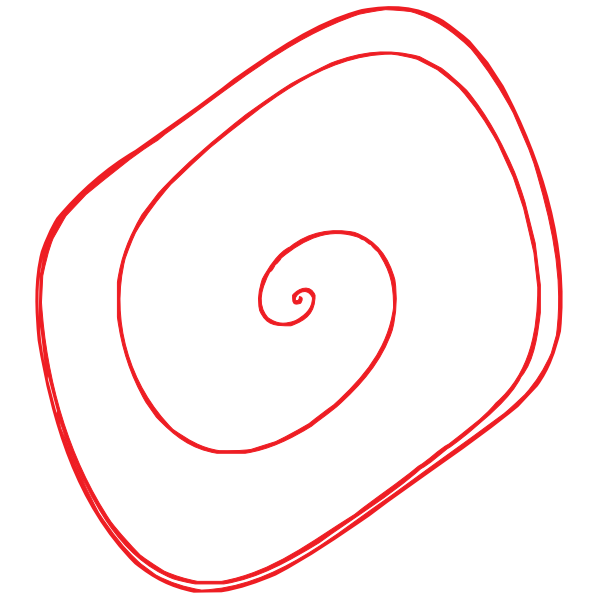


ORDER THRESHOLDING

Min Hee Kim and Michael G. Akritas

Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA



Abstract

A new thresholding method, based on L -statistics and called *order thresholding*, is proposed as a technique for improving the power when testing against high-dimensional alternatives. The new method allows great flexibility in the choice of the threshold parameter. This results in improved power over the soft and hard thresholding methods. Moreover, order thresholding is not restricted to the normal distribution. The performance of the basic order threshold statistic is evaluated with extensive simulations.

Background on Thresholding Methods

- Introduced in the context of nonparametric function estimation using wavelets by Donoho and Johnstone (1994).
- Johnstone and Silverman (2004) elaborate on the following additional applications of thresholding:
 - Image processing, model selection, data mining.
- Spokoiny (1996), Fan (1996), and Fan and Lin (1998) consider applications of thresholding methods to testing problems.
- Beran (2004) considered a one-way ANOVA design, but from the estimation point of view.

Hard Thresholding Method

- In the simple context where the X_i are independent $N(\theta_i, 1)$, $i = 1, \dots, n$, and we wish to test $H_0: \theta_1 = \dots = \theta_n = 0$ vs $H_a: H_0$ is false, the naive test statistic $\sum_{i=1}^n Y_i$, where $Y_i = X_i^2$, cannot detect alternatives of the order $\|\theta\|^2 = o(\sqrt{n})$.
- The hard threshold statistic is of the form

$$T_H(\delta_n) = \sum_{i=1}^n Y_i I\{Y_i > \delta_n\}, \quad (1)$$

where $\delta_n = 2 \log(na_n)$, with $a_n = c(\log n)^{-d}$, for $c > 0$ and $d > 0.5$.

- $T_H(\delta_n)$ is centered and scaled by

$$\mu_{n,H} = \sqrt{2/\pi} a_n^{-1} \delta^{1/2} (1 + \delta^{-1}), \quad \sigma_{n,H}^2 = \sqrt{2/\pi} a_n^{-1} \delta^{3/2} (1 + 3\delta^{-1}),$$

and

$$T_H^*(\delta_n) = \frac{T_H(\delta_n) - \mu_{n,H}}{\sigma_{n,H}} \xrightarrow{d} N(0, 1), \quad \text{as } n \rightarrow \infty. \quad (2)$$

Why Invent Another Threshold Method?

- The asymptotic theory for hard thresholding (Fan, 1996) is specific to the normality assumption.
 - The centering and scaling of $T_H(\delta_n)$ are specific to the normality assumption.
 - The choice of δ_n is specific to the normality assumption.
- Small departures from the value of δ_n recommended in Fan (1996) ($c = 1$, $d = 2$) have significant effect on the level of the test: (30,000 simulation runs)

	$\delta_n - 2.0$	$\delta_n - 1.6$	$\delta_n - 1.2$	$\delta_n - 0.8$	$\delta_n - 0.4$	δ_n	$\delta_n + 0.4$	$\delta_n + 0.8$
$n = 50$	0.0003	0.0099	0.0231	0.0341	0.0431	0.0493	0.0543	0.0588
$n = 100$	0.0101	0.0229	0.0324	0.0390	0.0461	0.0504	0.0552	0.0559
$n = 200$	0.0231	0.0316	0.0382	0.0439	0.0484	0.0507	0.0539	0.0562
$n = 500$	0.0327	0.0388	0.0422	0.0465	0.0502	0.0535	0.0540	0.0563

- For larger δ_n , a small gain in power can be achieved, but the asymptotic theory does not provide a good approximation to the distribution of $T_H^*(\delta_n)$.
- Even under normality, δ_n which attains the highest power is different from alternatives (Johnstone and Silverman, 2004).
- In problems where the random variables X_i have a bounded support, if δ_n goes to infinity, then this test does not work.

