

ESTATÍSTICA MULTIVARIADA APLICADA AO ESTUDO DE VARIÁVEIS MACROECONÔMICAS

Maria Emilia Camargo e Adriano Mendonça Souza

Departamento de Estatística - PPGMQ - Centro de Ciências Naturais e Exatas

UFSM - Santa Maria, RS

RESUMO

Este trabalho tem por objetivo, além de expor de forma estruturada a metodologia de Componentes Principais, mostrar a sua aplicabilidade em reduzir a dimensionalidade do número de variáveis sem perder as informações contidas no conjunto de dados originais. Foi realizada uma análise no período de janeiro de 1985 a dezembro de 1991, das variáveis macroeconômicas, Política Monetária, Política de Preços, Atividade Econômica, Agregados de Créditos e da Balança Comercial. Com os modelos ajustados através da metodologia de séries temporais, fez-se previsões para o período de janeiro a dezembro de 1992.

ABSTRACT

The objective of this research is to explain the methodology of principal components in a structured way, by showing its applicability of reducing the dimensionality of the number of variables without losing the originality of the set. An analysis was made of the economic variables which represent Monetary and Price Policies, Economical Activities, Investment Variables and the current account of the Balance of Payments, from January 1985 to December 1991. A forecast was made for the period from January to December of 1992, using the adjusted models obtained from Principal Components and Box-Jenkins methodologies.

1- INTRODUÇÃO

Apresenta-se neste trabalho, a Análise de Componentes Principais aplicada ao estudo de variáveis macroeconômicas (MORRISON, 1976). A Análise de Componentes Principais vem se tornando cada vez mais utilizada na formação de modelos macroeconômicos sem perder a originalidade do conjunto de dados (PINTO, 1981). Com a finalidade de analisar variáveis macroeconômicas brasileiras foram utilizadas as metodologias de Componentes Principais e de Séries Temporais para futuras previsões (BOX & JENKINS, 1976). As variáveis analisadas estão compreendidas no período amostral de janeiro de 1985 a dezembro de 1991.

A organização do trabalho é a seguinte: na seção 2, apresenta-se a metodologia de Componentes Principais e a metodologia de Box-Jenkins, na seção 3, apresenta-se as aplicações e discussão e, finalmente, na seção 4 é apresentado a conclusão.

2 - METODOLOGIAS

2.1 - O modelo de componentes principais (CP)

Normalmente trabalha-se em uma dimensão que é o caso univariado; quando necessita-se trabalhar com duas, três ou mais dimensões, recorre-se à estatística multivariada onde é possível trabalhar-se com p-dimensões, sendo necessário recorrer-se ao uso de matrizes para melhor compreensão.

As medidas de várias características de uma mesma unidade experimental de forma simultânea e em certos intervalos de tempo, geram uma série de dados que devem ser analisados com técnicas multivariadas. Os métodos estatísticos multivariados podem ser subdivididos em dois grupos:

- 1ª) Os que permitem extrair informações a respeito da independência entre as variáveis que caracterizam cada elemento tais como: Análise de Fatores, Análise de Clusters, Análise de Correlação Canônica, Análise de Ordenamento Multidimensional, Análise de Componentes Principais e alguns métodos não-paramétricos; (AFIFI, 1971; ANDERSON, 1984).
- 2ª) Os que permitem extrair informações a respeito da dependência entre uma ou mais variáveis, ou uma com a outra tais como: Análise de Regressão Multivariada, Análise de Contingência Múltipla e Análise Discriminante. (AFIFI, 1971; ANDERSON, 1984).

O estudo de CP consiste em encontrar autovetores e autovalores que darão origem as novas variáveis denominadas $Y(k)$, $k = 1, \dots, p$ que sejam combinações lineares das variáveis originais $X_{(j)}$, e impor a este sistema certas condições que permitam satisfazer os objetivos da Análise de Componentes Principais.

Seja a matriz de variância - covariância conhecida, admite-se que

$$Y1 = \alpha_{11}X_1 + \dots + \alpha_{1p}X_p,$$

desejando-se encontrar $\alpha_{11}, \dots, \alpha_{1p}$, de modo que a variância de $Y1$, $V(Y1)$

$$V(Y1) = \sum_{i=1}^p \sum_{j=1}^p \alpha_{1i} \alpha_{1j} \sigma_{ij},$$

seja maximizada sujeita a condição de que $\sum_{i=1}^p (\alpha_{1i})^2 = 1$. Esta condição garante a singularidade da solução.

