Testing normality in regression problems:
the estimated empirical process, Khmaladze transformation,
and power of Kolmogorov-Smirnov tests.

Ray Brownrigg

Report 08-09 SMSCS, Victoria University of Wellington. Revision December 17, 2008

1 Introduction

Consider the classical non-parametric regression problem: we observe a sequence of pairs (X_i, Y_i), where $X_i \in \mathbb{R}^d$ are d-dimensional *explanatory* random variables, and Y_i are corresponding *responses*, and assume that

$$Y_i = m(X_i) + e_i, \quad i = 1, \ldots, n$$

and the *errors* $e_1, \ldots, e_n$ are i.i.d. random variables. We do not know the *regression function* $m(x)$ and do not presume it has any given parametric form.

What we need is to test the hypothesis that the distribution function of e_i is a given $0-$mean distribution function F. Since we do not know $m(x)$ we can not observe e_i and will have to use some non-parametric estimator $\hat{m}_n(x)$ of the regression function and use *estimated errors*

$$\hat{e}_i = Y_i - \hat{m}_n(X_i)$$

to test our hypothesis. However, these $\hat{e}_1, \ldots, \hat{e}_n$, are neither independent nor identically distributed any more.

Here we will consider two ways of testing our hypothesis. One way is to use the *estimated* empirical process

$$\hat{v}_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\mathbb{I}_{\{\hat{e}_i \le x\}} - F(x)]$$

and use goodness of fit statistics based on this process. The other way is to use another version of the empirical process, called the *Khmaladze transformation*,

$$w_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\mathbb{I}_{\{\hat{e}_i \le x\}} - K_n(x)]$$

where we give the explicit form of the centering process $K_n(x)$ below, and to base goodness of fit statistics on this process.

The theoretical advantage of the process $w_n(x)$ is that under the usual time transformation $t = F(x)$ it converges, under the null hypothesis, to standard Brownian motion on the interval $[0, 1]$ (Khmaladze and Koul 2007). Hence, the null distribution of any goodness of fit statistic from $w_n(x)$, invariant under the time transformation $t = F(x)$,

will be free from the underlying F, $\hat{m}_n$ and m, which is a useful property in practice, as even though we know the hypothetical F, we have no knowledge of the underlying regression function $m(x)$ and a large variety of non-parametric estimators $\hat{m}_n$ could have been used. Contrary to this, the null distribution of $\hat{v}_n(x)$, in general, depends on all three functions F, $\hat{m}_n$ and m, and this inconvenient property is inherited by goodness of fit statistics, based on this process.

The aim of this work is

- to provide convenient computational formulae for the calculation of trajectories of the process $w_n(x)$ in the case of testing normality, i.e. when the hypothetical F is just a standard normal distribution function;
- to verify, through simulation experiments, that the convergence of w_n to its limit is sufficiently rapid and indeed is not affected by the choice of various regression functions $m(x)$ and its non-parametric estimators;
- to investigate to what degree the distribution of $\hat{v}_n$ depends on regression function m and its estimator $\hat{m}_n$;
- to compare the power of the tests based on $\hat{v}_n(x)$ and $w_n(x)$ for various alternatives.

2 Computational Formulae for $K_n(x)$

The formula for $K_n(x)$ above is defined as

$$K_n(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_{-\infty}^{x\wedge\hat{e}_i} \frac{(y-\mu(y))(\hat{e}_i-\mu(y))}{\sigma_f^2(y)} f(y)dy - ln[1-F(x\wedge\hat{e}_i)]\right)$$

where $f(x)$ is the standard normal density function, $F(x)$ is the standard normal distribution function,

$$\mu(y) = \frac{f(y)}{1-F(y)}$$

and

$$\sigma_f^2(y) = [1-F(x)](1+x\mu(x)-\mu^2(x))$$

3 Simulation Experiments under H_0

3.1 The regression structure

As a simple test, the data was chosen to be of the form:

$$Y_i = m(X_i) + e_i, \quad i = 1, \ldots, n$$

where e_i are i.i.d with distribution F. The null hypothesis is

$$H_0 : F = N(0,1)$$

i.e the e_i are from the standard normal distribution.

Initially the function $m(x)$ was chosen to be of the form $m(x) = cx$ with $c = 1$ or 5, the X_i were chosen as uniform $U[0, T]$ with $T = 2$ and the sample size n was chosen as 200 or 500. Thus we generate each sample as

$$Y_i = cX_i + e_i, \quad i = 1, \ldots, n$$

Later simulations used the exponential function $m(x) = e^{cx}$ as the regression function with the value of c chosen as 1 or 2.

3.2 The estimation procedure

Once the form of the explanatory random variable, the regression function and the error distribution have been decided to allow generation of the responses Y_i, the next step of the simulation is to find the residuals, i.e. to generate $\hat{e}_i = Y_i - \hat{m}(X_i)$. Thus we need a procedure to calculate an estimate $\hat{m}$ for the function $m(x)$.

The procedure chosen was the Nadaraya-Watson estimator which depends on a window width (or bandwidth) parameter a.

Thus we have:

$$\hat{m}_n(x) = \sum_{i=1}^{n} Y_i \mathbb{I}_{\{X_i \in [x-a, x+a]\}} / \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in [x-a, x+a]\}},$$

with various choices for window width (or bandwidth) parameter a.

This a was initially chosen to be 0.08, but further experimentation led to different values being chosen and to a minor modification.

The modification was implemented to reduce the end-effects of a fixed window width parameter. This was achieved by adjusting the window width parameter near the end-points so that at all times the total active window was symmetric about the point (the particular X_i) being analyzed. Thus in particular, when the extreme end-points were being analyzed, the window was effectively just that single point.

The initial sample size was selected as $n = 200$ since this is not too big, but is such that one could expect asymptotic results to be visible.

It is, of course, difficult to speak about the distribution of the processes w_n or $\hat{v}_n$ as such, and instead we speak about the distribution of the K-S statistics based on these processes, namely:

$$\hat{V}_n = \sup_x |\hat{v}_n(x)| \text{ and } W_n = \sup_x |w_n(x)|.$$

3.3 The null distribution of $\hat{V}_n$

We now proceed to generate the K-S statistics for $\hat{V}_n$ under H_0 for the different situations described above.

3.3.1 Linear regression

Figure 1 shows the empirical distributions of $\hat{V}_n$ under H_0 when $m(x) = cx$ with $c = 1$, for a selection of values of a.

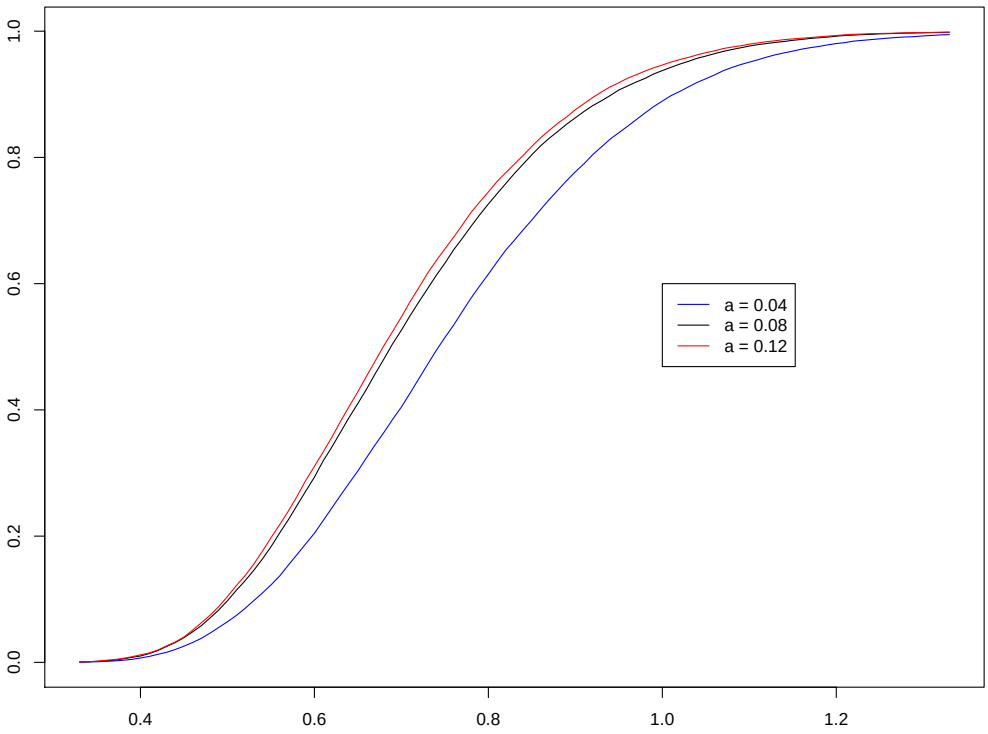

Figure 1: Empirical distributions of $\hat{V}_n$ under H_0 when $m(x) = x$.