From Order Statistics to Order Thresholding

- Let V_1, \dots, V_n be iid from the standard exponential distribution, let $V_{1,n} < \dots < V_{n,n}$ be the corresponding order statistics.
- The order threshold statistic is of the form

$$T_{E,L}(k_n) = \sum_{i=1}^n c_{in} V_{i,n} = \sum_{i=n-k_n+1}^n V_{i,n} \stackrel{d}{=} \sum_{j=1}^n \alpha_{E,jn}(k_n) V_j, \quad (3)$$

where $\alpha_{E,jn}(k_n) = \frac{1}{n-j+1} \sum_{i=j}^n c_{in}$ and $c_{in} = I(i > n - k_n)$.

Theorem 1 Let $T_{E,L}(k_n)$ be given in (3). Then, provided that $k_n \rightarrow \infty$, as $n \rightarrow \infty$,

$$T_{E,L}^*(k_n) = \frac{T_{E,L}(k_n) - \sum_{i=1}^n \alpha_{E,in}(k_n)}{\sqrt{\sum_{i=1}^n \alpha_{E,in}(k_n)^2}} \xrightarrow{d} N(0, 1), \quad \text{as } n \rightarrow \infty.$$

The Basic Order Threshold Statistic

- In the simple context where the X_i are independent $N(\theta_i, 1)$, $i = 1, \dots, n$, and we wish to test $H_0: \theta_1 = \dots = \theta_n = 0$ vs $H_a: H_0$ is false, the order threshold statistic is

$$T_L(k_n) = \sum_{i=1}^n c_{in} Y_{i,n}, \quad (4)$$

where $Y_i = X_i^2$, $Y_{1,n} < \dots < Y_{n,n}$ are the ordered Y_i s, and $c_{in} = I(i > n - k_n)$.

- The asymptotic null distribution: The approach of Chernoff, Gastwirth, and Johns (1967) is based on the representation

$$T_L(k_n) \stackrel{d}{=} \sum_{i=n-k_n+1}^n \tilde{H}(V_{i,n}),$$

where $V_{i,n}$ are the ordered observations from $Exp(1)$ random variables, and $\tilde{H}(v) = F^{-1} \circ G(v)$ with $F(y) = \frac{1}{\sqrt{2\pi}} \int_0^y t^{-1/2} e^{-t/2} dt$, $y > 0$, and $G(v) = 1 - e^{-v}$, $v \geq 0$.

