A multiscale variance stabilization for binomial sequence proportion estimation

Matthew A. Nunes* and Guy P. Nason
Department of Mathematics, University of Bristol, UK

January 29, 2008

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. We also explore the efficacy of our method in real applications.

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); Nason and Silverman (1994); Vidakovic (1999); for thorough coverage of wavelet shrinkage estimation, see Donoho and Johnstone (1994, 1995); Abramovich *et al.* (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: we observe the data, $\mathbf{v} = (v_0, v_1, \ldots, v_{N-1})$ at equally-spaced timepoints assumed to be in the unit interval, where $N = 2^J$. Our assumption is that the N observations $\{v_k\}$ are modelled as a sequence of binomial random variables X_k, where we assume the variables to be independent: $X_k \sim Bin(n_k, p_k)$ for $k \in \{0, \ldots, N-1\}$. Our aim is to try

*Corresponding author: Matt.Nunes@bristol.ac.uk

and estimate the proportion vector $\mathbf{p} = (p_0, p_1, \ldots, p_{N-1})$ from the observations $\{v_k\}$. We assume $p_k = P(k/N)$ 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, unlike the Gaussian situation: we have $\mathrm{var}(X_k) = n_k p_k (1 - p_k)$ and $\mathbb{E}(X_k) = n_k p_k$. 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 is 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 and χ^2 data in Fryźlewicz and Nason (2004, 2006). 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 also the Freeman-Tukey averaged inverse sine transformation (Freeman and Tukey, 1950) when investigating Gaussianizing and variance-stabilizing properties.

Our method exhibits many benefits, namely:

1. It is shown to possess good Gaussianizing and variance-stabilizing properties;
2. It outperforms traditional Gaussianizing transformations in difficult cases, for example, when the binomial size is small or the binomial proportion is extreme;
3. It is computationally simple and easy to code;
4. 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) in Section 2.4. Section 3 proposes a new Gaussianizing transform called the NN transform for binomially distributed random variables. Section 4 adapts our new transform for use on binomial data and explores its properties. We also 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 now give a brief outline of work in the literature for binomial process proportion estimation problems.

2.2 Wavelet methods for binomial processes

Antoniadis and LeBlanc (2000) considers 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 (see Section 2.3). A particular form of empirical wavelet coefficient is proposed to obtain smoother regression estimators than other coefficient estimators. The adaptive choice of resolution parameter

in resulting wavelet series expansions is implemented in the binary regression context by generalizing existing selection criteria. 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 and is computationally faster than traditional local polynomial estimators.

Wavelet shrinkage is used in the 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 to be a diagonal linear shrinker and using a cross-validation approach. The function estimate is then constructed using a minimizer of the risk estimate.

Sardy *et al.* (2004) proposes a generalization of the *WaveShrink* wavelet smoother (Donoho and Johnstone, 1994) to include a range of non-Gaussian distributions such as the binomial and Bernoulli distributions. 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.

2.3 Other techniques for binomial processes

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); Fan and Gijbels (1995) assume that the proportion function $P(x)$ follows the relation

$$g(P(x)) = s(x),$$

where g is a monotone smooth function called the *link function*, and $s(x)$ is a smooth function which is estimated by methods suitable for smooth (continuous) regression functions.

Different assumptions and estimation techniques for $s(x)$, and also link function choice are discussed in Fan and Gijbels (1995); Fan *et al.* (1995). For a more involved discussion of generalized linear models, see for example Hastie and Tibshirani (1990); McCullagh and Nelder (1989).

Antoniadis and LeBlanc (2000), mentioned in Section 2.2, uses a generalized linear model construction for their wavelet regression technique.

Kolaczyk and Nowak (2005) presents 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.

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

As previously mentioned, another 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. Suppose $\{x_i\}$ are realizations from i.i.d. binomial random variables $X_i \sim Bin(n, p)$. Then the transformed data given by

$$\mathcal{A}x_i = \sin^{-1}\sqrt{\left(\frac{x_i + c}{n + 2c}\right)} \tag{1}$$

will be distributed 'more normally'. Anscombe states that the value $c = \frac{3}{8}$ is optimal for μ and $n - \mu$ large (where μ is the mean of the binomial distribution). The variance will be stabilized at $\frac{1}{4}(n + \frac{1}{2})^{-1}$ for this value of c.

Donoho (1993) uses Anscombe's similar result for Poisson data, applying it to low light photon counts. Though computationally efficient, Anscombe's transformations used in conjunction with such traditional wavelet methods are reported to oversmooth and not perform well when intensities are low (Antoniadis and Sapatinas, 2001).

Freeman and Tukey (1950) discusses a similar transformation for binomial data which takes the form of an averaged inverse sine function:

$$\mathcal{B}x_i = \sin^{-1}\sqrt{\left(\frac{x_i}{n+1}\right)} + \sin^{-1}\sqrt{\left(\frac{x_i+1}{n+1}\right)}. \tag{2}$$

This is said to have variance stabilization around $(n+\frac{1}{2})^{-1}$ 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.

All of the above methods are suitable for binomial proportion estimation. However, the methods based on generalized linear models often have the decision of link choice to make; 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 *et al.* (2004) can be computationally expensive. The aim of the method presented in this paper is to take advantage of the computational efficiency and flexibility of transformations such as Anscombe but improve performance in cases of low intensity.

2.4 The Haar-Fisz transformation

In this section, we give a brief overview of the Haar-Fisz transform, introduced in Fryźlewicz and Nason (2004).

2.4.1 The Haar discrete wavelet transform

The Haar-Fisz transform combines a Gaussianizing transform with the Haar discrete wavelet transform. We now give the fast computational description of this wavelet transform by Mallat (1989).

The Haar discrete wavelet transform (DWT) is performed on an input data vector **v** by iterating the steps

$$c_{j,k} = (c_{j+1,2k} + c_{j+1,2k+1})/2$$
$$d_{j,k} = (c_{j+1,2k} - c_{j+1,2k+1})/2,$$

for $j = J-1, \ldots, 0$. [1]

The inverse DWT can be expressed in the two equations

$$\begin{aligned} c_{j+1,2k} &= c_{j,k} + d_{j,k} \\ c_{j+1,2k+1} &= c_{j,k} - d_{j,k}. \end{aligned}$$

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

2.4.2 The Fisz transform

The properties of the Haar-Fisz transform follow from a result by Fisz (1955), which asserts the asymptotic normality of a special ratio of random variables under certain conditions. We now give some notation which are used in the Fisz theorem and which we will use later.

Let $\xi_1(\lambda_1)$ and $\xi_2(\lambda_2)$ be two independent non-negative random variables based on distributions with respective parameters λ_r. We denote, for $r = 1, 2$,

$$m_r = \mathbb{E}(\xi_r), \quad \sigma_r^2 = \mathrm{var}(\xi_r), \quad \text{and} \quad \psi = \sqrt{\sigma_1^2 + \sigma_2^2}. \tag{3}$$

Theorem 1. *(Fisz, 1955). If*

(a) the variable $\xi(\lambda)/m(\lambda)$ converges in probability to the number 1,

(b) the variable $\xi(\lambda)$ is asymptotically normal $N(m(\lambda), \sigma^2(\lambda))$,

(c) the variables ξ_1 and ξ_2 are independent and

$$\lim_{\substack{\lambda_1 \to \infty \\ \lambda_2 \to \infty}} \frac{m_1}{m_2} = 1, \tag{4}$$

then the variable

$$\zeta^\alpha(\lambda_1, \lambda_2) = \frac{\xi_2 - \xi_1}{(\xi_2 + \xi_1)^\alpha}, \tag{5}$$

where α is an arbitrary positive number, is asymptotically normal

$$N\left(\frac{m_2 - m_1}{(m_1 + m_2)^\alpha}, \frac{\psi^2}{(m_1 + m_2)^{2\alpha}}\right), \quad \textit{when } \lambda_1 \to \infty, \lambda_2 \to \infty. \tag{6}$$

(We use the convention here that if ξ_1 and ξ_2 are both zero, then ζ^α takes the value zero as well).

We now explicitly define the *Fisz transform* of two random variables.

Definition 2. *The* **Fisz transform with exponent $\boldsymbol{\alpha}$** *of two non-negative random variables X_1 and X_2 is*

$$\zeta^\alpha(X_1, X_2) = \frac{X_2 - X_1}{(X_1 + X_2)^\alpha},$$

with the convention that $0/0 = 0$.

2.4.3 The Haar-Fisz transform

Suppose a positive data vector, **v**, of length $N = 2^J$ has been observed. The Haar-Fisz algorithm proposed in Fryźlewicz and Nason (2004) is as follows:

1. Perform the Haar discrete wavelet transform on the data, to transform **v=$\boldsymbol{c_J}$** into ($\boldsymbol{c_0}$,$\boldsymbol{d_0}$,$\boldsymbol{d_1}$,. . . ,$\boldsymbol{d_{J-1}}$), where as usual, $\boldsymbol{c_0}$ denotes the smooth component and the $\boldsymbol{d_j}$ represent the detail components in the transform. However, as each level is computed, perform the modification

$$f_{j,k} = \begin{cases} 0 & \text{if } c_{j,k} = 0, \\ d_{j,k}/\sqrt{c_{j,k}} & \text{otherwise} \end{cases} \tag{7}$$

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

These two steps are known as the *Haar-Fisz transform* of **v**. We denote the transform as an operator by $\mathbf{u}=\mathcal{F}\mathbf{v}$. Note that these steps can be easily inverted.

Fryźlewicz and Nason (2004) use the above observation to prove the following result for Poisson observations:

Proposition 3. *If* **v** *is a sequence of observations (of length N) of i.i.d Poisson random variables with mean* λ*. Let* $\mathbf{u}=\mathcal{F}\mathbf{v}$ *be the Haar-Fisz transform of* **v***. Then* $\forall k \in \{0, \ldots, N-1\}$

$$u_k - \lambda = \nu + Y_k,$$

where $\nu \to 0$ *as* $\lambda/k \to 0$ *and* $Y_k \to N(0,1)$ *as* $(\lambda, k) \to (\infty, \infty)$.

In other words, (asymptotically) the vector **u** is just the Poisson intensity, λ, with additional Gaussian noise. This result motivates Fryźlewicz and Nason (2004) to propose a method for Poisson intensity estimation as follows:

1. Perform the Haar-Fisz transform on a vector of Poisson observations, **v**, to bring the data closer to normality.

2. Use any denoiser suitable for Gaussian noise.

3. Invert the Haar-Fisz transform to obtain the estimate of the Poisson intensity.

The argument for the modification to the detail coefficients in the Haar-Fisz transform needs a few words of explanation.

Applying the Fisz transform to Poisson random variables $\xi_r(\lambda_r) \sim Poi(\lambda_r)$, $r = 1, 2$ the asymptotic distribution in (6) becomes

$$N\left(\frac{\lambda_2 - \lambda_1}{(\lambda_1+\lambda_2)^{\alpha}}, \frac{\lambda_1+\lambda_2}{(\lambda_1+\lambda_2)^{2\alpha}}\right), \tag{8}$$

when $\lambda_1 \to \infty, \lambda_2 \to \infty$.

The choice of $\alpha = 1/2$ in equation (8) demonstrates the variance stabilizing property of the Fisz transform – it causes the asymptotic normal distribution in (6) to have unit variance for Poisson random variables.

The Haar-Fisz transform is motivated by this observation: Fryźlewicz and Nason (2004) shows that each of the transformed points $\mathbf{u}=\mathcal{F}\mathbf{v}$ are expressed in terms of the original observations as linear combinations of ratios of the form of the Fisz theorem, with $\alpha = 1/2$ (see equations (9) – (16) in Section 2.2 of Fryźlewicz and Nason (2004)). Hence if the observations $\{v_k\}$ come from independent Poisson random variables, then all the terms in (7) will be approximately normal, provided that the Poisson means satisfy the conditions at the beginning of Theorem 1. Fryźlewicz and Nason (2006) also take advantage of the variance-stabilizing properties of the Fisz transform (with exponent $\alpha = 1$) for χ^2 random variables.

2.4.4 Fisz-transformed binomial random variables

Let us apply the Fisz theorem to two binomial random variables, as in Fisz (1955). Suppose $X_1 \sim Bin(n_1, p_1)$ and $X_2 \sim Bin(n_2, p_2)$ are independent random variables. Assuming that condition (c) holds with $m_r = n_r p_r$ for $r = 1, 2$, Fisz notes that the hypothesis of equal binomial probabilities p_r can be tested for binomial random variables. In this case ($p_1 = p_2 = p$), the asymptotic Gaussian distribution reduces to

$$N\left(\frac{n_2 - n_1}{(n_1 + n_2)^{\alpha}} p^{1-\alpha}, \frac{(1-p)}{(n_1 + n_2)^{2\alpha - 1}} p^{1-2\alpha}\right). \tag{9}$$

If we further impose that the random variables X_1 and X_2 have equal size, i.e. $n_1 = n_2$, this asymptotic distribution simplifies further to

$$N\left(0, \frac{(1-p)}{(2n)^{2\alpha-1}} p^{1-2\alpha}\right).$$

Note that even in this special case, the variance function of the asymptotic normal distribution depends on p; there is no choice of α in (9) which produces an asymptotic variance constant in p and so the variance *cannot* be stabilized by the usual Fisz transform (with any exponent).

For the Haar-Fisz transform to be effective for binomial variables, we would like to have variance stability on each decomposition level j in equation (7).

