Supplemental file

In this supplemental file, some technical details, including the Theorem 1 and its proof, and the approximation to the null steady-state distribution of the CUSUM statistic, are provided.

Appendix

Theorem 1 *The test D_{GOF} has the same detection boundary as the statistic HC_n^* for $\beta \in (0,1)$.*

Proof. We denote by $G(\cdot)$ the cumulative distribution function (CDF) of X_i under H_1. The proof of this theorem is divided into three parts which are stated by the following three lemmas respectively.

Lemma 1 *Under H_0, $Z_C/c_n \xrightarrow{p} 0$ provided $c_n/(\log n \log_2 n) \to \infty$.*

Proof. The idea of the proof comes essentially from the proof of Theorem 3.1 in Jager and Wellner (2007). By the probability integral transformation, we can, without loss of generality, suppose F_0 is the uniform on $[0,1]$. In this situation, Z_C becomes

$$Z_C = 2n \int_0^1 [u(1-u)]^{-1} K(F_n(u), u) du, \tag{A.1}$$

where

$$K(x,y) = \left\{ x \ln\left(\frac{x}{y}\right) + [1-x] \ln\left(\frac{1-x}{1-y}\right) \right\}.$$

Set $a_n = n^{-1} \log n$ and write

$$Z_C = 2n \left(\int_0^{a_n} + \int_{a_n}^{1-a_n} + \int_{1-a_n}^1 \right) [u(1-u)]^{-1} K(F_n(u), u) du \equiv \Delta_1 + \Delta_2 + \Delta_3.$$

Firstly, we show that $\Delta_1/c_n \xrightarrow{p} 0$. To show this, fix $\epsilon > 0$ and choose $\lambda = \lambda_\epsilon$ so large that

$$P(||F_n(u)/u||_0^1 > \lambda) < \epsilon,$$

where $||F_n(u)/u||_a^b \equiv \sup_{a\le u\le b}(F_n(u)/u)$ and the above result comes from Lemma 2 in Wellner (1978). On the event $||F_n(u)/u||_0^1 \le \lambda$ we have

$$\begin{aligned}\Delta_1/c_n &= \frac{2n}{c_n}\int_0^{a_n}(1-u)^{-1}\frac{F_n(u)}{u}\log\frac{F_n(u)}{u}du + o(1) \quad \text{a.s.}\\ &\le \frac{2n}{c_n}\int_0^{a_n}\frac{F_n(u)}{u}\log\frac{F_n(u)}{u}du(1+o(1))\\ &\le \lambda\log\lambda\cdot\frac{2n}{c_n}a_n \to 0.\end{aligned}$$

since $c_n/\log n \to \infty$.

By symmetry, $\Delta_3/c_n \xrightarrow{p} 0$ immediately. Thus, it remains to show $\Delta_2/c_n \xrightarrow{p} 0$. Note that $||K(F_n(u),u)||_0^1$ is the well-known Berk-Jones goodness-of-fit test statistic. By Theorem 1.1 in Wellner and Koltchinskii (2003), it can be easily seen that

$$\frac{n||K(F_n(u),u)||_0^1}{\log_2 n} \xrightarrow{p} 1. \tag{A.2}$$

Consequently,

$$\begin{aligned}\Delta_2/c_n &\le \frac{2n||K(F_n(u),u)||_0^1}{c_n}\int_{a_n}^{1-a_n}\frac{1}{u(1-u)}du\\ &= 4nc_n^{-1}||K(F_n(u),u)||_0^1(\log n - \log_2 n)(1+o(1)) \xrightarrow{p} 0, \end{aligned}\tag{A.3}$$

which completes the proof of this lemma. □

This lemma tells us that Z_C grows to infinity very slowly under the null hypothesis. Furthermore, by Lemma 5 in the Appendix A.3 of the supplemental file, we know $D_{\mathrm{GOF}} = Z_C + O(\log n \log_2 n)$. Thus, a convenient critical point for rejecting the null hypothesis is when $D_{\mathrm{GOF}} > \log^2 n$ and the test is accordingly $T_{\text{new}} = I(D_{\mathrm{GOF}} > \log^2 n)$.

