Nonparametric Threshold Estimation for Autocorrelated Monitoring Statistics

Taylor Grimm[1], Kathryn Newhart[2], and Amanda Hering[1]

August 7, 2024

This work is supported by the National Alliance for Water Innovation (NAWI), funded by the US Department of Energy (DOE), Energy Efficiency and Renewable Energy Office, Advanced Manufacturing Office under Funding Opportunity Announcement DE-FOA-0001905, Project 5.17 Data-Driven Fault Detection and Process Control for Potable Reuse with Reverse Osmosis.

[1] Department of Statistical Science, Baylor University
[2] Department of Geography and Environmental Engineering, United States Military Academy

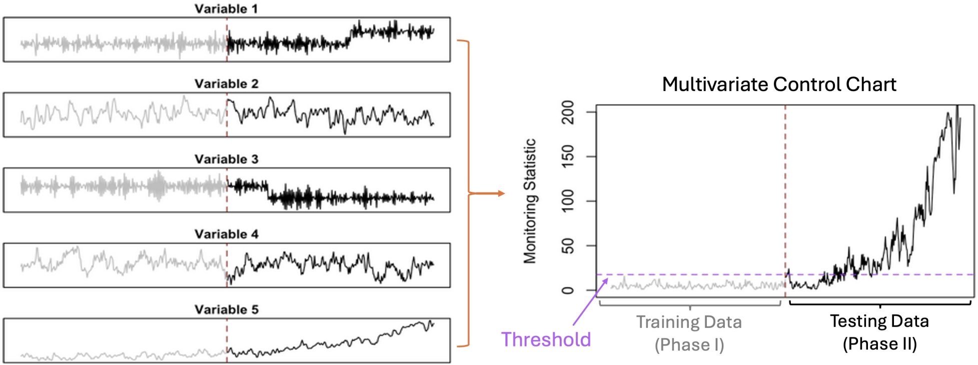
Presentation Outline

1. Multivariate Statistical Process Monitoring (MSPM)
2. Parametric and Nonparametric Thresholds
3. Simulation Study
4. Case Study: Wastewater Treatment

Multivariate Statistical Process Monitoring (MSPM)

MSPM allows us to

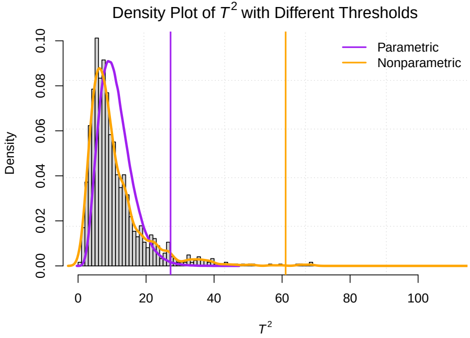
- monitor a multivariate process for abnormal behavior in real time
- condense information from many variables into one or two *monitoring statistics*



Parametric and Nonparametric Thresholds

Example monitoring statistic: T^2 (Hotelling, 1947): $T_t^2 = (\mathbf{x}_t - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_t - \boldsymbol{\mu})$

- **Parametric thresholds** require assumptions to be met
  $T_t^2 \sim \frac{p(n^2-1)}{n(n-p)} F_{1-\alpha}(p, n-p)$
- **Nonparametric thresholds** estimate the $(1-\alpha)$ quantile of the monitoring statistics during the training period



Issues with Nonparametric Thresholds

Nonparametric thresholds are most commonly obtained via

- **kernel density estimation**
- **bootstrapping**

These methods

- Do not assume multivariate normality of the original data.
- Both assume independence in the monitoring statistics.
  - often violated (Mason and Young, 2002; Vanhatalo and Kulahci, 2015; Rato et al., 2016; Vanhatalo et al., 2017)
  - adjusting for dependence in the original data (e.g., using dynamic PCA) can still result in autocorrelated monitoring statistics.

How does dependence in the monitoring statistics affect the performance of nonparametric thresholds?

Nonparametric Thresholds

Kernel Density Estimation (KDE)

For monitoring statistics $z_1, \ldots, z_n$ during the training period, we estimate the density $f(\cdot)$ by

$$\hat{f}_h(z) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{z - z_i}{h}\right), \tag{1}$$

where K is a kernel, and h is the bandwidth.

- Find q_α such that $\int_0^{q_\alpha} \hat{f}_h(t)dt = 1 - \alpha$.

Bootstrapping

Obtain random samples, with replacement, from $\mathbf{z} = (z_1, \ldots, z_n)'$.

- Calculate the average $1 - \alpha$ sample quantile across all bootstrap samples.