Unfortunately this cannot be achieved for binomial random variables with the Fisz transform, unlike the case of Poisson variables. Hence we propose a different Gaussianizing transform, similar to the Fisz transform, with which asymptotic normality with stabilized variance can be obtained.

3 The NN variance-stabilizing transform

3.1 The transform and its theoretical properties

In this section we introduce a new transform for binomial random variables. The idea stems from the Fisz theorem (Fisz, 1955). In our new 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 Theorem 1, the proof of which can be found in the appendix.

Theorem 4. *Let $X_r \sim Bin(n_r, p_r)$, for $r = 1, 2$ with $p_r \in (0, 1)$ (fixed). Let m_r and ψ be as in Theorem 1. If the random variables X_1 and X_2 are independent and*

$$\lim_{\substack{\lambda_1 \to \infty \\ \lambda_2 \to \infty}} \frac{m_1}{m_2} = 1, \tag{10}$$

then the random variable defined by

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

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

$$m_B = \frac{m_2 - m_1}{\left(\frac{m_1+m_2}{n_1+n_2}(n_1+n_2-(m_1+m_2))\right)^{1/2}} \tag{11}$$

and

$$\sigma_B = \frac{\psi}{\left(\frac{m_1+m_2}{n_1+n_2}(n_1+n_2-(m_1+m_2))\right)^{1/2}}. \tag{12}$$

In the definition of ζ_B, we assume that the random variable takes the value zero when both X_1 and X_2 are zero.

We now give an example of Theorem 4. Suppose $X_r \sim Bin(n_r, p)$ for $r = 1, 2$, i.e. the binomial random variables have equal trial probabilities. Due to the theorem, the random variable $\zeta_B(X_1, X_2)$ will be asymptotically normal

$$N\left(\frac{(n_2 - n_1)p^{1/2}}{((n_1+n_2)(1-p))^{1/2}}, 1\right), \tag{13}$$

when $n_1 \to \infty$, $n_2 \to \infty$. In other words, using the transform $\zeta_B(X_1, X_2)$ will stabilize the variance of the asymptotic distribution.

Note also, that if in addition we impose the constraint that the binomial sample sizes are equal (i.e. $n_1 = n_2$), the asymptotic distribution will be 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 of the data. We might also like to know how fast the mean of ζ_B converges to the asymptotic normal mean. Even though the asymptotic normal distribution the theorem only holds when the means of the two binomial random variables are close (and large) through condition (10), it is interesting to study these properties in the finite sample case.

In some of the simulations below, we compare properties of our transform with that of Anscombe's angular transformation (1) and the Freeman-Tukey transformation (2) outlined in Section 2.

We follow a similar approach to these simulations as 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 of these parameters feature 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 binomial sizes $n = 1, 2, 4, 32, 128$ 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 the random variable $\zeta_B(X_1, X_2)$, denoted $z_n(p_1, p_2)$.

For the 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, doing this is favourable to Anscombe and Freeman-Tukey.

3.2.1 Mean simulations

To investigate the convergence of the samples of $\zeta_B(X_1, X_2)$ to the asymptotic mean m_B in equation (11), we computed their difference $|\bar{z}_n(p_1, p_2) - m_B|$, where the mean $\bar{z}$ is taken over the 10^5 samples.

Figure 1 shows the surface plots across the lattice of binomial probabilities (p_1,p_2) for increasing binomial size, n. The surfaces show that for larger n, the difference approaches zero across the whole lattice, with only a slight difference at the lattice boundary.

3.2.2 Variance simulations

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 2 gives a series of contour plots of the sample variance for each of the binomial sizes $n = 1, 2, 4, 32, 128$, renormalized so that the asymptotic distribution will 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 happens most near the line $p_1 = p_2$. This reflects the observation that equal binomial probabilities will result in an asymptotic distribution with unit variance.

To further examine the case when the two binomial proportions are equal, we display this 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 plots the squared residual of the variance from one against the (equal) binomial proportion. From this plot, it is more obvious that for small binomial sizes, our transform has variance closer to one for low and high proportions, especially when compared against Anscombe's transformation, although the Freeman-Tukey comes quite close to our transform. It is comparable to the two competitors for the middle half interval (0.25,0.75). For larger n, all three transforms do well at stabilizing the variance at one.

3.2.3 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 which are more Gaussian.

Figure 4 shows contour plots of the difference in Kolmogorov-Smirnov statistics between Anscombe's transform and ζ_B. A 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 very 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 has better Gaussianization properties than both Anscombe and the Freeman-Tukey transformation. As expected, as the binomial size becomes high, the differences between the Kolmogorov-Smirnov statistics becomes negligible, due to both transforms having good Gaussianizing properties. However, examining the statistics further, the means of the statistics for ζ_B are lower compared to those of its competitors (for all values of the binomial size, n). This indicates that the transformed data using our transform is more Gaussian than those of the Anscombe or Freeman-Tukey transforms.

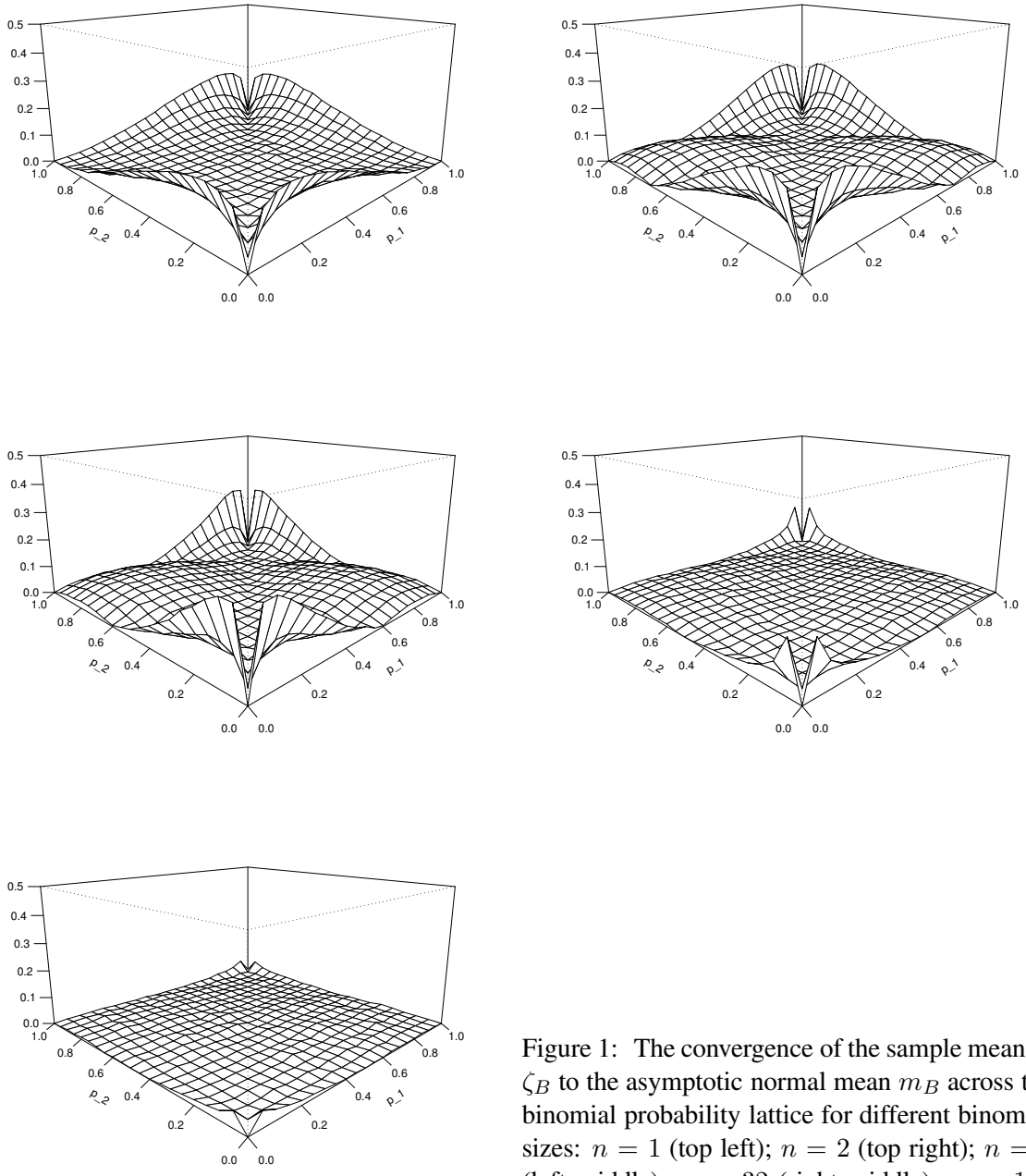


Figure 1: The convergence of the sample mean of ζ_B to the asymptotic normal mean m_B across the binomial probability lattice for different binomial sizes: $n = 1$ (top left); $n = 2$ (top right); $n = 4$ (left middle); $n = 32$ (right middle); $n = 128$ (bottom).

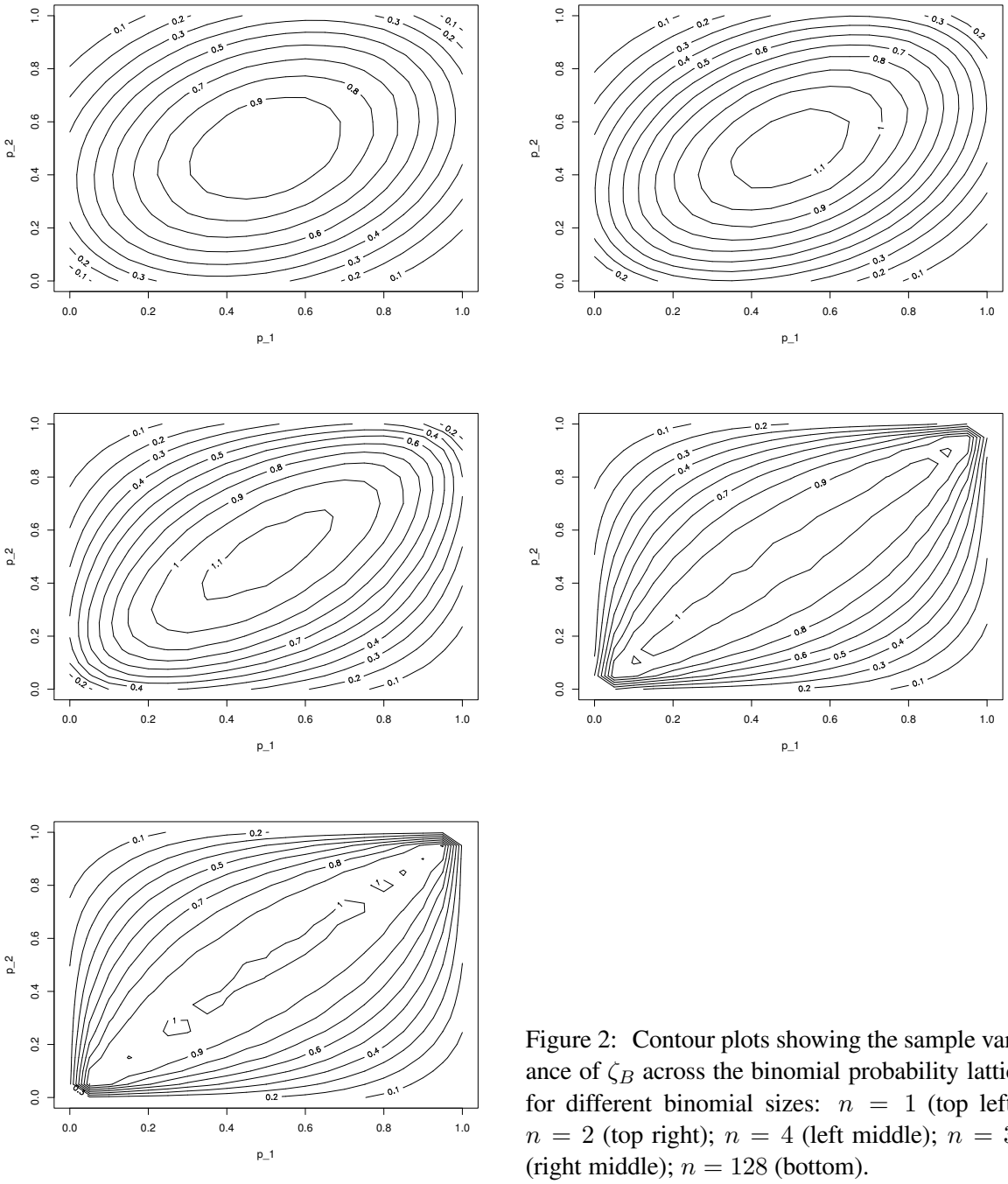


Figure 2: Contour plots showing the sample variance of ζ_B across the binomial probability lattice for different binomial sizes: $n = 1$ (top left); $n = 2$ (top right); $n = 4$ (left middle); $n = 32$ (right middle); $n = 128$ (bottom).

