Multiscale Poisson data smoothing

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

October 2005, final version

Abstract

This paper introduces a framework for nonlinear, multiscale decompositions of Poisson data that possess piecewise smooth intensity curves. The key concept is conditioning on the sum of the observations that are involved in the computation of a given multiscale coefficient. Within this framework, most classical wavelet thresholding schemes for data with additive, homoscedastic noise can be used. Any family of wavelet transforms (orthogonal, biorthogonal, second generation) can be incorporated into this framework. Our second contribution is to propose a Bayesian shrinkage approach with an original prior for coefficients of this decomposition. As such, the method combines the advantages of the Haar-Fisz transform with wavelet smoothing and (Bayesian) Multiscale Likelihood models, with additional benefits, such as extendibility towards arbitrary wavelet families. Simulations show an important reduction in average squared error of the output, compared to the present techniques of Anscombe or Fisz variance stabilisation or Multiscale Likelihood Modelling.
Keywords: Wavelet, smoothing, Poisson count data, Bayesian estimation, Dirichlet

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 kernel estimators, 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 be able to locate (directly, or indirectly, as in the case of wavelet thresholding) the points of discontinuity, since they carry the essential information of the data (signal, image). The combination of a wavelet transform, linear in itself, with the nonlinear threshold method is a fast and efficient approach for catching the singularities.

The subject of this paper is Poisson data. More specifically, we want to estimate, for a given vector of Poisson count observations, the corresponding vector of Poisson intensities. Poisson noise has a multiplicative aspect. This means: the more intense the signal is, the more variable are the fluctuations (the noise). This type 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 of such a real data example. The main difficulty we address is that Poisson noise with time varying intensity cannot be homoscedastic, since the variance equals the expected value.

Many wavelet thresholding techniques assume a constant noise behaviour: noise is typically supposed to be additive, homoscedastic, and uncorrelated. Any orthogonal 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 subject to noise with a large variance and another coefficient where the noise has a small variance. The basic idea behind thresholding is *sparsity*, i.e., a wavelet transform maps a (digital) signal onto a set of

wavelet *coefficients*. The classical additive model (not used in this paper) assumes that the input $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\eta}$ is a vector of noisy observations from unknown function values $\boldsymbol{f}$ *plus* noise $\boldsymbol{\eta}$. The observed (or empirical) wavelet coefficients inherit the additive model, i.e., $\boldsymbol{w} = \boldsymbol{v} + \boldsymbol{\omega}$. Most noise-free values $\boldsymbol{v}$ are close to zero, while only a limited subset of large coefficients carries the essential information. Since noise $\boldsymbol{\omega}$ is spread out evenly (at least if the observations are statistically independent and homoscedastic), a smooth reconstruction of the underlying signal can be obtained if all coefficients with magnitude below a certain *threshold* value are replaced by zero. The multiresolution nature of a wavelet transform offers a solution for data with homoscedastic but correlated 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 normal density is stable under a linear transform: the wavelet coefficients are again normally distributed.

Unlike the model of observations with additive, jointly normal noise, the multiplicative Poisson model is not stable under a linear transform, such as a wavelet transform: wavelet coefficients are not Poisson distributed. A straightforward method to deal with this problem is a preprocessing, normalising step. Examples of this strategy are the Anscombe transformation (**??**) and the Fisz normalisation (**??**). Most papers (**????**) use the simple Haar transform. These specific properties allow for an exact closed form of the scaling coefficient densities. This exact expression can be used in a general Bayesian multiscale model (**??**). A major theme of this paper is to extend ideas from Haar-Fisz decompositions (**?**) and Bayesian Multiscale models (**??**) 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.

A third direction of existing research is an asymptotic study of the applicability of classical wavelet thresholding when the Poisson intensities tend to infinity, for example, bursts against a homogeneous background. This paper however considers signals with low intensities as well as signals with a mixture of both low-count and high-count intervals.

Other methods apply general wavelet domain filtering (such as shrinkage and modulation) based on unbiased variance estimation in the wavelet domain (**???**, page 136–137). Instead of filtering wavelet coefficients, some methods look for a generalisation of the universal threshold to Poisson data (**?**) or wider classes of distributions (**?**).

For a more complete overview of the literature, we refer to an extensive comparative study of several existing methods in **?**.

This paper is organised as follows: Section 2 introduces a conditional variance stabilisation. This is extended to an empirical Bayesian shrinkage approach in Section 3. A simulation study is presented in Section 4. This simulation compares the proposed method with the normalisations by Anscombe (**?**) and Fisz (**?**), and with the Bayesian multiscale likelihood model (**?**). These methods are considered among the state-of-the-art of the currently available methods (**?**). The proposed method is found to have a superior performance in most settings.

2 The proposed conditional variance stabilisation (CVS)

2.1 Definitions

Suppose $\boldsymbol{X}$ is a vector of n Poisson observations X_i with intensities $\boldsymbol{\lambda} = [\lambda_i]_{i=1\ldots n}$. The wavelet transform, $\boldsymbol{W}$, of these data is given by

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

where $\widetilde{W} = [W_{ij}]_{i,j=1,\ldots,n}$ denotes the $n \times n$ 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:

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

where $\mathcal{I}_{j,k}$ is the set of indices with nonzero entries $\gamma_{j,k,i} \neq 0$ in the wavelet transform matrix $\widetilde{W}$ on the row corresponding to the coefficient $W_{j,k}$. 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 and not on location k within that scale. Note that the doubley (j,k) indexed wavelet coefficients can be stored in single vector of length n. The wavelet coefficients $W_{j,k}$ are clearly heteroscedastic. For Poisson data, we can look for a normalisation factor, so that the variances of the normalised coefficients are approximately constant and independent of the input intensities. We would like the coefficients with noise-free value equal to zero, i.e., the negligible ones, to have constant variances, so that they can be removed with a single threshold.

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

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

It is clear that $N_{j,k}$ is Poisson distributed with intensity $\lambda_{j,k}$, equal to $\lambda_{j,k} = \sum_{i \in \mathcal{I}_{j,k}} \lambda_i$. We then define the *normalised* or *variance stabilised* wavelet coefficient $Z_{j,k}$ as:

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

When applied to a Haar transform, this variance stabilisation reduces to the Haar-Fisz decomposition (**?**). Our first intention is to make this variance stabilisation method applicable to any wavelet basis function.