Nonparametric Estimation Methods: KDE

Baseline

- Sample quantile

KDE Plug-in Bandwidth Estimators

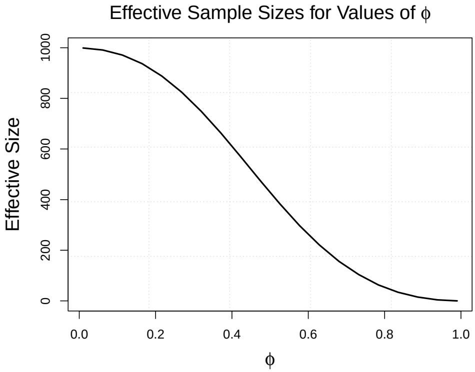
- Silverman (SLVM)

$$h = 0.9n^{-1/5} \min\left\{\mathrm{sd}(\mathbf{z}), \frac{\mathrm{IQR}(\mathbf{z})}{1.34}\right\} \tag{2}$$

- Scott (SCOTT)
    - Replace 0.9 with 1.06 in (2).
- Adjusted Silverman (ADJ-SLVM)
    - Replace n in (2) with the estimated effective sample size $\hat{n}_{\mathrm{eff}}$.
- Adjusted Scott (ADJ-SCOTT)
    - Replace 0.9 with 1.06 and n with $\hat{n}_{\mathrm{eff}}$ in (2).

Plots: Effective Sample Sizes

Autoregressive model of order 1 (AR(1)): $z_t = \phi z_{t-1} + \epsilon_t$.



$$\hat{n}_{\text{eff}}^{\text{AR}(1)} = \frac{n}{1 + \frac{1}{n}\sum_{i \neq j} \sum \frac{\hat{\phi}^{|i-j|}}{1-\hat{\phi}^2}}$$

Nonparametric Estimation Methods: Bootstrap

Bootstrap

- Standard bootstrap (BOOT)
- Moving block bootstrap (MB-BOOT)
  - Randomly sample blocks of size ℓ instead of individual observations.
- Random block bootstrap (RB-BOOT)
  - Same as MB-BOOT, but randomly generate block sizes according to a Geometric distribution with mean ℓ.

Block sizes ℓ are selected based on an estimate of the dependence.

Simulation Study Design

Goal: Evaluate the performance of different nonparametric methods when the monitoring statistic is autocorrelated.

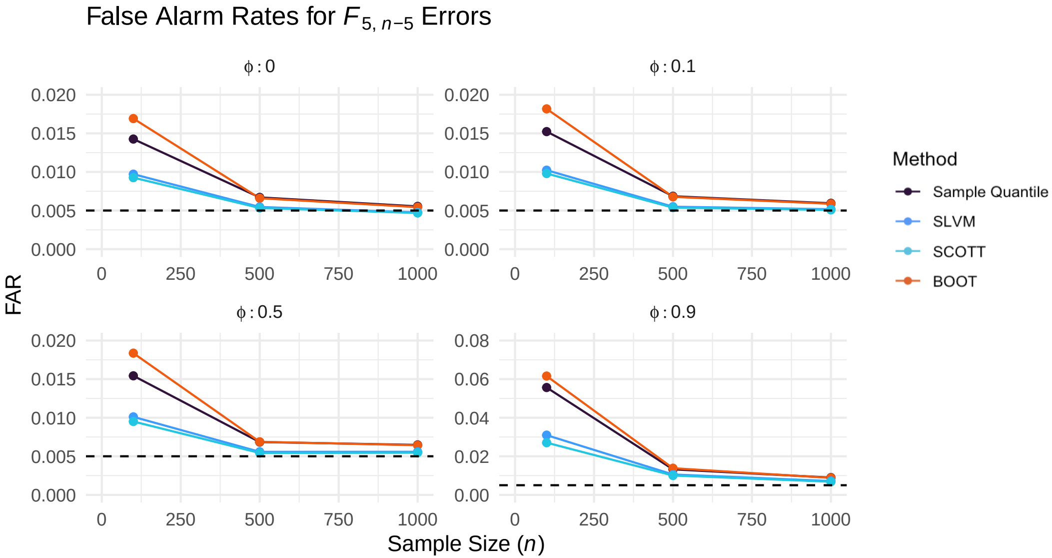
- Do nonparametric thresholds actually yield false alarm rates (FARs) equal to α?
- Which method yields FARs closest to α?

1. Generate IC monitoring statistics under different conditions
2. Evaluate FAR for each threshold

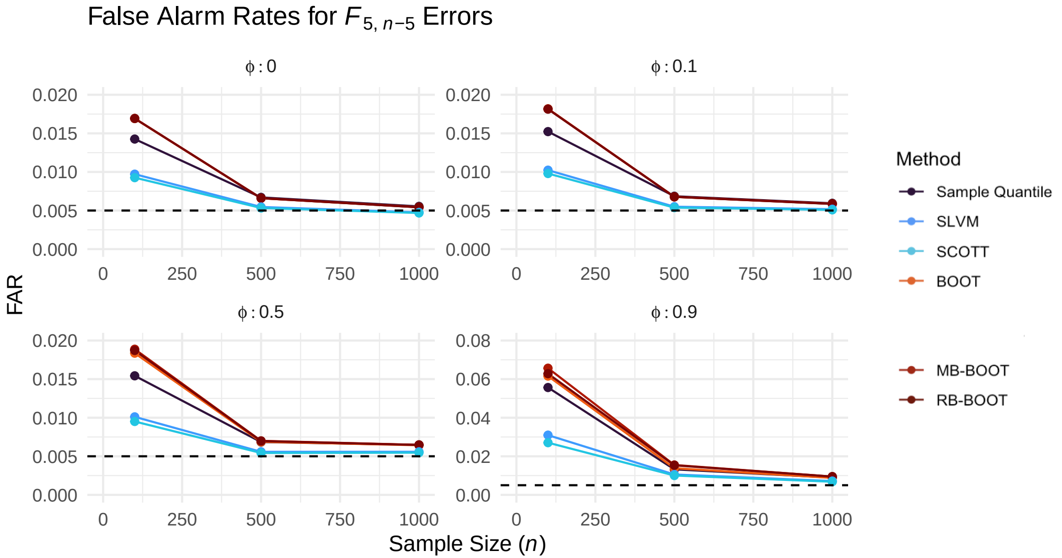
Factor	Levels
ϕ	$0, 0.1, 0.5, 0.9$
Sample Size (n)	100, 500, 1000, 5000
Error Distribution	$F_{5,20}$, $F_{5,n-5}$, Gamma$(0.05, 0.05)$