Lemma 2 *Suppose that ε_n and μ_n satisfy the sparse regime and $r > \rho^*(\beta)$ in (**??**). Then the test T_{new} satisfies*

$$P_{H_0}(T_{new}=1) + P_{H_1}(T_{new}=0) \to 0, \quad \textit{as} \quad n\to\infty.$$

Proof. By Lemma 1, $P_{H_0}(T_{\text{new}} = 1) \to 0$ is obvious. As in Donoho and Jin (2004), we examine the cases $r > (1-\sqrt{1-\beta})^2$ and $r < \beta/3$ separately; those two cases overlap and together cover the full region$1/2 < \beta < 1, r > \rho^*(\beta)$.

For the first case, since $(r+\beta)/(2\sqrt{r}) < 1$, we can pick a constant $q < 1$ such that $\max\{(r+\beta)/(2\sqrt{r}), \sqrt{r}\} < \sqrt{q} < 1$. As argued by Donoho and Jin (2004), under H_1, $\#\{i : p_i \leq n^{-q}\} \sim \text{Binomial}(n, L_n n^{-[\beta+(\sqrt{q}-\sqrt{r})^2]})$, where $L_n n^{-[\beta+(\sqrt{q}-\sqrt{r})^2]} >> n^{-q}$; here L_n is a logarithmic term that does not significantly contribute to the argument. Accordingly, $[(i-1)/n]/p_{(i)} >> 1$ for those p-values $p_{(i)} \leq n^{-q}$.

Note that

$$
\begin{aligned}
D_{\text{GOF}} &\geq \sum_{i=1}^{n} \left\{\log\left[\frac{F_0(X_{(i)})^{-1}-1}{n/i-1}\right]\right\}^2 I(F_0(X_{(i)}) \geq i/n) \\
&= \sum_{i=1}^{n} \left\{\log\left[\frac{p_{(i)}}{\frac{i-1}{n}}\right] - \log\left[\frac{1-p_{(i)}}{1-\frac{n-i+1}{n}}\right]\right\}^2 I(p_{(i)} < (i-1)/n) \\
&\geq \sum_{i=1}^{\#\{i:p_i\leq n^{-q}\}} \left\{\log\left[\frac{p_{(i)}}{\frac{i-1}{n}}\right] + O_p(n^{-[\beta+(\sqrt{q}-\sqrt{r})^2]})\right\}^2 \sim L_n \text{Binomial}(n, L_n n^{-[\beta+(\sqrt{q}-\sqrt{r})^2]}).
\end{aligned}
$$

By Chebyshev's inequality, $P_{H_1}(D_{\text{GOF}} < \log^2 n) \to 0$ as $1-\beta-(\sqrt{q}-\sqrt{r})^2 > 0$. We conclude that D_{GOF} is able to separate H_0 and H_1 in this range.

Now, suppose that $r < \beta/3$ and then $r < 1/4$. Similar to (A.1), we can rewrite Z_C^+ as

$$
Z_C^+ = 2\int_0^1 [u(1-u)]^{-1} nK^+(F_n(u), u)du, \tag{A.4}
$$

where $F_n(u) = n^{-1}\sum_{i=1}^{n} I(Y_i \leq u)$, $Y_i \equiv 1-\Phi(X_i)$ i.i.d. $F = 1-G(\Phi^{-1}(1-\cdot))$ (with the Y_i's taking values in $[0,1]$), and $K^+(F_n(u), u) = K(F_n(u), u)I(F_n(u) \geq u)$. From Lemma 7.3 in Jager and Wellner (2007), the convergence

