Multiscale Poisson data smoothing

Multiscale Poisson data smoothing

Maarten Jansen
TU Eindhoven, dept. of Math. and Comp. Sci.
K.U.Leuven, dept. of Comp. Sci.

December 2003

Abstract

This paper introduces a framework for nonlinear, multiscale decompositions of Poisson data with piecewise smooth intensity curves. The key concept is conditioning on the sum of the observations that are involved in the computation of a given coefficient. Within this framework, most classical wavelet thresholding schemes for data with additive, homoscedastic noise apply. Any family of wavelet transforms (orthogonal, biorthogonal, second generation) can be incorporated into this framework. The second contribution is a Bayesian shrinkage with an original prior for coefficients of this decomposition. As such, the method combines the advantages of the Fisz-wavelet transform and (Bayesian) Multiscale Likelihood models, with additional benefits, such as the extendibility towards arbitrary wavelet families.

Keywords

Wavelet, smoothing, Poisson count data, Bayesian estimation, Dirichlet distribution

1 Introduction

Wavelet thresholding has proven to be a successful method in non-parametric smoothing or estimation of piecewise smooth functions. Such intermittent data occur in a wide variety of applications, such as medical signals, biology, astronomy, geology, physics and all kinds of electrical signals. The most visual application is probably image denoising, where the edges correspond to the discontinuities (or jumps). Classical linear smoothing techniques, e.g. using Fourier or kernels, are doomed to fail: the output shows Gibbs phenomena and/or an often inacceptable bias — blur in image processing terminology — near the jumps. A good smoothing algorithm should start from the locations of discontinuities, since they carry the essential information of the signal. The combination of a wavelet transform, linear in itself, with the nonlinear threshold method is a fast and efficient approach to catch the singularities. As a matter of fact, a wavelet decomposition spends less asymptotical effort to the singularities than what the representation of the smooth parts in between requires from a wavelet or any other basis decomposition.

The basic idea of thresholding is *sparsity*, i.e., a wavelet transform maps a (digital) signal onto a set of wavelet *coefficients*, in which most values are close to zero, while only a limited subset of large coefficients carries the essential information. Since noise is spread out evenly, a smooth reconstruction of the underlying signal can be obtained if all coefficients with magnitude below a certain *threshold* value are replaced by zero.

While a wavelet transform automatically adapts to time (or space) varying signal characteristics, most thresholding techniques assume a constant noise behaviour: noise is typically supposed to be additive and homoscedastic (i.e., with constant variance) and white (uncorrelated), since any linear transform of uncorrelated, homoscedastic noise is again uncorrelated and homoscedastic. Homoscedastic wavelet coefficients are desirable for thresholding, since a single threshold cannot be optimal for both a coefficient with little noise and another coefficient with a large variance. The multiresolution nature of a wavelet transform offers a solution for data with homoscedastic but correlated (coloured) noise: they are mapped onto coefficients that are homoscedastic within each resolution, i.e., coefficients that correspond to basis functions with equal support width, have equal variance.

The concept of thresholding is appropriate for any noise distribution with finite variance. Most threshold assessment procedures however have been designed with additive, normal noise in mind.

The subject of this paper is Poisson data. Poisson noise has a multiplicative nature. This means: the more intense the signal is, the heavier are the fluctuations (the noise). This sort of noise results from counting processes of 'particles' which independently hit the observer. Typical examples of such Poisson processes in practice are web statistics (number of hits on a web page), (internet) traffic data, observations in astronomy, and tomographical imaging. This paper includes a discussion on two such real data examples. Poisson noise with moving average (time varying intensity) cannot be homoscedastic, since the variance equals the expected value.

Unlike the additive, normal case, a Poisson model is not stable under a linear transform, such as a wavelet transform. A straightforward method to deal with this problem a preprocessing, normalising step. An example of this strategy is the Anscombe transformation (Donoho 1993; Anscombe 1948) and the Fisz normalisation (Fisz 1955; Fryźlewicz and Nason 2003). The latter can also be considered as an example of a second direction of research: some papers (Fryźlewicz and Nason 2003; Kolaczyk and Nowak 2004, 2002; Kolaczyk 1996, 1997, 1999; Timmerman and Nowak 1999) concentrate on properties of the Haar transform and a few other wavelet bases. These specific properties allow for an exact closed form of the scaling or wavelet coefficient densities. This exact expression can be used in a general Bayesian multiscale model (Kolaczyk 1999; Timmerman and Nowak 1999). A major theme of this paper is to extend ideas from multiscale Fisz decompositions (Fryźlewicz and Nason 2003) and Bayesian Multiscale models (Kolaczyk 1999; Timmerman and Nowak 1999) to any family of wavelet transforms. The proposed procedure can thus deal with any degree of smoothness in between sharp transitions, just as in classical wavelet shrinkage. In this perspective, the proposed method can be seen as a wavelet like alternative for a multiscale partitioning with piecewise polynomials (Kolaczyk and Nowak 2003).

A third direction of research is an asymptotic study of the applicability of classical wavelet thresholding when the Poisson intensities tend to infinity. A similar point of view is looking at specific types of signals, like bursts against a homogeneous background (Kolaczyk 1996, 1997). This paper however considers signals with low intensities as well as signals with a mixture of both low-count and high-count intervals.

The method presented in this paper is also somehow related to wavelet domain filtering (shrinkage, modulation) based on unbiased variance estimation in the wavelet domain (Antoniadis and Sapatinas 2001; Antoniadis et al. 2001; Nowak and Baraniuk 1999; Jansen 2001, page 136–137). As explained below, this idea applies to a broader spectrum of noise models. Several contributions typically consider exponential families (Antoniadis and Sapatinas 2001; Antoniadis et al. 2001; Sardy et al. 2003). Instead of filtering wavelet coefficients, some methods look for a generalisation of the universal threshold to Poisson data (Charles and Rasson 2003) or wider classes of distributions (Sardy et al. 2003).

For a more complete overview overview of the literature, we refer to an extensive comparative study of several existing methods in (Besbeas et al. 2004).

This paper is organised as follows: Section 2 introduces a conditional variance stabilisation. This is followed by a discussion in Section 3 before proceeding, in Section 4 to a Bayesian shrinkage approach embedded in this framework of conditional variance stabilisation. Several aspects of this Bayesian model are further elaborated in Section 5. A simulation study is presented in Section 6. This simulation compares the method with the normalisations by Anscombe (Donoho 1993) and Fisz (Fryźlewicz and Nason 2003), and with the existing Bayesian multiscale model (Kolaczyk 1999). Not only are these methods most closely related to the method presented in this paper, they are also considered among the state-of-the-art of the currently available methods (Besbeas et al. 2004).

2 The proposed conditional variance stabilisation (CVS)

2.1 Definitions

Suppose $\boldsymbol{X}$ is a vector of n Poisson data X_i with intensities $\boldsymbol{\lambda} = [\lambda_i]_{i=1\ldots n}$. We call $\boldsymbol{W}$ the wavelet transform of these data:

$$\boldsymbol{W} = \widetilde{W}\boldsymbol{X},$$

where $\widetilde{W}$ stands for the forward wavelet transform matrix. Any wavelet coefficient $W_{j,k}$ at scale j and location k can be written as a linear combination of input data:

$$W_{j,k} = \sum_{i \in \mathcal{I}_{j,k}} \gamma_{j,k,i} X_{j,k,i},$$

where $\gamma_{j,k,i} \neq 0$ for $i \in \mathcal{I}_{j,k}$ is the set of nonzero entries in $\widetilde{W}$ on the row of $W_{j,k}$, and by $X_{j,k,i}$ we denote the corresponding entries of $\boldsymbol{X}$. We let $r_{j,k} = \#\mathcal{I}_{j,k}$ be the number of these non-zero entries. In most applications, and certainly in the case of classical wavelet transforms on equispaced data, $r_{j,k}$ only depends on scale j, not on location k within that scale. Note that the double (j,k) indexed wavelet coefficients can in fact be stored in single vector of length n. These wavelet coefficients $W_{j,k}$ are clearly heteroscedastic. We look for a normalisation factor, so that the variances of the normalised coefficients are approximately constant and independent of the input intensities. At least we want that the coefficients with zero mean, i.e., the non-important ones, have constant variances, in order to remove them with a single threshold.

We introduce a *normalisation factor* $N_{j,k}$:

$$N_{j,k} = \sum_{i \in \mathcal{I}_{j,k}} X_{j,k,i}.$$

It is clear that $N_{j,k}$ is Poisson distributed with intensity $\lambda_{j,k}$, equal to:

$$\lambda_{j,k} = \sum_{i \in \mathcal{I}_{j,k}} \lambda_{j,k,i}.$$

We then define the *normalised* or *variance stabilised* wavelet coefficient $Z_{j,k}$ as:

$$Z_{j,k} = \begin{cases} \dfrac{W_{j,k}}{\sqrt{N_{j,k}}} = \dfrac{\sum_{i \in \mathcal{I}_{j,k}} \gamma_{j,k,i} X_{j,k,i}}{\sqrt{\sum_{i \in \mathcal{I}_{j,k}} X_{j,k,i}}} & \text{if } N_{j,k} \neq 0 \\ 0 & \text{if } N_{j,k} = 0 \end{cases}$$

When applied to a Haar transform, this variance stabilisation reduces to the Multiscale Fisz decomposition, which is discussed further below in this section. The remainder of this section concentrates on a single wavelet coefficient, and therefore we can omit the subscripts j and k for notational convenience, so we write W, N, Z, λ to indicate a wavelet coefficient's value, its normalisation factor, its normalised value and the intensity of its normalisation factor. The length of the set of non-zero entries in the wavelet transform matrix is denoted by r and we renumber these entries such that $\mathcal{I} = \{1, \ldots, r\}$.

We then define for each wavelet coefficient the *relative intensities* ρ_i of input X_i as:

$$\rho_i = \lambda_i / \lambda$$

2.2 Variance and expectation of the normalised coefficients

The joint distribution of $X_i, i = 1, \ldots, r$ conditioned on N is multinomial. Using the expressions for mean and covariance of a multinomial vector, we can write for expected value and variance of Z given N:

$$\begin{aligned} E(Z|N=n) &= \sqrt{n} \sum_{i=1}^{r} \gamma_i \rho_i & (1) \\ V(Z|N=n) &= \sum_{i=1}^{r} \gamma_i^2 \rho_i (1-\rho_i) - 2 \sum_{i=1}^{r} \sum_{j=1}^{i-1} \gamma_i \gamma_j \rho_i \rho_j \\ &= \sum_{i=1}^{r} \gamma_i^2 \rho_i - \left(\sum_{i=1}^{r} \gamma_i \rho_i \right)^2 \quad \text{for } n \neq 0. & (2) \end{aligned}$$

Note that $V(Z|N = n)$ does not depend on n and both expressions only depend on the *relative* intensities of X_i, not on the absolute intensities.

