ADRIANO MENDONÇA SOUZA

COMPONENTES PRINCIPAIS: APLICAÇÃO NA REDUÇÃO DE VARIÁVEIS ECONÔMICAS PARA O ESTUDO DE SÉRIES TEMPORAIS

DISSERTAÇÃO DE MESTRADO

Santa Maria, RS – BRASIL

1993

UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

COMPONENTES PRINCIPAIS: APLICAÇÃO NA REDUÇÃO DE VARIÁVEIS ECONÔMICAS PARA O ESTUDO DE SÉRIES TEMPORAIS

Adriano Mendonça Souza

Santa Maria – RS

1993

COMPONENTES PRINCIPAIS: APLICAÇÃO NA REDUÇÃO DE VARIÁVEIS ECONÔMICAS PARA O ESTUDO DE SÉRIES TEMPORAIS

por

Adriano Mendonça Souza

Dissertação de mestrado apresentada ao curso de Pós-graduação em Engenharia de Produção, da Universidade Federal de Santa Maria – RS, como requisito parcial para a obtenção do grau de Mestre em Engenharia de Produção.

Santa Maria

1993

UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

A BANCA EXAMINADORA, ABAIXO ASSINADA, APROVA A DISSERTAÇÃO

COMPONENTES PRINCIPAIS: APLICAÇÃO NA REDUÇÃO DE VARIÁVEIS ECONÔMICAS PARA O ESTUDO DE SÉRIES TEMPORAIS

ELABORADA POR

ADRIANO MENDONÇA SOUZA

COMO REQUISITO PARCIAL PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA DE PRODUÇÃO

BANCA EXAMINADORA:

Orientadora - Profª. Drª. Maria Emília Camargo

Prof. Msc. Odorico A. Bortoluzzi

Prof. Dr. Ramaswami Ramaswami

Santa Maria, RS, 1993.

AGRADECIMENTOS

A Profª. Drª. Maria Emília Camargo, pela orientação segura e disponível dispensada a este trabalho.

Aos professores do curso de Pós-graduação em Engenharia de Produção, pelos conhecimentos transmitidos.

Aos colegas do Departamento de Estatística – UFSM e aos colegas de aula, pessoas com quem sempre pude contar nas horas mais difíceis.

Aos familiares pelo incentivo e apoio, que se fizeram sempre presentes.

Ao CNPq pelo apoio financeiro.

Aos meus pais, a quem devo toda esta caminhada, a eles o meu reconhecimento e carinho.

À Márcia, a quem dedico esta caminhada, pelo estímulo, compreensão e força nos momentos mais necessários.

RESUMO

COMPONENTES PRINCIPAIS: APLICAÇÃO NA REDUÇÃO DE VARIÁVEIS ECONÔMICAS PARA O ESTUDO DE SÉRIES TEMPORAIS

Autor: Adriano Mendonça Souza
Orientadora: Profª. Drª. Maria Emilia Camargo

Este trabalho tem por objetivo, em essência, além de expor de forma estruturada a metodologia de componentes principais procurando-se mostrar a sua aplicabilidade em reduzir a dimensionalidade do número de variáveis sem perder a finalidade de realizar previsões. Foi feita uma análise empírica de variáveis econômicas no período de janeiro de 1985 a dezembro de 1991 representativas da Política Monetária, Política de Preços, da Atividade Econômica, dos Agregados de Crédito e da Balança Comercial. Utilizando-se os modelos ajustados através da Metodologia de séries temporais, fez-se previsões para o período de janeiro a dezembro de 1992.

UNIVERSIDADE FEDERAL DE SANTA MARIA
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
AUTOR: Adriano Mendonça Souza
ORIENTADORA: Profª. Drª. Maria Emilia Camargo
TÍTULO: Componentes Principais: Aplicação na Redução de Variáveis Econômicas para o Estudo de Séries Temporais
Dissertação de Mestrado em Engenharia de Produção
Santa Maria, 03 de maio de 1993.