- Use $\alpha = 0.005$, which corresponds to estimating the 0.995 quantile.

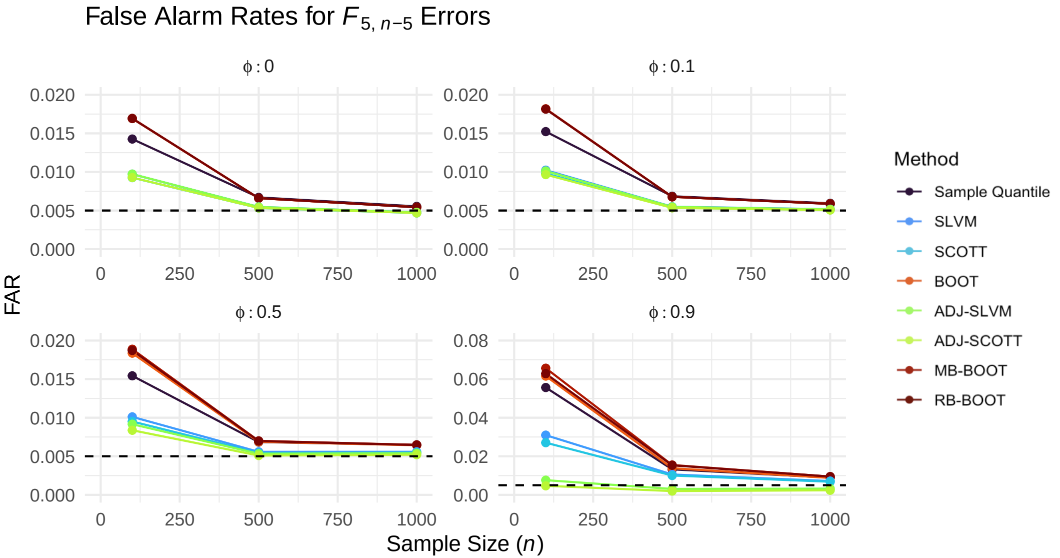
Simulation Study: AR(1) FAR Results



Simulation Study: AR(1) FAR Results



Simulation Study: AR(1) FAR Results



Simulation Study: Summary

Overall Results

- Estimator performance depends on sample size and autocorrelation strength.
- Adjusted KDE methods yield FAR values closest to 0.005 under almost all scenarios.

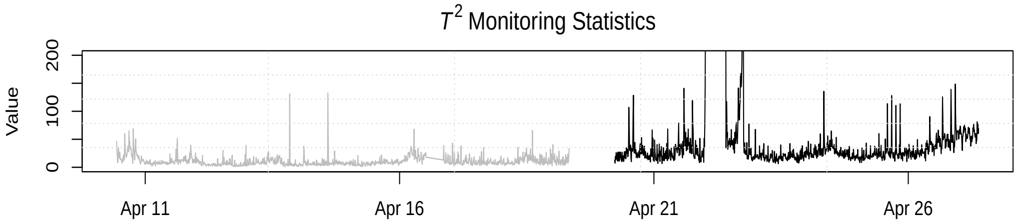
Best methods for autocorrelation strengths and sample sizes.

		Autocorrelation	
		Weak	Strong
Sample Size	Small	Any KDE	Adjusted KDE
	Large	Any	Any

Case Study: Wastewater Treatment

Demonstration-scale wastewater treatment facility in Golden, Colorado, in April 2010.

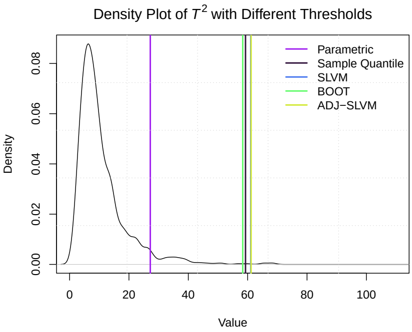
- Monitor 10 minute averages of 20 process variables from April 10 to May 10, 2010.
- The training period contains 1235 observations, and the testing period contains 1031 observations.