The remainder of this section concentrates on a single wavelet coefficient, and therefore we omit the subscripts j and k for notational convenience, and write W, N, Z, λ to indicate, respectively, a wavelet coefficient's value, its normalisation factor, its normalised value and the intensity of its normalisation factor. The cardinality 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 define for each wavelet coefficient the *relative intensities* ρ_i of input X_i, with $i \in \mathcal{I}$ as:

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

where $\lambda = \sum_{i \in \mathcal{I}_{j,k}} \lambda_i$.

The variance stabilisation property of this normalisation is formalised in the following lemma. All proofs follow in an appendix.

Lemma 1 *The variance stabilised wavelet coefficient Z has, conditional on the normalisation factor N being non-zero, the following moments:*

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

$$\mathrm{Var}(Z|N \neq 0) = \sum_{i=1}^{r} \gamma_i^2 \rho_i - \left(\sum_{i=1}^{r} \gamma_i \rho_i\right)^2 + V\left(\sqrt{N}|N \neq 0\right)\left(\sum_{i=1}^{r} \gamma_i \rho_i\right)^2, \tag{2}$$

$$E(Z^2|N \neq 0) = \sum_{i=1}^{r} \gamma_i^2 \rho_i - \left(\sum_{i=1}^{r} \gamma_i \rho_i\right)^2 + \frac{\lambda}{1 - e^{-\lambda}}\left(\sum_{i=1}^{r} \gamma_i \rho_i\right)^2. \tag{3}$$

In the expression of the conditional variance $\mathrm{Var}(Z|N \neq 0)$, the only factor depending on the absolute intensities λ_i is $\mathrm{Var}(\sqrt{N}|N \neq 0)$. N is Poisson distributed with intensity $\lambda = \sum_{i=1}^{r} \lambda_i$. It is well known that, for $\lambda \to \infty$, the variance of the square root of a Poisson count, N, converges quickly to $1/4$. This means that $\mathrm{Var}(Z|N \neq 0)$ is nearly independent of the absolute intensities λ_i. This is an interesting observation, since it motivates the application of a threshold procedure on the normalised coefficients Z.

The normalisation property of the proposed approach follows from a generalisation of Fisz's theorem (**?**).

Proposition 1 *Consider a vector of r independent random variables $\boldsymbol{X}$ with finite mean $\mu_i(\lambda_i)$ and variance $\sigma_i^2(\lambda_i)$, where the density of X_i depends parametrically on λ_i. The parametric dependence is the same for all variables (i.e., if $\lambda_i = \lambda_j$, then X_i and X_j are i.i.d.). Suppose that, for each i, when $\lambda_i \to \infty$, X_i/μ_i converges in probability to 1 and $(X_i - \mu_i)/\sigma_i$ converges in distribution to a standard normal variable. Assume also that, if $\lambda_i \to \infty$ for each i in a given subset of $\{i = 1, \ldots, r\}$, then we have*

$$\frac{\sum_{i=1}^r \gamma_i \mu_i}{\sum_{i=1}^r \mu_i} \to 0.$$

Then for the same subset, and for any positive number p, the variable

$$Z = \frac{\sum_{i=1}^r \gamma_i X_i}{\left(\sum_{i=1}^r X_i\right)^p}$$

converges in distribution to normal variable with mean and variance

$$\mu_Z = \frac{\sum_{i=1}^r \gamma_i \mu_i}{\left(\sum_{i=1}^r \mu_i\right)^p} \tag{4}$$

$$\sigma_Z^2 = \frac{\sum_{i=1}^r \gamma_i^2 \sigma_i^2}{\left(\sum_{i=1}^r \mu_i\right)^{2p}}. \tag{5}$$

Applying this proposition to our case, with $\mu_i = \sigma_i = \lambda_i$ and $p = 1/2$, leads to an asymptotic variance of $\sigma_Z^2 = \sum_{i=1}^r \gamma_i^2 \rho_i$ and an asymptotic mean of $\mu_Z = \sqrt{\lambda} \cdot \sum_{i=1}^r \gamma_i \rho_i$. This asymptotic mean tends to zero if the corresponding noise-free wavelet coefficient vanishes, and tends to infinity otherwise. Actually, Lemma 1 shows that the mean of the normalised coefficient Z is *exactly* zero if the corresponding noise-free wavelet coefficient vanishes. The variance of a normalised coefficient with zero mean equals $\sigma_Z^2 = \frac{1}{r}\sum_{i=1}^r \gamma_i^2$, which is $1/r$ for orthogonal wavelet decompositions. Again, Lemma 1 says that this value is exact, at least after conditioning on $N \neq 0$. The factor $1/r$ is of order $\mathcal{O}(2^{-j})$, with j the scale of the coefficient.

2.2 Thresholding and reconstruction

The above conditional variance stabilisation (CVS) can be implemented via the following threshold algorithm.

1. Apply a wavelet transform to input $\boldsymbol{y}$. The transform may be the standard fast decomposition, or, alternatively, a redundant, shift-invariant representation. Denote by $\boldsymbol{w}$ the vector of coefficients in this decomposition.
2. For all coefficients, compute the normalisation factor $\sqrt{\boldsymbol{N}}$.
3. Find a vector of appropriate level dependent thresholds $\boldsymbol{\theta}$ for the normalised coefficients $\boldsymbol{z} = \boldsymbol{w}/\sqrt{\boldsymbol{N}}$ (using coordinatewise division). Apply this threshold (using soft, hard or any intermediate threshold approach) and denote by $\boldsymbol{z_\theta}$ 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. An inverse wavelet transform yields an estimate of the Poisson intensities.

In Step 2, 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. It is important to note that the introduction of bookkeeping does not cause a bottleneck in the smoothing algorithm.

When applied to the Haar transform, the proposed algorithm coincides with the Haar-Fisz approach when used with Haar wavelet shrinkage (**?**). 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 transform filter coefficients are normalised to 1 and -1 (instead of the more common ℓ_2 normalisation $\pm 1/\sqrt{2}$). Then the variance $\mathrm{Var}(Z|N \neq 0) = 1$, regardless of the resolution level of the normalised coefficient Z. We propose a Bayesian shrinkage and threshold procedure in Section 3.