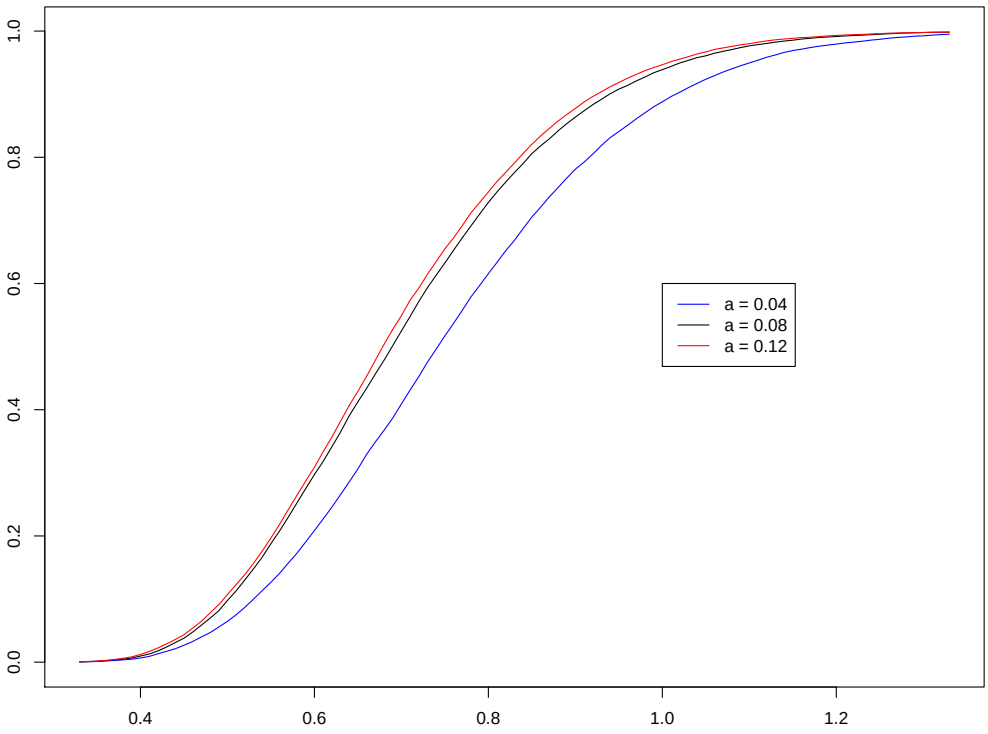

Figure 2: Empirical distributions of $\hat{V}_n$ under H_0 when $m(x) = 5x$.

When the value of the slope c is chosen as 5, i.e. $m(x) = 5x$, there is no discernible change in the plotted e.d.f.s. Figure 2 shows the empirical distributions of $\hat{V}_n$ under H_0 when $m(x) = 5x$ for the same selection of values of a as for Figure 1.

Indeed, in the case of any linear regression, the behaviour of $\hat{m}(x)$ is more or less the same. In particular it contains no bias.

3.3.2 Exponential regression function

Now we choose $m(x) = e^{cx}$, initially with $c = 1$. Figure 3 shows the empirical distribution of $\hat{V}_n$ under H_0 for a selection of values of a. Note that these e.d.f. curves are similar but not identical to those for the case of a linear regression function.

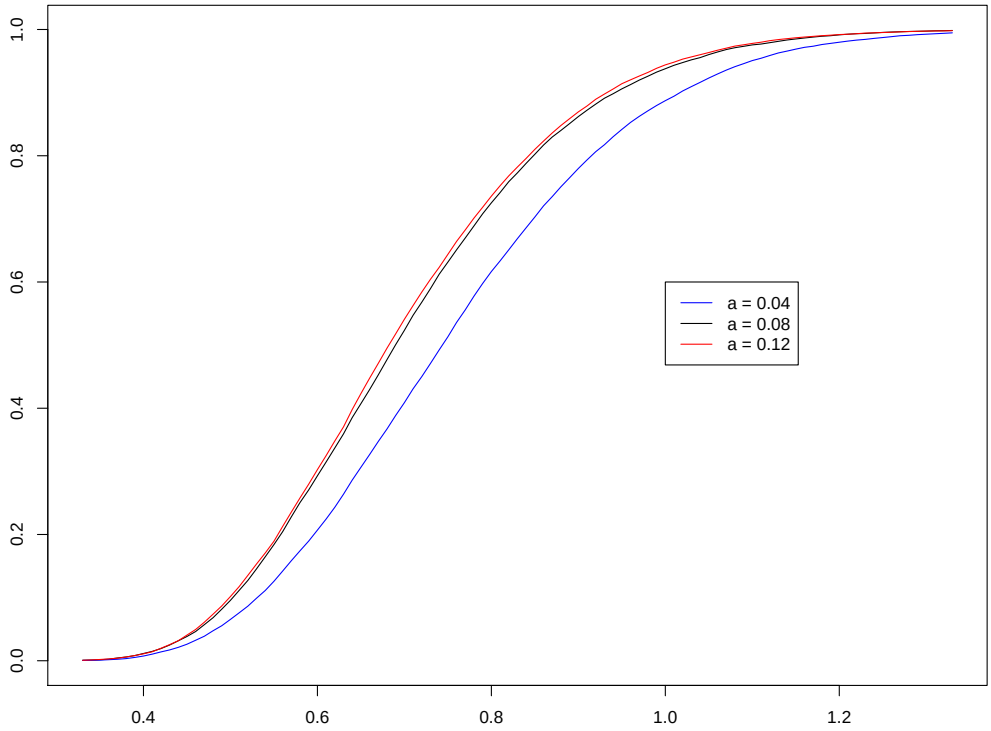

Figure 3: Empirical distributions of $\hat{V}_n$ under H_0 when $m(x) = e^x$.

When the value of c is chosen as 2, i.e. $m(x) = e^{2x}$, there are more noticeable differences than for the case of a linear regression function. Figure 4 shows the empirical distribution of $\hat{V}_n$ under H_0 when $m(x) = e^{2x}$ for the same values of a as Figure 3.

The overall impression delivered by this set of four figures is that the distribution of $\hat{V}_n$ under the null hypothesis shows noticeable variation with respect to both the choice of window width parameter a for the Nadaraya-Watson estimator of the regression function $m(x)$ and the actual regression function used.

