Sharp Fixed n Bounds and Asymptotic Expansions for the Mean and the Median of a Gaussian Sample Maximum, and Applications to Donoho-Jin Model

Anirban DasGupta, S.N. Lahiri, Jordan Stoyanov

January 23, 2014

Purdue University, North Carolina State University, and Newcastle University, UK

Dedicated to the Fond Memories of Kesar Singh

Abstract We are interested in the sample maximum $X_{(n)}$ of an iid standard normal sample of size n. First, we derive two-sided bounds on the mean and the median of $X_{(n)}$ that are valid for any fixed $n \geq n_0$, where n_0 is 'small', e.g. $n_0 = 7$. These fixed n bounds are established by using new very sharp bounds on the standard normal quantile function $\Phi^{-1}(1-p)$. The bounds found in this paper are currently the best available explicit nonasymptotic bounds, and are of the correct asymptotic order up to the number of terms involved.

Then we establish exact three term asymptotic expansions for the mean and the median of $X_{(n)}$. This is achieved by reducing the extreme value problem to a problem about sample means. This technique is general and should apply to suitable other distributions. One of our main conclusions is that the popular approximation $E[X_{(n)}] \approx \sqrt{2 \log n}$ should be discontinued, unless n is fantastically large. Better approximations are suggested in this article. An application of some of our results to the Donoho-Jin sparse signal recovery model is made.

The standard Cauchy case is touched on at the very end.

Key words: standard normal, maximum, mean, median, bounds, false discovery, sparse, mixture

MSC: 60E05, 62E20

1 Introduction

Let $X_{(n)}$ denote the maximum of an iid sample $X_1, \cdots, X_n$ from a univariate standard normal distribution. A knowledge of distributional properties of $X_{(n)}$, and especially of its mean value $E[X_{(n)}]$, became important in several frontier areas in theory as well as applications. A few instances are the widespread use of properties of $X_{(n)}$ for variable selection in sparse high dimensional regression, in studying the hard and soft thresholding estimators of Donoho and Johnstone (1994) for sparse signal detection and false discovery, and in analyzing or planning for extreme events of a diverse nature in practical enterprises, such as climate studies, finance, and hydrology.

We do know quite a bit about distributional properties of $X_{(n)}$ already. For example, we know the asymptotic distribution on suitable centering and norming; see Galambos (1978). We know that up to the first order, the mean, the median, and the mode of $X_{(n)}$ are all asymptotically of order $\sqrt{2\log n}$, and that convergence to the asymptotic distribution is very slow; see Hall (1979). There is also very original and useful nonasymptotic work by Lai and Robbins (1976) on the mean of $X_{(n)}$. The Lai-Robbins bounds were generalized to other order statistics, including the maximum, in Gascuel and Caraux (1992) and in Rychlik (1998), who considers general L-estimates.

In this article, we first provide the currently best available nonasymptotic bounds on the mean and the median of $X_{(n)}$. The bounds resemble the Edgeworth expansions of sample means. However, here the bounds are valid for fixed n. For example, our Theorem 4.1 shows that for $n \geq 7$, we have

$$E[X_{(n)}] \leq \sqrt{2\log n} - \frac{\log 4\pi + \log\log n - 2}{2\sqrt{2\log n}} + \frac{K(\log 4\pi + \log\log n)}{(2\log n)^{3/2}},$$

where the constant K is explicit and can be taken to be 1.5 (or anything bigger). If simpler nonasymptotic bounds, although numerically less accurate, were desired, they could be easily extracted out from the above bound. In particular, it follows that for all $n \geq 10$,

$$E[X_{(n)}] \leq \sqrt{2\log n} - \frac{\log 4\pi + \log\log n - 6}{6\sqrt{2\log n}}.$$

While valid for fixed n, the successive terms in the above bound go down at increasing powers of $1/\sqrt{2\log n}$, just as the successive terms in an Edgeworth expansion for the CDF of a sample mean go down at increasing powers of $1/\sqrt{n}$. However, the greatest utility of our bounds is that they are nonasymptotic, the sharpest ones available to date, and moreover, they cover both the mean and the median of $X_{(n)}$. Bounds on the median are stated in Section 3, and the bounds on the mean in Section 4.

On the technical side, there are a few principal ingredients for the bounds on the mean and the median. One is the following bound for the standard normal distribution function Φ and its density φ : There are nonnegative constants a, b, c, d such that

$$1 - \Phi(x) \leq \frac{a\varphi(x)}{bx + \sqrt{c^2x^2 + d}} \quad \text{for all} \quad x > 0. \quad \text{(See Szarek and Werner (1999).)}$$

The next important result we need is the following general inequality: If $X_{(n)}$ is the maximum of n iid random variables drawn from a distribution F, whose inverse function is denoted by F^{-1}, then

$$E[X_{(n)}] \leq F^{-1}\left(1 - \frac{1}{n}\right) + n\int_{F^{-1}(1-\frac{1}{n})}^{\infty} (1 - F(x))dx. \quad \text{(See Lai and Robbins (1976).)}$$

We are going to use also several analytic inequalities, e.g. Jensen, etc.

Our principal strategy is to first analytically bound the standard normal quantile function $z_p = \Phi^{-1}(1-p)$ and then transform those to bounds on the mean and the median of $X_{(n)}$. For example, it is proved in this article (Corollary 2.1) that for $p \leq 0.1$,

$$z_p \leq \sqrt{2\log t} - \frac{\log 4\pi + \log\log t}{2\sqrt{2\log t}}\left(1 - \frac{K}{\log t}\right),$$

where $t = 1/p$ and the constant K can be taken to be 0.6403. Once again, if one wishes, simpler bounds can be extracted out from this bound, at the cost of some loss in accuracy.

In fact, the bounds on z_p, stated in Section 5, are two-sided, and may quite possibly be of some independent interest.

In Section 5, we move on to deriving higher order asymptotic expansions for the mean and the median of $X_{(n)}$. This is achieved by first accomplishing a three term asymptotic expansion for z_p as $p \to 0$ (eq. (23)). We then use the celebrated Rényi's representation to reduce the problem about $X_{(n)}$ to a problem about sample means of an iid standard exponential sequence. The integral representation (20) is the key, and it is applicable in general, not only in the normal case alone. Very careful collection of terms is then needed to produce the ultimate three term asymptotic expansions for the mean and the median of $X_{(n)}$ up to an error of order $O\left(\frac{(\log\log n)^2}{(\log n)^{5/2}}\right)$ (see Theorems 5.1 and 5.2). For example, Theorem 5.2 gives the result that the median of $X_{(n)}$ admits the three term asymptotic expansion

$$\begin{aligned}\mathrm{med}(X_{(n)}) = \sqrt{2\log n} - \frac{\log\log n + \log 4\pi + 2\log\log 2}{2\sqrt{2\log n}} + \frac{1}{8(2\log n)^{3/2}} \times \\ \times \left[(\log\log n - 2(1+\log\log 2))^2 + (\log 4\pi - 2)^2 - 4\log\log 2(\log 4\pi + 2\log\log 2)\right] \\ + O\left(\frac{(\log\log n)^2}{(\log n)^{5/2}}\right).\end{aligned}$$

In the standard Cauchy case, the same methods lead to the three term asymptotic expansion

$$\mathrm{med}(X_{(n)}) = \frac{n}{\pi\log 2} + \frac{1}{2\pi} - \frac{(4\pi^2-1)\log 2}{12n\pi} + O(n^{-2}).$$

Apart from the difficulty of the intermediate calculation, the asymptotic expansions can be practically useful when n is truly fantastically large and therefore the mean or the median cannot be reliably calculated.

The article ends with a topical application. For identifying the sparse and weak true signals in the Donoho-Jin sparse Gaussian mixture model (Donoho and Jin (2004)), the $\sqrt{2\log n}$ threshold used in the Donoho-Johnstone hard thresholding estimate is replaced by some substitutes guided by the theorems in this article. Our theorems show that these adjustments lead to an improvement in the fraction of true signals that get discovered, at the expense of somwhat looser control of the total number of false discoveries. The overall conclusion we have arrived at is that although theoretically $E[X_{(n)}] \sim \sqrt{2\log n}$, the $\sqrt{2\log n}$ proxy for $E[X_{(n)}]$ is indeed a poor one practically, being too much of an overestimate (see Table 3 below). We urge that this approximation should be discontinued unless n is fantastically large.

One may wonder if the fixed n bounds we have derived could have been obtained by using *chaining techniques* of empirical process theory. We believe that the answer is *no*. Our fixed n bounds are so tight, that chaining, which is a versatile tool but weak in sharpness, cannot produce the correction terms with the accuracy we have provided. However, while our methods do not obviously apply to dependent situations, chaining will provide something in well controlled dependent situations with a subgaussian tail. They may be too weak; we do not know at this time.

All proofs are deferred to Section 7, for easy reading of the nontechnical parts.

2 New Two-sided Analytic Bounds on the Normal Quantile

In this section we derive upper and lower bounds on the standard normal quantile function $z_p = \Phi^{-1}(1-p)$. These bounds are a precursor for finding bounds on the mean and the median of $X_{(n)}$ in the next sections. The analytic bounds on z_p could be of some independent theoretical interest.

We need a result from Szarek and Werner (1999) stating that there exist nonnegative constants a, b, c, d, such that the following inequality is satisfied:

$$1 - \Phi(z) \leq \frac{a\varphi(z)}{bz + \sqrt{c^2z^2 + d}} \quad \text{for all} \quad z > 0. \tag{1}$$

We assume that $p \leq p_0$ (p_0 will be specified), and use the following notations:

$$t_0 = \frac{1}{p_0},\ \beta_0 = -\frac{\log(z_{p_0}/\sqrt{2\log t_0})}{1 - z_{p_0}/\sqrt{2\log t_0}},\ z_{p_0} = \Phi^{-1}(1-p_0). \tag{2}$$

In this case we introduce more notations which will be used further on:

$$\alpha = b^2 + c^2,\ \delta = \frac{d}{2\alpha z_{p_0}^2}. \tag{3}$$

We are going to establish a new upper bound on the standard normal quantile z_p.