From this, we can easily average over all possible N:

$$\begin{aligned} V(Z|N \neq 0) &= E(V(Z|N)|N \neq 0) + V(E(Z|N)|N \neq 0) \\ &= \sum_{i=1}^{r} \gamma_i^2 \rho_i - \left(\sum_{i=1}^{r} \gamma_i \rho_i\right)^2 + V\left(\sqrt{N}|N \neq 0\right) \left(\sum_{i=1}^{r} \gamma_i \rho_i\right)^2 . \end{aligned} \tag{3}$$

The only factor depending on the absolute intensities λ_i is $V(\sqrt{N}|N \neq 0)$. N is Poisson distributed with intensity $\lambda = \sum_{i=1}^{r} \lambda_i$. It is well known that for $\lambda \to \infty$, this variance of the square root of a Poisson count converges quickly to $1/4$. This means that $V(Z|N \neq 0)$ is nearly independent from the absolute intensities λ_i. This is an interesting observation, since it motivates the application of a threshold procedure on the normalised coefficients Z. This near-homoscedasticity is further explored in Section 3.

For the expected value, we have

$$E(Z|N \neq 0) = E\left(\sqrt{N}|N \neq 0\right) \left(\sum_{i=1}^{r} \gamma_i \rho_i\right), \tag{4}$$

so the expected squared value is

$$\begin{aligned} E(Z^2|N \neq 0) &= (E(Z|N \neq 0)^2 + V(Z|N \neq 0) \\ &= \sum_{i=1}^{r} \gamma_i^2 \rho_i - \left(\sum_{i=1}^{r} \gamma_i \rho_i\right)^2 \\ &\quad + \left(E\left(\sqrt{N}|N \neq 0\right)^2 + V\left(\sqrt{N}|N \neq 0\right)\right) \left(\sum_{i=1}^{r} \gamma_i \rho_i\right)^2 \\ &= \sum_{i=1}^{r} \gamma_i^2 \rho_i - \left(\sum_{i=1}^{r} \gamma_i \rho_i\right)^2 + E(N|N \neq 0) \left(\sum_{i=1}^{r} \gamma_i \rho_i\right)^2 \\ &= \sum_{i=1}^{r} \gamma_i^2 \rho_i - \left(\sum_{i=1}^{r} \gamma_i \rho_i\right)^2 + \frac{\lambda}{1 - e^{-\lambda}} \left(\sum_{i=1}^{r} \gamma_i \rho_i\right)^2 \end{aligned} \tag{5}$$

2.3 Thresholding and reconstruction

A threshold algorithm based on conditional variance stabilisation (CVS) proceeds as follows:

1. Apply a wavelet transform to input $\boldsymbol{y}$. The transform can be the standard fast decomposition, but also a redundant, shift-invariant representation. Call $\boldsymbol{w}$ the coefficients in this decomposition.

2. For all coefficients, compute the normalisation factors $\sqrt{\boldsymbol{N}}$. The computation is basically bookkeeping of the support of the filter operations in each step of the multiscale wavelet decomposition. This bookkeeping is as fast as the actual transform, and if we take into account that it actually requires no additional *floating point* operations, it is even faster. It anyhow induces no bottleneck in the smoothing algorithm.

3. Find a vector of appropriate level dependent thresholds $\boldsymbol{\theta}$ for the normalised coefficients $\boldsymbol{z} = \boldsymbol{w}/\sqrt{\boldsymbol{N}}$ (coordinatewise division). The threshold has to be level dependent, because, in general, the variance is not constant across scales. The Haar transform is an exception, provided that the coefficients are normalised to 1 and -1. Then the variance $V(Z|N \neq 0) = 1$, regardless of the resolution level of the normalised coefficient Z. We propose a Bayesian shrinkage and threshold procedure in Section 4. Apply this threshold (using soft, hard or any intermediate threshold approach) and call $\boldsymbol{z_\theta}$ the thresholded normalised coefficients.

4. Re-multiply the thresholded normalised coefficients to obtain estimates for the noise-free wavelet coefficients: $\widehat{\boldsymbol{v}} = \boldsymbol{z_\theta} \cdot \sqrt{\boldsymbol{N}}$ (coordinatewise multiplication).

5. Inverse wavelet transform yields an estimate of the Poisson intensities.

2.4 The Multiscale Fisz transform

As mentioned before, the proposed variance stabilisation procedure applied to a Haar transform, coincides with the Multiscale Fisz decomposition. On the other hand, so does the Fisz-Wavelet decomposition (Fryźlewicz and Nason 2003). Reconstruction from smoothed (shrunk and/or thresholded) coefficients however yields a slightly different output. Indeed, if the wavelet transform used in our approach is a Haar-transform with scaling filters normalised as $[1,1]$, then the scaling coefficients coincide with the normalisation factors $N_{j,k}$. There is no need to compute this normalisation separately: the Multiscale Fisz decomposition relies on the scaling coefficients. Scaling coefficients however are further decomposed into wavelet coefficients and scaling coefficients at coarser levels. As a consequence, the values of these scaling coefficients are modified by the thresholds at coarser scales. This means that in the Multiscale Fisz approach, the normalisation factor upon decomposition is different from the multiplicative factor upon reconstruction. Simulations suggest however that all this has little impact on the overall output.

2.5 An unbiased variance estimator: diagonal covariance stabilisation

Multiscale Fisz can also be seen as a special case of a procedure based on the diagonal of an unbiased covariance estimator (Jansen 2001, page 136). We first observe that the input values X_i are unbiased estimates of the variances:

$$V(X_i) = \lambda_i = E(X_i).$$

If the input is a vector of indecent count data, the diagonal matrix with these values is an unbiased estimate $\widehat{C}_{\boldsymbol{X}}$ of the covariance matrix $C_{\boldsymbol{X}}$. The covariance matrix $C_{\boldsymbol{W}}$ of the wavelet coefficients is then

$$C_{\boldsymbol{W}} = \widetilde{W} C_{\boldsymbol{X}} \widetilde{W}^T,$$

and this can be estimated in an unbiased way as:

$$\widehat{C}_{\boldsymbol{W}} = \widetilde{W} \widehat{C}_{\boldsymbol{X}} \widetilde{W}^T.$$

The diagonal of this matrix has unbiased estimates of the wavelet coefficient variances. The computation of $\widehat{C}_{\boldsymbol{W}}$ is in principle a 2-dimensional, rectangular (i.e., fully tensor product) wavelet transform, which requires $\mathcal{O}(n^2)$ computations (n is the length of the input data). A 2-dimensional square wavelet transform (the version mostly used in image processing) however generates the same diagonal elements and is much cheaper. Because of the sparse structure of the input covariance, one can even further economise the computations and end up with a linear complexity. In the case of classical wavelet transforms on equispaced grids, a fast computation follows from realising that

$$(C_{\boldsymbol{W}})_{ll} = \sum_{i \in \mathcal{I}_l} \gamma_{l,i}^2 X_{l,i},$$

where $\gamma_{l,i}$ are the nonzero entries of the forward wavelet transform matrix on the row corresponding to the $l$the coefficient. (Index l corresponds to scale j and location k). The entries $\gamma_{l,i}$ can be computed quickly as pointed out in Section 3.

The wavelet coefficients W_l are then normalised with the square root of these variance estimates:

$$Z_l^{\mathrm{DC}} = W_l / \sqrt{(C_{\boldsymbol{W}})_{ll}}.$$

This normalisation is related to the normalisation by conditioning on N_l through

$$Z_l^{\mathrm{DC}} = Z_l / \sqrt{\sum_{i \in \mathcal{I}_l} \gamma_{l,i}^2 \frac{X_{l,i}}{N_l}}.$$

This unbiased coefficient variance estimation can also be used in a wavelet domain filtering approach (Antoniadis and Sapatinas 2001; Antoniadis et al. 2001; Nowak and Baraniuk 1999).

A similar approach can be applied to instances from other members of the exponential family (Antoniadis and Sapatinas 2001; Morris 1982), provided that we have an unbiased (or other 'good') variance estimator of the input. For instance, if the input has an exponential density, i.e., $X_i \sim exp(\lambda_i)$, we know that $V(X_i) = 1/\lambda_i^2 = \mu_{X_i}^2$, so $E(X_i^2) = E(X_i)^2 + V(X_i) = 2/\lambda_i^2$. We find that $E(X_i^2/2) = V(X_i)$. Another example is the χ^2 density, where $E(2X_i) = 2k = V(X_i)$, with k the number of degrees of freedom. The normalisation for χ^2 is thus essentially the same as for Poisson.

When applied to a Haar transform, this again coincides with the Multiscale Fisz decomposition.

The resulting normalised wavelet coefficients do not have entirely constant variances, even if the noise-free values are all zero.

3 Discussion on the variance stabilisation

3.1 Fast computation of the forward wavelet transform matrix

The expressions for expectation and variance of (normalised) coefficients use the values of the entries γ_i of the forward wavelet transform matrix. These entries could be found by applying a forward wavelet transform to the identity matrix, but a fast and simple trick (at least in the case of equispaced data) is to apply an *inverse* wavelet transform with the *forward* transform filters (i.e., the dual, analysis, or decomposition filters). This is a subdivision algorithm. It follows that the values of γ_i converge to the dual wavelet functions.

3.2 Non-equispaced data

The conditional variance stabilisation approach is immediately extendible to second generation wavelets (Sweldens 1997) for observations on non-equispaced point sets. The computation of the forward wavelet transform matrix is now a bit more complicated, since the matrix rows within one resolution are no longer shifts of each other. Careful implementation with bandlimited matrices however leads to an algorithm with linear complexity.

3.3 Near-homoscedasticity

Expressions (2) and (3) show that the variance of the normalised coefficients only depends on the *relative* intensities. Homoscedasticity is an important property for a threshold procedure to work properly. In particular, it should hold for coefficients that carry no information, i.e., coefficients with zero expected value. If the expected coefficient is zero, we have that $\sum_{i=1}^{r} \gamma_i \rho_i = 0$ and hence $V(Z|N \neq 0) = \sum_{i=1}^{r} \gamma_i^2 \rho_i$. This equals $(1/r)\|\gamma\|_2^2$ if the intensities λ_i are constant. In that case, the noise variance is completely independent of the intensity. If however the intensity curve is a polynomial with degree smaller than the number of vanishing moments of the analysis wavelet, the expected coefficient is still zero, while its variance slightly depends on the *relative* intensity curve. In order to eliminate this dependency, one should further normalise the coefficients by:

$$Z^{\text{new}} = Z / \sqrt{\sum_{i=1}^{r} \gamma_i^2 \rho_i}.$$

Since the exact relative intensities ρ_i are unknown, they could be replaced by their observed counterparts X_i/N. This results in the diagonal covariance stabilisation.

