ТЕХНИЧЕСКИ УНИВЕРСИТЕТ - ВАРНА
Technical University of Varna
е-списание ТУ-Варна
ISSN 3033-2559
OPEN ACCESS
DOI : https://doi.org/10.29114/ejtuv.vol1.iss3.89

ИДЕНТИФИКАЦИЯ НА ДИНАМИЧНИ СИСТЕМИ ЧРЕЗ КОРЕЛАЦИОНЕН АНАЛИЗ ЗА УСЪВЪРШЕНСТВАНО УПРАВЛЕНИЕ И АВТОМАТИЗАЦИЯ НА МНОГОМЕРНИ ПРОМИШЛЕНИ ПРОЦЕСИ

Илиян Василев Василев
Катедра Автоматизация на производството
Химикотехнологичен и металургичен университет - София

Резюме: Автокорелацията на сигналите и динамичните взаимодействия между променливите в промишлените процеси могат да изместят пиковете на взаимнокорелационните функции и да доведат до неточно оценяване на времевите закъснения. В статията се изследва възпроизводима методика за идентификация на закъсненията, която съчетава предварителна времева декорелация чрез авторегресионно филтриране (prewhitening), взаимнокорелационен анализ по двойки и векторно авторегресионно (VAR) моделиране с отчитане на условните зависимости между наблюдаваните променливи. Методиката е проверена върху синтетични данни за термохидравличен процес с четири променливи, три преки връзки с предварително зададени закъснения и една косвена връзка. Наборът от 3200 наблюдения е разделен хронологично на обучаваща и тестова извадка, а надеждността на идентификацията е оценена допълнително чрез 200 независими реализации по метода Монте Карло. При демонстрационната реализация взаимнокорелационният анализ на нефилтрираните сигнали дава оценки за две от трите закъснения с отклонение от един период на дискретизация спрямо зададените стойности. След времева декорелация и трите закъснения са идентифицирани правилно. Този резултат се запазва във всичките 200 реализации на експеримента Монте Карло, докато при анализа без времева декорелация правилните оценки за трите закъснения са получени съответно в 2, 158 и 13 реализации. За модела VAR(3), чийто ред е избран чрез Байесовия информационен критерий (BIC), не се установява статистически значим допълнителен принос на управляващия сигнал към нагревателя за прогнозиране на налягането при отчитане на предходните стойности на температурата и управляващия сигнал към клапана ($p = 0{,}496$). Върху тестовата извадка грешките при едностъпково прогнозиране на температурата и налягането са съответно с 39,2% и 40,9% по-ниски спрямо тези на отделни авторегресионни модели от първи ред - AR(1). Резултатите за изследвания процес показват, че времевата декорелация подобрява точността на идентификацията на времевите закъснения, а условният анализ във VAR модела позволява да се оцени допълнителната прогностична информация, която дадена променлива предоставя при отчитане на останалите наблюдавани променливи.

Ключови думи: промишлена автоматизация, идентификация на времеви закъснения, автокорелация, векторна авторегресия, мониторинг на процеси, прогнозно моделиране, метод Монте Карло, времева декорелация

ТЕХНИЧЕСКИ УНИВЕРСИТЕТ - ВАРНА
Technical University of Varna
е-списание ТУ-Варна
ISSN 3033-2559
OPEN ACCESS
DOI : https://doi.org/10.29114/ejtuv.vol1.iss3.89

CORRELATION-BASED DYNAMIC SYSTEM IDENTIFICATION FOR ADVANCED CONTROL AND AUTOMATION OF MULTIVARIABLE INDUSTRIAL PROCESSES

Iliyan Vasilev Vasilev
Department of Industrial Automation
University of Chemical Technology and Metallurgy, Sofia

Abstract: *Industrial processes often contain persistent signals and delayed interactions, which can shift the peaks of conventional cross-correlation functions and affect delay estimates. This paper investigates a reproducible method for identifying delays using autoregressive prewhitening, pairwise correlation analysis and conditional vector autoregressive (VAR) modeling. The study uses a synthetic thermal–hydraulic process dataset with four variables, three known direct delays and one indirect interaction. A chronological split separates the 3,200 observations into estimation and test sets, while 200 independent Monte Carlo realizations assess the reliability of lag identification. In the illustrative dataset, raw correlation places two direct delays one sampling interval away from their actual values, whereas prewhitening identifies all three correctly. Across the 200 realizations, prewhitened correlation identifies all three delays in every run. In comparison, raw correlation gives correct results in only 2, 158 and 13 runs for the three delays, respectively. In the BIC-selected VAR(3) model, heater input did not contribute significantly to the prediction of pressure once lagged temperature and valve input were included ($p = 0.496$). On the held-out data, one-step prediction errors for temperature and pressure were 39.2% and 40.9% lower than those of separate AR(1) models. Prewhitening corrected the position of the delay peaks, and conditioning on the other measured variables.*