Theorem 2.1. *For $p \leq p_0$ (p_0 will be specified later) and $t = 1/p$, we have*

$$z_p \leq \sqrt{2\log t} - \frac{\log(4\alpha\pi/a^2) + \log\log t}{2\sqrt{2\log t}}\left(1 - \frac{\beta_0}{2\log t}\right) - \frac{d(1-\delta)}{2\alpha(2\log t)^{3/2}}\left(1 - \frac{\beta_0}{2\log t}\right),$$

where $t_0, \beta_0, \alpha, \delta$ are defined by (2) and (3) and a, b, c, d are chosen according to (1).

An important consequence of this general upper bound is the following corollary.

Corollary 2.1. *For $p \leq 0.1$ and $t = 1/p$,*

$$z_p \leq \sqrt{2\log t} - \frac{\log 4\pi + \log\log t}{2\sqrt{2\log t}}\left(1 - \frac{0.6403}{\log t}\right).$$

The upper bound in Corollary 2.1 comes out in its best colors when p is 'small', i.e., when we want a fully analytic razor sharp upper bound on an extreme quantile. The table below testifies to it.

Table 1

p	z_p	Upper bound of Corollary 2.1
10^{-2}	2.326	2.459
10^{-3}	3.090	3.172
10^{-4}	3.719	3.777
10^{-5}	4.265	4.309
10^{-6}	4.753	4.789
10^{-7}	5.199	5.229
10^{-9}	5.998	6.019
10^{-11}	6.706	6.723
10^{-13}	7.349	7.362
10^{-15}	7.949	7.953

To attack the corresponding lower bound problem correctly, one has to restart with a lower bound on the tail of the standard normal distribution function. If we use the lower bound $1-\Phi(z)\geq\varphi(z)[\frac{1}{z}-\frac{1}{z^3}],\ z>0$, and carry out calculations similar to the ones we use in proving Theorem 2.1, then the following lower bound on z_p can be derived.

Theorem 2.2. *For $p\leq p_0$ with p_0 specified such that $z_{p_0}>1$, we have*

$$z_p\geq\sqrt{2\log t}-\frac{\log 4\pi+\log\log t-2\log k_0}{2\sqrt{2\log t}}\left(1-\frac{1}{4\log t}+\frac{1}{16(\log t)^2}\right),$$

where $k_0=(z_{p_0}^2-1)/z_{p_0}^2$.

3 Fixed n Bounds on the Median of $X_{(n)}$

The bounds in Section 2 on the standard normal quantile function are used here to obtain bounds on the median of $X_{(n)}$. We remember that $X_{(n)}$ is the maximum of an iid standard normal sample. Let us note that the bounds on the median do not follow immediately from the bounds in Section 2. A fair amount of new work is still needed, although the bounds in Section 2 form a critical ingredient. For brevity, we present only the upper bounds on the median, although explicit lower bounds may also be derived. We let θ_n denote the median of $X_{(n)}$, i.e. $\theta_n=\mathrm{med}(X_{(n)})$.

Theorem 3.1. *Let p_0 be specified, and β_0 is as defined in (2). Let*

$$n_0=\frac{\log 2}{-\log(1-p_0)},\ a_0=\log\frac{2^{1/n_0}}{2^{1/n_0}-1},\ b_0=\frac{4}{\log a_0+\log 4\pi-2},$$

$$n_1=\frac{\log 2}{\beta_0(3+b_0)-\log[e^{\beta_0(3+b_0)}-1]}.$$

Then for any integer $n\geq\max\{n_0,n_1\}$, one has

$$\begin{aligned}
\theta_n &\leq \sqrt{2\log n}\left(1-\frac{2\log\log 2-\frac{\log 4}{n}}{4\log n}\right)\\
&- \frac{\log 4\pi+\log\log n}{2\sqrt{2\log n}}\left(1-\frac{\beta_0}{\log n}\right)\left(1-\frac{\frac{\log 2}{n}-\log\log 2}{2\log n}\right) \qquad (4)\\
&= \sqrt{2\log n}-\frac{\log 4\pi+\log\log n+2\log\log 2-\frac{\log 4}{n}}{2\sqrt{2\log n}}\\
&+ \frac{(\log 4\pi+\log\log n)(\frac{\log 2}{n}-\log\log 2+2\beta_0)}{2(2\log n)^{3/2}}\\
&- \frac{\beta_0(\log 4\pi+\log\log n)(\frac{\log 2}{n}-\log\log 2)}{(2\log n)^{5/2}}. \qquad (5)
\end{aligned}$$

A bound which is less accurate, but simpler and hence easier to use, is given bellow.

Corollary 3.1. *For $n \geq 31$, one has*

$$\begin{aligned}\theta_n &\leq \sqrt{2\log n} - \frac{\log 4\pi + \log\log n + 2\log\log 2 - \frac{\log 4}{n}}{2\sqrt{2\log n}}\left(1 - \frac{1.29}{\log n}\right) \\ &\leq \sqrt{2\log n} - \frac{\log 4\pi + \log\log n - 1}{6\sqrt{2\log n}}. \end{aligned} \tag{6}$$

Once again, we first provide a small table on the accuracy of the simple bound (6). The values of θ_n in the table below are exact; they are obtained by using the exact formula for θ_n.

Table 2

n	θ_n	Upper bound (6)
50	2.204	2.625
500	2.992	3.367
5000	3.636	3.979
25000	4.031	4.358
10^5	4.346	4.660
10^7	5.267	5.551

We see, as expected, that the upper bound in (6) becomes accurate, in terms of both absolute and relative error, for large n.

4 Fixed n Bounds on the Mean $E[X_{(n)}]$

Intuitively, the mean of $X_{(n)}$ is approximately equal to $\Phi^{-1}(1 - \frac{1}{n})$, in the sense that $E[X_{(n)}]/\Phi^{-1}(1 - \frac{1}{n}) \to 1$ as $n \to \infty$. Therefore, our bounds on the standard normal quantile function derived in Section 2 should be useful in deriving nonasymptotic bounds on $E[X_{(n)}]$. This is what is done here. In addition to the bounds in Section 2, an inequality from Lai and Robbins (1976) will be very useful, and it is stated as a lemma.

Lemma 4.1. *Let F be the CDF of a real valued random variable X with $E[|X|] < \infty$, and let $a_n = F^{-1}(1 - \frac{1}{n})$. Then, with F^{-1} denoting the inverse function to F, we have*

$$E[X_{(n)}] \leq a_n + n\int_{a_n}^{\infty}(1 - F(x))dx, \;\; n \geq 2. \tag{7}$$

A consequence of the lemma is the following bound on $E[X_{(n)}]$ in the Gaussian case. We will refine this to a more explicit form later in this section.

Corollary 4.1. *Let a, b, c, d be nonnegative constants for which bound (1) holds and as before, $X_{(n)}$ be the maximum of iid standard normal random variables $X_1, \ldots, X_n$. Then, for all $n \geq 7$,*

$$E[X_{(n)}] \leq z_{1/n} + \frac{az_{1/n}}{(b+c)z_{1/n}^2 + \sqrt{c^2 z_{1/n}^2 + d} - c}. \tag{8}$$

In particular, for all $n \geq 7$, one has the inequality

$$E[X_{(n)}] \leq z_{1/n} + \frac{1}{z_{1/n}}, \tag{9}$$

and even the stronger inequality

$$E[X_{(n)}] \leq z_{1/n} + \frac{2z_{1/n}}{2z_{1/n}^2 + 1}. \tag{10}$$

Our next task is to convert Corollary 4.1 into explicit bounds for the mean as announced in the introduction. This is the content of the next theorem.

Theorem 4.1. *Let $X_1, \cdots, X_n$ be iid standard normal. For all $n \geq 7$,*

$$E[X_{(n)}] \leq \sqrt{2\log n} - \frac{\log 4\pi + \log\log n - 2}{2\sqrt{2\log n}} + C\frac{\log 4\pi + \log\log n}{(2\log n)^{3/2}}, \tag{11}$$

where the constant C may be taken to be 1.5 *(or larger).*

Comment: There is something interesting and important about the bound in (11). Indeed, it is known that in the maximally dependent case, the mean $E[X_{(n)}]$ satisfies the following relation, see eq. (17) in Lai and Robbins (1976):

$$E[X_{(n)}] \leq \sqrt{2\log n} - \frac{\log 4\pi + \log\log n - 2}{2\sqrt{2\log n}} + O\left(\frac{(\log\log n)^2}{(\log n)^{3/2}}\right) \quad \text{for large} \quad n.$$

Compare this with (11). It is interesting to see that the two leading terms of this *asymptotic representation* in the maximally dependent case are exactly the same as in the *nonasymptotic bound* on the mean of $X_{(n)}$ in Theorem 4.1, while Theorem 4.1 gives a better error term than the Lai-Robbins result.

5 Asymptotic Expansions and Application to Donoho-Jin Model

We mentioned in the introduction that the common practice of approximating $E[X_{(n)}]$ by $\sqrt{2\log n}$ is not practically sound. The upper bound in Theorem 4.1 offers an alternative approximation. A third approach is outlined below, and then the three approximations are compared in Table 3, which shows without any doubt how poor the $\sqrt{2\log n}$ term is. Here is the third approach.

The idea is to approximate $E[X_{(n)}]$ by a suitable standard normal quantile $\Phi^{-1}(1-\delta_n)$. One may first consider using $\delta_n = 1/n$ for at least two reasons. First, it follows from suitable strong laws for sample maxima that in the normal case $X_{(n)}/\sqrt{2\log n} \stackrel{a.s.}{\to} 1$ as $n \to \infty$ (see DasGupta (2008), Chapter 8 for a proof). A second reason is that we can write $X_n = \Phi^{-1}(1-\xi_n)$ with $\xi_n = 1 - \Phi(X_{(n)})$, and the quantile transformation shows that $n\xi_n$ converges in distribution to a standard exponential random variable, say ξ. Thus, heuristically,

$$X_{(n)} = \Phi^{-1}(1-\xi_n) \approx \Phi^{-1}(1 - \frac{1}{n}\xi) \approx \Phi^{-1}(1 - \frac{1}{n}).$$

However, these heuristics are just not sufficiently effective. We can do better. In fact, the approximation $E[X_{(n)}] \approx \Phi^{-1}(1 - \frac{1}{n}e^{-\gamma})$, where γ denotes Euler's constant, is a *far better* *practical approximation to* $E[X_{(n)}]$ *than either* $\sqrt{2\log n}$ *or* $\Phi^{-1}(1-\frac{1}{n})$. Here is a precise result.

