Statistica Sinica **28** (2018), 2885-2901
doi:https://doi.org/10.5705/ss.202017.0029

LIKELIHOOD RATIO HAAR VARIANCE STABILIZATION AND NORMALIZATION FOR POISSON AND OTHER NON-GAUSSIAN NOISE REMOVAL

Piotr Fryzlewicz

London School of Economics

Abstract: We propose a methodology for denoising, variance-stabilizing, and normalizing signals whose varying mean and variance are linked via a single parameter, such as Poisson or scaled chi-squared. Our key observation is that the signed and square-rooted generalized log-likelihood ratio test for the equality of the local means is approximately distributed as standard normal under the null. We use these test statistics within the Haar wavelet transform at each scale and location, referring to them as the *likelihood ratio Haar (LRH) coefficients* of the data. In the denoising algorithm, the LRH coefficients are used as thresholding decision statistics, which enables the use of thresholds suitable for i.i.d. Gaussian noise. In the variance-stabilizing and normalizing algorithm, the LRH coefficients replace the standard Haar coefficients in the Haar basis expansion. We prove the consistency of our LRH smoother for Poisson counts with a near-parametric rate, and various numerical experiments demonstrate the good practical performance of our methodology.

Key words and phrases: Anscombe transform, Box-Cox transform, Gaussianization, Haar-Fisz, log transform, variance-stabilizing transform.

1. Introduction

The popularity of wavelets and their potential for useful applications in data science did not escape the attention of Peter Hall, who wrote, amongst others, on threshold choice in wavelet curve estimation (Hall and Patil (1996b,a)), wavelet methods for functions with many discontinuities (Hall, McKay and Turlach (1996)), wavelets for regression with irregular design (Hall and Turlach (1997)) and block-thresholded wavelet estimators (Hall, Kerkyacharian and Picard (1999)). I learned of Peter through wavelets by reading some of his papers on the topic during my doctoral study. I remember my surprise at discovering that both my then PhD supervisor, Guy Nason, and someone else I knew, Prakash Patil, had co-authored papers with Peter Hall. When I shared my surprise with Guy, he responded by saying that he did not know many people who were *not* Peter's co-authors!

Even though I can unfortunately count myself in this "minority" category, I have learned and am still learning a great deal from Peter, especially by appreciating the careful and elegant way in which he used mathematics to support his arguments.

Traditional wavelet transformations are orthonormal transformations of the input data into coefficients that carry information about the local behaviour of the data at a range of dyadic scales and locations. They tend to offer sparse representation of the input data, with a small number of wavelet coefficients often being able to encode much of the energy of the input signal, and are computable and invertible in linear time via recursive pyramid algorithms (Mallat (1989); Daubechies (1992)). Reviews of the use of wavelets in statistics can be found, for example, in Vidakovic (1999) and Nason (2008). One canonical task facilitated by wavelets is the removal of noise from signals, which usually proceeds by taking a wavelet transform of the data, thresholding away the (typically many) wavelet coefficients that are small in magnitude, preserving those few that are large in magnitude, and taking the inverse wavelet transform. Since the seminal paper by Donoho and Johnstone (1994) in which the general idea was first proposed, several other methods for wavelet smoothing of one-dimensional signals have appeared, but the vast majority make the i.i.d. Gaussian noise assumption. By contrast, the focus of this article is the treatment of signals in which the variance of the noise is a function of its mean; this includes Poisson- or scaled-chi-squared-distributed signals. (Throughout the paper, we refer to a distribution as a 'scaled chi-squared', or simply 'chi-squared', if it takes the form $\sigma^2 m^{-1} \chi^2_m$.)

The simplest example of a wavelet transform, and the focus of this article, is the Haar transform, which can be described as a sequence of symmetric scaled differences of consecutive local means of the data, computed at dyadic scales and locations and naturally forming a binary tree consisting of 'parents' and 'children'. Its local difference mechanism means that it offers sparse representations for (approximately) piecewise-constant signals. Our starting point is the observation that testing whether or not each Haar coefficient of a signal exceeds a certain threshold (in the denoising task described above) can be interpreted as the *likelihood ratio test* for the equality of the corresponding local means of the signal in the i.i.d. Gaussian noise model. In this paper, we take this observation further and propose similar multiscale likelihood ratio tests for other distributions, most notably those in which the variance is a function of the mean, such as Poisson or scaled chi-squared. The proposed multiscale likelihood ratio tests reduce to the traditional thresholding of Haar wavelet coefficients for Gaussian data, but take entirely new forms for other distributions. This leads to a new, unified class of algorithms useful for problems such as e.g. Poisson intensity estimation, Poisson image denoising, spectral density estimation in time series, or time-varying volatility estimation in finance. (Extension of our methodology to images is as straightforward as the extension of the standard one-dimensional Haar wavelet transform to two dimensions.)

The new multiscale likelihood ratio tests naturally induce a new construction, *likelihood ratio (Haar) wavelets*, which have the benefit of producing (equivalents of) Haar wavelet coefficients that are asymptotically standard normal under the null hypothesis of the corresponding local means being equal, even for inhomogeneous non-Gaussian signals. This makes it much easier to choose a single threshold parameter in smoothing these kinds of data, and serves as a basis for new normalizing transformations for these kinds of data that bring their distribution close to Gaussianity. We demonstrate both these phenomena. The device that enables these results is Wilks' theorem, according to which the signed square-rooted likelihood ratio statistic is often approximately distributed as standard normal, a fact that, we believe, has not been explored in a variance-stabilization context before.