- Let

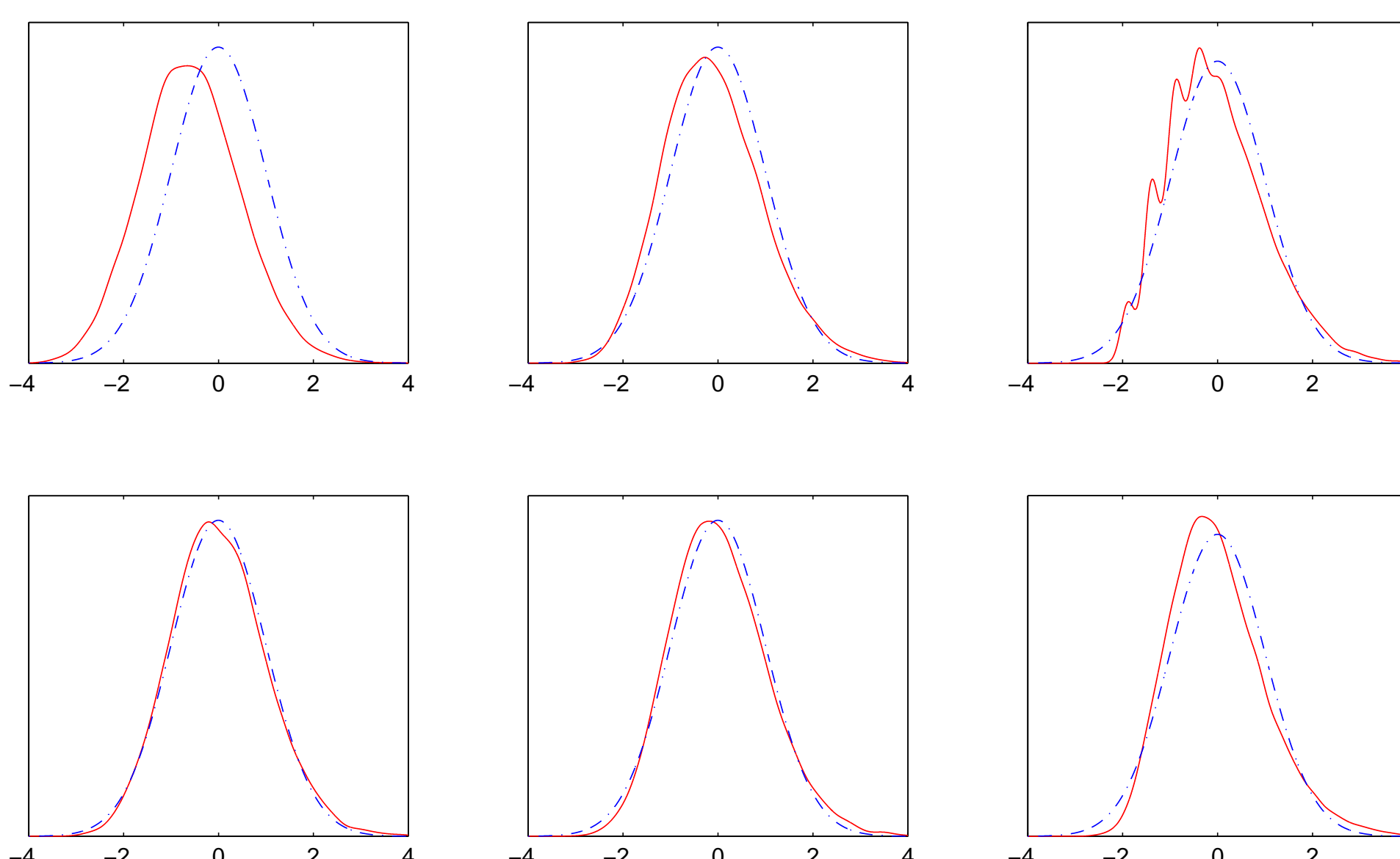
$$\mu_n(k_n) = \frac{1}{n} \sum_{i=1}^n c_{in} \tilde{H}(\tilde{v}_{in}), \quad \sigma_n^2(k_n) = \frac{1}{n} \sum_{i=1}^n \alpha_{in}^2(k_n), \quad (5)$$

where $\alpha_{in}(k_n) = \frac{1}{n-i+1} \sum_{j=i}^n c_{jn} \tilde{H}'(\tilde{v}_{jn})$ and $\tilde{v}_{in} = \sum_{j=1}^i \frac{1}{n-j+1}$.

Theorem 2 Let $\mu_n(k_n)$ and $\sigma_n^2(k_n)$ be as in (5) with $c_{in} = I(i > n - k_n)$, and let $T_L(k_n)$ be given in (4). Then, provided that $k_n \rightarrow \infty$, as $n \rightarrow \infty$, we have

$$T_L^*(k_n) = \frac{T_L(k_n) - n\mu_n(k_n)}{\sqrt{n\sigma_n(k_n)}} \xrightarrow{d} N(0, 1), \quad \text{as } n \rightarrow \infty. \quad (6)$$

Simulations



Top panel: Estimated densities of $T_H^*(\delta_{200})$ for $\delta_{200} = 1.842, 3.927$, and 5.672 . Bottom panel: Estimated densities of $T_L^*(k_{200})$ for $k_{200} = 35, 10$, and 3 . Interpretation: 1. Estimated densities of the order threshold statistic are closer to the standard normal density than those of the hard threshold statistic. 2. Upper panel show the rapid deterioration of the quality of the normal approximation to the distribution of $T_H^*(\delta_{200})$ as δ_{200} shifts away from recommended value of 3.927 .

