Statistica Sinica **19** (2009), 1491-1510

A MULTISCALE VARIANCE STABILIZATION FOR BINOMIAL SEQUENCE PROPORTION ESTIMATION

Matthew A. Nunes and Guy P. Nason

Imperial College and University of Bristol

Abstract: There exist many different wavelet methods for classical nonparametric regression in the statistical literature. However, techniques specifically designed for binomial intensity estimation are relatively uncommon. In this article, we propose a new technique for the estimation of the proportion of a binomial process. This method, called the Haar-NN transformation, transforms the data to be approximately normal with constant variance. This reduces the binomial proportion problem to the usual 'function plus normal noise' regression model and thus any wavelet denoising method can be used for the intensity estimation. We demonstrate that our methodology possesses good Gaussianization and variance-stabilizing properties through extensive simulations, comparing it to traditional transformations. Further, we show that cycle-spinning can improve the performance of our technique. We also explore the efficacy of our method in an application.

Key words and phrases: Binomial random variable, Gaussianization, Haar-Fisz, sequence probability estimation, variance stabilization.

1. Introduction

Wavelet transforms are now widely used as mathematical tools for applications such as data compression, density estimation and nonparametric regression. In particular, they can be used to estimate underlying signals from noisy observations, with many of these shrinkage techniques assuming that the corrupting noise is Gaussian. For detailed discussions of the mathematical aspects of wavelets, see Mallat (1989), Daubechies (1992) and Vidakovic (1999); for thorough coverage of wavelet shrinkage estimation, see Donoho and Johnstone (1994), Donoho and Johnstone (1995) and Abramovich, Bailey and Sapatinas (2000).

This article investigates the problem of estimating the proportion parameter associated with a sequence of binomial random variables (a binomial process) using a wavelet-based transform. The usual regression model takes the following form: observe the data $\mathbf{v} = (v_0, v_1, \ldots, v_{N-1})$ at equally-spaced timepoints in the unit interval, where $N = 2^J$; assume the N observations $\{v_k\}$ are a sequence of independent binomial random variables X_k, $X_k \sim Bin(n_k, p_k)$, for $k \in \{0, \ldots, N-1\}$, where n_k is assumed known for each k. The aim is to

estimate the unknown proportion vector $\mathbf{p} = (p_0, p_1, \ldots, p_{N-1})$ from the observations $\{v_k\}$. We take $p_k = P(k/N)$ for $k \in \{0, \ldots, N-1\}$, where P denotes an underlying binomial proportion function.

In practice, this type of problem is difficult since the noise is not Gaussian, but more importantly the variance of the 'noise' depends on the mean; we have $E(X_k) = n_k p_k$ and $\mathrm{Var}\,(X_k) = n_k p_k q_k$ (with $q_k = 1 - p_k$).

One approach is to transform the data so that it is variance-stabilized and approximately normal; a denoiser suitable for Gaussian noise is then applied and the data are transformed back to obtain an estimate of the proportion. One such transform is Anscombe's inverse sine transformation (Anscombe (1948)), reviewed in the next section. Existing methodology for Haar-Fisz variance stabilization and Gaussianization has been successful for Poisson data (see for example Fryźlewicz and Nason (2004) and Nason and Bailey (2008)). The Haar-Fisz transform cannot be used directly on binomial data as the variance is not stabilized. However, we introduce a modified transform that does.

In our simulations, we will compare the algorithm with Anscombe's inverse sine transformation (Anscombe (1948)) and the Freeman-Tukey averaged inverse sine transformation (Freeman and Tukey (1950)) when investigating Gaussianizing and variance-stabilizing properties.

Our method exhibits many benefits.

- It has good Gaussianizing and variance-stabilizing properties.
- It outperforms traditional Gaussianizing transformations in difficult cases, for example, when the binomial size is small or the binomial proportion is extreme.
- It is computationally simple and easy to code.
- Since it is an effective variance-stabilizing 'Gaussianizer', a wide range of smoothing methods can be used to obtain a proportion estimate.

This article is organized as follows. Section 2 reviews estimation methods for binomial processes, including a discussion of the Haar-Fisz transform and its motivation from the Fisz transform (Fisz (1955)). Section 3 proposes a new Gaussianizing transform called the NN transform for binomially distributed random variables. Section 4 adapts our transform for use on binomial data and explores its properties. We propose a technique for proportion estimation from a binomial sequence in Section 5. Section 6 concludes and outlines ideas for further work.

2. Review of Work on Binomial Proportion Estimation

We give a brief outline of work in the literature for binomial process proportion estimation problems.

One approach to the binomial problem is to transform the observations so that the transformed data can be assumed to be (at least approximately) normally distributed. For the binomial distribution, Anscombe (1948) suggests the following: if $\{x_i\}$ are realizations from i.i.d. binomial random variables $X_i \sim Bin(n, p)$, then the transformed data $\mathcal{A}x_i = \sin^{-1}\sqrt{(x_i + c)/(n + 2c)}$ will be distributed 'more normally'. Anscombe states that the value $c = 3/8$ is optimal for μ and $n-\mu$ large (where μ is the mean of the binomial distribution). The asymptotic mean of the transform is approximately $\mathcal{A}\mu = \sin^{-1}\sqrt{(8m + 3)/(8n + 6)}$, and the variance is stabilized at $(n + 1/2)^{-1}/4$ for this value of c. Though computationally efficient, Anscombe's transformations used in conjunction with traditional wavelet methods are reported to oversmooth and to not perform well when intensities are low (Antoniadis and Sapatinas (2001)).

Freeman and Tukey (1950) discuss a similar transform, the averaged inverse sine function $\mathcal{B}x_i = \sin^{-1}\sqrt{x_i/(n + 1)} + \sin^{-1}\sqrt{(x_i + 1)/(n + 1)}$, with asymptotic mean approximately $\mathcal{B}\mu$. This transform has variance stabilization around $(n + 1/2)^{-1}$ for almost all cases when the binomial mean is at least one, though it is difficult to use as a pre- and postprocessor since it does not have a unique inverse function.

Nonparametric regression techniques for proportions usually assume that the underlying proportion function has a certain degree of smoothness. For example, recent work on generalized linear models (Hastie and Tibshirani (1990) and Fan and Gijbels (1995)) assume that the proportion function $P(x)$ satisfies $g(P(x)) = s(x)$, where g is a monotone smooth function called the *link function*; $s(x)$ is assumed smooth and estimated by methods suitable for smooth (continuous) regression functions. Different assumptions and estimation techniques, as well as link function choice, are discussed in Fan and Gijbels (1995) and Fan, Heckman and Wand (1995). For a more involved discussion of generalized linear models, see Hastie and Tibshirani (1990) and McCullagh and Nelder (1989).

Altman and MacGibbon (1998) use cross-validation for bandwidth selection in kernel estimators for either fixed or random design binary regression. The asymptotic risk of the kernel estimators has good convergence properties under certain smoothness conditions on the regression function.

Antoniadis and LeBlanc (2000) consider linear wavelet smoothers for the irregular design binary regression situation. A generalized linear model with identity link function is imposed on the regression function and, via usual wavelet projection, an estimator of the smooth model function $s(x)$ is obtained. A particular form of empirical wavelet coefficient is proposed to obtain smoother regression estimators, and the adaptive choice of resolution parameter in resulting wavelet series expansions is implemented in the binary regression context. The estimator is then modified to give a suitable estimator of the regression function $P(x)$. The estimator is shown to have good asymptotic properties.

Wavelet shrinkage is used in modulation estimator methodology by Antoniadis and Sapatinas (2001), extending the idea to obtain smooth estimates for data from exponential families with quadratic variance functions, including the binomial distribution. An estimator of the risk is formed by assuming the function estimate is a diagonal linear shrinker and using a cross-validation approach. The function estimate is then constructed using a minimizer of the risk estimate.