A solução de $\alpha_1 = (\alpha_{11}, \dots, \alpha_{1p})$ é chamada de primeiro autovetor e é associado com o maior autovalor da matriz .

Este valor é igual para a variância $V(Y1)$. A combinação linear $Y1 = \alpha_{11}X_1 + \dots + \alpha_{1p}X_p$, é a primeira Componente Principal de X_1, \dots, X_p que explica $\left(\frac{100 V(Y1)}{V} \right)$ por cento da variância total, denotada por (V) .

Sendo $Y2 = \alpha_{21}X_1 + \dots + \alpha_{2p}X_p$ a segunda componente, deseja-se encontrar $\alpha_{21}, \dots, \alpha_{2p}$ de modo que a variância $V(Y2)$:

$$V(Y2) = \sum_{i=1}^p \sum_{j=1}^p \alpha_{2i} \alpha_{2j} \sigma_{ij}$$

seja maximizada sujeita às condições de que $\sum_{i=1}^p (\alpha_{2i})^2 = 1$ e,

$$\text{COV}(Y_1, Y_2) = \sum_{i=1}^p \sum_{j=1}^p \alpha_{1i} \alpha_{2j} \sigma_{ij} = 0$$

A primeira condição garante uma singularidade da solução e a segunda condição garante que $Y1$ e $Y2$ sejam não correlacionados.

A solução de $\alpha_2 = (\alpha_{21}, \dots, \alpha_{2p})$ é chamado de segundo autovetor e é associado com o segundo maior autovalor da matriz . Este valor é igual para a variância $V(Y2)$. A combinação

linear $Y_2 = \alpha_{21}X_1 + \dots + \alpha_{2p}X_p$ é a segunda Componente Principal de X_1, \dots, X_p e explica $\frac{100 V(Y_2)}{V}$ por cento da variância total.

Os dois componentes explicam a variância total em termos percentuais $\frac{100[V(Y_1) + V(Y_2)]}{V}$, de maneira análoga deve-se encontrar os Y_i 's componentes restantes.

Os componentes principais apresentam as seguintes características:

1º) Os componentes não estão correlacionados e além disso, pode-se supor multinormalidade nos dados originais, mostrando assim a sua independência;

2º) Cada componente sintetiza a máxima variabilidade residual contida nos dados;

A partir dos elementos da matriz S é possível calcular R (matriz de correlação), de mesma dimensão de S, cujos os elementos são os coeficientes de correlação entre a j-ésima e k-ésima variável.

$$R(jk) = \frac{S(jk)}{\sqrt{S(jj) \cdot S(kk)}}$$

A matriz S de covariância é uma maneira de expressar a dispersão dos dados ao redor da média.

A transformação linear que sintetiza a máxima variabilidade corresponderá a gerada pelo valor (j) que seja maior. Convencionalmente esta solução máxima é denominada de α_1 , notação que se utiliza de maneira que se cumpra:

$$\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_p$$

Assim, a primeira transformação linear, ou a primeira variável geradora ou o primeiro componente, sintetiza a máxima variabilidade possível no conjunto de dados originais. A segunda transformação linear, ou o segundo componente sintetiza a máxima variabilidade residual, sujeita a condição de não correlação com o primeiro componente, e assim até o p-ésimo componente.

O número de CP a serem utilizados é mostrado pelos seguintes métodos:

i) recomenda-se fazer um gráfico onde se representa a porcentagem de variação explicada por cada componente nas ordenadas e dos componentes em ordem decrescente nas abscissas, onde deve-se considerar os componentes anteriores ao ponto de inflexão da curva;

ii) o critério de KAISER (citado por MARDIA et al, 1979), inclui os componentes cujos valores próprios sejam maiores que 1. O outro critério consiste em incluir só aqueles cujos

valores próprios sejam superiores média. Em geral, utiliza-se aqueles componentes que conseguem sintetizar uma variância acumulada em torno de 70%.