ABSTRACT

PRINCIPAL COMPONENTS: APPLICATION IN THE REDUCTION OF ECONOMIC VARIATES TO THE STUDY OF TIME SERIES

Author: Adriano Mendonça Souza
Adviser: Profª. Drª. Maria Emilia Camargo

The objective of this research is to explain the methodology of the 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 Polycy and Price, Economical Activites, Investiment Variables and Comercial Balance, from January 1985 to December 1991. A prevision was made for the period from January to December of 1992, using the adjusted models by the methodology of time series.

UNIVERSIDADE FEDERAL DE SANTA MARIA
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
AUTOR: Adriano Mendonça Souza
ORIENTADORA: Profª. Drª. Maria Emilia Camargo
TÍTULO: Componentes Principais: Aplicação na Redução de Variáveis Econômicas para o Estudo de Séries Temporais
Dissertação de Mestrado em Engenharia de Produção
Santa Maria, 02 de maio de 1993.

SUMÁRIO

RESUMO v

ABSTRACT vi

LISTA DE TABELAS ix

LISTA DE FIGURAS xi

LISTA DE ANEXOS xii

1 - INTRODUÇÃO 1

1.1 - OBJETIVOS 1
1.1.1 OBJETIVO GERAL 1
1.1.2 OBJETIVOS ESPECÍFICOS 1

1.3 - ESTRUTURA DO TRABALHO 1

2 - REVISÃO DA LITERATURA 3
2.1 - TRABALHOS COM ENFOQUE GERAL 3
2.2 - TRABALHOS COM ENFOQUE ECONÔMICO 6

3 - O MÉTODO DE ANÁLISE DE COMPONENTES PRINCIPAIS 7
3.1 - INTRODUÇÃO 7
3.2 - ORIGENS 9
3.3 - GERAÇÃO DOS COMPONENTES PRINCIPAIS 10
3.3.1 - NÃO CORRELAÇÃO ENTRE OS COMPONENTES 11
3.3.2 - MÁXIMA VARIABILIDADE 17

3.4 - NOVA EXPRESSÃO DOS DADOS 23
3.5 - USO DA MATRIZ DE CORRELAÇÃO 25
3.6 - INTERPRETAÇÃO DOS COMPONENTES PRINCIPAIS 29
3.7 - SELEÇÃO DO NÚMERO DE COMPONENTES 31
3.8 - CORRELAÇÃO ENTRE VARIÁVEIS ORIGINAIS E COMPONENTES PRINCIPAIS 33
3.9 - TESTE DE HIPÓTESE PARA OS VALORES PRÓPRIOS 35

4 - MODELOS UNIVARIADOS DE BOX & JENKINS 39
4.1 - CRITÉRIOS DE AJUSTAMENTO 40
4.2 - CRITÉRIOS DE PREVISÃO 42

5 - APLICAÇÕES E DISCUSSÕES 43
5.1 - INTRODUÇÃO 43
5.2 - VARIÁVEIS UTILIZADAS NA ANÁLISE EMPÍRICA 43
5.3 - ANÁLISE DA POLÍTICA MONETÁRIA 45
5.4 - ANÁLISE DA POLÍTICA DE PREÇOS 49
5.5 - ATIVIDADE ECONÔMICA 53
5.6 - ANÁLISE DOS AGREGADOS DE CRÉDITO 56
5.7 - ANÁLISE DA BALANÇA COMERCIAL 59
5.8 - AJUSTAMENTO E PREVISÃO 63

5.8.1 - AJUSTAMENTO 63
5.8.2 - PREVISÃO PARA O PERÍODO DE JANEIRO A DEZEMBRO DE 1992 71

6 - CONCLUSÃO 75

REFERÊNCIAS BIBLIOGRÁFICAS 76

LISTA DE TABELAS

TABELA 01 - Estatísticas univariadas 46

TABELA 02 - Relação dos autovalores da política monetária 46

TABELA 03 - Coeficientes dos cinco componentes principais 46

TABELA 04 - Percentual da variância explicada por cada componente e o percentual acumulado 48

TABELA 05 - Estatísticas univariadas 49

TABELA 06 - Relação dos autovalores da variável política de preços 50

TABELA 07 - Coeficientes das seis componentes principais 50

TABELA 08 - Percentual da variância explicada por cada componente e o percentual acumulado 51