Sardy, Antoniadis and Tseng (2004) propose a generalization of the wavelet smoother *WaveShrink* (Donoho and Johnstone (1994)) to include a range of non-Gaussian distributions such as the binomial and the Bernoulli. The procedure uses interpoint algorithms to find the solution to a penalized log-likelihood problem based on the l^1-norm of the wavelet coefficients in a wavelet estimator representation.

Fryźlewicz and Nason (2004) introduce the Haar-Fisz transformation, combining the Haar wavelet transform and a result by Fisz (1955), which asserts the asymptotic normality of a special ratio of Poisson random variables. The result motivates the authors to propose a method for Poisson intensity estimation using the Haar-Fisz technique as a pre- and postprocessing tool for Poisson data. The algorithm consists of applying the Haar wavelet transform to count data, and then modifying the wavelet coefficients according to Fisz (1955). Inverting the Haar wavelet transform after the modification creates a variance stabilizing transform. The algorithm has been used successfully for its gaussianizing and variance stabilizing properties.

Kolaczyk and Nowak (2005) present a multiscale generalized linear model for the estimation of functions in a general one-dimensional nonparametric regression setting. Piecewise polynomials, defined on recursive partitionings of the unit interval, are used to construct estimators of the regression function, optimizing a penalized likelihood criterion to choose a piecewise polynomial fit.

All the above methods are suitable for binomial proportion estimation. However, the methods based on generalized linear models often require a link choice; others assume some degree of regularity of the underlying proportion function, or produce estimates belonging to a certain smoothness class. The use of interpoint algorithms in Sardy, Antoniadis and Tseng (2004) can be computationally expensive. Here we aim to take advantage of the computational efficiency and flexibility of transformations, such as Anscombe's, to improve performance in cases of low intensity.

3. The NN Variance-stabilizing Transform

3.1. The transform and its theoretical properties

The Haar-Fisz transform (outlined in Section 2) is motivated by the observation that applying the Fisz theorem (Fisz (1955)) to pairs of Poisson random variables results in an asymptotic normal distribution with unit variance.

However, for binomial variables, there is no choice of Fisz exponent that produces a constant asymptotic variance, so the variance *cannot* be stabilized by the usual Fisz transform (even in the limited case of equal trial probabilities and equal binomial sizes). We propose a different Gaussianizing transform, similar to the Fisz transform, with which asymptotic normality with stabilized variance can be obtained. We then use this result, similar to the Haar-Fisz technique, to propose an algorithm for binomial proportion estimation.

In our transform, we divide the Haar difference $X_2 - X_1$ by its standard error, $\sqrt{\mathrm{Var}\,(X_1) + \mathrm{Var}\,(X_2)}$. This essentially uses the observations from X_1 and X_2 as estimates for the individual binomial means $n_r p$ $(r = 1, 2)$ and combines them in the expression for the standard error. We first state our alternative theorem to the Fisz theorem (Fisz (1955)), the proof of which can be found in Appendix A of Nunes and Nason (2008).

Theorem 3.1. *Let $X_r \sim Bin(n_r, p_r)$, $r = 1, 2$, with $p_r \in (0, 1)$ (fixed). Let*

$$m_r = E(\xi_r), \quad \sigma_r^2 = \mathrm{Var}\,(\xi_r), \quad \textit{and} \quad \psi = \sqrt{\sigma_1^2 + \sigma_2^2}. \tag{3.1}$$

If X_1 and X_2 are independent and

$$\lim_{n_1, n_2 \to \infty} \frac{m_1}{m_2} = 1, \tag{3.2}$$

then

$$\zeta_B(n_1, n_2) = \frac{X_2 - X_1}{\big([(X_1 + X_2)/(n_1 + n_2)]\big(n_1 + n_2 - (X_1 + X_2)\big)\big)^{1/2}}$$

is asymptotically normal $N(m_B, \sigma_B^2)$ when $n_1, n_2 \to \infty$, where

$$m_B = \frac{m_2 - m_1}{\big([(m_1 + m_2)/(n_1 + n_2)]\big(n_1 + n_2 - (m_1 + m_2)\big)\big)^{1/2}}, \tag{3.3}$$

$$\sigma_B = \frac{\psi}{\big([(m_1 + m_2)/(n_1 + n_2)]\big(n_1 + n_2 - (m_1 + m_2)\big)\big)^{1/2}}. \tag{3.4}$$

We take ζ_B to be zero when both X_1 and X_2 are zero.

Consider the specific case $X_r \sim Bin(n_r, p)$ for $r = 1, 2$, i.e., the binomial random variables have equal trial probabilities. From Theorem 3.1, $\zeta_B(X_1, X_2)$ is asymptotically normal with mean $\big((n_2 - n_1)/(n_1 + n_2)\big)(p/q)^{1/2}$, but with *unit variance* as $n_1, n_2 \to \infty$. In other words, $\zeta_B(X_1, X_2)$ stabilizes the variance of the asymptotic distribution. Note also that if, in addition, the sample sizes are equal, the asymptotic distribution is N(0,1).

3.2. Gaussianization and variance-stabilization properties of the NN transform

In this section we demonstrate, through simulations, how well the transform ζ_B can bring binomial data closer to normality, whilst stabilizing the variance.

In some of the simulations below, we compare properties of our transform with that of Anscombe's angular transformation and the Freeman-Tukey transformation outlined in Section 2. We follow a similar approach to simulations as in Fryźlewicz and Nason (2004). However, since the size of the binomial means depends on the trial success probability, p, as well as the binomial size, n, the effect of both parameters must figure in our simulations.

Let $X_r \sim B(n, p_r)$ for $r = 1, 2$. For each experiment, we sampled 10^5 values of X_r for various binomial sizes and for each probability lattice point (p_1, p_2), where p_r ranged from 0 to 1 in steps of 0.05. The binomial samples were then used to compute 10^5 values of $\zeta_B(X_1, X_2)$, denoted $z_n(p_1, p_2)$. For comparisons with the Anscombe and Freeman-Tukey inverse sine transformations, the values of the binomial variable corresponding to the larger of the two probabilities p_r was used. Since these transformations work better for larger means,this favors the Anscombe and Freeman-Tukey transforms.

Simulations of the variance. The sample variance was computed over the 10^5 samples of ζ_B arising from the samples of X_1 and X_2 for each point (p_1, p_2). Figure 3.1 gives a series of contour plots of the sample variance for each of the binomial sizes $n = 1, 25, 100$, renormalized so that the asymptotic distributions have unit variance. The plots show a "flattening" of the surface peaks as the binomial size increases, with the variance of the peak approaching one. In fact, this feature is prominent near the line $p_1 = p_2$; this reflects the observation that equal binomial probabilities result in an asymptotic distribution with unit variance.

To further examine the case of two equal binomial proportions, we display this feature graphically for ζ_B, Anscombe's transformation $\mathcal{A}$, and the Freeman-Tukey transformation $\mathcal{B}$, on the interval $p_1 = p_2 \in (0, 1)$ for increasing n. Figure 3.2 plots the squared residual of the variance from one against the (equal) binomial proportion. From this plot, for small binomial sizes, our transform had variance closer to one for low and high proportions, especially when compared against Anscombe's transformation, although the Freeman-Tukey came close. The three were comparable in the interval (0.25,0.75). For larger n, all three transforms do well at stabilizing the variance at one.