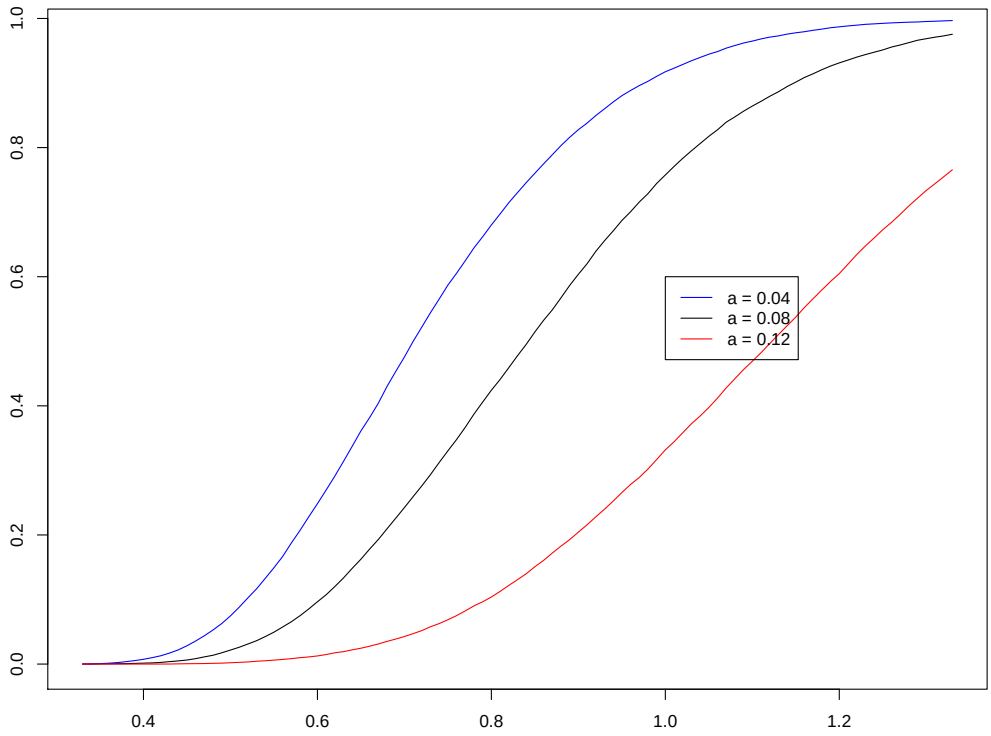

Figure 4: Empirical distributions of $\hat{V}_n$ under H_0 when $m(x) = e^{2x}$.

3.4 The null distribution of W_n

We now proceed to generate the K-S statistics for W_n under H_0 for the same situations as analysed for $\hat{V}_n$.

Unlike the case of $\hat{V}_n$, the distribution of W_n is fixed, and is free from the hypothesis H_0, the regression function m and its estimator $\hat{m}_n$. As a matter of fact

$$P(W_n < x) = \frac{4}{\pi} \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} \exp(-\frac{\pi^2(2n+1)^2}{8x^2})$$

(Shorak 1987 p. 34).

3.4.1 Linear regression

Figures 5 and 6 show the empirical distributions of W_n under H_0 for a selection of values of a when $m(x) = x$ and $m(x) = 5x$ respectively.

3.4.2 Exponential regression function

Figures 7 and 8 show the empirical distributions of W_n under H_0 for a selection of values of a when $m(x) = e^x$ and $m(x) = e^{2x}$ respectively.

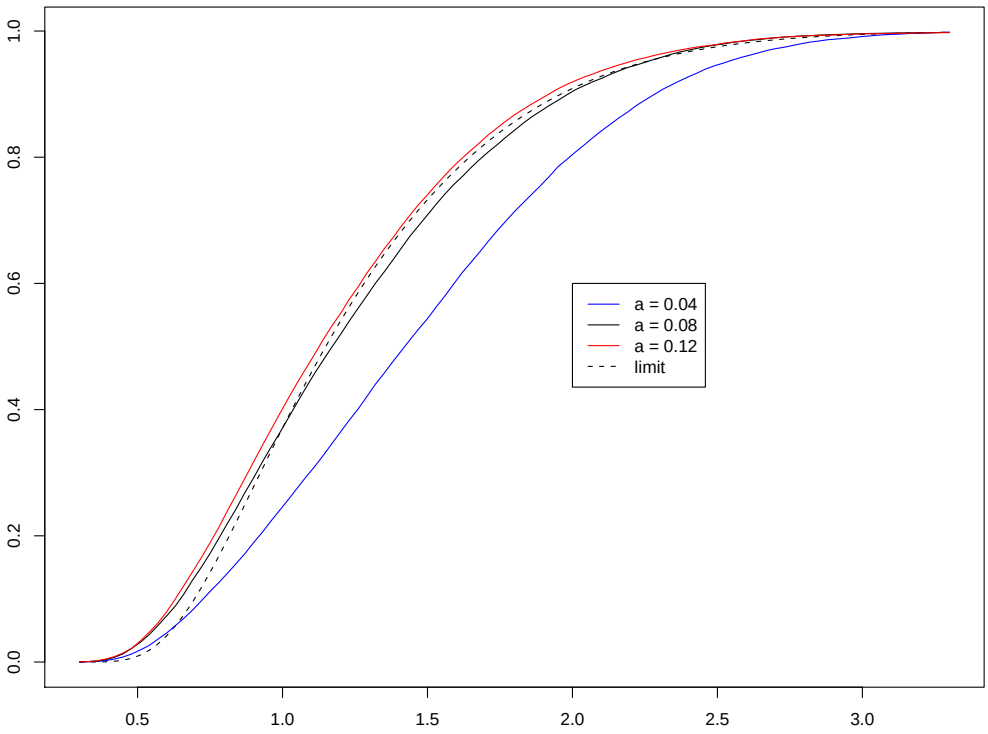

Figure 5: Empirical distributions of W_n under H_0 when $m(x) = x$.

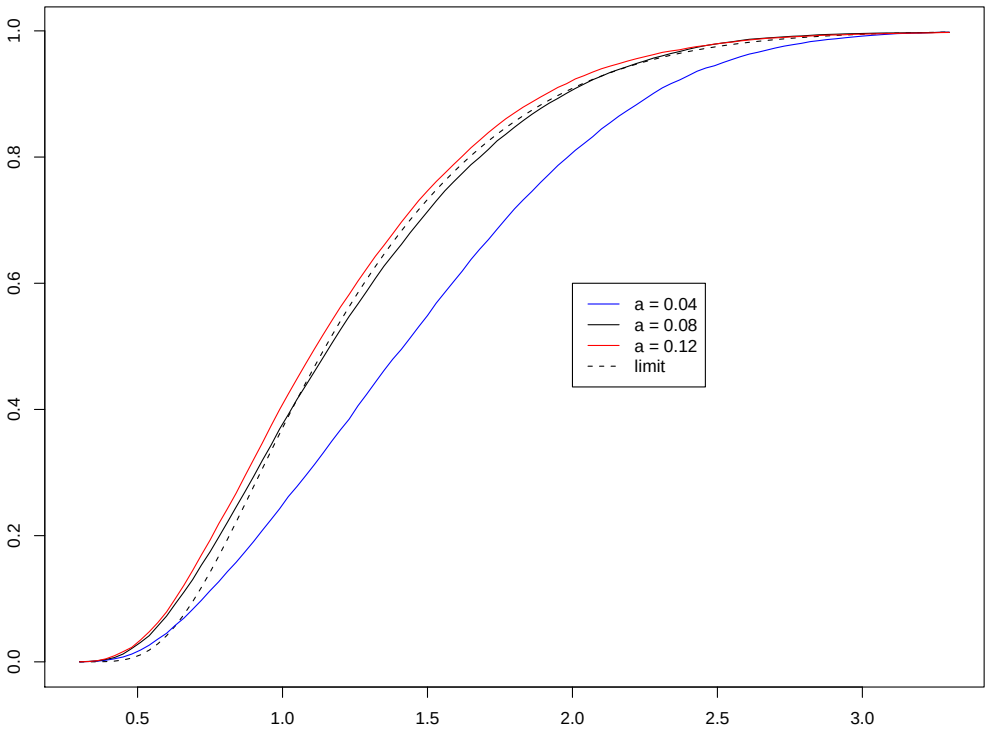

Figure 6: Empirical distributions of W_n under H_0 when $m(x) = 5x$.

