Generalization of the mechanisms of cross-correlation analysis in the case of a multivariate time series

O Ja Kravets¹, G V Abramov², S Ju Beletskaja¹

¹ Department of automated and computer systems, Voronezh State Technical University, Voronezh, 394077 Russian Federation
² Department of Mathematics and Applied Analysis, Voronezh State University, Voronezh, 394077 Russian Federation

Abstract. The article describes a generalization of the mechanisms of cross-correlation analysis in the case of a multivariate time series and how this allows the optimal lags to be identified for each of the independent variables (IV) using a number of algorithms. The use of generalized mechanisms will allow variables to be analysed and predicted based on the retrospective analysis of multidimensional data. In the available literature, cross-correlation has been defined only for pairs of time series. However, the study of dependent variable (DV) dependencies on multidimensional independent variables that takes into account the vector of specially selected time lags will significantly improve the quality of models based on multiple regression. The idea of multiple cross-correlation lies in the sequential forward shift of each IV row with respect to DV (it transpires that DV is delayed relative to IV) until we obtain a minimum error or the best test of multiple regression. After the completion of all stages of multiple cross-correlation, the synthesis of the model is not a difficult process.

Introduction
The article describes a generalization of the mechanisms of cross-correlation analysis in the case of a multivariate time series and how this allows the optimal lags to be identified for each of the independent variables using a number of algorithms.

The use of the generalized mechanisms will allow variables to be analysed and predicted based on the retrospective analysis of multidimensional data.

Initial data
There are M time series \(X_1, X_2, \ldots, X_M\) each of N observations. There is a study variable \(Y\) - the same time series with parameters, depending, as we believe, on a series of observations.

Multiple regression
The purpose of multiple regression [1, 5] is to provide (1) a calculation of the independent effect of changes in the values of each dependent variable (DV) on an independent variable (IV) and (2) on an empirical basis to predict the values of a dependent variable on the basis of knowledge of the joint effect of IV. The analysis starts by writing an equation that accurately describes the investigated causal relationship. Since this equation can be viewed as a process model, this step is regarded as the developing of the model. The general formula of multiple regression is as follows:

\[ Y' = a_0 + b_1X_1 + b_2X_2 + \ldots + b_MX_M + e. \] (1)

Here the unknown coefficients \(a_0, b_1, b_2, \ldots, b_M\) are evaluated in such a way as to minimize the average value of error \(e\). In order to ensure that the application of the method of multiple regression is successful, it is necessary for the model and data to meet five requirements which provide the basis of the use of regression [4, 7].

1. The model must accurately describe the real relationships being studied. This requires that (a) the relation between the variables is linear, (b) that none of the important independent variables are excluded and (c) that no irrelevant variables are included.
2. There should be no errors in the measurement of variables.
3. As for the error - the following conditions are required:
4. Variables must be measured on an interval scale.
   a) the mean (the estimated value for each observation) is 0;
   b) errors for each observation are not correlated,
   c) IVs are not correlated with the error;
   d) the deviation of the error remains constant for all values of IV; this condition is called homoscedasticity;
   e) the error has a normal distribution.
5. None of the IVs are correlated clearly with any other IV, or with any linear combination of other IVs. If so, then we can say that there is no clear multicollinearity.

If the study adequately satisfies these conditions, we can substitute specific values for \( Y', \) \( X_i \) and solve the regression equation which describes assumptions about the unknown values of \( a_0, b_i, \) using, for example, the least-squares procedure or standard statistical packages.

**Interpretation of the results of multiple regression**

Consider the coefficient of multiple determination, or \( R^2 \) [9]:

\[
R^2 = \frac{\sum(Y' - \overline{Y})^2}{\sum(Y - \overline{Y})^2} = \frac{\text{regression sum of squares}}{\text{initial sum of squares}}. \tag{2}
\]

This ratio indicates how closely the points that represent data are located to the "straight" line \( \overline{Y} \) provided by the model; it is usually called the measure of deviations of DV, which can be explained by the oscillation of all IVs. \( R^2 \) varies between 0 and 1; the closer it is to one, the more principal the model is. The \( R^2 \) value can always be increased by introducing additional IVs to the model.