Gaussianization simulations. For judging the relative Gaussianizing properties of the transform ζ_B, we computed the Kolmogorov-Smirnov statistics for ζ_B and for the two competitor transformations over the binomial proportion lattice. Lower Kolmogorov-Smirnov statistics are representative of samples that are

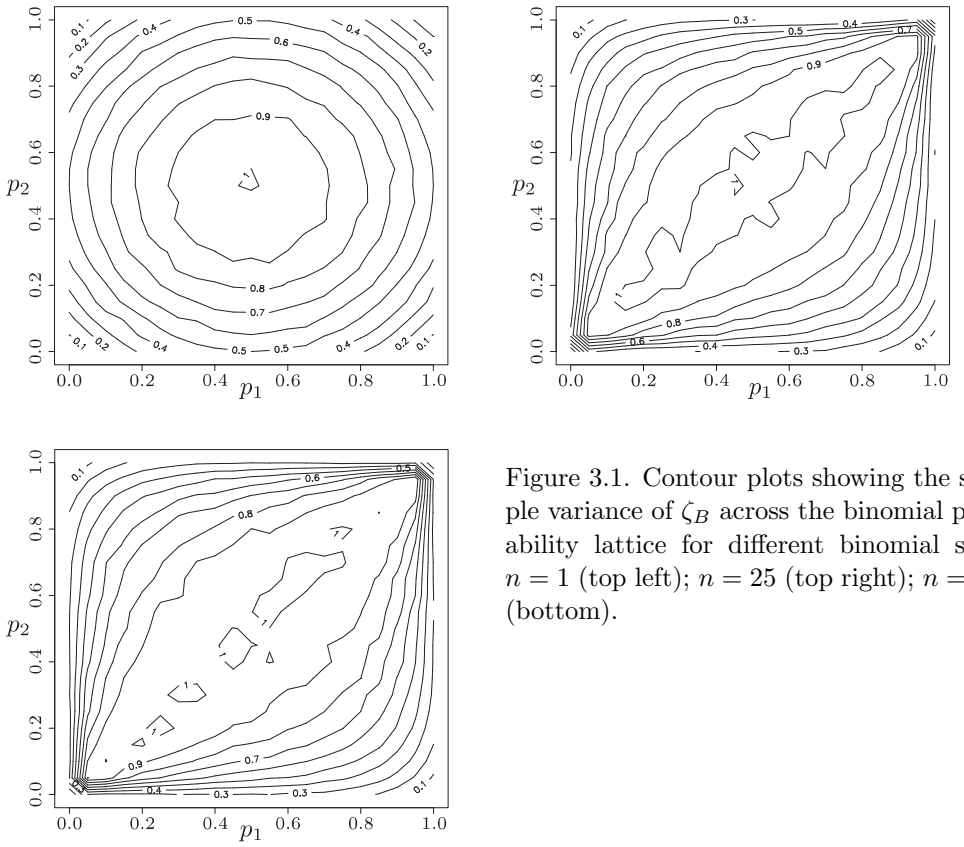


Figure 3.1. Contour plots showing the sample variance of ζ_B across the binomial probability lattice for different binomial sizes: $n = 1$ (top left); $n = 25$ (top right); $n = 100$ (bottom).

more Gaussian. Figure 3.3. shows contour plots of the difference in Kolmogorov-Smirnov statistics between Anscombe's transform and ζ_B; positive difference in these plots corresponds to our transform being more Gaussian. The corresponding plot for the difference between the Freeman-Tukey transform and ζ_B is similar.

The overall trend is that the difference in Kolmogorov-Smirnov statistics is positive for small and moderate binomial sizes, irrespective of the binomial proportions p_1 and p_2. This demonstrates that our transform had better Gaussianization properties than both the Anscombe and the Freeman-Tukey transformation. As expected, as the binomial size grew, the differences between the Kolmogorov-Smirnov statistics became negligible, due to both transforms having good Gaussianizing properties. However, examining the statistics further, the means of the statistics for ζ_B were lower compared to those of its competitors (for all values of the binomial size, n). This indicates that the transformed data using our transform was more Gaussian than those of the Anscombe or Freeman-Tukey

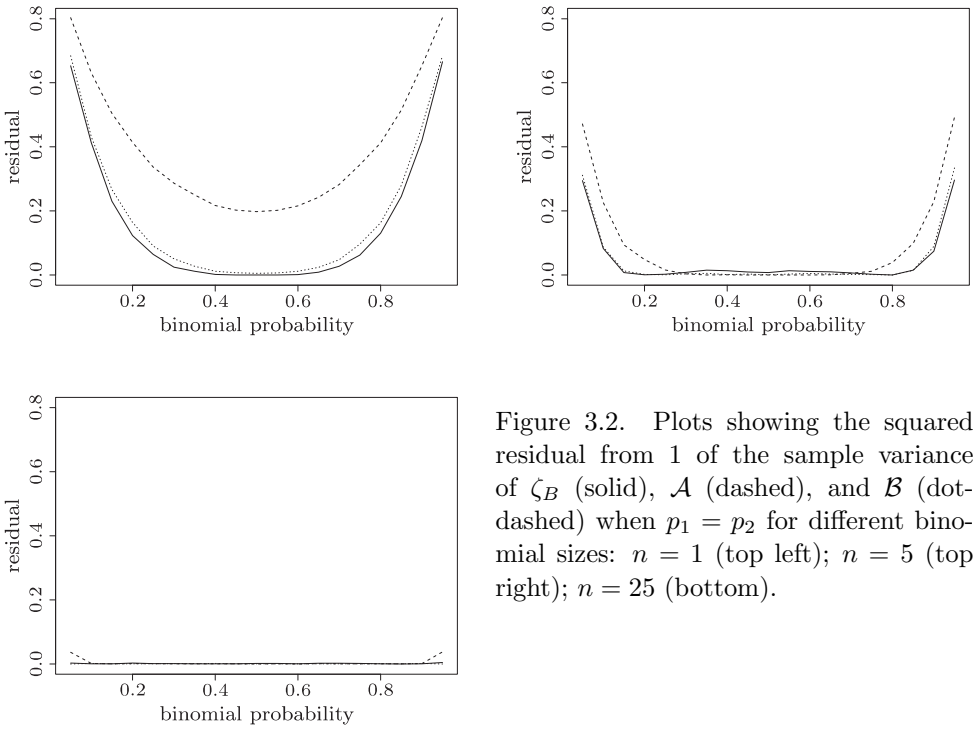


Figure 3.2. Plots showing the squared residual from 1 of the sample variance of ζ_B (solid), $\mathcal{A}$ (dashed), and $\mathcal{B}$ (dot-dashed) when $p_1 = p_2$ for different binomial sizes: $n = 1$ (top left); $n = 5$ (top right); $n = 25$ (bottom).

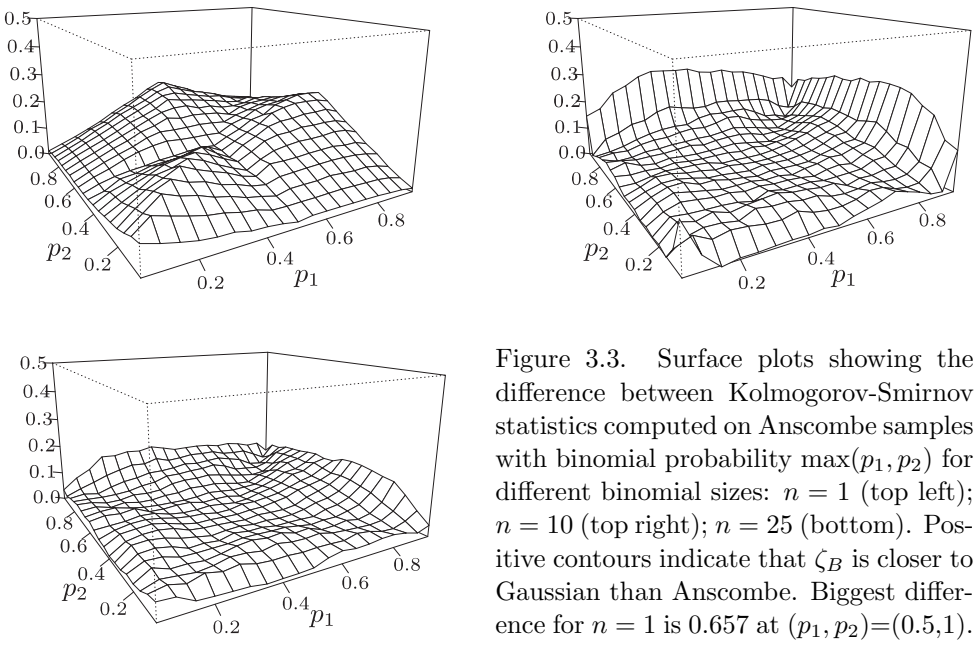


Figure 3.3. Surface plots showing the difference between Kolmogorov-Smirnov statistics computed on Anscombe samples with binomial probability $\max(p_1, p_2)$ for different binomial sizes: $n = 1$ (top left); $n = 10$ (top right); $n = 25$ (bottom). Positive contours indicate that ζ_B is closer to Gaussian than Anscombe. Biggest difference for $n = 1$ is 0.657 at (p_1, p_2)=(0.5,1).

transforms. More graphical evidence from the simulations in Section 3.2 can be found in Nunes and Nason (2008).