3.4 Variance of normalised wavelet coefficients

If one accepts the value of $V(Z|N \neq 0)$ for constant input intensities as an approximation for the variance of all non-significant coefficients (i.e., coefficients with expected value equal to zero), then this value can replace an estimation of that variance. This estimation is often a first step in a threshold assessment procedure, explicitly, as in SURE-thresholding (Donoho and Johnstone 1995), where a Median Absolute Deviation (MAD) estimate is used, or implicitly, for instance when using cross validation. In our case, such an estimation is not needed, since the (approximating) value of $V(Z|N \neq 0)$ can be computed exactly and quickly, using the entries of the forward wavelet transform matrix. This is most interesting at coarse scales, where MAD is not sufficiently robust. An overestimation of the variance is particularly undesirable in an empirical Bayes approach, as we discuss below, since it not only affects the empirical estimation of the hyperparameters but also makes the final classification of coefficients more difficult.

3.5 Sensitivity and selectivity

A coefficient selection procedure based on thresholding requires that data carrying coefficients have significantly higher expected squared values than coefficients with only noise.

From the expression (5) for the expected squared value, and since trivially $E(N|N \neq 0) \geq 1$, we have that

$$E(Z^2|N \neq 0) \geq \sum_{i=1}^{r} \gamma_i^2 \rho_i.$$

Similarly, $V\left(\sqrt{N}|N \neq 0\right) \leq 1$, so from the expression (3) for the variance, we deduce that

$$V(Z|N \neq 0) \leq \sum_{i=1}^{r} \gamma_i^2 \rho_i.$$

Both inequalities become equalities if $\sum_{i=1}^{r} \gamma_i \rho_i = 0$.

In the case of a Haar transform, all γ_i^2 at a given scale are equal. They all equal one, independently of the scale, if the filters are appropriately normalised, i.e., if the analysis filters are set to $(1, 1)$ and $(1, -1)$. Indeed, this case leads to

$$\sum_{i=1}^{r} \gamma_i^2 \rho_i = \sum_{i=1}^{r} \rho_i = 1.$$

The right hand sides of the inequalities above are then absolute bounds, they do not depend on $\boldsymbol{\rho}$ anymore. As a consequence, the noise variance $V(Z|N \neq 0)$ reaches an absolute maximum where the expected value of the coefficient equals zero and at the same moment, the coefficient's expected squared value reaches a minimum. In other words: the smallest coefficients carry the highest amount of noise. This is even more favourable than the classical wavelet thresholding assumption that the noise is spread out *evenly* over all coefficients.

Unfortunately, the same property no longer holds for general wavelet transforms. It may happen that $\sum_{i=1}^{r} \gamma_i^2 \rho_i$ is larger for a coefficient with non-zero mean than for a coefficient which is only noise.

Once again, a further normalisation with factor $\sqrt{\sum_{i=1}^{r} \gamma_i^2 \rho_i}$ could remedy this problem, if only the relative intensities ρ_i were known. Replacing them with observed values again results in a diagonal covariance normalisation.

4 A Bayesian threshold scheme

The above discussion on near-homoscedasticity and selectivity illustrates that a more subtle and adaptive coefficient selection could be interesting. Also, the exact distribution of the stabilised coefficient still depends on the absolute intensities. A Bayesian procedure can deal with this complete distribution in a natural way. The variance stabilisation of Section 2 can be seen as just an instance of the main underlying idea of this paper, which is the conditioning on $N_{j,k}$, the total number of counts that participate in the expression of a wavelet coefficient $W_{j,k}$ at scale j and location k. This section constructs a Bayesian model within this framework of given $N_{j,k}$. This approach can then be seen as a generalisation of the multiscale likelihood method (Kolaczyk and Nowak 2000) to wavelet families beyond Haar wavelets.

4.1 Prior model for relative intensities

As before, we concentrate on a single wavelet coefficient, which we denote by W, thereby omitting the indices j for scale and k for location, or l for its position in the wavelet transform matrix.

$$W = \sum_{i=1}^{r} \gamma_i X_i$$

We call $N = \sum_{i=1}^{r} X_i$ and

$$Z = \frac{W}{\sqrt{N}} = \frac{1}{\sqrt{N}} \cdot \sum_{i=1}^{r} \gamma_i X_i.$$

By ζ we denote the expected value of Z, conditioned on N. Restating expression (1), this becomes:

$$\zeta = E(Z|\boldsymbol{\rho}, N) = \sqrt{N} \cdot \sum_{i=1}^{r} \gamma_i \rho_i.$$

We make the dependence on the vector of relative intensities $\boldsymbol{\rho}$ explicit, since this vector is now modelled as a random variable.

Inspired by similar approaches in wavelet denoising (Kolaczyk 1999; Johnstone and Silverman 2004; Chipman et al. 1997; Abramovich et al. 1998) the proposed prior model for the noise-free value ζ is a mixture of a point mass at zero and a continuous density away from zero. We call p the prior probability of a coefficient having a non-zero noise-free value:

$$p = P(\zeta \neq 0|N). \tag{6}$$

This value p is a model parameter. An empirical method for choosing it, is presented after we calculate the marginal probabilities for the observed coefficients. We set $q = 1 - p$. It reasonable to assume that the event $\{\zeta \neq 0\}$ is independent of N.

For ζ-values away from zero, we assume that the relative intensities come from a *Dirichlet* distribution with parameter vector $\boldsymbol{a}$:

$$f_{\boldsymbol{\rho}}(\boldsymbol{\rho}|\zeta \neq 0) = \frac{\Gamma(A)}{\prod_{i=1}^{r} \Gamma(a_i)} \prod_{i=1}^{r} \rho_i^{a_i - 1}, \tag{7}$$

where

$$A = \sum_{i=1}^{r} a_i. \tag{8}$$

If we call $\alpha_i = a_i/A$, we have:

$$\begin{aligned}
E(\rho_i|\zeta \neq 0) &= \alpha_i = \frac{a_i}{A} \\
V(\rho_i|\zeta \neq 0) &= \frac{a_i(A - a_i)}{A^2(A+1)} \\
\mathrm{cov}(\rho_i, \rho_j|\zeta \neq 0) &= \frac{-a_i a_j}{A^2(A+1)} \\
\mathrm{corr}(\rho_i, \rho_j|\zeta \neq 0) &= -\sqrt{\frac{a_i a_j}{(A - a_i)(A - a_j)}}
\end{aligned}$$

For symmetry, there is no reason to assume that the prior on ρ_i is different from the prior on ρ_j, so we can take all parameters a_i equal to a single a. This simplifies the above expressions. In particular, the prior correlation becomes independent of a.

It has been shown (Karlin et al. 1986; Ignatov and Kaishev 1989; Dahmen and Micchelli 1986) that a linear combination $\zeta = \sqrt{N}\sum_{i=1}^{r}\gamma_i\rho_i$ of a Dirichlet vector has a B-spline density. The knots are in γ_i and have multiplicity a_i. If not all a_i are integer, the density becomes a so called generalised B-spline (Kaishev 1991). Both for B-splines and generalised B-splines, there exist cubature formulae to find integrals, mean values, etc., but we will use a simple normal approximation, based on the following results for mean and variance:

$$\begin{aligned}
E(\zeta|\zeta \neq 0, N) &= \sqrt{N}\cdot\sum_{i=1}^{r}\gamma_i\alpha_i && (9)\\
&= 0 \qquad \text{if all } \alpha_i = 1/r \text{ are equal.} \\
V(\zeta|\zeta \neq 0, N) &= \frac{N}{A+1}\cdot\left[\sum_{i=1}^{r}\gamma_i^2\alpha_i - \left(\sum_{i=1}^{r}\gamma_i\alpha_i\right)^2\right] && (10)\\
&= \frac{N}{ra+1}\cdot\frac{1}{r}\cdot\sum_{i=1}^{r}\gamma_i^2 \qquad \text{if all } \alpha_i \text{ are equal.}
\end{aligned}$$

4.2 Posterior distributions

The conditional probability of $\boldsymbol{X}$, given $\boldsymbol{\rho}$ and N is multinomial. The conditional expectation $E(Z|\boldsymbol{\rho}, N)$ and variance $(Z|\boldsymbol{\rho}, N)$ are given by expressions (1) and (2). The expressions did not explicitly state the conditioning on $\boldsymbol{\rho}$, as they were presented in a non-Bayesian context.

The Dirichlet distribution is a conjugate prior for the multinomial distribution (Robert 2001). This means that the posterior density

$$f_{\boldsymbol{\rho}|\boldsymbol{X}}(\boldsymbol{\rho}|\boldsymbol{x}, \zeta \neq 0)$$

is again a Dirichlet distribution. The posterior parameter vector becomes $\boldsymbol{a} + \boldsymbol{x}$.

As a consequence, the posterior density $f_{\zeta|\boldsymbol{X}}(\zeta|\boldsymbol{x}, \zeta \neq 0)$ of non-zero ζ is again a (generalised) B-spline function with knots in γ_i and multiplicity $a_i + x_i$. Since the observations $\boldsymbol{x}$ are always integers, this posterior density is still a classical (i.e., not generalised) B-spline if the prior density $f_\zeta(\zeta|\zeta \neq 0)$ is a classical spline.

The posterior distribution of the noise-free value ζ can be written as:

$$\begin{aligned}
F_{\zeta|\boldsymbol{X}}(\zeta|\boldsymbol{x}) &= P(\zeta \neq 0|\boldsymbol{X} = \boldsymbol{x})\cdot F_{\zeta|\boldsymbol{X}}(\zeta|\boldsymbol{x}, \zeta \neq 0)\\
&\quad + P(\zeta = 0|\boldsymbol{X} = \boldsymbol{x})\cdot I_{\mathbb{R}^+}(\zeta), && (11)
\end{aligned}$$

with $I_{\mathbb{R}^+}(x)$ the indicator function on the positive real numbers (including 0).