2.2 - Modelos Univariado de Box & Jenkins

A modelagem de Box & Jenkins (1976), é baseada num ciclo iterativo, no qual a escolha da estrutura do modelo é baseada nos próprios dados. Os estágios da metodologia são os seguintes:

1º) Identificação da estrutura do modelo, através da função de autocorrelação e autocorrelação parcial, podendo ser dos seguintes tipos:

$$\text{AR}(p) \quad \tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \dots + \phi_p \tilde{Z}_{t-p} + a_t$$

$$\text{MA}(q) \quad \tilde{Z}_t = a_t + \theta_1 a_{t-1} + \dots + \theta_q a_{t-q}$$

$$\text{ARIMA}(p,d,q) \quad \Phi(B)\Delta^d Z_t = \Theta(B)a_t;$$

2º) Estimação dos parâmetros do modelo, isto é, 's, 's e σ_a^2 (ou 's; 's no caso sazonal) através da maximização da função de verossimilhança condicional;

3º) Verificação do modelo ajustado, através da análise dos resíduos;

4º) Previsão dos valores futuros, através do modelo ajustado.

Foram utilizados os seguintes critérios de ajustamento:

i) Critério de informação de AKAIKE (AIC);

$$\text{AIC} = -2 \log (\text{máxima verossimilhança}) + 2 (n^\circ \text{ de parâmetros})$$

ii) Critério de SCHWARTZ, também conhecido como BIC (Bayesian Information Criterion);

$$\text{BIC} = \log (\hat{\sigma}_a^2) + \frac{\log(N) \cdot k}{N},$$

onde:

k = número de parâmetros;

N = número de resíduos;

$\hat{\sigma}_a^2$ = estimador de máxima verossimilhança residual.

iii) Coeficiente de determinação (R^2);

O critério para verificar a qualidade de previsão utilizado foi o erro percentual médio absoluto (MAPE).

3 - APLICAÇÕES E DISCUSSÃO

Neste item, apresenta-se a análise empírica dos agregados macroeconômicos, através da metodologia de Componentes Principais e de Séries Temporais.

As variáveis utilizadas foram:

1) Política Monetária (PM), composta pelas seguintes variáveis: M1 - Meios de pagamentos de ordem 1 (X1); M2 - Meios de pagamentos de ordem 2 (X2); M3 - Meios de pagamentos de ordem 3 (X3); M4 - Meios de pagamentos de ordem 4 (X4) e Base Monetária (X5).

A Política Monetária, representa a moeda em circulação no país, desde a ordem mais restrita que é o M1, que representa o papel moeda em poder do público até a ordem mais ampla o M4 que representa o saldo dos títulos públicos federais em circulação. A Base Monetária é o somatório do saldo em papel moeda mais os recursos em moeda dos bancos comerciais do Brasil e Caixas Econômicas, abrangendo somente as contas do Banco Central do Brasil.

2) Política de Preços (PP), composta pelas seguintes variáveis: IGP-DI - Índice Geral de Preços Disponibilidade Interna (X1); IGP-OG - Índice Geral de Preços Oferta Global (X2); IPA-DI - Índice de Preços por Atacado Disponibilidade Interna (X3); IPA-OG - Índice de Preços por Atacado Oferta Global (X4); IPA-OG-PA - Índice de Preço por Atacado Oferta Global de Produtos Agrícolas (X5); IPA-OG-PI - Índice de Preço por Atacado Oferta Global de Produtos Industrializados (X6).

O Índice Geral de Preços - IGP, representa as operações em geral, preços a varejo e o índice da construção civil, o Índice de Preço por Atacado - IPA, representa cada tipo de operação na formação da despesa interna bruta que são: a produção, transporte, comercialização a grosso modo de bens de consumo e de produção.

3) Balança Comercial (BC), composta pelas seguintes variáveis: Saldo da Balança Comercial (X1); Importações de Combustíveis e Lubrificantes (X2) e Exportações de Produtos Industrializados (X3).

Para o desenvolvimento deste estudo, foram utilizados os seguintes cálculos:

- i) estatísticas descritivas;
- ii) matriz de covariância, utilizando os dados originais;

- iii) matriz R a partir de S;
- iv) valores e vetores próprios com a sua respectiva proporção de variação total explicada, para cada um dos componentes;
- vi) coeficiente de correlação entre as variáveis e os componentes principais;
- vii) proporção da variação original aplicada para cada componente principal;
- viii) seleção dos componentes principais a serem utilizados na combinação linear;
- ix) correlação das variáveis originais com os componentes selecionados.

Na TABELA 1, apresenta-se a relação dos autovetores das variáveis em estudo.