Keywords: *industrial automation, time-delay identification, cross-correlation, vector autoregression, process monitoring, predictive modeling, Monte Carlo simulation, prewhitening.*

1. Introduction

Industrial automation systems continuously collect measurements from actuators, sensors and supervisory controllers. A change in an input signal may influence a downstream measurement only after transport, thermal inertia, actuator dynamics or computational delays. Recognizing such temporal relations is important for model-based monitoring, fault localization, digital-twin calibration and the selection of control sampling intervals. In multivariable installations, however, simultaneous signals can be highly autocorrelated and connected through intermediate process stages [4, 8, 9].

The sample cross-correlation function (CCF) is a convenient first tool for describing relationships between two time series. Its largest peak is not, in general, the first physically meaningful response delay: autocorrelated excitation and the dynamics of the response can shift or broaden the peak. Furthermore, a large cross-correlation may result from a common

DOI : https://doi.org/10.29114/ejtuv.vol1.iss3.89

driver or from an indirect path rather than a direct interaction. Classical transfer-function identification addresses these issues through prewhitening, while vector autoregression offers a way to condition each predictive relationship on the other observed variables [2, 7].

In this article, we examine whether a relatively simple procedure for identifying time delays can be used at an early stage, before a detailed control-oriented model is developed. The approach rests on three elements. First, defining the lag convention explicitly, so that the direction and size of each estimated delay can be interpreted without ambiguity. Second, it is used the cross-correlation function of prewhitened series together with a conditional VAR model, allowing each method to serve as a check on the other. Third, we test the procedure on a simulated thermal-hydraulic system in which the true delays are known in advance, so the estimates can be compared directly against them and the analysis can be reproduced. The study should therefore be read as a methodological validation.

2. Methodology

2.1. Autocorrelation and cross-correlation analysis

If x_t and y_t are sampled, approximately covariance-stationary process variables. For a nonnegative lag k, the lead–lag correlation is then:

$$r_{xy}(k) = \mathrm{cov}(x_t, y_{t+k}), \qquad k = 0,1, \dots, K. \tag{1}$$

A positive value of k means that the source signal x is observed before the response y. Autocorrelation is obtained by setting x = y. The empirical screening criterion used here is the largest absolute sample correlation among k = 0, …, 10. This criterion is intentionally simple and serves only to propose delays for further testing. It can overestimate a transport delay when the driver itself is persistent or the output retains the effect of previous inputs [2].

2.2. Prewhitening of candidate input–output pairs

For each candidate driver x, an autoregressive model of order m is estimated using the training data. The fitted model is then used to apply the same linear filter to x and its candidate response y:

$$\tilde{x}_t = x_t - \sum_{j=1}^{m} \hat{\phi}_j \, x_{t-j}, \qquad \tilde{y}_t = y_t - \sum_{j=1}^{m} \hat{\phi}_j \, y_{t-j}. \tag{2}$$

The variable $\hat{\varphi_j}$ are estimated autoregressive coefficients and the tildes stand for the filtered series. Applying the same filter to both variables is essential: filtering only the driver changes the apparent transfer relation. The autoregressive order is selected by the Bayesian information criterion (BIC) among orders 1–8; the CCF of the two filtered series is examined over the same nonnegative lag interval [2, 10]. Prewhitening suppresses correlations arising mainly from persistence in the candidate input, but it does not remove links mediated by other process variables.