Wavelet-based Poisson noise removal, with or without the use of a variance-stabilizing and/or normalizing transform, has a long history. For a Poisson variable X, the Anscombe (1948) transform $2(X + 3/8)^{1/2}$ brings its distribution to approximate normality with variance one. Donoho (1993) proposes to preprocess Poisson data via the Anscombe transform, and then use wavelet-based smoothing techniques suitable for i.i.d. Gaussian noise. This and a number of other wavelet-based techniques for denoising Poisson-contaminated signals are reviewed and compared in Besbeas, De Feis and Sapatinas (2004). These include the translation-invariant multiscale Bayesian techniques by Kolaczyk (1999a) and Timmermann and Nowak (1997, 1999), shown to outperform earlier techniques in Kolaczyk (1997, 1999b) and Nowak and Baraniuk (1999). Willett and Nowak (2003) propose the use of "platelets" in Poisson image denoising. The Haar-Fisz methodology of Fryzlewicz and Nason (2004), drawing inspiration from earlier work by Fisz (1955) outside the wavelet context, proceeds by decomposing the Poisson data via the standard Haar transform, then variance-stabilizing the Haar coefficients by dividing them by the MLE of their own standard deviation, and then using thresholds suitable for i.i.d. Gaussian noise with variance one. Closely related ideas appear in Luisier et al. (2010) and Reynaud-Bouret and Rivoirard (2010). Jansen (2006) extends the Haar-Fisz idea to other wavelets. As an alternative to Anscombe's transform, which is known not to work well for low Poisson intensities, Zhang, Fadili and Starck (2008) introduce a more involved square-root-type variance-stabilizing transform for (filtered) Poisson data. Hirakawa and Wolfe (2012) propose Bayesian Haar-based shrinkage for Poisson signals based on the exact distribution of the difference of two Poisson variates (the Skellam distribution).

In multiplicative set-ups, such as signals distrubuted as $X_k = \sigma_k^2 m^{-1} \chi_m^2$, the logarithmic transform stabilizes the variance exactly, but does not bring the distribution of the transformed X_k close to normality, especially not for small values of m such as 1 or 2. In the context of spectral density estimation in time series, in which the signal is approximately exponentially distributed, wavelet shrinkage for the logged (and hence variance-stabilized) periodogram is studied, amongst others, in Moulin (1994), Gao (1997), Pensky, Vidakovic and De Canditis (2007) and Freyermuth, Ombao and von Sachs (2010). An alternative route, via pre-estimation of the variance of the wavelet coefficients (rather than via variance stabilization) is taken in Neumann (1996). Haar-Fisz or wavelet-Fisz estimation for the periodogram or other (approximate) chi-squared models is developed in Fryzlewicz, Sapatinas and Subba Rao (2006), Fryzlewicz and Nason (2006) and Fryzlewicz, Nason and von Sachs (2008). In more general settings, wavelet estimation for exponential families with quadratic or cubic variance functions is considered in Antoniadis and Sapatinas (2001), Antoniadis, Besbeas and Sapatinas (2001) and Brown, Cai and Zhou (2010). The Haar-Fisz or wavelet-Fisz transformations for unknown distributions are studied in Fryzlewicz (2008), Fryzlewicz, Delouille and Nason (2007), Motakis et al. (2006) and Nason (2014). Variance-stabilizing transformations are reviewed in the (unpublished) manuscript by Foi (2009).

Our approach departs from the existing literature in that our variance-stabilization and normalization device does not involve either the pre-estimation of the variance (as, effectively, in the Haar-Fisz transform) or the application of a Box-Cox-type transform (as in the Anscombe variance stabilization for Poisson data or the logarithmic transform in multiplicative models). By contrast, we use the entire likelihood for the purpose of variance-stabilization and normalization. As a result, the thresholding decision in our proposed smoothing methodology is not based on the usual wavelet detail coefficients, but on the newly-proposed likelihood ratio Haar coefficients. For completeness, we mention that Kolaczyk and Nowak (2004) construct multiscale decompositions of the Poisson likelihood, which leads them to consider binomial likelihood ratio tests for the purpose of thresholding; however, this is done in a context that does not use the signed and square-rooted generalized log-likelihood ratio tests or utilize their variance-stabilizing or normalizing properties.

The paper is organized as follows. Section 2 motivates and introduces the concept of likelihood ratio Haar coefficients and outlines our general methodology for smoothing and variance stabilization/normalization. Section 3 describes our method in two special cases, those of the Poisson and the scaled chi-squared distribution. Section 4 formulates and discusses a consistency result for the Poisson smoother. Section 5 provides a numerical study illustrating the practical performance of our smoothing and variance stabilization/normalization algorithms.

2. General Methodology

Let $X_1, \ldots, X_n$ be a sequence of independent univariate random variables such that $X_k \sim F(\theta_k)$, where $F(\theta)$ is a family of distributions parameterized by a scalar parameter θ such that $\mathbb{E}(X_k) = \theta_k$. Our two running examples are: $X_k \sim \mathrm{Pois}(\lambda_k)$, and $X_k \sim \sigma_k^2 m^{-1} \chi_m^2$ (throughout the paper, we refer to the latter example as 'scaled chi-squared' or simply 'chi-squared'). Extensions to higher-dimensional parameters are possible, but certain aspects of the asymptotic normality are then lost.

We recall the traditional Haar transform and the fundamentals of signal smoothing via (Haar) wavelet thresholding. In the following, we assume that $n = 2^J$, where J is an integer. Extensions to non-dyadic n are possible, see e.g. Wickerhauser (1994). Given the input data $\mathbf{X} = (X_1, \ldots, X_n)$, we define $\mathbf{s}_0 = (s_{0,1}, \ldots, s_{0,n}) = \mathbf{X}$. The Haar transform recursively performs the steps

$$s_{j,k} = 2^{-1/2}(s_{j-1,2k-1} + s_{j-1,2k}), \quad d_{j,k} = 2^{-1/2}(s_{j-1,2k-1} - s_{j-1,2k}), \qquad (2.1)$$

for $j = 1, \ldots, J$ and $k = 1, \ldots, 2^{J-j}$. The indices j and k are thought of as "scale" and "location" parameters, respectively, and the coefficients $s_{j,k}$ and $d_{j,k}$ as the "smooth" and "detail" coefficients (respectively) at scale j, location k. It is easy to express $s_{j,k}$ and $d_{j,k}$ as explicit functions of $\mathbf{X}$:

$$s_{j,k} = 2^{-j/2} \sum_{i=(k-1)2^j+1}^{k2^j} X_i, \; d_{j,k} = 2^{-j/2} \left(\sum_{i=(k-1)2^j+1}^{(k-1)2^j+2^{j-1}} X_i - \sum_{i=(k-1)2^j+2^{j-1}+1}^{k2^j} X_i \right).$$