TABELA 09 - Estatísticas univariadas 53

TABELA 10 - Relação dos autovalores da atividade econômica 53

TABELA 11 - Coeficientes das três componentes principais 54

TABELA 12 - Percentual da variância explicada por cada componente e o percentual acumulado 55

TABELA 13 - Estatísticas univariadas 56

TABELA 14 - Relação dos autovalores da balança comercial 57

TABELA 15 - Coeficientes das três componentes principais 57

TABELA 16 - Percentual da variância explicada por cada componente e o percentual acumulado 58

TABELA 17 - Estatísticas univariadas 60

TABELA 18 - Relação dos autovalores da balança comercial 60

TABELA 19 - Coeficientes das duas componentes principais 60

TABELA 20 - Percentual da variância explicada por cada componente e o percentual acumulado 62

TABELA 21 - Valores previstos e observados para a variável M1 e os valores previstos para a Variável Referência em 1992 72

TABELA 22 - Valores previstos e observados para a variável IGP-DI e os valores previstos para a Variável Referência em 1992 72

TABELA 23 - Valores previstos e observados para a variável PRODUÇÃO INDUSTRIAL (PI) e os valores previstos para a Variável Referência em 1992 73

TABELA 24 - Valores previstos e observados para a variável CRÉDITO e os valores previstos para a Variável Referência em 1992 73

TABELA 25 - Valores previstos e observados para a variável SALDO DA BALANÇA COMERCIAL (SBC) e os valores previstos para a Variável Referência em 1992... 74

LISTA DE FIGURAS

FIGURA 01 - Elipse representando dois componentes principais em um espaço bidimensional 30

FIGURA 02 - Filtros lineares estacionários e não estacionários..... 39

LISTA DE ANEXOS

ANEXO 1 - GRÁFICOS 79

ANEXO 2 - GRÁFICOS DAS PREVISÕES 90

ANEXO 3 - DEFINIÇÕES BÁSICAS 96

1 - INTRODUÇÃO

Com a grande complexidade de variáveis que envolvem o universo econômico e com a preocupação de agilizar-se as tomadas de decisões com base em uma fonte segura pode-se recorrer a Análise de Componentes Principais, que tem como objetivo reduzir o número de variáveis, mas mantendo o mesmo nível de informação oferecido pelo conjunto total das variáveis originais.

Para o desenvolvimento desta dissertação, foram definidos os seguintes objetivos:

1.1 - OBJETIVOS

1.1.1 – OBJETIVO GERAL

Analisar variáveis macroeconômicas brasileiras através da metodologia de componentes principais e da metodologia de Box & Jenkins.

1.1.2 – OBJETIVOS ESPECÍFICOS

- Apresentar de forma estruturada a metodologia de Componentes Principais;
- Gerar variáveis que possam expressar a informação contida no conjunto dos dados originais;
- Reduzir a dimensionalidade do problema que se está estudando, como passo prévio para futuras análises;
- Eliminar as variáveis originais, que possuem pouca informação;
- Determinar a variável referência de: Política Monetária, Política de Preços, Atividade Econômica, Agregados de Crédito e Balança Comercial;
- Fazer previsões a médio prazo utilizando a metodologia de Box & Jenkins, com a dimensionalidade dos dados originais reduzido pelo método de Componentes Principais.

1.3 – ESTRUTURA DO TRABALHO

A estrutura deste trabalho, segue a seguinte forma: serão analisadas separadamente as séries macroeconômicas que compõem: a Política Monetária, a Política de Preços, a Atividade Econômica, os Agregados de Crédito e a Balança Comercial.

Para cada conjunto de variáveis macroeconômicas foi feita uma análise de componentes principais, verificando-se assim quantos componentes serão necessários para representar a série original.

A eliminação de algumas variáveis quando possível foi feita com base na proporção de variância explicada por cada componente, onde as que possuíram maiores variâncias foram as que possuirão maior grau de representatividade através do componente.

Após feita a análise de componentes principais em cada grupo de variáveis macroeconômicas, foi montado uma variável de referência para cada grupo de variáveis macroeconômicas selecionadas através dos componentes principais e da Análise de Correlação.

