

Matched-Calibration of Changepoint Detectors for Biopharmaceutical Quality Monitoring

Halil İbrahim Özdemir

halilibrahim307@marun.edu.tr

Marmara University

Berfin Dogan

Istanbul Technical University

Pemra Ozbek

Marmara University

Article

Keywords: matched false-positive rate, changepoint detection, self-normalization, penalty calibration, process drift

Posted Date: August 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10498205/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Matched-Calibration of Changepoint Detectors for Biopharmaceutical Quality Monitoring

Halil İbrahim Özdemir^{1,2*}, Berfin Dogan^{2,3†} and Pemra Ozbek^{1†}

¹*Department of Bioengineering, Marmara University, İstanbul, 34854, Türkiye.

²R&D Department, AbdiBIO, Abdi İbrahim Pharmaceuticals, İstanbul, 34538, Türkiye.

³Department of Molecular Biology, Genetics & Biotechnology, Istanbul Technical University, İstanbul, 34469, Türkiye.

*Corresponding author(s). E-mail(s): halilibrahim307@marun.edu.tr;
Contributing authors: doganberfl6@itu.edu.tr;
pemra.ozbek@marmara.edu.tr;

[†]These authors contributed equally to this work.

Abstract

Monitoring a critical quality attribute (CQA) for process drift requires a changepoint detector. Two facts confound the choice: the record available at a decision point ranges from a few dozen measurements to over a thousand, and reported comparisons routinely place methods at different, uncontrolled false-positive rates. Non-Gaussian CQA data are meanwhile taken to call for a robust method, a recommendation never tested under controlled conditions. We compared a variance-based penalty, a robust penalty scaled by the median absolute deviation, and a permutation-calibrated test under two matched conditions: a common achieved false-positive rate and a common search-space ablation. They were evaluated across sample sizes from 25 to 10,000, under three noise distributions, and on an industrial fermentation dataset with documented regime changes. At default settings the false-positive rate of each formula-based penalty drifted with sample size, making raw comparison uninterpretable. Once calibrated to a common 5% rate, the variance-based penalty matched or exceeded the robust one at every sample size in all three distributions, reversing the usual expectation. The cause is a forfeited self-normalization, not a loss of efficiency: a robust denominator does not move with the outliers that inflate the numerator. Restricting the search to a recency window gave the largest power gain, but a matched ablation attributed it to the search restriction, not the scale estimator. Changepoint

detectors should therefore be compared only at a matched false-positive rate and under a matched ablation. Mild autocorrelation degraded every method far more than any noise-shape effect, and should be checked first.

Keywords: matched false-positive rate, changepoint detection, self-normalization, penalty calibration, process drift

1 Introduction

Monitoring critical quality attributes (CQAs) for process drift is a routine statistical task throughout a biologic product’s lifecycle, and the sample size it must work with is not fixed. Comparability exercises test a few dozen lots, whereas ongoing process verification accumulates hundreds to thousands of measurements over the commercial life of the product, an expectation made explicit by ICH Q2 [1], by the lifecycle view of ICH Q14 [2], and by Stage 3 of the FDA process validation framework [3], all elements of the quality-by-design paradigm [4]. Because a CQA history does not stop growing at validation, the number of measurements available at a decision point is whatever the lifecycle has supplied by then, which is why a detection method should be compared across the range a real CQA history spans rather than at one convenient value of n . Throughout we use “drift” in the applied sense of an abrupt departure from an established baseline, the alternative a changepoint method targets, rather than of a gradual ramp, which poses a different problem; the question also connects to multivariate statistical process control and batch-process monitoring [5–7], although we remain univariate.

Changepoint detection is a mature field, and none of the three detectors we compare is new, the problem dating to the earliest work on detecting a shift in a sequence of observations [8–10]. The offline literature is organized by cost function, search method and constraint on the number of changepoints [11, 12], spanning binary-segmentation and sparse-projection variants [13–15] and multiscale or econometric formulations of the same problem [16, 17], and our detectors occupy established positions on that map: a penalized-cost method with a BIC-type penalty [18–22], the moving-sum procedures that bound the search region [23], the outlier-resistant branch [24], and the distribution-free branch [25–28]. Even the self-normalization we invoke to explain our central result is an established device [29]. What this body of work does not settle is how these families should be compared with one another, since methods are typically demonstrated at a single sample size and reported at their nominal rather than their achieved false-positive rate, and penalty calibration remains among the least settled parts of the pipeline [11]. A comparison conducted on those terms cannot separate detection ability from calibration, and it is that gap this paper addresses.

Two features of the problem are easily overlooked when a method is validated at a single sample size. First, quality attribute measurements are frequently non-Gaussian, being heavy-tailed through assay and batch-pooling variability or skewed through multiplicative concentration effects, and non-normality is routinely taken to motivate the

robust and nonparametric alternatives the literature supplies [30–32], a recommendation that is precisely what we set out to test. Second, a method’s calibration, meaning whether it achieves its intended false-positive rate, is not guaranteed to be stable across the range of n a lifecycle spans, an instability well documented for estimated control-chart limits more generally [33–39], and where it is unstable an uncontrolled comparison measures the difference in achieved false-positive rates as much as any genuine difference in detection ability.

The contribution of this paper is therefore primarily methodological. We argue that changepoint detectors for CQA monitoring should be compared only under two matched conditions, a common achieved false-positive rate and a common search-space ablation, and we show that imposing this discipline changes the practical conclusion. To make the argument concrete we use three detectors spanning the choices a practitioner faces: a variance-based penalty, a representative robust penalty built as a fair stand-in for the “use a robust method” recommendation rather than as a new method of our own, and a distribution-free permutation test. Section 3 establishes that the achieved false-positive rate of each formula-based penalty drifts with sample size at default settings; that under a matched rate the variance-based penalty matches or exceeds the robust one across all sample sizes and distributions, a reversal we trace to self-normalization [29] rather than to any loss of efficiency; and that windowing, under a matched ablation, delivers its gain through the search restriction rather than the scale estimator. Section 4 draws out the implications, chief among them that autocorrelation is the dominant practical threat.

2 Materials and Methods

2.1 Problem Setup and Notation

Write $y_1, \dots, y_n$ for the CQA series under test and $y_{a:b}$ for the contiguous block from its a -th measurement to its b -th, with mean $\bar{y}_{a:b}$. We consider the single-changepoint problem, in which measurements before a changepoint τ fluctuate about a mean μ_0 and those from τ onward about $\mu_0 + \delta$, the noise having standard deviation σ common to both segments, which lets us quote shift sizes in units of σ ; Section 3.7 reports one real variable that violates this common-scale assumption and says so. A candidate changepoint is a split point t , cutting the series into $y_{1:t}$ and $y_{t+1:n}$, each required to hold at least h observations, so the admissible splits are $t \in \{h, \dots, n-h\}$ with $h = 3$ throughout (Supplementary Section S8 sweeps it). Table 1 collects the symbols used from here on.

Let $C(y_{a:b}) = \sum_{i=a}^b (y_i - \bar{y}_{a:b})^2$ be the cost of a block, its within-block sum of squared deviations from its own mean. The drop in total cost that splitting at t buys,

$$\Delta C(t) = \underbrace{C(y_{1:n})}_{\text{cost of describing the series with one mean}} - \underbrace{[C(y_{1:t}) + C(y_{t+1:n})]}_{\text{cost of describing it with two means}}, \quad (1)$$

is the evidence that a changepoint sits there, and is large exactly when two means fit the series much better than one. By the analysis-of-variance identity it has the closed

form

$$\Delta C(t) = \frac{t(n-t)}{n} (\bar{y}_{1:t} - \bar{y}_{t+1:n})^2, \quad (2)$$

the squared gap between the two block means weighted by the two block sizes; the weight $t(n-t)/n$ keeps $\Delta C(t)$ comparable across splits that divide the series very unevenly.