4. The Haar-NN Transform for Binomial Random Variables

4.1. The transform

The Haar discrete wavelet transform. The Haar-Fisz transform combines a Gaussianizing transform with the Haar discrete wavelet transform. The Haar discrete wavelet transform (DWT) is performed on an input data vector $\mathbf{v}$ by iterating the steps

$$c_{j,k} = \frac{c_{j+1,2k} + c_{j+1,2k+1}}{2} \quad \text{and} \quad d_{j,k} = \frac{c_{j+1,2k} - c_{j+1,2k+1}}{2}, \quad j = J-1, \ldots, 0.$$

The inverse DWT can be expressed in the two equations

$$c_{j+1,2k} = c_{j,k} + d_{j,k} \quad \text{and} \quad c_{j+1,2k+1} = c_{j,k} - d_{j,k}.$$

Note that the forward and inverse steps described above translate into using wavelet filters $(1,1)/2$ and $(1,-1)/2$. This differs from the Haar filters used in many descriptions of the Haar transform, that make the Haar basis orthonormal.

We now introduce an algorithm similar to the Haar-Fisz transform described in Section 2, based on the asymptotic result from the preceding section. Suppose we have an observed vector $\mathbf{v}=(v_0, v_1, \ldots, v_{N-1})$ of length $N = 2^J$, with $0 \le v_i \le n_k$, for some integers n_k.

1. Perform the Haar DWT on $\mathbf{v}$ to obtain the vector $(\boldsymbol{c_0},\boldsymbol{d_0},\boldsymbol{d_1},\ldots,\boldsymbol{d_{J-1}})$. As each level is produced, modify the coefficients as

$$f_{j,k} = \frac{d_{j,k}}{\sqrt{\big(c_{j,k}(n_{j+1,k-1} + n_{j+1,k} - 2^{J-1}c_{j,k})\big)/(n_{j+1,k-1} + n_{j+1,k})}}. \tag{4.1}$$

2. Perform the inverse Haar DWT on the vector $(\boldsymbol{c_0}, \boldsymbol{f_0}, \boldsymbol{f_1}, \ldots, \boldsymbol{f_{J-1}})$, with the result $\mathbf{u}$.

In the above, $n_{j+1,k-1}$ and $n_{j+1,k}$ are the recursive pairwise sums of the binomial sizes as the DWT levels are produced, and $f_{j,k} = 0$ if the denominator in (4.1) is zero. When $n_k = n\ \forall k$, the modification in Step 1 simplifies to

$$f_{j,k} = \begin{cases} 0 & \text{if } c_{j,k} = 0 \text{ or } c_{j,k} = n, \\ \dfrac{d_{j,k}}{\sqrt{\big(c_{j,k}(n-c_{j,k})\big)/n}} & \text{otherwise} \end{cases} \tag{4.2}$$

We denote this transform by $\mathbf{u}:=\mathcal{F}_B\mathbf{v}$. As with the usual Haar-Fisz transform, $\mathcal{F}_B$ can be inverted by "undoing" Steps 1 and 2.

Let us examine the effect of the modification in Step 1 of the above procedure. For clarity, we use the simplified modification (4.2). Consider the coefficients v_0 and v_1. The modified detail coefficient $d_{J-1,0}$ is

$$f_{J-1,0} = \frac{(v_1 - v_0)/2}{\left(\left((v_0+v_1)\left(n-\frac{v_0+v_1}{2}\right)\right)/2n\right)^{1/2}} = \frac{(v_1 - v_0)}{\left(\left((v_0+v_1)\left(2n-(v_0+v_1)\right)\right)/n\right)^{1/2}}.$$

Similarly the next coarsest level coefficient is

$$f_{J-2,0} = \frac{((v_0 + v_1) - (v_3 + v_4))}{\left(\left((v_0 + \cdots + v_3)\big(4n - (v_0 + \cdots + v_3)\big)\right)/n\right)^{1/2}}.$$

This computation is similar for every coefficient within a level, and for each DWT decomposition level. If the data vector $\mathbf{v}$ is representative of observations from i.i.d. $X_k \sim Bin(n, p)$, then the modified detail coefficients can be expressed as $f_{j,k} = 2^{-(J-j)/2}\zeta_B(Y_1, Y_2)$, where Y_1 and Y_2 are both sums of 2^{J-j-1} of the X_k, and thus are binomially distributed as well. Since the application of the inverse Haar transform is identical for $\mathcal{F}_B\mathbf{v}$ as for $\mathcal{F}\mathbf{v}$, after performing the transform $\mathcal{F}_B\mathbf{v}$, the original data can be expressed as linear combinations of quantities of the form $\zeta_B(Y_1, Y_2)$ for binomial Y_1 and Y_2. This also applies to the more general (4.1). It is analogous to the Haar-Fisz transform (see Section 2.2 in Fryźlewicz and Nason (2004)). Thus $\mathcal{F}_B\mathbf{v}$ represents a *diagonal* transformation of $\mathbf{v}$, that is, there is one transformed value for each v_i.

4.2. Finite sample Gaussianization and variance stabilization properties of the Haar-NN transform

The following investigation compares the Gaussianization and variance-stabilizing properties of the transform $\mathcal{F}_B$ introduced in Section 4, with Anscombe's transformation, the Freeman-Tukey transformation, and the identity transformation. Again, we follow an approach similar to Fryźlewicz and Nason (2004).

For these simulations, we chose a binomial proportion vector $\boldsymbol{p}$ of length $N = 1,024$, sampled from a (normalized and stretched) version of the well-known *Blocks* test signal of Donoho and Johnstone (1994) (see Figure 5.1). For each binomial size we denote by $\boldsymbol{\lambda} := n\boldsymbol{p}$ the mean intensity vector corresponding to n. It should be noted that although the mean vector depends on the binomial size, n, this is not included in the notation explicitly, since it will be obvious from the context which value of n we use. A sample path generated from binomial random variables with mean vector $\boldsymbol{\lambda}$ is denoted by $\mathbf{v}$. As expected, a sample path takes the value 1 more often when $\boldsymbol{p}$ is near 1, and hits zero more frequently when $\boldsymbol{p}$ is near zero.