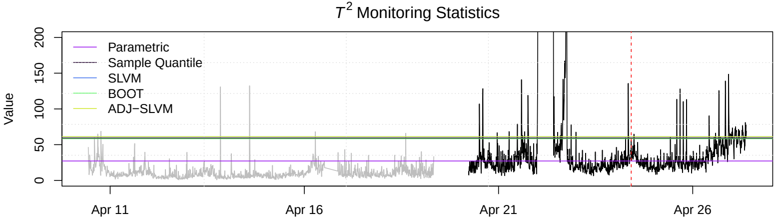


Here, $\hat{\phi} = 0.40$, $\hat{n}_{\text{eff}} = 529$, $\ell = 4$.

Case Study: T^2 Values

- Parametric threshold results in too many false alarms
- Nonparametric thresholds are similar because $\hat{\phi}$ is moderate and $\hat{n}_{\text{eff}}$ is large





Conclusion

- Monitoring statistics, such as T^2, are often autocorrelated and skewed.
  - Using parametric thresholds degrades control chart performance.
  - Nonparametric threshold estimators assume independence of the monitoring statistics.
- Adjusting for dependence can improve nonparametric threshold performance.
  - Threshold selection is most important when n is small and dependence is strong.
- Adjusted KDE methods are recommended for general use.

Conclusion

- Monitoring statistics, such as T^2, are often autocorrelated and skewed.
  - Using parametric thresholds degrades control chart performance.
  - Nonparametric threshold estimators assume independence of the monitoring statistics.
- Adjusting for dependence can improve nonparametric threshold performance.
  - Threshold selection is most important when n is small and dependence is strong.
- Adjusted KDE methods are recommended for general use.

Thank you!

Contact: taylor_grimm1@baylor.edu

References

Hotelling, H. (1947) "Multivariate quality control" In Eisenhart, C., Hastay, M. W., and Wallis, W., editors, Techniques of Statistical Analysis, pages 111–184. McGraw-Hill, New York.

Jackson, J. E. and Mudholkar, G. S. (1979) "Control procedures for residuals associated with principal component analysis," Technometrics, 21(3):341–349.

Kazor, K., Holloway, R., Cath, T., and Hering, A. S. (2016) "Comparison of linear and nonlinear dimension reduction techniques for automated process monitoring of a decentralized wastewater treatment facility," Stochastic Environmental Research and Risk Assessment, 30: 1527-1544.

Klanderman, M., Newhart, K.B., Cath. T.Y., Hering, A.S. (2020) "Fault isolation for a complex decentralized wastewater treatment facility," Journal of the Royal Statistical Society, Series C., 69, 931-951.

Klanderman, M. C., Newhart, K. B., Cath, T. Y., and Hering, A. S. (2020) "Case studies in real-time fault isolation in a decentralized wastewater treatment facility," Journal of Water Process Engineering, 38: 101556.

Ku, W., Storer, R. H., and Georgakis, C. (1995). "Disturbance detection and isolation by dynamic principal component analysis," Chemometrics and Intelligent Laboratory Systems, 30(1):179–196.

Ma, X., Zhang, L., Hu, J., and Palazoglu, A. (2018). "A model-free approach to reduce the effect of autocorrelation on statistical process control charts," Journal of Chemometrics, 32(12).

References

Mason, R. L. and Young, J. C. (2002). "Multivariate statistical process control with industrial applications," Society for Industrial and Applied Mathematics.

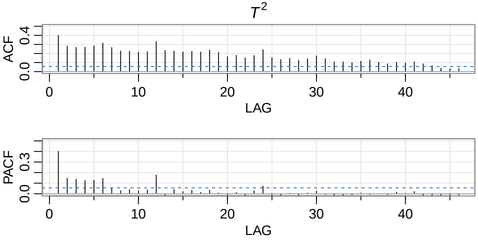
Phaladiganon, P., Kim, S. B., Chen, V. C. P., and Jiang, W. (2013). "Principal component analysis-based control charts for multivariate nonnormal distributions," Expert Systems with Applications, 40(8):3044–3054.

Rato, T., Reis, M., Schmitt, E., Hubert, M., and De Ketelaere, B. (2016). "A systematic comparison of PCA-based statistical process monitoring methods for high-dimensional, time-dependent Processes," AIChE Journal, 62(5):1478–1493.

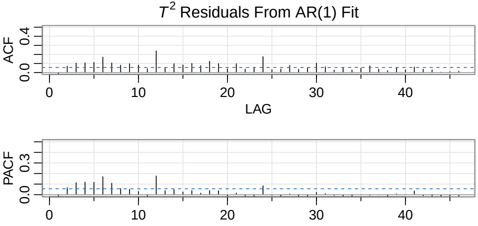
Vanhatalo, E. and Kulahci, M. (2015). "The effect of autocorrelation on the Hotelling T^2 control chart," Quality and Reliability Engineering International, 31(8):1779–1796.

Vanhatalo, E., Kulahci, M., and Bergquist, B. (2017). "On the structure of dynamic principal component analysis used in statistical process monitoring," Chemometrics and Intelligent Laboratory Systems, 167:1–11.

Original Monitoring Statistic



Residuals of AR(1) Fit



Effective Sample Size Equations

Consider $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$, where z_i are iid $\mathrm{N}(\mu, \sigma^2)$.
To create a confidence interval around $\bar{z}$, we need to determine $\mathrm{E}(\bar{z})$ and $\mathrm{Var}(\bar{z})$. We know that $\mathrm{E}(\bar{z}) = \mu$, and $\mathrm{Var}(\bar{z}) = \frac{\sigma^2}{n}$. So, $\bar{z} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$.