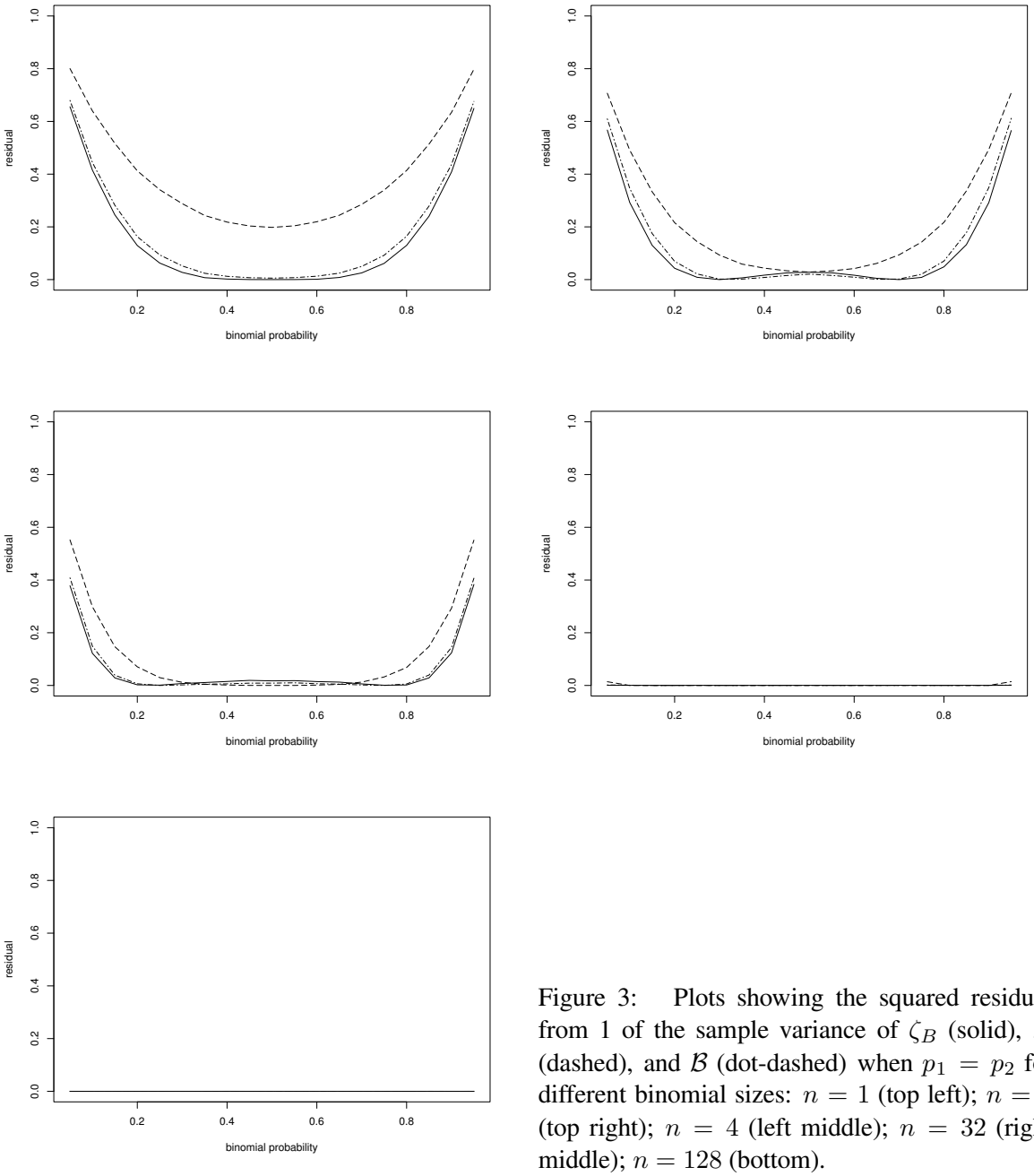


Figure 3: Plots showing the squared residual from 1 of the sample variance of ζ_B (solid), $\mathcal{A}$ (dashed), and $\mathcal{B}$ (dot-dashed) when $p_1 = p_2$ for different binomial sizes: $n = 1$ (top left); $n = 2$ (top right); $n = 4$ (left middle); $n = 32$ (right middle); $n = 128$ (bottom).

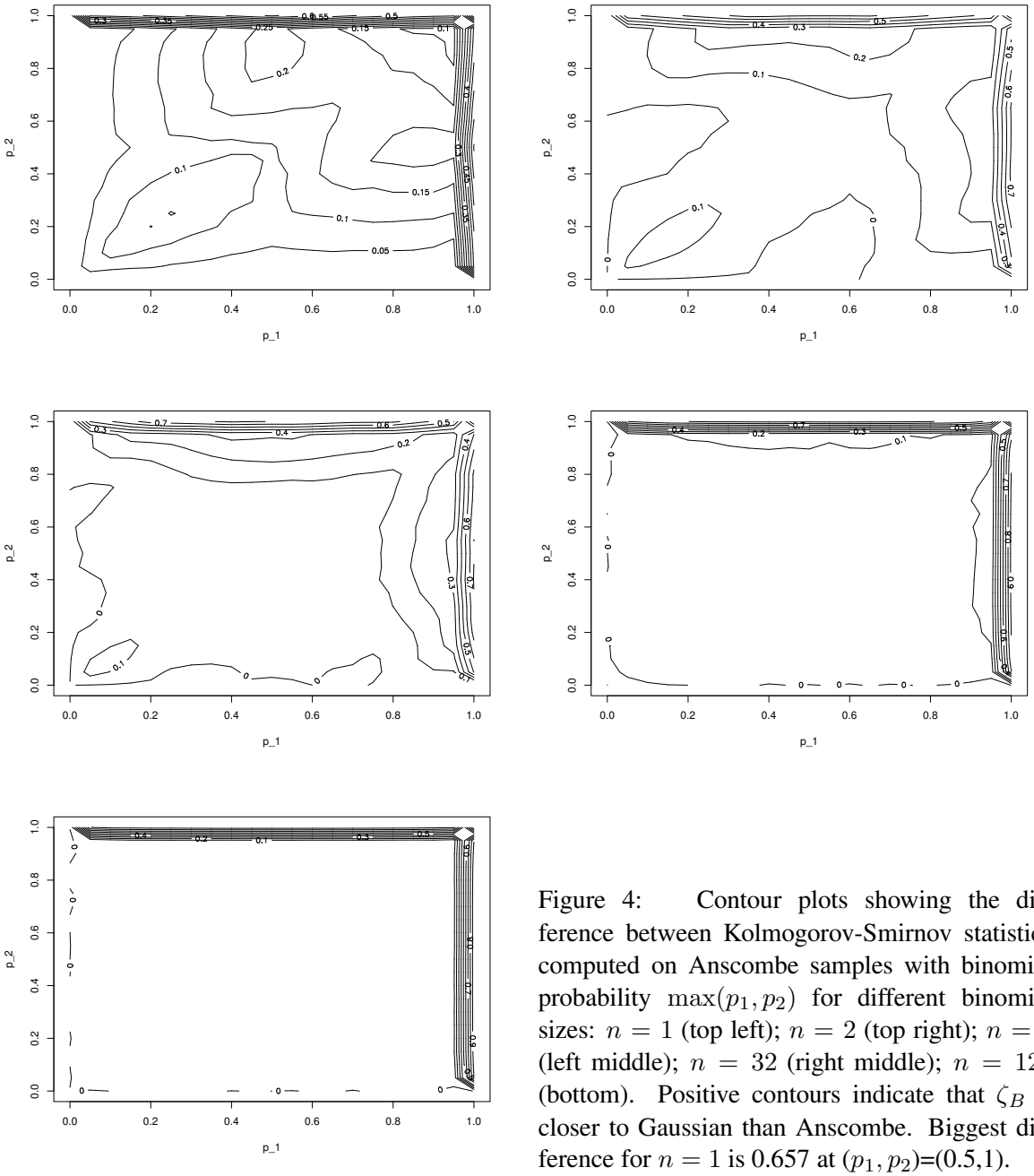


Figure 4: Contour plots showing the difference between Kolmogorov-Smirnov statistics computed on Anscombe samples with binomial probability $\max(p_1, p_2)$ for different binomial sizes: $n = 1$ (top left); $n = 2$ (top right); $n = 4$ (left middle); $n = 32$ (right middle); $n = 128$ (bottom). Positive 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).

4 The Haar-NN transform for binomial random variables

4.1 The transform

We now introduce an algorithm similar to the Haar-Fisz transform described in Section 2.4, 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 $0 \le v_i \le n$, for some integer n. The algorithm is as follows.

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 by defining

$$f_{j,k} = \begin{cases} 0 & \text{if } c_{j,k} = 0 \text{ or } c_{j,k} = n, \\ d_{j,k}/\sqrt{c_{j,k}(n - c_{j,k})/n} & \text{otherwise} \end{cases} \tag{14}$$

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

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" the steps 1 and 2.

Let us examine the effect of the modification in step 1 of the above procedure. Consider the coefficients v_0 and v_1. The modified detail coefficient $d_{J-1,0}$ is produced by

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

Similarly, for the next coarsest level coefficient, we have

$$\begin{aligned} f_{J-2,0} &= \frac{\frac{1}{2}(c_{J-1,1} - c_{J-1,0})}{\left(\frac{1}{2}(c_{J-1,0} + c_{J-1,1})\left(n - \frac{c_{J-1,0}+c_{J-1,1}}{2}\right)/n\right)^{1/2}} \\ &= \frac{((v_0 + v_1) - (v_3 + v_4))}{\left((v_0 + v_1 + v_2 + v_3)(4n - (v_0 + v_1 + v_2 + v_3))/n\right)^{1/2}}. \end{aligned}$$

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. binomial random variables $X_k \sim (n, p)$, 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 random variables 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 random variables Y_1 and Y_2, 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.

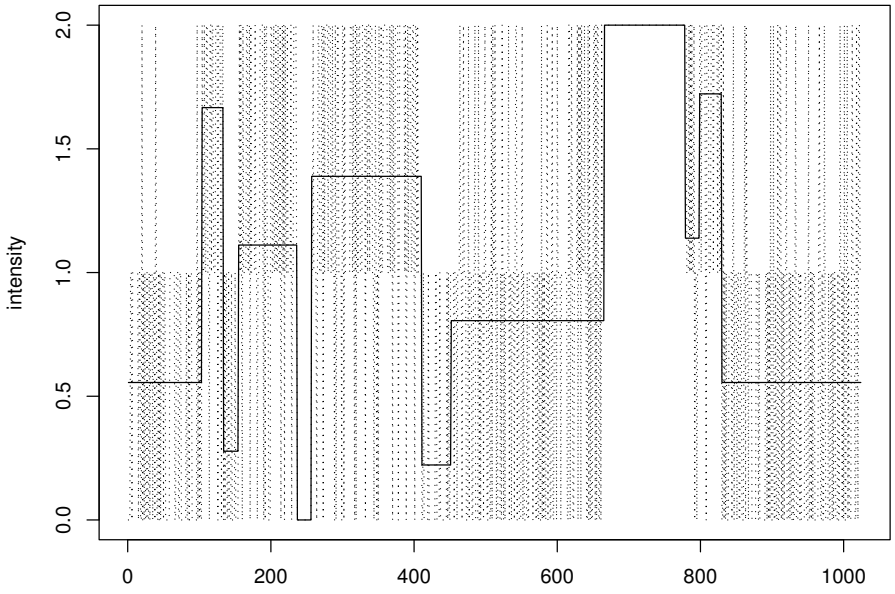


Figure 5: *Blocks* mean intensity vector $\boldsymbol{\lambda}$ based on $\boldsymbol{p}$ and binomial size $n = 2$, together with an example sample path (dotted).

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 (1), the Freeman-Tukey transformation (2) and the identity transformation. Again, we follow an approach similar to Fryźlewicz and Nason (2004).

For these simulations, we have chosen a binomial proportion vector, $\boldsymbol{p}$ of length $N = 1024$ sampled from a (normalized and stretched) version of the well-known *Blocks* test signal of Donoho and Johnstone (1994). For each binomial size $n = 1, 2, 4, 32, 128$, we will 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 will use. A sample path generated from binomial random variables with the mean vector $\boldsymbol{\lambda}$ will be denoted by $\mathbf{v}$. Figure 5 shows the (mean) intensity vector for $n = 2$, overlaid with a sample path generated from it. As expected, the 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.