DOI : https://doi.org/10.29114/ejtuv.vol1.iss3.89

2.3. Conditional multivariable modeling

The observed signals are assembled into the vector $z_t = [u_t, v_t, T_t, P_t]^T$. A stable VAR model with p previous observations of each signal is fitted by least squares:

$$\mathbf{z}_t = \mathbf{c} + \sum_{i=1}^{p} \mathbf{A}_i \, \mathbf{z}_{t-i} + \boldsymbol{\varepsilon}_t. \tag{3}$$

The coefficient matrices A_i describe linear predictive relationships at each lag. We choose the model order with the Bayesian information criterion (BIC), considering values of p from 1 to 8 [7, 10]. For each response equation, a conditional block F-test is used to judge whether a chosen source variable carries useful predictive information. Under the null hypothesis, every coefficient attached to the lags of that source equals zero. Because the lagged values of the remaining variables are kept in the equation, a rejection indicates that the source improves prediction beyond what the other variables already explain. A significant result should be read in the Granger sense: it shows predictive information, not physical causation or the effect of an intervention [5]. Finally, we check the residuals with a multivariate portmanteau test to make sure no serial correlation remains [6].

2.4. Evaluation protocol

Model estimation and order selection use the first 2,400 observations, leaving the final 800 for testing. The forecasting process follows a one-step-ahead approach, where each prediction relies on actual values from previous observations rather than earlier forecasts. This firmly avoids recursive multistep forecasting. For clear and exhaustive comparison, the benchmark includes two separate AR(1) models, one for temperature and one for pressure, both trained on the same data interval. The root-mean-square error (RMSE) serves as the measure of forecast accuracy:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{\mathrm{test}}} \sum_{t=1}^{N_{\mathrm{test}}} (y_t - \hat{y}_t)^2}. \tag{4}$$

In a second experiment, we generate 200 independent synthetic datasets, each with 1,200 retained samples. For every dataset and every known direct link, we record whether the raw and the prewhitened CCF identify the correct lag, and from this we compute the correct-lag recovery rate for each. Both screening rules search over k = 0, …, 10 and pick the lag where the absolute correlation is largest. Repeating the evaluation over many datasets makes it less likely that one favourable realisation is mistaken for typical behaviour.

ТЕХНИЧЕСКИ УНИВЕРСИТЕТ - ВАРНА
Technical University of Varna
е-списание ТУ-Варна
ISSN 3033-2559
OPEN ACCESS
DOI : https://doi.org/10.29114/ejtuv.vol1.iss3.89

3. Simulated industrial process

A simplified heated-circulation process provides reference delays without the need for undisclosed industrial data. The model uses four dimensionless variables to describe deviations from nominal operating conditions. Heater input u and valve input v serve as the two command variables, while outlet temperature T and line pressure P represent the measured outputs. The sampling interval is $\Delta t = 10$s.:

$$\begin{aligned} u_t &= 0.82u_{t-1} + \varepsilon_{u,t}, \\ v_t &= 0.70v_{t-1} + \varepsilon_{v,t}, \\ T_t &= 0.60T_{t-1} + 0.70u_{t-3} + \varepsilon_{T,t}, \\ P_t &= 0.55P_{t-1} + 0.60v_{t-2} + 0.40T_{t-2} + \varepsilon_{P,t}. \end{aligned} \tag{5}$$

The four innovations ε are mutually independent, zero-mean Gaussian disturbances with standard deviations (1.00, 1.00, 0.65, 0.65). The autoregressive coefficients are below one and the lagged coupling is feedforward, giving a definite stable stationary generating process. The imposed direct paths are u → T at three samples (30 s), v → P at two samples (20 s), and T → P at two samples (20 s). No direct u → P term is present. Any association between them is transmitted through temperature. The diagram in Fig. 1 summarizes these dependencies.

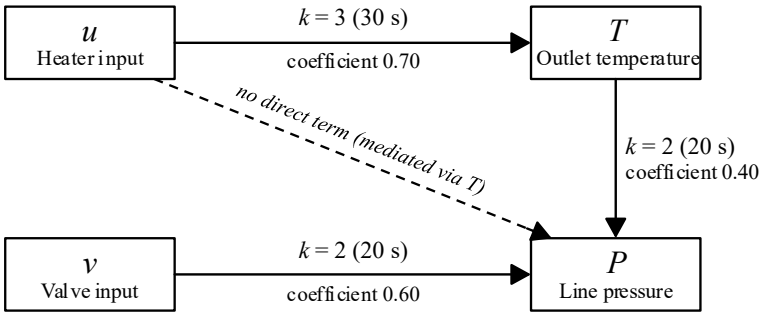


Fig. 1. Direct lag structure of the synthetic thermal–hydraulic process.