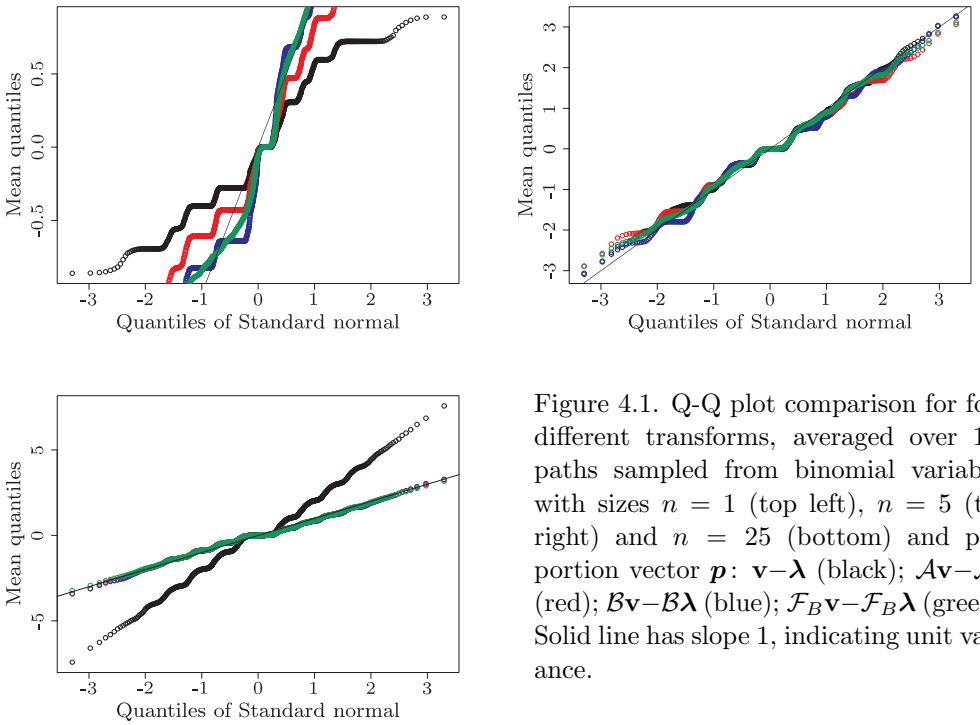


Figure 4.1. Q-Q plot comparison for four different transforms, averaged over 100 paths sampled from binomial variables with sizes $n = 1$ (top left), $n = 5$ (top right) and $n = 25$ (bottom) and proportion vector $\boldsymbol{p}$: $\mathbf{v}-\boldsymbol{\lambda}$ (black); $\mathcal{A}\mathbf{v}-\mathcal{A}\boldsymbol{\lambda}$ (red); $\mathcal{B}\mathbf{v}-\mathcal{B}\boldsymbol{\lambda}$ (blue); $\mathcal{F}_B\mathbf{v}-\mathcal{F}_B\boldsymbol{\lambda}$ (green). Solid line has slope 1, indicating unit variance.

Gaussianizing simulations. We compared the Gaussianizing properties of the different transforms by considering the Q-Q plots of $\mathbf{v}-\boldsymbol{\lambda}$ (identity transform), $\mathcal{A}\mathbf{v}-\mathcal{A}\boldsymbol{\lambda}$ (Anscombe), $\mathcal{B}\mathbf{v}-\mathcal{B}\boldsymbol{\lambda}$ (Freeman-Tukey), and $\mathcal{F}_B\mathbf{v}-\mathcal{F}_B\boldsymbol{\lambda}$ (Haar-NN), averaged over 100 sample paths. These paths were created from the mean vector $\boldsymbol{\lambda}$ for various binomial sizes. Figure 4.1 shows comparisons for the binomial sizes $n = 1, 5$ and 25.

For $n = 1$ and 2, the raw data (marked in black) was quite "stepped" since the data are discrete. The Anscombe-transformed data, and those transformed by Freeman-Tukey transformation, still exhibited this characteristic, whilst for our transform, $\mathcal{F}_B$, they lost most of this stepped character; the data were closer to a straight line, indicating more Gaussian data. Moreover, closer to the solid line (which has a slope of 1), indicates a variance of one. As n increases, the Q-Q lines become similar, although it can be said that our transform displayed slightly better Gaussianization (and also variance-stabilization), since the quantile points do not deviate from the (solid) straight line as much as the other transforms, especially at the tails. For large n, all three transforms did very well at bringing the data to normality. Furthermore, the variance was very close to one, as expected due to large n and the Central Limit Theorem.

Variance simulations To assess how well the transformations $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{F}_B$ force the data to have variance nearer to one, we plotted the squared residuals

$|\mathcal{A}\mathbf{v}-\mathcal{A}\boldsymbol{\lambda}|^2$, $|\mathcal{B}\mathbf{v}-\mathcal{B}\boldsymbol{\lambda}|^2$ and $|\mathcal{F}_B\mathbf{v}-\mathcal{F}_B\boldsymbol{\lambda}|^2$ rescaled by their respective asymptotic variances. The residuals were averaged over 1,000 sample paths generated from the mean intensity vector $\boldsymbol{\lambda}$ for a range of binomial sizes. With optimal performance, squared residuals stabilize at one when the proportion is nonzero, since the squared residuals form an estimate of the variance. Examples of the squared residuals for the three transforms are given in Figures 4.2 and 4.3 for $n = 1$ and 25.

For small n, our transform did much better than the competitors, $\mathcal{A}$ and $\mathcal{B}$, at stabilizing the sample path variances. For example, for $n = 1$, the Anscombe transform had squared residuals in the range 0.2 to 0.6, and the Freeman-Tukey transform had squared residuals in the range 0.4 to 0.9, whereas for our transform the residual was nearer 1 for most of the sample path range. Further, our transform did relatively well compared to Anscombe, and slightly better than Freeman-Tukey, when the binomial proportion was small, that is, in the three non-zero 'troughs'. However, there was a degree of erratic behaviour near the discontinuities in the proportion vector. Moderate binomial sizes had the competitor transformations beginning to achieve similar stabilization as ours; for large n, all three transforms did very well at variance stabilization, though Anscombe did slightly better in performance in this case, due to the occasional downward spikes in the Haar-NN transform (see Figure 4.3).

5. Binomial Proportion Estimation

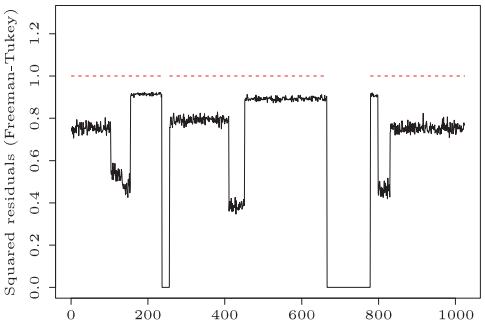
Motivated by these observations about the properties of the transform $\mathcal{F}_B$, we now propose an algorithm for probability curve estimation for a binomial sequence.

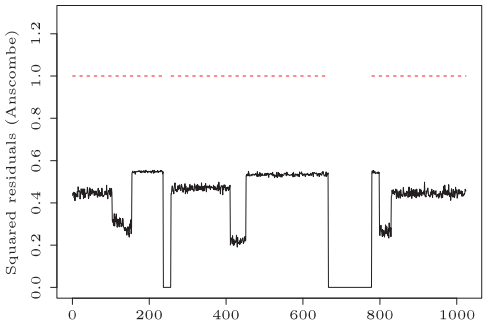
Suppose $\mathbf{v}=(v_0,\ldots,v_{N-1})$ is a vector of observations of length $N = 2^J$ from a binomial process with size n and unknown probability vector $\mathbf{p}$.

1. Perform the transform $\mathcal{F}_B$ on $\mathbf{v}$ to produce $\mathbf{u}=\mathcal{F}_B\mathbf{v}$. The vector $\mathbf{u}$ should be approximately normally distributed with constant variance.
2. Use any denoiser suitable for handling Gaussian noise with constant variance.
3. Invert the Haar-NN transform to obtain the estimate of the binomial probability vector.