Proposition 5.1. *Define the number* δ_n *by the equation*

$$E[X_{(n)}] = \Phi^{-1}(1-\delta_n) \iff \delta_n = 1 - \Phi(E[X_{(n)}]).$$

Then $n\delta_n \to e^{-\gamma}$ *as* $n \to \infty$, *where* γ *denotes Euler's constant.*

This suggests that we may consider approximating $E[X_{(n)}]$ by $\Phi^{-1}(1-\frac{1}{n}e^{-\gamma})$. Below is a small table that demonstrates the poor performance of the $\sqrt{2\log n}$ approximation, the much better performance of the analytic bound in Theorem 4.1, and the amazingly good approximation obtained by using $\Phi^{-1}(1-\frac{1}{n}e^{-\gamma})$. The long standing practice of approximating $E[X_{(n)}]$ by $\sqrt{2\log n}$ should probably be discontinued. In the table below, $E[X_{(n)}]$ has been computed by numerical integration of its defining formula $E[X_{(n)}] = n\int_{-\infty}^{\infty} x\phi(x)\,\Phi^{n-1}(x)dx$.

Table 3

n	$E[X_{(n)}]$	$\sqrt{2\log n}$	Bound of Theorem 4.1	$\Phi^{-1}(1-\frac{1}{n}e^{-\gamma})$
50	2.25	2.80	2.73	2.28
500	3.04	3.53	3.34	3.06
10^4	3.85	4.29	4.06	3.86
10^6	4.86	5.26	5.01	4.87
10^8	5.71	6.07	5.82	5.71
10^9	6.09	6.44	6.19	6.09

A higher order asymptotic expansion for the sequence δ_n defined above may be derived by very carefully putting together the terms in the asymptotic expansion for $E[X_{(n)}]$. This derivation, in turn, requires higher order asymptotic expansions with remainder terms for z_p as an intermediate step, and these, we believe deserve to be more well known than they are. Additionally, the derivation of the asymptotic expansion for $E[X_{(n)}]$ to three terms is quite nontrivial. Also, we can imagine the higher order expansion being useful in some way in extreme value theory, or for high dimensional Gaussian inference. These reasons motivate us to present the following first theorem in this section.

Theorem 5.1. *Let, as before,* $X_1, X_2, \cdots$ *be an iid standard normal sequence and* γ *denote Euler's constant. Then the mean* $E[X_{(n)}]$ *admits the following three term asymptotic expansion:*

$$\begin{aligned} E[X_{(n)}] &= \sqrt{2\log n} - \frac{\log\log n + \log 4\pi - 2\gamma}{2\sqrt{2\log n}} + \frac{1}{8(2\log n)^{3/2}} \times \\ &\times \left[(\log\log n - 2(1-\gamma))^2 + (\log 4\pi - 2)^2 + 4\gamma(\log 4\pi - 2\gamma) - \frac{2}{3}\pi^2\right] \\ &+ O\left(\frac{(\log\log n)^2}{(\log n)^{5/2}}\right) \quad \textit{for large } n. \end{aligned}$$

Our method of proof of Theorem 5.1 is such that it automatically delivers a three term asymptotic expansion also for the median of $X_{(n)}$. Below is the corresponding result. Notice the remarkable similarity between the asymptotic expansion for the mean in Theorem

5.1 and that for the median in Theorem 5.2 below. The asymptotic expansions should also be compared with the bounds for fixed n in Theorem 4.1 and Corollary 3.1. The conclusion is that the nonasymptotic bounds in Theorem 4.1 and Corollary 3.1 are rather tight, and that there is not much scope for improvement in those nonasymptotic bounds.

Theorem 5.2. *Let $X_1, X_2, \cdots$ be an iid standard normal sequence. Then the median of $X_{(n)}$ admits the following three term asymptotic expansion:*

$$\begin{aligned}
\theta_n &= \mathit{med}\ (X_{(n)}) = \sqrt{2\log n} - \frac{\log\log n + \log 4\pi + 2\log\log 2}{2\sqrt{2\log n}} + \frac{1}{8(2\log n)^{3/2}} \times \\
&\times \Big[(\log\log n - 2(1+\log\log 2))^2 + (\log 4\pi - 2)^2 - 4\log\log 2(\log 4\pi + 2\log\log 2)\Big] \\
&+ O\Big(\frac{(\log\log n)^2}{(\log n)^{5/2}}\Big) \quad \textit{for large} \ \ n.
\end{aligned}$$

An important potential application below ends this subsection.

Example 5.1. Adjusting the Donoho-Johnstone Estimate. The sparse signal detection problem is one of detecting a very small percentage of relatively weak signals hidden among a large number of pure noises. It is evidently a very difficult problem, and usually we fail to recognize most of these weak signals. The theoretical development has largely concentrated on Gaussian signals. A popular model is the following mixture of two normals:

$$X_{i,n} \overset{indep.}{\sim} (1-\varepsilon_n)\, N(0,1) + \varepsilon_n\, N(\mu_{i,n}, 1).$$

Here $N(0,1)$ and $N(\mu_{i,n},1)$ are normal random variables both with variance 1 and with mean values 0 and $\mu_{i,n}$, respectively, while ε_n is taken to be a numerical sequence converging to zero at a suitable power rate. See Donoho and Jin (2004) for more details on this model. Donoho and Johnstone (1994), motivated by Gaussian extreme value theory, gave the *hard thresholding estimate*

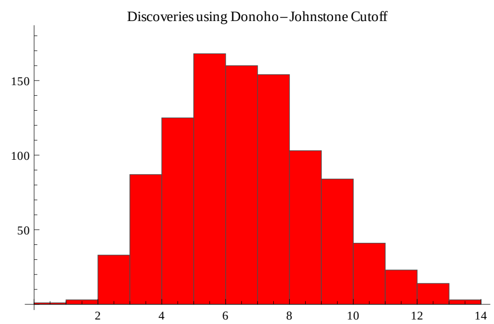
$$\hat{\mu}_i = X_i I_{\{|X_i| > \sqrt{2\log n}\}}.$$

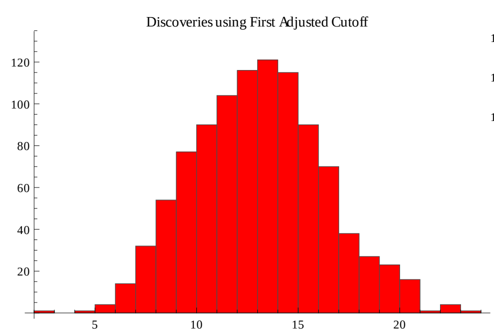
The rough motivation is that if the recorded signal exceeds the value $E[X_{(n)}]$, which is the average value of the maximum among n pure noises, only then we treat it as a possible true signal.

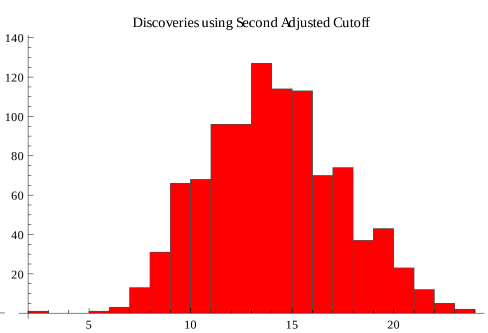
We have performed simulations which suggest that the $\sqrt{2\log n}$ proxy for $E[X_{(n)}]$ in the Donoho-Johnstone estimate is practically speaking too conservative, in the sense that a great many of the true signals will not be picked up by using $\sqrt{2\log n}$ as the yardstick. In contrast, if we use, for example, $\Phi^{-1}(1-\frac{1}{n}e^{-\gamma})$ as the threshold, then generally,every signal picked up by the original Donoho-Johnstone estimate is picked up, and about twice as many true signals get picked up by adjusting the threshold. Formally, we look at the following three adjustments of the usual Donoho-Johnstone estimate:

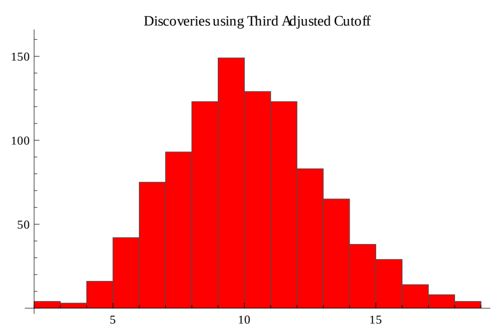
$$\begin{aligned}
\hat{\mu}_{i,1} &= X_i I_{A_{i,1}}, \ \text{ where } \ A_{i,1} = \{|X_i| > \Phi^{-1}(1-\frac{1}{n}e^{-\gamma})\}, \\
\hat{\mu}_{i,2} &= X_i I_{A_{i,2}}, \ \text{ where } \ A_{i,2} = \{|X_i| > \Phi^{-1}(1-\frac{1}{n}\log 2)\}, \\
\hat{\mu}_{i,3} &= X_i I_{A_{i,3}}, \ \text{ where } \ A_{i,3} = \{|X_i| > \sqrt{2\log n} - (\log\log n)/2\sqrt{2\log n}\}.
\end{aligned}$$

The improved performance of the three adjusted estimates is indicated below by graphics. The important conclusion is that the original Donoho-Johnstone estimate picks up on an average 6.17 of a total of 60 true signals, and the adjusted Donoho-Johnstone estimates pick up up to about 14. The second adjusted estimate with an adjusted threshold of $\Phi^{-1}(1 - \frac{1}{n}\log 2)$, the choice being motivated by Theorem 5.2, picks up the most (13.64 on the average). The common value of the signals was taken to be $\mu_i \equiv \sqrt{\log n} = 3.03$, with $n = 10,000$. The simulation was repeated 1000 times and the histograms report the number of signals discovered over those 1000 simulations. There is scope for demonstrating this advantage of the adjusted estimates theoretically, and we provide below some concrete theorems that quantify this advantage.