Defining $\mathbf{d}_j = (d_{j,k})_{k=1}^{2^{J-j}}$, the Haar transform H of $\mathbf{X}$ is $H(\mathbf{X}) = (\mathbf{d}_1, \ldots, \mathbf{d}_J, s_{J,1})$. The "pyramid" algorithm at (2.1) enables the computation of $H(\mathbf{X})$ in $O(n)$ operations. $H(\mathbf{X})$ is an orthonormal transform of $\mathbf{X}$ and can be inverted by undoing (2.1). If the mean signal $\Theta = (\theta_1, \ldots, \theta_n)$ is piecewise-constant, then those coefficients $d_{j,k}$ that correspond to the locally constant segments of Θ are zero-centered. This justifies a procedure for estimating the mean vector Θ: take the Haar transform of $\mathbf{X}$, retain those coefficients $d_{j,k}$ for which $|d_{j,k}| > t$ for a certain threshold t and set the others to zero, then take the inverse Haar transform of the thus-"hard"-thresholded vector. In the i.i.d. Gaussian noise model, in which $X_k = \theta_k + \varepsilon_k$, where $\varepsilon \sim N(0, \sigma^2)$ with σ^2 assumed known, the operation $|d_{j,k}| > t$ is the likelihood ratio test for the local constancy of Θ in the following sense.

1. Assume $(\theta_u)_{u=(k-1)2^j+1}^{(k-1)2^j+2^{j-1}} = \theta^{(1)}$ for all u, and $(\theta_v)_{v=(k-1)2^j+2^{j-1}+1}^{k2^j} = \theta^{(2)}$ for all v. The indices u (respectively v) are the same as those corresponding to the X_u's (X_v's) with positive (negative) weights in $d_{j,k}$.
2. Test $H_0 : \theta^{(1)} = \theta^{(2)}$ against $H_1 : \theta^{(1)} \neq \theta^{(2)}$; the Gaussian likelihood ratio test reduces to $|d_{j,k}| > t$, where t is naturally related to the desired significance level. H_0 can alternatively be phrased as $E(d_{j,k}) = 0$, and H_1 as $E(d_{j,k}) \neq 0$.

Because under each H_0, the variable $d_{j,k}$ is distributed as $N(0, \sigma^2)$ due to the orthonormality of the Haar transform, the same t can meaningfully be used across different scales and locations (j, k).

In models other than Gaussian, the operation $|d_{j,k}| > t$ can typically no longer be interpreted as the likelihood ratio test for the equality of $\theta^{(1)}$ and $\theta^{(2)}$. Moreover, the distribution of $d_{j,k}$ is not generally the same under each H_0 but will, in many models, vary with the local (unknown) parameters $(\theta_i)_{i=(k-1)2^j+1}^{k2^j}$, which makes the selection of t in the operation $|d_{j,k}| > t$ challenging. This is, for example, the case in our running examples, $X_k \sim \text{Pois}(\lambda_k)$ and $X_k \sim \sigma_k^2 m^{-1} \chi_m^2$, both of which are such that $\text{Var}(X_k)$ is a non-trivial function of $E(X_k)$, which translates into the dependence of $d_{j,k}$ on the local means vector $(\theta_i)_{i=(k-1)2^j+1}^{k2^j}$, even under the null hypothesis $E(d_{j,k}) = 0$.

In the (non-Gaussian) model under consideration, our proposal is to remedy this by replacing the operation $|d_{j,k}| > t$ with a likelihood ratio test for $H_0 : \theta^{(1)} = \theta^{(2)}$ against $H_1 : \theta^{(1)} \neq \theta^{(2)}$ suitable for the distribution at hand. More specifically, denoting by $L(\theta \,|\, X_{k_1}, \ldots, X_{k_2})$ the likelihood of the constant parameter θ given the data $X_{k_1}, \ldots, X_{k_2}$, and by $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}$ the MLEs of $\theta^{(1)}, \theta^{(2)}$, respectively, we design a new Haar-like transform, in which we replace the "test statistic" $d_{j,k}$ by

$$g_{j,k} = \operatorname{sign}\left(\hat{\theta}^{(1)} - \hat{\theta}^{(2)}\right) \left\{ 2\log \frac{\sup_{\theta^{(1)}} L(\theta^{(1)} \mid X_{(k-1)2^j+1}, \ldots, X_{(k-1)2^j+2^{j-1}}) \, \sup_{\theta^{(2)}} L(\theta^{(2)} \mid X_{(k-1)2^j+2^{j-1}+1}, \ldots, X_{k2^j})}{\sup_{\theta} L(\theta \mid X_{(k-1)2^j+1}, \ldots, X_{k2^j})} \right\}^{1/2}, \tag{2.2}$$

the signed and square-rooted generalized log-likelihood ratio statistic for testing H_0 against H_1. The rationale is that by Wilks' theorem, under H_0, this quantity is asymptotically distributed as $N(0,1)$ for a class of models that includes, amongst others, our two running examples. We refer to $g_{j,k}$ as the *likelihood ratio Haar coefficient* of $\mathbf{X}$ at scale j and location k. By performing this replacement, we tailor-make a new Haar transform suitable for the distribution of the input vector.

2.1. General methodology for smoothing

We now outline the general methodology for signal smoothing (denoising) involving likelihood ratio Haar wavelets. The problem is to estimate Θ from $\mathbf{X}$. Let $\mathbb{I}$ be the indicator function. The basic smoothing algorithm proceeds as follows.

1. With $\mathbf{X}$ on input, compute the coefficients $s_{j,k}$, $d_{j,k}$, and $g_{j,k}$ as defined by (2.1) and (2.2).
2. Estimate each $\mu_{j,k} := E(d_{j,k})$ by
$$\hat{\mu}_{j,k} = \begin{cases} 0 & j = 1, \ldots, J_0, \\ d_{j,k}\mathbb{I}(|g_{j,k}| > t) & j = J_0 + 1, \ldots, J. \end{cases} \tag{2.3}$$
3. Defining $\hat{\mu}_j = (\hat{\mu}_{j,k})_{k=1}^{2^{J-j}}$, compute the inverse Haar transform of the vector $(\hat{\mu}_1, \ldots, \hat{\mu}_J, s_{J,1})$ and use it as the estimate $\hat{\Theta}$ of Θ.