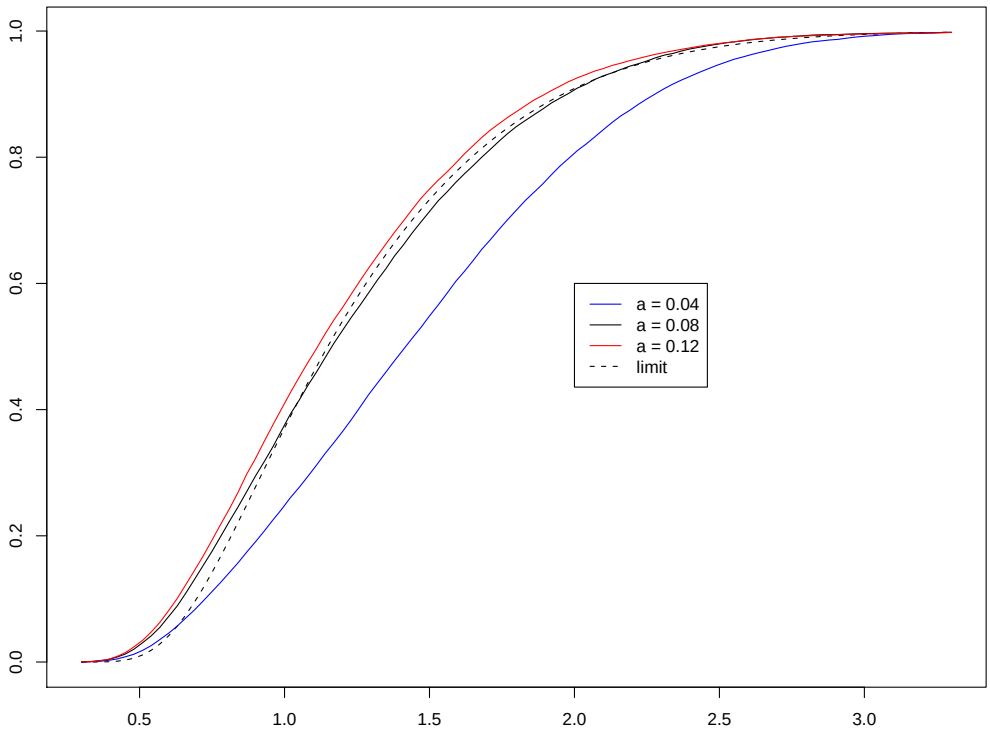

Figure 7: Empirical distributions of W_n under H_0 when $m(x) = e^x$.

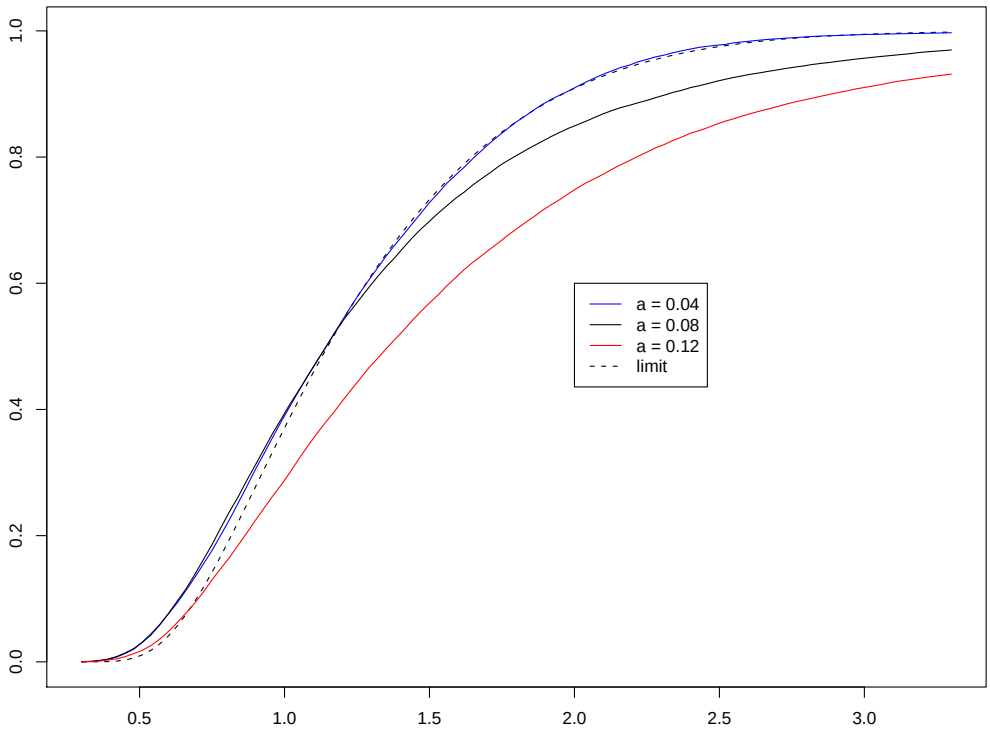

Figure 8: Empirical distributions of W_n under H_0 when $m(x) = e^{2x}$.

4 Power of K-S tests

The alternative distribution chosen for determining the power of various test statistics is a mixture distribution of 80% standard normal distribution and 20% "double exponential" distribution with density $0.5e^{-|x|}$. The density of the mixture, compared with a standard normal density, is shown in Figure 9. Different proportions of the mixture were tested in the early stages of the following analyses, but in general this did not affect the overall picture. Further, some different sample sizes were also tested, but as can be seen below, these do not affect the effects demonstrated either.

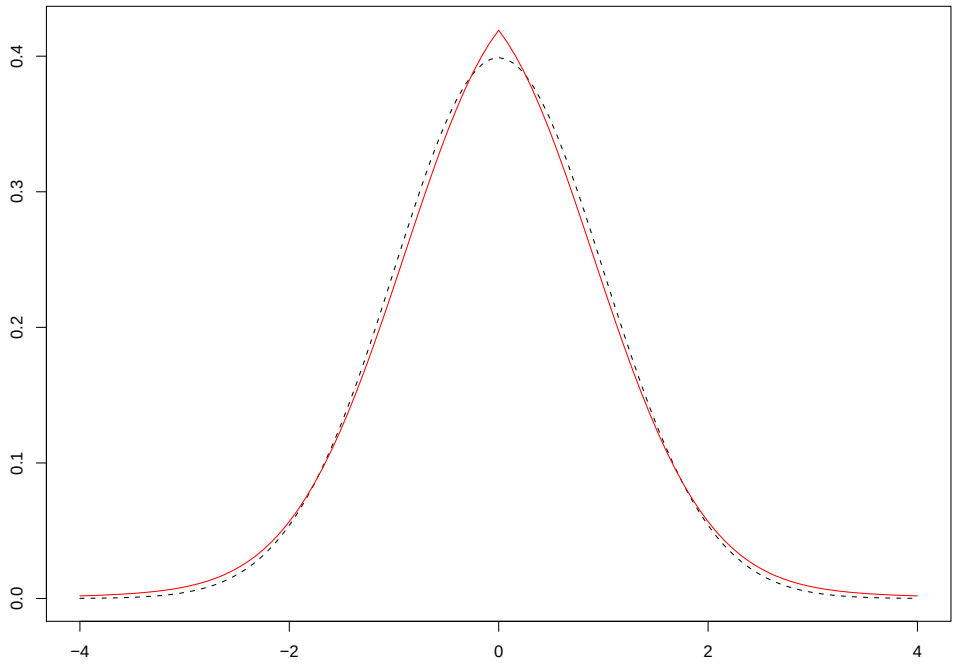

Figure 9: Standard normal density and the density of the mixture of 4/5 standard normal and 1/5 double exponential densities.

4.1 Power of $\hat{V}_n$ compared to $\bar{V}_n$.

The paper Khmaladze and Koul (2007) contains a section which explains why there is an intrinsic loss of power in statistics based on the process $\hat{v}_n(x)$. Graphs below illustrate this point.