4.2.1 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, $\mathbf{v}$. These paths were created from the mean vector $\boldsymbol{\lambda}$ for the

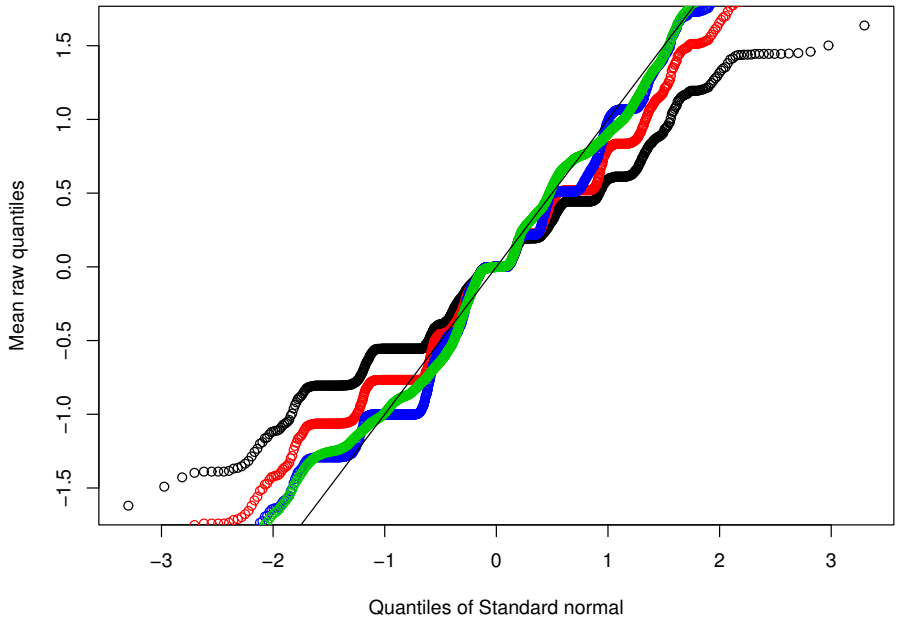


Figure 6: Q-Q plot comparison for four different transforms, averaged over 100 paths sampled from binomial variables with size $n = 2$ and proportion vector $\boldsymbol{p}$: $\mathbf{v}-\boldsymbol{\lambda}$ (black); $\mathcal{A}\mathbf{v}-\mathcal{A}\boldsymbol{\lambda}$ (red); $\mathcal{B}\mathbf{v}-\mathcal{B}\boldsymbol{\lambda}$ (blue); $\mathcal{F}_B\mathbf{v}-\mathcal{F}_B\boldsymbol{\lambda}$ (green). Solid line has slope 1, indicating unit variance.

binomial sizes $n = 1, 2, 4, 32, 128$. Figures 6 – 8 show this comparison for the binomial sizes $n = 2, 4$ and 128.

For the lowest binomial sizes, namely $n = 1$ and 2, the raw data (marked in black) is quite "stepped". This is expected since the data are discrete.

The Anscombe-transformed data and those transformed by Freeman-Tukey transformation still exhibit this characteristic, whilst for our transform, $\mathcal{F}_B$, they have lost most of this stepped character; the data lies closer to a straight line, showing that the data is more Gaussian. Moreover, the data is closer to the solid line (which has a slope of 1), which indicates a variance of one.

As n increases, the Q-Q lines become similar, although it can be said that our transform displays 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 do very well at bringing the data to normality. Furthermore, the variance is very close to one. However, this is mostly expected due to the high value of n, since at this large binomial size, the Central Limit Theorem comes into effect.

4.2.2 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 residual $|\mathcal{A}\mathbf{v}-\mathcal{A}\boldsymbol{\lambda}|^2$, $|\mathcal{B}\mathbf{v}-\mathcal{B}\boldsymbol{\lambda}|^2$ and $|\mathcal{F}_B\mathbf{v}-\mathcal{F}_B\boldsymbol{\lambda}|^2$ for the Anscombe transform, Freeman-Tukey transform and our transform (respectively), rescaled by their respective asymptotic variances. The residuals were averaged over 1000 sample paths, which were generated from the mean

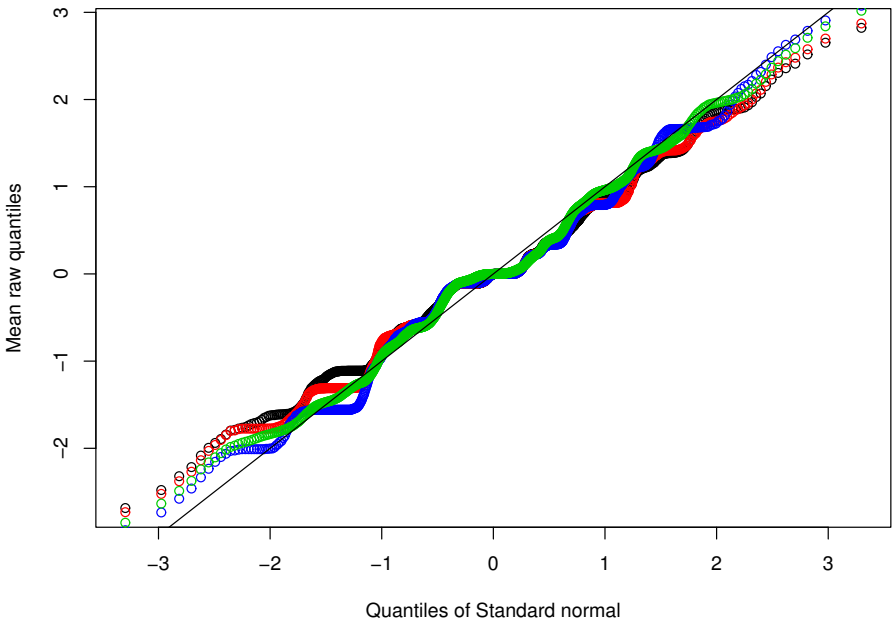


Figure 7: Q-Q plot comparison for four different transforms, averaged over 100 paths sampled from binomial variables with size $n = 4$ and proportion vector $\boldsymbol{p}$: $\mathbf{v}-\boldsymbol{\lambda}$ (black); $\mathcal{A}\mathbf{v}-\mathcal{A}\boldsymbol{\lambda}$ (red); $\mathcal{B}\mathbf{v}-\mathcal{B}\boldsymbol{\lambda}$ (blue); $\mathcal{F}_B\mathbf{v}-\mathcal{F}_B\boldsymbol{\lambda}$ (green). Solid line has slope 1, indicating unit variance.

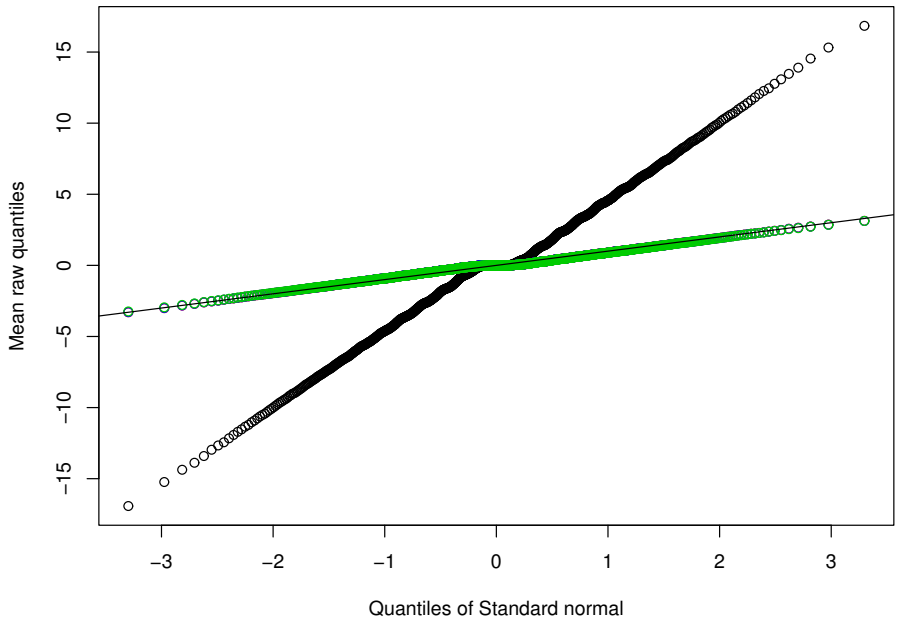


Figure 8: Q-Q plot comparison for four different transforms, averaged over 100 paths sampled from binomial variables with size $n = 128$ and proportion vector $\boldsymbol{p}$: $\mathbf{v}-\boldsymbol{\lambda}$ (black); $\mathcal{A}\mathbf{v}-\mathcal{A}\boldsymbol{\lambda}$ (red); $\mathcal{B}\mathbf{v}-\mathcal{B}\boldsymbol{\lambda}$ (blue); $\mathcal{F}_B\mathbf{v}-\mathcal{F}_B\boldsymbol{\lambda}$ (green). Solid line has slope 1, indicating unit variance.

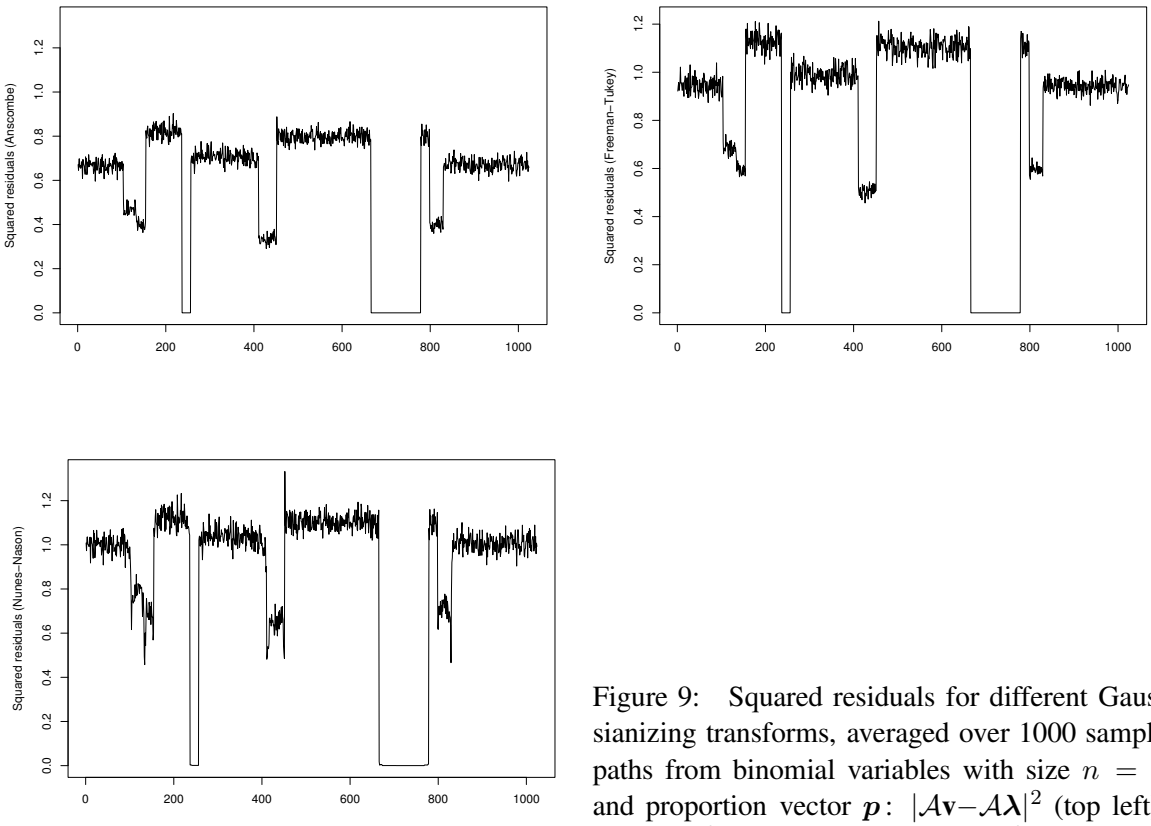


Figure 9: Squared residuals for different Gaussianizing transforms, averaged over 1000 sample paths from binomial variables with size $n = 2$ and proportion vector $\boldsymbol{p}$: $|\mathcal{A}\mathbf{v}-\mathcal{A}\boldsymbol{\lambda}|^2$ (top left); $|\mathcal{B}\mathbf{v}-\mathcal{B}\boldsymbol{\lambda}|^2$ (top right); $|\mathcal{F}_B\mathbf{v}-\mathcal{F}_B\boldsymbol{\lambda}|^2$ (bottom).

intensity vector $\boldsymbol{\lambda}$ for binomial sizes $n = 1, 2, 4, 32, 128$. When performance is optimal, the squared residuals stabilize at one when the proportion is nonzero, since the squared residuals form an estimate of the variance. The squared residuals for the three transforms are given in Figures 9 – 11 for $n = 2, 4$ and 128.

When the binomial size is small, the simulations show that our transform does much better than the competitors, $\mathcal{A}$ and $\mathcal{B}$, at stabilizing the sample path variances. For example, for $n = 2$, the Anscombe transform has the squared residual in the range 0.6 to 0.8, and the Freeman-Tukey transform has the squared residual in the range 0.9 to 1.1, whereas for our transform, the residual is nearer 1 for most of the sample path range. Further, our transform does relatively well compared to Anscombe and slightly better than Freeman-Tukey when the binomial proportion is small, that is in the three non-zero 'troughs'. However, there is a degree of erratic behaviour near the discontinuities in the proportion vector.

Moderate binomial sizes have the competitor transformations beginning to achieve similar stabilization as our transform; when $n = 128$, all three transforms do very well at variance stabilization, though Anscombe can be considered to do slightly better in performance in this case, due to the occasional downward spikes in the Haar-NN transform (see Figure 11).