E, finalmente foi feita uma modelagem de cada variável referência e da variável mais representativa de cada grupo de variáveis através da metodologia de Box & Jenkins, procurando verificar a estrutura destas variáveis com a variável referência. Foi feita previsão proporcionando assim conhecimento com antecedência do movimento das variáveis macroeconômicas.

A dissertação se subdivide da seguinte forma:

O capítulo 2 aborda a revisão da literatura. O capítulo 3 trata do método de Componentes Principais, neste capítulo apresentamos a fundamentação teórica sobre componentes principais. No capítulo 4 apresentamos a metodologia de Box & Jenkins para séries temporais.

O capítulo 5 traz a aplicação das metodologias de Componentes Principais e de Séries Temporais a variáveis macroeconômicas brasileiras. A conclusão é apresentada no capítulo 6.

No anexo 1 apresenta-se os gráficos de todas variáveis macroeconômicas utilizadas nesta dissertação maior representatividade e o gráfico da variável referência.

No anexo 2 apresenta-se os gráficos da previsão para a série de maior representatividade e da variável referência em comparação com o valor observado da variável mais representativa.

No anexo 3 apresenta-se as definições básicas sobre estatística multivariada.

2 – REVISÃO DA LITERATURA

A análise multivariada, principalmente a metodologia de componentes principais, vem sendo estudada há muito tempo e por vários autores, assim a revisão será apresentada em dois itens:

2.1 – TRABALHOS COM ENFOQUE GERAL

Em 1901 KARL PEARSON publicou um trabalho sobre o ajuste de um sistema de pontos em um multiespaço a uma linha e a um plano. Este enfoque foi retomado em 1933 por HOTELLING, que foi o primeiro a formular a análise de componentes principais tal como se tem difundido até hoje.

KENDALL (1957), utilizou a Análise de Componentes Principais para ajustar uma equação de regressão multivariada e posteriormente MARQUARDT em 1970 também ajustou uma equação de regressão para variáveis independentes colineares.

JOLICOER e MOSIMANN (1960), apud MORRISON (1976) investigaram os Componentes Principais do comprimento, largura e altura da carapaça de tartarugas pintadas num esforço de dar o significado de "tamanho" e de "forma". O primeiro Componente Principal respondeu por quase toda a variância nas três dimensões. Assim, a nova média ponderada das mensurações das carapaças foi dada por,

$Y1 = 0.81(comprimento) + 0.50(largura) + 0.31(altura)$, logo o tamanho dos cascos das tartarugas poderia ser caracterizado por esta variável simples com pouca perda de informação.

Os autores denominaram a segunda e terceira medida dos componentes de "forma" da carapaça, porque elas parecem ser comparações do comprimento versus largura e altura, e altura versus comprimento e largura, respectivamente.

GNANADESIKAN e KETTENRING (1972), demonstraram que para valores extremos os primeiros componentes servem para detectar aquelas observações que contribuem para aumentar o alto grau de variância e de covariância (ou a correlação se na análise foi utilizada matriz de correlação), e que os últimos componentes são sensíveis para detectar observações que agregam dimensões espúrias aos dados. Sugeriram que para detectar visualmente os valores marginais se faça um diagrama de dispersão entre pares dos primeiros e dos últimos Componentes Principais.

HAWKINS (1974), propõe três critérios baseados em Componentes Principais e os compara com o teste de um componente por vez (um Qui-Quadrado) em duas hipóteses alternadas diferentes. Nos três casos, os primeiros componentes tem menor capacidade para detectar marginalidade, e por isso, os testes se baseiam nos últimos componentes.

GREENBERG (1975), examinou as propriedades dos estimadores gerados e chegou a conclusão de que ao incluir no modelo os últimos componentes principais (pequenos valores próprios) aumenta a variância dos estimadores, que tem uma elevada correlação destes com a variável dependente (ou com as variáveis dependentes) diminuindo-se o erro.

GUNST et alli (1976), compararam o método de mínimos quadrados com o de Componentes Principais e concluíram que quando existir multilinearidade é preferível o método de Componentes Principais, tanto para estimar parâmetros como para selecionar variáveis.