3 A Bayesian threshold scheme

The preliminary algorithm of Section 2.2 adopts one of the available threshold or shrinkage procedures in Step 3. This shrinkage procedure could also be Bayesian. Bayesian shrinkage rules, if carefully designed, can be a compromise between the classical hard- and soft-thresholding rules. Indeed, while hard thresholding is a discontinuous operation, possibly leading to instability and large variance in the result, soft thresholding suffers from bias in large coefficients (**?**). A good compromise requires that the shrinkage is bounded for large coefficients; see Proposition 3. Bayesian shrinkage also has excellent theoretical properties (**??**). This section tailors a new Bayesian model specifically designed for Poisson data. This Bayesian model has the following properties.

1. It is constructed within the framework of conditioning on $N_{j,k}$, the total number of counts that contribute to wavelet coefficient $W_{j,k}$ at scale j and location k.
2. Inspired by similar approaches in wavelet denoising (**???**), the proposed prior model for the noise-free wavelet coefficients is a *mixture* of a point mass at zero and a continuous density. A point mass at zero models the *sparsity* of a vector a wavelet coefficients with lots of zeros.
3. In the literature, prior models are specified directly on wavelet coefficients. In our approach, however, the prior for the non-zero coefficients is specified indirectly by a model for the *original data intensities*. The prior model for the significant (i.e., non-zero) wavelet coefficients follows in a second step, as a result of taking linear combinations of the intensities; see Section 3.1. This approach is a generalisation of the multiscale likelihood method (**?**) to wavelet families beyond Haar wavelets.
4. Although the prior model is specified in terms of the intensities, we demonstrate that careful design of this model allows all computations to be performed in the wavelet domain.

Since this paper adopts a mixture model, computation of the posterior distribution requires two applications of Bayes' rule: once for the computation of the mixture parameter, called p^*, and a second time for the computation of the posterior distribution under the assumption that a coefficient is significant. Section 3.1 presents the full prior model. Section 3.2 computes the posterior distribution and posterior mixture parameter. Special attention is paid to the posterior mean and median, since these values can be used as shrinkage rules. In order to approximate the posterior median, we also need the posterior variance. As explained in Section 3.2, its computation is slightly more complicated than for the posterior mean. In order to compute the posterior values, we also need expressions for the marginal probabilities, discussed in Section 3.3. The expressions for the marginal distributions also appear in a threshold based on Bayes factors (marginal likelihood ratios). Section 3.4 proves that the posterior mean leads to bounded shrinkage. Finally, Sections 3.5 and 3.6 provide empirical procedures to estimate the hyperparameters of the model, i.e., the prior mixture (or sparsity) parameter p and the parameter $\boldsymbol{a}$ defined below.

3.1 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. This wavelet coefficient can be written as a linear combination of observations $\boldsymbol{X}$: $W = \sum_{i=1}^{r} \gamma_i X_i$. We define $N = \sum_{i=1}^{r} X_i$ 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. Thus $\zeta = E(Z|\boldsymbol{\rho}, N) = \sqrt{N} \cdot \sum_{i=1}^{r} \gamma_i \rho_i$. The dependence of ζ on $\boldsymbol{\rho}$ is now explicitly specified, as in our Bayesian formulation the components of $\boldsymbol{\rho}$ become random variables with their own prior, specified later.

We denote by 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 p is presented after we calculate the marginal probabilities for the observed coefficients. Set $q = 1 - p$. It is 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 the following *Dirichlet* distribution with parameter vector $\boldsymbol{a}$ where $A = \sum_{i=1}^{r} a_i$:

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

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.

It has been shown (**?**) that a linear combination such as $\zeta|N = \sqrt{N}\sum_{i=1}^{r}\gamma_i\rho_i$ 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 (**?**). Both for B-splines and generalised B-splines, there exist quadrature formulae to find integrals, mean values, etc., but we will use a simple normal approximation, based on the following results for mean and variance.

Lemma 2 *Writing $\alpha_i = a_i/A$, the mean and variance of a noise-free wavelet coefficient ζ under the prior specified in equation (7) satisfy:*

$$\begin{aligned} E(\zeta|\zeta \neq 0, N) &= \sqrt{N} \cdot \sum_{i=1}^{r} \gamma_i \alpha_i & (8)\\ &= 0 \qquad \textit{if all } \alpha_i = 1/r \textit{ are equal.} \\ \mathrm{Var}(\zeta|\zeta \neq 0, N) &= \frac{N}{A+1} \cdot \left[\sum_{i=1}^{r}\gamma_i^2\alpha_i - \left(\sum_{i=1}^{r}\gamma_i\alpha_i\right)^2\right] & (9)\\ &= \frac{N}{A+1}\cdot\frac{1}{r}\cdot\sum_{i=1}^{r}\gamma_i^2 \qquad \textit{if all } \alpha_i \textit{ are equal.} \end{aligned}$$

The prior model for significant wavelet coefficients is approximately normal, where, according to Lemma 2, the variance depends on the hyperparameter a. This way, the proposed model for the original intensities has the same descriptive power as a normal model specified directly for the wavelet coefficients. Moreover, as we are working within the framework of conditioning on N, the variance also depends on N. As illustrated in Proposition 3, this feature undoes the drawbacks of a normal prior compared to a heavy tailed prior on wavelet coefficients (**?**).

3.2 Posterior distributions

The conditional probability of $\boldsymbol{X}$ given $\boldsymbol{\rho}$ and N is multinomial. The Dirichlet distribution is a conjugate prior for the multinomial distribution (**?**). 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 is $\boldsymbol{a} + \boldsymbol{x}$.