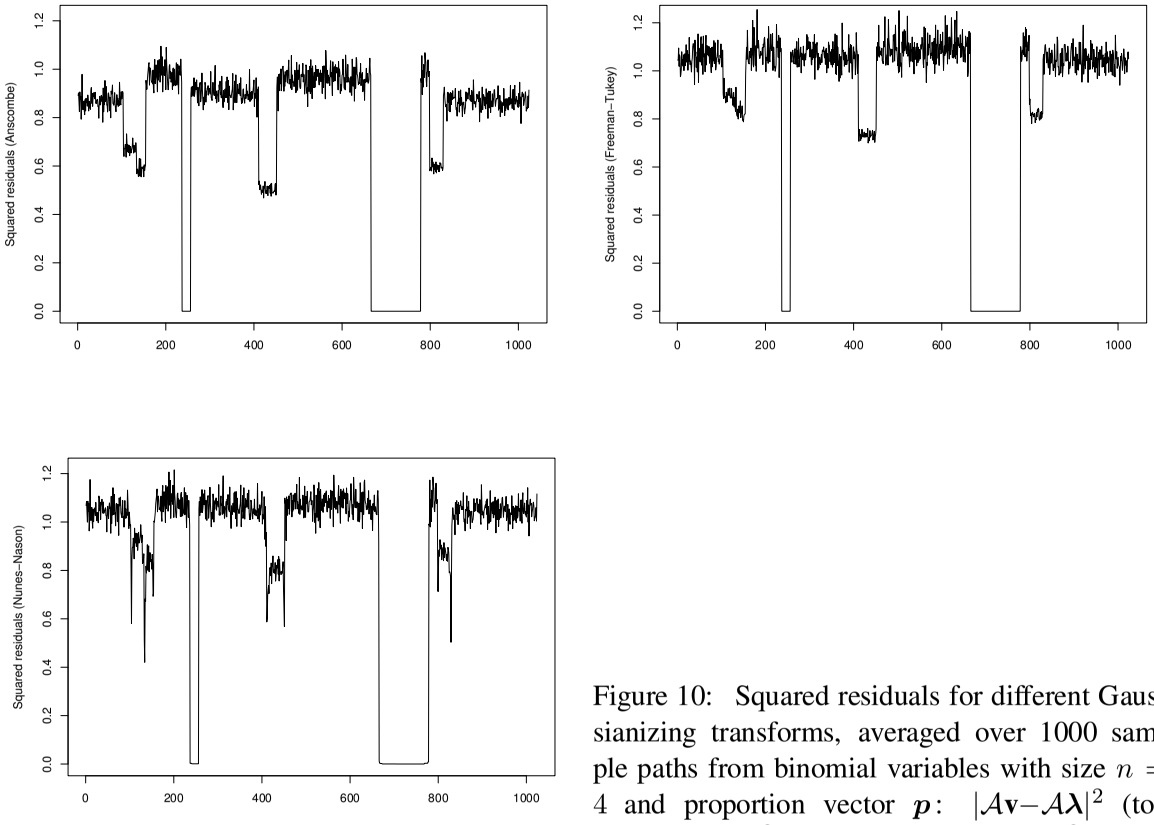


Figure 10: Squared residuals for different Gaussianizing transforms, averaged over 1000 sample paths from binomial variables with size $n = 4$ and proportion vector $\boldsymbol{p}$: $|\mathcal{A}\mathbf{v}-\mathcal{A}\boldsymbol{\lambda}|^2$ (top left);$|\mathcal{B}\mathbf{v}-\mathcal{B}\boldsymbol{\lambda}|^2$ (top right); $|\mathcal{F}_B\mathbf{v}-\mathcal{F}_B\boldsymbol{\lambda}|^2$ (right).

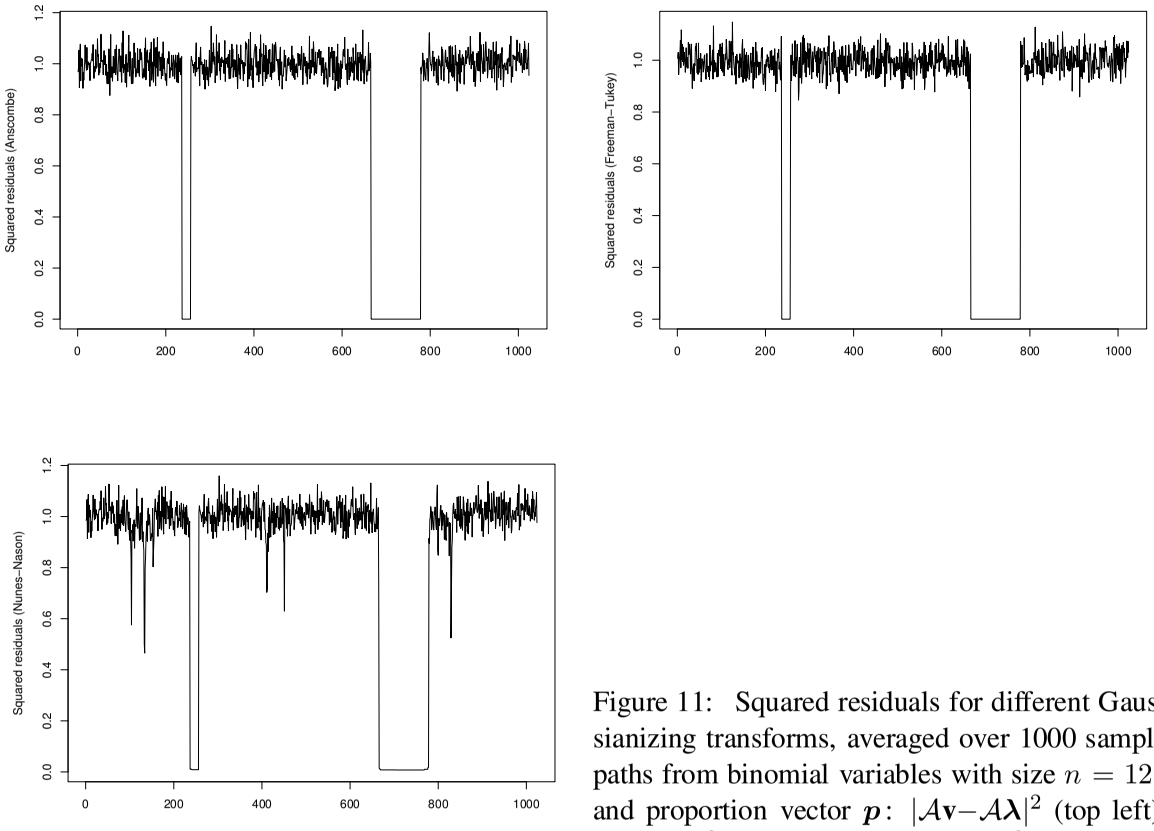


Figure 11: Squared residuals for different Gaussianizing transforms, averaged over 1000 sample paths from binomial variables with size $n = 128$ and proportion vector $\boldsymbol{p}$: $|\mathcal{A}\mathbf{v}-\mathcal{A}\boldsymbol{\lambda}|^2$ (top left); $|\mathcal{B}\mathbf{v}-\mathcal{B}\boldsymbol{\lambda}|^2$ (top right); $|\mathcal{F}_B\mathbf{v}-\mathcal{F}_B\boldsymbol{\lambda}|^2$ (right).

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, similar to that in Section 2.4:

Suppose **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 **v** to produce **u**=$\mathcal{F}_B$**v**. The vector **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 above. Several proportion functions were chosen to be estimated, each exhibiting different properties. These were the *Sinlog* function in Antoniadis and LeBlanc (2000):

$$\begin{aligned} P_1(t) = \Big(0.7 \quad &+ \quad \big[\sin(5\pi(t-0.4)) + \sin(6\pi(t-0.4-1/60)) \\ &+ \quad \sin(7\pi(t-0.4-1/35)) + \sin(8\pi(t-0.4-3/80)) \\ &+ \quad \sin(9\pi(t-0.4-2/45)) + \sin(10\pi(t-0.4-1/20)) \\ &+ \quad \sin(11\pi(t-0.4-3/55))\big]/7 \\ &+ \quad \cos((6/5)\pi(t-0.5))\Big)/3; \end{aligned}$$

a scaled and reflected version of the P_2 function described in Antoniadis and LeBlanc (2000):

$$P_3(t) = \begin{cases} 1.4\ P_2(2t) & t \in [0, \frac{1}{2}) \\ 1.4\ P_2(1-2t) & t \in [\frac{1}{2}, 1] \end{cases} \tag{15}$$

where

$$P_2(t) = (t+0.01)^{1/4}\, \mathrm{e}^{\left(-\frac{t^2}{1-t^2}\right)} \quad t \in [0,1);$$

and the modified *Blocks* proportion from Section 4.2.

These functions were sampled on regular grids of length $N = 128, 256, 512$ and 1024. The sampled vectors were then used to create binomial sample paths (from binomial sizes $n = 1, 4, 8$ and 16) 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, 1000 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, truncating the proportion estimates if necessary to lie within [0,1]. In the denoising step,

the DWT was used with Daubechies' Least Asymmetric wavelets. The thresholding implemented was the *SureShrink* procedure of Donoho and Johnstone (1995). The AMSE of the 1000 simulations for each wavelet and processor was recorded. For each binomial size and signal length, Table 1 shows the percentage difference in errors between $\mathcal{F}_B$ and $\mathcal{A}$ for the primary resolution level/vanishing moment combination with best performance (for both methods), for the proportion functions $P_1(t)$, $P_3(t)$ and *Blocks*. Positive differences show percentage average error improvement of our transform over that of Anscombe.

Table 1: Percentage improvement of $\mathcal{F}_B$ over $\mathcal{A}$ for binomial sizes $n = 1, 4, 8, 16$ and signal lengths $N = 128, 256, 512, 1024$ for primary resolution/vanishing moment combinations with best performance.

	Sinlog				P_3				*Blocks*			
Binomial	Signal length (N)				Signal length (N)				Signal length (N)			
size (n)	128	256	512	1024	128	256	512	1024	128	256	512	1024
1	1.76	2.51	3.44	4.44	5.73	12.17	13.54	7.38	0.42	0.44	1.77	2.45
4	7.50	9.22	12.30	18.69	11.36	6.52	6.50	9.76	3.44	4.22	4.37	5.05
8	5.57	6.67	10.00	15.52	1.13	0.21	3.47	3.59	1.85	3.30	2.25	4.71
16	3.49	3.24	5.02	6.73	−1.42	−4.57	−3.21	3.18	1.01	−7.37	−5.13	−3.32

The results of the simulation study are very encouraging. Overall, the algorithm with our method outperforms the algorithm when used with Anscombe nearly all of the time, especially with medium binomial sizes. The relative performance of the Haar-NN transform seems to increase as the signal length increases. The error improvement over Anscombe is in some cases quite substantial ($\geq 15\%$).

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 also of great importance.

In particular, DNA sequencing and gene expression methods are a couple of the topics in this area. One important problem in these areas is the modelling and prediction of isochore clusters in DNA sequence data (Bernardi, 2000). This information is useful to know for a range of biological applications. In this section we hope to use the Gaussianizing and variance stabilizing properties of the random variable $\zeta_B(X_1, X_2)$ for this application.

5.2.1 Biological background to the isochore problem

Before expressing the problem in a mathematical context, we now outline the problem in a biological setting.

DNA sequences are strings (polymers) of nucleotides, which store genetic information. Nucleotides are chemical compounds which play important rôles, for example in cellular behaviour and enzyme regulation.

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 groups, namely *purines* (adenine and thymine) and *pyrimidines* (cytosine and guanine), differing in structure. The nucleotides from a specific compound group are referred to as *base pairs*. For a more detailed discussion of the structure of DNA, see any introductory text on genomics, for example Brown (2002); Dale and von Schantz (2002); Cooper and Hausman (2004).

A *DNA isochore* is a long DNA segment which is (fairly) homogeneous in G+C content (Oliver *et al.*, 2004). 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, which asserts that genomes (chromosome DNA sequences) are mosaics of long DNA segments with different G+C content in adjacent segments; under this model, the G+C content mosaics differs for different organisms, especially between warm- and cold-blooded vertebrates (Bernardi, 2000), and so these features of DNA G+C content could be used, for example, in organism classification applications. Although the isochore features of certain vertebrates has already been investigated, an effective prediction method is of obvious interest.

5.2.2 IsoFinder: an existing approach to the isochore problem

In Oliver *et al.* (2004) and Zhang and Chen (2004), a procedure of sequential hypothesis testing is implemented to attempt 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 method 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 obtained clusters are predictions of isochores of the original DNA sequence. This method is known as the *IsoFinder* procedure.

5.2.3 Haar-NN transform approach to the isochore problem

Let us 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:

$$\begin{aligned} \text{DNA sequence:} \quad & \text{ATGCGCTACGTGCATGCAGTACCATGGACG}\ldots \\ \text{Converted sequence:} \quad & 00111100110110011010011001101 1\ldots \end{aligned}$$

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 (independent) Bernoulli random variables on 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 so

$$\begin{aligned} \mathbb{P}(\text{nucleotide } k \text{ has G+C content}) &= \mathbb{P}(X_k = 1) = p_k \\ \mathbb{P}(\text{nucleotide } k \text{ has A+T content}) &= \mathbb{P}(X_k = 0) = 1 - p_k = q_k. \end{aligned}$$

Estimating equal p_k for long consecutive sequences of k indicate regions of equal G+C content, and is representative of an isochore.

5.2.4 Examples

To test the G+C proportion estimation procedure, two chromosome strands were acquired from the Wellcome Trust Sanger Institute Human Genome Sequencing Group, namely the chromosome 6 MHC

strand (examined in Oliver *et al.* (2004)) as well as chromosome 20 of the human genome[2]. To make it feasible to process this data with our method, the sequence strands were cropped to $2^{21} = 2097152$ 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 $2^{11} = 2048$ bases do not feature in our estimates of G+C content produced after inversion of the wavelet transform.

To assess our isochore map estimates, the IsoFinder method was applied to the cropped nucleotide sequences, using the online IsoFinder implementation. Figures 12 and 14 were created using this web interface[3].