5.1. Simulation study

A simulation study was performed to assess the curve estimation procedure. Several proportion functions were chosen to be estimated, each exhibiting different properties. These were the *Sinlog* function in Antoniadis and LeBlanc (2000); a scaled and reflected version of the P_2 function described in Antoniadis and LeBlanc (2000) (denoted here by P_3), and the modified *blocks* proportion





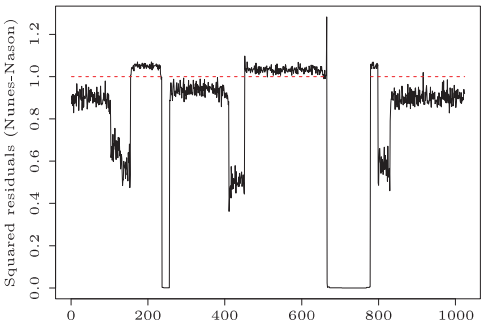
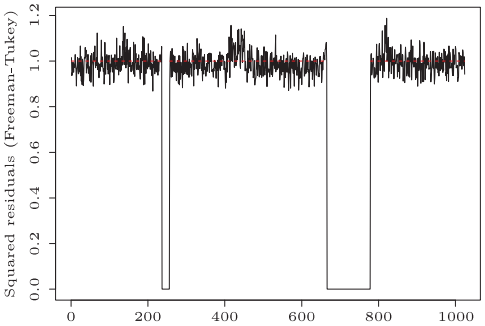
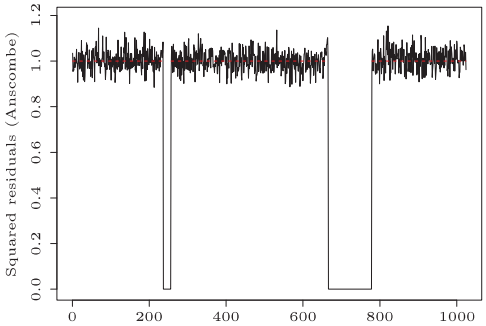


Figure 4.2. Squared residuals for different Gaussianizing transforms, averaged over 1,000 sample paths from binomial variables with size $n = 1$ and proportion vector $\boldsymbol{p}$: $|\mathcal{A}\mathbf{v}-\mathcal{A}\boldsymbol{\lambda}|^2$ (top left); $|\mathcal{B}\mathbf{v}-\mathcal{B}\boldsymbol{\lambda}|^2$ (top right); $|\mathcal{F}_B\mathbf{v}-\mathcal{F}_B\boldsymbol{\lambda}|^2$ (bottom). Dashed line shows ideal (unit) residual where intensity $\in$ (0,1).





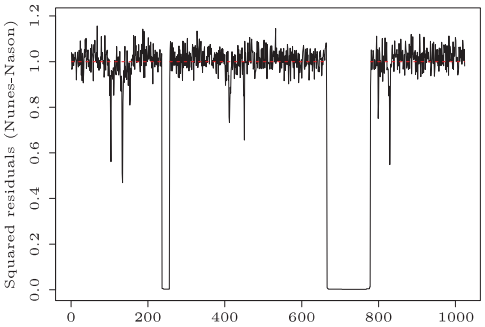


Figure 4.3. Squared residuals for different Gaussianizing transforms, averaged over 1,000 sample paths from binomial variables with size $n = 25$ and proportion vector $\boldsymbol{p}$: $|\mathcal{A}\mathbf{v}-\mathcal{A}\boldsymbol{\lambda}|^2$ (top left); $|\mathcal{B}\mathbf{v}-\mathcal{B}\boldsymbol{\lambda}|^2$ (top right); $|\mathcal{F}_B\mathbf{v}-\mathcal{F}_B\boldsymbol{\lambda}|^2$ (right). Dashed line shows ideal (unit) residual where intensity $\in$ (0,1).

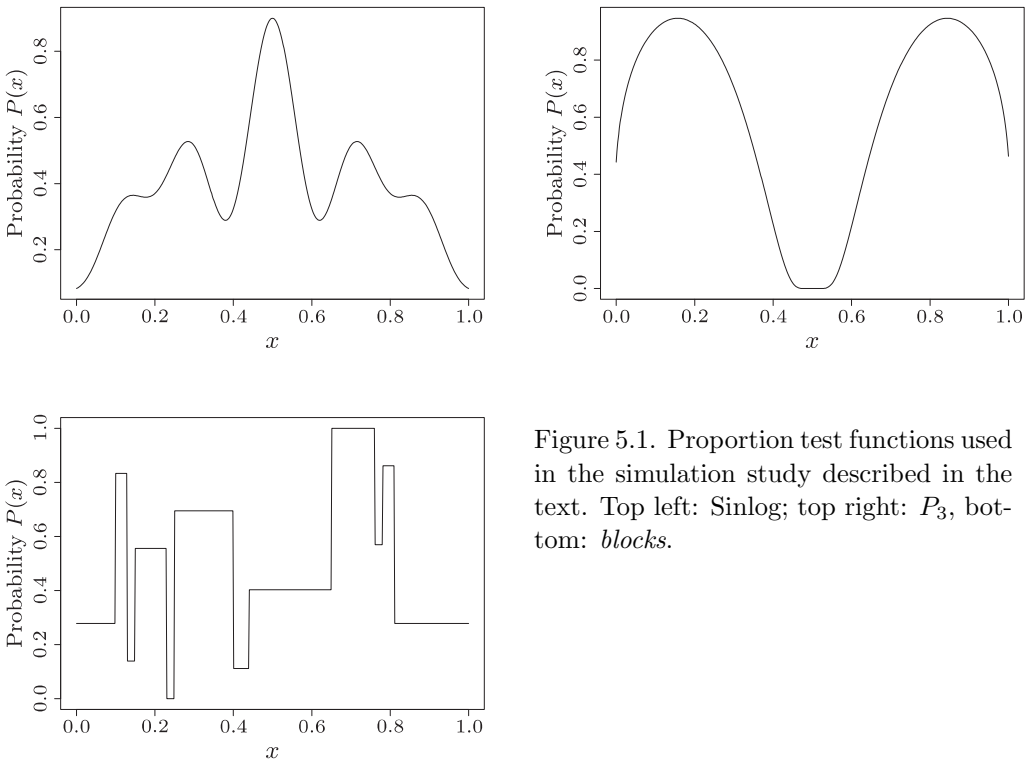


Figure 5.1. Proportion test functions used in the simulation study described in the text. Top left: Sinlog; top right: P_3, bottom: *blocks*.

from Section 4.2. These test functions are shown in Figure 5.1. More details of these functions can be found in Nunes and Nason (2008).

These functions were sampled on regular grids of length $N = 128, 256, 512$, and $1,024$. The sampled vectors were then used to create binomial sample paths using the sample vectors to define the binomial trial probabilities, i.e., $p_i = P_j(t_i)$ for $i = 1, \ldots, n$ and each proportion function P_j (*Sinlog*, P_3 and *blocks*). For each grid length/binomial size combination, 1,000 sample paths were created. These sample paths were then denoised using the estimation procedure described at the beginning of this section (transform-denoise-invert) with both $\mathcal{F}_B$ and $\mathcal{A}$ as pre- and post-processors in Steps 1 and 3 of the procedure (note that in the rest of this section we abuse notation by keeping the same symbols $\mathcal{F}_B$ and $\mathcal{A}$ for the estimation procedure as in Section 4). In the denoising step, the DWT was used with Daubechies' Least Asymmetric wavelet with 8 vanishing moments and universal soft thresholding. A comparison was also made to the wavelet shrinkage method of Antoniadis and Sapatinas (2001), denoted *AS*. All methods were optimized over the resolution level.

Cycle-spinning. We also implemented a variant of our method using cycle-spinning. Since the Haar-NN transform is translation invariant, cycle-spinning can be used to gain performance improvement (Fryźlewicz and Nason (2004)).