HOCKING (1976), demonstrou como pode-se relacionar a regressão em cadeia com a regressão em Componentes Principais, utilizando a caracterização da regressão em cadeia sugerida por ALLEN (1974).

CHATTERJEE & PRICE (1977), mostrou um exemplo em que as correlações existentes entre os coeficientes de regressão calculados a partir dos dados originais e dos calculados a partir dos Componentes Principais. O exemplo ilustra como a eliminação de algum Componente Principal do modelo de regressão equivale a impor uma restrição a dos coeficientes do modelo gerado com as variáveis originais. De tal forma que ao examinar os contrastes que menos contribuíram para explicar a variação total, não só se simplifica o modelo, mas que se conhece novas relações entre as variáveis originais.

WEISBERG (1980), diz que partindo-se de um modelo de regressão em função dos Componentes Principais nos quais se identifica como Z, obtém-se um novo modelo da forma:

$$Y = \beta Z + \varepsilon$$

sendo Y a variável dependente, β o vetor dos coeficientes de regressão, Z a matriz dos Componentes Principais e ε o vetor dos componentes aleatórios do modelo. Ainda afirma que escrevendo-se mais de um grupo de variáveis independentes e realizando a Análise e Componentes

Principais dentro de cada grupo e, logo, ajustando o modelo de regressão em função dos grupos de Componentes Principais, elimina-se completamente a multicolinearidade, garantindo-se a não-correlação dentro de cada grupo.

KUBRUSLY (1982), utilizou Componentes Principais e/ou Análise Fatorial para analisar parte de um questionário, mais precisamente a que diz respeito às questões da difusão de novas tecnologias, barreiras encontradas no caminho da adoção dessas novas tecnologias, e mudanças nas condições e na organização do trabalho, sendo que em alguns casos, recorre-se diretamente a informações contidas nas matrizes de correlação, ou mesmo a análises univariadas, quando isto se mostrar indispensável para a interpretação dos resultados.

MATTEUCCI & COLMA (1982), aplicaram o método de Componentes Principais ao estudo da cobertura vegetal do estado de Falcón no nordeste da Venezuela. Utilizaram uma parte da informação florística gerada pelo projeto "Análise Regional da Vegetação e Ambiente do estado de Falcón" e concluíram que existem seis pares de espécies altamente relacionadas. Identificaram através dos primeiros componentes que não existia uma diferença tão grande entre os censos anteriores e os componentes, e que os valores mais elevados deve-se a presença da Castela erecta que em poucos lugares ela é abundante, o que foi possível detectar visualmente.

VELAZQUEZ (1984), utilizou dados de um projeto para o diagnóstico leitero realizado pela Universidade de Francisco de Miranda e sua empresa de serviços INUFALCA, conjuntamente com o fundo de crédito agropecuário (UNEFM – FCA) correspondente ao distrito federal do estado de Falcón, Venezuela, com o objetivo de conhecer a situação leitera neste distrito. Assim, realizaram uma pesquisa, durante a qual visitaram produtores em suas propriedades rurais, coletaram dados à respeito de uma série de variáveis que influenciam a produção total e a produtividade por propriedade e por vaca. Esta análise foi feita através do método de Componentes Principais. Os resultados encontrados foram utilizados posteriormente para organizar a assistência técnica que deveria formular um plano de evolução técnico-econômica que faz parte do projeto global.

HENRIQUEZ (1985), analisou através do método de Componentes Principais as notas obtidas por três grupos de estudantes da Faculdade de Agronomia da Universidade Central da Venezuela nas chamadas matérias básicas que compreendem 17 disciplinas de departamentos diferentes, tendo como objetivo detectar as disciplinas que contribuem com maior variabilidade no conjunto total das disciplinas e verificar se realmente são as que mais contribuem para a formação

do profissional, tendo em vista que ao término de sua carreira um estudante é julgado e classificado pela média geral entre as disciplinas.