The posterior probability p^* of a noise-free coefficient ζ being non-zero follows from:

$$\begin{aligned}
p^* &= P(\zeta \neq 0|\boldsymbol{X} = \boldsymbol{x})\\
&= \frac{P(\zeta \neq 0|N)P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta \neq 0, N)}{P(\zeta \neq 0|N)P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta \neq 0, N) + P(\zeta = 0|N)P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta = 0, N)}\\
&= \frac{p}{p + q\cdot\dfrac{P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta = 0, N)}{P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta \neq 0, N)}}. && (12)
\end{aligned}$$

This expression uses the values of the marginal probabilities $P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta = 0, N)$ and $P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta \neq 0, N)$. The marginal under $\zeta \neq 0$ is fixed by the prior $f_{\boldsymbol{\rho}}(\boldsymbol{\rho}|\zeta \neq 0)$ and the multinomial conditional $P_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\rho}, \zeta \neq 0, N) = P_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\rho}, N)$. The same conditional applies under $\zeta = 0$, but the prior under $\zeta = 0$ has not been specified yet. We now use this freedom to compute the posterior p^* based on the single wavelet coefficient Z instead of the whole set of observed $\boldsymbol{X}$:

$$\begin{aligned}
p^* &= P(\zeta \neq 0|Z = z, N)\\
&= \frac{p}{p + q\cdot\dfrac{P_Z(z|\zeta = 0, N)}{P_Z(z|\zeta \neq 0, N)}}. && (13)
\end{aligned}$$

We want the Bayes factor (marginal likelihood ratio) for the observed $\boldsymbol{x}$ in expression (12) such that it depends on $z = \frac{1}{\sqrt{N}} \sum_{i=1}^{r} \gamma_i x_i$ only. Indeed, it is more convenient and intuitive to construct a model for $P_Z(z|\zeta = 0, N)$ than for $P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta = 0, N)$. The details of this model are filled in later, when we discuss marginal probabilities. As for now we suppose that the model is fully specified, and we state that for all K possible configurations of $\boldsymbol{x}$, where

$$K = \binom{N + r - 1}{r},$$

and for $z = \frac{1}{\sqrt{N}} \sum_{i=1}^{r} \gamma_i x_i$,

$$\frac{P_Z(z|\zeta = 0, N)}{P_Z(z|\zeta \neq 0, N)} = \frac{P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta = 0, N)}{P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta \neq 0, N)} \tag{14}$$

This leads to K conditions on $f_{\boldsymbol{\rho}}(\boldsymbol{\rho}|\zeta = 0)$:

$$\int_{\boldsymbol{\rho}} f_{\boldsymbol{\rho}}(\boldsymbol{\rho}|\zeta = 0) \cdot P_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\rho}, N)\, d\boldsymbol{\rho} = \frac{P_Z(z|\zeta = 0, N)}{P_Z(z|\zeta \neq 0, N)} \cdot P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta \neq 0, N)$$

In order to find a prior that satisfies these conditions, one could for instance write

$$f_{\boldsymbol{\rho}}(\boldsymbol{\rho}|\zeta = 0) = \sum_{k=1}^{K} c_k f_k(\boldsymbol{\rho}),$$

for some basis functions $f_k(\boldsymbol{\rho})$ and then solve a set of K linear equations in the coefficients c_k.

As mentioned before, modelling the wavelet coefficient, given that its noise-free version is zero, is more intuitive than a model for the corresponding input data. In particular, it is easy to construct a continuous, normal, approximation for the probability function of the discrete variable Z.

4.3 Posterior mean, variance and median

It is interesting to have a closer look at the posterior mean and variance. For the posterior mean, we have:

$$E(\zeta|\boldsymbol{X}, \zeta \neq 0) = \sqrt{N} \sum_{i=1}^{r} \gamma_i \frac{a_i + X_i}{A + N}.$$

And if all a_i are equal, the first vanishing moment of the dual (analysis) wavelet reduces this to:

$$E(\zeta|\boldsymbol{X}, \zeta \neq 0) = \frac{\sqrt{N}}{N + A} \sum_{i=1}^{r} \gamma_i X_i = \frac{N}{N + A} \cdot Z.$$

Since right hand side only depends on Z, we can equally write:

$$E(\zeta|Z, \zeta \neq 0, N) = E(\zeta|\boldsymbol{X}, \zeta \neq 0) = \frac{N}{N + A} \cdot Z, \tag{15}$$

and, consequently,

$$E(\zeta|Z, N) = P(\zeta \neq 0|Z, N) \cdot \frac{N}{N + A} \cdot Z. \tag{16}$$

The posterior variance is a bit more complicated. It has the form of expression (10), with $\alpha_i = (a_i + X_i)/(A + N)$. The first vanishing moment of the dual wavelet is no longer sufficient to eliminate all dependence on individual X_i's. As a consequence, the posterior variance $V(\zeta|\boldsymbol{X}, \zeta \neq 0)$ is not (necessarily) equal to $V(\zeta|Z, \zeta \neq 0, N)$.

From posterior mean and variance, and using a normal approximation for the posterior density $f_{\zeta|\boldsymbol{X}}(\zeta|\boldsymbol{x}, \zeta \neq 0)$ (which is spline or generalised spline function), it is possible to derive the *posterior median*. Since for small observed coefficients z, the posterior mixture p^* is much smaller than $1/2$, this posterior median must be exactly zero. A posterior median therefore leads to threshold scheme (Abramovich et al. 1998). This threshold erases small posterior means and therefore leads to a smoother reconstruction.

4.4 Marginal probabilities

The marginal probability functions of $\boldsymbol{X}$ and Z already appeared in expressions (12) and (13 and is also interesting in estimating the model's parameters in an empirical Bayes approach.

It is easy to verify (for instance by marginal = prior · conditional / posterior) that the marginal under $\zeta \neq 0$ equals:

$$P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta \neq 0, N) = \frac{\Gamma(N+1) \cdot \Gamma(A) \cdot \prod_{i=1}^{r} \Gamma(a_i + x_i)}{\left[\prod_{i=1}^{r} \Gamma(x_i+1)\Gamma(a_i)\right] \Gamma(A+N)}. \tag{17}$$

If all $a_i = 1$, this, somehow remarkably, reduces to a uniform distribution on all configurations $\boldsymbol{x}$ (note that the conditional distribution is multinomial):

$$P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta \neq 0, N) = \frac{N!(r-1)!}{(N+r-1)!}.$$

The subsequent expressions for marginal mean and variance include this special case of a uniform distribution on the discrete simplex $\{\boldsymbol{x}\}$. We are interested in the expectation and variance of a linear combination of the elements of this configuration. The rules of conditional expectation lead to:

$$E(X_i|\zeta \neq 0, N) = E(E(X_i|\rho_i, \zeta \neq 0, N)) = E(N\rho_i|\zeta \neq 0, N) = N\alpha_i,$$

so the marginal expectation of Z equals:

$$E(Z|\zeta \neq 0, N) = N \cdot \sum_{i=1}^{r} \gamma_i \alpha_i = 0 \qquad \text{if } \alpha_i = \frac{1}{r}.$$

For the variance and covariance, we have:

$$\begin{aligned}
V(X_i|\zeta \neq 0, N) &= V(E(X_i|\rho_i, \zeta \neq 0, N)) + E(V(X_i|\rho_i, \zeta \neq 0, N)) \\
&= V(N\rho_i|\zeta \neq 0, N) + E\left(N\rho_i(1-\rho_i)|\zeta \neq 0, N\right) \\
&= N(N+A) \cdot V(\rho_i|\zeta \neq 0, N) \\
\operatorname{cov}(X_i, X_j|\zeta \neq 0, N) &= \operatorname{cov}\Big(E(X_i|\boldsymbol{\rho}, \zeta \neq 0, N), E(X_j|\boldsymbol{\rho}, \zeta \neq 0, N)\Big) \\
&\quad + E\Big(\operatorname{cov}(X_i, X_j|\boldsymbol{\rho}, \zeta \neq 0, N)\Big) \\
&= E(-N\rho_i\rho_j|\zeta \neq 0, N) + \operatorname{cov}(N\rho_i, N\rho_j|\zeta \neq 0, N) \\
&= N(N+A) \cdot \operatorname{cov}(\rho_i, \rho_j|\zeta \neq 0, N)
\end{aligned}$$

This allows to conclude that

$$V(Z|\zeta \neq 0, N) = \frac{N+A}{N} \cdot V(\zeta|\zeta \neq 0, N). \tag{18}$$

Note that the Bayesian shrinkage factor from (15) for non-zero coefficients, $N/(N+A)$, equals the ratio of the prior and marginal variances, just as in the case of a sum of two normals.

If all $\alpha_i = 1/r$, the marginal variance of the significant (i.e., with non-zero noise-free value) coefficients becomes

$$V(Z|\zeta \neq 0, N) = \frac{N+A}{1+A} \cdot \frac{1}{r} \sum_{i=1}^{r} \gamma_i^2.$$

The marginal variance of the coefficients with zero noise-free value is not uniquely fixed by our model, since it depends on the prior density of the relative intensities under $\zeta = 0$. From expression (14), we see that we actually still have the freedom to *impose* a Bayes factor in z. We assume that

$$V(Z|\zeta = 0, N) \approx V(Z|\rho_i = 1/r, \forall i, N) = \frac{1}{r}\sum_{i=1}^{r} \gamma_i^2,$$

so we can write

$$V(Z|\zeta \neq 0, N) \approx \frac{N+A}{1+A} \cdot V(Z|\zeta = 0, N).$$

We now *impose* that the Bayes factor equals the ratio:

$$\frac{P_Z(z|\zeta = 0, N)}{P_Z(z|\zeta \neq 0, N)} = \frac{\phi_{\sigma_0}(z)}{\phi_{\sigma_1}(z)},$$

where ϕ_σ stands for the normal density function with zero mean and standard deviation σ and

$$\sigma_1 = \sqrt{\frac{N+A}{1+A}}\sigma_0.$$

4.4.1 Thresholds and bounded shrinkage

Given the expressions for the Bayes factor, we have all the elements for the posterior probability p^*. We can compute the threshold value θ_{BF} such that if $|z| > \theta_{\mathrm{BF}}$, we have $p^* > 1/2$. This threshold equals:

$$\theta_{\mathrm{BF}} = \sqrt{2\,\frac{N+A}{N-1}\,\log\left(\frac{q}{p}\sqrt{\frac{N+A}{1+A}}\right)} \cdot \sigma_0. \tag{19}$$

The threshold induced by a posterior median can be found by solving

$$F_{\zeta|\boldsymbol{X}}(0|\boldsymbol{x}) = 1/2.$$

Since the complete posterior distribution in expression (11) depends on all the observations $\boldsymbol{x}$, and not just on the coefficient z, the threshold is not a constant for a given coefficient, but in any case, it is only slightly larger than the Bayes factor threshold θ_{BF}. If the sum of the observations, N, grows larger, so does the threshold, although at a quite slow rate. There is no upper bound for the threshold value, which means that there is no upper bound for the difference between input value and its shrinkage.

This unbounded shrinkage is a consequence of unbounded thresholds. It is more interesting to investigate whether shrinkage is bounded for finite threshold values. In particular, a Gaussian prior with Gaussian noise would lead to undesirable unbounded shrinkage for large input coefficients. In our case, we have for the posterior mean:

$$Z - E(\zeta|Z, N) = Z - p^* \cdot \frac{N}{N+A} \cdot Z = \frac{A}{N+A} \cdot Z + (1 - p^*) \cdot \frac{N}{N+A} \cdot Z. \tag{20}$$

For large coefficients, the factor $1 - p^* = P(\zeta = 0|Z, N)$ is close to 0 and the shrinkage is approximately proportional to the input value of Z. This input value is however bounded by the normalisation factor N:

$$|Z| \leq ||\gamma||_\infty \sqrt{N}.$$

As a consequence, the first term in (20) is bounded by

$$\frac{A}{N+A} \cdot Z \leq \frac{||\gamma||_\infty \sqrt{N}}{N+A} \leq \frac{||\gamma||_\infty \sqrt{A}}{2}.$$

The second term in (20) is arbitrarily small for sufficiently large values of z. The condition that $(1-p^*)\cdot z < \epsilon$ leads to

$$\left(\frac{z}{\epsilon}-1\right)\frac{q\,\sigma_1}{p\,\sigma_0} < \exp\left(z^2\cdot\frac{\sigma_1^2-\sigma_0^2}{2\sigma_1^2\sigma_0^2}\right).$$

This is satisfied if

$$\frac{z}{\epsilon}\frac{q\,\sigma_1}{p\,\sigma_0} < \frac{\exp(z)}{\epsilon}\frac{q\,\sigma_1}{p\,\sigma_0} < \exp\left(z^2\cdot\frac{\sigma_1^2-\sigma_0^2}{2\sigma_1^2\sigma_0^2}\right).$$

Solving the last inequality reduces to a quadratic form in z. We find that $(1-p^*)\cdot z < \epsilon$ if

$$z \geq \frac{N+A}{N-1}\cdot\sigma_0^2 + \sqrt{\left(\frac{N+A}{N-1}\right)^2\sigma_0^4 + 2\frac{N+A}{N-1}\cdot\sigma_0^2\cdot\log\left(\frac{q}{p}\sqrt{\frac{N+A}{1+A}}\right) + \log\frac{1}{\epsilon}},$$

This expression shows the same asymptotic behaviour as the Bayes factor threshold (19), which is much slower than the maximal value of z for a given N.