We set $\hat{\mu}_{j,k} = 0$ at the finest scales because of a certain strong asymptotic normality argument; see the proof of Theorem 1. This theorem also specifies the permitted magnitude of J_0. The operation in the second line of (2.3) is referred to as hard thresholding; soft thresholding, in which the surviving coefficients are shrunk towards zero, is also possible. The threshold t is a tuning parameter of the procedure and we discuss its selection later. The above algorithm differs from the standard smoothing using Haar wavelets in that we use $g_{j,k}$, rather than $d_{j,k}$, as the thresholding statistic.

2.2. General methodology for variance stabilization and normalization

Due to the fact that $g_{j,k}$ will typically be distributed as close to $N(0,1)$ under each H_0 (that is, for the majority of scales j and locations k), replacing the coefficients $d_{j,k}$ with $g_{j,k}$ can be viewed as "normalizing" or "Gaussianizing" the input signal in the Haar wavelet domain. The standard inverse Haar transform will then yield a normalized version of the input signal. We outline the basic algorithm.

1. With $\mathbf{X}$ on input, compute the coefficients $s_{j,k}$ and $g_{j,k}$ as defined by (2.1) and (2.2).

2. Defining $\mathbf{g}_j = (g_{j,k})_{k=1}^{2^{J-j}}$, compute the inverse Haar transform of the vector $(\mathbf{g}_1, \ldots, \mathbf{g}_J, s_{J,1})$ and denote the resulting vector by $G(\mathbf{X})$.

Throughout the paper, we refer to $G(\mathbf{X})$ as the *likelihood ratio Haar transform* of X. In the online supplement, we show that the likelihood Haar transform is invertible, at least in the Poisson and chi-squared cases. An invertible variance-stabilization transformation such as $G(\mathbf{X})$ is useful as it enables the smoothing of X in a modular way: (i) apply $G(X)$, (ii) use any smoother suitable for i.i.d. standard normal noise, (iii) take the inverse of $G(X)$ to obtain a smoothed version of X.

3. Specific Examples: Poisson and Chi-Squared

For $X_i \sim \text{Pois}(\lambda)$, we have $P(X_i = k) = \exp(-\lambda)(\lambda^k/k!)$ for $k = 0, 1, \ldots$, and if $X_s, \ldots, X_e \sim \text{Pois}(\lambda)$, then the MLE $\hat{\lambda}$ of λ is $\bar{X}_s^e = \{1/(e-s+1)\}\sum_{i=s}^{e} X_i$. This, after straightforward algebra, leads to

$$\begin{aligned} g_{j,k} = \operatorname{sign}&\left(\bar{X}_{(k-1)2^j+1}^{(k-1)2^j+2^{j-1}} - \bar{X}_{(k-1)2^j+2^{j-1}+1}^{k2^j}\right) 2^{j/2} \\ &\times \left\{\log\left(\bar{X}_{(k-1)2^j+1}^{(k-1)2^j+2^{j-1}}\right) \bar{X}_{(k-1)2^j+1}^{(k-1)2^j+2^{j-1}} \right. \\ &+ \log\left(\bar{X}_{(k-1)2^j+2^{j-1}+1}^{k2^j}\right) \bar{X}_{(k-1)2^j+2^{j-1}+1}^{k2^j} \\ &\left. - \; 2 \log\left(\bar{X}_{(k-1)2^j+1}^{k2^j}\right) \bar{X}_{(k-1)2^j+1}^{k2^j}\right\}^{1/2}, \end{aligned} \tag{3.1}$$

using the convention $0 \log 0 = 0$. For $X_i \sim \sigma_i^2 m^{-1}\chi_m^2 = \Gamma(m/2, m/(2\sigma_i^2))$, if $X_s, \ldots, X_e \sim \Gamma(m/2, m/(2\sigma^2))$, then the MLE $\hat{\sigma}^2$ of σ^2 is $\bar{X}_s^e = \{1/(e-s+1)\}$ $\sum_{i=s}^{e} X_i$. This gives

$$g_{j,k} = \operatorname{sign}\left(\bar{X}_{(k-1)2^j+1}^{(k-1)2^j+2^{j-1}} - \bar{X}_{(k-1)2^j+2^{j-1}+1}^{k2^j}\right) 2^{j/2}$$

$$\times\left[m\left\{\log\left(\bar{X}^{k2^j}_{(k-1)2^j+1}\right)-\frac{1}{2}\log\left(\bar{X}^{(k-1)2^j+2^{j-1}}_{(k-1)2^j+1}\right)\right.\right.$$
$$\left.\left.-\frac{1}{2}\log\left(\bar{X}^{k2^j}_{(k-1)2^j+2^{j-1}+1}\right)\right\}\right]^{1/2}. \tag{3.2}$$

Up to the multiplicative factor $m^{1/2}$, the form of the transform in (3.2) is the same for any m, which means that the chi-squared likelihood ratio Haar coefficients $g_{j,k}$ (computed with an arbitrary m) also achieve variance stabilization if m is unknown (but possibly to a constant different from one). In both the Poisson and the chi-squared cases, $g_{j,k}$ is a function of the local means $\bar{X}^{(k-1)2^j+2^{j-1}}_{(k-1)2^j+1}$ and $\bar{X}^{k2^j}_{(k-1)2^j+2^{j-1}+1}$, which is unsurprising as these are sufficient statistics for the corresponding population means in both these distributions. These local means and, therefore, the coefficients $g_{j,k}$, can be computed in computational time $O(n)$ using the pyramid algorithm at (2.1).

4. L_2 Theory for the Likelihood Ratio Haar Poisson Smoother

In this section, we provide a theoretical mean-square analysis of the performance of the signal smoothing algorithm involving likelihood ratio Haar wavelets, described in Section 2.1. Although we focus on the Poisson distribution, the statement of the result and the mechanics of the proof are similar for certain other distributions, including scaled chi-squared.