2.2 – TRABALHOS COM ENFOQUE ECONÔMICO

HADDAD (1977), para a formação do indicador do produto apresenta os resultados sobre indicadores do comportamento do produto real e nível de emprego, também é feita uma análise da sondagem conjuntural como indicador preditivo do comportamento da produção industrial. É estudado o problema de como estimar rapidamente o produto em períodos próximos passados através do emprego de variáveis obtidas com uma certa rapidez. É apresentado a previsão do produto real. Para esta análise empírica foi utilizada a técnica de componentes principais.

PINTO (1981), aplicou a técnica de Componentes Principais para prever o comportamento de variáveis econômicas e sua capacidade de identificar as reversões cíclicas dos fenômenos econômicos, tentando atingir o maior grau de confiabilidade nas estimativas para taxas de crescimento futuro do produto real no Brasil, a técnica foi utilizada agregando informações contidas em grande número de variáveis explicativas do comportamento de uma variável referência.

KUBRUSLY & GOUVÊA (1988), estudaram a análise de estabilidade no tempo das matrizes do balanço energético nacional para o período (1976 – 1980), utilizando a Análise de Componentes Principais para reduzir a dimensão das matrizes. A matriz do Balanço Energético Nacional fornece informações sobre a produção e o consumo de energia (fonte primária e secundária) nos diversos setores da economia. Foram selecionados 23 fontes e 35 setores econômicos. Quanto as fontes energéticas, a análise indicou que 83% da variância total é explicada através dos cinco primeiros componentes relativos a cada ano da série. Além disso, as variâncias se apresentam bastante estáveis ao longo do tempo.

BOFF (1992), apresenta um estudo sobre variáveis padronizadas múltiplas proposta inicialmente por AMATO (1976 a 1988), ao mesmo tempo que examina a possibilidade de definir as variáveis padronizadas múltiplas com restrições sobre os coeficientes de ajustamento. Mostra os novos aspectos da relação entre as variáveis padronizadas múltiplas e as correlações parciais e também explicita a inserção desta relação na Análise de Componentes Principais restritos. Apresenta uma aplicação das variáveis padronizadas múltiplas com a definição das variáveis instrumentais padronizadas para a estimação paramétrica do modelo linear geral e os critérios para a escolha das variáveis de padronização e as propriedades dos estimadores.

3 – O MÉTODO DE ANÁLISE DE COMPONENTES PRINCIPAIS

Neste item será apresentada a fundamentação teórica sobre componentes principais, baseando-se em AFIFI, 1971; KENDALL, 1980; ARNOLD, 1981; ANDERSON, 1984; FLURY, 1988.

3.1 – INTRODUÇÃO

Normalmente trabalha-se em 1 dimensão no caso univariado e quando necessitamos trabalhar com 2 ou três dimensões recorremos a análise estatística multivariada onde é possível trabalhar-se com p-dimensões, sendo que neste caso recorre-se ao uso das matrizes para melhor compreensão.

A medição 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.

Quando trabalha-se com populações univariadas, quase sempre é possível caracterizar completamente a distribuição de probabilidade a partir dos parâmetros μ *e* σ^2. A interferência estatística exige que seja tomada uma amostra aleatória e que sejam feitos cálculos dos melhores estimadores destes parâmetros, terminando a análise com a interpretação dos parâmetros encontrados.

No caso multivariado estuda-se populações p-variadas e tem-se um conjunto de indivíduos onde se tenha observado e medido p características e propriedades, dispondo-se de p médias, p variâncias e ½ p(p-1) parâmetros, e se reduzem as dimensões de p a (p-1), se passa de ½ p(p+3) parâmetros populacionais a ser estimados e interpretados a ½ p(p+3) – ½ (p-1) (p+2) = (p+1) parâmetros estimados e interpretados.

Se existe interesse em todos os parâmetros, eles podem ser estimados para se encontrar uma interpretação mais compreensível com a estrutura original da amostra.

Cada situação exige um estudo particular para utilizar o método de análise multivariada mais adequado, permitindo extrair a máxima informação do conjunto de dados, para garantir a validade de sua aplicabilidade. As técnicas multivariadas são muito potentes e podem levar o

investigador a encontrar uma justificativa que não se sustente na análise do objetivo da informação reorganizada.

