УДК 519.86

ПРИМЕНЕНИЕ МЕТОДА МНОГОСТУПЕНЧАТОЙ ИДЕНТИФИКАЦИИ В ФАКТОРНОМ АНАЛИЗЕ

Дургарян И.С., Лясковская И.В., Пащенко А.Ф., Пащенко Ф.Ф.

Институт проблем управления им. В.А. Трапезникова РАН, Москва, e-mail: feodor@ipu.ru

Использование корреляционных функций для нелинейных стохастических систем, а также для систем, на входах которых действуют сигналы с нелинейной структурой, часто не приводит к желательным результатам. Корреляционные функции не являются исчерпывающими характеристиками связи между случайными процессами и могут обращаться в нуль даже тогда, когда существует детерминированная зависимость между входным и выходным процессами системы. Работа посвящена состоятельному методу, основанному на использовании обобщенных корреляционных и дисперсионных функций. В работе рассматривается статический многомерный объект, выходная переменная которого зависит от вектора наблюдаемых входных факторов и вектора ненаблюдаемых или наблюдаемых с большим запаздыванием входных факторов. В постановке задачи и при выводе конечных результатов предполагалось, что прогнозируемые входные факторы зависят от разных векторов косвенных показателей.

Ключевые слова: предикторные факторы, корреляционный момент, дисперсионные функции, прогнозируемые факторы

APPLICATION OF THE MULTISTAGE IDENTIFICATION METHOD IN THE FACTOR ANALYSIS

Durgaryan I.S., Lyaskovskaya I.V., Pashchenko A.F., Pashchenko F.F.

Institute of Control Sciences, Moscow, e-mail: feodor@ipu.ru

Use of correlation functions for nonlinear stochastic systems, and also for systems on inputs of which signals with nonlinear structure are acting, often doesn't lead to desirable results. Correlation functions are not exhaustive characteristics of dependence between random processes and can turn into zero even when there is a determined dependence between input and output system's processes. Paper is devoted to the consistent method based on use of generalized correlation and variance functions. In current work the static multidimensional object is considered, output variable of which depends on a vector of observable input factors and a vector of unobservable or observable with large delay input factors. In a problem definition and at a conclusion of the final results it was supposed that predicted input factors depend on different vectors of indirect indicators.

Keywords: predictive factors, correlation moment, variance functions, predicted factors

Известно, что для нелинейных стохастических систем, а также для систем, на входах которых действуют сигналы с нелинейной структурой, использование корреляционных функций часто не приводит к желательным результатам, поскольку корреляционные функции не являются исчерпывающими характеристиками связи между случайными процессами [4, 5] и могут обращаться в нуль даже тогда, когда существует детерминированная зависимость между входным и выходным процессами системы.

Для устранения негативных явлений, возникающих в этих случаях, предлагается использовать аппарат дисперсионных функций [4, 5].

Однако, как показано в [2, 3], дисперсионные меры связи хотя и являются более мощным статистическим аппаратом, чем корреляционные функции, также как и корреляционные функции не являются состоятельными мерами связи между случайными процессами. Поэтому в работе предлагается состоятельный метод, основанный на использовании обобщенных корреляционных функций и функциональной корреляции и идеях статистической линеаризации.

Метод многоступенчатой идентификации

Рассмотрим линейный статический многомерный объект, выходная переменная Y которого зависит от вектора наблюдаемых входных факторов $\overline{X} = (X_1,..., X_n)$ и вектора ненаблюдаемых или наблюдаемых с большим запаздыванием входных факторов $Z = (Z_1, ..., Z_m)$. Согласно поставленной задаче, будем предполагать, что значения ненаблюдаемых входных факторов Z_1, ..., Z_m, соответствующие синхронным значениям сигнала на выходе объекта, достаточно хорошо представляются в виде некоторых функций от наборов косвенных показателей или же процессами авторегрессии.

Ввиду того, что процесс авторегрессии является частным случаем регрессии одного случайного процесса относительно других случайных процессов, будем считать в дальнейшем, что значения ненаблюдаемых входных переменных представимы в виде функций от некоторых наборов наблюдаемых косвенных показателей.

Как известно, наилучшим приближением зависимой случайной величины через независимые переменные в смысле кри-

терия минимума средней квадратической ошибки является условное математическое ожидание. Поэтому будем полагать, что ненаблюдаемые (наблюдаемые с запаздыванием) входы Z_1, ..., Z_m достаточно хорошо представляются своими условными математическими ожиданиями относительно векторов косвенных показателей $\overline{\xi}_1,...,\overline{\xi}_m$ т.е.

$$\hat{Z}_i = M\left\{ Z_i / \xi_i,...,\xi_{i\,k_i} \right\}. \qquad (1)$$

Ограничения типа линейности на регрессию $M\left\{Z_i / \xi_i,...,\xi_{i\,k_i}\right\}$ не накладываются.