All three detectors then work identically, evaluating $\Delta C(t)$ at every admissible split, taking the best one, $t^* = \arg \max_t \Delta C(t)$, and declaring a changepoint there if $\Delta C(t^*) > \beta$ for a threshold β . They differ only in how β is obtained and, for the windowed variants of Section 2.3, in which range of t they may search; the cost function and the search machinery are otherwise the same, which is what allows a difference in performance to be attributed to the calibration strategy rather than to an incidental difference of implementation.

Table 1 Notation used throughout.

Symbol	Meaning
$y_1, \dots, y_n$	the CQA series; y_i is the i -th measurement, in order
n	number of measurements available at the decision point
$y_{a:b}$	the block of measurements $(y_a, \dots, y_b)$
$\bar{y}_{a:b}$	the mean of that block
τ	location of the true changepoint
μ_0	baseline mean, i.e. the level before τ
δ	size of the mean shift; $\delta = 0$ means no change
σ	noise standard deviation, assumed common before and after τ
t	a candidate split, cutting the series into $y_{1:t}$ and $y_{t+1:n}$
h	minimum segment length; every block must hold $\geq h$ points ($h = 3$)
$C(y_{a:b})$	cost of a block: its sum of squared deviations from its own mean
$\Delta C(t)$	evidence for a split at t : the drop in total cost it buys
t^*	the best split, $\arg \max_t \Delta C(t)$
β	detection threshold; a change is declared if $\Delta C(t^*) > \beta$
$\hat{\sigma}^2$	sample variance of the full series (used by Classical BIC)
$\hat{\sigma}_{\text{MAD}}^2$	robust, MAD-based variance estimate (used by ROBIN)
c	multiplier of the variance-based penalty
k, λ	multiplier and growth-rate mixing weight of ROBIN
$g(n)$	ROBIN's penalty growth rate, $\lambda \log n + (1 - \lambda)\sqrt{n}$
W	recency window: the windowed methods search only the last W points
B	number of permutations used by the permutation test
M	how many of those B permutations beat the observed statistic
$\hat{p}$	permutation p -value, $(1 + M)/(B + 1)$
α	nominal significance level (0.05 throughout)

2.2 Candidate Detection Methods

The first two detectors derive β from a formula, the third estimates it by resampling. Because $\Delta C(t^*)$ carries the units of a sum of squares, every threshold below is a multiple of an estimate of the noise variance times a term that grows with the number

of candidate splits searched: the more splits a method examines, the larger the largest $\Delta C(t)$ will be even when nothing has changed, so the bar must rise with n simply to hold the false-positive rate fixed.

A variance-based penalty (Classical BIC). The standard choice sets

$$\beta = c \hat{\sigma}^2 \log n, \quad (3)$$

where $\hat{\sigma}^2$ is the sample variance of the whole series and c sets how conservative the rule is. We use $c = 2$, the convention that charges a parameter for the changepoint location as well as for the post-change mean [18, 19], although because c is calibrated empirically nothing downstream depends on that choice. Note already that $\hat{\sigma}^2$ is inflated by the same extreme observations the penalty is meant to guard against; that coupling turns out to matter (Section 3.3).

A representative robust penalty (ROBIN). ROBIN is our label for the representative robust penalty defined here, and it is neither a previously published method nor one we propose. To represent the recommendation that non-Gaussian CQA data call for a robust penalty, we modify Eq. (3) in the two ways such a recommendation typically does, replacing the sample variance by the median absolute deviation $\hat{\sigma}_{\text{MAD}} = 1.4826 \times \text{median}_i |y_i - \text{median}(y)|$, which is built from medians and so barely moved by a few extreme observations, and replacing the growth term by a tunable mixture of the classical $\log n$ rate and a faster $\sqrt{n}$ rate,

$$\beta = k g(n) \hat{\sigma}_{\text{MAD}}^2, \quad g(n) = \lambda \log n + (1 - \lambda) \sqrt{n}, \quad \lambda \in [0, 1], \quad k > 0, \quad (4)$$

in which k plays the role c plays in Eq. (3). We fix $\lambda = 0.5$, with a sensitivity check in Supplementary Section S6; the $\sqrt{n}$ term is not derived from an optimality argument but is simply an available, more conservative candidate. Because k is calibrated empirically, our conclusions concern the class of MAD-scaled robust penalties calibrated to a common false-positive rate rather than one arbitrary member of it.

Permutation test. Rather than deriving β from a formula, this method estimates the null distribution of $\Delta C(t^*)$ directly from the series under test, on the reasoning that if there is no changepoint the order of the measurements carries no information, so shuffling them should produce an equally large best-split statistic. We permute $y_{1:n}$ at random B times; with M the number of permutations whose best-split statistic is at least as large as the observed $\Delta C(t^*)$, the Monte Carlo p -value is $\hat{p} = (1 + M)/(B + 1)$, and a changepoint is declared if $\hat{p} < \alpha$ ($\alpha = 0.05$ throughout). Because permutation destroys any true changepoint while preserving the marginal distribution, this calibration is exact under the null for any exchangeable noise distribution, the same logic that underlies circular binary segmentation for copy-number variant detection [40]. Exchangeability is nonetheless a meaningful restriction, since it holds for i.i.d. noise but is violated by autocorrelation, which permutation destroys along with any genuine changepoint; biopharmaceutical production series are plausibly autocorrelated through assay drift, batch pooling and correlated sequential runs, and Section 3.6 quantifies the damage.

We set $B = 99$ in the simulations, because false-positive rate and power are averaged over thousands of independent series per cell, so the granularity of any single

$\hat{p}$ washes out; that does not license $B = 99$ in practice, where a practitioner acts on one verdict from one series and needs B in the thousands, which costs seconds. The discreteness of $\hat{p}$ at $B = 99$ also makes the test’s exact level 0.04, so its achieved rate sits slightly below the nominal target.

2.3 Restricting the Search to a Recency Window

The CQA question is frequently not whether there is a changepoint somewhere in the full history but whether the attribute has drifted recently, within the last W measurements, and searching the full range pays an unnecessary price for that question, since the more candidate splits a method searches the higher its threshold must be to hold the false-positive rate. Penalizing according to the length of the search actually performed follows the segment-length-dependent BIC modification of Zhang and Siegmund [19], and moving-sum procedures already bound the search with their own size-control theory [23]. Our design differs in two ways: it is asymmetric, comparing a W -length recent window against a baseline mean and scale estimated from the entire history rather than against a matching window on the other side; and the window is anchored at the most recent observations rather than scanned across the series, since a scanning window would reintroduce the very multiple-looks penalty that windowing is meant to remove.