**Cross-correlation and time lag. The idea of multiple cross-correlation**

Sometimes we find that one event affects the other only after some time. The value of this time is called the time lag \( \Delta \) and it can be determined by using a cross-correlation value [1, 8, 9], defined for two functions \( Z(t) \) and \( V(t) \) as

\[
R_\Delta = \frac{\int Z(t)V(t+\Delta)dt}{\sqrt{\int Z(t)dt} \sqrt{\int V(t+\Delta)dt}} \leq 1. \tag{3}
\]

The closer \(|R_\Delta|\) is to 1, the better the optimal time lag \( \Delta \) adjusts the formula model of the pair regression.

In the available literature, cross-correlation has been defined only for pairs of time series [3, 6, 7, 10]. However, the study of DV dependencies on multidimensional IVs that takes into account the vector of specially selected time lags \( \Delta = (\Delta_1, \Delta_2, \ldots, \Delta_M) \) will significantly improve the quality of models based on multiple regression. Fig. 1 shows the idea of multiple cross-correlation. Its essence lies in the sequential forward shift of each IV row with respect to DV (it transpires that DV is delayed relative to IV) until we obtain a minimum error or the best test of multiple regression. The maximum value of the time lag \( \Delta \) for a specific IV should not exceed \( N/2 \), otherwise, there will be difficulties in filling the released elements in the vector of IV.
Possible qualitative behaviours of the criterion are shown in Fig. 2-6. Fig. 2 shows the growth of the criterion, which may indicate a big lag, its belonging to a quasi-periodic process with a long period as a fragment of Fig. 3, or the ascending branch of the parabola (Fig. 5-6). In any case, in this situation we are not able to achieve the optimum value $\Delta$.

Fig. 3 corresponds to the quasiperiodic behaviour of the criterion – we can see the minimum value $\Delta$ corresponding to the best criteria.

Fig. 4 reflects the situation of the local reduction of the criterion. As with Fig. 2, it is difficult to obtain the optimum value of $\Delta$ without additional information about the behaviour of the time lag - we have a situation of negative or zero time lag. Perhaps we are seeing only a part of a parabola or a batch process - but the sample size is too small for definitive conclusions.

Fig. 5 reflects the obvious maximum, Fig. 6 indicates an inefficient sample size or belonging to a quasi-periodic case with a long period. See also the legend to Fig. 2 and 4.
Figure 3. "Found" in the period.

Figure 4. The scale is too small.
That is what we were looking for.

The situations of criterion behaviour shown in Fig. 2, 3 and 5 are useful for practical purposes.

**Tactics for selecting the next (the first) variable**

The beginning of the process of determining the optimal time lags of the multivariate model may be different and it is not yet possible to predict which of them is the most optimal one. Here are the main ones:

1. The maximum of the local criterion – we compute the multiple regression parameters and select the IV with the best explanatory properties. Then, for this particular variable, we undertake the procedure of finding the optimal $\Delta$.
2. The minimum of the local criterion (IV with the worst explanatory properties).
3. The random selection of the IV.

After the procedure of determining the optimal $\Delta$ for a selected variable, there are two possible tactics to continue the process.

3.1. Reduction of the set of IVs
   - searching for paired cross-correlation and remembering $b_i x_i$,
   - subtracting $b_i x_i$ and the average explanatory error from the index,
   - deleting IV $i$ and repeating the search-cycle (as long as there are variables that make a significant contribution to the model).

3.2. Saving of the processed IV
   - adding IV $i$ to the list of variables prohibited for later selection (initially the list is empty);
   - repeating the search-cycle (as long as there are variables that make a significant contribution to the model which are not in the list of variables prohibited for later selection).

An outline flowchart for the process of determining the vector of time lags $\Delta=(\Delta_1, \Delta_2, \ldots, \Delta_M)$ is shown in Fig. 7.
Tactics of backward extrapolation

In the case of saving the processed independent variable with an index i (3.2 tactics to continue the process) after the determining of the next element $\Delta_i$ of time lags vector $\Delta=\left(\Delta_1, \Delta_2, ..., \Delta_M\right)$ immediately before the following multiple regression analysis of the resulting model, which will be the next step, it is necessary to fill the elements of the vector $X_i$ for the accuracy of the following analysis - but the filling must be directed "upward" or "leftward" with respect to the vector tail - in the elements with indices $0, -1, -2, ..., -\Delta_i+1$. We call this process - "backward extrapolation". The complexity of backward extrapolation is not less than the complexity of the traditional method [2], and it is therefore proposed to use one of the following tactics while filling the elements of the vector $X_i$:

1. zero filling,
2. filling with the mean value of the vector $X_i$,
3. the use of the "moving average" method,
4. the complete analysis of the time series $X_i$ as a separate object.