Table 5.1. AMSE ($\times 10^4$) simulation results for binomial size regimes and test functions described in the text, signal length $N = 256$: $\mathcal{A}$ (Anscombe); $\mathcal{F}_B$ (Haar-NN); AS (Antoniadis-Sapatinas); $\mathcal{F}_B(CS)$ (Haar-NN with cycle-spinning).

Binomial	*P3*				*Blocks*				*Sinlog*			
regime	$\mathcal{F}_B$	$\mathcal{A}$	AS	$\mathcal{F}_B$ (CS)	$\mathcal{F}_B$	$\mathcal{A}$	AS	$\mathcal{F}_B$ (CS)	$\mathcal{F}_B$	$\mathcal{A}$	AS	$\mathcal{F}_B$ (CS)
1	9.3	11.7	5.5	7.6	32.8	33.4	30.8	30.5	12.6	12.9	11.4	11.7
5	2.1	2.3	2.0	1.7	12.4	13.5	20.6	11.7	2.7	3.2	3.0	2.4
10	1.3	1.4	1.3	1.0	7.9	8.3	15.7	7.5	1.4	1.6	1.5	1.2
r_1	66.6	54.8	1518.6	54.1	91.8	77.9	1122.7	75.0	96.0	80.4	942.8	75.3
r_2	34.1	26.5	105.4	13.9	48.6	39.9	86.2	32.2	47.0	31.1	66.7	15.0

Table 5.2. AMSE ($\times 10^4$) simulation results for binomial size regimes and test functions described in the text, signal length $N = 1,024$: $\mathcal{A}$ (Anscombe); $\mathcal{F}_B$ (Haar-NN); AS (Antoniadis-Sapatinas); $\mathcal{F}_B(CS)$ (Haar-NN with cycle-spinning).

Binomial	*P3*				*Blocks*				*Sinlog*			
regime	$\mathcal{F}_B$	$\mathcal{A}$	AS	$\mathcal{F}_B$ (CS)	$\mathcal{F}_B$	$\mathcal{A}$	AS	$\mathcal{F}_B$ (CS)	$\mathcal{F}_B$	$\mathcal{A}$	AS	$\mathcal{F}_B$ (CS)
1	3.3	5.7	2.0	2.9	16.2	16.9	22.8	14.2	3.3	3.6	3.4	3.0
5	0.7	0.9	0.7	0.7	6.7	7.6	15.4	6.0	0.7	1.0	0.9	0.6
10	0.5	0.5	0.5	0.4	4.4	4.7	11.5	4.0	0.4	0.5	0.5	0.3
r_1	67.0	55.4	1303.6	54.5	90.2	77.0	1300.4	65.1	95.9	80.5	1053.5	65.4
r_2	34.6	26.8	129.7	12.0	45.7	33.4	86.9	18.8	47.1	31.3	68.6	10.1

In this case, the binomial vector is shifted before denoising and the estimate is shifted back afterwards. Estimates from different shifts are then averaged to obtain an overall estimator of the proportion function. We used 50 shifts as suggested in Fryźlewicz and Nason (2004). Performing cycle-spinning with Anscombe's transformation would not give any performance gain since this transformation commutes with the shift operator. The cycle-spinning variant is denoted by $\mathcal{F}_B(CS)$.

For each method, and for different binomial sizes and signal lengths, the averaged mean square error (x 10^4) between the estimates and the sampled proportion function were recorded. For the binomial size regimes, we used different *fixed* binomial sizes of $n_k \equiv n \equiv 1, 5, 10$, as well as randomly generated binomial sizes; r_1 then indicates where binomial sizes were uniformly generated from $n_k \in \{1, 5, 10\}$, and r_2 denotes random generation of binomial sizes from $n_k \in \{2, 4, 6\}$. For brevity, only signal lengths $N = 256$ and $N = 1,024$ are shown in Tables 5.1 and 5.2.

The results of the simulation study are very encouraging. When the binomial size was fixed, our algorithm always outperformed Anscombe, and quite often beat the Antoniadis and Sapatinas (2001) method. For the randomly-generated

binomial sizes (regimes r_1 and r_2), the AS technique did not perform well, and Anscombe performed better than our regular method. However, the performance gain from using our method combined with cycle-spinning for these regimes was clear: there was an improvement over our regular method and over Anscombe. The relative performance of the Haar-NN transform seemed to increase as the signal length increased. Initial investigation into other thresholding techniques indicate that there could be further improvements with our method.

5.2. Application: DNA Isochore detection

There has been substantial work in the field of bioinformatics in recent years, and the quest to improve existing methods and computational techniques is of great importance, in particular, for DNA sequencing and gene expression data. One important problem is the modelling and prediction of isochore clusters in DNA sequence data (Bernardi (2000)); in this section we apply the Gaussianizing and variance stabilizing properties of the Haar-NN method to it.

Biological background to the isochore problem. DNA sequences are strings (polymers) of nucleotides, chemical compounds which play important biological rôles. Each nucleotide is characterized by its nitrogen base, represented by a letter: A (adenine); C (cytosine); G (guanine); and T (thymine). These four nucleotide bases come from two compound *base pair* groups, namely *purines* (adenine and thymine) and *pyrimidines* (cytosine and guanine), differing in structure. For a more detailed discussion of the structure of DNA, see any introductory text on genomics, for example Brown (2002). G+C content can be seen as the ratio between the number of pyrimidine nucleotides to the total number of nucleotides in a DNA segment.

A school of thought in bioinformatics accepts an *isochore* model for DNA that asserts that chromosome DNA sequences are mosaics of long DNA segments of (fairly) homogenous G+C content in adjacent segments (Oliver, Carpena, Hackenberg and Bernaola-Galvan (2004)). Under this model, the G+C content mosaics differ for different organisms, especially between warm- and cold-blooded vertebrates (Bernardi (2000)); so prediction is of obvious interest, for example, in organism classification applications.

IsoFinder: an existing approach to the isochore problem. In Oliver, Carpena, Hackenberg and Bernaola-Galvan (2004) and Zhang and Chen (2004), sequential hypothesis testing is used to model the distribution of G+C cluster sizes of a DNA sequence.

The procedure works as follows. The G+C content of the sequence is counted, and a t-statistic is used to assess the significance of the difference in mean G+C values on either side of a sliding pointer moving along the DNA sequence. After heterogeneity is filtered out, the information is used to split the original sequence

into two distinct regions of differing G+C mean value. This is then repeated on successive blocks until the original sequence is divided into a number of regions with significantly different mean G+C levels. These clusters are predictions of isochores of the original DNA sequence, and the method is known as the *IsoFinder* procedure.

Haar-NN transform approach to the isochore problem. Consider a DNA sequence. Since we are interested in the sections of the strand containing G+C content, we can view the DNA section as a binary sequence with a corresponding sequence of indicator values at each nucleotide site showing whether or not a particular nucleotide comes from the pyrimidine (G or C) base pair; for an unseen strand, if we assume each molecule along the sequence is from one of the two nucleotide base pairs independently, we can assign Bernoulli random variables at the nucleotide sites. Suppose we have a DNA sequence of length $n = 2^J$. Let X_k indicate the type of nucleotide k. Then $X_k \sim$Bernoulli(p_k) and $\mathbb{P}$(nucleotide k has G+C content)$= \mathbb{P}(X_k = 1) = p_k$, $\mathbb{P}$(nucleotide k has A+T content) $= \mathbb{P}(X_k = 0) = 1 - p_k = q_k$. Estimating equal p_k for long consecutive sequences of k indicates regions of equal G+C content, and is representative of an isochore.

Example. To test the G+C proportion estimation procedure, a chromosome strand was acquired from the Wellcome Trust Sanger Institute Human Genome Sequencing Group, namely chromosome 20 of the human genome (available online from the website `http://www.sanger.ac.uk/HGP/`). To make it feasible to process these data with our method, the sequence strands were cropped to $2^{21} = 2,097,152$ bases, and then converted into binary sequences indicating G+C content as outlined above.