All together, we have shown that for any value of N, sufficiently significant values of z have bounded shrinkage and there exist an upperbound independent of N. Actually, the concept of conditional variance stabilisation turns a situation without bounded shrinkage (normal prior with normal noise) into a more favourable situation with bounded shrinkage.

4.4.2 Empirical Bayes

The expressions for marginal variances also allow for the computation of the marginal likelihood of parameter p, the probability for a coefficient being significant. We assume that this parameter is scale dependent, and denote the value at scale j by p_j. If we call $\sigma_0^2 = V(Z|\zeta \neq 0, N_{j,k})$ and $\sigma_{1,j,k}^2 = V(Z|\zeta = 0, N_{j,k})$, we can approximate the likelihood in p_j for an observed vector of normalised coefficients $\boldsymbol{z_j}$ as:

$$\log L(p_j) = \sum_{k=1}^{2^j}\log\left(p_j\cdot\phi_{\sigma_{1,j,k}}(z_{j,k}) + (1-p_j)\cdot\phi_{\sigma_0}(z_{j,k})\right),$$

where, again, ϕ_σ stands for the normal density function with zero mean and standard deviation σ. Note that $\sigma_{1,j,k}$ depends on $N_{j,k}$, so the likelihood expression is different in every observed coefficient.

The values for σ_0 are independent from the observed $N_{j,k}$, and could be estimated from the data using Median Absolute Deviation (MAD). As mentioned before, this works fine on fine scales, but MAD is not sufficiently robust on coarse scales. Therefore, we use the exact expression for σ_0 in terms of the wavelet transform coefficients γ_k.

At fine scales, p_j is generally quite small, and it might show difficult to capture the few significant coefficients at those fine scales, leading to $\widehat{p_j} = 0$. We therefore impose that the posterior median threshold should be below the universal threshold (Johnstone and Silverman 2004), i.e.,

$$P(\zeta > 0|Z = \lambda_{\text{univ}}, N) \geq 1/2.$$

This implies a condition on the posterior probability p^* for an observation equal to the universal threshold:

$$P(\zeta = 0|Z = \lambda_{\text{univ}}, N) \leq 1/2.$$

Since $\lambda_{\text{univ}} = \sqrt{2\log n}\sigma_0$ is a large value, the probability $P(\zeta < 0|Z = \lambda_{\text{univ}}, N)$ is small and both conditions are practically equal. Elaboration of the latter condition leads to

$$p > \frac{C}{C+n^D},$$

where

$$
\begin{aligned}
C &= \frac{\sigma_{1,j,k}}{\sigma_0} = \sqrt{\frac{N_{j,k}+A_{j,k}}{1+A_{j,k}}} \\
D &= 1-1/C^2 = \frac{N_{j,k}-1}{N_{j,k}+A_{j,k}}
\end{aligned}
$$

This minimum value for the prior p therefore also depends on the observed $N_{j,k}$ in each coefficient.

5 Discussion on the Bayesian approach

5.1 Estimation of the Dirichlet parameter vector

The relative intensities $\rho_{j,k,i}$ involved in the computation of a coefficient $\zeta_{j,k}$ at scale j and location k can be expressed as a combination of the relative intensities for the single coefficient at the coarsest scale $j = 0$:

$$\rho_{j,k,i} = \frac{\rho_{0,0,i}}{\sum_{i\in\mathcal{I}_{j,k}} \rho_{0,0,i}}. \tag{21}$$

This implies that the whole prior is fully specified by the model for the relative intensities at coarsest scale. In particular,

$$\boldsymbol{\rho_{0,0}} \sim \text{Dirichlet}(\boldsymbol{a}) \Rightarrow \boldsymbol{\rho_{j,k}} \sim \text{Dirichlet}(\boldsymbol{a}_{\mathcal{I}_{j,k}}).$$

This can be verified by constructing a vector of independent Gamma distributed variables $\boldsymbol{V} \sim \Gamma(\lambda, \boldsymbol{a})$ (for some λ), such that

$$\boldsymbol{\rho_{0,0}} \overset{\mathrm{d}}{=} \frac{\boldsymbol{V}}{\sum V_i} \text{ and } \boldsymbol{\rho_{j,k}} \overset{\mathrm{d}}{=} \frac{\boldsymbol{V}_{\mathcal{I}_{j,k}}}{\sum_{\mathcal{I}_{j,k}} V_i}.$$

The parameter vector $\boldsymbol{a}$ is therefore scale invariant. If all a_i are assumed to be equal, then this single parameter a can be estimated from the expression for the marginal variance of the observations X_i:

$$V(X_i|\zeta \neq 0, N) = N(N+A)V(\rho_i|\zeta \neq 0) = N(N+A)\frac{a(A-a)}{A^2(A+1)}.$$

At the coarsest scale, we can assume that indeed $\zeta \neq 0$. Using the sample variance from the input data, we can then construct the following estimator for a:

$$\hat{a} = \frac{N^2(n-1) - n^2\widehat{\sigma}^2}{n^3\widehat{\sigma}^2 - Nn(n-1)}.$$

In this expression n is the sample size, N is the observed sum of counts and $\widehat{\sigma}^2$ is the estimated marginal variance.

This parameter vector $\boldsymbol{a}$ is thus closely related to the variance of the underlying intensity function. If this function shows a clear heterogeneous behaviour, one could consider a non-constant vector $\boldsymbol{a}$ and estimate it from the variances at different subsets of the input sample.

5.2 Multiscale likelihood

The Dirichlet distribution is a multivariate extension of a Beta distribution. A Beta prior for relative intensities has already been proposed (Kolaczyk 1999) for the relative intensities,

$$\theta_{j,k} = \frac{\lambda_{j+1,2k}}{\lambda_{j+1,2k} + \lambda_{j+1,2k+1}}$$

of two neighbouring accumulated count data at scale j. These accumulated intensities can be computed as the scaling coefficients at successive scales of a forward (non-normalised) Haar transform of the input intensities: $\lambda_{j,k} = \lambda_{j+1,2k} + \lambda_{j+1,2k+1}$.

Since this is a (non-normalised) Haar transform, these scaling coefficients coincide with the sum of the input intensities (at finest scale) involved in the computation of the ratio $\theta_{j,k}$. This observation is quite similar to a key observation in the discussion on the multiscale Fisz decomposition, and it reveals an essential obstacle for extending the concept of multiscale likelihood methods to a general wavelet transforms (Kolaczyk and Nowak 2002, 2004).

This multiscale likelihood model can however be mapped into the framework presented in this paper, if only the hyperparameters are appropriately chosen. Indeed, we observed that, due to the specific properties of the Haar transform, the denominators in the expression of $\theta_{j,k}$ coincide with the denominators in the expression of $\rho_{j,k,i}$. More specifically, we have that

$$\theta_{j,k} = \sum_{i \in \mathcal{I}_{j+1,2k}} \rho_{j,k,i} = \frac{\sum_{i \in \mathcal{I}_{j+1,2k}} \rho_{0,0,i}}{\sum_{i \in \mathcal{I}_{jk}} \rho_{0,0,i}},$$

where $\mathcal{I}_{j+1,2k}$ coincides with the left half of $\mathcal{I}_{j+1,k}$, i.e., the left half of coefficients involved in the computation of $\zeta_{j,k}$. Note that, again because of the structure of a Haar transform, we have that

$$\mathcal{I}_{j+1,k} = \mathcal{I}_{j+1,2k} \cup \mathcal{I}_{j+1,2k+1}.$$

If $\boldsymbol{\rho_{j,k}} \sim \mathrm{Dirichlet}(\boldsymbol{a}_{\mathcal{I}_{j,k}})$, then the marginal prior for $\theta_{j,k}$ must be

$$\theta_{j,k} \sim \mathrm{Beta}\left(\sum_{i \in \mathcal{I}_{j+1,2k}} a_i, \sum_{i \in \mathcal{I}_{j+1,2k+1}} a_i \right).$$

From this, it follows that, unlike the parameters in the model of this paper, the parameters of the corresponding Beta-densities for $\theta_{j,k}$ do depend on scale.

If the model is constructed from a specification in $\theta_{j,k}$, the $\rho_{0,0,i}$ follow from:

$$\rho_{0,0,i} = \prod_{j=0}^{J-1} \overline{\theta_{j,k(j,i)}},$$

where $J = \log_2(n)$ is the total number of dyadic scales, $k(j,i) = \lfloor i/2^{J-j} \rfloor$ and

$$\overline{\theta_{j,k(j,i)}} = \begin{cases} \theta_{j,k(j,i)} & \text{if} \quad k(j+1,i) \text{ is even} \\ 1 - \theta_{j,k(j,i)} & \text{if} \quad k(j+1,i) \text{ is odd.} \end{cases}$$

Specification of the model for $\theta_{j,k}$ fixes the prior for $\rho_{0,0,i}$, and hence for all $\rho_{j,k,i}$. Since a Dirichlet prior on $\rho_{0,0,i}$, implies a Beta with specific parameters on $\theta_{j,k}$, only a consistent choice of Beta-parameters leads to a Dirichlet prior on $\rho_{0,0,i}$. A generalisation of the prior model on $\rho_{0,0,i}$ that includes more (Beta) priors on multiscale likelihood $\theta_{j,k}$, for general classes of wavelets (i.e., beyond Haar) is subject of current research.

This discussion illustrates that the idea to consider intensities and observations within the framework of *total* intensity at sample resolution ($N_{j,k}$) is probably easier to extend beyond Haar than directly relating to successive scales.

5.3 A hidden Markov tree model

So far, the prior mixture probabilities $p_l = P(\zeta_l \neq 0)$, have been modelled separately, i.e., the events $\{\zeta_l \neq 0\}$ for different l are independent. We also implemented a version where the observation at coarser scales was taken into account for the computation of posterior probabilities. This approach somehow helps in removing one-scale artifacts and accentuating important features. Though the improvement was not spectacular in this simple model, more elaborated Hidden Markov Trees (Romberg et al. 2001), are easy to incorporate into the presented framework as well, and possibly add some substantial improvement.

