Multiple testing in rolling-window analysis via p-value combination

Xing Ling, Qian Wu[*] and Kun Chen

School of Statistics and Data Science, Southwestern University of Finance and Economics, China

Abstract

Rolling-window methods are widely used to explore time variation in economic relationships. Testing across overlapping windows induces a multiple-testing problem and can inflate the familywise error rate. Existing corrections are often either overly conservative or computationally intensive. We propose a harmonic mean p-value procedure that provides a global test together with window-level diagnostics. We establish asymptotic familywise error rate control under conditions that accommodate overlap dependence. Simulations using several standard time-series tests show accurate size, higher power, and substantially lower computational cost than bootstrap-based extreme-statistic procedures.

JEL classification: C12; C32; C58.

Keywords: Bootstrap; Familywise error rate; Hypothesis testing; Time series.

1 Introduction

Structural change and parameter instability are common in macroeconomic and financial time series (e.g. Andrews, 1993; Bai and Perron, 1998; Cheng et al., 2016; Hanson, 2002; Stock and Watson, 1996). A common way to explore such time variation is to estimate a model using rolling windows and track how coefficients or test statistics evolve over time (e.g. Diebold and Yilmaz, 2014; Giacomini and White, 2006; Giacomini and Rossi, 2010; Inoue et al., 2017; Yousuf and Ng, 2021). Rolling-window methods are widely used in applications such as Granger-causality analysis, market-efficiency tests, and time-varying dependence measures.

A key statistical complication of rolling-window analysis is multiplicity, which is caused by testing across overlapping windows. Interpreting each window-level test at its nominal level can inflate the familywise error rate (FWER) (e.g. Harvey et al., 2016; Romano and Wolf, 2005; White, 2000). Some studies interpret nominal significance in individual windows as evidence of time variation, without adjusting for multiplicity.

[*] Correspondence to: School of Statistics and Data Science, Southwestern University of Finance and Economics, Chengdu, China; E-mail address: wuqj@swufe.edu.cn (Q. Wu)

A convenient option is to apply generic multiplicity corrections such as Bonferroni or Benjamini–Hochberg (Benjamini and Hochberg, 1995), which can be conservative under the strong dependence induced by overlapping windows. Another approach resamples under the global null to calibrate an extreme rolling statistic, as in sup-type structural-break tests (e.g. Andrews, 1993; Shi et al., 2018, 2020). It provides a global decision, but obtaining window-level diagnostics typically requires an additional step, such as a stepdown rule (Romano and Wolf, 2005). This approach can be computationally demanding, since each draw repeats the rolling estimation across all windows.

We frame rolling-window inference as a multiple-testing problem and focus on FWER control. We make three contributions. First, we combine window-wise p-values via the harmonic mean p-value (HMP; Wilson, 2019). This yields a global decision and a window-level diagnostic rule. Second, we establish verifiable conditions under which the HMP calibration controls the FWER asymptotically in the presence of overlap dependence. The argument is general and applies to any local test satisfying these conditions. Third, simulations show reasonable finite-sample properties with substantial computational savings relative to bootstrap-based extreme-statistic procedures. An empirical application further illustrates how the global decision and the diagnostic rule can be used together.

2 Multiple-testing correction in rolling windows

Rolling-window analysis turns a single time-series question into a sequence of related hypothesis tests. Let $\{Z_t\}_{t=1}^T$ be the observed sample. Use windows of length $m = m_T$ and define $W_k = \{k, k+1, \ldots, k+m-1\}$ for $k = 1, \ldots, K$ with $K = T-m+1$. Let $H_0^{(k)}$ be the window-wise null based on observations $\{Z_t : t \in W_k\}$, and let $p_{k,m}$ be the corresponding p-value. The global null of interest is

$$\boldsymbol{H}_0 = \bigcap_{k=1}^{K} H_0^{(k)},$$

which states that the null holds in every window. If we reject $\boldsymbol{H}_0$ whenever any $H_0^{(k)}$ is rejected at level α, the FWER can substantially exceed α when K is large. This is the multiplicity problem induced by rolling windows.