We therefore restrict candidate splits to $t \in \{\max(h, n - W + h), \dots, n - h\}$ and replace n by W in the growth rate, applying this to both penalties so their effects can be separated:

$$\text{windowed Classical BIC: } \beta_W = c \hat{\sigma}^2 \log W, \quad (5)$$

$$\text{windowed ROBIN: } \beta_W = k g(W) \hat{\sigma}_{\text{MAD}}^2. \quad (6)$$

These are Eqs. (3) and (4) with W substituted for n inside the growth term and nothing else, so the scale estimates are still computed from the full series and the cost $C(y_{1:t})$ still uses all t preceding observations. Only the number of splits searched and the rate at which the threshold grows are tied to W , which separates two roles a global- n penalty conflates: estimating the nuisance scale, for which more data always helps, and correcting for the number of hypotheses the search implicitly tests, which for a recency-restricted question should depend on W and not on how much unrelated baseline happens to exist. The trade-off is explicit: a windowed method can only detect a shift inside its window. Building the windowed variance-based penalty alongside the robust one is what lets us ask whether any power gain comes from the search restriction or from the scale estimator alongside which windowing is usually introduced. We fix $W = 20$ throughout: the window must exceed the horizon of the change to contain it, yet a much wider one pays an avoidable multiple-testing cost for splits that cannot hold the shift, so we set W just above the 15-observation shift the simulations inject, a choice the window-size sweep of Supplementary Section S4 confirms by placing peak power at the shift length.

A fixed window carries a further consequence: in an unrestricted search the null distribution of $\max_t \Delta C(t) / \hat{\sigma}^2$ grows on the order of $\log \log n$ under Darling–Erdős-type asymptotics [41], so a $\sqrt{n}$ penalty term grows far faster and at fixed k must

eventually dominate any bounded signal, whereas with W fixed both $\log W$ and $\sqrt{W}$ are constants whatever n may be.

2.4 Simulation Design

We evaluate all methods at thirteen sample sizes spanning $n = 25$ to $n = 1000$ at roughly even log-spacing, under three noise distributions standardized to zero mean and unit variance: Gaussian, Student’s t_3 (heavy-tailed), and log-normal with shape 0.5 (skewed). The log-normal noise is standardized before the additive shift δ is applied, so on the original scale a $+\delta\sigma$ shift acts on a different part of the right-skewed distribution than a $-\delta\sigma$ shift, an asymmetry we return to in Section 4.

The changepoint follows a fixed end-window design, in which the mean, where a shift is present, changes 15 observations before the end regardless of n ; this encodes the applied question of whether a fixed amount of recent history has drifted, and differs from a midpoint shift, which would let apparent power grow with n . Because binary segmentation can return more than one changepoint, we score any detection within 8 observations of the truth as correct, and we use a single-split search rather than an exact multiple-change algorithm such as PELT [20] because every comparison here concerns the at-most-one-change question a CQA decision poses. The false-positive rate is estimated from shift-free sequences and power from a 1σ shift, with $30 \text{ seeds} \times 400$ replications per cell unless noted; differences are assessed on the paired difference, and since the standard error of a cell mean is about 0.4 points, gaps above roughly 1 point are statistically resolved.

This design also fixes the shape of the power curves in advance. A true shift δ gives the standardized statistic a noncentral distribution whose noncentrality parameter is $n_1 n_2 \delta^2 / n$ in the two block sizes, and because $n_2 = 15$ is fixed while $n_1 = n - 15$ grows, that quantity tends to $15\delta^2$: the detectable signal is bounded, rising at $\delta = 1$ only from 6 at $n = 25$ to 14.8 at $n = 1000$, since the factor $n_1 n_2 / n$ is dominated by the smaller block and the smaller block never grows. That is why the unrestricted methods do not keep gaining power with n , their threshold rising against a bounded signal, and equally why the windowed methods keep improving, their threshold being fixed once $n > W$ so additional data only sharpens the scale estimate.

3 Results

3.1 Why Raw Comparison Fails: False-Positive Drift at Default Settings

Figure 1 shows the false-positive rate at a single fixed default configuration ($c = 2$; $\lambda = 0.5$, $k = 2$; $\alpha = 0.05$) across the full sample-size range. The permutation test stays close to its nominal target throughout, by construction, whereas the two formula-based penalties both drift substantially: the robust penalty ranges from near 0% to 33% depending on n and noise distribution, worst under heavy-tailed and log-normal noise at small n , while the variance-based penalty drifts less severely but is not stable either. This is the core reason a raw comparison is uninterpretable, since an apparent difference in power at a given n conflates a genuine difference in detection ability with

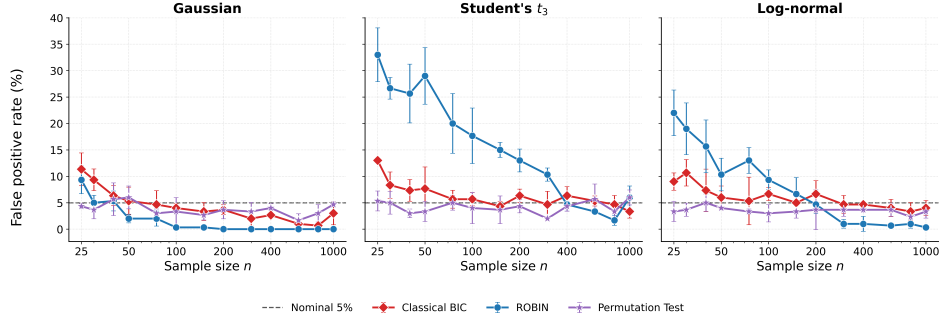


Fig. 1 False-positive rate versus sample size at default, uncalibrated settings, for three noise distributions. Only the permutation test stays close to the nominal 5% level (dashed line) across the full range.

a difference in how conservative each method happens to be there, and the sections that follow remove the confound by calibrating every method to a common rate.

3.2 Comparing Under a Matched False-Positive Rate

To compare detection ability on equal terms we calibrated c (Classical BIC) and k (ROBIN) to a common 5% false-positive rate in each (n, noise) cell, and only then measured power, using bisection with 2000 null replications per evaluation, which brings each method’s achieved rate within about half a point of 5%. That precision matters, because a coarse grid can leave the two mismatched, as an earlier version of this comparison did when it left ROBIN at 7.8% against 4.0% for Classical BIC, inflating ROBIN’s apparent power on its own.

Table 2 Power (% , mean $\pm$ SD across 30 seeds) at bisection-matched 5% false-positive rate (achieved rates within 0.5 points of target for both methods), Classical BIC versus ROBIN. The SE of each cell mean is approximately 0.4 points; bold marks gaps that are both statistically resolved and ≥ 3 points (a materially meaningful advantage).

Noise	$n = 50$		$n = 200$		$n = 1000$	
	BIC	ROBIN	BIC	ROBIN	BIC	ROBIN
Gaussian	65.5$\pm$2.2	55.9$\pm$2.4	73.0 $\pm$ 2.2	71.2 $\pm$ 2.2	72.1 $\pm$ 2.4	71.9 $\pm$ 2.7
Student’s t_3	67.9$\pm$3.0	16.3$\pm$1.5	69.4$\pm$2.3	56.4$\pm$2.2	64.7$\pm$1.9	58.0$\pm$2.0
Log-normal	60.6$\pm$2.3	24.0$\pm$2.2	65.6$\pm$2.4	55.0$\pm$2.7	62.8$\pm$2.7	59.8$\pm$2.7

The pattern reverses the usual argument for robust penalties, because with false-positive rates genuinely matched Classical BIC matches or exceeds ROBIN in every cell. Under heavy-tailed and log-normal noise the advantage is large at $n = 50$ (51.6 points under t_3 and 36.6 under log-normal) and shrinks with n , yet remains statistically clear at $n = 1000$ (6.7 and 3.0 points, both far exceeding the ≈ 0.6 -point SE of the difference), while under Gaussian noise it narrows to a practical tie. Convergence to

parity therefore occurs only under Gaussian noise, whereas under the non-Gaussian distributions the classical penalty retains a residual edge across the whole range: the opposite of what the raw comparison of Section 3.1 suggests, where ROBIN’s shifting false-positive rate obscures whether its power is really changing or merely tracking its own moving threshold. The supplementary checks use 15 seeds rather than 30 and a coarser calibration search, so every conclusion drawn from them rests on a within-section comparison rather than on an absolute figure set against Table 2.