$$
\sup_{n^{-4r}\leq u\leq n^{-4r_0}} \left|\frac{F_n(u)}{u} - 1\right| \xrightarrow{p} 0
$$

holds, provided that $0 < \rho^*(\beta) < r < \beta/3$ and $0 < r_0 < r < 1/4$. Thus, by part (ii) of Lemma 7.2 in Jager and Wellner (2007), it follows that for $n^{-4r} \leq u \leq n^{-4r_0}$, we have

$$
2nK^+(F_n(u), u) = n\left(\frac{(F_n(u)-u)^+}{\sqrt{u(1-u)}}\right)^2 (1+o_p(1)),
$$

Hence, by letting $r_0 = r - [4\log n]^{-1}$

$$\begin{aligned} Z_C^+ &\geq \int_{n^{-4r}}^{n^{-4r_0}} [u(1-u)]^{-1} n \left(\frac{(F_n(u)-u)^+}{\sqrt{u(1-u)}} \right)^2 (1+o_p(1)) du \\ &\equiv n \left(\frac{(F_n(u^*)-u^*)^+}{\sqrt{u^*(1-u^*)}} \right)^2 (1+o_p(1)) \end{aligned}$$

for some u^* satisfying $u^* \in (n^{-4r}, n^{-4r_0})$. The remaining proof is essentially the same as the proof of Theorem 1.2 in Donoho and Jin (2004; pp. 976) and the only difference is that we use some $r^* \in (r_0, r)$ rather than r. However, by noting that $r^* = r + o(1)$ and the result follows immediately with the fact that $Z_C \geq Z_C^+$ and Lemma 8. $\square$

Lemma 3 *Suppose that ε_n and μ_n satisfy the dense regime and $r < \rho^*(\beta)$ in (**??**). Then the test T_{new} satisfies*

$$P_{H_0}(T_{new} = 1) + P_{H_1}(T_{new} = 0) \to 0, \quad as \;\; n \to \infty.$$

Proof. The proof of this lemma is essentially similar to that of Theorem 7 in Cai et al. (2011) but some details may differ much. Recall the expression of D_{GOF} in (A.4).

Let $u_n = \bar{\Phi}(1) + [\log n]^{-1}$. It can be seen that

$$\begin{aligned} E\left[F_n(\bar{\Phi}(1)) - \bar{\Phi}(1)\right] &= \varepsilon_n \left[\bar{\Phi}\left(1-\mu_n\right) - \bar{\Phi}(1)\right] \sim C n^{-\beta}\mu_n \\ \mathrm{var}\left[F_n(\bar{\Phi}(1)) - \bar{\Phi}(1)\right] &= O(n^{-1}), \end{aligned}$$

and hence $F_n(\bar{\Phi}(1))/\bar{\Phi}(1) \xrightarrow{p} 1$. By the second-order Taylor expansion (similar to the proof of Lemma A.4 in Donoho and Jin 2004), we have for $\bar{\Phi}(1) \leq u \leq u_n$,

$$2nK^+(F_n(u), u) = n(1-\bar{\Phi}(1)) \left(\frac{F_n(u)-u}{\sqrt{u(1-u)}} \right)^2 (1+o_p(1)),$$

and correspondingly

$$\begin{aligned} Z_C^+ &\geq \int_{\bar{\Phi}(1)}^{u_n} [u(1-u)]^{-1} n(1-\bar{\Phi}(1)) \left(\frac{F_n(u)-u}{\sqrt{u(1-u)}} \right)^2 (1+o_p(1)) du \\ &\sim L_n n \left(\frac{F_n(u^*)-u^*}{\sqrt{u^*(1-u^*)}} \right)^2 (1+o_p(1)) \equiv L_n W_n^2(u^*)(1+o_p(1)) \end{aligned}$$

for some u^* satisfying $u^* \in (\bar{\Phi}(1), u_n)$. Now it suffices to show $P\left(W_n(u^*) < L_n\right) \to 0$. By direct calculation, when $r < 1/2 - \beta$