For comparison's sake, consider the problem where Y_i-s are i.i.d. with unknown constant mean value m:

$$Y_i = m + e_i,\ i = 1, \ldots, n\ .$$

We estimate m by the average of Y_i-s, and consider the empirical process based on the estimated errors $\bar{e}_i = Y_i - \bar{Y}_n, i = 1, \ldots, n$. This process we denote

$$\bar{v}_n(x) = \sqrt{n}(\bar{F}_n(x) - F(x)),$$

where again the hypothetical F is a standard normal distribution function. Finally, we denote

$$\bar{V}_n = \sup_x |\bar{v}_n(x)|$$

as the K-S statistic from $\bar{v}_n(x)$.

The next set of graphs, Figures 10 – 17, show the simulated distributions of both $\bar{V}_n$ and $\hat{V}_n$ under the null and under the alternative distribution for two different sample sizes, two different regression functions and two different window widths. In all these graphs, the upper (black) and lower (red) dashed lines are respectively the null and alternative distributions for $\bar{V}_n$ and the upper (black) and lower (red) solid lines are respectively the null and alternative distributions for $\hat{V}_n$.

We see that indeed $\bar{V}_n$ has greater power, but sometimes not by much - as was discussed in Khmaladze and Koul (2007).

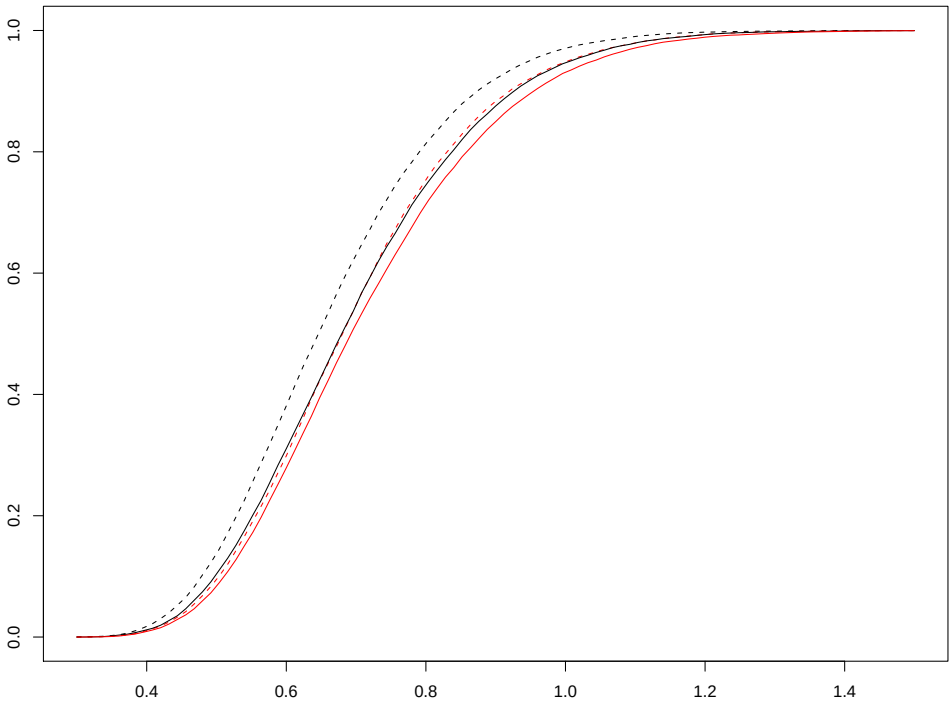

Figure 10: Empirical distributions of $\bar{V}_n$ and $\hat{V}_n$ for null and mixture distribution when $a = 0.12$ and $m(x) = x$. Sample size 200.

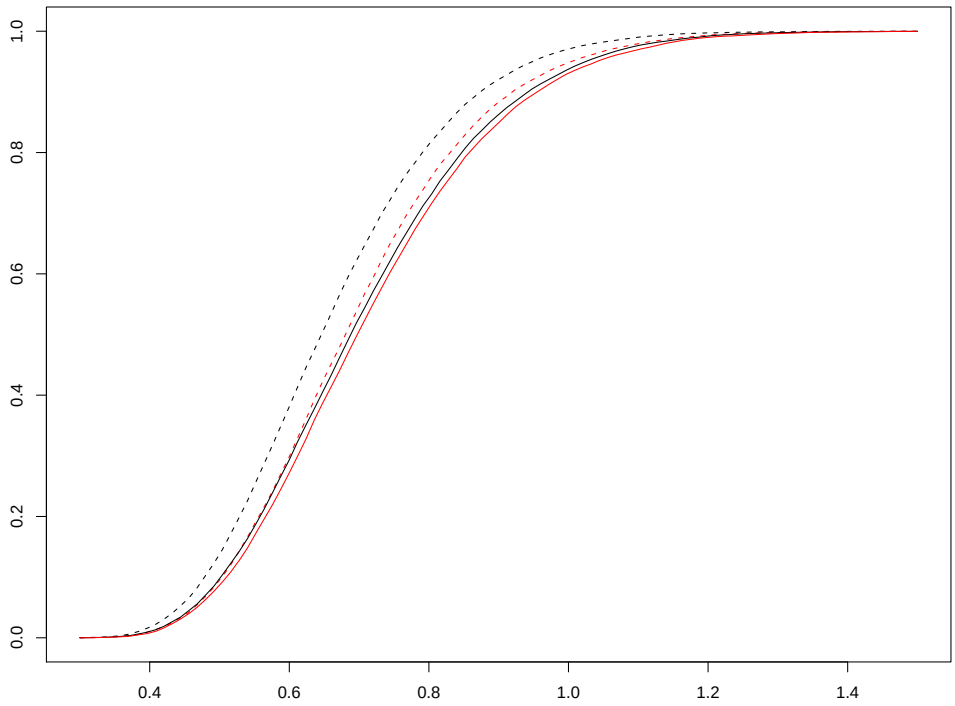

Figure 11: Empirical distributions of $\bar{V}_n$ and $\hat{V}_n$ for null and mixture distribution when $a = 0.08$ and $m(x) = x$. Sample size 200.

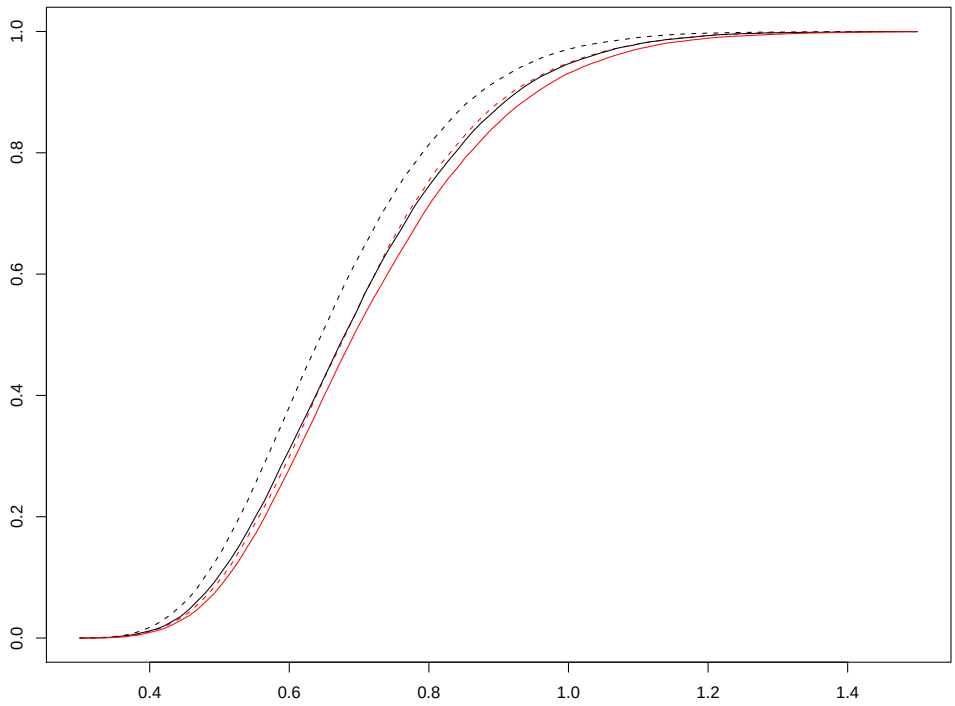

Figure 12: Empirical distributions of $\bar{V}_n$ and $\hat{V}_n$ for null and mixture distribution when $a = 0.12$ and $m(x) = e^x$. Sample size 200.