Table 1

Variables and autoregressive characteristics of the generating model

Variable	Illustrative role	Own lag-1 coefficient	Innovation SD
u	Heater input	0.82	1.00
v	Valve input	0.70	1.00
T	Outlet temperature	0.60	0.65
P	Line pressure	0.55	0.65

A pseudorandom generator with seed 20261029 produces one series of 3,200 usable observations after a 300-sample burn-in. For the Monte Carlo experiment, independent seeds

ISSN 3033-2559
OPEN ACCESS
DOI : https://doi.org/10.29114/ejtuv.vol1.iss3.89

130000–130199 and the same burn-in are used. The innovations are resampled independently in each replicate. The data represent a controlled numerical experiment; they should not be interpreted as measured temperature, pressure or actuator data from a specific installation.

4. Results and discussion

4.1. Delay estimation from the cross-correlation functions

For the main realisation, the raw CCF between heater input and outlet temperature had its peak at k = 4. The true delay in the simulation was k = 3, so the raw estimate was off by one sample. The same problem showed up between temperature and pressure, where the raw peak was at k = 3 but the direct effect had been imposed at k = 2. After prewhitening, both peaks were found at the right lags (Fig. 2). No such error occurred on the valve-to-pressure path: the raw and prewhitened CCFs both gave the correct lag of two samples. BIC chose prewhitening orders of 1 for u and v, and 3 for T, in the representative dataset.

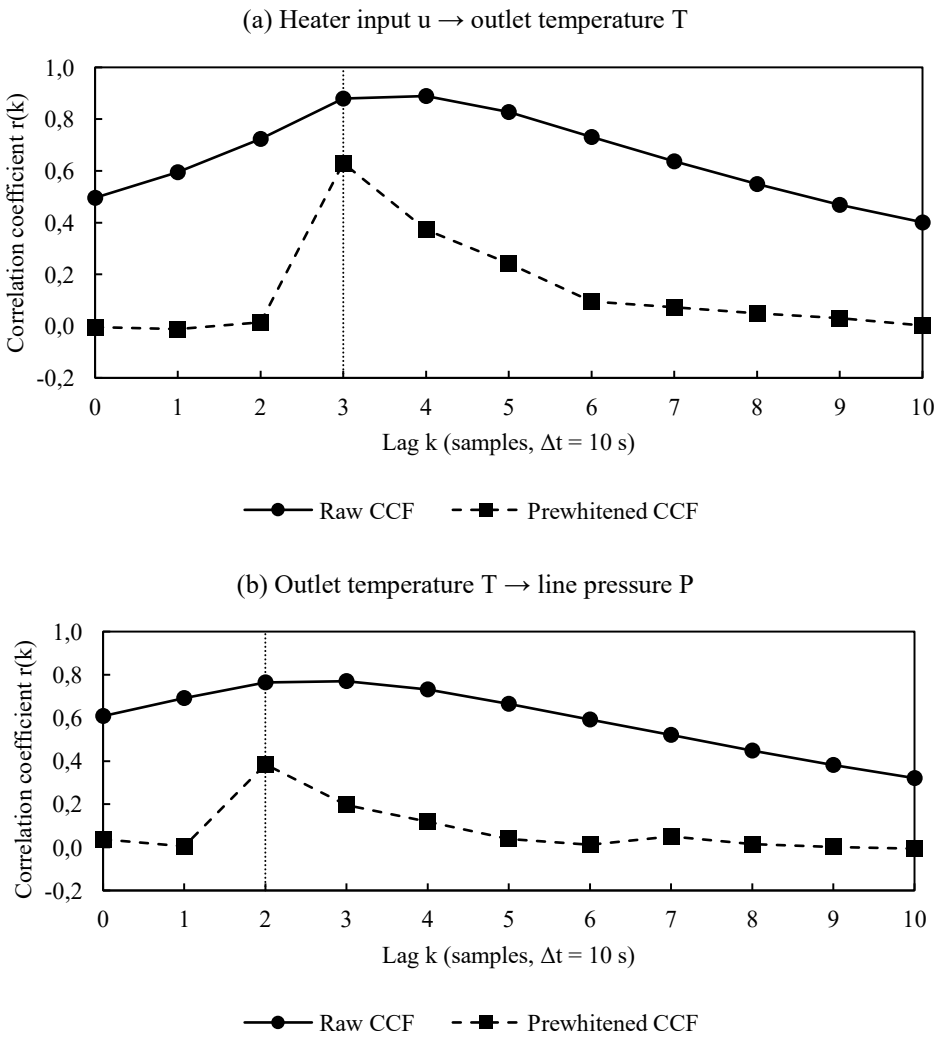


Fig. 2. Raw and prewhitened cross-correlations for two direct paths; dotted lines mark the imposed lag.