Estes autovetores deram origem aos autovalores que formaram a componente principal, os autovalores fornecem a contribuição de cada variável na combinação linear.

| Variáveis | Autovetores | | | | | |
|--------------|-------------|--------|--------|--------|--------|--------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Pol. Monet. | 3.1181 | 1.5228 | 0.1792 | 0.1316 | 0.0294 | --- |
| Pol. Preço | 4.6637 | 0.9879 | 0.2650 | 0.0769 | 0.0049 | 0.0016 |
| Pol. Comerc. | 1.3175 | 1.0820 | 0.6004 | --- | --- | --- |

TABELA 1 - Relação dos autovetores das variáveis estudadas

3.1 - Combinação linear da primeira e da segunda componente principal

Apresenta-se a seguir as duas primeiras componentes principais para a Política Monetária Y1(PM) e Y2 (PM); Política Comercial Y1(PC) e Y2(PC); Política de Preços Y1(PP) e Y2(PP) onde os autovalores deram origem aos autovalores, que fornecem a contribuição de cada variável.

$$Y1(PM) = 0.4593 X1 + 0.0583 X2 + 0.4599 X3 + 0.4603 X4 + 0.4143 X5 + 0.4364 X6$$

$$Y2(PM) = - 0.0238 X1 + 0.9980 X2 - 0.0203 X3 - 0.0205 X4 - 0.0458 X5 - 0.0223 X6$$

$$Y1(PP) = 0.3294 X1 + 0.2952 X2 + 0.5265 X3 + 0.5161 X4 + 0.5106 X5$$

$$Y2(PP) = - 0.6091 X1 - 0.6485 X2 + 0.1514 X3 + 0.2920 X4 + 0.3167 X5$$

$$Y1(PC) = 0.6811 X1 + 0.1054 X2 + 0.7245 X3$$

$$Y2(PC) = -0.9018 X1 + 0.0918 X2 + 0.2191 X3$$

Para se determinar quais as variáveis que apresentam maior contribuição para cada componente principal, utilizou-se a seguinte relação:

$$\left| \alpha_{ij} (V(y_j))^{1/2} \right|; i = j = 1, \dots, 5$$

desta maneira encontrou-se o valor da correlação absoluta, onde:

$$\alpha_{ij} = \text{valor do autovetor e } V(y_j) = \text{valor do autovalor.}$$

Assim, pode-se observar que a variável de maior representatividade da Política de Preços é o IGP-DI; para a Política Monetária a variável de maior representatividade é o M1 e o Saldo da Balança Comercial é a variável que mais representa a Balança Comercial.

Na TABELA 2 apresenta-se a variância explicada para cada componente e a variável que mais contribuiu para a sua formação.

| Variáveis | V(Y1)% | V(Y2)% | V(Y1+Y2)% | Maior Contribuição | | Redução |
|--------------|--------|--------|-----------|--------------------|--------|----------|
| | | | | Y1 | Y2 | |
| Pol. Monet. | 62.34 | 30.85 | 93.19 | M1 | M1 | 5 para 2 |
| Pol. Preço | 77.73 | 16.46 | 94.16 | IGP-DI | IGP-OG | 6 para 2 |
| Pol. Comerc. | 43.91 | 36.07 | 79.98 | SBC | ICL | 3 para 2 |

TABELA 2 - Variância explicada pelo primeiro e segundo CP, variáveis que mais contribuíram para a formação da 1ª e 2ª combinação linear e o número de redução cada variável macroeconômica

3.2 - Ajustamento e previsão pelos modelos ARIMA

Neste item, apresenta-se o modelo ARIMA com intervenção para o componente principal e para a variável de maior representatividade através da metodologia de Box & Jenkins (1976). Nas expressões a seguir os valores entre parênteses representam a estatística "t" calculada.

$$Z_t(\text{IGP-DI}) = \frac{4.0806}{(11.36)} + \frac{1.3399}{(13.30)} Z_{t-1} - \frac{0.30748}{(-5.05)} Z_{t-2} + a_t + \frac{4.0598}{(10.84)} X_{63} + \frac{1.3621}{(4.33)} X_{40} - \frac{1.5615}{(-4.94)} X_{15} + \frac{1.9578}{(5.38)} X_{62} + \dots + \frac{1.2802}{(4.11)} X_{61}$$

$$Z_t(\text{CP}) = \frac{2.2511}{(7.30)} + \frac{0.84033}{(13.42)} Z_{t-1} + a_t - \frac{1.2573}{(-4.26)} X_{16} - \frac{1.1623}{(-3.92)} X_{64} + \frac{0.96529}{(2.89)} X_{52}$$

$$Z_t(M1) = 31.562 + 0.28563 Z_{t-1} + a_t + 168.83 X_{63} + 62.405 X_{62} + 88.562 X_{16} + 30.282 X_{12} + 11.196 X_{34}$$