Ненаблюдаемые входы Z_1, ..., Z_m, прогнозируемые с помощью уравнений (1), будем называть факторными переменными, а уравнения (1) – промежуточными факторами. Факторы Z_1, ..., Z_m, используя терминологию предикторного управления, можно называть предикторными факторами.

Частными случаями уравнения (1) являются уравнения линейной регрессии:

$$\hat{Z}_i = \sum_{i=1}^{k} a_i \xi_i.$$

Уравнения регрессий (1) можно получить, используя известные методы.

Выбор наборов косвенных переменных для прогноза соответствующих ненаблюдаемых параметров осуществляется на основе алгоритмов выбора информативных переменных методов факторного анализа. В наборы косвенных переменных могут входить и наблюдаемые входные факторы – $\overline{X}$ и наоборот.

Уравнение основной математической модели для прогнозирования выходной переменной объекта будем искать в классе линейных моделей вида

$$\hat{Y} = b_0 + \sum_{i=1}^{n} b_i X_i + \sum_{i=1}^{m} b_{i+n} \hat{Z}_i, \qquad (2)$$

где b_0, b_i, $i = 1, ..., n, ..., n + m$ – неизвестные параметры.

При решении практических задач во многих случаях удобнее пользоваться нормированными статистическими характеристиками анализируемых случайных величин и процессов. При этом упрощаются вычисления и становится более наглядным анализ влияния отдельных входных факторов на прогнозируемую выходную величину.

Выразим все переменные и зависимости между ними в стандартизованном масштабе по формулам

$$t_y = \frac{Y - M\{y\}}{\sigma_y};$$

$$t_i = \begin{cases} \dfrac{X_i - M\{X_i\}}{\sigma_{x_i}}, & I = 1,...,n \\ \dfrac{\hat{Z}_i - M\left\{\hat{Z}_i\right\}}{\sigma_z}, & i = n+1,...,n+m. \end{cases} \qquad (3)$$

При этом уравнение модели (2) примет вид

$$\hat{t}_y = \sum_{i=1}^{n+m} \tilde{\beta}_i t_i, \qquad (4)$$

где $\tilde{\beta}_i$ – коэффициенты стандартизированной модели – находятся из условия квадратичного минимума функционала

$$J = M\left\{ \left[\sum_{i=1}^{n+m} \tilde{\beta}_i t_i - t_y \right]^2 \right\}, \qquad (5)$$

которое приводит к системе из $n + m$ линейных уравнений относительно $n + m$ неизвестных параметров модели (4)

$$\begin{gathered}
\tilde{\beta}_1 + ... + \tilde{\beta}_n \rho_{x_1 x_n} + \tilde{\beta}_{n+1} \eta_{x_1 z_1 \overline{\xi}_1} + ... + \tilde{\beta}_{m+n} \eta_{x_1 z_m \overline{\xi}_m} = \rho_{y x_1}; \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\tilde{\beta}_1 \rho_{x_n x_1} + ... + \tilde{\beta}_n + \tilde{\beta}_{n+1} \eta_{x_n z_1 \overline{\xi}_1} + ... + \tilde{\beta}_{m+n} \eta_{x_n z_m \overline{\xi}_m} = \rho_{y x_n}; \\
\tilde{\beta}_1 \eta_{z_1 x_1 \overline{\xi}_1} + ... + \tilde{\beta}_n \eta_{z_1 x_1 \overline{\xi}_1} + \tilde{\beta}_{n+1} \eta^2_{z_1 \overline{\xi}_1} + ... + \tilde{\beta}_{n+1} \eta_{z_1 z_m \overline{\xi}_1 \overline{\xi}_m} = \eta_{y z_1 \overline{\xi}_1}; \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\tilde{\beta}_1 \eta_{z_m x_1 \overline{\xi}_m} + ... + \tilde{\beta}_n \eta_{z_m x_n \overline{\xi}_m} + \tilde{\beta}_{n+1} \eta_{z_m z_1 \overline{\xi}_m \overline{\xi}_1} + ... + \tilde{\beta}_{n+m} \eta^2_{z_m \overline{\xi}_m} = \eta_{y z_m \overline{\xi}_m},
\end{gathered} \qquad (6)$$