Theorem 1. *Let $\Lambda = (\lambda_1, \ldots, \lambda_n)$ be a positive piecewise-constant vector, there exist up to N indices $\eta_1, \ldots, \eta_N$ for which $\lambda_{\eta_i} \neq \lambda_{\eta_i - 1}$. Let $n = 2^J$, where J is a positive integer. Assume Λ is bounded from above and away from zero, and let $\bar{\Lambda} := \max_i \lambda_i$, $\underline{\Lambda} := \min_i \lambda_i$, $\Lambda' = \bar{\Lambda} - \underline{\Lambda}$ and $\bar{\lambda}^e_s = \{1/(e-s+1)\}\sum_{i=s}^e \lambda_i$. Let $X_k \sim Pois(\lambda_k)$ for $k = 1, \ldots, n$. Let $\hat{\Lambda}$ be obtained as in the algorithm of Section 2.1, using threshold t and with a fixed $\beta \in (0,1)$ such that $J_0 = \lfloor \log_2 n^\beta \rfloor$. Then, with $d_{j,k}$ and $\mu_{j,k}$ defined in the algorithm of Section 2.1, and with $\bar{X}^e_s = \{1/(e-s+1)\}\sum_{i=s}^e X_i$, on set $\mathcal{A} \cap \mathcal{B}$, where*

$$\mathcal{A} = \{\forall\ j = J_0+1, \ldots, J,\ k = 1, \ldots, 2^{J-j} \quad (\bar{\lambda}^{k2^j}_{(k-1)2^j+1})^{-1/2}|d_{j,k} - \mu_{j,k}| < t_1\},$$
$$\mathcal{B} = \{\forall\ j = J_0, \ldots, J,\ k = 1, \ldots, 2^{J-j}$$
$$2^{j/2}(\bar{\lambda}^{k2^j}_{(k-1)2^j+1})^{-1/2}|\bar{X}^{k2^j}_{(k-1)2^j+1} - \bar{\lambda}^{k2^j}_{(k-1)2^j+1}| < t_2\},$$

whose probability approaches 1 as $n \to \infty$ if $t_1 = C_1 \log^{1/2} n$ and $t_2 = C_2 \log^{1/2} n$ with $C_1 > \{2(1-\beta)\}^{1/2}$ and $C_2 > \{2(1-\beta)\}^{1/2}$, if threshold t is such that

$$t \geq \frac{t_1}{(1 - t_2 2^{-(J_0+1)/2}\underline{\Lambda}^{-1/2})^{1/2}}, \tag{4.1}$$

we have

$$n^{-1}\|\hat{\Lambda}-\Lambda\|^2 \le \frac{1}{2}n^{-1}N(\Lambda')^2(n^{\beta}-1) + 2n^{-1}N\bar{\Lambda}^{1/2}\left\{(J-J_0)(t^2+t_1^2)\bar{\Lambda}^{1/2}+t^2t_2(2+2^{1/2})n^{-\beta/2}\right\} + n^{-1}t_1^2\bar{\lambda}_1^n,$$

where $\|\cdot\|$ *is the* l_2*-norm of its argument.*

Bearing in mind the magnitudes of t_2 and J_0, we can see that the term $t_2 2^{-(J_0+1)/2}\underline{\Lambda}^{-1/2}$ becomes arbitrarily close to zero for large n, and therefore, from (4.1), the threshold constant t can become arbitrarily close to t_1. In particular, it is safe to set t to be the "universal" threshold suitable for iid $N(0,1)$ noise (Donoho and Johnstone (1994)), $t=\{2\log n\}^{1/2}$. It is in this sense that our likelihood ratio Haar construction achieves variance stabilization and normalization: in order to denoise Poisson signals in which the variance of the noise depends on the local mean, we make it possible to use the universal Gaussian threshold, as if the noise were Gaussian with variance one. In classical Haar wavelet thresholding with $|d_{j,k}|>\tilde{t}$ as the thresholding decision, $\tilde{t}$ would have to depend on the level of the Poisson intensity Λ over the support of $d_{j,k}$, which is unknown; our approach circumvents this.

If the number N of change-points does not increase with the sample size n, then the dominant term in the mean-square error is of order $O(n^{\beta-1})$. This suggests that β should be set to be "arbitrarily small", in which case the MSE arbitrarily close to the parametric rate $O(n^{-1})$.

5. Practical Performance

In the online supplement, we demonstrate that the likelihood ratio Haar coefficients appear to offer better normalization and variance stabilization than the Fisz coefficients. In this section, we show that this translates into better MSE properties of the likelihood ratio Haar smoother than the analogous Haar-Fisz smoother, in both the Poisson and the exponential models, on the examples considered. For comprehensive comparison of the performance of the Haar-Fisz smoother to that of other techniques, see Fryzlewicz and Nason (2004), Besbeas, De Feis and Sapatinas (2004) and Fryzlewicz (2008), among others. Our test signals are [1] Donoho and Johnstone's (1994) `blocks` and [2] `bumps` functions, scaled to have (min, max) of [1] $(0.681, 27.029)$ and [2] $(1, 12.565)$, both of length $n=2{,}048$. We consider the following models: **(1a), (2a):** Poisson models, in which the signals [1], [2] (respectively) play the role of the Poisson intensity Λ,

Table 1. MSE over 1,000 simulations for the two methods and four models described in Section 5.