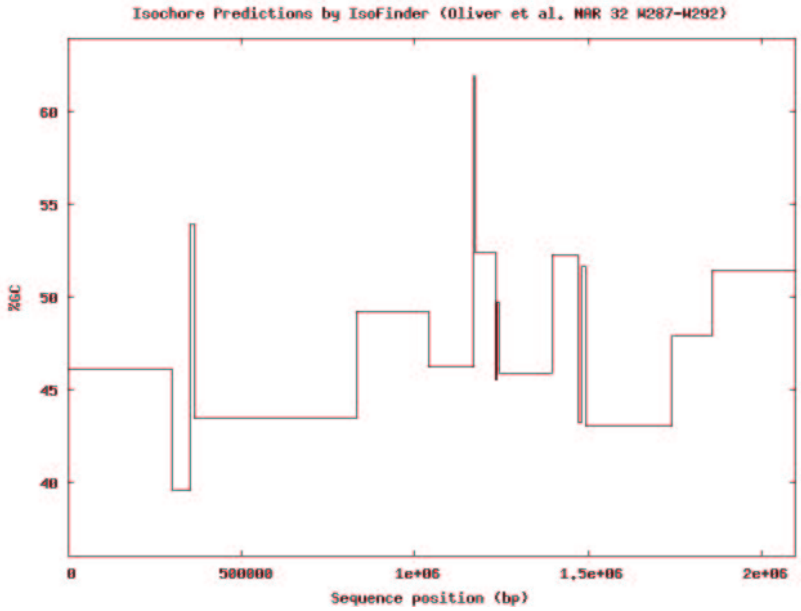


Figure 12: Isochore map of the chromosome 6 MHC nucleotide sequence, as estimated by the Isofinder procedure (with 3 kilobase filtering).

Figures 12 and 13 show the isochore maps of the MHC nucleotide sequence for the two estimation procedures, whereas Figures 14 and 15 give the corresponding estimates for the chromosome 20 of the human genome. 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 accurate estimates, for example, *EbayesThresh* (Johnstone and Silverman, 2005a, 2004, 2005b), which is known to be more insensitive to wavelet primary resolution choice. Also, our method uses a low kilobase

[2]All sequences produced by the Sanger Institute are available online from the website http://www.sanger.ac.uk/HGP/ .

[3]This can be found at http://bioinfo2.ugr.es/IsoF/isofinder.html.

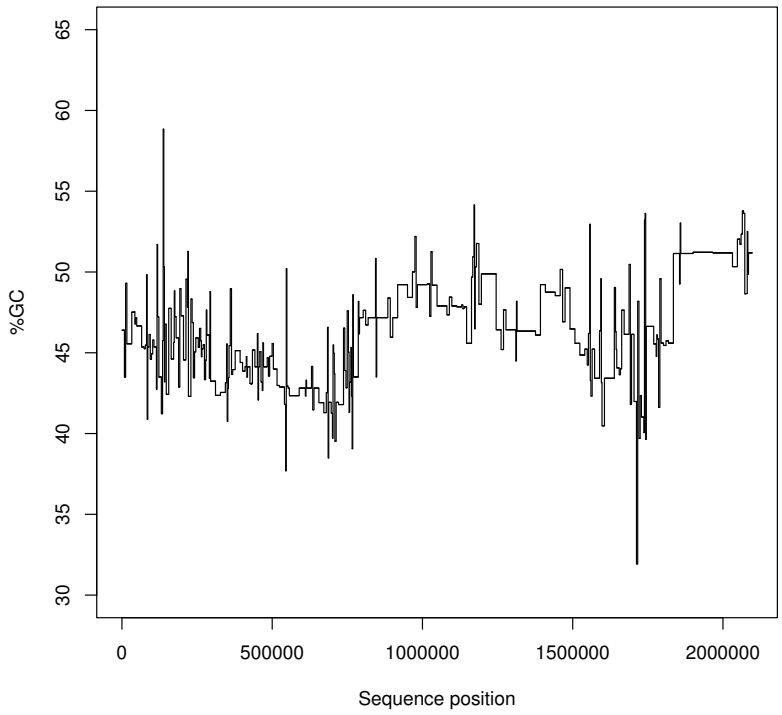


Figure 13: Isochore map of chromosome 20 of the human genome, as estimated by our Haar-Fisz Gaussianizing procedure (with 11 finest detail coefficient levels set to zero).

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

6 Conclusions

This article has proposed a new transform, ζ_B, that possesses variance-stabilizing properties for binomial random variables, which the Fisz transform (Fisz, 1955) cannot achieve.

An asymptotic result was established about this transform for binomial random variables, and simulations for different binomial sizes and probabilities were performed to investigate how well it Gaussianizes and stabilizes the variance compared to Anscombe's transformation. The results indicate that our transform does very well for smaller binomial sizes, n, and/or for extreme binomial proportions. As n is increased, the two transforms are comparable.

Section 4 introduced a new modified Haar transform using our Gaussianizing transform. This was compared to the Anscombe transform also, and it was found to again outperform the traditional transformation for smaller binomial sizes and/or binomial proportions nearer the boundaries of the interval (0,1). This improvement for small n and extreme proportions is important, since in practice, large binomial sizes and "nice" success probabilities could be unrealistic. Both methods perform well when n is large.

The evidence of good properties from the simulations lead us to suggest an algorithm for binomial

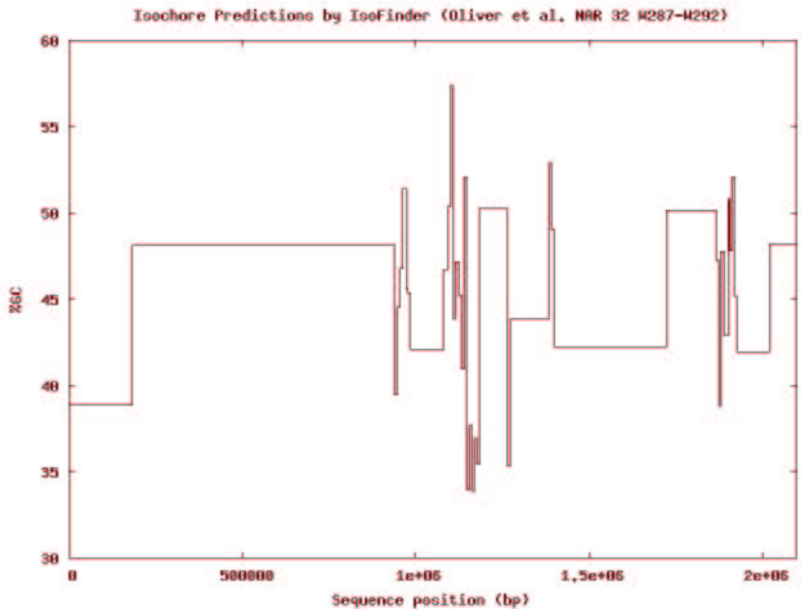


Figure 14: Isochore map of chromosome 20 of the human genome, as estimated by the Isofinder procedure (with 3 kilobase filtering).

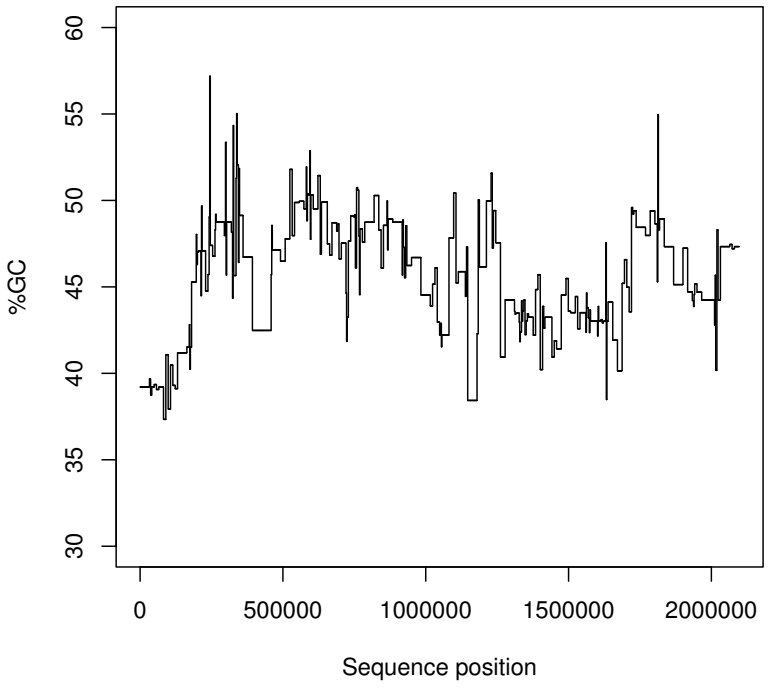


Figure 15: Isochore map of chromosome 20 of the human genome, as estimated by our Haar-Fisz Gaussianizing procedure (with 11 finest detail coefficient levels set to zero).

proportion curve estimation. Investigations show error improvements over a competitor in all but a few cases, with improvements in some cases being large.

7 Software and Acknowledgements

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`

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

A Proof of Theorem 4

The proof of this theorem follows the ideas used for the proof of Theorem 1 of Fisz (1955). We begin by presenting some introductory lemmas.

Lemma 5. *If ξ_1 and ξ_2 are independent and $\xi_r(\lambda_r)/m_r(\lambda_r)$ converges in probability to 1, then*

$$\lim_{\substack{\lambda_1\to\infty\\ \lambda_2\to\infty}} \mathbb{P}\left(\left|\frac{\xi_1+\xi_2}{m_1+m_2}-1\right|>\varepsilon\right)=0, \tag{16}$$

where ε is an arbitrary positive number.

For the proof of this lemma, see Fisz (1955).

Due to the Law of Large Numbers, X_r/m_r converges in probability to 1 for $p_r \in (0,1)$ and $r=1,2$. Thus, taking ξ_r to be the binomial random variables X_r, it follows that $R(n_1,n_2)=\frac{X_1+X_2}{m_1+m_2}$ also converges to 1 in probability when $n_1\to\infty$ and $n_2\to\infty$.

Lemma 6. *For the random variable $R_1(n_1,n_2)=\frac{n_1+n_2-(X_1+X_2)}{n_1+n_2-(m_1+m_2)}$,*

$$\lim_{\substack{n_1\to\infty\\ n_2\to\infty}} \mathbb{P}\left(|R_1(n_1,n_2)-1|>\varepsilon\right)=0, \tag{17}$$

where ε is an arbitrary positive number.

Proof of lemma 6. For $\varepsilon_1>0$, Lemma 5 implies that for sufficiently large values of n_1 and n_2, the inequality

$$-\varepsilon_1 < 1-R(n_1,n_2) < \varepsilon_1 \tag{18}$$

occurs with probability greater than $1-\delta$, for $\delta>0$ an arbitrarily small positive number. Then using the definition of R and R_1,

$$\begin{aligned}
-\varepsilon_1 < \quad & 1-R(n_1,n_2) && < \varepsilon_1\\
\Leftrightarrow \frac{-(m_1+m_2)\varepsilon_1}{n_1+n_2-(m_1+m_2)} < \quad & \tfrac{(m_1+m_2)-(X_1+X_2)}{n_1+n_2-(m_1+m_2)} && < \frac{(m_1+m_2)\varepsilon_1}{n_1+n_2-(m_1+m_2)}\\
\Leftrightarrow \frac{-(m_1+m_2)\varepsilon_1}{n_1+n_2-(m_1+m_2)} < \quad & R_1(n_1,n_2)-1 && < \frac{(m_1+m_2)\varepsilon_1}{n_1+n_2-(m_1+m_2)}.
\end{aligned}$$

Now let $\varepsilon = \frac{p\,\varepsilon_1}{1-p}$, where $p = \max\{p_1, p_2\}$. Since

$$0 \leq \frac{(m_1+m_2)}{n_1+n_2-(m_1+m_2)} \leq \frac{m_1+m_2}{(n_1+n_2)(1-p)} \leq \frac{(n_1+n_2)p}{(n_1+n_2)(1-p)} = \varepsilon/\varepsilon_1,$$

for the values of n_1 and n_2 such that the inequality (18) holds, we have

$$\begin{aligned} \mathbb{P}\left(|R_1(n_1,n_2)-1|<\varepsilon\right) &= \mathbb{P}\left(|R_1(n_1,n_2)-1|<\frac{p\,\varepsilon_1}{1-p}\right) \\ &\geq \mathbb{P}(|1-R(n_1,n_2)|<\varepsilon_1) \geq 1-\delta, \end{aligned}$$

where the numbers ε, ε_1 and δ are arbitrarily small. Hence

$$\lim_{\substack{n_1\to\infty \\ n_2\to\infty}} \mathbb{P}\left(|R_1(n_1,n_2)-1|>\varepsilon\right) = 0, \tag{19}$$

i.e. $R_1(n_1, n_2)$ converges in probability to 1.

□

Lemma 7. *(Fisz, 1955) If $\xi_r(\lambda_r)$ is asymptotically normal $N(m_r(\lambda_r), \sigma_r^2(\lambda_r))$, then the random variable $\xi_2 - \xi_1$ is asymptotically normal $N(m_2 - m_1, \psi^2)$ when $\lambda_1 \to \infty$, $\lambda_2 \to \infty$.*

For the proof of this lemma, see Fisz (1955).

We now state a theorem by Cramér (1946)[4] which we will also use in the proof of our theorem. Theorem 8 and its proof can be found in Cramér (1946), Section 20.6, p.254.