What if the z_i's are dependent?

$$\begin{aligned}
\mathrm{Var}(\bar{z}) &= \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} z_i) \\
&= \frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n} z_i) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{Var}(z_i) + \sum_{i \neq j}\sum \mathrm{Cov}(z_i, z_j)\right)
\end{aligned}$$

Effective Sample Size Equations: AR(1) Dependence

The AR(1) covariance function is $\mathrm{Cov}(z_i, z_j) = \frac{\sigma^2 \phi^{|i-j|}}{1-\phi^2}$.

$$\begin{aligned}
\mathrm{Var}(\bar{z}) &= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathrm{Var}(z_i) + \sum_{i \neq j}\sum \mathrm{Cov}(z_i, z_j)\right) \\
&= \frac{1}{n^2}\left(n\sigma^2 + \sigma^2 \sum_{i \neq j}\sum \frac{\phi^{|i-j|}}{1-\phi^2}\right) \\
&= \frac{\sigma^2}{n}\left(1 + \frac{1}{n}\sum_{i \neq j}\sum \frac{\phi^{|i-j|}}{1-\phi^2}\right),
\end{aligned}$$

and the effective sample size is

$$n_{\mathrm{eff}} = \frac{n}{1 + \frac{1}{n}\sum_{i \neq j}\sum \frac{\phi^{|i-j|}}{1-\phi^2}}.$$

Dependent Bootstrap Block Sizes

For MB-BOOT and RB-BOOT, we select ℓ by finding ℓ such that

$$|\hat{\rho}(\ell)| \leq 0.05,$$

where the AR(1) autocorrelation function (ACF) is $\rho(\ell) = \phi^{\ell}$.

AR(1)

$$|\hat{\rho}(\ell)| = |\hat{\phi}^{\ell}| \leq 0.05$$

ARMA(1, 1)

$$|\hat{\rho}(\ell)| = \left| \frac{(1 + \hat{\theta}\hat{\phi})(\hat{\phi} + \hat{\theta})}{1 + 2\hat{\theta}\hat{\phi} + \hat{\theta}^2} \hat{\phi}^{\ell-1} \right| \leq 0.05$$

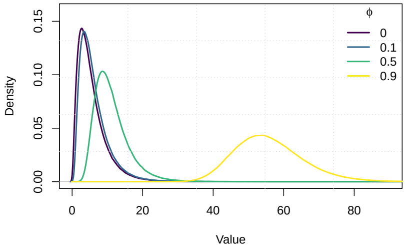
Restrict ℓ to be no larger than $n/2$.

Simulation Study Steps

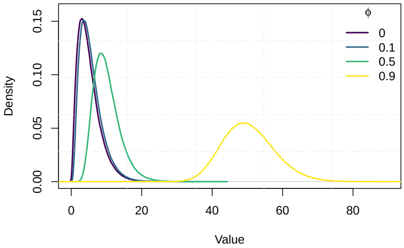
1. Generate n IC monitoring statistics.
2. Compute each threshold using the sample from 1.
3. Generate 2000 more IC monitoring statistics. Record the FAR and RL for each threshold from 2.
4. Repeat 3 5000 times.
5. Compute the mean of the 5000 FARs and RLs.
6. Repeat 1-5 1000 times. Compute the mean of the 1000 FARs and $\mathrm{ARL}_{\mathrm{IC}}$s.