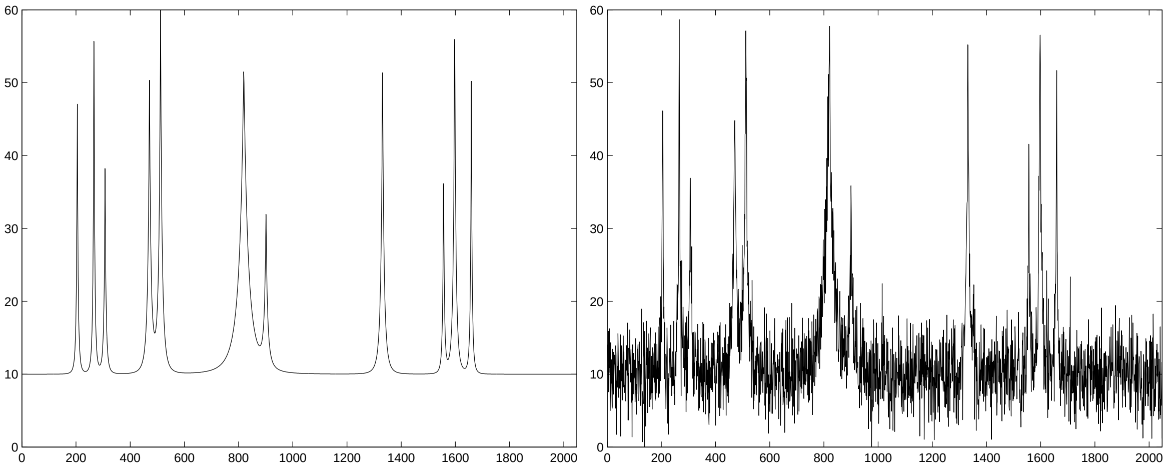

Figure 1: The ‘Bumps’ test signal and a realisation.

6 Simulations

6.1 Conditional Variance Stabilisation versus Anscombe

A first important competitor for the method proposed in this paper is the normalisation procedure for Poisson data by Anscombe (Anscombe 1948; Donoho 1993). The Anscombe procedure is quite straightforward:

1. For every observed count x_i, define
$$y_i = \sqrt{x_i + c},$$
with some constant c. For asymptotic reasons, this constant is generally given the value $c = 3/8$, although simulations indicate that $c = 0$ might be at least equally interesting for small intensities.

2. Apply any wavelet (or other) smoothing technique for additive normal data to the vector $\boldsymbol{y}$. Call $\widehat{\mu}_i$ the output for the i-th data point. It estimates $\mu_i = EY_i$.

3. Estimate the Poisson intensity λ_i of the observation x_i as
$$\widehat{\lambda}_i = \widehat{\lambda}_i^{\,*} + V\left(\sqrt{\xi + c}\,\middle|\,\xi \sim \text{Poisson}\big(\hat{\lambda}_i^{\,*}\big)\right),$$
where
$$\widehat{\lambda}_i^{\,*} = \widehat{\mu}_i^{\,2} - c.$$
The term $V\left(\sqrt{\xi + c}\,\middle|\,\xi \sim \text{Poisson}\big(\hat{\lambda}_i^{\,*}\big)\right)$ corrects for the bias due to squaring an estimation. Indeed, if $\mu_i = EY_i$, and $X_i = Y_i^2 - c$, then
$$\lambda_i = EX_i = E(Y_i^2 - c) = EY_i^2 - c = \mu_i^2 + V(Y_i) - c.$$

We ran 100 simulations on the ‘Bumps’ test signal (Donoho and Johnstone 1994) in Figure 1, with 2048 observations. As measure of quality, we use Signal-to-Noise ratio (SNR), defined as:

$$\text{SNR}(\widehat{f}) = 10 \log_{10} \left(\frac{\|f\|^2}{\|\widehat{f} - f\|^2} \right). \tag{22}$$

	CVS	Anscombe
mean output SNR	13.96	13.08
st.err. output SNR	0.051	0.047

Table 1: Mean output SNR values and standard errors of output SNR values for ‘bumps’ test signal in Figure 1, using Conditional Variance Stabilisation (CVS) and Anscombe’s normalisation.

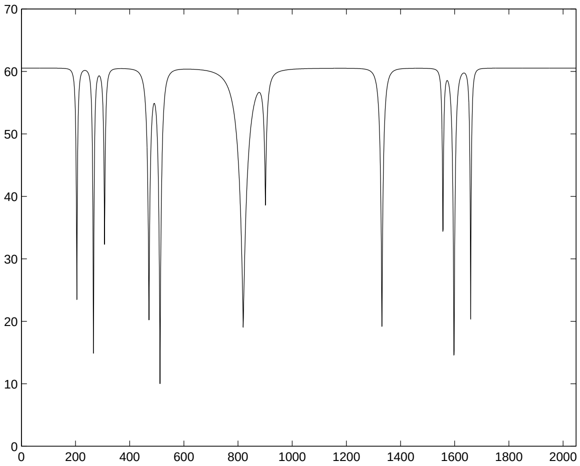

Figure 2: The ‘Bumps’ test signal upside down.

An increase in SNR of 1dB corresponds to reduction of the mean squared error (MSE $= \|\widehat{f} - f\|^2$) of 21% (i.e., a multiplication with factor $10^{-1/10}$). We run the simulations with a non-decimated wavelet transform, using Daubechies' orthogonal wavelets with 3 vanishing moments, and for now, we apply simple thresholds with the exact minimum mean squared error thresholds. (In practical applications, such a threshold has to be approximated, e.g., using SURE or cross validation.) Table 1 compares mean and standard error of the output SNR-values. Anscombe's approach has at least two disadvantages: first, it considers smoothness of $\sqrt{f}$ rather than smoothness of f itself. Second, taking square roots makes bumps less prominent against background noise. This results in a SNR which is almost 1dB lower than obtained with CVS. CVS outperformed Anscombe in all of our 100 runs. Taking square roots however also accentuates downward bumps. We therefore flipped the bumps signal upside down, as in Figure 2 and rerun the 100 simulations. As Table 2 illustrates, Anscombe is now slightly better than CVS, but the difference is less outspoken. In most applications, upward bumps are probably more important than dips. The same conclusions hold for other types of wavelets and also if one compares the Bayesian algorithm proposed in this paper with the Bayesshrink procedure proposed by Johnstone and Silverman (Johnstone and Silverman 2004).

	CVS	Anscombe
mean output SNR	9.83	10.22
st.err. output SNR	0.040	0.043

Table 2: Mean output SNR values and standard errors of output SNR values for upside down ‘bumps’ in Figure 2, using Conditional Variance Stabilisation (CVS) and Anscombe’s normalisation.

6.2 Fisz-Wavelet versus Conditional Variance Stabilisation

The Fisz-Wavelet transform uses the Multiscale Fisz transform as preprocessing step, whereas Conditional Variance Stabilisation can be seen as an extension of Multiscale Fisz from a Haar-like decomposition towards arbitrary types of wavelets. Fisz-Wavelet smoothing proceeds as follows (Fryźlewicz and Nason 2003):

1. Apply a Multiscale Fisz (MF) decomposition. This is equivalent to:
   (a) Apply a Haar transform.
   (b) Apply Conditional Variance Stabilisation to these Haar transform coefficients.
2. To these MF coefficients, apply an inverse Haar transform.
3. Apply a forward wavelet transform, using the basis and filters that best match with the signal at hand.
4. Apply any smoothing (threshold, shrinkage) technique for wavelet coefficients with additive, normal noise.
5. Apply an inverse wavelet transform.
6. Apply a forward Haar transform
7. Reconstruct the data with an inverse Multiscale Fisz transform, i.e., undo the variance stabilisation and apply an inverse Haar transform.

The separation of stabilisation from the actual multiscale processing creates a few disadvantages:

1. A separate multiscale preprocessing leads to a global algorithm which is slightly more computationally complex than doing everything in one single decomposition.
2. It is a bit unclear whether the underlying signal keeps the same smoothness characteristics after applying the Haar-like preprocessing. Also, upon reconstruction, the undoing of the normalisation happens in a Haar-basis, and therefore may partly destroy the initial smoothness of the reconstruction from the inverse wavelet transform with non-Haar filters: although this last step has probably less impact than a threshold, it does operate on coefficients in a Haar-basis, so the output will show some Haar-like artifacts.
3. A fully redundant implementation of a Fisz-Wavelet transform is impossible. A cycle spinning version of the actual wavelet transform is of course straightforward, but the Multiscale Fisz variance stabilisation is intrinsically based on a decimated decomposition. Indeed: a non-decimated multiscale Fisz decomposition is very unlikely to be an exact redundant Haar decomposition of any signal. Any reconstruction from this multiscale Fisz decomposition using an inverse redundant Haar transform is therefore an irreversible process, unless the reconstruction is based on only one of the cycles. In that case, there is no point in a redundant multiscale Fisz in the first place. So, the only way to perform a cycle-spinning Multiscale Fisz variance stabilisation is by averaging all possible cycles explicitly. Although this is time consuming, it may help in reducing artifacts from the post-processing step in a Haar basis.

The benefits from the integrated approach of CVS are more outspoken if the intensity function has long smooth intervals. Indeed, artifacts due to thresholding and postprocessing become most visual on such intervals. To illustrate this, we run 100 simulations on the ‘Heavisine’ test signal (Donoho and Johnstone 1994), depicted in Figure 3. Once again, we adopt Daubechies’ orthogonal wavelets with three vanishing moments, this time in a decimated transform, and we apply level-dependent minimum MSE thresholds. The results for the 100 simulations are summarised in Table 3. The differences in output SNR are small, compared to the standard errors of the the mean values. This is not so surprising, since the issue is smoothness of the output, and MSE is not well suited to measure smoothness. Moreover, a closer look at the coefficients at the input of the final step, show that these are non-zero but relatively small (compared to significant coefficients). Nevertheless, CVS performed systematically better than FW in all of the 100 runs.

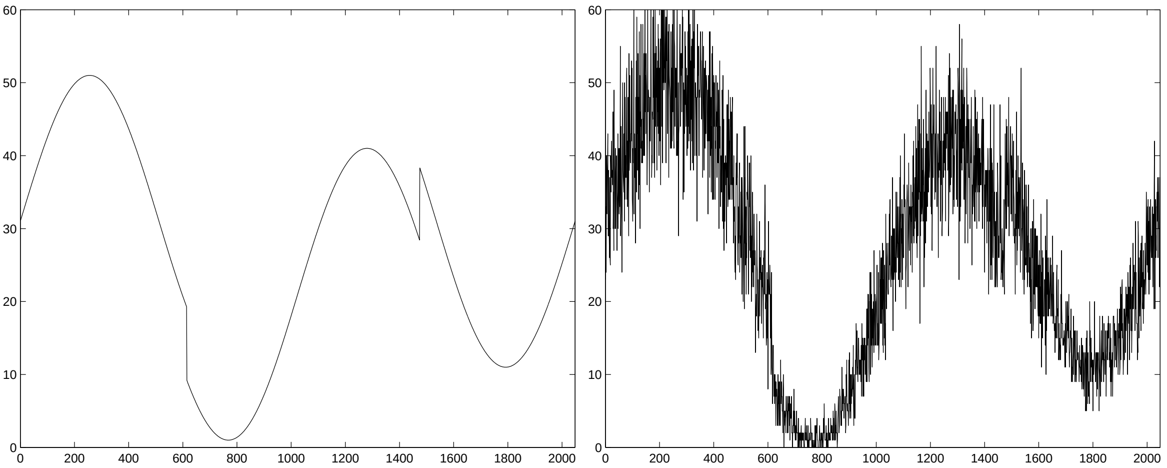

Figure 3: The ‘heavisine’ test signal and a realisation.