5.1 Theoretical Advantage of Adjusted Estimates

For theoretically demonstrating the advantage of estimates that adjust the $\sqrt{2\log n}$ threshold, we use the Donoho-Jin (2004) model. In the context of doing formal inference with microarray data, Donoho and Jin (2004) formulated a mixture model for gene expression levels which has since been seriously studied and adapted or innovated by various authors;

a few specific references are Hall and Jin (2008, 2010), Cai, Xu, and Zhang (2009), and Cai and Wang (2011). The model in the 2004 article considered iid data with a null model that articulates the case that absolutely nothing is going on, and a particular alternative model that says that a small fraction of observations contain a possibly detectable signal. The Donoho-Jin model is easily seen to be mathematically the same as our description here: we have a sequence of iid standard normals, $Z_1, Z_2, \cdots$, a contemplated signal at the level $\mu_n = \sqrt{2r \log n}, 0 < r < 1$, and for each n, a collection of iid Bernoullis, $B_1, \cdots, B_n \sim Ber(\epsilon_n)$, where $\epsilon_n = n^{-\beta}, 1/2 < \beta < 1$. These Bernoullis, assumed independent of the Z's, remain hidden from our eyes, but in conjunction with the standard normals, produce for us our sample values, $X_{i,n} = Z_i + \mu_n B_i, i = 1, \cdots, n$; so, we have, at hand a triangular array of row iid random variables, rather than one long string of iids. The assumptions $1/2 < \beta < 1$ and $r < 1$ are there for excellent reasons; they are there to keep the problem from falling into the realm of trivialities. We record the canonical Donoho-Jin paradigm for future reference:

$$
\begin{gathered}
Z_1, Z_2, \cdots \overset{iid}{\sim} N(0,1); \\
\text{For}\, n \geq 1, \epsilon_n = n^{-\beta},\ \mu_n = \sqrt{2r \log n},\ 1/2 < \beta < 1,\ 0 < r < 1; \\
\text{For each given}\, n, B_1, \cdots, B_n \overset{iid}{\sim} Ber(\epsilon_n); \\
\text{For each given}\, n,\ \{B_1, \cdots, B_n\} \text{ are independent of } \{Z_1, \cdots, Z_n\}; \\
\text{For each given}\, n,\ X_{i,n} = Z_i + \mu_n B_i, i = 1, 2, \cdots, n. \qquad (12)
\end{gathered}
$$

The original Donoho-Jin rule flags an observation X_i if it exceeds $c_n = \sqrt{2 \log n}$. Our intention is to document that we are bound to miss essentially all the signals if we use this rule without adjustments. The results below show how we can adjust the threshold sequence c_n so that asymptotically we pick up some, or at least more, of the genuine signals.

We need some notation, which is standard, but needs to be given. We partition the n observations into the familiar four categories, observations that get flagged and happen to be genuine signals (discoveries D), observations that get flagged but are actually noises (false discoveries F), observations that happen to be signals but we don't flag them (missed discoveries M), and observations that we do not flag and also happen to be noises (the junk J). We let $D + M = S$, the number of true signals, $F + J = N$, the number of pure noises, $D + F = L$, the number that we think deserves a look because we flagged them, and $M + J = I$, the number that we think may be ignored because we didn't flag them. Note that obviously, $S + N = L + I = n$; D, F, M, J, S, N, L, I are all sequences, although the sequence notation is suppressed here purely for brevity. Often, these are presented in the form of a familiar table as below:

Table 4

	Signal	Noise	
Flagged	D	F	L
Unflagged	M	J	I
	S	N	n

The popular index *FDR* is the ratio $\frac{F}{L}$ and the missed discovery rate is the ratio $\frac{M}{S}$; we like these to be small. On the other hand, the discovery rate is $\frac{D}{S}$; we like $\frac{D}{S}$ to be large. Generally, we have placed much more emphasis on keeping $\frac{F}{L}$ small over keeping $\frac{M}{S}$ small, although there is an obvious tension between these two demands. It is in fact counterproductive to make either one too small, because then the other one soars. We document that with the $\sqrt{2\log n}$ threshold,

$$\frac{F}{L} \xrightarrow{P} 0, \ \frac{D}{S} \xrightarrow{P} 0.$$

Thus, with the $\sqrt{2\log n}$ cutoff, a perfect score on the false discovery front is won in return for missing *all the signals* asymptotically. We show that adjusting the threshold helps. Here are a few theorems that quantify this advantage.

Theorem 5.3. *Consider the Donoho-Jin model with the cutoff* $c_n = \sqrt{2\log n}$. *Then,*

$$(a) F \xrightarrow{P} 0.$$

$$(b) \frac{F}{L} \xrightarrow{P} 0.$$

$$(c) \frac{D}{S} \xrightarrow{P} 0; \quad \frac{M}{S} \xrightarrow{P} 1.$$

$$(d) \frac{D}{S} = \frac{n^{-(1-\sqrt{r})^2}}{2\sqrt{\pi}(1-\sqrt{r})\sqrt{\log n}}(1+o_p(1)).$$

This theorem tells us that use of the $\sqrt{2\log n}$ *cutoff results in unacceptably low values for the discovery rate* $\frac{D}{S}$. *We need less conservative cutoffs to improve the true discovery rate. If we allow some false discoveries, in return it will buy us an improved value for* $\frac{D}{S}$. *A useful question to ask is how we should adjust the* $\sqrt{2\log n}$ *cutoff to something less conservative in order to allow a tolerable number of false discoveries. An answer to this question will guide us to finding adjusted cutoffs for improving our discovery rate* $\frac{D}{S}$. *In the following theorem, we can keep things general by using a general sequence* ϵ_n *for the proportion of observations that are not pure noises; the only assumptions necessary are that* $\epsilon_n \to 0$.

Theorem 5.4. *Consider the Donoho-Jin model with a general* ϵ_n *and a general* μ_n *(not necessarily* $n^{-\beta}$ *and* $\sqrt{2r\log n}$*). Let* $0 < \lambda < \infty$ *be a fixed constant, and let*

$$c_n = \sqrt{2\log n} - \frac{\log\log n + \log(4\pi\lambda^2)}{2\sqrt{2\log n}}. \tag{13}$$

Then, the number of false discoveries F *satisfies*

$$F \stackrel{\mathcal{L}}{\Rightarrow} Poi(\lambda).$$

Remark: *In effect, Theorem 5.4 is saying that if we replace the traditional thresholding sequence* $\sqrt{2\log n}$ *by the sharper thresholding sequence essentially as in (13), then* F *will admit Poisson asymptoctics, instead of collapsing to zero. A net result of this will be the desired outcome that the discovery rate* $\frac{D}{S}$ *will also improve; by adjusting* c_n *as proposed above, we sarifice some in the false discovery front, but gain on the true discovery front.* Exactly how much is the gain on the true discovery front by adjusting the $\sqrt{2\log n}$ cutoff to the less conservative cutoff of Theorem 5.4? The next result quantifies the order of that gain.

Theorem 5.5. *Consider the Donoho-Jin model (12) with* $\mu_n = \sqrt{2r \log n}, 0 < r < 1$, *and a general sequence* ϵ_n. *Let* c_n *be as in (13). Then,*

$$\frac{D}{S} = (4\pi\lambda^2 \log n)^{\frac{1-\sqrt{r}}{2}} \times \frac{n^{-(1-\sqrt{r})^2}}{2\sqrt{\pi}(1-\sqrt{r})\sqrt{\log n}}(1 + o_p(1)).$$

Remark: Comparing with part (d) of Theorem 5.3, we find that by adjusting c_n to the level as in (13), we improve the true discovery rate $\frac{D}{S}$ by the factor $(4\pi\lambda^2 \log n)^{\frac{1-\sqrt{r}}{2}}$. *So, adjusting the thresholding level causes a relative improvement in* $\frac{D}{S}$, *although* $\frac{D}{S}$ *will still converge in probability to zero.*

6 Other Possibilities

The methods of Theorem 5.1 and Theorem 5.2 may be useful for writing asymptotic expansions for the mean and the median of a more general extreme order statistic $X_{(n-k+1)}, 1 \leq k < \infty$; we need to use the corresponding Réyni representation. It may also be possible to derive formal expansions for the median (and when applicable, the mean) of $X_{(n)}$ for nonnormal parents, e.g., the double exponential or Cauchy.

To be specific, for example, consider the standard Cauchy case, and let as usual $X_{(n)}$ denote the maximum of a sample of size n. Since $E[X_{(n)}]$ does not exist for any n, it is not sensible to talk about an asymptotic expansion for it. However, the methods of Theorem 5.2 lead to the following correct asymptotic expansion to the median of $X_{(n)}$:

$$\text{med}(X_{(n)}) = \frac{n}{\pi \log 2} + \frac{1}{2\pi} - \frac{(4\pi^2 - 1)\log 2}{12n\pi} + O(n^{-2}). \tag{14}$$

In (14), the very first term $\frac{n}{\pi \log 2}$ follows from the weak convergence fact that $\frac{\pi}{n} X_{(n)} \stackrel{\mathcal{L}}{\Rightarrow} T^{-1}$, where T is a standard exponential; but the entire asymptotic expansion does not follow from mere weak convergence. It is really quite surprising how fantastically accurate the expansion in (14) is. With $n = 50$, the exact value of the median of $X_{(n)}$ is 23.1063; the one term expansion gives the approximation 22.9612, the two term 23.1204, and the three term expansion gives the approximation 23.1062, almost the exact value.

Acknowledgement. From the first author: Kesar Singh and I came to the ISI in Calcutta the same year. He came from Allahabad and I was already from Calcutta. We were both staying in the dorms, and I used to see him a few times on most days at the dining room. The news had rapidly spread what a smart and sweet person he is. I remember looking up to him as a 15 year old. That respect only increased as the years passed. Kesar was a tower among my friends. He was also about the gentlest soul I ever met. His work and personality continue to have my greatest admiration; he was an icon in my eyes. I am just so sorry that the good leave young.

From the second author: Kesar Singh and I shared the ISI connection, but he had already left for Stanford when I joined the ISI as a student in 1981. I first met him in person at the Bootstrap conference at the Michigan State University in 1990. He had always been an inspiring figure for me, for his deep contribution to the theory of Bootstrap and to several other areas of Statistics. He was a gentle man in the true sense of the word and particularly nice to his younger colleagues. It is very sad that he left us so early.

It is a pleasure to thank Tony Cai, Holger Drees, Peter Hall, Dominique Picard and Sara van de Geer for their gracious input in preparing this article.

