft wavelet thresholding (for example see Cohen *et al.* [4]). Combining the above upper bound with inequality (7.9) and using again Lemma 3.2 concludes the proof of the theorem. □

References

[1] Abramovitch F. and Silverman B.W. (1998), wavelet decomposition approaches to statistical inverse problems, *Biometrika*, **85**, 115–129.

[2] Barron A.R and Sheu C.H. (1991). Approximation of density functions by sequences of exponential families, *Ann. Statist.*, **19**, No. 3, 1347-1369.

[3] Cavalier L. and Koo J.Y. (2002) Poisson intensity estimation for tomographic data using a wavelet shrinkage approach.*IEEE Trans. on Information Theory* **48**, 2794-2802.

[4] Cohen A., DeVore R., Kerkyacharian G. and Picard D. (2001). Maximal spaces with given rate of convergence for thresholding algorithms, *Appl. Comp. Harm. Anal.* , **11**, 167–191.

[5] Cohen A. Hoffmann M. and Reiss M. (2003). Adaptive wavelet-galerkin methods for inverse problems, to appear in *SIAM J. Numerical Analysis*.

[6] Csiszar I. (1975) I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* Vol. 3, No. 1, 146-158.

[7] Daubechies I. (1992) Ten Lectures on Wavelets. *CBMS-NSF Lecture Notes nr. 61, SIAM.*

[8] Donoho D. (1995) Nonlinear solution of linear inverse problems by Wavelet-Vaguelette decomposition, *Appl. Comput. Harmon. Anal.*, **2**,101-126.

[9] Donoho D.L., Johnstone I.M., Kerkyacharian, G. and Picard, D. (1996), Density estimation by wavelet thresholding, *Ann. Statist.*, **24**, 508-539.

[10] Figueiredo M. and Nowak R. (2003) An EM Algorithm for Wavelet-Based Image Restoration. Rice University Technical report to appear in *IEEE Transactions on Image Processing*.

[11] Johnstone, I.M. and Silverman, B.W. (1990). Speed of estimation in positron emission tomography and related inverse problems, *The Annals of Statistics*, **18** 251?280.

[12] Hall P. and Smith R.L. (1988) The kernel method for unfolding sphere size distributions *J. Comput. Phys.* **74**, 409-421.

[13] Koo J.Y and Kim W.C. (1996) Wavelet density estimation by approximation of log-densities, *Stat. Proba. Letters.*, **26**, 271-278.

[14] Koo J.Y and Kim W.C. (2002) Inhomogeneous Poisson intensity estimation via information projections onto wavelet subspaces. *Journal of Korean Statistical Society* **31**, 5, 343-357.

[15] Kovac A. and Silverman B.W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Am. Statist. Assoc.* **95**, 172–83.

[16] Mase S. (1995) Stereological estimation of particle size distributions, *Adv. Appl. Prob.* **27**, 350-366

[17] Meyer Y. (1992) Wavelets and operators. *Advanced mathematics. Cambridge University Press.*

[18] Nowak, R.D. and Kolaczyk, E.D. (2000) A Bayesian multiscale framework for Poisson inverse problems. *IEEE Transactions on Information Theory*, 46:5, 1811-1825.

[19] Nychka, D. and D. D. Cox (1989) Convergence rates for regularized solutions of integral equations from discrete noisy data, *The Annals of Statistics*, **17**(5), 556-572.

[20] Nychka D., Wahba G., Goldfarb S. and Pugh T. (1984). Cross-validated spline methods for the estimation of three-dimensional tumor size distributions from observations on two-dimensional cross sections. *J. Amer. Statist. Assoc.*, **79**, 832–846.

[21] O'Sullivan F. (1986) A statistical perspective on ill-posed problems. *Statist. Sci.*, **1**, 502–518.

[22] Reynaud-Bouret P. (2003) Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields* 126:1, 103-153.

[23] Silverman B.W., Jones M.C., Wilson J.D. and Nychka D. (1990). A smoothed EM approach to indirect estimation problems with particular reference to stereology and emission tomography. *J. R. Statist. Soc., B*, **52**, 271–324.

[24] Szkutnik Z. (2000) Unfolding intensity function of a Poisson process in models with approximately specified folding operator, *Metrika* **52** 1-26.

[25] Taylor C. (1983) A new method for unfolding sphere size distributions *J. Microsc.* **132**, 57-66

[26] Vannucci M. and Corradi F. (1999). Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. *J. R. Statist. Soc.*, B **61**, 971–86.

[27] Weese J., Korat E., Maier D. and Honerkamp J. (1997). Unfolding Sphere Size Distributions with a Density estimator based on Tikhonov regularization. *J. of Comput. Physics*, **138**, 331–353.