Simulation Study: True AR(1) Distributions

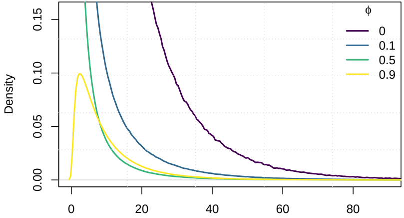
AR(1) Series with $F_{5,\,20}$ Errors



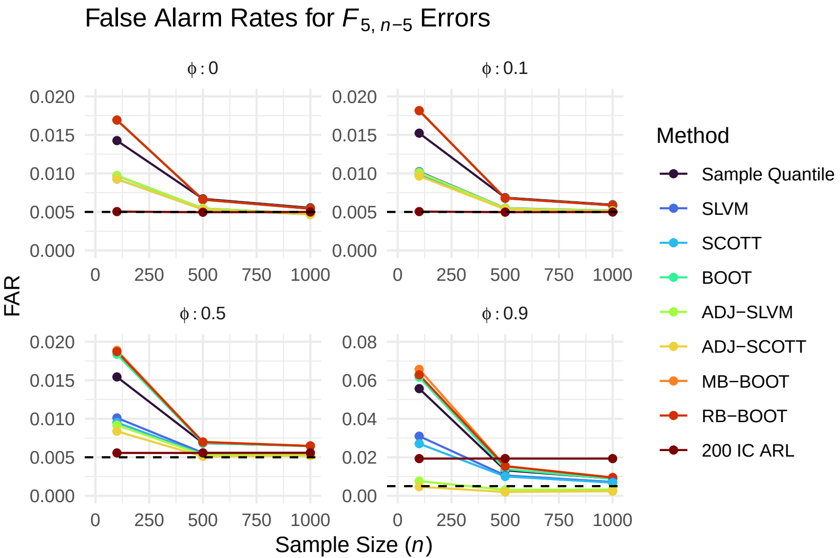
AR(1) Series with $F_{5,\,n-5}$ Errors



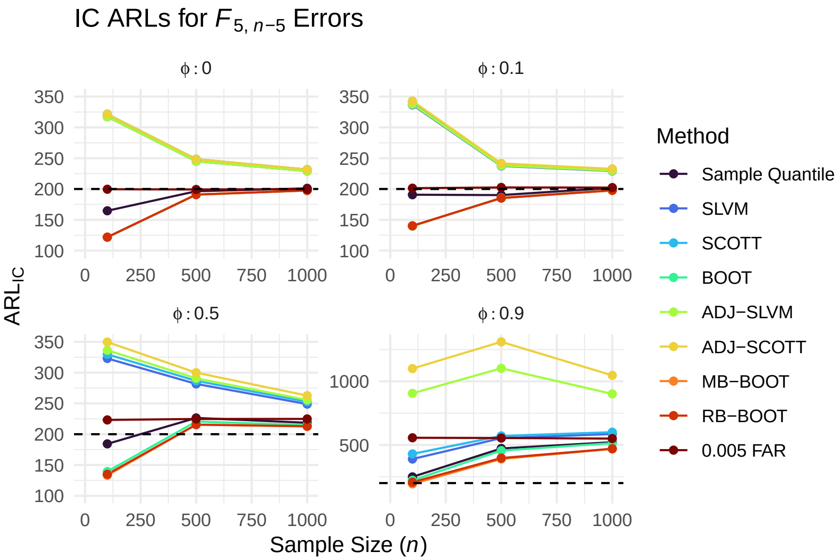
AR(1) Series with Gamma(0.05, 0.05) Errors



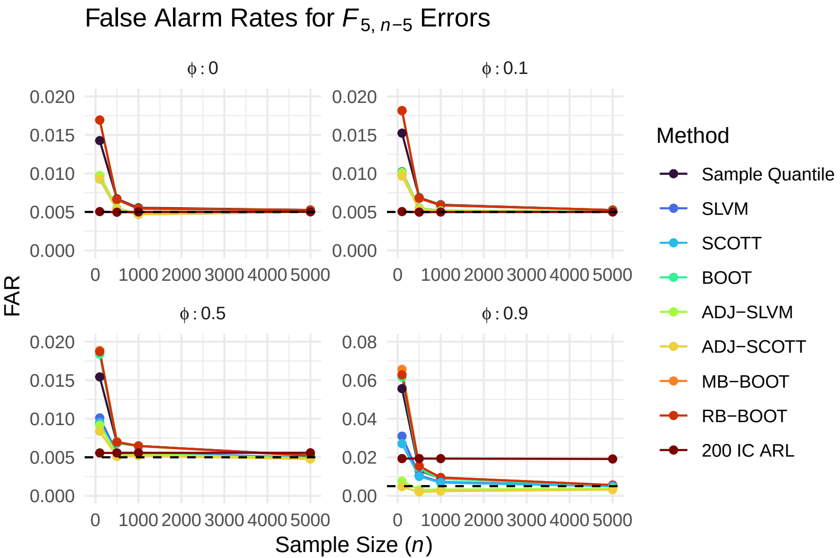
Simulation Study: AR(1) FAR Results



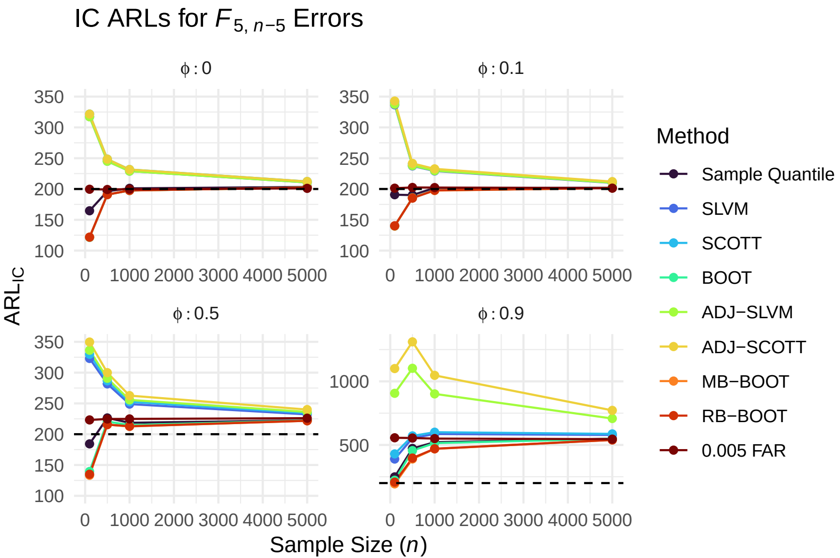
Simulation Study: AR(1) ARL_{IC} Results



Simulation Study: AR(1) FAR Results



Simulation Study: AR(1) ARL_{IC} Results



Simulation Study: Additional AR(1) Results

False alarm rates for each estimator for AR(1) monitoring statistics with $F_{5,20}$ errors. The value closest to 0.005 in each row is shown in bold.