3.3 The Mechanism: Self-Normalization, Not Efficiency

Both methods reduce to a self-normalized rule, rejecting if $\Delta C^*/\hat{\sigma}^2 > c \log n$ for Classical BIC or if $\Delta C^*/\hat{\sigma}_{\text{MAD}}^2 > k g(n)$ for ROBIN. Under heavy-tailed noise a single extreme observation can substantially inflate the numerator ΔC^* , and in Classical BIC that same observation also inflates the denominator $\hat{\sigma}^2$, so the ratio is self-normalizing, the established property [41, 42] that makes studentized statistics resistant to heavy tails without any explicit robustness design. What we add is specific to this setting, namely that the same self-normalization can be *lost* by substituting a robust denominator that no longer moves with the numerator, which is exactly what happens when a MAD-based scale replaces the sample variance. Because $\hat{\sigma}_{\text{MAD}}^2$ is by construction insensitive to that observation, the ratio inherits the heavy right tail of the numerator undiluted, and to hold the false-positive rate at 5% ROBIN’s threshold must rise high enough to exclude these outlier-driven excursions in the null; the real signal, of bounded noncentrality (Section 2.4), is often smaller than that inflated threshold, so power collapses precisely where the mismatch is worst. On the alternative side self-normalization works mildly against Classical BIC, since a genuine shift also inflates $\hat{\sigma}^2$ and so depresses its ratio while leaving ROBIN’s untouched, but that effect is second-order and the null-tail effect dominates it in every cell we measured.

This is more accurate than the familiar claim that MAD-based scaling trades away efficiency. Under t_3 noise the sample variance has infinite mean squared error as a scale estimator, whereas that of the MAD is always finite, so the MAD is if anything the more efficient estimator here, and the power loss comes not from a worse scale estimate but from a denominator that does not share the numerator’s sensitivity to the same extreme observations. The fix is accordingly not to seek a more efficient robust scale, the MAD already being one, but either to accept a self-normalizing design or to restrict the search so the numerator’s outlier-driven excursions are themselves bounded, which is what windowing does. The choice of MAD constant is immaterial here, since multiplying $\hat{\sigma}_{\text{MAD}}$ by a constant a and rescaling k by $1/a^2$ leaves ROBIN’s decision rule unchanged.

3.4 Not Knowing the Noise Distribution

Everything above assumes an oracle who knows the noise distribution and calibrates against it in every cell. Two results say what happens when that assumption fails. The first is what a wrong calibration costs. Calibrating both formula-based methods on Gaussian null data only, then applying those parameters to heavy-tailed and log-normal noise without recalibration, separates them sharply: Classical BIC holds its

false-positive rate under roughly 8% across all distributions and sample sizes, whereas ROBIN reaches 60% at $n = 1000$ under t_3 , a twelvefold inflation at which it would flag a nonexistent drift on the majority of in-control series (Supplementary Section S12). This is the self-normalization mismatch of Section 3.3 operating on the calibration rather than on the power, since the MAD denominator stays insensitive to the extreme observations that inflate the numerator under heavy tails, whereas the variance denominator is inflated in tandem, keeping the false-positive rate roughly calibrated even though calibration was performed on a Gaussian null.

The second is what distribution-freeness costs. The permutation test needs no distributional calibration at all, and its power cost is real but modest, and smallest exactly where it matters most: at $n = 50$ it comes within 2 to 6 points of Classical BIC and beats ROBIN by a very large margin under both non-Gaussian distributions (65.2% against 16.3% under t_3), so a practitioner facing a short, possibly heavy-tailed series gives up little by refusing to assume a distribution. The cost grows with n , and fastest where the tail is heaviest, the test trailing Classical BIC at $n = 1000$ by 2.5 points under Gaussian noise, 5.8 under log-normal and 9.8 under t_3 (Supplementary Section S13), an ordering that tracks tail weight because the permutation test thresholds the raw cost reduction rather than a studentized one and so does not inherit self-normalization. A practitioner who does not know the noise distribution and must calibrate once is therefore far safer with Classical BIC than with ROBIN, and safer still with the permutation test.

3.5 A Matched Ablation: The Search Restriction, Not the Scale Estimator, Drives the Gain

We calibrated the tunable constant of both windowed methods once per noise distribution, at $n = 200$ and window $W = 20$, and applied that calibration unchanged across the full range. Figure 2 compares both windowed variants against the unrestricted methods, each near but not exactly at a common 5% false-positive rate.

A single calibration keeps both variants usable across the whole range, since each stays within roughly 1% to 10% of the nominal false-positive rate without retuning and both substantially outperform their own unrestricted counterparts once $n \gtrsim 100$ to 200. This is the ablation result, and it does not depend on the scale estimator: restricting the search to a recency window, rather than choosing a robust scale, is what removes the false-positive drift of Section 3.1 and restores the lost power.

That 1–10% spread is loose enough for the ablation, which compares each windowed method against its own unrestricted counterpart, but not tight enough to rank the two windowed variants against each other, because wherever their achieved rates differ that comparison is confounded in exactly the way this paper argues against. We therefore recalibrated both to an exact 5% rate at every n (Table 3), and the result depends on sample size in a way the single calibration obscures. At small n the windowed variance-based penalty leads by a wide margin under all three distributions and not only the heavy-tailed one, which is the self-normalization mismatch of Section 3.3 at its most extreme: with little baseline outside the window, ROBIN’s scale estimate is both noisier and structurally unable to inflate alongside the extreme observations that

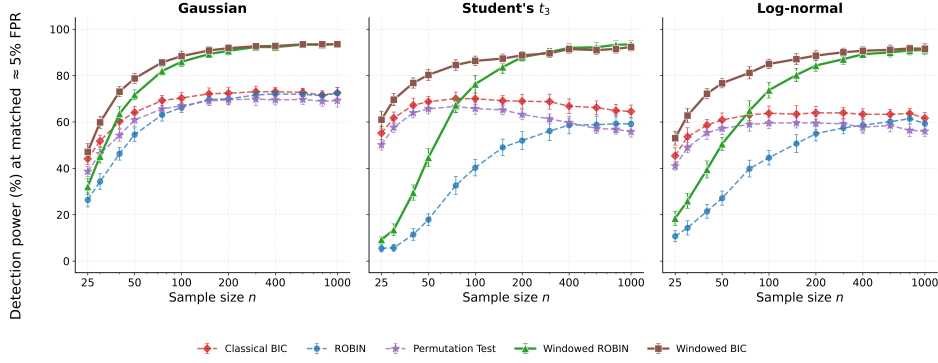


Fig. 2 Detection power versus sample size. Both windowed variants (bold lines) use a single calibration across the entire sample-size range and overtake every unrestricted method by $n \approx 100$ to 200 in all three noise distributions, which is the ablation result: the gain comes from the search restriction, and it appears for the variance-based penalty just as it does for the robust one. Each unrestricted method is calibrated to a matched 5% false-positive rate at every n ; the windowed variants, carrying one calibration across the range, sit near but not exactly at that level (roughly 1% to 10%), so the comparison between the two windowed variants should be read from Table 3, where both are recalibrated to an exact 5% at every n , rather than from the vertical gap between the two bold lines here.

inflate the numerator, leaving its threshold disproportionately conservative exactly where power is hardest to come by.