7 Proofs

7.1 Proof of Corollary 2.1

Consider the inequality $1-\Phi(z) \le \frac{\phi(z)}{z}$, $z>0$. In the notation of Theorem 2.1, this makes $a=1, b=1, c=d=0, \alpha=1, \delta=0$, and since $t_0 = 1/p_0 = 10$, we have

$$\beta_0 = -\log \frac{z_{0.1}}{\sqrt{2\log 10}} / \left(1 - \frac{z_{0.1}}{\sqrt{2\log 10}}\right) = 1.28052 \ \Rightarrow\ \frac{1}{2}\beta_0 = 0.64026 < 0.6403.$$

If we plug these values into the bound of Theorem 2.1, the corollary follows.

7.2 Proof of Theorem 2.1

Since the constants a, b, c, d are chosen such that $1-\Phi(z) \le a\varphi(z)/(bz + \sqrt{c^2z^2+d})$ for all $z>0$, see also (1), by using $z = z_p$, we have the following relations:

$$\begin{aligned} 1-\Phi(z) &= p \le \frac{a}{\sqrt{2\pi}e^{z^2/2}(bz+\sqrt{c^2z^2+d})} \\ &\Rightarrow \quad e^{z^2/2}(bz+\sqrt{c^2z^2+d}) \le \frac{a}{\sqrt{2\pi}p} = \frac{at}{\sqrt{2\pi}} \\ &\Rightarrow \quad e^{z^2/2}(\sqrt{b^2z^2+c^2z^2+d}) \le \frac{at}{\sqrt{2\pi}}. \end{aligned}$$

Thus, recalling the notation $\alpha = b^2+c^2$, we obtain the inequality

$$2\log t + \log\frac{a^2}{2\pi} \ge z^2 + \log(\alpha z^2 + d) = z^2 + \log\alpha + \log z^2 + \log\left(1 + \frac{d}{\alpha z^2}\right). \tag{15}$$

Using the inequality $\log(1+y) \ge (1-\delta)y$ for $y \le 2\delta$ with $y = \frac{d}{\alpha z^2}$, we have from (12),

$$2\log t + \log\frac{a^2}{2\pi\alpha} \ge z^2 + 2\log z + (1-\delta)\frac{d}{\alpha z^2}. \tag{16}$$

Clearly, (13) is a transcendental inequality for z, hence not easy to deal with. To avoid this difficulty we use a critical technical trick of writing $z = z_p$ in the form

$$z = \sqrt{2\log t} - \frac{\rho}{\sqrt{2\log t}},\ t > 1,$$

where $\rho = \rho(t) > 0$. This will enable us to find first bounds for the variable ρ and transform them to bounds for z. Indeed, from this representation of z, we get

$$z^2 = 2\log t + \frac{\rho^2}{2\log t} - 2\rho = 2\log t\left(1 - \frac{\rho}{2\log t}\right)^2,$$

$$\log z = \frac{1}{2}\log 2 + \frac{1}{2}\log\log t + \log(1 - \frac{\rho}{2\log t}),$$

and write the inequalities

$$\frac{1}{z^2} \ge \frac{1}{2\log t},\ \log z \ge \frac{1}{2}\log 2 + \frac{1}{2}\log\log t - \beta_0\frac{\rho}{2\log t}.$$

We have used above the number β_0 as introduced by (2) and the assumption that $p \le p_0$, so $t = 1/p \ge t_0 = 1/p_0$. Therefore, from (15),

$$2\log t + \frac{\rho^2}{2\log t} - 2\rho + \log 2 + \log\log t - \beta_0 \frac{\rho}{\log t} + \frac{d(1-\delta)}{2\alpha \log t} \le 2\log t + \log \frac{a^2}{2\pi\alpha}. \tag{17}$$

For convenience, let us introduce two notations:

$$A = \beta_0 + 2\log t, \;\; B = 2\log t \left(\log \frac{4\alpha\pi}{a^2} + \log\log t + \frac{d(1-\delta)}{2\alpha \log t} \right).$$

Thus (16) is equivalent to the following inequality:

$$\rho^2 - 2A\rho + B \le 0. \tag{18}$$

As a function of ρ, the expression $\rho^2 - 2A + B$ is a quadratic convex function. The quadratic equation $\rho^2 - 2A\rho + B = 0$ has two real and positive roots,

$$\rho_1 = A - \sqrt{A^2 - B}, \;\; \rho_2 = A + \sqrt{A^2 - B}.$$

This implies that in (17) the values of ρ are between the two roots, i.e. $\rho_1 < \rho < \rho_2$. Consider the lower bound:

$$\begin{aligned} \rho &\ge \rho_1 = A - \sqrt{A^2 - B} = \frac{B}{A + \sqrt{A^2 - B}} \ge \frac{B}{A} \\ &\ge \frac{2\log t \left(\log \frac{4\alpha\pi}{a^2} + \log\log t + \frac{d(1-\delta)}{2\alpha \log t} \right)}{2(\beta_0 + 2\log t)} \\ &\ge \frac{1}{2} \left(\log \frac{4\alpha\pi}{a^2} + \log\log t + \frac{d(1-\delta)}{2\alpha \log t} \right) \left(1 - \frac{\beta_0}{2\log t} \right). \end{aligned}$$

We use this lower bound on ρ to derive an upper bound of z :

$$\begin{aligned} z = z_p &= \sqrt{2\log t} - \frac{\rho}{\sqrt{2\log t}} \\ &\le \sqrt{2\log t} - \frac{1}{2\sqrt{2\log t}} \left(\log \frac{4\alpha\pi}{a^2} + \log\log t + \frac{d(1-\delta)}{2\alpha \log t} \right) \left(1 - \frac{\beta_0}{2\log t} \right). \end{aligned}$$

This is exactly the upper bound for z_p as stated in Theorem 2.1.

Comment. It is clear that if we start with the upper bound for ρ, namely, $\rho \le \rho_2$, and follow almost the same steps as above, at the end we will derive a lower bound for the quantile z_p. If we denote by z^* the upper bound and z_* the lower bound for the quantile z_p, then of interest is the interval (z_*, z^*) and its length $\Delta = z^* - z_*$. We easily find that $\Delta = (\rho_1 + \rho_2)/\sqrt{2\log t}$, where ρ_1, ρ_2 are the two real roots used above. Since $\rho_1 + \rho_2 = 2A = 2\beta_0 + 4\log t$ we obtain an explicit expression for Δ and can see its behavior as a function of t and hence of p.

7.3 Proof of Corollary 3.1

First note that if we choose $p_0 = 0.1$, then we get $\beta_0 = 0.6403$ and the conditions $n \geq n_0$ and $n \geq n_1$ work out respectively to $n \geq 6.58$ and $n \geq 30.55$. This explains why we need the condition $n \geq 31$ for the corollary.

Since the term corresponding to $(2\log n)^{-5/2}$ in Theorem 3.1 is negative, we can ignore it and get the upper bound

$$\begin{aligned}
\theta_n &\leq \sqrt{2\log n} - \frac{1}{2\sqrt{2\log n}}\left(\log 4\pi + \log\log n + 2\log\log 2 - \frac{1}{n}\log 4\right) \\
&+ \frac{1}{2(2\log n)^{3/2}}(\log 4\pi + \log\log n)\left(\frac{\log 2}{n} - \log\log 2 + 2\beta_0\right) \\
&= \sqrt{2\log n} - \frac{1}{2\sqrt{2\log n}}\left(\log 4\pi + \log\log n + 2\log\log 2 - \frac{\log 4}{n}\right)\left(1 - \frac{1}{\log n}h(n)\right),
\end{aligned}$$

where

$$h(n) = \frac{1}{2}\frac{(\log 4\pi + \log\log n)(\frac{\log 2}{n} - \log\log 2 + 2\beta_0)}{\log 4\pi + \log\log n + 2\log\log 2 - \frac{\log 4}{n}}.$$

Now we use the fact that as a function of n, $h(n)$ is increasing for $n \geq 7$ and is lower bounded by 1.29, and this gives the first statement of the corollary. For the second statement, use the bounds

$$2\log\log 2 - \frac{\log 4}{n} \geq 2\log\log 2 - \frac{\log 4}{7} > -1, \text{ and } 1 - \frac{1.29}{\log n} \geq 1 - \frac{1.29}{\log 7} > \frac{1}{3}.$$

Plug these two bounds into the first statement of the corollary and get the second one.

7.4 Proof of Theorem 3.1

We start with noting that the median θ_n satisfies the relation $\Phi(\theta_n) = \frac{1}{2}$, and hence, $\theta_n = z_p = \Phi^{-1}(1-p)$, where $p = p_n = 1 - 2^{-1/n}$. Thus, in the notation of Theorem 2.1, we have $t = \frac{1}{p} = \frac{2^{1/n}}{2^{1/n}-1}$. Theorem 3.1 will be established by using the bound in Theorem 2.1 on z_p as the principal tool, although additional specific algebraic inequalities will also be needed. Referring to Theorem 2.1, with $t = \frac{2^{1/n}}{2^{1/n}-1}$, we will separately upper bound the terms $\sqrt{2\log t}$ and $-\frac{\log 4\pi + \log\log t}{2\sqrt{2\log t}}(1 - \frac{\beta_0}{\log t})$. We combine then these two bounds to produce the upper bound on θ_n as stated in Theorem 3.1.

The following inequality will be used in the proof of this theorem:

$$\log t = \log\left(\frac{2^{1/n}}{2^{1/n}-1}\right) = \frac{\log 2}{n} - \log(e^{\frac{\log 2}{n}} - 1) \leq \frac{\log 2}{n} - \log\log 2 + \log n \tag{19}$$

Hence

$$\begin{aligned}
\sqrt{2\log t} &= \sqrt{2\log n - 2\log\log 2 + \frac{\log 4}{n}} = \sqrt{2\log n}\sqrt{1 - \frac{2\log\log 2 - \frac{\log 4}{n}}{2\log n}} \\
&\leq \sqrt{2\log n}\left(1 - \frac{2\log\log 2 - \frac{\log 4}{n}}{4\log n}\right).
\end{aligned} \tag{20}$$

We now turn our attention to bounding the expression $-\frac{\log 4\pi+\log\log t}{2\sqrt{2\log t}}(1-\frac{\beta_0}{\log t})$. For this, we will use the following calculus fact: The function

$$u(x) := \frac{\log 4\pi + \log x}{\sqrt{x}}\left(1-\frac{\beta_0}{x}\right) \tag{21}$$

is monotone decreasing for $x \geq x_1$, where

$$x_1 := \beta_0(3+y_0),\ \beta_0 \text{ is from (2)},\ y_0 = \frac{4}{\log x_0 + \log 4\pi - 2} \quad \text{and} \quad x_0 = \log \frac{2^{1/n_0}}{2^{1/n_0}-1}.$$

This can be proved by a direct differentiation of the function $u(x)$ and checking that its derivative is negative.