Method \ Model	(1a)	(1b)	(2a)	(2b)
Haar-Fisz	0.615	8.647	0.357	1.053
Likelihood ratio Haar	0.605	7.958	0.341	0.905

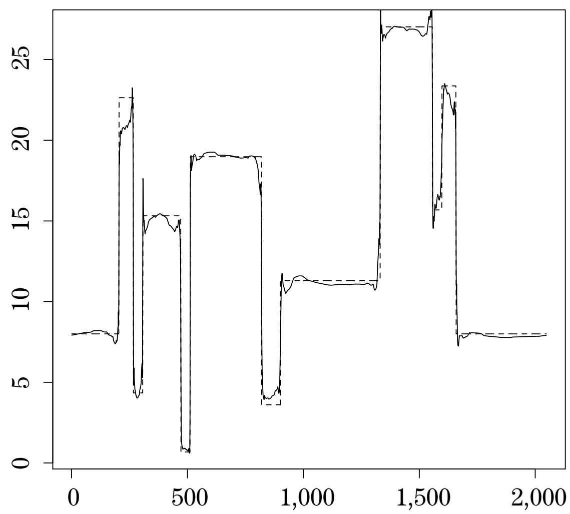

Figure 1. Sample likelihood ratio Haar reconstruction in model (1a), see Section 5 for details.

so that $X_k \sim \text{Pois}(\lambda_k)$; **(1b), (2b):** Exponential models, in which the signals [1], [2] (respectively) play the role of the exponential mean σ^2, so that $X_k \sim \sigma_k^2 \,\text{Exp}(1) = \sigma_k^2 2^{-1}\chi_2^2$.

For all models, we compared the MSE performance of "like-for-like" likelihood ratio Haar and Haar-Fisz smoothers, both constructed as described in Section 2.1, except the Haar-Fisz smoother used the corresponding coefficients $f_{j,k}$ in place of $g_{j,k}$. We used the non-decimated (translation invariant, stationary, maximum overlap) Haar transform (Nason and Silverman (1995)) to achieve fast averaging over all possibly cyclic shifts of the input data. For better comparison of the effects of thresholding alone, we used $J_0 = 0$. We used the universal threshold $t = \{2 \log n\}^{1/2}$. Figures 1 and 2 show sample reconstructions for the likelihood ratio Haar method in the Poisson models (1a), (2a).

Table 1 shows that the likelihood ratio Haar smoother outperforms Haar-Fisz across all the models considered. For the Poisson models, the improvement is fairly modest (2% for `blocks`, 4% for `bumps`) but for the exponential models,

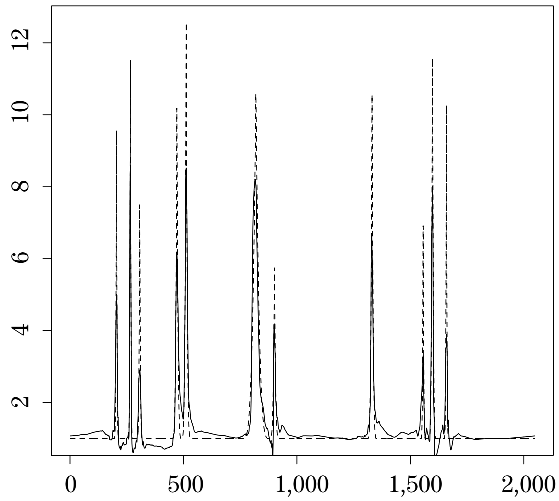

Figure 2. Sample likelihood ratio Haar reconstruction in model (2a), see Section 5 for details.

it is more significant (8% for `blocks`, 14% for `bumps`). One important reason for this improved performance is that, as demonstrated in the online supplement, the likelihood ratio Haar coefficients have a higher magnitude than the corresponding Fisz coefficients, and therefore more easily survive thresholding. This implies that the likelihood ratio Haar smoother lets through "more signal" compared to the Haar-Fisz smoother if both use the same threshold, however chosen. Another possible reason is that, as shown in the online supplement, the likelihood ratio Haar coefficients are closer to variance-one normality than the Fisz coefficients and therefore the use of thresholds designed for standard normal noise may be more suitable for them.

We now briefly illustrate the normalizing and variance-stabilizing properties of the likelihood ratio Haar transform $G(\cdot)$ described in Section 2.2, using data simulated from models (1a) and (1b). We used the non-decimated version of the Haar transform.

Figure 3 illustrates the results for the Poisson case. In both the Poisson and the exponential examples, the likelihood ratio Haar transform is a very good normalizer and variance-stabilizer: the transformed data minus the transformed signal shows good agreement with an i.i.d. normal sample; its sample variance equals 1.07 for the Poisson model and 1.14 for the exponential model. Particularly for the exponential model, the likelihood ratio Haar transform is a significantly better normalizer than the Haar-Fisz transform (not shown here).