A common response is to calibrate an extreme rolling statistic, such as $\min_k p_{k,m}$, by bootstrap under the global null. A typical implementation estimates a restricted model imposing $\boldsymbol{H}_0$, generates a pseudo-sample by resampling residuals or simulating innovations, reruns the rolling estimations, and records the extreme statistic. Repeating this for B draws yields an empirical null distribution of the extreme statistic and a global bootstrap p-value. One limitation is interpretability. This procedure delivers a global decision, but does not directly provide FWER-controlled window-level diagnostics without an additional stepdown rule or other decision rules. Another concern is computational cost. Each draw repeats the rolling computation over all K windows, so the cost grows quickly with B, K, and model complexity.

Instead, we use HMP to aggregate the window-wise p-values $p_{1,m}, \dots, p_{K,m}$. With equal weights, the harmonic mean p-value is

$$\mathring{p}_H = \left(\frac{1}{K} \sum_{k=1}^{K} \frac{1}{p_{k,m}} \right)^{-1}.$$

Under marginal validity and lower-tail independence of the p-values, the Landau approximation in Wilson (2019) yields a calibrated combined p-value:

$$p_H = 1 - F_L \left(\mathring{p}_H^{-1} \right),$$

where F_L is the corresponding Landau CDF. When a non-negligible subset of window-wise p-values is small, p_H becomes small as well. This behavior makes HMP particularly suitable for rolling-window settings, where departures from the null typically persist over several adjacent windows rather than appearing as a single isolated event.

To justify this calibration, we work in an asymptotic regime where $T \to \infty$, $m = m_T \to \infty$, and $K = T - m + 1 \to \infty$. Assumptions 1–4 formalize the requirements in a rolling-window setting. Let $\{Z_t\}_{t \in \mathbb{Z}}$ be the underlying process generating the observed sample $\{Z_t\}_{t=1}^{T}$. Let $h : \mathcal{Z} \to \mathbb{R}^d$ be measurable, define $\theta_0 := \mathbb{E}[h(Z_0)]$, and set $X_t := h(Z_t) - \theta_0$. For each window, define the window mean $\widehat{\theta}_{k,m} = m^{-1} \sum_{t \in W_k} h(Z_t) \in \mathbb{R}^d$ and the window fluctuation $U_{k,m} := \sqrt{m}\,(\widehat{\theta}_{k,m} - \theta_0) \in \mathbb{R}^d$.

Assumption 1. *(Data). Under $\boldsymbol{H}_0$, $\{X_t\}$ is strictly stationary and strongly mixing with coefficients $\alpha(r)$. There exists $\eta > 0$ such that $\mathbb{E}\|X_0\|^{2+\eta} < \infty$ and $\sum_{r=1}^{\infty} \alpha(r)^{\eta/(2+\eta)} < \infty$. The long-run covariance $\Gamma := \sum_{j \in \mathbb{Z}} \operatorname{Cov}(X_0, X_j)$ exists, is finite, and is nondegenerate along the directions relevant for the local statistic.*

Assumption 2. *(Window separation). Consider two windows $k \neq \ell$ whose separation satisfies $|\ell - k|/m \to \delta_{k\ell} \in (0, \infty)$. Define the limiting overlap fraction $\kappa_{k\ell} := \max\{1 - \delta_{k\ell}, 0\} \in [0, 1)$.*

Assumption 3. *(Local statistic). The window-wise statistic has the form $T_{k,m} = \tau_m\big(U_{k,m}, \widehat{\eta}_{k,m}\big)$, where nuisance estimator $\widehat{\eta}_{k,m} \xrightarrow{p} \eta_0$ under $\boldsymbol{H}_0$. There exists a measurable map $\tau(\cdot, \eta_0)$, continuous in its first argument, such that jointly for any pair (k, ℓ),*

$$\big(T_{k,m}, T_{\ell,m}\big) = \big(\tau(U_{k,m}, \eta_0), \tau(U_{\ell,m}, \eta_0)\big) + o_p(1).$$

Here $o_p(1)$ is with respect to $T \to \infty$ and $m \to \infty$ under $\boldsymbol{H}_0$. In addition, there exist constants $C > 0$, $q \geq 1$, and $M < \infty$ such that, for all $\|u\| \geq M$, $\tau(u, \eta_0) \leq C\big(1 + \|u\|^q\big)$.