где $\rho_{x_i x_j} = \operatorname{cov}\left(X_i, X_j\right) \Big/ \sigma_i \sigma_j$, $i, j = 1, ..., n$ – коэффициент корреляции между случайными величинами X_i, X_j; $\rho_{x_i y}$ – коэффициент корреляции между X_i и Y;

$$\eta^2_{z_i \overline{\xi}_i} = \frac{\theta_{z_i \overline{\xi}_i}}{\sigma^2_{z_i}}; \quad \eta_{x_i z_j \overline{\xi}_j} = \frac{\theta_{x_i z_j \overline{\xi}_j}}{\sigma_{x_i} \cdot \sigma_{z_j}};$$

$$\eta_{y z_j \overline{\xi}_j} = \frac{\theta_{y z_j \overline{\xi}_j}}{\sigma_y \cdot \sigma_{z_j}}; \quad \eta_{z_i z_j \overline{\xi}_i \overline{\xi}_j} = \frac{\theta_{z_i z_j \overline{\xi}_i \overline{\xi}_j}}{\sigma_{z_i} \cdot \sigma_{z_j}}$$

– нормированные значения соответствующих дисперсионных функций;

$K_{x_i x_j} = \mathrm{cov}\left(X_i, X_j\right)$ – корреляционный момент случайных величин X_i и X_j, а – корреляционный момент между сигналами на выходе и i-м входе объекта; $K_{yx_i} = \mathrm{cov}\left(Y, X_i\right)$; $\theta_{z_i\overline{\xi}_i}, \theta_{z_i x_j \overline{\xi}_i}, \theta_{z_i z_j \overline{\xi}_i \overline{\xi}_j}$ – различные типы дисперсионных функций (моментов).

$$\theta_{z_i\overline{\xi}_i} = M\left\{\left[M\left\{Z_i / \overline{\xi}_i\right\} - M\left\{Z_i\right\}\right]^2\right\};$$

$$\theta_{z_i x_j \overline{\xi}_i} = M\left\{\left[M\left\{Z_i / \overline{\xi}_i\right\} - M\left\{Z_i\right\}\right]\left[x_j - M\left\{X_j\right\}\right]\right\};$$

$$\theta_{z_i z_j \overline{\xi}_i \overline{\xi}_j} = M\left\{\left[M\left\{Z_i / \overline{\xi}_i\right\} - M\left\{Z_i\right\}\right]\left[M\left\{Z_j / \overline{\xi}_j\right\} - M\left\{Z_j\right\}\right]\right\}.$$

Решение системы (6) может быть записано в виде

$$\tilde{\beta}_i = \frac{\tilde{\Delta}_i}{\tilde{\Delta}}, \quad i = 1, ..., n + m, \qquad (7)$$

где

$$\tilde{\Delta} = \det \left\| \begin{array}{c|c} K & \theta \\ \hline \theta^T & \theta^* \end{array} \right\| \qquad (8)$$

– определитель системы (6); $\tilde{\Delta}_i$ – определитель, получающийся из $\tilde{\Delta}$ заменой в нем соответствующего столбца столбцом свободных членов системы (6).

В (8) T – знак транспонирования, а матрицы K, θ и θ^* равны

$$K = \left\| \begin{array}{cccc} 1 & \rho_{x_1x_2} & ... & \rho_{x_1x_n} \\ & ..... & & \\ \rho_{x_nx_1} & \rho_{x_nx_2} & ... & 1 \end{array} \right\|;$$

$$\theta = \left\| \begin{array}{ccc} \eta_{x_1 z_1 \overline{\xi}_1} & ... & \eta_{x_1 z_m \overline{\xi}_m} \\ ..... & & \\ \eta_{x_n z_1 \overline{\xi}_1} & ... & \eta_{x_n z_m \overline{\xi}_m} \end{array} \right\|;$$

$$\theta^* = \left\| \begin{array}{ccc} \eta^2_{z_1\overline{\xi}_1} & ... & \eta_{z_1 z_m \overline{\xi}_1 \overline{\xi}_m} \\ ..... & & \\ \eta_{z_m z_1 \overline{\xi}_m \overline{\xi}_1} & ... & \eta_{z_m \overline{\xi}_m} \end{array} \right\|.$$

При записи определителя $\tilde{\Delta}$ в форме (8) учтено то обстоятельство, что матрицы системы уравнений (6), а также матрицы K и θ^* являются симметричными. Этот факт легко следует из определений и свойств корреляционных и дисперсионных функций.

