Autocorrelated non-normal data in control charts

Claudio G. Borroni, Manuela Cazzaro and Paola M. Chiodini

Abstract A common problem in control chart analyses is dealing with autocorrelated data. This problem is very often faced by fitting a suitable time-dependence model to data and by building the chosen control chart on its residuals based on the assumption that the stochastic process, of which the observed dataset is considered as a finite realization, is gaussian. In this paper the chance of dealing with autocorrelated non-gaussian data in control charts is analyzed. In particular, two cases will be considered: the case of uncorrectly modeled autocorrelated data and the case where time-dependence in data is disregarded.

1 Introduction

A common problem in SPC (Statistical Process Control) is dealing with autocorrelated data [4,5]. When a standard or a wrong analysis is applied, control charts can be highly unreliable. More specifically, [1] highlights that the rate of false positives (RFP), as measured by the conditional probability of a false out-of-control signal, is likely to increase due to unmodeled correlated data; similarly, a larger rate of false negatives (RFN) may be experienced. This problem is very often faced by fitting a suitable time-dependence model to data and by building the chosen control chart on its residuals. To get into details, the following steps are needed: i) identify the presence of dependence of data upon time; ii) identify a correct equation to model such a dependence; iii) estimate the parameters of such a model; iv) compute the estimated residuals and check if the dependence upon time has been eliminated; v) build the chosen control chart. To implement step i), the sample autocorrelation function (acf) is usually computed and the significance of its values is discussed up to a suitable lag. Such a task is usually accomplished by drawing suitable confidence bands on the acf graph. These bands are

[1] Claudio G. Borroni, Manuela Cazzaro, Paola M. Chiodini, Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali, Università degli Studi di Milano-Bicocca; email: claudio.borroni@unimib.it, manuela.cazzaro@unimib.it, paola.chiodini@unimib.it.

based, however, on the assumption that the stochastic process, of which the observed dataset is considered as a finite realization, is gaussian. The same assumption bases often steps ii) and iii): in the former the Box-Jenkins procedure is usually applied and a correct model is identified by comparing the observed sample acf with its theoretical counterpart corresponding to a suitable gaussian process in the ARMA class; in the latter, the gaussian assumption is used to compute the likelihood and to get the ML estimates of parameters (such estimates being however equivalent to the ones got from the OLS method under the same assumption). Finally, the normality of the process, or at least of the innovation term as below detailed, is often essential to step iv), where the significance of the estimated values of the residual acf is to be discussed.

In this paper the chance of dealing with autocorrelated non-gaussian data in control charts is analyzed.

2 Focus on the problem

To be consistent with more situations of non-normality, a general ARMA(p,q) model will be assumed for the process X_t which generated data: $\Phi_p(\mathrm{B}) \cdot X_t = \Theta_q(\mathrm{B}) \cdot \varepsilon_t$, $t = 1,2,\ldots$ where $\Phi_p(\mathrm{B}) = 1 - \phi_1\mathrm{B} - \phi_2\mathrm{B}^2 - \cdots - \phi_p\mathrm{B}^p$, $\Theta_q(\mathrm{B}) = 1 - \theta_1\mathrm{B} - \theta_2\mathrm{B}^2 - \cdots - \theta_q\mathrm{B}^q$ and B denotes the backward operator, i.e. $\mathrm{B} \cdot X_t = X_{t-1}$. The noise ε_t in model (1) is often called *innovation term*. Classically the ε_t's are assumed to be independently identically distributed as $N(0,\ \sigma^2)$; this assumption, along with suitable constraints on the values of the parameters of the model, let the process X_t be stationary gaussian. When the innovation term fails to be gaussian, the normality of X_t is not necessarily prejudiced as X_t can be often regarded as a linear combination of the ε_t's (like in the simple AR(1) case). Provided the latter are iid, the central limit works and X_t turns out to be marginally approximately normal. Of course this fact does not give strictly a Gaussian process, but the significance of the values of the sample acf can be meaningfully discussed. In the following, we will emphasize however that, due to the non-normality of the innovation term, some problem could still arise when the control chart is built on residuals of an incorrectly specified model. There are other cases where the non-normality of the innovation term causes X_t not to have a gaussian distribution, not even marginally or approximately. A simple example is when a MA(1) process with non-normal innovation is faced. In this chance, the significance of the estimated acf is harder to discuss and hence the dependence of data upon time is likely to be disregarded or at least incorrectly modeled. This chance will be further discussed.

It has to be emphasized that an unidentified time dependence of data, along with an incorrectly specified time-dependence model, can lead to meaningless control charts, as detailed in the Introduction. Consider first the case when the innovation term is non-gaussian, but X_t is marginally approximately normal. Despite step i) in the above-reported sequence poses specifically no problems, some difficulties could be experienced in the development of step ii). In [3], it is shown that, when the innovation is not normal, a correct identification of a model (even in the simpler AR class) needs to consider deeper tools than the acf, such as moments of the second and of the third order. More than the problem of misspecification, the non-normality of the error term could pose further problems in step v) above: due to the non-normality of the innovation, the estimated residuals are likely not to be gaussian. Hence the classical test to check if the dependence upon time has been eliminated, based on Pearson's product-