The gap then closes monotonically: it is within about a point by $n = 200$ under Gaussian and t_3 noise and gone by $n = 1000$, where windowed ROBIN even edges ahead under t_3 . Small as it is, that reversal is what the mechanism predicts: under t_3 the MAD is the more efficient scale estimator, and once the search is windowed and the baseline is long enough to pin it down, the excursions self-normalization was needed to absorb are themselves bounded and that efficiency advantage surfaces. The robust penalty is not intrinsically weaker, then, but weaker exactly where self-normalization is doing work, which is a statement about search design and baseline length rather than about robustness. Supplementary Section S8 confirms this from a different direction, since raising h from 3 to 8, which likewise denies extreme observations the very short segments through which they inflate ΔC^* , brings ROBIN level with Classical BIC under t_3 without any windowing.

The scale estimator therefore matters for the windowed search only when the baseline is short. We also verified windowing’s design commitment directly, by placing a 1σ shift at the series midpoint: both unrestricted methods retain high power there (78% to 95%), whereas both windowed methods score exactly 0%, a complete mechanical miss, so a user who cannot rule out a changepoint outside the assumed recency window should use an unrestricted method whatever Figure 2 shows for the end-window scenario.

3.6 The Cost of Autocorrelation

Every result so far assumes i.i.d. noise, whereas autocorrelation, plausible in production data through sequential runs, shared reagent lots and slow instrument drift, is a distinct and probably larger threat than noise shape, and is documented as a

Table 3 Windowed Classical BIC versus windowed ROBIN ($W = 20$), power (%) at an exact 5% false-positive rate recalibrated at every n (30 seeds $\times$ 400 replications; achieved rate 4.7% to 5.4% throughout). The advantage of the variance-based penalty is large at small n in every noise distribution and vanishes by $n = 1000$.

Noise	Method	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 1000$
Gaussian	windowed BIC	48.4	78.9	88.8	91.2	93.5
	windowed ROBIN	32.1	71.2	85.8	90.0	93.5
	difference	+16.3	+7.7	+3.0	+1.3	+0.1
Student's t_3	windowed BIC	60.7	80.8	86.4	88.9	91.9
	windowed ROBIN	9.3	43.4	75.5	87.5	93.5
	difference	+51.4	+37.4	+10.9	+1.4	-1.6
Log-normal	windowed BIC	52.9	76.3	84.5	88.1	91.5
	windowed ROBIN	17.8	48.1	73.4	83.9	90.6
	difference	+35.1	+28.1	+11.1	+4.2	+0.9

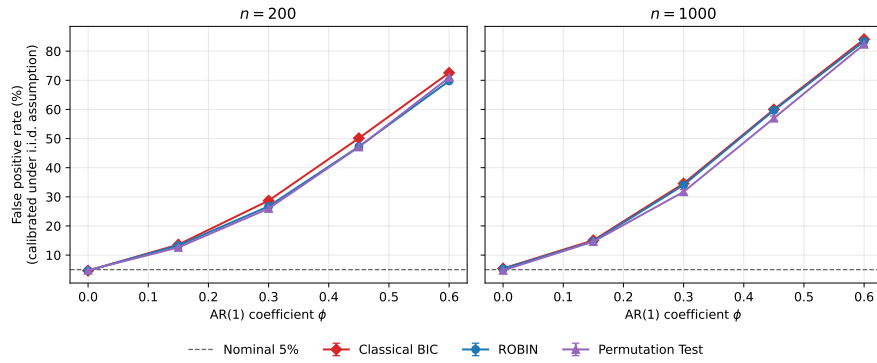


Fig. 3 False-positive rate versus AR(1) coefficient ϕ , for all three methods calibrated under the i.i.d. assumption ($\phi = 0$) and left unchanged. Even mild autocorrelation inflates the false-positive rate far more than any noise-distribution effect studied elsewhere in this paper.

recurring feature of biopharmaceutical process-monitoring series [43]. We calibrated all three methods under the i.i.d. assumption, then measured the actual false-positive rate when that calibration is left unchanged but the true series is AR(1) with $\phi \in \{0, 0.15, 0.3, 0.45, 0.6\}$, the situation of a practitioner who calibrates once without verifying the assumption. The direction of the result agrees with a recent preprint in a different monitoring domain, in which Li [44] reports that moderate autocorrelation costs BIC-based binary segmentation about 40 percentage points of power.

The degradation is severe and immediate (Figure 3). At $n = 200$ the false-positive rate rises from a correctly calibrated 3–5% at $\phi = 0$ to 26–29% at $\phi = 0.3$ and 70–73% at $\phi = 0.6$ for all three methods, and at $n = 1000$ it is the same or worse, with no method meaningfully more robust than the others; the distribution-free exactness of the permutation test, which held across every noise shape, gives no protection

whatever here, as its exchangeability requirement predicts. For scale, the largest false-positive inflation attributable to noise shape anywhere in this paper is roughly six- to sevenfold, which $\phi = 0.3$ alone approaches and $\phi = 0.6$ more than doubles. Because the formula-based methods were calibrated while ignoring autocorrelation, this measures the cost of an unverified i.i.d. assumption rather than any claim that they could not be recalibrated for a known ϕ .

That failure is remediable, and we show it. Building an estimated ϕ into the calibration, either by prewhitening each series through a lag-1 AR(1) fit or by recalibrating the threshold on AR(1) surrogates, returns the false-positive rate of Classical BIC to within a point or two of 5% at every ϕ we tested, and the permutation test, broken on the raw series, is likewise repaired when run on prewhitened residuals (Supplementary Section S14). This is a proof of concept at representative n rather than a full re-validation, and the formula-based fix is only as good as the $\hat{\phi}$ estimated from a possibly short in-control record, so autocorrelation should still be checked, and corrected, before any finding here is applied to a real series.

Eight further analyses in the Supplementary Information (Sections S1–S8) extend the effect to $n = 10,000$, generalize it to milder heavy-tailed models, and test sensitivity to the robust loss, window size, effect size, penalty form, scoring rule and minimum segment length; none changes a conclusion drawn here.

3.7 Case Study: Comprehensive IndPenSim Bioprocess Data

To ask whether these trade-offs surface on a real process, we use 100 consecutive batches of the IndPenSim V3 dataset, a first-principles simulation of an industrial-scale penicillin fermentation [45, 46], taking the six laboratory assays it records at batch end. The dataset supplies its own ground truth, since it runs batches 1–30 on a fixed recipe, 31–60 under operator control, 61–90 under Raman-based advanced process control (APC) and 91–100 as deliberate fault batches, its fault-reference channel flagging exactly batches 91–100 and so documenting a changepoint at $t = 90$. Every series is standardized and every formula-based detector calibrated to an exact 5% rate on a Gaussian null at $n = 100$.

Every detector fires on every variable (Table 4, Figure 4), so with regime changes this large the test is not one of power but of localization: four variables break at the documented boundary $t = 90$, viscosity breaks at $t = 97$, and NH_3 breaks at $t = 60$, which is not an error but the documented operator-to-APC boundary, an event of its own and a larger one. NH_3 is therefore where the windowing trade-off surfaces, because the windowed detectors may search only batches 81–100 and so cannot see the batch-60 event, returning batch 90 instead. That is not a failure but the design commitment of Section 2.3, a windowed detector answering the narrower recency question it was built for; the choice between them is a choice of question rather than of quality.