DOI : https://doi.org/10.29114/ejtuv.vol1.iss3.89

Table 2

Pairwise lag peaks and conditional VAR evidence (main realization)

Candidate link	Direct lag in model	Raw peak (k; r)	Filtered peak (k; r)	VAR block-test p
u → T	3	4; 0.889	3; 0.630	< 0.001
v → P	2	2; 0.478	2; 0.482	< 0.001
T → P	2	3; 0.770	2; 0.385	< 0.001
u → P	None	7; 0.695	7; 0.264	0.496

The final row of Table 2 highlights a limitation of pairwise analysis. Heater input and pressure show a strong correlation at a lag of seven samples, even though the generating equations contain no direct link between them. Prewhitening reduces this peak but does not remove it completely, as the indirect path through temperature (u → T → P) remains. The conditional VAR model takes a different approach by assessing the heater's contribution while also accounting for lagged temperature and valve signals. With these variables included, the test does not reject the null hypothesis that heater input provides no additional direct predictive information ($p = 0.496$).

4.2. Multivariable model diagnostics and prediction

The BIC minimum among VAR orders 1–8 occurs at $p = 3$, matching the largest imposed direct lag. The fitted direct-path coefficients are 0.696 for u_{t-3} in the temperature equation, 0.626 for v_{t-2} in the pressure equation and 0.362 for T_{t-2} in the pressure equation, compared with generating coefficients 0.700, 0.600 and 0.400. Conditional block tests for these three source–response pairs yield $p < 0.001$. The multivariate residual portmanteau test through lag 12 does not reject residual whiteness ($p = 0.231$), although this does not verify every physical assumption of the model.

Table 3

One-step prediction on the held-out 800 observations

Response	AR(1) RMSE	VAR(3) RMSE	RMSE reduction
Temperature T	1.069	0.650	39.2%
Pressure P	1.070	0.633	40.9%

Table 3 compares predictive errors in the dimensionless units used by the generator. The VAR improvement reflects the availability of informative lagged inputs, not a demonstrated gain in plant-level energy efficiency, fault-detection accuracy or controller performance. In operation, forecasts must additionally be validated under closed-loop behavior, operating-state changes and sensor faults.

DOI : https://doi.org/10.29114/ejtuv.vol1.iss3.89

4.3. Repeated-simulation assessment

Table 4

Correct direct-lag identification in 200 independent realizations

Direct link	True lag	Raw CCF	Prewhitened CCF
u → T	3	2 / 200 (1.0%)	200 / 200 (100%)
v → P	2	158 / 200 (79.0%)	200 / 200 (100%)
T → P	2	13 / 200 (6.5%)	200 / 200 (100%)

All 200 prewhitened CCF runs recover the three imposed direct delays in this particular linear and stationary setting (Table 4). In contrast, raw CCF frequently selects an adjacent lag because command persistence and output dynamics spread a delayed response across neighboring correlation coefficients. These are empirical success frequencies conditional on the chosen generating equations, noise levels, sample sizes and search window. They are not general detection probabilities for arbitrary industrial processes, and perfect recovery in this simulation should not be transferred to field deployments.

5. Engineering interpretation and limitations

The text explains how sensor readings are synchronised and how the model is organised, which may later support operator diagnostics and controller design. In the given numerical example, changes in heater input affect the outlet temperature after about 30s, while changes in valve input affect pressure after about 20s. With an assumed measurement interval of 10 s, these delays span several samples, so they should appear explicitly in any dynamic model identified from the data. This conclusion applies only to the simulated installation. A controller for a real plant would also require documented hardware timing, knowledge of the actuators and trials on the plant itself. Timestamps must be reliable, missing values handled consistently, and start-up periods and disturbances separated from steady operation, because nonstationarity and structural breaks can distort the correlations. Only then does it make sense to screen candidate delays: the same lead–lag convention should be used throughout, the inputs prewhitened where needed, and the results compared with a multivariable model that includes the other measured inputs. Model order and residual behaviour should be checked rather than assumed [1, 6, 7, 10].