Since $X_i$ is already processed, it does not appear necessary to carry out a scrupulous analysis of its autonomous behaviour and in most cases we can use tactics 2 and 3 only.
Figure 7. Outline flowchart for the process of determining the vector of time lags $\Delta = (\Delta_1, \Delta_2, \ldots, \Delta_M)$. 

- **Determining the IV selection tactics for optimal time lag search**
  - TV=1 ➔ Local Maximum ➔ IV ranking ➔ $\Omega := \emptyset$
  - TV=2 ➔ Local Minimum ➔ IV ranking ➔ $\Omega := \emptyset$
  - TV=3 ➔ Random Selection ➔ IV ranking ➔ $\Omega := \emptyset$

- **Selecting the next $X_i \in \Omega$ using TV tactics**
  - $\delta := 0$; Criterion := 0; $\Delta_i := 0$

- **Cross-correlational analysis of the pair $Y^*[\delta], X^*[\delta+\delta]$**
  - Yes ➔ Criterion $\Delta_i$ ≤ Criterion?
    - Yes ➔ Criterion := Criterion $\Delta_i$
    - No ➔ $\delta := \delta + 1$
  - No ➔ $\delta > [N/2]$?
    - Yes ➔选用战术继续过程
    - No ➔ Reduction

  **Cross-correlational analysis of the pair $Y^*[\Delta], X^*[\Delta+\Delta]$**

  **Saving**
  - Backward extrapolating $X_i$ to $\Delta_i$
  - Excluding the contribution of $X_i$ from $Y$
  - $\Omega := \Omega \cup X_i$

  **Time lags vector is constructed** $\Delta = (\Delta_1, \Delta_2, \ldots, \Delta_M)$

- **No** ➔ {X_i} $\cup \Omega = \emptyset$?
  - Yes ➔ **Time lags vector is constructed** $\Delta = (\Delta_1, \Delta_2, \ldots, \Delta_M)$.
Synthesis of the model

After the completion of all the stages of multiple cross-correlation, the synthesis of the model is not a difficult process. We select

$$\Delta_{\text{max}} = \max \{ \Delta_1, \Delta_2, \ldots, \Delta_M \},$$

then the initial set of $X_i$ is shifted with respect to $Y$ into accordance with $\Delta_i$.

From the resulting set of variables – the independent "shifted" $X_i$ and the dependent $Y$ – we construct a matrix for the final multiple regression analysis in accordance with Fig. 8. The idea of constructing this is to shorten the overall time series by $\Delta_{\text{max}}$ taking into account the "virtual" filled-in values.

![Diagram](image)

**Figure 8.** The idea of forming a final matrix of original data: $\Delta_1=0, \Delta_2=1, \Delta_3=3, \ldots, \Delta_{M-1}=3; \Delta_M=2; \Delta_{\text{max}}=3$.

Conclusion

Using the reduced data matrix which is constructed in such a way, multiple regression is carried out and a linear model with a high degree of adequacy is formed.

Thus, the considered idea can be used for detection of difficult multiple-factor dependences of one time series from several ones. At the same time it is possible to set optimum temporal lags. It is important for detection of the postponed response and prediction at least in medicine and ecology.

References

[1] Eliseeva I I, Yuzbashev M M 2005 General Theory of Statistics 656 p
[2] Lukashin Y P 2005 Adaptive methods of short-term time series forecasting 416 p
[3] Mannheim J B, Rich R K 1997 Political science. Research Methods 544 p
[4] Mkhitaryan V S 2005 Statistics 671 p
[5] Bohrnstedt G W, Knoke D 1982 Statistics for Social Data Analysis
[6] Couch J V 1982 Fundamentals of Statistics for the Behavioral Sciences
[7] Kachigan S K 1982 *Multivariate Statistical Analysis*
[8] Lindeman R, Merenda P P, Gold R Z 1980 *Introduction to Bivariate Multivariate Analysis*
[9] Moore D S 1991 *Statistics*
[10] Tabachnick B G, Fidell L S 1983 *Using Multivariate Statistics*