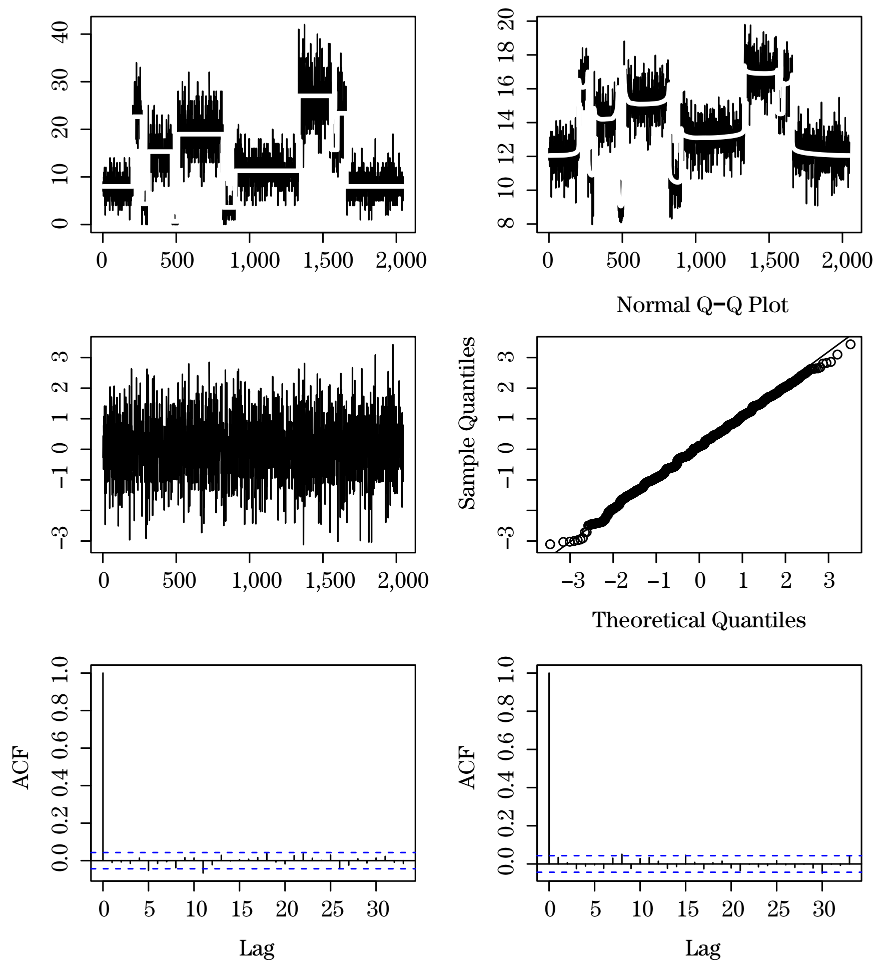

Figure 3. The Poisson model. Top left: Poisson intensity Λ (white) and simulated data $\mathbf{X}$ (black). Top right: the likelihood ratio transform $G(\Lambda)$ (white) and $G(\mathbf{X})$ (black). Middle left: $G(\mathbf{X}) - G(\Lambda)$. Middle right: Q-Q plot of $G(\mathbf{X}) - G(\Lambda)$ against the normal quantiles. Bottom left: sample acf plot of $G(\mathbf{X}) - G(\Lambda)$. Bottom right: sample acf plot of $(G(\mathbf{X}) - G(\Lambda))^2$.

5.1. California earthquake data

We revisited the Northern California earthquake dataset, analysed in Fryzlewicz and Nason (2004) and available from `http://quake.geo.berkeley.edu/ncedc/catalog-search.html`. We analyzed the time series N_k, $k = 1, \ldots, 1{,}024$, where N_k is the number of earthquakes of magnitude 3.0 or more which occurred in the kth week, the first week under consideration starting April 22nd, 1981 and the final ending December 5th, 2000. We took $N_k \sim \text{Pois}(\lambda_k)$ and estimated Λ

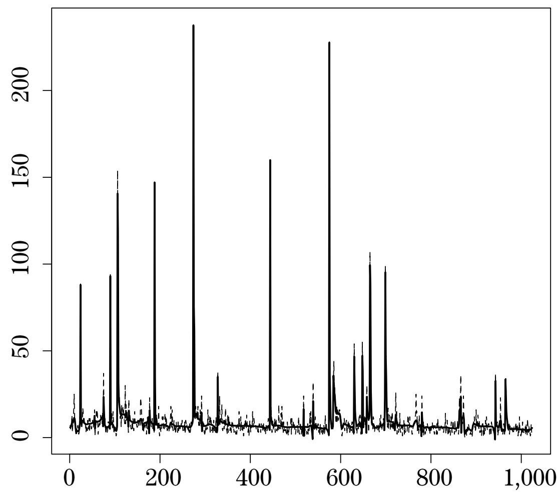

Figure 4. Northern California earthquake data: N_k (dashed) and the likelihood ratio Haar estimate (thick solid). See Section 5.1 for details.

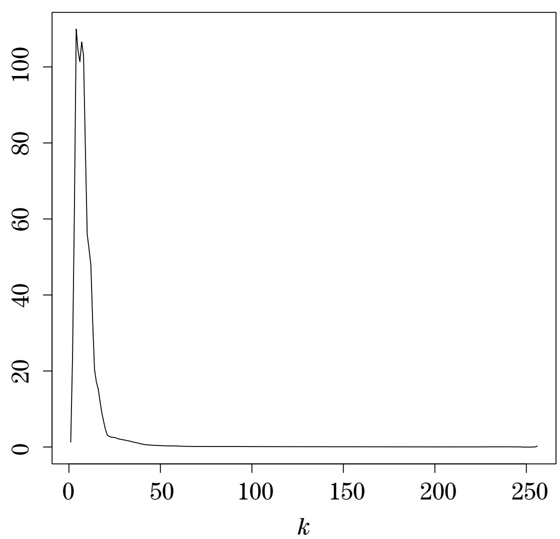

Figure 5. The likelihood ratio Haar smooth of M_k under the Poisson assumption.

using our likelihood ratio Haar smoother, as described in Section 5.

The estimate and the data are shown in Figure 4. The appearance of the estimator reveals an interesting phenomenon, not necessarily easily seen in the noisy data: for many of the intensity spikes observed in this dataset, the intensity in the time period just before the spike appears to be much lower than the intensity in the period following the spike, which may point to a degree of persistence in the seismic activity following the major spikes in activity observed in these data.

Further, we analysed the histogram of counts M_k, $k = 0, \ldots, 255$, defined as the number of weeks in which k earthquakes of magnitude 3.0 or more which occurred. The raw data (not shown here) show an apparent bimodality with modes at 4 and 6. To verify whether this is a spurious or "real" effect, we smoothed M_k using our likelihood ratio Haar smoother suitable for Poisson data (note that M_k, being a histogram, can approximately be modelled as Poisson-distributed). Figure 5 reveals that our fit preserves the bimodality, which gives support to the argument that this is a genuine, rather than spurious, effect. This finding points towards a mixture model with two components: one corresponding to "quieter" periods (i.e. those with a low intensity of earthquakes of magnitude 3.0 or more) and the other to periods with high earthquake intensity.

Supplementary Materials

The supplementary materials contain the proof of our theoretical result and further technical details.

References

Anscombe, F. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35**, 246–254.

Antoniadis, A., Besbeas, P. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with cubic variance functions. *Sankhya A Mathematical Statistics and Probability* **63**, 309–327.

Antoniadis, A. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with quadratic variance functions. *Biometrika* **88**, 805–820.

Besbeas, P., De Feis, I. and Sapatinas, T. (2004). A comparative study of wavelet shrinkage estimators for Poisson counts. *International Statistical Review* **72**, 209–237.

Brown, L., Cai, T. and Zhou, H. (2010). Nonparametric regression in exponential families. *The Annals of Statistics* **38**, 2005–2046.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia, Pa.: SIAM.