Os métodos estatísticos multivariados, podem agrupar-se em dois grupos:

i) Os que permitem extrair informações a respeito da independência entre as variáveis que caracterizam a cada um dos indivíduos e;

ii) Os que permitem extrair informações a respeito da dependência entre uma ou várias variáveis, ou umas com as outras.

O método de análise multivariada para detectar a interdependência entre variáveis e também entre indivíduos se incluem a análise de fatores, análise por conglomerados "clusters", análise de correlação canônica, a análise de ordenamento multidimensional "scaling", a Análise de Componentes Principais e alguns métodos não-paramétricos.

Os métodos para detectar a dependência compreendem a análise de regressão multivariada, análise de contingência múltipla e análise discriminante.

No caso univariado tem-se a média, a variância, o desvio-padrão e outras estatísticas que ajudam a fazer a interpretação de uma população, analogamente tem-se no caso multivariado, estas mesmas características.

O método de Análise de Componentes Principais é um dos mais difundidos, que permite a estruturação de um conjunto de dados multivariados obtidos de uma população, cuja distribuição de probabilidade não necessita ser conhecida.

Trata-se de uma técnica que não requer um modelo estatístico para explicar a estrutura de probabilística dos erros. Sem problemas, se pode supor que a população amostrada tem distribuição multinomial. Poder-se-á estudar a significação estatística e será possível utilizar a amostra efetivamente observada para efetuar-se testes de hipóteses que contribuam para conhecer-se a estrutura da população original, com um certo grau de confiabilidade fixado a priori ou a posteriori.

As novas variáveis geradas se denominam Componentes Principais e possuem algumas características estatísticas, tais como independência (quando assume multinormalidade) e em todos os casos não correlacionados. Isto significa que as variáveis originais não estão correlacionadas, a Análise de Componentes Principais não oferece vantagem alguma, ela se baseia em uma

transformação das observações originais, esta transformação linear é conhecida no campo da álgebra vetorial como geração de vetores e valores próprios.

3.2 – ORIGENS

O trabalho original de PEARSON (1901), se centrava naqueles componentes, ou combinações lineares de variáveis originais, para os quais a variância não explicada foi mínima. Estas combinações geram um plano, função das variáveis originais, no qual o ajuste do sistema de pontos é "o melhor", por ser mínima a soma das distancias de cada ponto ao plano de ajuste.

O enfoque de HOTELLING se concentra em analise dos componentes que sintetizam a maior variabilidade do sistema de pontos, ele explica assim o qualificativo de "principal". Por inspeção destes componentes, que resumem a maior proporção possível de variabilidade total entre o conjunto de pontos, pode-se encontrar um meio para classificar e detectar relações entre os pontos.

Cada ponto em um multiespaço p-dimensional é o extremo de um vetor X tal que cada um de seus elementos $X_{(j)}$, para j = 1, ..., p, é uma medida da j-ésima variável em um dado individual. Medindo-se n indivíduos, se obtém n vetores X e n pontos no espaço de p dimensões.

Desde suas origens, a Análise de Componentes Principais tem sido aplicada em situações muito variadas: em psicologia, medicina, meteorologia, geografia, ecologia, agronomia e outras ciências.

Esta análise se aplica quando se dispõe de um conjunto de dados multivariados e não se pode postular, sobre a base de conhecimento prévio sobre o universo em estudo, uma estrutura particular das variáveis.

Quando se conhece a existência de uma ou várias variáveis dependentes, pode-se aplicar as técnicas de regressão múltipla e de regressão multivariada. Sabendo-se que não existe nenhuma relação entre as variáveis (existe independência ou, ao menos, não há correlação), deverá abster-se de procurar uma explicação de "relação" entre as variáveis, ou entre os indivíduos a partir dessas variáveis em forma conjunta. Neste último caso, em estudos do tipo unidimensional se obterá os mesmos resultados com técnicas mais potentes e em forma menos trabalhosa, tanto do ponto de vista computacional, como do ponto de vista da interpretação.