	CVS	FW
mean output SNR	24.18	23.54
st.err. output SNR	0.071	0.062

Table 3: Mean output SNR values and standard errors of output SNR values for ‘heavisine’ test signal in Figure 3, using Conditional Variance Stabilisation (CVS) and the Fisz-Wavelet (FW) approach.

	CVS	FW
mean output SNR	25.63	24.40
st.err. output SNR	0.098	0.078

Table 4: Mean output SNR values and standard errors of output SNR values for ‘heavisine’ test signal in Figure 3, using non-decimated wavelet transforms and Bayesian Conditional Variance Stabilisation (BCVS) and the Bayesian Multiscale Shrinkage Models (BMSMS) (Kolaczyk 1999). BCVS is not limited to Haar-like decompositions.

6.3 Multiscale likelihood versus Bayesian Conditional Variance Stabilisation

We now proceed to the Bayesian framework and compare with the Bayesian multiscale likelihood model in (Kolaczyk 1999). As explained in Section 4, the two models coincide when Bayesian CVS is applied to the Haar transform. The only essential difference might be the choice of the hyperparameters, since the models are specified in a somehow different way. This is confirmed in simulations, where the two variants show comparable performance.

The benefit from going for a higher number of vanishing moments is illustrated in the following setup: a non-decimated transform of the ‘heavisine’ test data, using Daubechies’ orthogonal filters with three vanishing moments, and a Bayesian posterior median thresholding. The results are summarised in Table 4. Once again, BCVS was the better choice in each of the 100 runs.

7 Applications

7.1 Gamma burst data

The first illustrative data set, available for download from Theofanis Sapatinas’ website (Besbeas et al. 2004), is the gamma-burst signals, as observed by the BATSE instruments on board NASA’s Compton Gamma Ray Observatory (Kolaczyk 1999; Besbeas et al. 2004).

This signal is a typical example of peaks against a low intensity background. Since the predominant part of the noise is situated in the background, this is close to the situation of additive, homoscedastic noise. Simple wavelet thresholding works reasonably well (Besbeas et al. 2004). In finding sharp transitions between background and significant bursts, it will be hard to beat a (non-decimated) version of Haar-based decompositions. Wider filters have a tendency towards Gibbs phenomena, i.e., wiggly structures, especially near sharp transitions. Nevertheless, we illustrate that a careful coefficient selection in a smooth wavelet basis leads to smooth reconstruction that preserves all important features.

Referring to the signal in Figure 4, any method designed for Poisson data has some difficulties in detecting the two dips between $t = 0.2s$ and $t = 0.25s$. This is because these two features live in an environment where the variance is higher than in the background noise. In order to save these features, we proceed in two steps: we first reconstruct the intensities using the Bayesian shrinkage procedure on a non-decimated decomposition using symmlet filters with 8 vanishing moments. Next, we replace all shrunk coefficients corresponding to locations where this pilot estimator is significantly above the background level, except in the finest three resolution levels. The background level can be estimated using a robust median estimator. The result is depicted in Figure 5. At first sight, there seems to be some Gibbs effect near the first burst (between $0.1s$ and $0.15s$). A closer look at the input however reveals that the small dips seem to be present in the original as well.

7.2 Hits on an internet domain

A second real data example comes from the weekly web statistics on my personal web site. The series has been running since the first week of February 1997 and it shows some remarkable properties, see Figure 6. The two peaks (indicated with a 1 and 2 in the figure) are due to announcements in news groups. The data of week 46 are missing (the software replaced it by 0). A human viewer also immediately recognises the annual “Christmas dips”. The smoothing algorithm

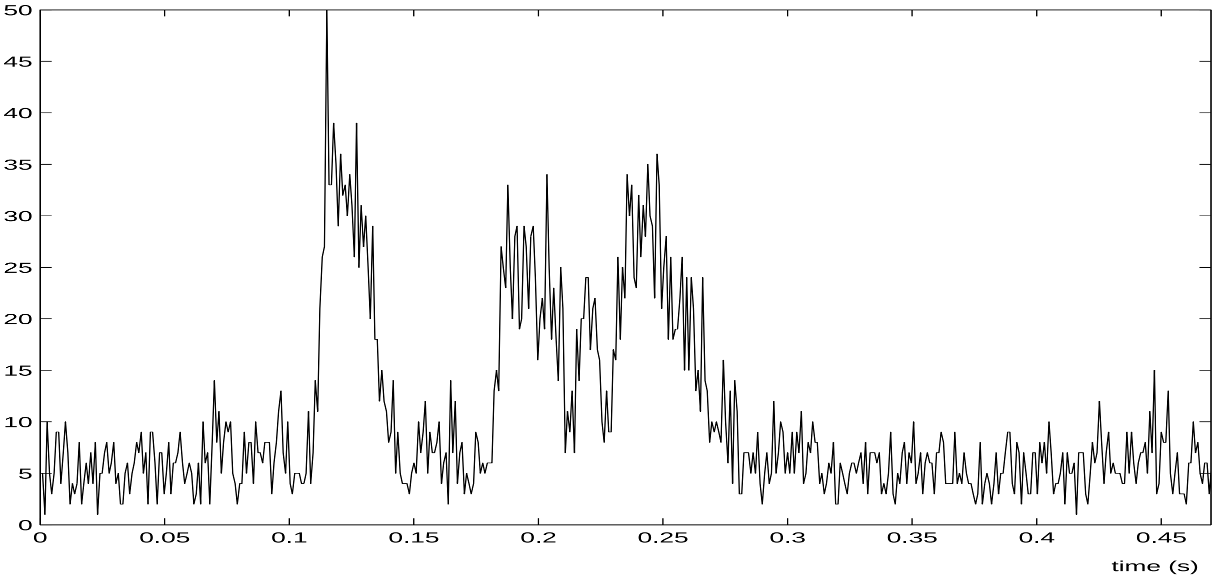

Figure 4: The Gamma-ray burst data. Only the first half of the 1024 observations is plotted. The second half consists of background only.

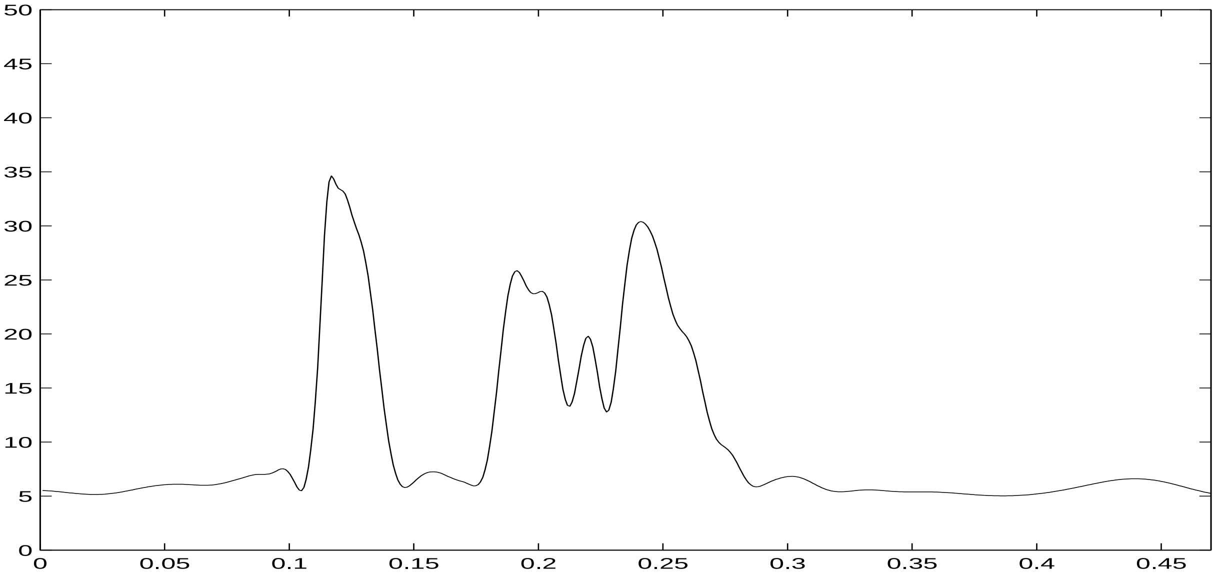

Figure 5: The estimated intensities of the Gamma-ray burst data in Figure 4. We used Bayesian shrinkage on a variance stabilised non-decimated decomposition using symmlet filters with eight vanishing moments. This was followed by some postprocessing on the data in areas with significant intensities (see text).

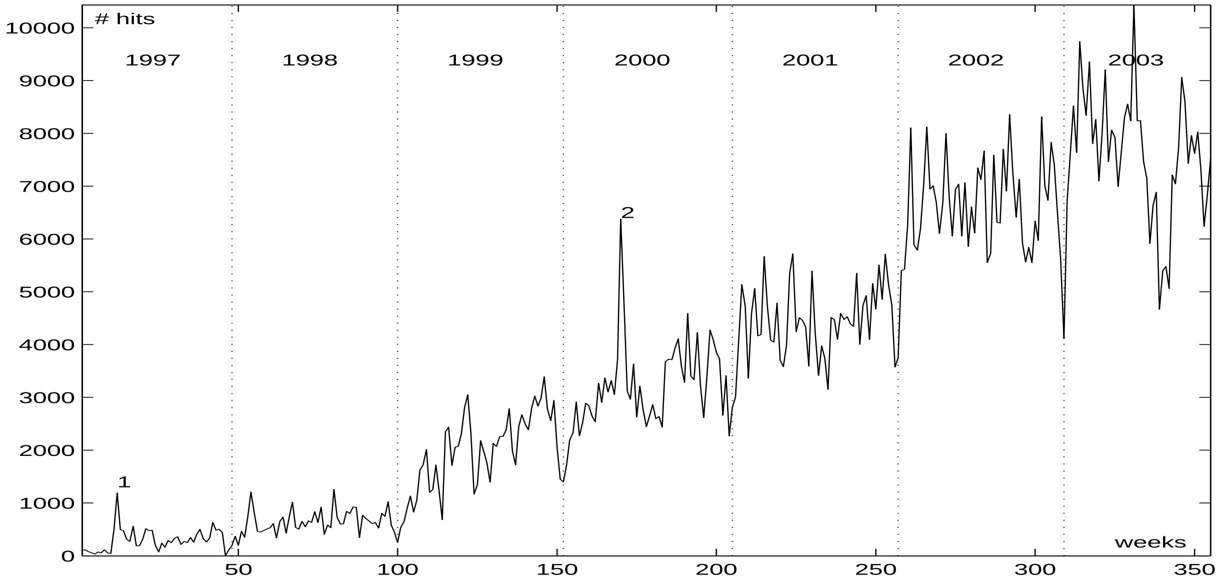

Figure 6: The web domain hits data.

finds those dips little or even not significant, as we discuss below, since it cannot take their annual character into account (as a human viewer does). Another striking, and yet unexplained, feature appears to be the sudden increase to a higher level after Christmas, especially in January 1999, 2001, 2002 and also 2003. In some years (2001, 2002, and especially 2003), that initial gain was lost (partly) after a few months. Since the underlying intensity seems to have discontinuous changes, a wavelet decomposition is an appropriate tool for analysis of these data.