Assumption 4. *(p-value mapping). Define the window-wise p-value $p_{k,m} := \pi_m(T_{k,m})$, where $\pi_m : \mathbb{R} \to (0, 1)$ is continuous and strictly decreasing on the relevant upper tail. Under $\boldsymbol{H}_0$, each $p_{k,m}$ is marginally valid, in the sense that for any $u \in (0, 1)$ and any fixed k, $\Pr(p_{k,m} \leq u) \leq u + o(1)$, where $o(1)$ is as $T \to \infty$*

and $m \to \infty$.

Assumption 1 yields a functional central limit theorem for the partial-sums process of $\{X_t\}$ (Herrndorf, 1985), which implies a Gaussian limit for window fluctuations $U_{k,m}$. Assumption 2 excludes asymptotic full overlap between distinct windows at the window-length scale. Assumption 3 links the extremes of the statistic $T_{k,m}$ to the extremes of $U_{k,m}$ through a continuous map with a polynomial growth control. Finally, Assumption 4 requires that the mapping π_m yields marginally valid window-wise p-values under $\boldsymbol{H}_0$.

Proposition 1 (Pairwise lower-tail independence)**.** *Under Assumptions 1–4, the rolling-window p-values $p_{k,m}$ and $p_{\ell,m}$ are asymptotically lower-tail independent in the sense that*

$$\lim_{u \downarrow 0} \limsup_{m \to \infty} \frac{\mathbb{P}(p_{k,m} \le u,\ p_{\ell,m} \le u)}{u} = 0.$$

Proposition 1 shows that overlap dependence is compatible with asymptotic lower-tail independence of the window-wise p-values, which follows as long as the limiting overlap fraction between two windows is strictly below one. The proof in the Supplementary Material proceeds by establishing a bivariate Gaussian limit for $(U_{k,m}, U_{\ell,m})$ with correlation strictly below one and then applying a Gaussian tail bound to translate this property into lower-tail independence for the induced p-values.

Proposition 1 provides the dependence condition required for HMP calibration, and Assumption 4 ensures marginal validity. The general results in Wilson (2019, 2020) then imply asymptotic FWER control.

Theorem 1. *Under $\boldsymbol{H}_0$ and Assumptions 1–4, the combined p-value p_H satisfies*

$$\mathbb{P}(p_H \le \alpha) \le \alpha + o(1)$$

for any fixed α as $T \to \infty$, $m = m_T \to \infty$ and $K = T - m + 1 \to \infty$. Therefore, the test that rejects $\boldsymbol{H}_0$ when $p_H \le \alpha$ controls the FWER at level α asymptotically.

Beyond the global test, HMP also supports a multilevel diagnostic rule for identifying windows with strong local evidence. For each window k, define the adjusted p-value

$$\tilde{p}_{k,m} = 1 - F_L\left[(K p_{k,m})^{-1}\right].$$

Under the same conditions, the rule that rejects $H_0^{(k)}$ when $\tilde{p}_{k,m} \le \alpha$ controls the FWER over the family $\{H_0^{(k)}\}_{k=1}^{K}$. The rule is conservative in the sense that rejecting $H_0^{(j)}$ implies rejecting the intersection null $\bigcap_{k \in J} H_0^{(k)}$ for any J with $j \in J$, including the global null $\boldsymbol{H}_0$. Hence, it provides only a sufficient condition for global rejection and may flag no windows even when $\boldsymbol{H}_0$ is rejected.