As a consequence, the posterior density $f_{\zeta|\boldsymbol{X}}(\zeta|\boldsymbol{x}, \zeta \neq 0)$ of a 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:

$$F_{\zeta|\boldsymbol{X}}(\zeta|\boldsymbol{x}) = P(\zeta \neq 0|\boldsymbol{X} = \boldsymbol{x}) \cdot F_{\zeta|\boldsymbol{X}}(\zeta|\boldsymbol{x}, \zeta \neq 0) + P(\zeta = 0|\boldsymbol{X} = \boldsymbol{x}) \cdot I_{\mathbb{R}^+}(\zeta), \tag{10}$$

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:

$$p^* = P(\zeta \neq 0|\boldsymbol{X} = \boldsymbol{x}) = \frac{p}{p + q \cdot \dfrac{P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta = 0, N)}{P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta \neq 0, N)}}. \tag{11}$$

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 probability $P_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\rho}, \zeta \neq 0, N) = P_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\rho}, N)$. The same multinomial conditional probability applies under $\zeta = 0$, but the prior $f_{\boldsymbol{\rho}}(\boldsymbol{\rho}|\zeta = 0)$ 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 $\boldsymbol{X}$, i.e., we want p^* to be equal to:

$$p^* = P(\zeta \neq 0|Z = z, N) = \frac{p}{p + q \cdot \dfrac{P_Z(z|\zeta = 0, N)}{P_Z(z|\zeta \neq 0, N)}}. \tag{12}$$

Thus, we want the Bayes factor (marginal likelihood ratio) for the observed $\boldsymbol{x}$ in expression (11) to be such that it depends on $z = \frac{1}{\sqrt{N}} \sum_{i=1}^{r} \gamma_i x_i$ only. The following proposition ensures that this is possible:

Proposition 2 *Suppose that we have a Dirichlet model (7) for joint relative intensities $\boldsymbol{\rho}$ under the hypothesis that the corresponding noise-free wavelet coefficient is non-zero, i.e., under the hypothesis that $\zeta = \sqrt{N} \sum_i \gamma_i \rho_i \neq 0$.*
Then there exists a prior model for these joint relative intensities $\boldsymbol{\rho}$ under the hypothesis that the corresponding noise-free wavelet coefficient equals zero, i.e., under the hypothesis that $\zeta = 0$, such that the Bayes factor (i.e., the likelihood ratio of both hypotheses) in the observations $\boldsymbol{x}$ depends only on $\boldsymbol{x}$ through the empirical (observed) wavelet coefficient $z = (1/\sqrt{N}) \sum_i \gamma_i x_i$.

This property is useful for two reasons. First, it is more convenient and intuitive to construct a model for $P_Z(z|\zeta = 0, N)$ than for $P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta = 0, N)$. In particular, it is easy to construct a continuous, normal, approximation for the probability function of the univariate (discrete) variable Z. Second, it enables us to perform all the computations for p^* in the wavelet domain.

In order to perform Bayesian shrinkage, we need expressions for the posterior mean and variance. For the posterior mean, we have the following result:

Lemma 3 *Consider a vector of Poisson counts $\boldsymbol{X}$ with relative intensities $\boldsymbol{\rho}$ that have a mixture distribution of a point mass and a joint Dirichlet prior as specified in expression (7). Assuming that the components a_i of the vector of prior hyperparameters $\boldsymbol{a}$ are equal, the posterior expected value for a noise-free wavelet coefficient $\zeta = \sqrt{N} \sum_{i=1}^{r} \gamma_i \rho_i$, equals*

$$E(\zeta|Z, N) = P(\zeta \neq 0|Z, N) \cdot \frac{N}{N + A} \cdot Z. \tag{13}$$

The posterior variance is a bit more complicated. It has the form of expression (9), with $\alpha_i = (a_i + X_i)/(A + N)$. Even if all prior a_i are equal, the posterior values of α_i are no longer equal. Hence, unlike in the proof of Lemma 2, 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 $\mathrm{Var}(\zeta|\boldsymbol{X}, \zeta \neq 0)$ is not necessarily equal to $\mathrm{Var}(\zeta|Z, \zeta \neq 0, N)$.

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 a 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 a threshold scheme (**?**). This threshold erases small posterior means and therefore leads to a smoother reconstruction.

3.3 Marginal probabilities

The marginal probability functions of $\boldsymbol{X}$ and Z already appeared in expressions (11) and (12) for the computation of posterior probabilities. They are also relevant to the estimation of the model's parameters in an empirical Bayes approach. We also need the marginal distribution of $Z|(\zeta \neq 0, N)$ to fill in the still open details of the model for $\zeta = 0$.

Lemma 4 *If $\boldsymbol{X}|(N, \boldsymbol{\rho})$ has a multinomial distribution and the relative intensities $\boldsymbol{\rho}|\zeta \neq 0$ have a joint Dirichlet density, with hyperparameters $\boldsymbol{a}$ and $A = \sum_{i=1}^{r} a_i$, then*

$$P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta \neq 0, N) = \Gamma(N+1) \cdot \Gamma(A) \cdot \prod_{i=1}^{r} \Gamma(a_i + x_i) \cdot \left(\left[\prod_{i=1}^{r} \Gamma(x_i + 1)\Gamma(a_i) \right] \Gamma(A+N) \right)^{-1} \tag{14}$$

If all $a_i = 1$, this, remarkably, reduces to a uniform distribution on all configurations $\boldsymbol{x}$ (note that the conditional distribution is multinomial). The subsequent expressions for marginal mean and variance include this special case of a uniform distribution on the discrete simplex $\{\boldsymbol{x}\}$.

Lemma 4 gives an expression for one of the marginal probabilities that appear in the right-hand side of Expression (11). It also leads to the forthcoming Lemma 5 about the marginal probabilities in Expression (12). In Section 3.2, we required that the two Expressions (11) and (12) for p^* are equal. This leads to a condition on the marginal probabilities $P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta = 0, N)$, and $P_Z(z|\zeta = 0, N)$, which have not been specified yet. As explained in Section 3.2, it is interesting to specify the marginal in the wavelet domain $P_Z(z|\zeta = 0, N)$ first, and let the marginal $P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta = 0, N)$ follow from the equality of the two expressions for p^*. We specify the marginal $P_Z(z|\zeta = 0, N)$, after we state a result for $P_Z(z|\zeta \neq 0, N)$.