(9.07) (2.50) (12.10) (6.14) (6.47) (5.57) (2.53)

$$Z_t(CP) = 28.978 + 0.52301 Z_{t-1} + a_t + 102.18 X_{63} + 27.876 X_{36} - 14.946 X_{13} + 35.627 X_{62} + 33.767 X_{16}$$

(11.62) (3.32) (10.69) (6.19) (-3.39) (3.08) (3.72)

$$Z_t(SBC) = 25.184 + a_t + 149.03 X_6 + 47.666 X_{31} - 24.184 X_{70} + 5.1133 X_{25} + 1.3753 X_{73}$$

(541.11) (358.0) (114.51) (-58.89) (12.28) (3.30)

$$Z_t(CP) = 17.587 + a_t + 32.694 X_{30} - 16.587 X_{69} + 3.4336 X_{24} + 1.0576 X_{67} + 1.0019 X_{72}$$

(465.5) (96.76) (-49.1) (10.16) (3.13) (2.97)

Na TABELA 3, apresenta-se o modelo ARIMA utilizado na previsão da variável referência (componente principal) e a variável de maior representatividade, no período de janeiro a dezembro de 1992 e as estatísticas de ajuste e de previsão.

| Variáveis | Modelo | Estatísticas de Ajuste | | | Estatísticas de Previsão MAPE (%) |
|-----------|--------|------------------------|--------|--------|--------------------------------------|
| | | R2 (%) | AIC | BIC | |
| IGP-DI | AR(2) | 90.90 | --- | --- | 308.25 |
| CP | AR(1) | 76.73 | --- | --- | 159.75 |
| M1 | AR(1) | 77.89 | 0.537 | 0.557 | 28.33 |
| CP | AR(1) | 78.03 | 0.473 | 0.494 | 15.85 |
| SBC | Const. | 99.94 | -1.696 | -1.523 | 13.17 |
| CP | Const. | 99.35 | -2.113 | -1.940 | 11.70 |

TABELA 3 - Ajuste das Previsões através dos modelos ARIMA

Pode-se verificar que tanto o componente principal e a variável de maior representatividade apresentam a mesma estrutura de modelagem, nos modelos autorregressivos e modelo constante. Através das estatísticas de ajuste R², AIC e BIC, verifica-se que todos os modelos apresentam um bom coeficiente de explicação, servindo assim, para se realizar previsões. Através da estatística MAPE verifica-se que o erro médio de previsão utilizando o método dos componentes principais é menor do que o modelo univariado.

4 - CONCLUSÃO

Neste trabalho, encontrou-se através da metodologia de CP a variável referência para o conjunto das variáveis analisadas bem como a variável que apresenta maior contribuição na

formação da variável referência. A metodologia de componentes principais apresentou um melhor desempenho na redução do número de variáveis, mantendo o mesmo nível de informação do conjunto das variáveis originais, bem como proporcionou que se identificasse qual a variável mais representativa em cada conjunto de variáveis econômicas, a Política de Preços (IGP-DI), a Política Monetária (M1), a Balança Comercial (SBC). Tanto a variável referência (o componente principal), como a variável de maior representatividade apresentaram a mesma estrutura (autorregressiva) de modelagem, portanto, a metodologia de Componentes Principais pode ser utilizada em estudos macroeconômicos, que envolvem muitas variáveis.

5 - REFERÊNCIAS BIBLIOGRÁFICAS

- AFIFI, A. A.; AZEN, S. P. (1971). **Statistical Analysis. A computer oriented approach.** Los Angeles, California. 2ª ed. 1971.
- ANDERSON, T. W. (1984). **An introduction to multivariate analysis.** Wiley, New York, N. Y. 2ª ed.
- BOX, G. E.; JENKINS, G. M. (1976) **Time séries analysis, forecasting and control.** San Francisco. Holden Day.
- MARDIA, K. W.; KENT, J. T.; BIBBY, J. M. (1979). **Multivariate analysis.** Academic, Londres, p. 521.
- MORRISON, D. F. (1976). **Multivariate statistical methods.** McGraw, New York, 2ª ed. 515 pag. 1976.
- PINTO, L. A. M. S. (1981). Componentes Principais: Um teste para a eficiência de previsão a curto prazo. **Dissertação de Mestrado.** Departamento de Engenharia Industrial da PUC/RJ. Rio de Janeiro.