Now in (18) we identify x with $\log t$. After some algebra, we see that the condition $x \geq x_1$ is equivalent to

$$n \geq n_1 = \frac{\log 2}{\beta_0(3+y_0) - \log[e^{\beta_0(3+y_0)}-1]}.$$

This explains why we have the condition $n \geq n_1$ in the statement of Theorem 3.1.

The monotone decreasing property of the function $u(x)$ (under the condition $n \geq n_1$) together with the bound (16) above, give us

$$\begin{aligned}
&\frac{\log 4\pi + \log\log t}{2\sqrt{2\log t}}\left(1-\frac{\beta_0}{\log t}\right) \\
\geq\ & \frac{\log 4\pi + \log(\log n + \frac{\log 2}{n} - \log\log 2)}{2\sqrt{2}\sqrt{\log n + \frac{\log 2}{n} - \log\log 2}}\left(1 - \frac{\beta_0}{\log n + \frac{\log 2}{n} - \log\log 2}\right) \\
=\ & \frac{\log 4\pi + \log(\log n + \frac{\log 2}{n} - \log\log 2)}{2\sqrt{2\log n}(1+\frac{\frac{\log 2}{n}-\log\log 2}{\log n})^{1/2}}\left(1 - \frac{\beta_0}{\log n + \frac{\log 2}{n} - \log\log 2}\right) \\
\geq\ & \frac{\log 4\pi + \log(\log n + \frac{\log 2}{n} - \log\log 2)}{2\sqrt{2\log n}(1+\frac{\frac{\log 2}{n}-\log\log 2}{\log n})^{1/2}}\left(1 - \frac{\beta_0}{\log n}\right) \\
\geq\ & \frac{\log 4\pi + \log(\log n + \frac{\log 2}{n} - \log\log 2)}{2\sqrt{2\log n}}\left(1 - \frac{\frac{\log 2}{n} - \log\log 2}{2\log n}\right)\left(1 - \frac{\beta_0}{\log n}\right) \\
\geq\ & \frac{\log 4\pi + \log\log n}{2\sqrt{2\log n}}\left(1 - \frac{\frac{\log 2}{n} - \log\log 2}{2\log n}\right)\left(1 - \frac{\beta_0}{\log n}\right).
\end{aligned}$$

From our findings (16) and (17) we first get assertion (4) of Theorem 3.1 and after expanding (4), we arrive at assertion (5). This completes the proof of Theorem 3.1.

7.5 Proof of Corollary 4.1

We note that the inequality (7) is the key Lai and Robbins (1976) bound, and that (8) is an improvement on (7) in the Gaussian case. To prove Corollary 4.1, observe that by

Lemma 4.1,

$$\begin{aligned}
E[X_{(n)}] &\le z_{1/n} + n\int_{z_{1/n}}^{\infty}[1-\Phi(x)]dx = z_{1/n} + n\int_{z_{1/n}}^{\infty}\frac{1-\Phi(x)}{\varphi(x)}\varphi(x)dx \\
&\le z_{1/n} + n\int_{z_{1/n}}^{\infty}\frac{a}{bx+\sqrt{c^2x^2+d}}\varphi(x)dx \\
&\le z_{1/n} + n\frac{a}{bz_{1/n}+\sqrt{c^2z_{1/n}^2+d}}\int_{z_{1/n}}^{\infty}\varphi(x)dx \\
&\le z_{1/n} + \frac{a}{bz_{1/n}+\sqrt{c^2z_{1/n}^2+d}}.
\end{aligned}$$

We use now the inequality $\sqrt{c^2z_{1/n}^2+d} \ge cz_{1/n} + (\sqrt{c^2z_{1/n}^2+d}-c)/z_{1/n}$ which is valid for all n such that $z_{1/n} \ge 1$, i.e., for $n \ge 7$. This is where the assumption $n \ge 7$ is needed.

Plugging this bound on $\sqrt{c^2z_{1/n}^2+d}$ into (8), we get

$$\begin{aligned}
E[X_{(n)}] &\le z_{1/n} + \frac{a}{bz_{1/n}+cz_{1/n}+\left(\sqrt{c^2z_{1/n}^2+d}-c\right)/z_{1/n}} \\
&= z_{1/n} + \frac{az_{1/n}}{(b+c)z_{1/n}^2+\sqrt{c^2z_{1/n}^2+d}-c},
\end{aligned}$$

which is the assertion (8). Assertion (9) follows from (8) by using $a=b=1, c=d=0$ and the inequality $1-\Phi(z) \le \frac{\varphi(z)}{z}, z>0$. Finally, (10) follows from (8) by using $a = 4, b=3, c=1, d=8$, by virtue of the inequality $1-\Phi(z) \le 4\varphi(z)/(3z+\sqrt{z^2+8}),\ z>0$ (Szarek and Werner (1999)), see also relation (1).

7.6 Proof of Theorem 4.1

Let us mention that the statement in this theorem is based on inequality (8). However somewhat better results can be obtained by using the stronger inequality (9) instead.

The starting point for the proof is the result in Corollary 2.1 that

$$z_{1/n} \le \bar{z}_{1/n}, \quad \text{where} \quad \bar{z}_{1/n} := \sqrt{2\log n} - \frac{\log 4\pi + \log\log n}{2\sqrt{2\log n}}\left(1-\frac{\beta_0}{\log n}\right),$$

where β_0 as defined in (2) may be taken to be 0.6403. Note now that the function $x \mapsto x+\frac{1}{x}$ is increasing in x for $x \ge 1$. Since $z_{1/n} \ge 1$ for $n \ge 7$, this implies that

$$E[X_{(n)}] \le z_{1/n} + \frac{1}{z_{1/n}} \le \bar{z}_{1/n} + \frac{1}{\bar{z}_{1/n}}.$$

Hence the statement in this theorem will follow by showing that $\bar{z}_{1/n} + \frac{1}{\bar{z}_{1/n}}$ is smaller than the expression in the right-hand-side of (10). This is what is done below.

We define the number v by the equation $1-\frac{1}{v} = \sup_{n\ge 7}\frac{\log 4\pi+\log\log n}{4\log n}$. This gives the approximate value $v = 1.69693$. The constant v is involved below when using the inequality $\frac{1}{1-u} \leq 1 + vu$ whenever $u \leq 1 - \frac{1}{v}$, and by identifying u with $u = \frac{\log 4\pi + \log\log n}{4\log n}$. The proof will be completed after seeing the following steps:

$$\bar{z}_{1/n} + \frac{1}{\bar{z}_{1/n}} = \sqrt{2\log n} - \frac{\log 4\pi + \log\log n}{2\sqrt{2\log n}}\left(1 - \frac{\beta_0}{\log n}\right) + \frac{1}{\sqrt{2\log n} - \frac{\log 4\pi + \log\log n}{2\sqrt{2\log n}}(1 - \frac{\beta_0}{\log n})}$$

$$\begin{aligned}
&= \sqrt{2\log n} - \frac{\log 4\pi + \log\log n}{2\sqrt{2\log n}}\left(1 - \frac{\beta_0}{\log n}\right) + \frac{1}{\sqrt{2\log n}[1 - \frac{\log 4\pi + \log\log n}{4\log n}\left(1 - \frac{\beta_0}{\log n}\right)]} \\
&\leq \sqrt{2\log n} - \frac{\log 4\pi + \log\log n}{2\sqrt{2\log n}}(1 - \frac{\beta_0}{\log n}) + \frac{1}{\sqrt{2\log n}(1 - \frac{\log 4\pi + \log\log n}{4\log n})} \\
&\leq \sqrt{2\log n} - \frac{\log 4\pi + \log\log n}{2\sqrt{2\log n}}(1 - \frac{\beta_0}{\log n}) + \frac{1}{\sqrt{2\log n}}\left(1 + v\frac{\log 4\pi + \log\log n}{4\log n}\right) \\
&= \sqrt{2\log n} - \frac{\log 4\pi + \log\log n - 2}{2\sqrt{2\log n}} + \frac{\log 4\pi + \log\log n}{(2\log n)^{3/2}}(\beta_0 + \frac{v}{2}).
\end{aligned}$$

If writing now $\beta_0 + \frac{v}{2} = C$, we get the desired bound (11). Since β_0 can be taken to be 0.6403 and v to be 1.69693, it follows that we may take C to be $C = 0.6403 + \frac{1.69693}{2} = 1.48877 < 1.5$, as claimed.

7.7 Proof of Proposition 5.1

Let

$$a_n := \sqrt{2\log n} - \frac{\log\log n + \log 4\pi}{2\sqrt{2\log n}},\ n \geq 2.$$

Let also Z be a random variable with the standard Gumbel distribution, i.e. its CDF is $G(z) = e^{-e^{-z}},\ z \in \mathbb{R}$. Then, it is well known that

$$Z_n = \sqrt{2\log n}(X_{(n)} - a_n) \overset{\mathcal{L}}{\Rightarrow} Z \quad \text{as} \quad n \to \infty.$$

Moreover, $\{Z_n\}$ is uniformly integrable, therefore $E[Z_n] \to E[Z] = \gamma$, the Euler constant. By transposition, we have the implication

$$E[X_{(n)}] = \sqrt{2\log n} - \frac{\log\log n + \log 4\pi - 2\gamma}{2\sqrt{2\log n}} + o\left(\frac{1}{\sqrt{\log n}}\right) = b_n + o\left(\frac{1}{\sqrt{\log n}}\right), \quad (22)$$

where

$$b_n := \sqrt{2\log n} - \frac{\log\log n + \log 4\pi - 2\gamma}{2\sqrt{2\log n}}.$$

Therefore,

$$\delta_n = 1 - \Phi\left(b_n + o\left(\frac{1}{\sqrt{\log n}}\right)\right).$$

Now, using the fact that $1 - \Phi(x) = \frac{\varphi(x)}{x}(1 + O(x^{-2}))$ as $x \to \infty$, we get

$$\begin{aligned}
\delta_n &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{2\log n}(1 + o(1))}\exp\left(-\frac{1}{2}[2\log n - (\log\log n + \log 4\pi - 2\gamma) + o(1)]\right) \\
&= \frac{1}{\sqrt{2\pi}}\frac{1}{n}\sqrt{\log n}\,(2\sqrt{\pi})e^{-\gamma}(1 + o(1)) = \frac{e^{-\gamma}(1 + o(1))}{n}.
\end{aligned}$$

This is the claim in the proposition.