The six panels of Figure 4 are chosen rather than exhaustive, on the criterion that near-copies must not be presented as independent witnesses: penicillin is recorded on two channels correlating at $|r| = 1.00$, and the four gas-exchange channels carry a participation ratio of 1.16, so we plot one representative of each group.

Table 4 also gives each detector’s margin above threshold, and ROBIN’s exceeds Classical BIC’s on all six variables. That is not evidence it would catch a shift Classical

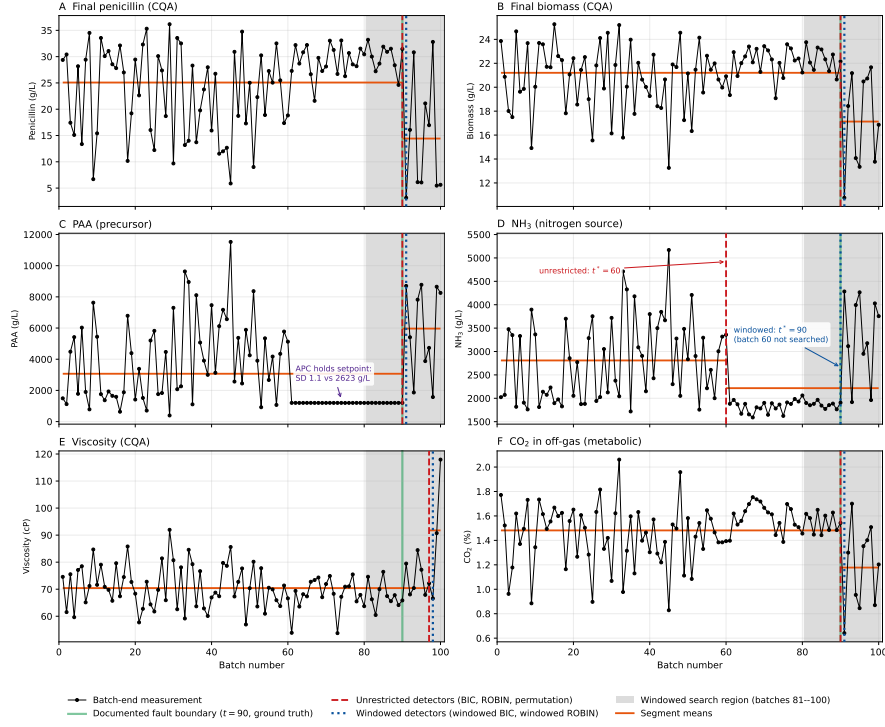


Fig. 4 Detection on the IndPenSim V3 dataset [45, 46], 100 chronological batches. Green: the documented fault boundary at batch 90 (ground truth, from the dataset’s own fault-reference channel). Red dashed: the changepoint returned by the unrestricted detectors, which agree on the location in every panel. Blue dotted: the changepoint returned by the windowed detectors, drawn just beside the red line in every panel where the two coincide so that its firing is visible; only in panel D do the two localize differently. Orange: the resulting segment means. Gray band: the only region the windowed detectors may search (batches 81–100). Panel D carries the point: NH_3 really breaks at batch 60, outside the windowed search region, so the windowed detectors return batch 90 instead. Panels are chosen from all 2,224 measurement channels to be non-redundant (see text); every remaining channel is plotted in the supplementary material.

BIC would miss, but the same self-normalization asymmetry seen from the alternative side: a large step inflates the numerator for every method, yet it also inflates the variance denominator that Classical BIC uses and the MAD denominator resists, and ranking the variables by $\hat{\sigma}/\hat{\sigma}_{\text{MAD}}$ reproduces their ranking by ROBIN’s margin exactly (Supplementary Section S11). Two variables also strain our assumptions: under APC the controller pins PAA to a tight setpoint (SD 1.1 g/L against a baseline near 2600), so its batch-90 event is a change of scale as much as of mean; and viscosity’s changepoint at $t^* = 97$ leaves only three post-change batches, the least secure localization of the six.

Repeating the comparison on AR(1) surrogates parameterized from each variable’s own in-control stretch reproduces the simulation findings on plant-derived rather than synthetic noise: Classical BIC matches or exceeds ROBIN on all six variables and windowing gains 19 to 23 points everywhere, but autocorrelation dominates both

Table 4 All five detectors on all six IndPenSim offline variables ($n = 100$ batches). Every formula-based method is calibrated to an exact 5% false-positive rate on a Gaussian null at $n = 100$ (thresholds: Classical BIC 8.59, ROBIN 9.63, windowed BIC 6.60, windowed ROBIN 6.95); the permutation test self-calibrates and is reported as a p -value ($B = 9999$; see text on the choice of B). Every detector fires on every variable, so the informative columns are the localization t^* and, descriptively, the margin above threshold. $\text{SNR} = (\text{test statistic})/(\text{threshold})$ is a descriptive margin measured on a series that already contains the shift; it is not a power comparison, and it is inflated for the MAD-scaled penalties by construction (see text). The windowed search covers batches 81–100 only, which is why it cannot return the NH_3 event at $t = 60$.

Variable	Unrestricted (SNR)			Windowed, $W = 20$ (SNR)			Permutation
	t^*	BIC	ROBIN	t^*	BIC	ROBIN	p
Penicillin (process)	90	1.60	1.99	90	2.09	2.75	0.006
Penicillin (assay)	90	1.61	2.02	90	2.09	2.80	0.005
Biomass	90	2.13	3.23	90	2.77	4.48	0.002
PAA (precursor)	90	1.22	6.20	90	1.59	8.60	0.028
NH_3	60	1.26	4.31	90	1.30	4.71	0.024
Viscosity	97	1.92	2.78	97	2.50	3.85	0.008

effects, since PAA carries the largest lag-1 coefficient ($\hat{\phi} = 0.293$) and is the worst cell for every calibrated method at 40.9%, against 61–81% where $\hat{\phi} \approx 0$, so one real CQA with mild positive autocorrelation costs more power than any distributional effect we studied (Table S5). The permutation test cannot be calibrated on these nulls at all, since permutation destroys the autocorrelation that defines them, so its achieved rate tracks $\hat{\phi}$ instead of sitting at 5%, reaching 24.0% on PAA: the cell where it appears to beat Classical BIC most convincingly is exactly the cell where it runs at that rate. A sweep of all 2,224 channels the dataset records, and the screening lessons it forced on us, are reported in Supplementary Section S10.

4 Discussion

The primary message of this paper is methodological. Comparisons of changepoint detectors frequently place methods at different achieved false-positive rates, the changepoint-detection analogue of a problem long documented for estimated control-chart limits [33–35], and credit a modification to its most visible component without an ablation; both practices can invert the apparent conclusion. Here the robust penalty appeared to lose power at large n in the raw comparison, when what was really falling was its false-positive rate, and windowing appeared to be a benefit of the robust construction, when the matched ablation showed the search restriction did the work. We therefore advocate two requirements: that the achieved false-positive rate be matched by calibration rather than by nominal setting, and that any search-space modification be matched across every method it could apply to.