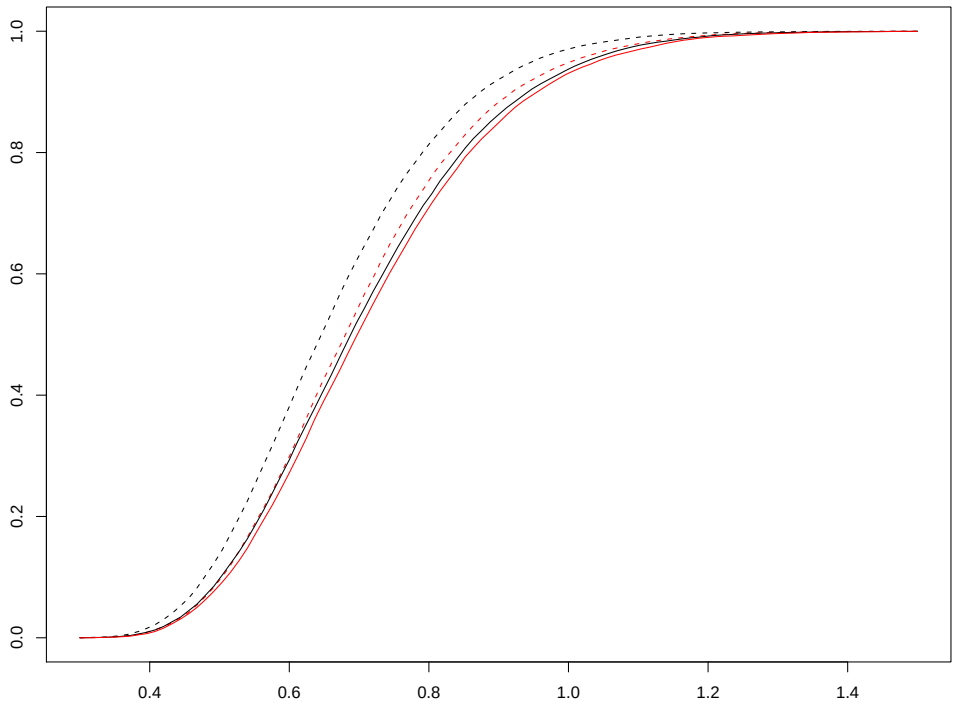

Figure 13: Empirical distributions of $\bar{V}_n$ and $\hat{V}_n$ for null and mixture distribution when $a = 0.08$ and $m(x) = e^x$. Sample size 200.

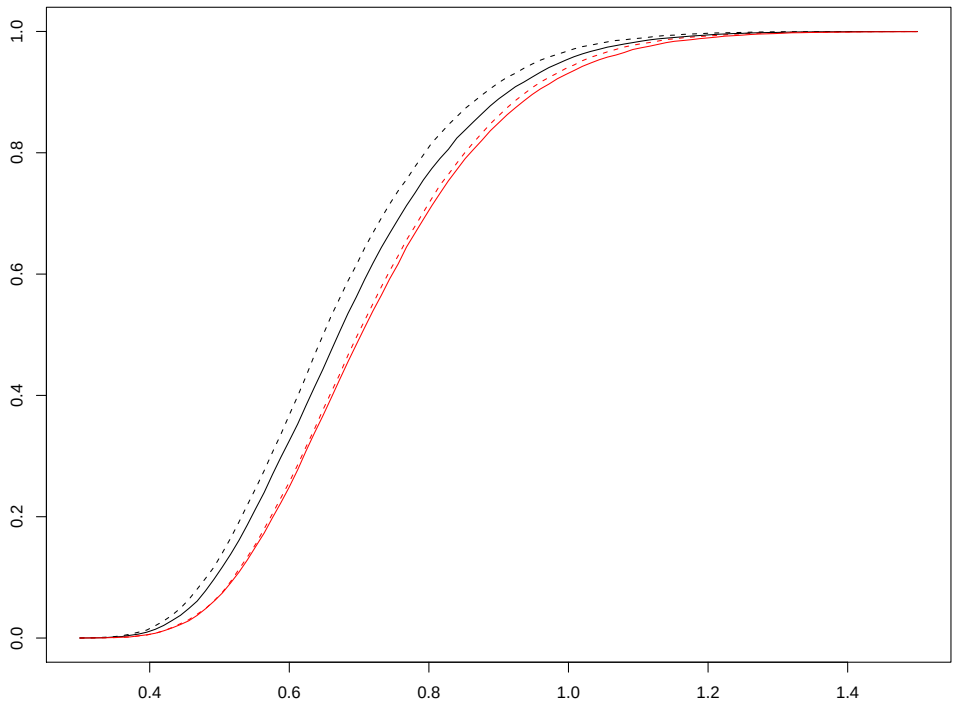

Figure 14: Empirical distributions of $\bar{V}_n$ and $\hat{V}_n$ for null and mixture distribution when $a = 0.12$ and $m(x) = x$. Sample size 500.

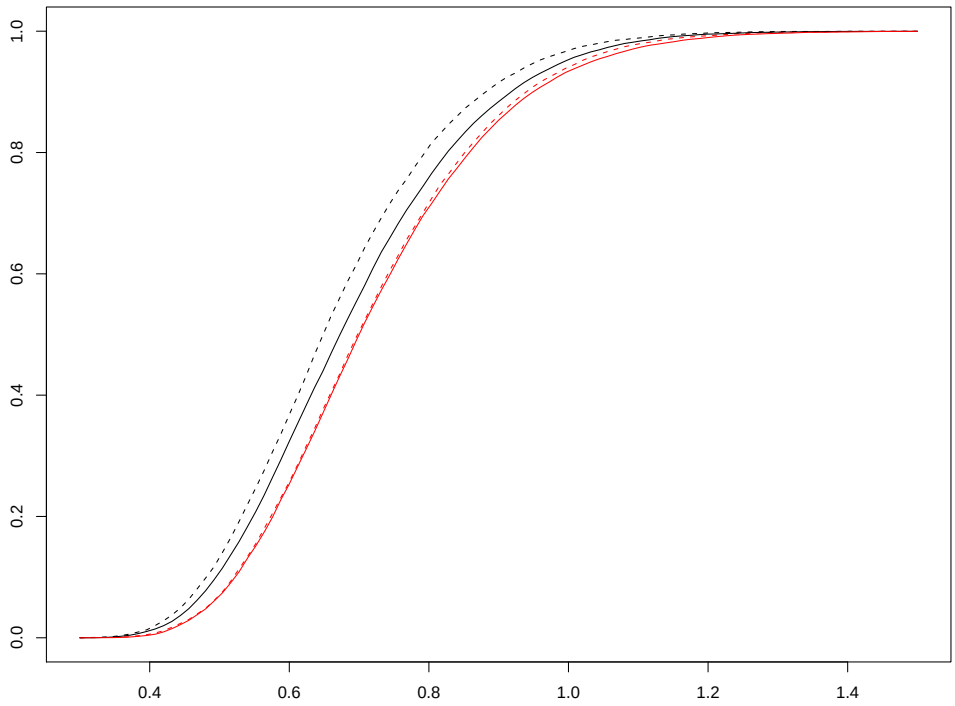

Figure 15: Empirical distributions of $\bar{V}_n$ and $\hat{V}_n$ for null and mixture distribution when $a = 0.08$ and $m(x) = x$. Sample size 500.