Lemma 5 *The mean and variance of a normalised empirical wavelet coefficient, given that its noise-free value is non-zero, satisfy $E(Z|\zeta \neq 0, N) = N \cdot \sum_{i=1}^{r} \gamma_i \alpha_i$, which is zero if $\alpha_i = 1/r$, and*

$$\mathrm{Var}(Z|\zeta \neq 0, N) = (N + A)/N \cdot \mathrm{Var}(\zeta|\zeta \neq 0, N).$$

Remark 1 *Note that the Bayesian shrinkage factor from (13) 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.*

We now specify the marginal probability of coefficients that do not carry information, i.e., for which $\zeta = 0$. The idea is to extend a Bayes factor that holds for typical, specific values of the parameters $\boldsymbol{\rho}$ to all situations. More precisely, we require that the variance of the wavelet coefficient under $\zeta = 0$ is well approximated by the variance in one specific instance of that hypothesis, namely the case of constant intensities:

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

On the other hand, if all $\alpha_i = 1/r$, a combination of Lemmata 2 and 5 yields for significant coefficients that

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

so we can write that

$$\mathrm{Var}(Z|\zeta \neq 0, N) \approx \frac{N + A}{1 + A} \cdot \mathrm{Var}(Z|\zeta = 0, N). \tag{15}$$

The approximation of the general by a specific case now motivates us to *require* that the Bayes factor equals the ratio:

$$\frac{P_Z(z|\zeta = 0, N)}{P_Z(z|\zeta \neq 0, N)} = \frac{\phi_{\sigma_0}(z)}{\phi_{\sigma_1}(z)}, \tag{16}$$

where ϕ_σ stands for the normal density function with zero mean and standard deviation σ, and $\sigma_1 = \sqrt{(N + A)/(1 + A)}\sigma_0$. In these expressions, $\sigma_0 = (1/r) \sum_{i=1}^{r} \gamma_i^2$.

Remark 2 *This choice in terms of the Bayes factor fixes $P_Z(z|\zeta = 0, N)$ for all possible z. This specification is necessary in the computation of the posterior probability p^*. It is, however, unlikely that the marginal probabilities $P_Z(z|\zeta = 0, N)$ sum to one. In order to remedy this problem, it is in principle necessary to allow a Bayes factor different from (16) for (at least) one value of z, for instance, the value $z = 0$ if that value has a non-zero probability.*

In practice, $P_Z(z|\zeta = 0, N)$ is not an exact normal density, so Expression (15) for the variances is still an approximation. The computation of the posterior probability p^* however relies on Bayes factors only: under the now fully specified model, those calculations are exact.

Remark 3 *This section describes marginal probabilities, i.e., likelihoods of* a single wavelet coefficient*. Likewise, the Bayesian approach finds posterior probabilities in every coefficient separately. Since the wavelets we use are not limited to the Haar basis, adjacent coefficients are mutually dependent, so the overall (full) likelihood is not just the product of the individual likelihoods in every coefficient. The proposed algorithm in this paper processes every coefficient separately. Such an approach is also referred to as a pseudo-likelihood.*

3.4 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} for which $|z| > \theta_{\mathrm{BF}}$ implies that $p^* > 1/2$. In other words, coefficients above this threshold, are qualified by the model as 'more likely signal than noise'. This threshold equals:

$$\theta_{\mathrm{BF}} = \sqrt{2\,\frac{N+A}{N-1}\,\log\left(\frac{q}{p}\sqrt{\frac{N+A}{1+A}}\right)}\cdot\sigma_0. \tag{17}$$

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 (10) 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}.

It is interesting to investigate how large coefficients are treated in the given model. In particular, we prove that shrinkage is bounded for given, finite threshold values. This is in contrast to the case of normal noise in combination with a Gaussian prior for significant coefficients: such a model leads to undesirable unbounded shrinkage for large input coefficients.

Proposition 3 *Using the posterior mean as a shrinkage rule for the mixture prior model with point mass in $\zeta = 0$ and a joint Dirichlet away from zero, as specified in (7) and (16), there exists a constant $C < \infty$ such that for $N \to \infty$, i.e. for $\theta_{\mathrm{BF}} \to \infty$,*

$$|Z - E(\zeta|Z, N)| \leq C \cdot \theta_{\mathrm{BF}}. \tag{18}$$

This proposition shows that, for any value of N, sufficiently significant values of z have bounded shrinkage and there exists an upper bound 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, as for heavy tailed priors (**?**).

3.5 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. This hyperparameter has to be estimated. We assume that this parameter is scale dependent, and denote the value at scale j by p_j. If we write $\sigma_0^2 = \mathrm{Var}(Z|\zeta \neq 0, N_{j,k})$ and $\sigma_{1,j,k}^2 = \mathrm{Var}(Z|\zeta = 0, N_{j,k})$, and use the same model details as described in Section 3.3, we can express the likelihood in p_j for an observed vector of normalised coefficients $\boldsymbol{z_j}$ as:

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

Note that $\sigma_{1,j,k}$ depends on $N_{j,k}$, so the likelihood expression is different for every observed coefficient.

The values for σ_0 are independent of the observed $N_{j,k}$, and could be estimated from the data using the Median Absolute Deviation (MAD). As mentioned before, this works satisfactorily 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_i.

At fine scales, p_j is generally quite small, and it might be difficult to capture the few significant coefficients at those fine scales, leading to $\widehat{p_j} = 0$. We therefore require that the posterior median threshold should be below the universal threshold (**?**), i.e. $P(\zeta > 0|Z = \theta_{\text{univ}}, N) \geq 1/2$. This implies a condition on the posterior probability p^* for an observation equal to the universal threshold, i.e., $P(\zeta = 0|Z = \theta_{\text{univ}}, N) \leq 1/2$. Since $\theta_{\text{univ}} = \sqrt{2 \log n}\sigma_0$ is a large value, the probability $P(\zeta < 0|Z = \theta_{\text{univ}}, N)$ is small and both conditions are practically equal. Elaboration of the latter condition leads to $p > C/(C + n^D)$, where $C = \sigma_{1,j,k}/\sigma_0$ and $D = 1 - 1/C^2$. This minimum value for the prior p therefore also depends, through $\sigma_{1,j,k}$, on the observed $N_{j,k}$ in each coefficient.

3.6 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 $j = 0$:

$$\rho_{j,k,i} = \rho_{0,0,i} / \sum_{i \in \mathcal{I}_{j,k}} \rho_{0,0,i}. \tag{19}$$

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 $\boldsymbol{\rho_{0,0}} \stackrel{\text{d}}{=} \boldsymbol{V} / \sum V_i$ and $\boldsymbol{\rho_{j,k}} \stackrel{\text{d}}{=} \boldsymbol{V}_{\mathcal{I}_{j,k}} / \sum_{\mathcal{I}_{j,k}} V_i$. 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:

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

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 clear heterogeneous behaviour, one could consider a non-constant vector $\boldsymbol{a}$ and estimate it from the variances of different subsets of the input sample.