moment autocorrelation coefficient r, could give misleading results at least if the sample size is not very large. To avoid such drawbacks, one can revert to alternative measures of time-dependence of residuals, such as the serial version of Spearman's rho. A simple example will give some evidence of such a need. Suppose that the process which generated data is AR(2) where the random innovation terms are iid with non-normal distribution (Logistic, Laplace and Cauchy). Suppose further that the researcher erroneously estimates an AR(1) model and wants to test the autocorrelation of residuals. To compare the ability of different tests to detect the dependence upon time which still characterizes data, Tables 1-3 report the simulated powers of the 5%-level tests based on Spearman's rho and on Pearson's *r*. More specifically, 10000 samples of size 10 were generated after setting different values of the parameter ϕ_2 in the AR(2) model (and fixing the parameter ϕ_1 to 0.5) and the percents of a correct rejection of the hypothesis of randomness was computed, when a test based on the residuals of an estimated AR(1) model is applied. As expected, the test based on Spearman's rho gives a substantial gain of power, even if there are cases where both tests have relatively small powers. To evaluate the effect on the performance of the related Shewhart control chart, Tables 1-3 report the simulated RFPs and RFNs. Being aware of an unmodeled second-order component is likely to reduce such rates; hence, a non-standard analysis of residuals, with a consequent different specification of the time-dependence model, is here to be advised.

Table 1: Simulated powers, RFPs and RFNs when the innovation term has a Logistic distribution.

ϕ_2	1/2	1/4	1/8	1/16	1/32
Spearman	18.13	8.31	5.31	4.45	3.82
Pearson	16.39	6.26	3.56	2.60	2.21
RFP / RFN	0.57/55.99	0.58/41.85	0.55/39.52	0.40/32.37	0.74/28.94

Table 2: Simulated powers, RFPs and RFNs when the innovation term has a Laplace distribution.

ϕ_2	1/2	1/4	1/8	1/16	1/32
Spearman	13.89	4.71	2.99	2.42	2.36
Pearson	10.86	2.26	1.08	0.70	0.58
RFP / RFN	1.26/44.07	1.30/25.28	1.23/17.47	1.17/14.31	1.32/16.42

Table 3: Simulated powers, RFPs and RFNs when the innovation term has a Cauchy distribution.

ϕ_2	1/2	1/4	1/8	1/16	1/32
Spearman	23.46	8.88	5.64	4.38	4.02
Pearson	11.22	1.34	0.62	0.30	0.38
RFP / RFN	0.72/23.52	0.79/11.75	0.97/8.94	0.77/7.49	0.88/9.10

The above reported second case of non-normality in autocorrelated data, that is the one where X_t in not even approximately gaussian, is surely more delicate. The considered problem of an incorrect significance analysis of the estimated acf is here faced in the very beginning, at step i). This fact could probably affect the whole procedure and hence the final performance of the control chart is likely to be very poor. A natural way is to apply a suitable transformation to data so that it can be reduced to normality [2]. Finding a good transformation however is not an easy task, especially when the source of non-normality and the actual distribution of the process are not completely known. Moreover, the researcher could not realize the need for a control chart based on residuals. Consider the following example where the underlying process

is MA(1) with innovation terms following non-normal distribution (Logistic, Laplace and Cauchy). Again, 10000 datasets of size 10 were generated and the parameter θ_1 in the MA(1) model was fixed to different values. Tables 4-6 report the percents of a correct rejection of the hypothesis of randomness, based on Spearman's rho and on Pearson's r. These results show that a preliminary analysis of data uniquely based on r could result in a poor performance of a Shewhart control chart, as further evidenced by the reported values of the simulated RFPs and RPNs.

Table 4: Simulated powers, RFPs and RFNs when the innovation term has a Logistic distribution.

θ_1	1/2	1/4	1/8	1/16	1/32
Spearman	12.36	7.52	5.28	5.45	4.96
Pearson	11.22	5.62	3.20	2.65	2.08
RFP / RFN	0.57/55.93	0.25/30.13	0.08/15.71	0.10/12.11	0.04/4.82

Table 5: Simulated powers, RFPs and RFNs when the innovation term has a Laplace distribution.

θ_1	1/2	1/4	1/8	1/16	1/32
Spearman	12.95	7.72	6.25	5.41	4.87
Pearson	10.90	4.95	2.78	2.07	1.86
RFP / RFN	0.67/51.58	0.23/27.03	0.07/15.57	0.07/9.93	0.01/4.21

Table 6: Simulated powers, RFPs and RFNs when the innovation term has a Cauchy distribution.

θ_1	1/2	1/4	1/8	1/16	1/32
Spearman	13.17	9.65	7.10	5.93	5.56
Pearson	6.26	2.64	1.39	0.97	1.06
RFP / RFN	2.48/26.72	0.62/14.55	0.10/6.80	0.10/4.13	0.04/1.60

Some final comments can now be given. First, the settings of the reported simulations are surely limited: different values of the parameters of the ARMA model should be tried, along with a wider set of values for the sample size. Nonetheless, an important warning could be given: a good development of control charts should never neglect the use of alternative measures of time-dependence along with the usual r. This fact could increase the awareness of the researcher of deeper unmodeled time-dependencies in data and hence the need for more advanced tools of analysis. An extended discussion of such a need will be the object of a future paper.

References

1. Alwan, L.C.: Effects of autocorrelation on control chart performance. Commun. Stat. Theory, 21, 4, 1025—1049, (1992).
2. Kugiumtzis, D., Bora-Senta, E.: Gaussian analysis of non-gaussian time series. Brussels Economic Review, 53, 2, 295—322 (2010).
3. Kumar, K.: Identification of Arma models with non-gaussian innovations. Commun. Stat. Theory, 21, 4, 1145—1161, (1992).
4. Montgomery, D.C.: Introduction to Statistical Quality Control. Wiley, New York (1985).
5. Woodhall, W.H., Faltin, F.W.: Autocorrelated data and SPC. American Soc. Of Quality Control Statistics Div. Newsletter, 13, 4, 18—21, (1993).