Table 1: Global rejection frequencies (size and power; $T = 200, m = 50, \alpha = 0.05$; bootstrap $B = 500$)

	Panel A. Size			**Panel B. Power**		
Test	HMP	Bonferroni	Bootstrap	HMP	Bonferroni	Bootstrap
Mean	0.056	0.018	0.054	0.803	0.529	0.793
Variance	0.054	0.012	0.058	0.542	0.333	0.532
Ljung–Box	0.051	0.020	0.058	0.661	0.520	0.641
ADF	0.061	0.032	0.060	0.483	0.415	0.464
Granger	0.055	0.016	0.049	0.738	0.348	0.551

3 Simulation study

We investigate the finite-sample size, power, and runtime of the proposed HMP global test in a rolling-window setting and compare it with standard benchmarks. We generate a time series of length $T = 200$ and apply rolling windows of length $m = 50$, resulting in $K = T - m + 1 = 151$ overlapping windows and a sequence of window-wise p-values.

We consider five representative local tests in empirical time-series analysis. Under the global null, the local null holds in every window. Under the alternative, the data depart from the null only over $t = 71, \ldots, 110$; let $I_t = 1$ on this interval and $I_t = 0$ otherwise. The DGPs are following:

1. *Mean* (two-sided t-test): $y_t \sim N(\mu_t, 1)$ with $\mu_t = 0$ under the null and $\mu_t = 0.6I_t$ under the alternative;

2. *Variance* (χ^2 test): $y_t \sim N(0, \sigma_t^2)$ with $\sigma_t^2 = 1$ under the null and $\sigma_t^2 = 0.8I_t + 1$ under the alternative;

3. *Serial Correlation* (Ljung-Box test): $y_t = \phi_t y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, 1)$ with $\phi_t = 0$ under the null and $\phi_t = 0.6I_t$ under the alternative;

4. *Unit root* (ADF test): $y_t = \phi_t y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, 1)$ with $\phi_t = 1$ under the null and $\phi_t = (1 - 0.4I_t)$ under the alternative;

5. *Granger causality* (F-test): $(x_t, y_t)'$ follows a first-order vector autoregression with i.i.d. Gaussian innovations. The coefficient on x_{t-1} in the y_t equation is zero under the null and $0.25I_t$ under the alternative.

We compare three global tests: (i) HMP; (ii) Bonferroni; and (iii) a parametric bootstrap calibration of the minimum rolling p-value with $B = 500$ draws. Results are based on 1,000 Monte Carlo replications at the nominal level $\alpha = 0.05$.

Table 1 reports the global rejection frequencies under the global null (size) and alternative (power). Under the null, both HMP and bootstrap are close to 0.05, while Bonferroni is conservative, as expected under strong overlap dependence. Under the alternative, HMP is substantially more powerful than Bonferroni and matches or exceeds the bootstrap across all designs; the advantage is especially pronounced in the Granger causality design.

Table 2: Computational time (in seconds; $T = 200, m = 50$; bootstrap $B = 500$).

Test	Baseline	HMP	Bootstrap
Mean	0.08	0.08	17.55
Variance	0.10	0.09	22.70
Ljung–Box	0.16	0.09	62.84
ADF	1.28	0.08	620.27
Granger	1.46	0.08	1,089.73

Notes: Baseline reports the time for rolling computation. HMP and Bootstrap report additional time beyond the baseline.

Table 3: Global p-values for directional Granger causality ($T = 276$, $m = 52$; bootstrap $B = 500$).

	IDEMV → returns	returns → IDEMV
HMP	0.0005	0.0170
Bootstrap	0.0380	0.0580

Table 2 reports the runtime of 1,000 Monte Carlo replications. The rolling computations take the same baseline time for both procedures. We then report the addtional time by HMP and the bootstrap, respectively. Relative to the baseline, the extra time for HMP is negligible, whereas the bootstrap dominates the runtime. The local test complexity has little effect on HMP, but it increases the bootstrap time sharply. For example, in the Granger causality design, HMP remains around 0.08 seconds, whereas the bootstrap runtime rises to 1,089.73 seconds. Overall, HMP maintains near-nominal size, delivers competitive power, and yields substantial computational savings relative to the bootstrap.

4 Empirical illustration

We illustrate our method by examining the time variation in Granger causality between weekly S&P 500 returns and the infectious disease equity market volatility index (IDEMV) of pandemic-related financial uncertainty (Baker et al., 2020). The sample runs from January 31, 2019 to May 11, 2024 (276 observations), covering the U.S. COVID-19 public health emergency (PHE) period and one year on either side.