4 Simulations

As in earlier papers (**??**), we ran 100 simulations on low and high intensity versions of four commonly used test signals, called 'Bumps', 'Blocks', 'Heavisine', and 'Doppler' (**?**). These four intensity curves f were rescaled and shifted along the y-axis, such that $\max f = 1/\min f = M$, with $M = 128$ for the high intensity version and $M = 8$ for the low intensity runs. We consider $n = 1024$ equidistant points on every intensity curve.

The simulations were run with a non-decimated wavelet transform. For the results in Table 1, the Daubechies least asymmetric orthogonal wavelets (also known as symmlets) with 10 vanishing moments were used. Table 2 repeats the simulations, this time with Daubechies orthogonal basis with 3 vanishing moments, and illustrates that the good performance of our CVS-method (conditional variance stabilisation) is independent of the particular wavelet transform used.

Rounded mean values of $10,000 \cdot \|\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\|^2 / \|\boldsymbol{\lambda}\|^2$								
Symmlet, 10 vanishing moments								
	Heavisine		Blocks		Bumps		Doppler	
$\lambda_{\max}$	8	128	8	128	8	128	8	128
Anscombe	55	6	219	30	2208	163	114	13
Fisz-Wav., Cycle Spin	29	6	194	30	1082	154	90	13
BMSMS	44	7	135	7	1824	184	147	20
CVS, Cycle Spin	28	5	196	30	910	118	92	12
CVS, non-decim. W.T.	30	6	203	31	930	119	98	13
Bayesian CVS	42	6	234	28	1142	126	114	10

Table 1: Standardised output Mean Average Squared Error (MASE) values (mean over 100 simulations, average over $n = 1024$ observations) for low and high intensity versions (peak intensities 8 and 128) of four test signals. The Anscombe normalisation, Cycled Spinned Haar-Fisz method with wavelet smoothing (abbreviated as Fisz-Wav.) and the Bayesian MultiScale Model Shrinkage (BMSMS) are three competitors discussed extensively in the main text. The Cycle Spin implementation of our method, CVS, as well as non-decimated implementations with and without Bayesian shrinkage are three variants of the new method we propose. All methods except the BMSMS use Daubechies Least Asymmetric orthogonal wavelets ('symmlets') with 10 vanishing moments. All non-Bayesian procedures adopt the exact level dependent minimum ASE thresholds, which explains why CVS without Bayes sometimes outperforms the empirical Bayes algorithm.

The method of Bayesian Multiscale Models (BMSMShrink) (**?**) is, however, inherently based on Haar wavelets. For all methods except the BMSMS and the Bayesian CVS, we apply simple level-dependent thresholds with the exact minimum average squared error. In practical applications, such a threshold has to be approximated, e.g., using SURE or cross validation. The use of exact minimum ASE (average squared error) thresholds explains the relative poor results for the Bayesian thresholding, especially on low intensity signals. Nevertheless, the Bayesian model succeeds very well in selecting the significant coefficients. For high intensity signals, Bayesian shrinkage may even outperform minimum ASE thresholding, because Bayesian shrinkage rules offer a transition between kill and keep which is smoother than the hard or soft thresholding rules.

All simulations can be reproduced using Matlab routines in the recently upgraded package PiefLab (**?**), which can be downloaded from `www.cs.kuleuven.ac.be/~maarten/software/pieflab.html`.

A first important competitor for the method proposed in this paper is the normalisation procedure for Poisson data by Anscombe (**??**). 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 an interesting alternative 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. The quantity $\widehat{\mu}_i$ estimates $\mu_i = EY_i$.

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

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. CVS outperformed Anscombe in all of our runs. 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 **?**.

Rounded mean values of $10,000 \cdot \|\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\|^2 / \|\boldsymbol{\lambda}\|^2$								
Daubechies wavelets, 3 vanishing moments								
	Heavisine		Blocks		Bumps		Doppler	
$\lambda_{\max}$	8	128	8	128	8	128	8	128
Anscombe	55	6	211	26	2190	151	132	15
Fisz-Wav., Cycle Spin	31	6	182	26	1071	142	106	16
BMSMS	44	7	135	7	1824	184	147	20
CVS, Cycle Spin	31	5	185	26	961	124	107	15
CVS, non-decim. W.T.	33	6	193	26	985	125	114	16
Bayesian CVS	39	5	216	21	1212	123	131	12

Table 2: Standardised output Mean Average Squared Error (MASE) values (mean over 100 simulations, average over $n = 1024$ observations) for low and high intensity versions (peak intensities 8 and 128) of four test signals. All methods except the BMSMS use Daubechies orthogonal wavelets with 3 vanishing moments.

The second competitor, Haar-Fisz normalisation with wavelet smoothing, proceeds as follows (**?**):

Step 1. Apply a Haar-Fisz (HF) decomposition. This is equivalent to a Haar transform, followed by a Conditional Variance Stabilisation applied to the Haar transform coefficients.

Step 2. To these HF coefficients, apply an inverse Haar transform.

Step 3. Apply any forward wavelet transform, using the basis and filters that best match with the signal at hand. (In our comparative simulation studies, we used the same filters as in the corresponding CVS method.)

Step 4. Apply any smoothing (threshold, shrinkage) technique for wavelet coefficients with additive, normal noise.