The method is not without limitations. What a VAR model and a Granger-type test actually capture is whether one signal helps to predict another. That is not the same as a physical cause, especially in a plant where some disturbances are unmeasured, feedback loops are active, or a relevant variable has been left out of the model. Responses that occur within the same sampling interval add a further difficulty, since their direction cannot be resolved from lagged terms alone. Correlation-based screening is also sensitive to effects that a linear model ignores, such as nonlinearity, transport times that change during operation, changes in sampling rate and strong collinearity between inputs. For large sensor networks, an unrestricted VAR quickly becomes impractical, and regularised estimation combined with cross-validation is likely to work better [3]. The most important gap, however, is the absence

DOI : https://doi.org/10.29114/ejtuv.vol1.iss3.89

of real data. Before the method can support operational decisions, it has to be tested on synchronised measurements from an actual process, ideally including planned interventions or experiments grounded in engineering knowledge.

6. Conclusion

The procedure examined in this study combines prewhitened cross-correlation for screening candidate delays with a conditional VAR model for checking them. It was tested on a four-variable thermal–hydraulic simulation with three known direct delays and one mediated path. In the main realisation, the raw CCF placed the heater-to-temperature and temperature-to-pressure peaks one sample later than the imposed lags, at k = 4 and k = 3 instead of k = 3 and k = 2. Over 200 independent realisations, the raw CCF identified these two delays correctly in only 2 and 13 runs, respectively, and the valve-to-pressure delay in 158 runs. The prewhitened CCF identified all three delays in every run. The displacement was largest on the paths whose source signal was strongly persistent, namely heater input and temperature, which inherits the heater dynamics. This agrees with the expected effect of input autocorrelation on the position of the correlation peak. Only a pairwise screening alone could not separate direct from mediated dependence- heater input and pressure remained correlated at a lag of seven samples after prewhitening, although the generating equations contain no direct term between them. In the BIC-selected VAR(3) model, this link was not significant ($p = 0.496$). The three direct coefficients were estimated at 0.696, 0.626 and 0.362, against generating values of 0.700, 0.600 and 0.400. On the 800 held-out samples, one-step prediction errors were 39.2% and 40.9% lower than those of separate AR(1) models for temperature and pressure. The two steps are thus complementary: prewhitening corrects the position of the peak, and conditioning on the other measured variables removes associations that pass through an intermediate process stage. All of these results were obtained for a linear, stationary, feedforward process with Gaussian disturbances, a fixed sampling interval and a search window of k = 0, …, 10, and the 100% recovery rate applies to that setting only. Feedback control, nonlinear behaviour, variable transport times and unmeasured disturbances, which are common in operating plants, were not represented in the simulation. The next step is to apply the procedure to synchronised, timestamped measurements from a real automated process and to compare the estimated delays with values obtained from step tests or other planned interventions.

References

1. Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705
2. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control (5th ed.). John Wiley & Sons.

ISSN 3033-2559
OPEN ACCESS
DOI : https://doi.org/10.29114/ejtuv.vol1.iss3.89

3. Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366), 427–431. https://doi.org/10.1080/01621459.1979.10482531
4. Ge, Z., Song, Z., & Gao, F. (2013). Review of recent research on data-based process monitoring. Industrial & Engineering Chemistry Research, 52(10), 3543–3562. https://doi.org/10.1021/ie302069q
5. Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3), 424–438. https://doi.org/10.2307/1912791
6. Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. Biometrika, 65(2), 297–303. https://doi.org/10.1093/biomet/65.2.297
7. Lütkepohl, H. (2005). New Introduction to Multiple Time Series Analysis. Springer. https://doi.org/10.1007/978-3-540-27752-1
8. MacGregor, J. F., & Kourti, T. (1995). Statistical process control of multivariate processes. Control Engineering Practice, 3(3), 403–414. https://doi.org/10.1016/0967-0661(95)00014-L
9. Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. Annual Reviews in Control, 36(2), 220–234. https://doi.org/10.1016/j.arcontrol.2012.09.004
10. Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461–464. https://doi.org/10.1214/aos/1176344136