			Assuming Independence			Accounting for Dependence			
ϕ	n	Sample Quantile	SLVM	SCOTT	BOOT	ADJ-SLVM	ADJ-SCOTT	MB-BOOT	RB-BOOT
0	100	0.0116	0.0078	**0.0077**	0.0138	0.0078	**0.0077**	0.0139	0.0138
	500	0.0066	0.0056	**0.0055**	0.0063	0.0056	**0.0055**	0.0063	0.0063
	1000	0.0057	**0.0051**	**0.0051**	0.0055	**0.0051**	**0.0051**	0.0055	0.0055
	5000	0.0052	**0.0050**	**0.0050**	0.0052	**0.0050**	**0.0050**	0.0052	0.0052
0.1	100	0.0147	0.0104	0.0102	0.0171	0.0104	**0.0101**	0.0170	0.0170
	500	0.0075	0.0064	**0.0063**	0.0072	0.0064	**0.0063**	0.0072	0.0072
	1000	0.0061	0.0054	**0.0053**	0.0059	0.0054	**0.0053**	0.0058	0.0059
	5000	0.0053	**0.0051**	**0.0051**	0.0053	**0.0051**	**0.0051**	0.0053	0.0053
0.5	100	0.0155	0.0110	0.0105	0.0182	0.0102	**0.0096**	0.0187	0.0185
	500	0.0077	0.0063	0.0061	0.0072	0.0061	**0.0059**	0.0072	0.0072
	1000	0.0065	0.0058	0.0057	0.0063	0.0057	**0.0056**	0.0064	0.0063
	5000	0.0052	0.0051	**0.0050**	0.0052	**0.0050**	**0.0050**	0.0052	0.0052
0.9	100	0.0656	0.0401	0.0357	0.0705	0.0123	**0.0084**	0.0732	0.0713
	500	0.0149	0.0118	0.0113	0.0150	**0.0046**	0.0033	0.0166	0.0164
	1000	0.0084	0.0072	0.0070	0.0084	**0.0046**	0.0038	0.0090	0.0090
	5000	0.0058	0.0054	**0.0053**	0.0058	0.0045	0.0041	0.0058	0.0058

Simulation Study: Additional AR(1) Results

False alarm rates for each estimator for AR(1) monitoring statistics with Gamma(0.05, 0.05) errors. The value closest to 0.005 in each row is shown in bold.

			Assuming Independence			Accounting for Dependence			
ϕ	n	Sample Quantile	SLVM	SCOTT	BOOT	ADJ-SLVM	ADJ-SCOTT	MB-BOOT	RB-BOOT
0	100	0.0110	0.0103	**0.0099**	0.0128	0.0103	0.0100	0.0129	0.0129
	500	0.0069	0.0078	0.0078	**0.0066**	0.0078	0.0078	**0.0066**	**0.0066**
	1000	0.0057	0.0062	0.0062	**0.0056**	0.0062	0.0062	**0.0056**	**0.0056**
	5000	0.0054	0.0064	0.0063	**0.0053**	0.0064	0.0063	**0.0053**	0.0053
0.1	100	0.0124	**0.0093**	**0.0093**	0.0145	**0.0093**	**0.0093**	0.0145	0.0145
	500	0.0072	**0.0063**	0.0063	0.0068	**0.0063**	**0.0063**	0.0068	0.0068
	1000	0.0061	0.0061	0.0061	**0.0059**	0.0061	0.0061	**0.0059**	**0.0059**
	5000	**0.0052**	0.0057	0.0055	**0.0052**	0.0056	0.0055	**0.0052**	**0.0052**
0.5	100	0.0172	**0.0128**	**0.0128**	0.0198	0.0128	**0.0128**	0.0204	0.0202
	500	0.0073	**0.0064**	**0.0064**	0.0071	**0.0064**	**0.0064**	0.0072	0.0072
	1000	0.0058	**0.0054**	**0.0054**	0.0057	**0.0054**	**0.0054**	0.0058	0.0058
	5000	0.0051	**0.0050**	**0.0050**	0.0051	**0.0050**	**0.0050**	0.0051	0.0051
0.9	100	0.0797	0.0705	0.0694	0.0846	0.0608	**0.0574**	0.0925	0.0901
	500	0.0131	0.0118	0.0117	0.0133	0.0108	**0.0104**	0.0147	0.0146
	1000	0.0092	0.0085	0.0084	0.0092	0.0081	**0.0079**	0.0100	0.0100
	5000	0.0053	0.0052	0.0052	0.0053	0.0051	**0.0050**	0.0054	0.0054

Simulation Study: Additional AR(1) Results

ARL_{IC} for each estimator for AR(1) monitoring statistics with $F_{5,20}$ errors.