The practical choice turns on what is known about the timing of a potential drift and on how much baseline exists. If a drift would be recent relative to the full history, restricting the search to that window is worth doing whichever penalty is used,

since both windowed variants substantially outperform their unrestricted counterparts once $n \gtrsim 100$ to 200; between the two, the variance-based one leads at small n and by $n = 1000$ they are interchangeable, so because the short-baseline case leaves least room for error, we would window a variance-based penalty first and add a robust scale estimate only if extreme outliers were a concrete concern. Where recency cannot be assumed an unrestricted method is required, and contrary to the usual argument the variance-based penalty then matches or exceeds the MAD-scaled robust one at matched false-positive rate at every sample size in all three distributions, with its largest advantages under exactly the heavy-tailed and log-normal noise a robust penalty targets. We qualify that finding to the class of MAD-scaled, unbounded-loss robust penalties ROBIN represents, since the bounded-loss biweight family that Fearnhead and Rigai’s own theory prefers was not tested and could narrow the gap.

Windowing is nonetheless a search restriction and not a replacement for unrestricted search, so a real changepoint outside the window will not be found however large it is, and W is not a free parameter either: it should be set close to the shortest relevant recency horizon rather than padded wide, since Supplementary Section S4 shows an asymmetric, size-dependent cost to over-widening.

The ablation generalizes into a rule: a modification must be tested against the same modification applied to the simplest baseline it could apply to, before the gain is credited to the more elaborate component alongside which it was introduced. Reporting only the improvement of the windowed robust penalty over its unrestricted counterpart would have implied that the fix required robustness, whereas the parallel windowed variance-based penalty revealed windowing to be doing essentially all of the work. The same holds for calibration, since comparisons that report raw power and raw false-positive rate side by side [24, 44] can mislead whenever the achieved rates differ. We include our own work in that stricture: our first pass at the windowed comparison carried a single calibration across the whole range, leaving the two variants at achieved rates as far apart as 1% and 10%, on which basis the variance-based penalty appeared to beat the robust one at every sample size, whereas recalibrating both to an exact 5% shrank the gap to nothing beyond $n \approx 400$, reversed its sign under heavy tails, and revealed a small- n advantage the loose calibration had hidden. A matched false-positive rate is accordingly not a standard to hold others to but a check to run on oneself.

Three limitations remain. First, autocorrelation rather than distributional shape is the single largest threat to everything reported here: classical sequential monitoring schemes developed for exactly this kind of drift [47, 48] have long been known to lose their nominal properties once the underlying series is autocorrelated [49–51], and validating these methods against an autocorrelated design is a higher priority than any other extension. Second, the self-normalization result held under two milder heavy-tailed alternatives, t_5 and a contaminated normal (Supplementary Section S2), as the mechanism predicts, but every check still assumes i.i.d. noise, a single changepoint, and the specific effect size and window studied. Third, because our log-normal model applies an additive shift to standardized right-skewed noise, power is not symmetric in the sign of the shift, so on right-skewed data practitioners should not assume symmetry

and should consider separate thresholds for upward and downward drift, or a variance-stabilising transform before detection.

5 Conclusion

The central lesson of this study is methodological: changepoint detectors for CQA monitoring should be compared only under a matched achieved false-positive rate and a matched search-space ablation, because either control alone can reverse the apparent conclusion. Applying that discipline across sample sizes from 25 to 1000 and three noise distributions, we found that the default calibration of neither formula-based penalty holds across the range, which makes raw comparison uninterpretable, and that once calibrated to a common rate a variance-based penalty matches or exceeds a representative robust MAD-scaled penalty at every sample size in all three distributions, and not only in the heavy-tailed case for which robust penalties are usually recommended. The cause is a forfeited self-normalization rather than any loss of efficiency, since under heavy tails the MAD is the more efficient scale estimator. Restricting the search to a recency window removes the false-positive drift and gives the best power of every method compared once enough baseline is available, at the explicit cost of being unable to detect a shift outside that window, and the matched ablation attributes that benefit to the search restriction rather than the scale estimator. Finally, even mild autocorrelation degrades the false-positive rate of every method far more than any noise-distribution effect studied here, with none more robust than the others, which makes it the single largest threat to practical application and the first thing to check before any method is deployed. We provide the simulation code, calibrated configurations and decision guidance needed to apply and extend this comparison.

Supplementary information.

- **Reproduction notebook** (ESM_1, .ipynb). A single self-contained Jupyter notebook that regenerates every table and figure of this paper from the raw inputs, in the order they appear, with the detectors defined from scratch. No reported value is hard-coded; its final section re-derives each of the paper’s claims and asserts that it holds, so a reader can audit the reproduction by running the notebook and reading one table. A flag switches between the paper’s replication counts and a faster smoke run over the identical code path.
- **All-channels figure** (ESM_2, .pdf). Every one of the 2,224 IndPenSim measurement channels not shown in Figure 4, plotted and labelled with the reason it was held back from the main figure (near-duplicate of a shown channel, member of the collinear gas-exchange group, a design variable rather than a quality attribute, or degenerate), with the per-class firing and localization summary that supports the case study’s channel-count caveat.
- **Supplementary Results** (ESM_3). Fourteen sensitivity, robustness and case-study analyses referenced from the main text (Supplementary Sections S1–S14, Tables S1–S8, Figure S1): persistence of the effect to $n = 10,000$; generalization to t_5 and contaminated-normal noise; a check against an established robust loss; the window-size sweep; sensitivity to effect size, penalty form, scoring rule and minimum segment length; power on AR(1) noise parameterized from the real in-control data; the

plant-wide channel sweep with the two screening lessons it forced; the SNR-ordering mechanism check on real data; cross-calibration under an unknown noise distribution; the permutation test’s power cost; and a prewhitening proof of concept that restores calibration under autocorrelation.

- **Code and data.** A public repository providing the five detectors as an importable module, every analysis and plotting script (including a standalone, seeded regeneration of the autocorrelation results and figure), the derived case-study data, and the reproduction notebook above. Every table and figure in this paper is regenerated from it; available at <https://github.com/bio-otto/changepoint-robustness-2026>.

Acknowledgements. No funding was received for conducting this study.

Declarations

- **Funding.** This research received no external funding.
- **Conflict of interest.** H.İ.Ö. and B.D. are employed by Abdi İbrahim Pharmaceuticals (AbdiBIO R&D). This study uses only simulated data and reports no results specific to any Abdi İbrahim product.
- **Ethics approval.** Not applicable.
- **Consent for publication.** Not applicable.
- **Data availability.** All results in this paper are produced from simulated data; the simulation and analysis code needed to regenerate every table and figure is provided as supplementary material.
- **Materials availability.** Not applicable.
- **Code availability.** The simulation and analysis code is provided as supplementary material and is publicly available at <https://github.com/bio-otto/changepoint-robustness-2026>.
- **Author contribution.** H.İ.Ö. conceived the study, developed the methodology, implemented the simulations, and wrote the original draft. B.D. contributed to the methodology and analysis, and reviewed and edited the manuscript. P.O. supervised the work, contributed to the interpretation of results, and reviewed and edited the manuscript.