Comparison Between Hard and Order Thresholding:

The simulation uses samples of size $n = 500$ generated from the normal distribution with variance 1. The threshold parameter k_{500} of the order threshold statistics takes values of 15, 40, 70, 100, 200 and 500. The hard threshold statistic we consider uses the recommended value of the threshold parameter which is $\delta_{500} = 2 \log(500 \log^{-2} 500) = 5.1216$. All results are based on 20,000 simulation runs. In particular, we consider the following sequence of alternatives indexed by r :

$$H_r: \theta_j = \eta_{j+r-1} \quad \text{for } j = 1, \dots, 500, \quad r = 1, \dots, 17,$$

where η_j , $j = 1, 2, \dots$, is a given sequence.

Example 1 We generate the values of η_j , $j = 1, \dots, 17$, from $Uniform(-5, 5)$. The remaining values of η_j are 0. The values different from 0 are as follows:

$$(-2.4644, 3.7345, 0.1340, 2.3265, -0.7777, 4.6137, -4.2794, 0.5341, -2.0802, 3.5796, -1.6424, 1.8020, -4.4656, -1.4334, -0.0170, -0.6556, 0.6246).$$

	$T_H(5.1216)$	$T_H(15)$	$T_H(40)$	$T_H(70)$	$T_H(100)$	$T_H(200)$	$T_H(500)$
H_1	0.9870	0.9993	0.9976	0.9939	0.9879	0.9667	0.9448
H_2	0.9822	0.9989	0.9963	0.9907	0.9823	0.9553	0.9275
H_3	0.9502	0.9971	0.9884	0.9712	0.9538	0.9036	0.8621
H_4	0.9501	0.9967	0.9875	0.9716	0.9542	0.9030	0.8600
H_5	0.9344	0.9952	0.9833	0.9626	0.9393	0.8760	0.8308
H_6	0.9323	0.9948	0.9812	0.9597	0.9355	0.8704	0.8225
H_7	0.7740	0.9597	0.9039	0.8401	0.7913	0.6925	0.6325
H_8	0.5373	0.8235	0.7039	0.6192	0.5657	0.4786	0.4313
H_9	0.5361	0.8146	0.6982	0.6173	0.5648	0.4728	0.4249
H_{10}	0.4814	0.7855	0.6522	0.5645	0.5126	0.4243	0.3803
H_{11}	0.3013	0.5807	0.4374	0.3675	0.3284	0.2746	0.2511
H_{12}	0.2722	0.5527	0.4039	0.3358	0.2999	0.2493	0.2271
H_{13}	0.2405	0.5175	0.3634	0.2976	0.2599	0.2110	0.1925
H_{14}	0.0640	0.0704	0.0698	0.0686	0.0675	0.0684	0.0673
H_{15}	0.0544	0.0607	0.0566	0.0563	0.0566	0.0566	0.0583
H_{16}	0.0554	0.0605	0.0575	0.0571	0.0574	0.0573	0.0578
H_{17}	0.0564	0.0608	0.0592	0.0591	0.0577	0.0563	0.0565

Discussion

The asymptotic theory of test statistics based on hard and soft thresholding pertain the normal distribution and require the threshold parameter to tend to infinity at a strictly prescribed rate. Order thresholding, a new thresholding method based on order statistics, is proposed. The asymptotic theory allows great flexibility in the choice of the threshold parameter. Simulation studies with normal data suggest that order thresholding far outperforms hard thresholding in terms of power.

References

- [1] BERAN, R. (2004). Hybrid Shrinkage Estimators Using Penalty Bases For the Ordinal One-Way Layout. *The Annals of Statistics*. **32** 2532-2558.
- [2] CHERNOFF, H., GASTWIRTH, J. L., and JOHNS, M. V. (1967). Asymptotic Distribution of Linear Combinations of Functions of Order Statistics with Applications to Estimation. *The Annals of Mathematical Statistics*. **38** 52-72.
- [3] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*. **81** 425-455.
- [4] FAN, J. (1996). Test of Significance Based on Wavelet Thresholding and Neyman's Truncation. *Journal of the American Statistical Association*. **91** 674-688.
- [5] FAN, J. and LIN, S. K. (1998). Test of Significance When Data Are Curves. *Journal of the American Statistical Association*. **93** 1007-1021.
- [6] JOHNSTONE, I. M. and SILVERMAN B. W. (2004). Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences. *The Annals of Statistics*. **32** 1594-1649.
- [7] SPOKOINY, V. G. (1996). Adaptive Hypothesis Testing Using Wavelets. *The Annals of Statistics*. **24** 2477-2498.