In the denoising step of the algorithm in Section 5, we used the Haar DWT with *Sureshrink* thresholding (Donoho and Johnstone (1995)), with primary resolution level 3. However, we modified the smoothing procedure. Recall that in the IsoFinder procedure there is an in-place heterogeneity filtering. This is usually applied to filter out isochores of less than 3 kilobases from the resulting isochore maps, so that these map estimates resemble mammalian genomes (Oliver et al. (2004)). To mimic this filtering, in the denoising step of the procedure we set the finest 11 detail coefficient levels to zero (after thresholding) before inverting the discrete wavelet transform. This has the effect of ensuring that isochore regions of less than $2^{11} = 2,048$ bases do not feature in our estimates of G+C content produced after inversion of the wavelet transform.

As a comparison to our procedure, the IsoFinder method was applied to the cropped nucleotide sequence using the online IsoFinder implementation (which can be found at `http://bioinfo2.ugr.es/IsoF/isofinder.html`). Figure 5.2 was created using this web interface.

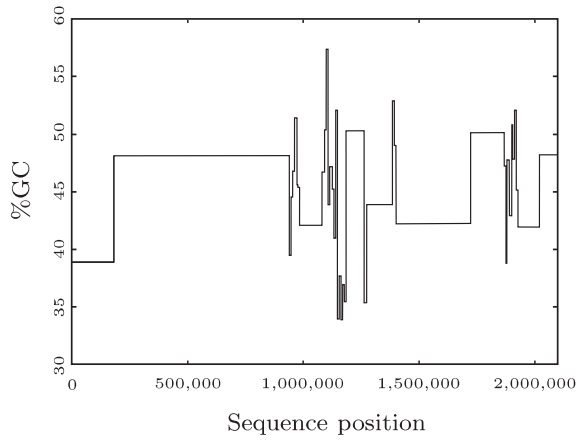


Figure 5.2. Isochore map of chromosome 20 of the human genome, as estimated by the Isofinder procedure (with 3 kilobase filtering).

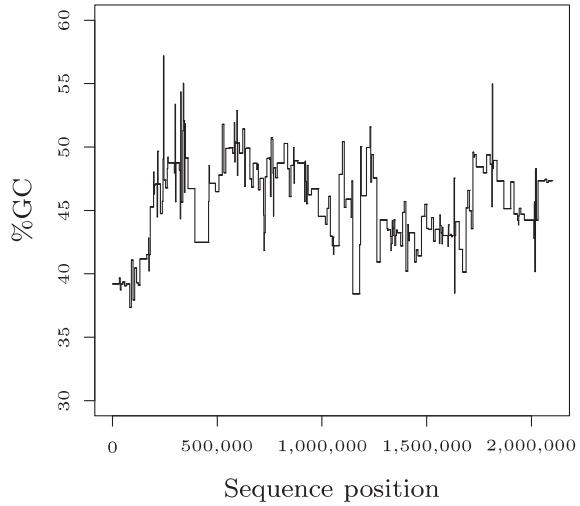


Figure 5.3. Isochore map of chromosome 20 of the human genome, as estimated by our Haar-Fisz Gaussianizing procedure (with 11 finest detail coefficient levels set to zero).

Figures 5.2 and 5.3 give estimates for the (unknown) isochore profile of chromosome 20 for the two procedures. Whilst the estimates produced using our method are more "spiky" and show shorter isochore regions, the estimates for both procedures exhibit similar overall features. It should be noted here that our estimates use *SureShrink* thresholding, with no consideration for the effect of the primary resolution level. More complex thresholding procedures could produce more homogeneous estimates. Also, our method uses a low kilobase filtering

compared to the IsoFinder procedure (due to being constrained to a power of two) and so is more likely to produce estimates which exhibit less homogeneity.

6. Conclusions

We proposed a transform, that possesses variance-stabilizing properties for binomial random variables. An asymptotic result was established about this transform, and simulations for different binomial sizes and probabilities were performed to investigate how well it Gaussianizes and stabilizes the variance compared to current normalizing transforms. The results indicate that our transform does very well for smaller binomial sizes, n, and/or for extreme binomial proportions.

Section 4 introduced a new modified Haar transform using our Gaussianizing transform. This was compared to the Anscombe transform and was found to outperform the traditional transformation for smaller binomial sizes and/or binomial proportions nearer the boundaries of the interval (0,1). This improvement is important since, in practice, large binomial sizes and "nice" success probabilities could be unrealistic. This evidence of good properties lead us to suggest an algorithm for binomial proportion curve estimation. Investigations show error improvements over competitors, especially when adopting cycle-spinning, with better performance in all but a few cases. The technique was then applied to isochore prediction.

Software code that implements our Haar-NN transform is freely available at the CRAN R software archive as an R package. It can also be found at `http://www.stats.bris.ac.uk/~maman/computerstuff/Binfisz.html`.

Acknowledgement

Nunes and Nason were both partially supported by EPSRC Grant D005221/1 and the UK Government.

References

Abramovich, F., Bailey, T. C. and Sapatinas, T. (2000). Wavelet analysis and its statistical applications. *J. Roy. Statist. Soc. D* **49**, 1–29.

Altman, N. S. and MacGibbon, B. (1998). Consistent bandwidth selection for kernel binary regression. *J. Stat. Planning and Inf.* **70**, 121–137.

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

Antoniadis, A. and LeBlanc, F. (2000). Nonparametric wavelet regression for binary response. *Statistics* **34**, 183–213.

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

Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. *Gene* **241**, 3–17.

Brown, T. A. (2002). *Genomes* (2nd edn). BIOS Scientific, Oxford.

Daubechies, I. (1992). *Ten Lectures On Wavelets*. SIAM, Philadelphia.

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

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

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Stat. Soc. B* **57**, 371–394.

Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Stat. Assoc.* **90**, 141–150.

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

Freeman, M. F. and Tukey, J. W. (1950). Transformations related to the angular and the square root. *Ann. Math. Stat.* **21**, 607–611.

Fryźlewicz, P. and Nason, G. P. (2004). A Haar-Fisz algorithm for poisson intensity estimation. *J. Comp. Graph. Stat.* **13**, 621–638.

Hastie, T. and Tibshirani, R. (1990). Generalized additive models. *Stat. Sci.* **1**, 297–318.

Kolaczyk, E. D. and Nowak, R. D. (2005). Multiscale generalised linear models for nonparametric function estimation. *Biometrika* **92**, 119–133.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn. Anal. Mach. Intell.* **11**, 674–693.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Nason, G. P. and Bailey, D. (2008). Estimating the intensity of conflict in Iraq. *J. Roy. Stat. Soc. A* **171**, 899–914.

Nunes, M. A. and Nason, G. P. (2008) A multiscale variance stabilization for binomial sequence proportion estimation. *Tech. Rep. 08:03*, Statistics Group, Department of Mathematics, University of Bristol, UK.

Oliver, J. L., Carpena, P., Hackenberg, M. and Bernaola-Galvan, P. (2004). Isofinder: computational prediction of isochores in genome sequences. *Nucleic Acids Research* **32**, w287-w292.

Sardy, S., Antoniadis, A. and Tseng, P. (2004). Automatic smoothing with wavelets for a wide class of distributions. *J. Comp. Graph. Stat.* **13**, 399–421.

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

Zhang, L. and Chen, J. (2004). Scaling behaviors of cg clusters in coding and noncoding dna sequences. *Chaos, Solitons and Fractals* **24**, 115–123.

Department of Epidemiology and Public Health, Imperial College, Norfolk Place, London W2 1PG, UK.

E-mail: m.nunes@imperial.ac.uk

School of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK.

E-mail: G.P.Nason@bristol.ac.uk

(Received January 2008; accepted April 2008)