The example illustrates that the immediate applications of the method presented in this paper is not strictly limited to Poisson data. Indeed, the weekly number of hits on an internet domain is certainly not Poisson distributed: it counts every attempt to downloaded any file including images, text and so on. Visitors usually cause more than one hit. If we call X_i the number of hits in week i, R_i the number of visitors that week, and $S_{i,k}$ the number of files downloaded by the k-th visitor in week i, we have

$$X_i = \sum_{k=1}^{R_i} S_{i,k}.$$

We assume that the number of visitors is Poisson distributed with intensity λ_i. The average number of downloads μ_{Si} and its variance σ_{Si}^2 are supposed to be little varying functions of time i. The average μ_{Si} depends, among others, on the number of files available on that domain.

We then have:

$$\begin{aligned} EX_i &= \lambda_i \, \mu_{Si} \\ V(X_i) &= \lambda_i^2 \sigma_{Si}^2 + \lambda_i \left(\mu_{Si}^2 + \sigma_{Si}^2 \right). \end{aligned}$$

We want to estimate EX_i and we assume that $\sigma_{Si}^2 \ll \mu_{Si}^2$, so that $V(X_i) = \kappa \cdot EX_i$, for some constant κ.

In order to capture the narrow peaks as much as possible, we opt for a wavelet with narrow support: the biorthogonal spline wavelet of Cohen, Daubechies and Feauveau (Cohen et al. 1992) with two primal and two dual vanishing moments. This basis (CDF 2,2) is well known in image processing (it is in the JPEG-2000 standard). The result in Figure 7 follows from a decimated Bayesian thresholding algorithm. The smoothing curve captures all characteristics that we discussed above. The piecewise linear CDF 2,2 basis functions are clearly reflected in this output. We also point out that other

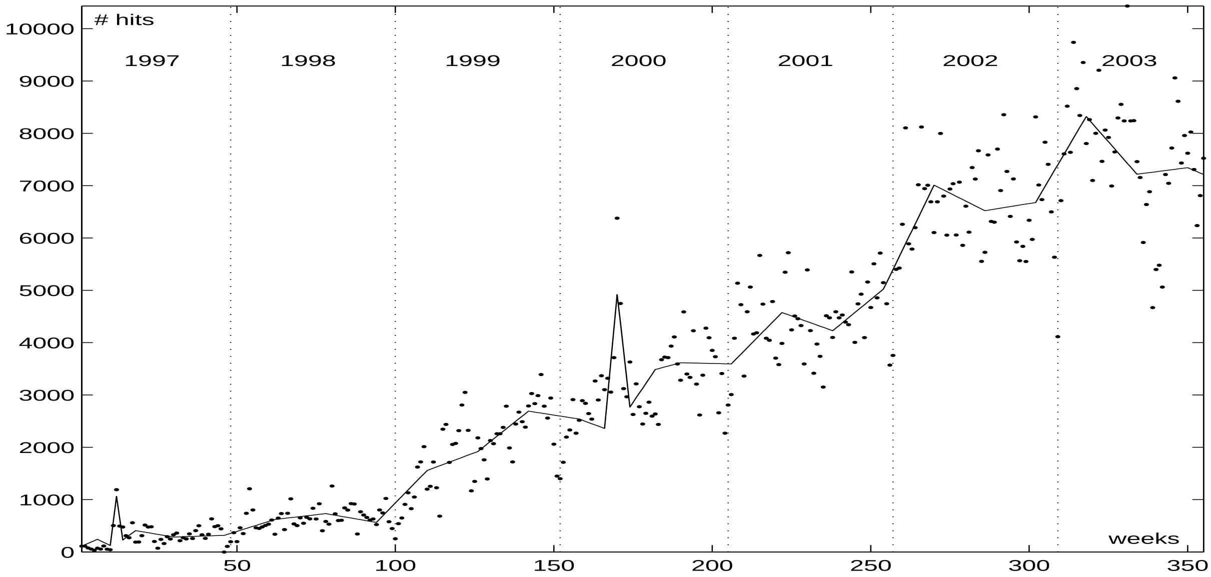

Figure 7: Estimation of the weekly expected number of hits, using a decimated wavelet transform, Bayesian shrinkage and CDF 2,2 wavelet basis.

wavelet bases might not reconstruct the features so well, or even skip some of them. This illustrates the importance of the ability to incorporate variance stabilisation into any wavelet basis.

8 Conclusions

This paper has introduced a novel framework for estimating the intensity curve of Poisson data with piecewise smoothly changing intensities. The key concept is the idea of conditioning on the sum of the observations involved in the computation of a wavelet coefficient. The proposed framework brings together the benefits from some existing procedures:

1. With the Anscombe preprocessing approach (Anscombe 1948), the proposed method shares the ability to incorporate any wavelet transform.
2. From the Fisz-Wavelet decomposition (Fryźlewicz and Nason 2003), it inherits the possibility to apply any threshold procedure for coefficients with additive, homoscedastic noise.
3. Just like the Bayesian multiscale model (Kolaczyk 1999), the method can also be implemented in a translation invariant way, and obviously, it also has a Bayesian component.

The proposed method automatically adapts to situations of low or high intensities, or data with areas of both low and high intensities. Beside these properties, the proposed method can also be extended into different directions:

1. Non-equispaced data, using the lifting scheme.
2. Models for interscale and intrascale dependencies: Hidden Markov Trees, Markov Random Fields.
3. Other statistical models for the noise, different from Poisson. The concept of Diagonal Covariance Stabilisation can be applied in a wider class of exponential distributions. Similar filtering techniques have already been studied (Antoniadis and Sapatinas 2001; Antoniadis et al. 2001).

References

Abramovich, F., Sapatinas, F., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, 60:725–749.

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 cubicariance functions. *Sankhya, Series A*, 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 simulation study of wavelet shrinkage estimators for poisson counts. *International Statistical Review*, 72.

Charles, C. and Rasson, J. P. (2003). Wavelet denoising of poisson-distributed data and applications. *Computational Statistics and Data Analysis*, 43:139–148.

Chipman, H., Kolaczyk, E., and McCulloch, R. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.*, 92:1413–1421.

Cohen, A., Daubechies, I., and Feauveau, J. (1992). Bi-orthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 45:485–560.

Dahmen, W. and Micchelli, C. A. (1986). Statistical encounters with b-splines. In *Function estimates (Arcata, CA, 1985)*, volume 59 of *Contemp. Math.*, pages 17–48. Amer. Math. Soc.

Donoho, D. L. (1993). Non-linear wavelet methods for recovery of signals, densities and spectra from indirect and noisy data. In Daubechies, I., editor, *Different perspectives on wavelets*, volume 47 of *Proceedings of Symposia in Applied Mathematics*, pages 173–205. American Mathematical Society.

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

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90:1200–1224.

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

Fryźlewicz, P. and Nason, G. (2003). A wavelet-Fisz algorithm for Poisson intensity estimation. *Journal of Computational and Graphical Statistics*.

Ignatov, Z. G. and Kaishev, V. K. (1989). A probabilistic interpretation of multivariate b-splines and some applications. *SERDICA, Bulgariae mathematicae publicationes*, 15:91–99.

Jansen, M. (2001). *Noise reduction by wavelet thresholding*, volume 161 of *Lecture Notes in Statistics*. Springer.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: empirical bayes estimates of possibly sparse sequences. *Annals of Statistics*, page to appear.

Kaishev, V. K. (1991). A gaussian cubature formula for the computation of generalized b-splines and its application to serial correlation. In *Statistical multiple integration (Arcata, CA, 1989)*, volume 115 of *Contemp. Math.*, pages 219–237. Amer. Math. Soc.

Karlin, S., Micchelli, C., and Rinott, Y. (1986). Multivariate splines: a probabilistic perspective. *J. Multivariate Analysis*, 20(1):69–90.

Kolaczyk, E. D. (1996). Estimation of intensities of burst-like poisson processes using haar wavelets. *Biometrika*, 46:352–363.

Kolaczyk, E. D. (1997). Non-parametric estimation of gamma-burst intensities using haar wavelets. *Astrophys. J.*, 483:340–349.

Kolaczyk, E. D. (1999). Bayesian multiscale models for Poisson processes. *J. Amer. Statist. Assoc.*, 94:920–933.

Kolaczyk, E. D. and Nowak, R. D. (2000). A multiresolution theory for likelihoods: Theory and methods. Technical report, Department of Mathematics and Statistics, Boston University.

Kolaczyk, E. D. and Nowak, R. D. (2002). Multiscale statistical models. In Denison, D. D., Hansen, M. H., Holmes, C. C., Mallick, B., and Yu, B., editors, *NNonlinear Estimation and Classification*, volume 171 of *Lecture Notes in Statistics*, pages 245–255. Springer-Verlag.

Kolaczyk, E. D. and Nowak, R. D. (2003). Multiscale generalized linear models for nonparametric function estimation. *Submitted*.

Kolaczyk, E. D. and Nowak, R. D. (2004). Multiscale likelihood analysis and complexity penalized estimation. *Annals of Statistics*, 32(2):to appear.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions: statistical theory. *Annals of Statistics*, 11:515–529.

Nowak, R. and Baraniuk, R. G. (1999). Wavelet-domain filtering for photon imaging systems. *IEEE Transactions on Image Processing*, 8(5):666–678.

Robert, C. P. (2001). *The Bayesian Choice*. Springer Texts in Statistics. Springer, second edition.

Romberg, J., Choi, H., and Baraniuk, R. G. (2001). Bayesian tree structured image modeling using wavelet-domain hidden markov models. *IEEE Transactions on Image Processing*, 10(7):1056–1068.

Sardy, S., Antoniadis, A., and Tseng, P. (2003). Automatic smoothing with wavelets for a wide class of distributions. *Journal of Computational and Graphical Statistics*.

Sweldens, W. (1997). The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546.

Timmerman, K. E. and Nowak, R. D. (1999). Multiscale modeling and estimation of poisson processes with application to photon-limited imaging. *IEEE Transactions on Information Theory*, 45(3):846–862.