$$E[W_n(\bar{\Phi}(1))] \sim n^{\gamma}$$

for some $\gamma > 0$ and thus $E[W_n(\bar{\Phi}(1))]/\log n \to \infty$. It follows immediately from Chebyshev's inequality

$$\begin{aligned} P\left(W_n(u^*) < L_n\right) &= P\big(W_n(\bar{\Phi}(1))(1 + o_p(1)) < L_n\big) \\ &\leq C\frac{\mathrm{var}(W_n(\bar{\Phi}(1)))}{E[W_n(\bar{\Phi}(1))]^2} \leq Cn^{-2\gamma} \to 0, \end{aligned}$$

which completes the proof of this lemma. $\square$

A.2 The empirical steady-State distribution of the CUSUM Statistic

Grigg and Spiegelhalter (2008) suggested the following approximation:

$$H(x;\mu) \approx \begin{cases} 0, & \text{if} \quad x = 0, \\ 1 - \exp\left\{ y_0 - \sqrt{r^2 - (x - x_0)^2} \right\}, & \text{if} \quad 0 < x \leq x_1, \\ 1 - \gamma\exp\{-x\}, & \text{otherwise}, \end{cases}$$

where

$$\begin{aligned} &\gamma_0 = \exp\{-0.651(\mu - 0.277)\} + 0.031\mu - 0.189, \quad \gamma = \exp\{-0.578(\mu + 0.024)\} + 0.006\mu, \\ &x_1 = 0.17\mu^2 + 1.052\mu - 0.02, \quad \rho_0 = \log(\gamma_0), \quad \rho = \log(\gamma) - \rho_0, \\ &u = \sqrt{x_1^2 + (\rho - x_1)^2}, \quad v = \sqrt{2}x_1 - \rho/\sqrt{2}, \quad \theta = \arcsin(v/u), \\ &r = u/(2\sin(\pi/2 - \theta)), \quad x_0 = -r\sin(2\theta - 0.75\pi), \quad y_0 = \rho_0 - r\cos(2\theta - 0.75\pi). \end{aligned}$$

A.3 Some technical Lemmas

Define

$$C_n = -2n\int_0^1 \frac{\log(1-u)}{u}\mathrm{d}u - \sum_{i=1}^{n}(b_{i-1} - b_i)^2,$$

where $b_i = i\log\frac{i}{n} + (n-i)\log(1-\frac{i}{n})$ for $i = 1, \ldots, n-1$ and $b_0 = b_n = 0$.

Lemma 4 *$C_n = O(\log n)$.*

Proof. Introduce a function

$$f(x) = x\log(x) + (1-x)\log(1-x)$$

with $f(0) = f(1) = 0$. Since

$$-2\int_0^1 \frac{\log(1-u)}{u}\mathrm{d}u = \int_0^1 \left(\log\frac{1-u}{u}\right)^2 \mathrm{d}u,$$

we have

$$\begin{aligned}
C_n =& n\int_0^1 \left(\log\frac{1-u}{u}\right)^2 \mathrm{d}u - b_1^2 - b_{n-1}^2 - n^2\sum_{i=2}^{n-1}\left(f(\frac{i-1}{n}) - f(\frac{i}{n})\right)^2 \\
=& 2\left[n\int_0^{\frac{1}{n}} \left(\log\frac{1-u}{u}\right)^2 \mathrm{d}u - b_1^2\right] + n\left[\int_{\frac{1}{n}}^{1-\frac{1}{n}} \left(\log\frac{1-u}{u}\right)^2 \mathrm{d}u - n\sum_{i=2}^{n-1}\left(f(\frac{i-1}{n}) - f(\frac{i}{n})\right)^2\right] \\
\equiv& C_{n1} + C_{n2} \qquad (A.5)
\end{aligned}$$