Следует заметить, что матрица системы (6) отличается от корреляционной матрицы системы нормальных уравнений, получающейся в результате применения МНК к стандартной задаче идентификации тем, что ее элементами являются не только коэффициенты корреляции, но и нормированные дисперсионные функции, а на главной диагонали, кроме единиц, стоят элементы $\eta^2_{z_i\overline{\xi}_i} \le 1$. Дело в том, что нормированная взаимная дисперсионная функция $\eta_{z\overline{\xi}}$ равна единице в том и только том случае, когда между случайными величинами Z и $\overline{\xi} = \left(\xi_{1,...,}\xi_k\right)$ существует точная функциональная зависимость. Очевидно, что при решении практических задач надо стремиться к тому, чтобы мера определенности прогноза случайной величины Z при помощи набора косвенных показателей ξ_1, ..., ξ_k была близка к единице, т.е. $\eta_{z\overline{\xi}} \approx 1$.

Используя матричные обозначения, уравнение модели представим в виде

$$Y_M = VB, \qquad (9)$$

где $Y_M = \left\| \begin{array}{c} Y_{M_1} \\ . \\ . \\ . \\ . \\ . \\ Y_{M_N} \end{array} \right\|$ – $[N\times1]$ матрица значений выходной переменной модели; N – число наблюдений;

$B = \left\| \begin{array}{c} b_1 \\ . \\ . \\ . \\ . \\ . \\ b_{n+m} \end{array} \right\|$ – $[(n + m)\times1]$ матрица параметров модели;

$V = \| X \mid \hat{Z} \|$ – $[N\times(n+m)]$ блочная матрица наблюдаемых и прогнозируемых значений входных сигналов;

$X = \|X_{ij}\|$ – $[N\times n]$ матрица значений наблюдаемых факторов $X_1, ..., X_n$;

$\hat{Z} = \|\hat{Z}_{ij}\|$ – $[N + m]$ матрица прогнозируемых значений входных факторов $Z_1, ..., Z_m$.

Функционал (4) можно записать в виде

$$J = E^T E, \qquad (10)$$

где $E = \|e_i\|$ – $[N\times 1]$матрица невязок,

$$E = Y_M - Y.$$

Минимизируя (10) по всем компонентам вектора параметров B и используя при этом стандартную процедуру минимизации квадратичного функционала, получим уравнение для определения вектора параметров модели (9).

$$V^T V B = V^T Y, \qquad (11)$$

где

$$Y = \begin{Vmatrix} Y_1 \\ . \\ . \\ . \\ . \\ . \\ Y_N \end{Vmatrix}$$ – $[N\times 1]$ матрица значений выходной переменной модели; N – число наблюдений; T – знак транспонирования.

Решение матричного уравнения (11) в предположении невырожденности матрицы $(V^T V)$ имеет вид

$$B = (V^T V)^{-1} V^T Y. \qquad (12)$$

Предикторные переменные в факторном анализе

В классическом факторном анализе основным предположением связи переменных является равенство

$$X = LF + E, \qquad (13)$$

где X – вектор-столбец наблюдаемых переменных размерности $p\times 1$; L – $p\times k$ матрица факторных нагрузок; F – $k\times 1$ вектор-столбец факторов $(k < p)$; E – $p\times 1$ вектор-столбец остатков, которые предполагаются независимыми как между собой, так и с факторами. Дисперсии остатков (или остаточные дисперсии) образуют матрицу V.

Уравнение (13) постулирует основные предположения факторного анализа о том, что множество наблюдаемых коррелированных переменных X, которые подчиняются многомерному нормальному распределению с корреляционной матрицей C размерности $p\times p$, можно описать меньшим числом гипотетических переменных или факторов F и множеством независимых остатков E.

Рассмотрим модель объекта с выходом Y и входом $X = (X_1, ..., X_p)$. Если p велико, возникает желание уменьшить размерность модели, выразив ее входы через меньшее количество $k < p$ некоторых переменных F. Таким образом, получаем схему факторного анализа. Построение модели Y непосредственно по переменным F невозможно, т.к. они являются гипотетическими (ненаблюдаемыми). Однако эти переменные могут быть выражены через наблюдаемые переменные X следующим образом: $\hat{F} = L^T C^{-1} X$ – для некоррелированных факторов и $\hat{F} = P L^T C^{-1} X$ – для коррелированных факторов, где P – оцененная корреляционная матрица факторов.

Модель объекта будем искать в виде

$$\hat{Y} = \hat{F}^T B, \qquad (14)$$

где B – вектор-столбец неизвестных коэффициентов размерности $k\times 1$. Коэффициенты вектора определим из условия минимума среднеквадратического критерия, т.е. таким образом, чтобы функционал $J = M\left\{\left(\hat{Y} - Y\right)^2\right\}$ принимал минимальное значение.