Donoho, D. (1993). Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In: *Proceedings of Symposia in Applied Mathematics*, 173–205.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

Fisz, M. (1955). The limiting distribution of a function of two independent random variables and its statistical application. *Colloquium Mathematicum* **3**, 138–146.

Foi, A. (2009). Optimization of variance-stabilizing transformations. *Preprint*.

Freyermuth, J.-M., Ombao, H. and von Sachs, R. (2010). Tree-structured wavelet estimation in a mixed effects model for spectra of replicated time series. *Journal of the American Statistical Association* **105**, 634–646.

Fryzlewicz, P. (2008). Data-driven wavelet-Fisz methodology for nonparametric function estimation. *Electronic Journal of Statistics* **2**, 863–896.

Fryzlewicz, P., Delouille, V. and Nason, G. (2007). GOES-8 X-ray sensor variance stabilization using the multiscale data-driven Haar-Fisztransform. *Journal of the Royal Statistical Society. Series C. Applied Statistics.*, to appear.

Fryzlewicz, P. and Nason, G. (2004). A Haar-Fisz algorithm for Poisson intensity estimation. *Journal of Computational and Graphical Statistics* **13**, 621–638.

Fryzlewicz, P. and Nason, G. (2006). Haar-Fisz estimation of evolutionary wavelet spectra. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **68**, 611–634.

Fryzlewicz, P., Nason, G. and von Sachs, R. (2008). A wavelet-Fisz approach to spectrum estimation. *Journal of Time Series Analysis* **29**, 868–880.

Fryzlewicz, P., Sapatinas, T. and Subba Rao, S. (2006). A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika* **93**, 687–704.

Gao, H.-Y. (1997). Choice of thresholds for wavelet shrinkage estimate of the spectrum. *Journal of Time Series Analysis* **18**, 231–252.

Hall, P., Kerkyacharian, G. and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statistica Sinica* **9**, 33–49.

Hall, P., McKay, I. and Turlach, B. (1996). Performance of wavelet methods for functions with many discontinuities. *The Annals of Statistics* **24**, 2462–2476.

Hall, P. and Patil, P. (1996a). Effect of threshold rules on performance of wavelet-based curve estimators. *Statistica Sinica* **6**, 331–345.

Hall, P. and Patil, P. (1996b). On the choice of smoothing parameter, threshold and truncation in nonparametric regression by non-linear wavelet methods. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **58**, 361–377.

Hall, P. and Turlach, B. (1997). Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *The Annals of Statistics* **25**, 1912–1925.

Hirakawa, K. and Wolfe, P. (2012). Skellam shrinkage: Wavelet-based intensity estimation for inhomogeneous Poisson data. *IEEE Transactions on Information Theory* **58**, 1080–1093.

Jansen, M. (2006). Multiscale Poisson data smoothing. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **68**, 27–48.

Kolaczyk, E. (1997). Non-parametric estimation of Gamma-ray burst intensities using Haar wavelets. *The Astrophysical Journal* **483**, 340–349.

Kolaczyk, E. (1999a). Bayesian multiscale models for Poisson processes. *Journal of the American Statistical Association* **94**, 920–933.

Kolaczyk, E. (1999b). Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. *Statistica Sinica* **9**, 119–135.

Kolaczyk, E. and Nowak, R. (2004). Multiscale likelihood analysis and complexity penalized estimation. *The Annals of Statistics* **32**, 500–527.

Luisier, F., Vonesch, C., Blu, T. and Unser, M. (2010). Fast interscale wavelet denoising of Poisson-corrupted images. *Signal Processing* **90**, 415–427.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation.

IEEE transactions on pattern analysis and machine intelligence **11**, 674–693.

Motakis, E., Nason, G., Fryzlewicz, P. and Rutter, G. (2006). Variance stabilization and normalization for one-color microarray data using a data-driven multiscale approach. *Bioinformatics* **22**, 2547–2553.

Moulin, P. (1994). Wavelet thresholding techniques for power spectrum estimation. *IEEE Transactions on Signal Processing* **42**, 3126–3136.

Nason, G. (2008). *Wavelet Methods in Statistics with R.* Springer, New York.

Nason, G. (2014). Multiscale variance stabilization via maximum likelihood. *Biometrika* **101**, 499–504.

Nason, G. and Silverman, B. (1995). The stationary wavelet transform and some statistical applications. In: *Lecture Notes in Statistics* **103** (Edited by A. Antoniadis and G. Oppenheim), 281–300. Springer-Verlag.

Neumann, M. (1996). Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series. *Journal of Time Series Analysis* **17**, 601–633.

Nowak, R. and Baraniuk, R. (1999). Wavelet domain filtering for photon imaging systems. *IEEE Transactions on Image Processing* **8**, 666–678.

Pensky, M., Vidakovic, B. and De Canditis, D. (2007). Bayesian decision theoretic scale-adaptive estimation of a log-spectral density. *Statistica Sinica* **17**, 635–666.

Reynaud-Bouret, P. and Rivoirard, V. (2010). Near optimal thresholding estimation of a Poisson intensity on the real line. *Electronic Journal of Statistics* **4**, 172–238.

Timmermann, K. and Nowak, R. (1997). Multiscale Bayesian estimation of Poisson intensities. In: *Proceedings of the Asilomar Conference on Signals, Systems and Computers*. 95–90. Pacific Grove, IEEE Computer Press, CA.

Timmermann, K. and Nowak, R. (1999). Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging. *IEEE Transactions on Information Theory* **45**, 846–862.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets.* Wiley, New York.

Wickerhauser, M. (1994). *Adapted Wavelet Analysis from Theory to Software.* AK Peters.

Willett, R. and Nowak, R. (2003). Platelets: a multiscale approach for recovering edges and surfaces in photon-limited medical imaging. *IEEE Transactions on Medical Imaging* **22**, 332–350.

Zhang, B., Fadili, J. and Starck, J.-L. (2008). Wavelets, ridgelets, and curvelets for Poisson noise removal. *IEEE Transactions on Image Processing* **17**, 1093–1108.

Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, UK.

E-mail: p.fryzlewicz@lse.ac.uk

(Received January 2017; accepted June 2017)