7.8 Proof of Theorem 5.1

The representation $E[X_{(n)}] = b_n + o(\frac{1}{\sqrt{\log n}})$ in (19) is not sharp enough to produce an asymptotic expansion for $E[X_{(n)}]$. We will derive such an higher order asymptotic expansion for $E[X_{(n)}]$ by using the trick of reducing a problem in extreme value theory to a problem about sums. This will be achieved by using the celebrated *Rényi's representation* for the distribution of the vector of uniform order statistics. This approach may be much more generally useful in extreme value theory than the particular purpose for which it is used in this article.

Since the proof of this theorem is quite long and intricate, we break it into smaller main steps.

Step 1: Let $U_1, U_2, \ldots$ be an iid $U[0,1]$ sequence and let $U_{(n)} = \max\{U_1, U_2, \ldots, U_n\}$. Let $X_1, X_2, \ldots$ denote an iid standard normal sequence, and let $\xi_0, \xi_1, \ldots$ denote an iid standard exponential sequence. Finally, let $Z_n := \frac{\xi_1+\ldots+\xi_n-n}{\sqrt{n}}$, and let $F_n(z)$ denote the CDF of Z_n.

By *Rényi's representation*, $U_{(n)} \overset{\mathcal{L}}{=} \frac{\xi_1+\ldots+\xi_n}{\xi_0+\xi_1+\ldots+\xi_n}$. Therefore, by the quantile transformation

$$X_{(n)} \overset{\mathcal{L}}{=} \Phi^{-1}(U_{(n)}) \overset{\mathcal{L}}{=} \Phi^{-1}\left(1 - \frac{\xi_0}{\xi_0+\xi_1+\ldots+\xi_n}\right).$$

Hence,

$$\begin{aligned} E[X_{(n)}] &= E\left[\Phi^{-1}\left(1 - \frac{\xi_0}{\xi_0+\xi_1+\ldots+\xi_n}\right)\right] \\ &= E\left[\Phi^{-1}\left(1 - \frac{\xi_0}{\xi_0+n+\sqrt{n}Z_n}\right)\right] \\ &= \int_{\mathbb{R}}\int_0^\infty \Phi^{-1}\left(1 - \frac{x}{x+n+\sqrt{n}z}\right) e^{-x} dx dF_n(z). \end{aligned} \tag{23}$$

Step 2: For any fixed x and z, the argument $\frac{x}{x+n+\sqrt{nz}}$ in (20) is small for large n. Thus, the idea now is to obtain and use sufficiently high order asymptotic expansions for $\Phi^{-1}(1-p)$ when $p \to 0$.

The asymptotic expansions for $\Phi^{-1}(1-p)$ are derived by inverting *Laplace's expansion* for the standard normal tail probability, which is

$$1 - \Phi(x) = \varphi(x)\left[\frac{1}{x} - \frac{1}{x^3} + \frac{1\times 3}{x^5} - \cdots + (-1)^k \frac{1\times 3\times(2k-1)}{x^{2k+1}} + R_k(x)\right], \tag{24}$$

where for any specific k, $R_k(x) = O(x^{-(2k+3)})$ as $x \to \infty$. By using (21) with $k = 1$, we obtain successive asymptotic expansions for $z_p = \Phi^{-1}(1-p)$ as $p \to 0$: Indeed, writing $t = 1/p$, we have the following:

$$\begin{aligned} z_p^2 &= 2\log t + O(\log\log t); \\ z_p^2 &= 2\log t - \log\log t + O(1); \\ z_p^2 &= 2\log t - \log\log t - \log 4\pi + O\left(\frac{\log\log t}{\log t}\right); \\ z_p^2 &= 2\log t - \log\log t - \log 4\pi + \frac{\log\log t}{2\log t} + O\left(\frac{1}{\log t}\right); \end{aligned}$$

$$z_p^2 \;=\; 2\log t - \log\log t - \log 4\pi + \frac{\log\log t + \log 4\pi - 2}{2\log t} + O\left(\frac{\log\log t}{(\log t)^2}\right). \qquad (25)$$

Step 3: From the last expression (22) we obtain:

$$\begin{aligned}
z_p &= \sqrt{2\log t}\left[1 - \frac{\log\log t + \log 4\pi}{2\log t} + \frac{\log\log t + \log 4\pi - 2}{(2\log t)^2} + O\left(\frac{\log\log t}{(\log t)^3}\right)\right]^{1/2} \\
&= \sqrt{2\log t}\Bigg[1 - \frac{\log\log t + \log 4\pi}{4\log t} - \frac{(\log\log t + \log 4\pi)^2 - 4\log\log t - 4\log 4\pi + 8}{32(\log t)^2} \\
&+ O\left(\frac{(\log\log t)^2}{(\log t)^3}\right)\Bigg] \\
&= T_1 - T_2 + T_3 + \sqrt{2\log t} \times O\left(\frac{(\log\log t)^2}{(\log t)^3}\right), \qquad (26)
\end{aligned}$$

where

$$T_1 := \sqrt{2\log t},\; T_2 := \frac{\log\log t + \log 4\pi}{2\sqrt{2\log t}},\; T_3 := \frac{(\log\log t - 2)^2 + (\log 4\pi - 2)^2}{\frac{32(\log t)^2}{\sqrt{2\log t}}}.$$

Step 4: Relation (23) in conjunction with the integral representation in (20) will be key in producing the desired asymptotic expansion for $E[X_{(n)}]$.

Toward this, with (20) in mind, we use the notations

$$p = \frac{x}{n + z\sqrt{n} + x} \quad \text{and} \quad t = \frac{1}{p} = \frac{n + z\sqrt{n} + x}{x} = \frac{n}{x}\left(1 + \frac{z}{\sqrt{n}} + \frac{x}{n}\right).$$

We will now use (23) and derive asymptotic expansions up to the needed order for each of the three terms T_1, T_2 and T_3. Then we combine them to produce the final asymptotic expansion for the mean $E[X_{(n)}]$.

The first term in (23) is

$$\begin{aligned}
T_1 : &= \sqrt{2\log t} = \sqrt{2(\log n - \log x) + O_p(\frac{1}{\sqrt{n}})} \\
&= \sqrt{2\log n}\sqrt{1 - \frac{\log x}{\log n} + O_p(\frac{1}{\sqrt{n}\log n})} \\
&= \sqrt{2\log n} - \frac{\log x}{\sqrt{2\log n}} - \frac{(\log x)^2}{2(2\log n)^{3/2}} + O_p((\log n)^{-5/2}). \qquad (27)
\end{aligned}$$

The next is

$$\begin{aligned}
T_2 : &= \frac{\log\log t + \log 4\pi}{2\sqrt{2\log t}} = \frac{\log(\log n - \log x + O_p(\frac{1}{\sqrt{n}})) + \log 4\pi}{2[\sqrt{2\log n} - \frac{\log x}{\sqrt{2\log n}} + O_p((\log n)^{-3/2})]} \\
&= \frac{\log\log n - \frac{\log x}{\log n} + \log 4\pi + O_p((\log n)^{-2})}{2\sqrt{2\log n}[1 - \frac{\log x}{2\log n} + O_p((\log n)^{-2})]} \\
&= \frac{\log\log n - \frac{\log x}{\log n} + \log 4\pi + \log x\frac{\log\log n}{2\log n} + \log x\frac{\log 4\pi}{2\log n} + O_p(\frac{\log\log n}{(\log n)^2})}{2\sqrt{2\log n}}. \qquad (28)
\end{aligned}$$

It remains to handle the third term in (23), T_3. For this, we use the following observations:

$$
\begin{aligned}
(\log\log t)^2 &= [\log\log n - \frac{\log x}{\log n} + O_p((\log n)^{-2})]^2 \\
&= (\log\log n)^2 - (2\log x)\frac{\log\log n}{\log n} + O_p\left(\frac{\log\log n}{(\log n)^2}\right); \qquad (29) \\
-4\log\log t &= -4\log\log n + 4\frac{\log x}{\log n} + O_p((\log n)^{-2}); \qquad (30) \\
\frac{1}{\frac{32(\log t)^2}{\sqrt{2\log t}}} &= \frac{1}{16\sqrt{2}(\log t)^{3/2}} = \frac{1}{8(2\log t)^{3/2}} \\
&= \frac{1 + \frac{3}{2}\frac{\log x}{\log n} + O_p((\log n)^{-2})}{8(2\log n)^{3/2}}. \qquad (31)
\end{aligned}
$$

The last three relations, (26), (27) and (28) give a bound for T_3 which then is combined with (24) and (25) thus obtaining

$$
\begin{aligned}
\Phi^{-1}\left(1 - \frac{x}{x + z\sqrt{n} + n}\right) &= \sqrt{2\log n} - \frac{\log\log n + \log 4\pi + 2\log x}{2\sqrt{2\log n}} \\
+\frac{1}{8(2\log n)^{3/2}} \times &\Big[(\log\log n)^2 - 4\log\log n - 4(\log x)\log\log n + (8 - 4\log 4\pi)(\log x) \\
&-4(\log x)^2 + (\log 4\pi)^2 - 4\log 4\pi + 8\Big] + O_p\left(\frac{(\log\log n)^2}{(\log n)^{5/2}}\right). \qquad (32)
\end{aligned}
$$

Notice the interesting fact that no terms involving z appear in this three term expansion for $\Phi^{-1}(1 - \frac{x}{x+z\sqrt{n}+n})$, although on introspection it is clear why the z terms do not appear.