Подставляя (14) в $J = M\left\{\left(\hat{Y} - Y\right)^2\right\}$ и дифференцируя полученное выражение по B, придем к уравнению

$$M\left\{\hat{F}\left(Y - \hat{F}^T B\right)\right\} = 0. \qquad (15)$$

Решая (15) с учетом $\hat{F} = L^T C^{-1} X$, получим

$$B = \left(Q K_{xx} Q^T\right)^{-1} Q K_{xy}, \qquad (16)$$

где $Q = L^T C^{-1}$ – матрица размерности $k\times p$, а матрицы K_{xx} и K_{xy} определяются соответственно формулами:

$$K_{xx} = M\left\{XX^T\right\}; \quad K_{xy} = M\left\{XY\right\}. \qquad (17)$$

Для коррелированных факторов получим

$$B = \left(PQK_{xx}Q^T P^T\right)^{-1} PQK_{xy}. \qquad (18)$$

Следует отметить, что (16) и (18) получены при условии линейной связи между факторами и входными переменными. Если эта связь нелинейна, то в (16) и (18) вместо (17) будут входить матрицы, элементами которых являются дисперсионные функции.

Заключение

Следует заметить, что в постановке задачи и при выводе конечных результатов предполагалось, что прогнозируемые входные факторы зависят от разных векторов косвенных показателей. В частном случае ненаблюдаемые входные сигналы могут определяться одним и тем же набором косвенных факторов.

Предложенный метод использовался для моделирования загрязнения и хронических заболеваний в Хабаровском крае [1].

Список литературы

1. Голяк И.В. Метод главных компонент в оценке влияния микроэлементного состава почв на распространение злокачественных новообразований в Хабаровском крае // Вестник Академии информатика, экология, экономика. – Т. 11. Ч. I. – М.: МАСИ, 2008. – С. 116–118.

2. Ефременко Ф.В., Пащенко А.Ф. О выборе информативных переменных в задаче структурной идентификации // Математические методы в технике и технологиях – ММТТ-20: сборник трудов XX Международной научной конференции. Т.7. – Ярославль: Изд-во Яросл. гос. техн. ун-та, 2007. – С. 236–239.

3. Пащенко Ф.Ф. Введение в состоятельные методы моделирования систем. – Ч.1. Математические основы моделирования систем. – М.: Финансы и статистика, 2006. – 328 с.

4. Durgarjan I.S., Pashchenko F.F. Non-parametric identification of nonlinear systems // Proc. of the 7-th IFAC/IFORS Symposium on Identification and system Parameters Estimation. – New-York Pergamon Press, 1985. – Vol 1. – P. 433–437.

5. Durgaryan I.S., Pashchenko F.F. Information methods in indentification // Trans. Of Ninth Prague Conference on Information theory, statistical decision functions, random processes. – Prague 1982. Czechoslovak academy of Sciences. Prague, 1983. – P. 207–214.

References

1. Golyak I.V. *Vestnik Mezhdunarodnoy akademii sistemnyh issledovanij. Informatika, ekologija, ekonomika* – Journal of the International academy of system studies. Moscow, 2008. Vol. 11, Part I, pp. 116–118.

2. Efremenko Ph.V, Pashchenko A.F. *Sbornyk trudov XX Mezhdunarodnoj konferencii «Matematicheskie metody v tehnike i tehnologiyah* (Proc. XX-th Int.conf.» Mathematical methods in science and technologies). Jaroslavl, 2007, Vol. 7, pp. 236–239.

3. Pashchenko F.F. *Vvedeniye v sostoyatelnye metody modelirovaniya system.* Ch.*1- Matematicheskie osnovy modelirovaniya system* [Introduction into consistent methods of systems modeling. P.1 Mathematical foundations of systems modeling] Moscow, Financy i Statistica, 2006.

4. Durgarjan I.S., Pashchenko F.F. Non-parametric identification of nonlinear systems // Proc. of the 7-th IFAC/IFORS Symposium on Identification and system Parameters Estimation. New-York Pergamon Press. 1985 Vol 1. pp. 433–437.

5. Durgaryan I.S., Pashchenko F.F. Information methods in indentification // Trans. Of Ninth Prague Conference on Information theory, statistical decision functions, random processes. Prague 1982. Czechoslovak academy of Sciences. Prague, 1983. pp. 207–214.

Рецензенты:

Гордеев Л.С., д.т.н., профессор кафедры «Кибернетика химико-технологических процессов» Российского химико-технологического университета им Д.И. Менделеева, г. Москва;

Комиссаров Ю.А., д.т.н., профессор, зав. кафедрой «Электротехника и электроника» Российского химико-технологического университета им Д.И. Менделеева, г. Москва.

Работа поступила в редакцию 06.08.2013.