Theorem 8. *(Cramér, 1946) Let $\xi_1, \xi_2, \ldots$ be a sequence of random variables with distribution functions $F_1, F_2, \ldots$. Suppose that $F_n(x)$ tends to a distribution function $F(x)$ as $n \to \infty$.*

Let $\eta_1, \eta_2, \ldots$ be another sequence of random variables and suppose that η_n converges in probability to a constant c. Put

$$X_n = \xi_n + \eta_n,\ Y_n = \xi_n\eta_n,\ Z_n = \frac{\xi_n}{\eta_n}.$$

Then the distribution function of X_n tends to $F(x-c)$. Further, if $c > 0$, the distribution function of Y_n tends to $F\left(\frac{x}{c}\right)$, while the distribution function of Z_n tends to $F(cx)$.

We can now prove our theorem, Theorem 4. Let

$$A = \frac{X_1+X_2}{n_1+n_2}, \qquad B = (n_1+n_2-(X_1+X_2)),$$

$$C = \frac{m_1+m_2}{n_1+n_2}, \qquad D = (n_1+n_2-(m_1+m_2)),$$

and

$$y = (AB)/(CD) = \left(\frac{X_1+X_2}{m_1+m_2}\right)\left(\frac{n_1+n_2-(X_1+X_2)}{n_1+n_2-(m_1+m_2)}\right).$$

[4]Theorem 8 and its proof can be found in Cramér (1946), Section 20.6, p.254.

Then

$$\begin{aligned}
\tau(n_1,n_2) &= \frac{\zeta_B - m_B}{\sigma_B} \\
&= \frac{\frac{X_2-X_1}{(AB)^{1/2}} - \frac{m_2-m_1}{(CD)^{1/2}}}{\frac{\psi}{(CD)^{1/2}}} \\
&= \frac{(CD)^{1/2}(X_2-X_1)-(m_2-m_1)(AB)^{1/2}}{(ABCD)^{1/2}} \times \frac{(CD)^{1/2}}{\psi} \\
&= \frac{(CD)^{1/2}(X_2-X_1-(m_2-m_1))+(CD)^{1/2}(m_2-m_1)-(m_2-m_1)(AB)^{1/2}}{(AB)^{1/2}\psi} \\
&= \frac{(CD)^{1/2}(X_2-X_1-(m_2-m_1))+(m_2-m_1)((CD)^{1/2}-(AB)^{1/2})}{(AB)^{1/2}\psi} \\
&= \frac{y^{-1/2}(X_2-X_1-(m_2-m_1))}{\psi} + \frac{(m_2-m_1)(y^{-1/2}-1)}{\psi} \\
&= \frac{\eta + \frac{m_2-m_1}{\psi}(1-y^{1/2})}{y^{1/2}}, \qquad (20)
\end{aligned}$$

where $\eta(n_1,n_2) = \frac{X_2-X_1-(m_2-m_1)}{\psi}$ is the random variable defined in the proof of Lemma 7 for our specific binomial case (see Fisz (1955)). Note that $y(n_1,n_2) = \left(\frac{X_1+X_2}{m_1+m_2}\right)\left(\frac{n_1+n_2-(X_1+X_2)}{n_1+n_2-(m_1+m_2)}\right) = R(n_1,n_2)R_1(n_1,n_2)$, where R and R_1 are as defined in Lemma 5 and Lemma 6.

Note also that $\tau(n_1,n_2)$ is the random variable $\zeta_B(n_1,n_2)$ standardized by the asymptotic normal mean (11) and standard deviation (12). To prove the theorem, we need to show that τ is asymptotically normal $N(0,1)$.

Due to Lemmas 5 and 6, the random variables $R(n_1,n_2)$ and $R_1(n_1,n_2)$ both converge in probability to 1. It follows from a proposition due to Slutsky[5], Section 20.6, p.255), a corollary to Theorem 8, that their product $y(n_1,n_2)$ also converges in probability to 1.

Using the same proposition again, this in turn implies that the function $y^{1/2}(n_1,n_2)$ converges in probability to $1^{1/2}=1$, since this is a rational function in $y(n_1,n_2)$.

Since X_1 and X_2 are asymptotically normal $N(m_r,\sigma_r^2)$, then Lemma 7 applies here; thus $\eta(n_1,n_2)$ is asymptotically normal $N(0,1)$, i.e. its distribution function converges to $\Phi(x)$.

Let us now consider the other expression in the numerator of τ. Note that when regarded as a function of y, we can write

$$(1-y^{1/2}) = (1/2+\theta)(1-y),$$

where $\theta(y) = \frac{1-\sqrt{y}}{2(1+\sqrt{y})}$. Note that $\theta \to 0$ as $y \to 1$, which means that

$$\exists\, \delta_0 > 0 \quad \text{such that } |y-1| < \delta_0 \Rightarrow |\theta| < \varepsilon_0, \qquad (21)$$

for any positive number ε_0. Since y converges in probability to 1, for sufficiently large n_1, n_2 we also have

$$\mathbb{P}(|y-1| < \delta_0) > 1-\gamma, \qquad (22)$$

[5]This proposition can be found in Cramér (1946), Section 20.6, p.255.

for the $\delta_0 > 0$ as in equation (21) and for any arbitrarily small number γ. Combining equations (21) and (22), we obtain

$$\mathbb{P}(|\theta| < \varepsilon_0) \geq \mathbb{P}(|y-1| < \delta_0) > 1 - \gamma,$$

for the values of n_1 and n_2 valid for the relation in equation (22); since γ can be arbitrarily small, θ converges in probability to 0.

Let us now write

$$z = \frac{m_2 - m_1}{\psi}(1 - y^{1/2}) = \frac{m_2 - m_1}{\psi}(1/2 + \theta)(1 - y).$$

Now $y = \left(\frac{X_1+X_2}{m_1+m_2}\right)\frac{B}{D}$, using the quantities B and D defined earlier. Thus

$$\begin{aligned} 1 - y &= 1 - \left(\frac{X_1 + X_2}{m_1 + m_2}\right)\frac{B}{D} \\ &= \frac{(m_1 + m_2)D - (X_1 + X_2)B}{(m_1 + m_2)D}. \end{aligned}$$

So

$$\begin{aligned} z &= -\frac{m_2 - m_1}{m_1 + m_2}(1/2 + \theta)\frac{(X_1 + X_2)B - (m_1 + m_2)D}{\psi D} \\ &= -\frac{m_2 - m_1}{m_1 + m_2}(1/2 + \theta)\varrho(n_1, n_2), \end{aligned}$$

where $\varrho(n_1, n_2) = \frac{(X_1+X_2)B-(m_1+m_2)D}{\psi D}$.

Note that ϱ can be expressed as $\varrho(n_1, n_2) = \frac{(X_1+X_2)R_1-(m_1+m_2)}{\psi}$.

A slight modification to the proof of Lemma 7 shows that $(X_1 + X_2)$ is asymptotically normal $N(m_1 + m_2, \psi^2)$. Due to Theorem 8, the random variable $(X_1 + X_2)R_1$ is also asymptotically normal $N(m_1 + m_2, \psi^2)$, since from Lemma 6, R_1 converges in probability to 1. It follows that ϱ is asymptotically normal $N(0, 1)$.

We want to show that z converges to zero in probability, in order to use Theorem 8 again to complete the proof of our theorem.

Let ε, $\delta > 0$ be arbitrary given numbers. Then

$$\mathbb{P}(|z| > \varepsilon) = \mathbb{P}\left(|z| > \varepsilon \middle| |\theta| > \delta\right)\mathbb{P}(|\theta| > \delta) + \mathbb{P}\left(|z| > \varepsilon \middle| |\theta| < \delta\right)\mathbb{P}(|\theta| < \delta). \tag{23}$$

Now

$$\mathbb{P}\left(|z| > \varepsilon \middle| |\theta| > \delta\right)\mathbb{P}(|\theta| > \delta) \to 0, \tag{24}$$

since θ converges in probability to zero, and thus it remains to show that the second summand in the expression (23) converges to zero. From the definition of z, we have

$$\begin{aligned} \mathbb{P}\left(|z| > \varepsilon \middle| |\theta| < \delta\right) &\leq \mathbb{P}\left(\left|\frac{m_2 - m_1}{m_1 + m_2}\right|(1/2 + |\theta|)|\varrho(n_1, n_2)| > \varepsilon \middle| |\theta| < \delta\right) \\ &\leq \mathbb{P}\left(\left|\frac{m_2 - m_1}{m_1 + m_2}\right|(1/2 + \delta)|\varrho(n_1, n_2)| > \varepsilon \middle| |\theta| < \delta\right). \end{aligned}$$

Let the first event in the last expression above be denoted by E. Then conditional on $|\theta| < \delta$, the probability of E^c can be expressed as

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{m_2 - m_1}{m_1 + m_2}\right| (1/2 + \delta)|\varrho(n_1, n_2)| < \varepsilon\right) \\ = \quad & \mathbb{P}\left(\left|\frac{m_2 - m_1}{m_1 + m_2}\right| \varrho(n_1, n_2) < \frac{\varepsilon}{1/2 + \delta}\right) - \mathbb{P}\left(\left|\frac{m_2 - m_1}{m_1 + m_2}\right| \varrho(n_1, n_2) < \frac{-\varepsilon}{1/2 + \delta}\right) \\ \to \quad & 1 - 0 = 1, \end{aligned}$$

since $\left|\frac{m_2 - m_1}{m_1 + m_2}\right| \leq \left|\frac{m_2 - m_1}{2\min\{m_1, m_2\}}\right| \to 0$ (due to the assumption (10) of the theorem) and the fact that $\varrho(n_1, n_2)$ is asymptotically normal $N(0, 1)$.

This implies that

$$\mathbb{P}\left(\left|\frac{m_2 - m_1}{m_1 + m_2}\right| (1/2 + \delta)|\varrho(n_1, n_2)| > \varepsilon \Big| |\theta| < \delta\right) \to 0. \tag{25}$$

The two relations (24) and (25) together imply that z converges in probability to 0.

Recall that we have

$$\tau(n_1, n_2) = \frac{\eta(n_1, n_2) + z(n_1, n_2)}{y(n_1, n_2)}.$$

Using the Cramér result, we see that the distribution of $(\eta + z)(n_1, n_2)$ tends to $\Phi(x - 0) = \Phi(x)$, since $\eta(n_1, n_2)$ is asymptotically normal $N(0, 1)$ and z converges in probability to 0.

Using the result again, the distribution of $\tau(n_1, n_2) = \left(\frac{\eta + z}{y}\right)(n_1, n_2)$ tends to $\Phi(1 \cdot x) = \Phi(x)$, since the distribution of $(\eta + z)(n_1, n_2)$ tends to $\Phi(x)$ and y converges in probability to 1. This completes the proof of the theorem.

□

B Simulation graphics

This appendix gives extra graphical representation of the findings in this article; the plots are related to the investigation into the Gaussianization and variance-stabilizing properties of ζ_B corresponding to Section 3.2 in the main text.

- Contour plot for convergence of the sample mean to the asymptotic mean (Figure 16)
- Perspective plot for the (normalized) sample variance (Figure 17)
- Comparison between ζ_B and Anscombe's transformation of sample variance for equal binomial proportions (Figure 18)
- Perspective plot for difference in Kolmogorov-Smirnov statistics between Anscombe and ζ_B (Figure 19).