Step 5. Apply an inverse wavelet transform, followed by a forward Haar transform

Step 6. Reconstruct the data with an inverse Haar-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 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 obtained from the inverse wavelet transform with non-Haar filters: although this last step probably has 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 (non-decimated) implementation of a Haar-Fisz method is impossible. A cycle spinning version of the actual wavelet transform is of course straightforward, but the Haar-Fisz variance stabilisation is intrinsically based on a decimated decomposition. Indeed, a non-decimated Haar-Fisz decomposition is very unlikely to be an exact redundant Haar decomposition of any signal. Any reconstruction from this Haar-Fisz decomposition using an inverse redundant Haar transform (Step 2 of the algorithm above) is therefore an irreversible process, unless the reconstruction is based on only one of the cycles. In that case, there is no point in using a redundant Haar-Fisz transform in the first place. So, the only way to perform a cycle-spinning Haar-Fisz variance stabilisation is by averaging all possible cycles explicitly. Although time consuming, such an external cycle spinning reduces

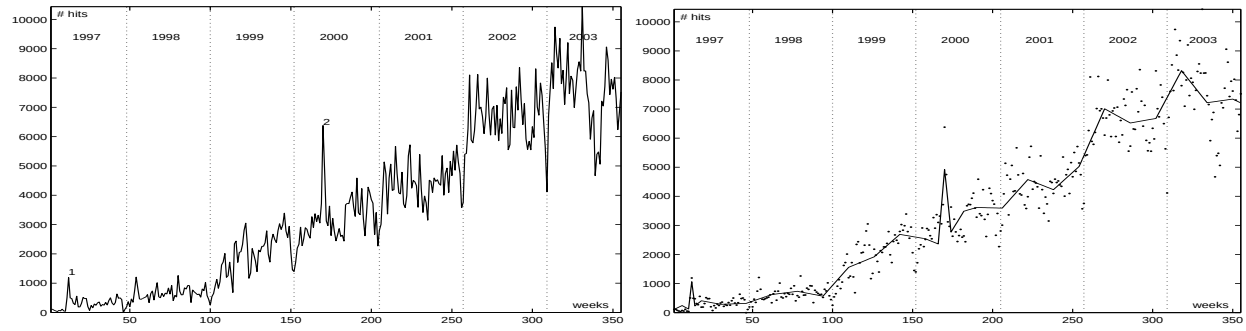

Figure 1: Left: The web domain hits data. Right: Estimation of the weekly expected number of hits, using a decimated wavelet transform, Bayesian shrinkage and CDF 2,2 wavelet basis.

artifacts from the post-processing step in a Haar basis. In principle, the cycles are found by shifting the n input observations over k positions, where $k = 1, \ldots, n$. The number of possible shifts therefore equals the data vector length. Experiments (**?**) indicate that in practice some 50 shifts are sufficient to remove most of the artifacts. We adopted this rule of thumb in our simulations when applying CVS or Haar-Fisz procedures with explicit (external) cycle spinning. We also included a non-decimated implementation of the CVS algorithm (sometimes referred to as internal cycle spinning).

The third competitor is the Bayesian multiscale likelihood model in (**?**). The two models coincide when Bayesian CVS is applied to the Haar transform. The only essential difference is the choice of the hyperparameters, since the models are specified in a different way. This is confirmed in simulations, where, for the special case of Haar wavelets, the two variants show comparable performance. Our method, however, has the important advantage of being applicable to any type of wavelet basis, via a classical wavelet analysis, as well as to second generation wavelets through a lifting scheme.

5 An application: Hits on an internet domain

A real data example (also available in the software package PiefLab used in the simulations) 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 1. 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 finds those dips modest or even insignificant, 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 (partially) lost after a few months. Since the underlying intensity seems to have discontinuous changes, a wavelet decomposition is an appropriate tool for the analysis of these data.

The example illustrates that the immediate applications of the method presented in this paper are 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 download 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 mean number of downloads μ_{Si} and its corresponding variance σ^2_{Si} are supposed to be modestly varying functions of time i. The average μ_{Si} depends on, among other things, the number of files available on that domain.

We then have $EX_i = \lambda_i \, \mu_{Si}$ and $\mathrm{Var}(X_i) = \lambda_i^2 \sigma^2_{Si} + \lambda_i \left(\mu^2_{Si} + \sigma^2_{Si} \right)$. We want to estimate EX_i and we assume that

$\sigma_{Si}^2 \ll \mu_{Si}^2$, as would be the case, for instance, if $S_{i,k}$ were Poisson with large mean, so that $\mathrm{Var}(X_i) \approx \kappa \cdot EX_i$, 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 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 1(Right) 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 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.

6 Conclusions and directions for further research

This paper 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 (**?**), the proposed method shares the ability to incorporate any wavelet transform.

2. From the Haar-Fisz normalisation (**?**), it inherits the possibility of applying any threshold procedure for coefficients with additive, homoscedastic noise.

3. As in the Bayesian multiscale model (**?**), the method can also be implemented in a translation invariant way, and it also has a Bayesian component.

The proposed method automatically adapts to situations of low or high intensities, and to data with areas of both low and high intensities. Beside these properties, extension towards non-equidistant data, using the lifting scheme, is straightforward.

An important subject of further research is the consistency analysis of the proposed estimator. This analysis involves a study of the maximum risk (i.e., expected MSE) over a function class, typically Besov function balls $B_{p,q}^{\alpha}(R)$, with parameters $R, p, q,$ and α. For a formal definition of these function classes, we refer to the literature. A short overview can be found in **?**, pages 76–78. Important for this discussion is the fact that a function is in a Besov ball $B_{p,q}^{\alpha}(R)$ if and only if the sequence of its wavelet coefficients $w_{j,k}$ is in a corresponding Besov sequence ball, $b_{p,q}^{\alpha}(R)$, meaning that

$$\|\boldsymbol{w}\|_{b_{p,q}^{\alpha}} := \left[\sum_{j=L}^{\infty} 2^{j\beta q} \left(\sum_{k=1}^{2^j} |w_{j,k}|^p \right)^{\frac{q}{p}} \right]^{\frac{1}{q}} \leq R, \tag{20}$$

with $\beta = \alpha + 1/2 - 1/p$. (The definition of the Besov sequence norm has to be slightly modified if $q = \infty$.) This Besov sequence norm can be interpreted as a mathematical formulation of multiscale sparsity. Indeed, the inner sum is the ℓ_p-norm of the coefficients at a given scale. Small values of p, i.e., $p < 2$ are of particular interest, since they favour sparse sequences: isolated large coefficients contribute little to the overall ℓ_p-norm. The sequence of the level-dependent ℓ_p norms at all scales is then measured with a weighted ℓ_q norm. The weights are $2^{j\beta q}$. Therefore, if $f \in B_{p,q}^{\alpha}(R)$, then the ℓ_p-norms at fine scales (i.e., with growing j) must decay at least as $\mathcal{O}(2^{-j\beta})$. Since the ℓ_p-norm, with $p < 2$, is dominated by the many small coefficients related to the intervals where f is smooth, this decay is related to the degree of smoothness of the function between its singularities. This observation is similar to the case of smooth functions in C^{α}, where α is related to the decay in the Fourier transform domain.