We estimate a bivariate vector autoregression using 52-week rolling windows, yielding 225 windows. In each window, we run Granger-causality F-tests in both directions with the lag order selected by BIC from one to five and obtain the corresponding window-wise F-statistics and p-values. We then apply HMP to the window-wise p-values and, for comparison, implement the residual-based sup-F bootstrap of Shi et al. (2018, 2020) with 500 draws. At the 5% level, HMP rejects the global null in both directions, whereas the bootstrap rejects in only one direction (Table 3).

Figure 1 reports the window-level evidence under both methods. In the direction from IDEMV to returns, the HMP diagnostics flag more early-sample windows than the bootstrap, which is consistent with the sharp rise in IDEMV at the start of the pandemic. In the reverse direction, the HMP test rejects globally, but flags no individual windows. This is not a contradiction. The HMP window-level rule is conservative because it is

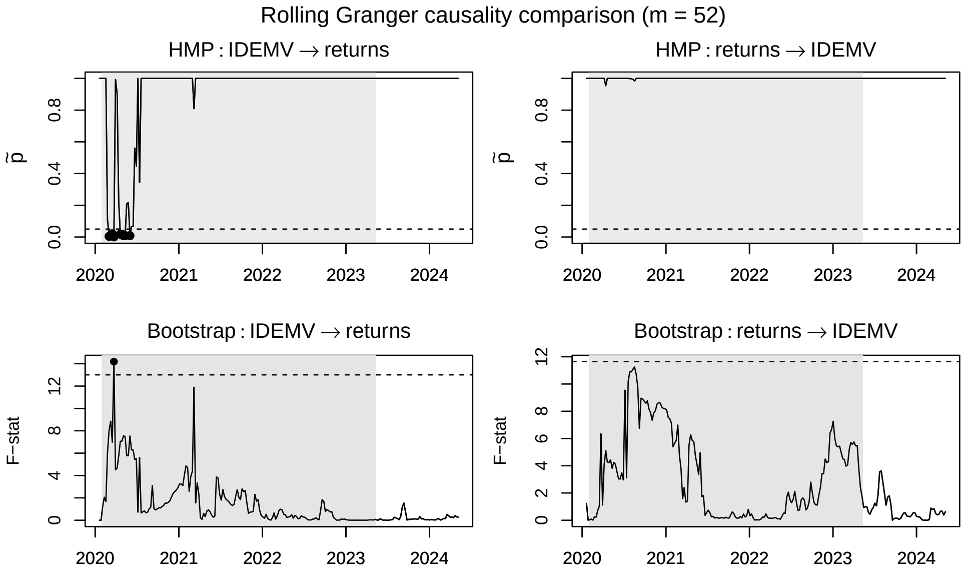

Figure 1: The top panels plot the multilevel HMP-adjusted p-values $\{\tilde{p}_{k,m}\}_{k=1}^{K}$ with $\alpha = 0.05$. The bottom panels plot rolling F-statistics with 5% bootstrap critical values. Exceedances are marked as window-level rejections. The x-axis denotes the window ending date; the PHE period is shaded.

only a sufficient condition for global rejection, and it becomes more conservative as the number of windows grows. Taken together, the global decision answers whether Granger causality is present in any window, while the diagnostics highlight where local evidence is strong.

5 Conclusion

We frame rolling-window tests as a multiple-testing problem and use HMP to construct an FWER-controlled global test and a multilevel diagnostic rule at minimal computational cost. Simulations show near-nominal size and competitive power for global inference relative to bootstrap calibration. The empirical illustration shows how the global decision and the multilevel diagnostic rule work together to summarize time variation. Overall, HMP provides a practical alternative to bootstrap-based extreme-statistic procedures in rolling-window applications. More broadly, the same idea extends to other subsample-based procedures, such as moving-window and recursive analyses, where many dependent tests are computed on overlapping data.

Appendix A. Supplementary Material

Supplementary material provides proofs for the theoretical results in the main text.