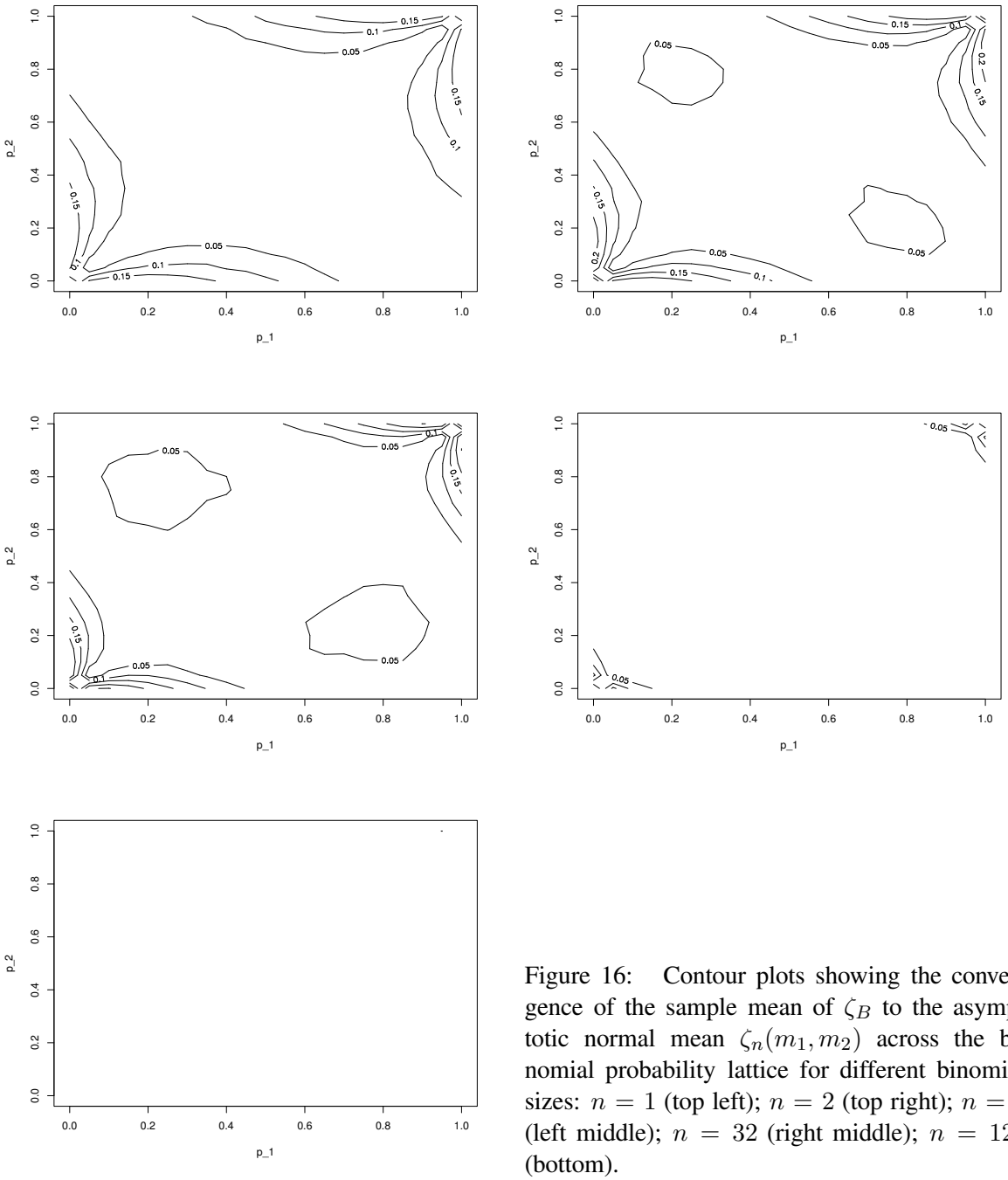
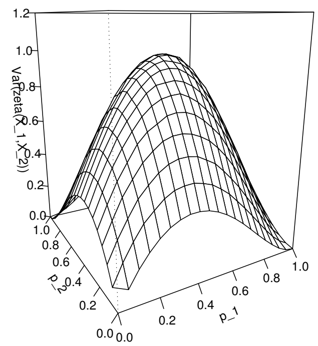
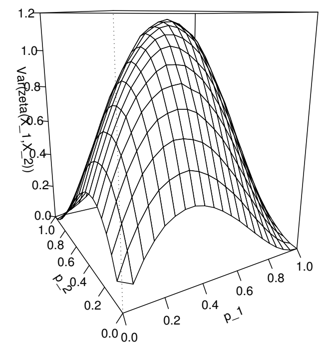
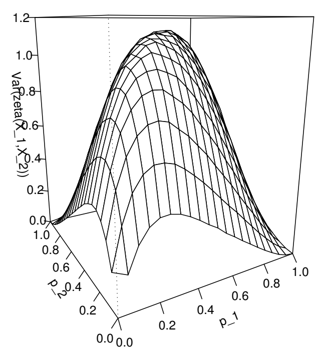
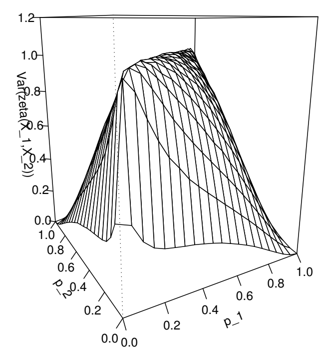


Figure 16: Contour plots showing the convergence of the sample mean of ζ_B to the asymptotic normal mean $\zeta_n(m_1, m_2)$ across the binomial probability lattice for different binomial sizes: $n = 1$ (top left); $n = 2$ (top right); $n = 4$ (left middle); $n = 32$ (right middle); $n = 128$ (bottom).









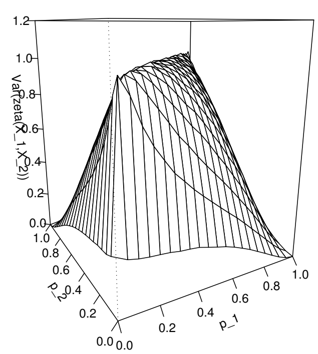


Figure 17: Plots showing the sample variance of ζ_B across the binomial probability lattice for different binomial sizes: $n = 1$ (top left); $n = 2$ (top right); $n = 4$ (left middle); $n = 32$ (right middle); $n = 128$ (bottom).

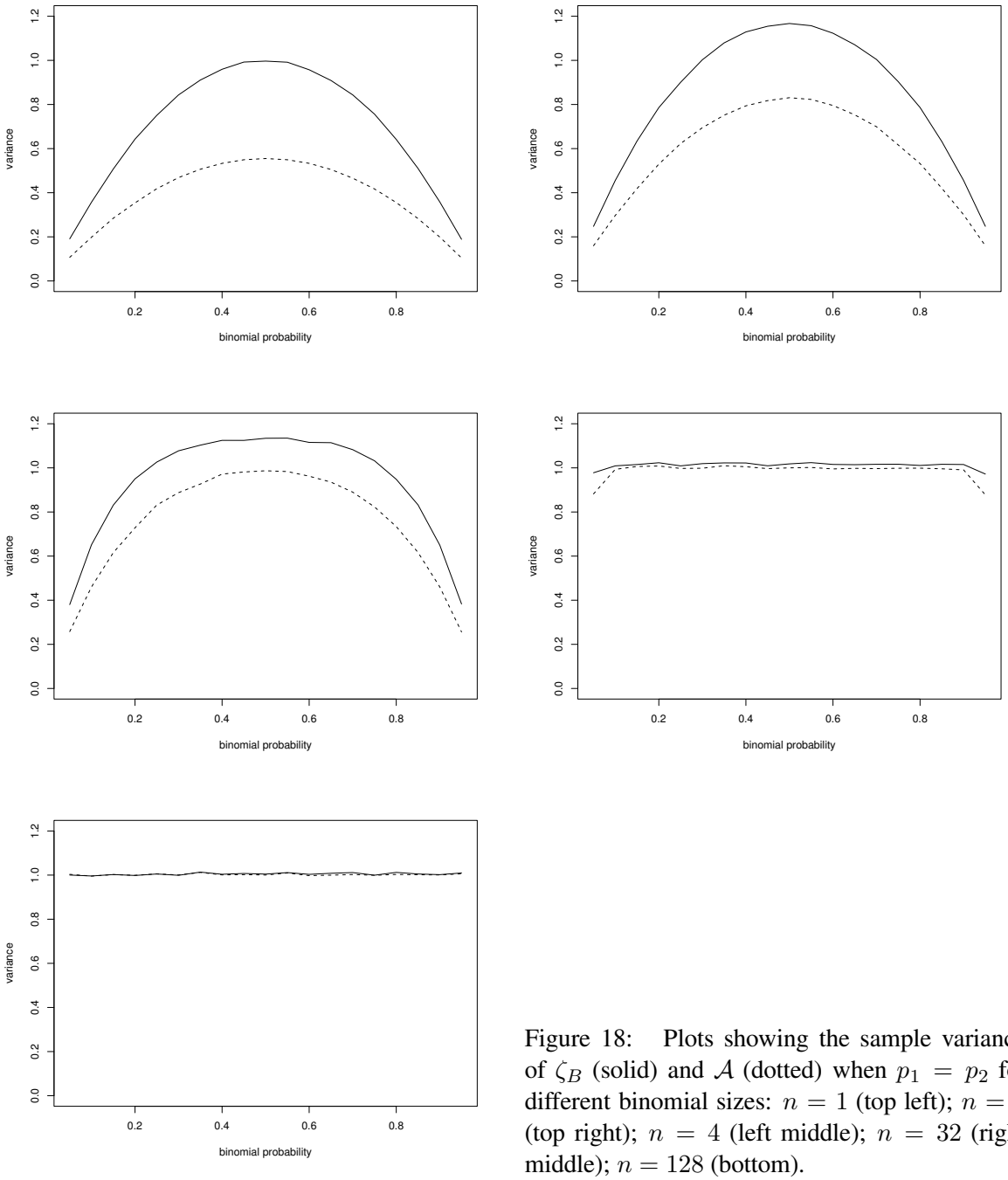
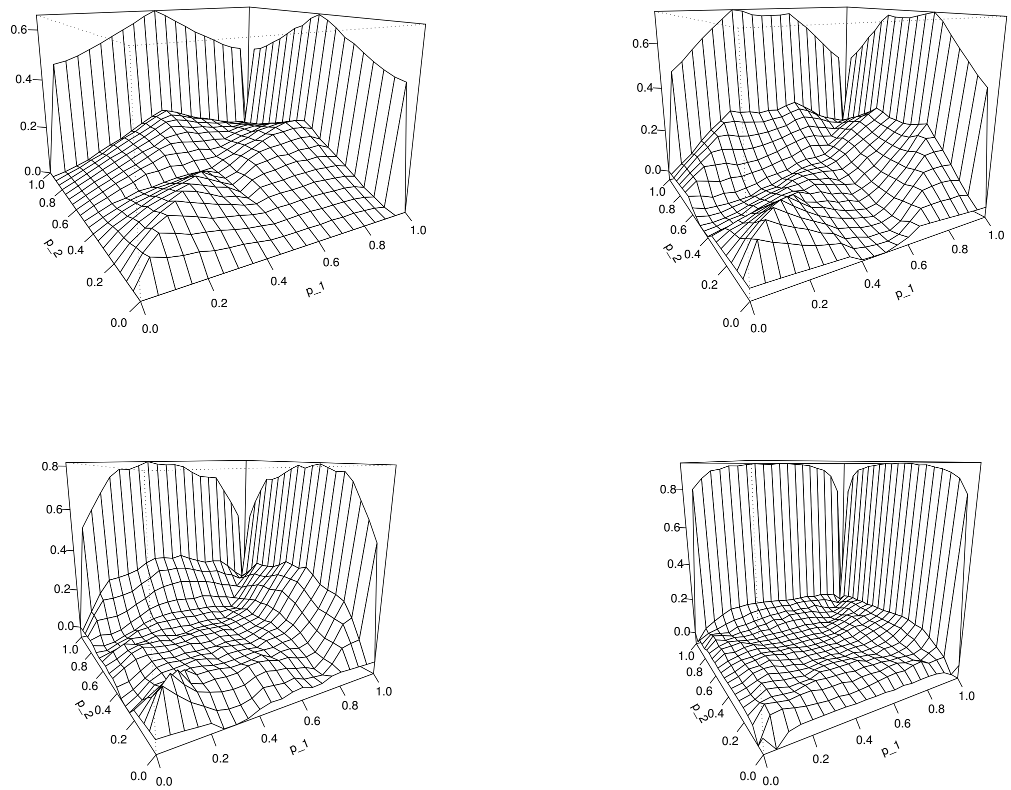


Figure 18: Plots showing the sample variance of ζ_B (solid) and $\mathcal{A}$ (dotted) when $p_1 = p_2$ for different binomial sizes: $n = 1$ (top left); $n = 2$ (top right); $n = 4$ (left middle); $n = 32$ (right middle); $n = 128$ (bottom).



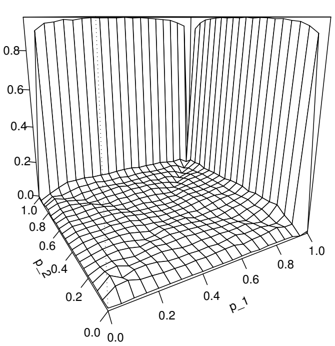


Figure 19: Plots showing the difference between Kolmogorov-Smirnov statistics computed on Anscombe samples with binomial probability $\max(p_1, p_2)$ for different binomial sizes: $n = 1$ (top left); $n = 2$ (top right); $n = 4$ (left middle); $n = 32$ (right middle); $n = 128$ (bottom). Positive difference indicates that ζ_B is closer to Gaussian.

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*. Oxford: BIOS Scientific, 2nd edn.

Cooper, G. M. and Hausman, R. E. (2004) *The Cell: a molecular approach*. ASM Press, 3rd edn.

Cramér, H. (1946) *Mathematical methods of statistics*. Princeton University Press.

Dale, J. W. and von Schantz, M. (2002) *From Genes to Genomes: concepts and applications of DNA technology*. Wiley.

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

Donoho, D. L. (1993) Nonlinear wavelet methods for recovery of signals, densities and spectra from indirect and noisy data. In *Proceedings of Symposia in Applied Mathematics: Different Perspectives on Wavelets* (ed. I. Daubechies), vol. 47, 173–205. American Mathematical Society.

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

— (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. Ass.*, **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.

— (2006) Haar-Fisz estimation of evolutionary wavelet spectra. *J. Roy. Stat. Soc. B*, **68**, 611–634.

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

Johnstone, I. M. and Silverman, B. W. (2004) Needles and hay in haystacks: Empirical bayes estimates of possibly sparse sequences. *Ann. Stat.*, **32**, 1594–1649.

— (2005a) Ebayesthresh: R programs for empirical bayes thresholding. *J. Stat. Soft.*, **12.8**, 1–38.

— (2005b) Empirical bayes selection of wavelet thresholds. *Ann. Stat.*, **33**, 1700–1752.

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*. London: Chapman and Hall.

Nason, G. P. and Silverman, B. W. (1994) The discrete wavelet transform in S. *J. Comp. Graph. Statist.*, **3**, 163–191.

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

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.