				Assuming Independence			Accounting for Dependence			
ϕ	n	True Quantile	Sample Quantile	SLVM	SCOTT	BOOT	ADJ-SLVM	ADJ-SCOTT	MB-BOOT	RB-BOOT
0	100	200	180	315	317	136	315	317	137	137
	500	200	201	247	248	204	247	248	204	204
	1000	200	202	225	226	203	225	226	203	203
	5000	200	200	205	206	200	205	206	200	200
0.1	100	202	182	313	315	139	313	315	138	139
	500	202	203	247	248	203	247	248	203	203
	1000	201	203	226	227	201	226	227	201	201
	5000	201	200	205	206	200	205	206	200	200
0.5	100	223	203	323	326	160	329	336	154	155
	500	223	223	273	274	227	276	279	222	222
	1000	223	223	248	250	223	251	254	221	221
	5000	223	225	231	232	224	232	234	224	224
0.9	100	557	237	345	374	214	727	879	190	201
	500	556	476	544	556	463	848	984	403	408
	1000	558	540	591	600	533	757	838	492	493
	5000	557	569	591	596	569	669	706	562	561

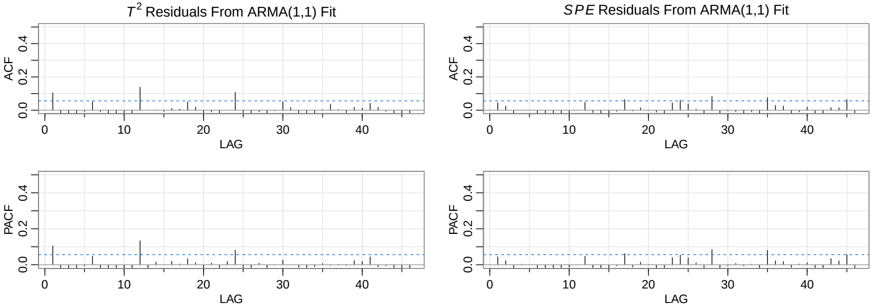
Simulation Study: Additional AR(1) Results

ARL_{IC} for each estimator for AR(1) monitoring statistics with Gamma(0.05, 0.05) errors.

				Assuming Independence			Accounting for Dependence			
ϕ	n	True Quantile	Sample Quantile	SLVM	SCOTT	BOOT	ADJ-SLVM	ADJ-SCOTT	MB-BOOT	RB-BOOT
0	100	201	189	258	271	145	258	272	145	144
	500	202	202	169	170	203	169	170	203	203
	1000	202	200	180	183	199	180	183	199	199
	5000	202	201	165	168	201	165	168	201	201
0.1	100	203	186	316	316	144	316	316	144	144
	500	203	189	225	225	190	225	226	190	190
	1000	203	199	199	200	198	199	200	198	198
	5000	203	200	183	187	201	184	188	201	201
0.5	100	233	206	313	313	165	314	314	156	157
	500	233	233	273	273	236	273	273	229	229
	1000	233	229	246	248	231	250	252	228	228
	5000	233	230	234	234	231	234	234	230	230
0.9	100	645	260	297	301	239	348	370	200	211
	500.	645	536	578	579	520	595	603	452	455
	1000	645	592	626	627	588	640	646	545	546
	5000	645	622	632	632	622	640	643	615	615

Case Study

Assessing Dependence in the Monitoring Statistics



ACF and PACF plots for the residuals of T^2 (left) and SPE (right) from an ARMA(1, 1) fit during the training period.

Case Study: Proportion of OC Observations

Proportion and number (in parentheses) of monitoring statistics flagged as OC during training and testing for selected methods when $\alpha = 0.005$.

Monitoring Statistic	Method	Training (# of obs / 1235)	Testing (# of obs / 1031)
T^2	Parametric	0.0478 (59)	0.481 (496)
	Sample Quantile	0.0057 (7)	0.153 (158)
	SLVM	0.0049 (6)	0.143 (147)
	BOOT	0.0057 (7)	0.157 (162)
	ADJ-SLVM	0.0049 (6)	0.142 (146)

Thresholds for T^2 with effective sample sizes and block sizes computed based on an AR(1) structure.

Monitoring Statistic	Sample Quantile	Assuming Independence			Accounting for Dependence			
		SLVM	SCOTT	BOOT	ADJ-SLVM	ADJ-SCOTT	MB-BOOT	RB-BOOT
T^2	59.3	61.0	61.3	58.3	61.2	61.3	57.8	58.2

Case Study: Training Period Sizes

Thresholds were also computed for different training period sample sizes.

- Parametric thresholds are mostly unchanged.
- Nonparametric thresholds vary greatly.
  - Smaller training periods may not capture the full range of normal operating behavior, resulting in major changes in threshold values.

Parametric and nonparametric thresholds for T^2 and SPE using different sizes of training periods.

Monitoring Statistic	Training Size	Parametric	Sample Quantile	SLVM	BOOT	ADJ-SLVM
T^2	100	28.4	23.4	24.5	22.7	24.6
	200	27.7	47.3	47.9	45.5	48.1
	300	26.8	67.5	68.9	68.8	69.1
	400	26.4	93.0	95.9	86.9	95.7
	500	26.2	105.7	107.2	96.9	107.3
SPE	100	11.8	11.2	12.0	10.6	12.0
	200	10.5	10.5	11.4	10.3	11.4
	300	12.9	13.3	13.8	13.9	13.8
	400	12.4	14.8	15.0	14.5	15.0
	500	14.2	18.9	19.1	18.0	19.1