References

- [1] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH): ICH harmonised guideline Q2(R2): Validation of analytical procedures. Technical report, ICH (2023)
- [2] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH): ICH harmonised guideline Q14: Analytical procedure development. Technical report, ICH (2023)
- [3] Food, U.S., Administration, D.: Guidance for industry. process validation: General principles and practices. Technical report, U.S. Department of Health and Human Services, FDA (2011)

- [4] Yu, L.X., Amidon, G., Khan, M.A., Hoag, S.W., Polli, J., Raju, G.K., Woodcock, J.: Understanding pharmaceutical quality by design. *The AAPS Journal* **16**(4), 771–783 (2014)
- [5] Montgomery, D.C.: *Introduction to Statistical Quality Control*, 8th edn. Wiley, ??? (2020)
- [6] Kourti, T., MacGregor, J.F.: Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems* **28**(1), 3–21 (1995)
- [7] Nomikos, P., MacGregor, J.F.: Monitoring batch processes using multiway principal component analysis. *AIChE Journal* **40**(8), 1361–1375 (1994)
- [8] Page, E.S.: Continuous inspection schemes. *Biometrika* **41**(1/2), 100–115 (1954)
- [9] Hinkley, D.V.: Inference about the change-point in a sequence of random variables. *Biometrika* **57**(1), 1–17 (1970)
- [10] Sen, A., Srivastava, M.S.: On tests for detecting change in mean. *The Annals of Statistics* **3**(1), 98–108 (1975)
- [11] Truong, C., Oudre, L., Vayatis, N.: Selective review of offline change point detection methods. *Signal Processing* **167**, 107299 (2020)
- [12] Aminikhanghahi, S., Cook, D.J.: A survey of methods for time series change point detection. *Knowledge and Information Systems* **51**(2), 339–367 (2017)
- [13] Fryzlewicz, P.: Wild binary segmentation for multiple change-point detection. *The Annals of Statistics* **42**(6), 2243–2281 (2014)
- [14] Cho, H., Fryzlewicz, P.: Multiple-change-point detection for high-dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B* **77**(2), 475–507 (2015)
- [15] Wang, T., Samworth, R.J.: High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B* **80**(1), 57–83 (2018)
- [16] Frick, K., Munk, A., Sieling, H.: Multiscale change point inference. *Journal of the Royal Statistical Society: Series B* **76**(3), 495–580 (2014)
- [17] Bai, J., Perron, P.: Estimating and testing linear models with multiple structural changes. *Econometrica* **66**(1), 47–78 (1998)
- [18] Schwarz, G.: Estimating the dimension of a model. *Annals of Statistics* **6**(2), 461–464 (1978)

- [19] Zhang, N.R., Siegmund, D.O.: A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63**(1), 22–32 (2007)
- [20] Killick, R., Fearnhead, P., Eckley, I.A.: Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* **107**(500), 1590–1598 (2012)
- [21] Yao, Y.-C.: Estimating the number of change-points via Schwarz’ criterion. *Statistics & Probability Letters* **6**(3), 181–189 (1988)
- [22] Lavielle, M.: Using penalized contrasts for the change-point problem. *Signal Processing* **85**(8), 1501–1510 (2005)
- [23] Eichinger, B., Kirch, C.: A MOSUM procedure for the estimation of multiple random change points. *Bernoulli* **24**(1), 526–564 (2018)
- [24] Fearnhead, P., Rigai, G.: Changepoint detection in the presence of outliers. *Journal of the American Statistical Association* **114**(525), 169–183 (2019)
- [25] Zou, C., Yin, G., Feng, L., Wang, Z.: Nonparametric maximum likelihood approach to multiple change-point problems. *Annals of Statistics* **42**(3), 970–1002 (2014)
- [26] Haynes, K., Fearnhead, P., Eckley, I.A.: A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing* **27**(5), 1293–1305 (2017)
- [27] Good, P.I.: *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, 3rd edn. Springer, New York (2005)
- [28] Hemerik, J., Goeman, J.: Exact testing with random permutations. *TEST* **27**(4), 811–825 (2018)
- [29] Shao, X., Zhang, X.: Testing for change points in time series. *Journal of the American Statistical Association* **105**(491), 1228–1240 (2010)
- [30] Hampel, F.R.: The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**(346), 383–393 (1974)
- [31] Rousseeuw, P.J., Croux, C.: Alternatives to the median absolute deviation. *Journal of the American Statistical Association* **88**(424), 1273–1283 (1993)
- [32] Maronna, R.A., Martin, R.D., Yohai, V.J.: *Robust Statistics: Theory and Methods*. Wiley, Hoboken, NJ (2006)
- [33] Jensen, W.A., Jones-Farmer, L.A., Champ, C.W., Woodall, W.H.: Effects of parameter estimation on control chart properties: A literature review. *Journal of*

Quality Technology **38**(4), 349–364 (2006)

- [34] Psarakis, S., Vyniou, A.K., Castagliola, P.: Some recent developments on the effects of parameter estimation on control charts. *Quality and Reliability Engineering International* **30**(8), 1113–1129 (2014)
- [35] Quesenberry, C.P.: The effect of sample size on estimated limits for $\bar{X}$ and x control charts. *Journal of Quality Technology* **25**(4), 237–247 (1993)
- [36] Jones, L.A., Champ, C.W., Rigdon, S.E.: The performance of exponentially weighted moving average charts with estimated parameters. *Technometrics* **43**(2), 156–167 (2001)
- [37] Jones, L.A.: The statistical design of EWMA control charts with estimated parameters. *Journal of Quality Technology* **34**(3), 277–288 (2002)
- [38] Jones, L.A., Champ, C.W., Rigdon, S.E.: The run length distribution of the CUSUM with estimated parameters. *Journal of Quality Technology* **36**(1), 95–108 (2004)
- [39] Jones-Farmer, L.A., Ezell, J.D., Hazen, B.T.: Applying control chart methods to enhance data quality. *Technometrics* **56**(1), 29–41 (2014)
- [40] Olshen, A.B., Venkatraman, E.S., Lucito, R., Wigler, M.: Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**(4), 557–572 (2004)
- [41] Csörgő, M., Szyszkowicz, B., Wang, Q.: Darling–Erdős theorem for self-normalized sums. *The Annals of Probability* **31**(2), 676–692 (2003)
- [42] Peña, V.H., Lai, T.L., Shao, Q.-M.: *Self-Normalized Processes: Limit Theory and Statistical Applications*. Springer, Berlin, Heidelberg (2009)
- [43] Bower, K.M.: Determining Shewhart control chart limits in biomanufacturing for continued process verification with autocorrelated data. *BioProcess International* **17**(4) (2019)
- [44] Li, A.A.: How short is too short? Power analysis for BIC-based changepoint detection in ecological monitoring. arXiv preprint arXiv:2603.21154 (2026)
- [45] Goldrick, S., Stefan, A., Lovett, D., Montague, G., Lennox, B.: The development of an industrial-scale fed-batch fermentation simulation. *Journal of Biotechnology* **193**, 70–82 (2015)
- [46] Goldrick, S., Duran-Villalobos, C.A., Jankauskas, K., Lovett, D., Farid, S.S., Lennox, B.: Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process. *Computers & Chemical Engineering* **130**, 106471 (2019) <https://doi.org/10.1016/j.compchemeng.2019.05.037>

- [47] Roberts, S.W.: Control chart tests based on geometric moving averages. *Technometrics* **1**(3), 239–250 (1959)
- [48] Lucas, J.M., Saccucci, M.S.: Exponentially weighted moving average control schemes: Properties and enhancements. *Technometrics* **32**(1), 1–12 (1990)
- [49] Alwan, L.C., Roberts, H.V.: Time-series modeling for statistical process control. *Journal of Business & Economic Statistics* **6**(1), 87–95 (1988)
- [50] Montgomery, D.C., Mastrangelo, C.M.: Some statistical process control methods for autocorrelated data. *Journal of Quality Technology* **23**(3), 179–193 (1991)
- [51] Woodall, W.H.: Controversies and contradictions in statistical process control. *Journal of Quality Technology* **32**(4), 341–350 (2000)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM1reproductionnotebook.ipynb](#)
- [ESM3supplementaryresults.tex](#)
- [ESM2allchannels.pdf](#)