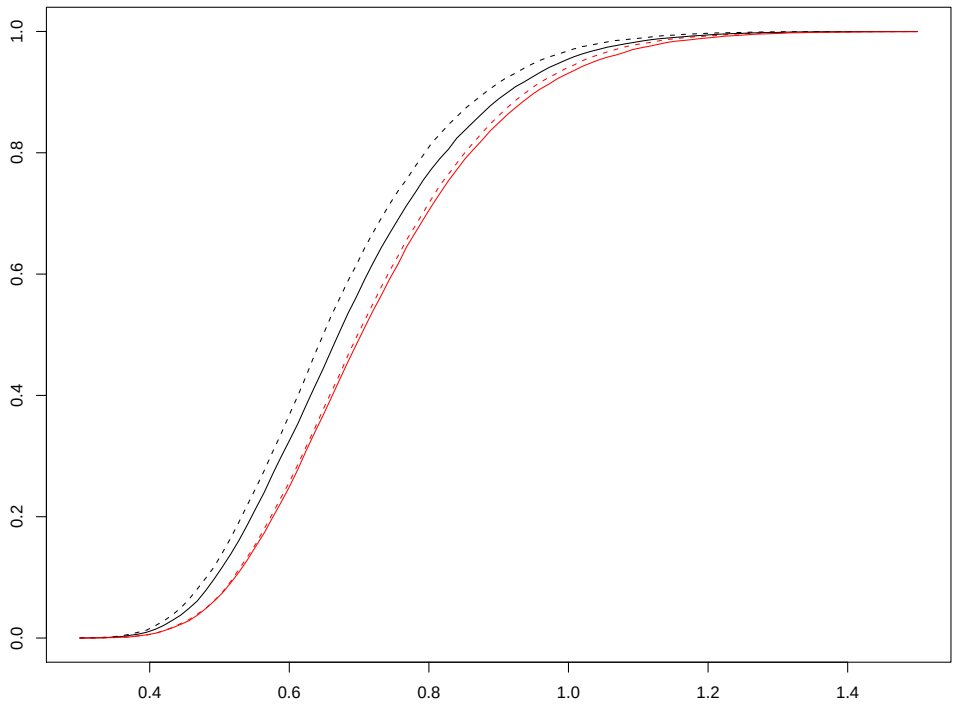

Figure 16: Empirical distributions of $\bar{V}_n$ and $\hat{V}_n$ for null and mixture distribution when $a = 0.12$ and $m(x) = e^x$. Sample size 500.

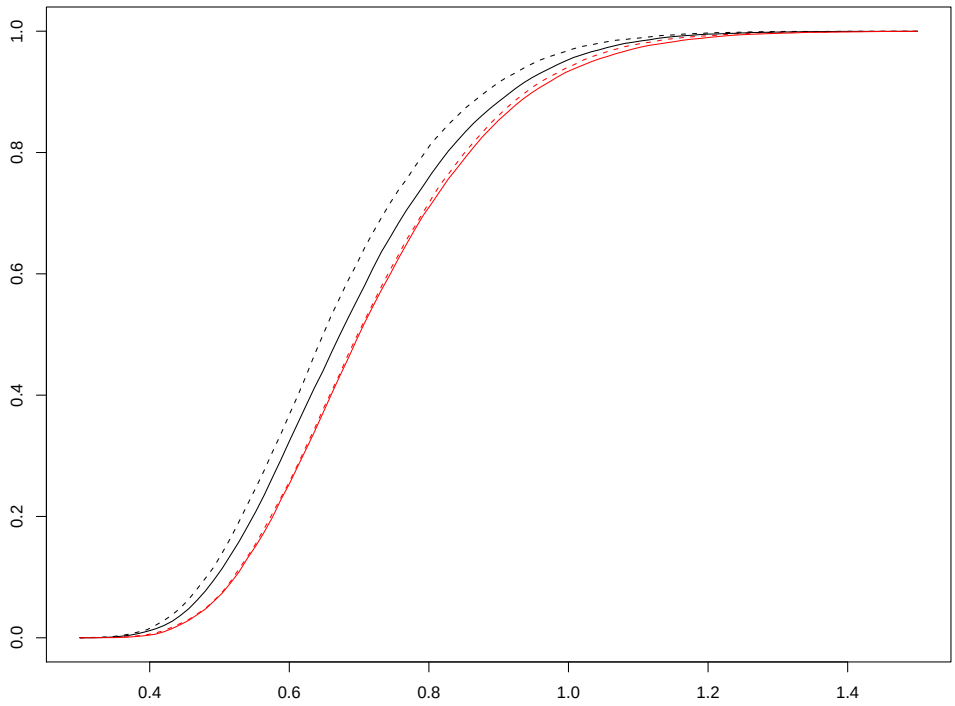

Figure 17: Empirical distributions of $\bar{V}_n$ and $\hat{V}_n$ for null and mixture distribution when $a = 0.08$ and $m(x) = e^x$. Sample size 500.

5 Comparison of Powers of w_n and $\hat{v}_n$

Figures 18 – 21 show, for two different sample sizes and two different regression functions, the limit and simulated null distribution of the statistic W_n and its distribution under the alternative as well as the respective graphs of $\bar{V}_n$ and $\hat{V}_n$. Now the increase of power is much more noticeable.

However, we note that this effect is the superposition of two different effects: one is that some loss of power in $\hat{v}_n$, which was mentioned above - and no such loss is associated with w_n as $\bar{v}_n$ and w_n are, asymptotically in a one-to-one relationship, and another is that the alternative distributions for e_i-s must, again, have mean 0, while K-S tests based on either $\hat{v}_n(x)$ or $\bar{v}_n(x)$ are not very powerful against such alternatives. Indeed, for alternatives with expected value 0, the largest (uniform) deviation from the hypothesis will occur on the "flanks" while the supremum is likely to occur at the point close to 0, where the deviation of the alternative distribution from F will not be sufficiently large (Janssen 2000).

Acknowledgements

I would like to thank Professor Estate Khmaladze for valuable discussions and comments made during the preparation of this report.

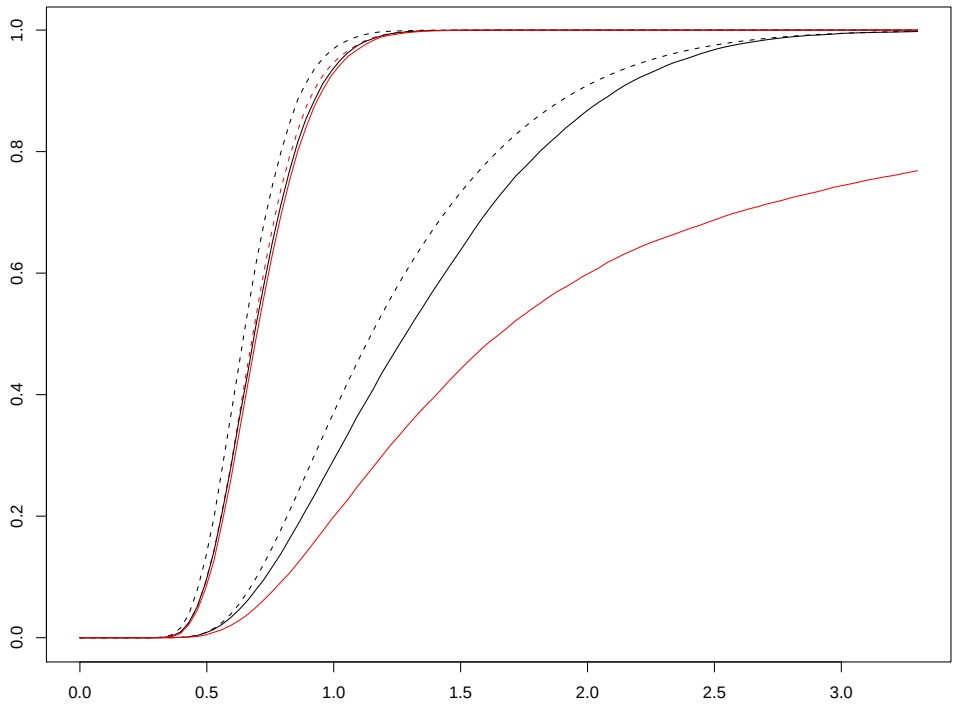

Figure 18: Empirical distributions of $\bar{V}_n$, $\hat{V}_n$ and W_n for null and mixture distribution when $a = 0.08$ and $m(x) = x$. Sample size 200.