Appendix: proofs

Proof of Lemma 1. The joint distribution of $(X_1, \ldots, X_r)$ conditional on N is multinomial. Using the expressions for mean and covariance of a multinomial vector, we write $E(Z|N=n) = \sqrt{n}\sum_{i=1}^r \gamma_i\rho_i$ and for the variance of Z given $N=n$:

$$\mathrm{Var}(Z|N=n) = \sum_{i=1}^r \gamma_i^2\rho_i(1-\rho_i) - 2\sum_{i=1}^r\sum_{j=1}^{i-1}\gamma_i\gamma_j\rho_i\rho_j = \sum_{i=1}^r \gamma_i^2\rho_i - \left(\sum_{i=1}^r \gamma_i\rho_i\right)^2 \text{ for } n \neq 0.$$

Note that $\mathrm{Var}(Z|N=n)$ 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 average over all non-zero N to obtain the results of the lemma. □

Proof of Proposition 1. This proposition is an extension of a theorem by Fisz (**?**) and the proof is almost completely similar. Referring to the paper by Fisz, the extension of the Lemma 1 in that paper is trivial. Lemma 2 in that paper can be extended towards more than two independent variables by letting $\xi_1 = \gamma_1 X_1$ and $\xi_i = -\gamma_i X_i$ for $i = 2, \ldots, r$. Then first consider $\xi_1 - \xi_2$, then $(\xi_1 - \xi_2) - \xi_3$ and so on. The proof of the actual theorem can be extended immediately by replacing $m_1 - m_2$ (in the notation of Fisz) by $\sum_{i=1}^r \gamma_i\mu_i$ (our notation).

Proof of Lemma 2. These results follow from combining the expressions for $E(\rho_i|\zeta \neq 0)$, $\mathrm{Var}(\rho_i|\zeta \neq 0)$, and $\mathrm{cov}(\rho_i, \rho_j|\zeta \neq 0)$ in a Dirichlet model. If all $\alpha_i = 1/r$, the first vanishing moment of the dual (analysis) wavelet (**?**, page 241) annihilates this constant, thereby simplifying the expressions. □

Proof of Proposition 2. Suppose that the model is fully specified in terms of wavelet coefficients z. In particular, $P_Z(z|\zeta = 0, N)$ is given, and we state that for all $K = \binom{N+r-1}{r}$ possible configurations of $\boldsymbol{x}$ which sum to given N, and for $z = (1/\sqrt{N})\sum_{i=1}^r \gamma_i x_i$, the Bayes factor in terms of $\boldsymbol{x}$ equals the Bayes factor in terms of the corresponding z: $P_Z(z|\zeta = 0, N)/P_Z(z|\zeta \neq 0, N) = P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta = 0, N)/P_{\boldsymbol{X}}(\boldsymbol{x}|\zeta \neq 0, N)$. This leads to K conditions on $f_{\boldsymbol{\rho}}(\boldsymbol{\rho}|\zeta = 0)$, one for each configuration $\boldsymbol{x}$ that sums up to N:

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

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.

Proof of Lemma 3. Since the posterior density $f_{\boldsymbol{\rho}|\boldsymbol{X}}(\boldsymbol{\rho}|\boldsymbol{x}, \zeta \neq 0)$ is a Dirichlet distribution, with parameter vector $\boldsymbol{a}+\boldsymbol{x}$, we have for $\zeta = \sqrt{N}\sum_{i=1}^r \gamma_i\rho_i$ that

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

And if all a_i are equal, the first vanishing moment of the dual (analysis) wavelet (**?**, page 241) annihilates this constant, reducing the expression 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 the right hand side only depends on Z, we can equivalently write: $E(\zeta|Z, \zeta \neq 0, N) = E(\zeta|\boldsymbol{X}, \zeta \neq 0) = N/(N + A)\cdot Z$, from which the lemma follows. □

Proof of Lemma 4. This is easy to verify, for instance by marginal = prior for $\boldsymbol{\rho}$ · conditional / posterior, using the fact that the posterior for $\boldsymbol{\rho}$ is also a Dirichlet distribution. □

Proof of Lemma 5. The proof follows from the rules of conditional expectation. For the variance, this leads to:

$$\mathrm{Var}(X_i|\zeta \neq 0, N) = \mathrm{Var}(E(X_i|\rho_i, \zeta \neq 0, N)) + E(\mathrm{Var}(X_i|\rho_i, \zeta \neq 0, N)) = N(N + A)\cdot \mathrm{Var}(\rho_i|\zeta \neq 0, N),$$

and a similar result holds for the covariance: $\mathrm{cov}(X_i, X_j|\zeta \neq 0, N) = N(N + A)\cdot \mathrm{cov}(\rho_i, \rho_j|\zeta \neq 0, N)$. Combining these results completes the proof. The computation of the mean is trivial. □

Proof of Proposition 3. We have for the posterior mean:

$$|Z - E(\zeta|Z,N)| = \left| Z - p^* \cdot \frac{N}{N+A} \cdot Z \right| = \frac{A}{N+A} \cdot |Z| + (1-p^*) \cdot \frac{N}{N+A} \cdot |Z|. \tag{21}$$

The value of $|Z|$ is bounded by the normalisation factor N, i.e. $|Z| \leq \|\boldsymbol{\gamma}\|_\infty \sqrt{N}$, and at the same time, the factor $N/(N+A)$ tends to one for $N \to \infty$. As a consequence, the first term in (21) is bounded by a constant: $A/(N+A) \cdot |Z| \leq \|\boldsymbol{\gamma}\|_\infty \sqrt{N}/(N+A) \leq \|\boldsymbol{\gamma}\|_\infty \sqrt{A}/2$. The second term in (21) 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 (17), which is of smaller order than the maximal value of $|z|$ for a given N. That maximum value depends linearly on N. This means that there exists a constant C^*, such that $(1-p^*) \cdot |z| < \epsilon$ if $|z| > C^* \theta_{\mathrm{BF}}$. The contribution of this second term to the total shrinkage $|Z - E(\zeta|Z,N)|$ is then bounded by $\max(\epsilon, C^* \theta_{\mathrm{BF}})$, since one can never shrink more than Z itself. □

Acknowledgement: The author thanks the editor and referees for their constructive suggestions.

References