First consider the term C_{n2} in (A.5). By using Lagrange mean value theorem

$$n\sum_{i=2}^{n-1}\left(f(\frac{i-1}{n}) - f(\frac{i}{n})\right)^2 = \sum_{i=2}^{n-1} n^{-1}\left(\log\frac{\xi_i}{1-\xi_i}\right)^2 = \int_{\frac{1}{n}}^{1-\frac{1}{n}} (\log\frac{1-u}{u})^2 \mathrm{d}u + O(n^{-1}),$$

where $\frac{i-1}{n} < \xi_i < \frac{i}{n}, 2 \le i \le n-2$. Then, we can conclude that the last term of C_{n2} should be of $O(1)$. Now, rewrite C_{n1} as follows:

$$\begin{aligned}
& n\int_0^{\frac{1}{n}} \left(\log\frac{1-u}{u}\right)^2 \mathrm{d}u - b_1^2 \\
&= n\int_0^{\frac{1}{n}} \left(\log\frac{1-u}{u}\right)^2 \mathrm{d}u - \left(\log\frac{1}{n} + (n-1)\log(1-\frac{1}{n})\right)^2 \\
&= -\frac{n\int_0^{\frac{1}{n}} \left(\log\frac{1-u}{u}\right)^2 \mathrm{d}u - \left(\log\frac{1}{n} + (n-1)\log(1-\frac{1}{n})\right)^2}{\log\frac{1}{n}} \times \log n
\end{aligned}$$

Introduce another function

$$g(t) = \frac{\int_0^t (\log\frac{1-u}{u})^2 \mathrm{d}u}{t\log t} - \frac{\left(\log t + \frac{1-t}{t}\log(1-t)\right)^2}{\log t}.$$

It can be easily checked that $g(t) \to 1$, as $t \to 0$. Hence $C_{n1} = O(\log n)$, which implies that $C_n = O(\log n)$. $\square$

Lemma 5 *Under the assumptions in Theorem 1 and* H_1*,* $D_{\mathrm{GOF}} = Z_C + O(\log n \log_2 n)$*.*

Proof. Under the conditions, we have $G(X_{(i)}) < \Phi(X_{(i)})$. Accordingly,

$$\begin{aligned}
D_{\mathrm{GOF}} &= \sum_{i=1}^{n} \left\{ \log \left[\frac{[\Phi(X_{(i)})]^{-1} - 1}{(n-1/2)/(i-3/4) - 1} \right] \right\}^2 I_{\{\Phi(X_{(i)}) > (i-3/4)/n\}} \\
&= Z_C - C_n - \sum_{i=1}^{n} \left\{ \log \left[\frac{[\Phi(X_{(i)})]^{-1} - 1}{(n-1/2)/(i-3/4) - 1} \right] \right\}^2 I_{\{\Phi(X_{(i)}) < (i-3/4)/n\}} \\
&\leq Z_C - O(\log n) - \sum_{i=1}^{n} \left\{ \log \left[\frac{[G(X_{(i)})]^{-1} - 1}{(n-1/2)/(i-3/4) - 1} \right] \right\}^2 ,
\end{aligned}$$

where we use Lemma 4. By Lemma 1 in Appendix A.1, the last term of the right side of the inequality above is of order $O_p(\log n \log_2 n)$ because it is essentially distributed as the same as Z_C under null hypothesis except the constant C_n. The assertion holds immediately. $\square$

References:

Wellner, J. A. (1978), "Limit Theorems for the Ratio of the Empirical Distribution Function to the True Distribution Function," *Z. Wahrsch. Verw. Gebiete*, 45, 73–88.

Wellner, J. A., and Koltchinskii, V. (2003), "A Note on the Asymptotic Distribution of Berk-Jones Type Statistics Under the Null Hypothesis, *High Dimensional Probability III* (J. Hoffmann-Jørgensen, M. B. Marcus and J. A. Wellner, eds.) 321–332. Birkhäser, Basel.