Step 5: Using now the integration facts

$$
\int_0^\infty (\log x)e^{-x}dx = -\gamma, \quad \text{and} \quad \int_0^\infty (\log x)^2 e^{-x}dx = \gamma^2 + \frac{\pi^2}{6},
$$

and (20), (29), and (20.46) in Corollary 20.3 in Bhattacharya and Rao (2010), we get the following three term asymptotic expansion for $E[X_{(n)}]$:

$$
\begin{aligned}
E[X_{(n)}] &= \sqrt{2\log n} - \frac{\log\log n + \log 4\pi - 2\gamma}{2\sqrt{2\log n}} + \frac{1}{8(2\log n)^{3/2}} \times \\
&\times \left[(\log\log n)^2 - 4(1-\gamma)\log\log n + \gamma(4\log 4\pi - 8) - 4\left(\gamma^2 + \frac{\pi^2}{6}\right)\right. \\
&+ \left.(\log 4\pi)^2 - 4\log 4\pi + 8\right] + O\left(\frac{(\log\log n)^2}{(\log n)^{5/2}}\right) \\
&= \sqrt{2\log n} - \frac{\log\log n + \log 4\pi - 2\gamma}{2\sqrt{2\log n}} + \frac{1}{8(2\log n)^{3/2}} \times \\
&\times \left[(\log\log n - 2(1-\gamma))^2 + (\log 4\pi - 2)^2 + 4\gamma(\log 4\pi - 2\gamma) - \frac{2}{3}\pi^2\right] \\
&+ O\left(\frac{(\log\log n)^2}{(\log n)^{5/2}}\right).
\end{aligned}
$$

This completes the proof of this theorem.

7.9 Proof of Theorem 5.2

Theorem 5.2 follows from our calculations in the proof of Theorem 5.1 after making a simple observation. Indeed, recall that $\theta_n = \Phi^{-1}(2^{-1/n})$, so $\theta_n = \Phi^{-1}(1-p)$, with

$$t = \frac{1}{p} = \frac{2^{1/n}}{2^{1/n}-1} = \frac{n}{\log 2} + O(1).$$

Therefore, the asymptotic expansion for θ_n up to an error of order $O\left(\frac{(\log\log n)^2}{(\log n)^{5/2}}\right)$ will follow in a straightforward manner from relation (23). The details are omitted.

7.10 Proof of Theorem 5.3

We give a sketch of parts (a) and (d); (b) follows readily from (a) and (c) can be derived from (d).
To prove part (a), observe the representation

$$F = \sum_{i=1}^{n} I_{X_i > c_n, B_i = 0} = \sum_{i=1}^{n} I_{Z_i > c_n, B_i = 0}. \tag{33}$$

Since in this theorem $c_n = \sqrt{2\log n}$, and the Bernoulli sequence $\{B_i\}$ and the standard normals $\{Z_i\}$ are independent, we get

$$E[F] = n^{1-\beta}[1 - \Phi(\sqrt{2\log n})]$$

$$= n^{1-\beta}\frac{\phi(\sqrt{2\log n})}{\sqrt{2\log n}}(1+o(1)) = \frac{n^{1-\beta}}{2n\sqrt{\pi\log n}}(1+o(1))$$

$$= \frac{1}{2\sqrt{\pi}}\frac{n^{-\beta}}{\sqrt{\log n}}(1+o(1)) = o(1).$$

Hence, $F \xrightarrow{P} 0$.
Next, to prove part (d), use the representations

$$D = \sum_{i=1}^{n} I_{X_i > c_n, B_i = 1} = \sum_{i=1}^{n} I_{Z_i > c_n - \mu_n, B_i = 1},$$

and

$$S = \sum_{i=1}^{n} I_{B_i = 1}. \tag{34}$$

Denote

$$p_D = P(Z_i > c_n - \mu_n, B_i = 1) = \epsilon_n[1 - \Phi(c_n - \mu_n)],$$

and

$$p_S = P(B_i = 1) = \epsilon_n.$$

Then, under the Donoho-Jin model, p_D, p_S both converge to zero, and it may be shown by use of the CLT that

$$D = np_D + \sqrt{np_D}\,O_p(1),$$

and

$$S = np_S + \sqrt{np_S}O_p(1).$$

This will give

$$\frac{D}{S} = \frac{p_d}{p_S}(1 + o_p(1)).$$

Now, use the facts that

$$\frac{p_D}{p_S} = 1 - \Phi(c_n - \mu_n),$$

and calculations will show that if $c_n = \sqrt{2\log n}$ and $\mu_n = \sqrt{2r\log n}$, then

$$\frac{p_D}{p_S} = \frac{n^{-(1-\sqrt{r})^2}}{2\sqrt{\pi}(1-\sqrt{r})\sqrt{\log n}}(1 + o(1)),$$

and this gives the stated result in part (d). We omit the intermediate calculations.

7.11 Proof of Theorem 5.4

Theorem 5.4 asserts convergence in distribution of F to a suitable Poisson. For this, define

$$p_F = P(Z_i > c_n, B_i = 0) = (1 - \epsilon_n)[1 - \Phi(c_n)]. \tag{35}$$

It is necessary and sufficient to show that with c_n as defined in the theorem,

$$p_F \to 0,\ np_F \to \lambda.$$

We give a sketch of the proof. It is clear that $p_F \to 0$, because $c_n \to \infty$. On the other hand, since $\epsilon_n \to 0$,

$$np_F = n(1 - \epsilon_n)[1 - \Phi(c_n)]$$

$$= n\frac{\phi(c_n)}{c_n}(1 + o(1)) = n\frac{e^{-\frac{c_n^2}{2}}}{c_n\sqrt{2\pi}}(1 + o(1)).$$

Use now the definition of the sequence c_n given in the theorem. Then, writing $K = \log(4\pi\lambda^2)$, on calculations, one gets

$$np_F = \frac{ne^{-\frac{1}{2}[2\log n - (\log\log n + K) + o(1)]}}{\sqrt{2\log n}\sqrt{2\pi}(1 + o(1))}$$

$$= \frac{n\sqrt{\log n}}{n}\frac{2\lambda\sqrt{\pi}}{\sqrt{2\log n}\sqrt{2\pi}}(1 + o(1)),$$

and this simplifies to

$$np_F = \lambda(1 + o(1)).$$

Again, we omit the intermediate calculations, which are not difficult.

7.12 Proof of Theorem 5.5

First note that, using the same arguments as in Theorem 5.3, one still has

$$\frac{D}{S} = \frac{p_D}{p_S}(1 + o_p(1)) = [1 - \Phi(c_n - \mu_n)](1 + o_p(1)). \tag{36}$$

Now, by using the definition of c_n in Theorem 5.5, and on using that $\mu_n = \sqrt{2r \log n}$, one gets, writing $K = \log(4\pi\lambda^2)$,

$$c_n - \mu_n = \sqrt{2 \log n}(1 - \sqrt{r}) - \frac{\log \log n + K}{2\sqrt{2 \log n}}$$

$$\Rightarrow (c_n - \mu_n)^2 = 2 \log n(1 - \sqrt{r})^2 - (1 - \sqrt{r})(\log \log n + K) + o(1). \tag{37}$$

This gives, with some intermediate calculations,

$$1 - \Phi(c_n - \mu_n) = \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}(c_n - \mu_n)^2}}{c_n - \mu_n}(1 + o(1))$$

$$= \frac{1}{\sqrt{2\pi}} \frac{e^{-(1-\sqrt{r})^2 \log n + \frac{1-\sqrt{r}}{2}(\log \log n + K) + o(1)}}{(1 - \sqrt{r})\sqrt{2 \log n}(1 + o(1))}$$

$$= \frac{n^{-(1-\sqrt{r})^2}(\log n)^{\frac{1-\sqrt{r}}{2}}(4\pi\lambda^2)^{\frac{1-\sqrt{r}}{2}}(1 + o(1))}{\sqrt{2\pi}(1 - \sqrt{r})\sqrt{2 \log n}(1 + o(1))}$$

$$= \frac{n^{-(1-\sqrt{r})^2}(4\pi\lambda^2 \log n)^{\frac{1-\sqrt{r}}{2}}}{2\sqrt{\pi}(1 - \sqrt{r})\sqrt{\log n}}(1 + o(1)),$$

and this establishes the claim of Theorem 5.5.

References

Bhattacharya, R. N. and Rao, R. R. (2010). Normal Approximation and Asymptotic Expansions. SIAM, Philadelphia.

Cai, T., Xu, G. and Zhang, J. (2009). On the recovery of sparse signals, IEEE Trans. Inform. Theory, 55, 3388-3397.

Cai, T. and Wang, L. (2011). Orthogonal matching pursuit for sparse signal recovery, IEEE Trans. Inform. Theory, 57, 4680-4688.

DasGupta, A. (2008). Asymptotic Theory of Statistics and Probability. Springer, New York.

DasGupta, A. (2013). How Many Signals Can We Detect: Sensitive Thresholding in the Ingster-Donoho-Jin Paradigm, In Preparation.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. Ann. Stat., 32, 962–994.

Donoho, D. and Johnstone, I. M. (1994). Minimax risk over ℓ_p balls with ℓ_q error. Probab. Theory Related Fields., 2, 277–303.

Galambos, J. (1978). Asymptotic Theory of Extreme Order Statistics. Wiley, New York.

Gascuel, O. and Caraux, G. (1992). Bounds on expectations of order statistics via extremal dependences. Statist. & Probab. Letters, 15, 2, 143–148.

Hall, P. (1979). On the rate of convergence of normal extremes. J. Appl. Probab., 16, 2, 433–439.

Hall, P. and Jin, J. (2008). Properties of higher criticism under strong dependence, Ann. Statist., 36, 381-402.

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise, Ann. Statist., 38, 1686-1732.

Lai, T. L. and Robbins, H. (1976). Maximally dependent random variables. Proc. Nat. Acad. Sci. USA, 73, 286–288.

Lai, T.L. and Robbins, H. (1978). A class of dependent random variables and their maxima. Probab. Theory Related Fields, 42, 2, 89–111.

Rychlik, T. (1998). Bounds on expectations of L-estimates. In Handbook of Statistics, 16, pp. 105–145. North-Holland, Amsterdam.

Szarek, S. and Werner, E. (1999). A nonsymmetric correlation inequality for Gaussian measures. J. Multivar. Anal., 68, 193–211.