Data availability

Data are available on https://www.policyuncertainty.com/infectious_EMV.html. Code is available on https://qianjoewu.github.io/.

References

Andrews, D.W., 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 61, 821–856. doi:https://doi.org/10.2307/2951764.

Bai, J., Perron, P., 1998. Estimating and testing linear models with multiple structural changes. Econometrica 66, 47–78. doi:https://doi.org/10.2307/2998540.

Baker, S.R., Bloom, N., Davis, S.J., Kost, K., Sammon, M., Viratyosin, T., 2020. The unprecedented stock market reaction to COVID-19. Rev. Asset Pricing Stud. 10, 742–758. doi:https://doi.org/10.1093/rapstu/raaa008.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300. doi:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Cheng, X., Liao, Z., Schorfheide, F., 2016. Shrinkage estimation of high-dimensional factor models with structural instabilities. Rev. Econ. Stud. 83, 1511–1543. doi:https://doi.org/10.1093/restud/rdw005.

Diebold, F.X., Yilmaz, K., 2014. On the network topology of variance decompositions: Measuring the connectedness of financial firms. J. Econometrics 182, 119–134. doi:https://doi.org/10.1016/j.jeconom.2014.04.012.

Giacomini, R., Rossi, B., 2010. Forecast comparisons in unstable environments. J. Appl. Econometrics 25, 595–620. doi:https://doi.org/10.1002/jae.1177.

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578. doi:https://doi.org/10.1111/j.1468-0262.2006.00718.x.

Hanson, B.E., 2002. Tests for parameter instability in regressions with I(1) processes. J. Bus. Econ. Stat. 20, 45–59. doi:https://doi.org/10.1198/073500102753410381.

Harvey, C.R., Liu, Y., Zhu, H., 2016. ... and the cross-section of expected returns. Rev. Financ. Stud. 29, 5–68. doi:https://doi.org/10.1093/rfs/hhv059.

Herrndorf, N., 1985. A functional central limit theorem for strongly mixing sequences of random variables. Z. Wahrsch. Verw. Gebiete 69, 541–550. doi:https://doi.org/10.1007/BF00532665.

Inoue, A., Jin, L., Rossi, B., 2017. Rolling window selection for out-of-sample forecasting with time-varying parameters. J. Econometrics 196, 55–67. doi:https://doi.org/10.1016/j.jeconom.2016.03.006.

Romano, J.P., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73, 1237–1282. doi:https://doi.org/10.1111/j.1468-0262.2005.00615.x.

Shi, S., Hurn, S., Phillips, P.C., 2020. Causal change detection in possibly integrated systems: Revisiting the money–income relationship. J. Financ. Econometrics 18, 158–180. doi:https://doi.org/10.1093/jjfinec/nbz004.

Shi, S., Phillips, P.C., Hurn, S., 2018. Change detection and the causal impact of the yield curve. J. Time Ser. Anal. 39, 966–987. doi:https://doi.org/10.1111/jtsa.12427.

Stock, J.H., Watson, M.W., 1996. Evidence on structural instability in macroeconomic time series relations. J. Bus. Econ. Stat. 14, 11–30. doi:https://doi.org/10.1080/07350015.1996.10524626.

White, H., 2000. A reality check for data snooping. Econometrica 68, 1097–1126. doi:https://doi.org/10.1111/1468-0262.00152.

Wilson, D.J., 2019. The harmonic mean p-value for combining dependent tests. Proc. Natl. Acad. Sci. U.S.A. 116, 1195–1200. doi:https://doi.org/10.1073/pnas.1814092116.

Wilson, D.J., 2020. Generalized mean p-values for combining dependent tests: comparison of generalized central limit theorem and robust risk analysis. Wellcome Open Res. 5, 55. doi:https://doi.org/10.12688/wellcomeopenres.15761.1.

Yousuf, K., Ng, S., 2021. Boosting high dimensional predictive regressions with time varying parameters. J. Econometrics 224, 60–87. doi:https://doi.org/10.1016/j.jeconom.2020.08.003.