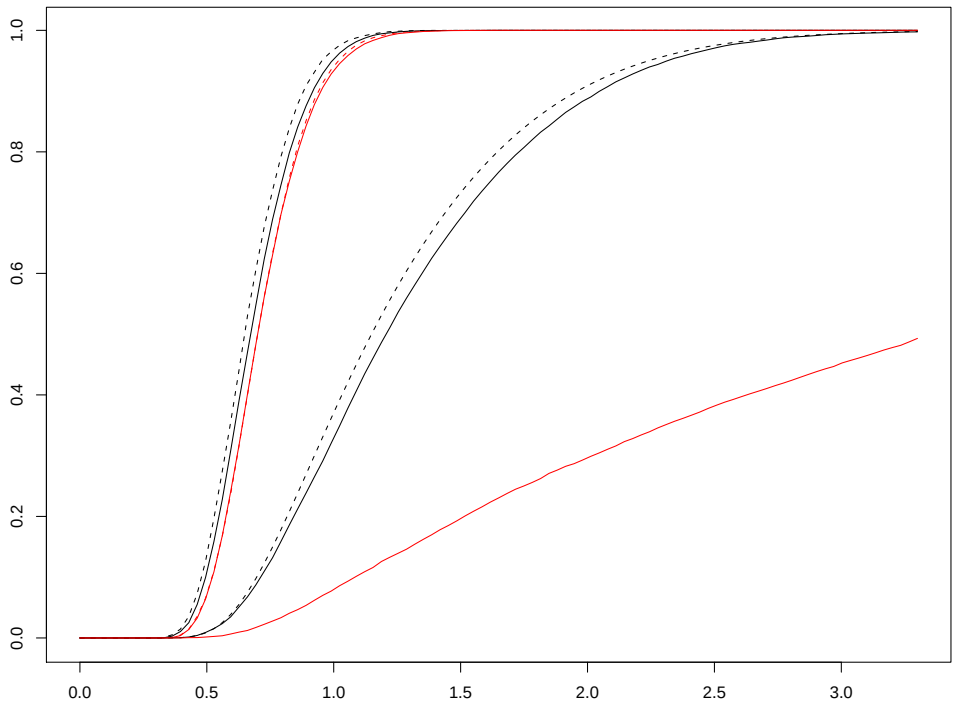

Figure 19: Empirical distributions of $\bar{V}_n$, $\hat{V}_n$ and W_n for null and mixture distribution when $a = 0.08$ and $m(x) = x$. Sample size 500.

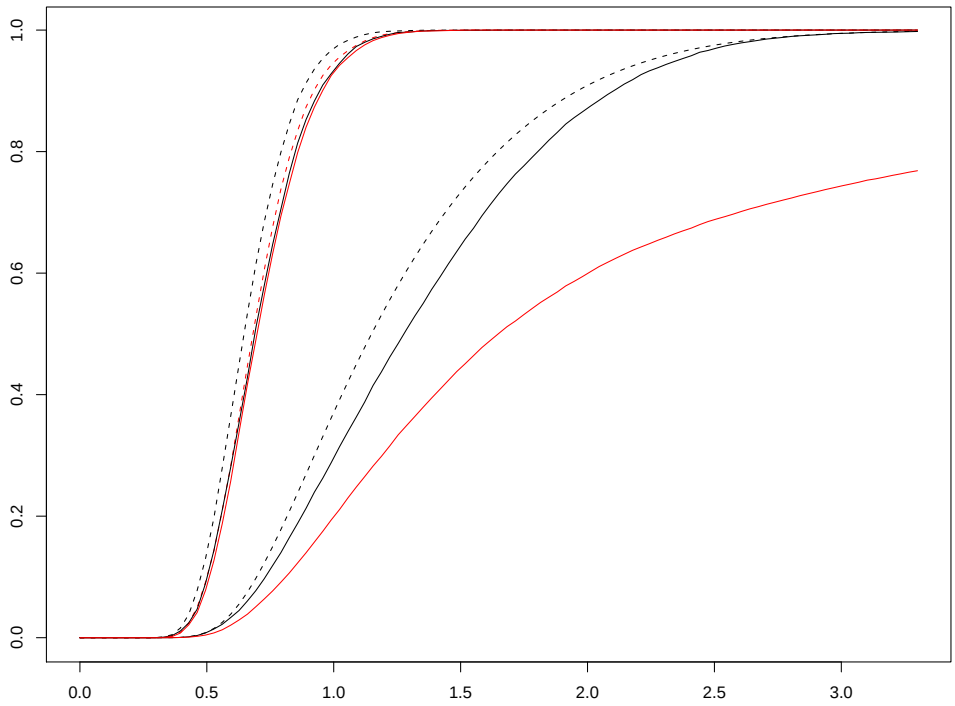

Figure 20: Empirical distributions of $\bar{V}_n$, $\hat{V}_n$ and W_n for null and mixture distribution when $a = 0.08$ and $m(x) = e^x$. Sample size 200.

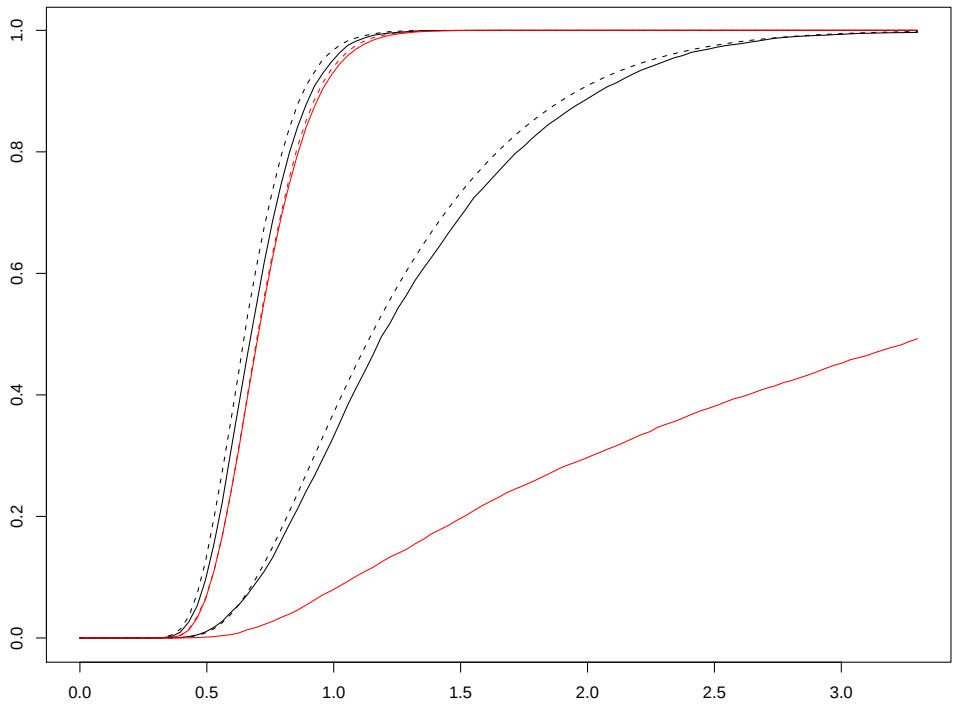

Figure 21: Empirical distributions of $\bar{V}_n$, $\hat{V}_n$ and W_n for null and mixture distribution when $a = 0.08$ and $m(x) = e^x$. Sample size 500.

References

Arnold Janssen, 2000, Global power functions of goodness of fit tests *Ann. Statist.* **28** 1, 239-253.

Estate Khmaladze and Hira L Koul, 2000, Goodness-of-fit problem for errors in non-parametric regression: distribution free approach *VUW MSCS Research Report Series* 2007-06

G.R. Shorak and J.A. Wellner, 1987, *Empirical processes with applications to statistics*, Wiley, New York