A Análise de Componentes Principais deverá ser aplicada quando se diz conhecer a relação entre os elementos de uma população e se suspeite que esta relação influencie de maneira desconhecida em um conjunto de variáveis e propriedades dos elementos.

3.3 – GERAÇÃO DOS COMPONENTES PRINCIPAIS

Os dados multivariados oferecem a possibilidade de serem expressos em combinações lineares das variáveis originais. Esta é a ferramenta mais poderosa para realizar este tipo de análise estatística, o qual não é fácil no campo univariado. Em um número reduzido de combinações é possível sintetizar a maior parte da informação contida nos dados originais, às vezes é muito complicado deduzir a distribuição exata de probabilidade, a respeito das combinações mais utilizadas se se conhecem resultados assintóticos.

Os Componentes Principais apresentam as seguintes características:

i) os Componentes Principais não estão correlacionados e além disso, pode-se supor multinormalidade nos dados originais, mostrando assim a sua independência;

ii) cada Componente Principal sintetiza a máxima variabilidade residual contida nos dados.

Ao estudar-se um conjunto de n indivíduos mediante p-variáveis é possível encontrar novas variáveis denominadas Y(k), k = 1, ..., 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.

Isto implica encontrar (p x p) constantes tais que:

$$Y(k) = \sum_{j=1}^{p} \alpha_{(jk)} \; X_{(j)} \qquad k = 1, ..., p \tag{1}$$

onde $\alpha_{(jk)}$ é cada uma dessas constantes. Observa-se que devido ao somatório, em cada nova variável $Y_{(k)}$ intervém todos os valores das variáveis originais $X_{(j)}$. O valor numérico de $\alpha_{(j)}$ indicará o grau de contribuição que cada variável original dará sobre a nova variável definida pela

transformação linear. É possível que $\alpha_{(j)}$ tenha em algum caso particular o valor zero, ou muito próximo a zero, o qual indica que essas variáveis não influem no valor da nova variável $Y_{(k)}$.

3.3.1 – NÃO CORRELAÇÃO ENTRE OS COMPONENTES

Sem perda de generalidade e para simplificar a apresentação, supôs-se que:

$$E < X_{(j)} > = 0, \quad j=1,..., p.$$

Para satisfazer a condição de não correlação entre as novas variáveis definidas na equação (1) tem-se que:

$$E < y_{(k)} \; y_{(m)} > = 0, \;\; k,m =1,..., p \; com \; k \neq m.$$

Substituindo cada nova variável por sua definição em função das variáveis originais se obterá:

$$E < \left(\sum_{j=1}^{p} \alpha_{(jk)} \;\; X_{(j)} \right) \left(\sum_{h=1}^{p} \alpha_{(hm)} \;\; X_{(h)} \right) > = 0 \qquad (2)$$

Dado que $\alpha_{(jk)}$ e $\alpha_{(hm)}$ são constantes, sua esperança matemática será a mesma constante e, por tanto, podem ser escritos fora da esperança e com somatório duplo.

$$E\left\langle y_{(k)} \; y_{(m)} \right\rangle = \sum_{j=1, h=1}^{p} \alpha_{(jk)} \; \alpha_{(hm)} \; E\left\langle X_{(j)} \; X_{(h)} \right\rangle \qquad (3)$$

Identifica-se a constante que multiplica a cada valor a variável original com os sub-índices (j e h) para destacar que a introdução do operador esperança matemática na equação (2) gera todos os produtos possíveis em dobro.

Ao reescrever a equação (3) a expressão $E\left\langle X_{(j)} \; X_{(h)} \right\rangle$ por seu valor, esta será a covariância entre as variáveis originais $X_{(j)}$ e $X_{(h)}$, em outras palavras, serão os termos que caem fora da diagonal principal da matriz variância-covariância.

Como se havia dito que $k \neq m$, se a condição de correlação não existisse entre os Componentes Principais, haverá então ½ p(p-1) restrições sobre as constantes $\alpha_{(jk)}$ que devem ser impostas para que o sistema tenha uma solução única.

Estas restrições se estabelecem ao aplicar a transformação linear definida em (1), que define as condições para que as novas variáveis originadas sejam ortogonais. Conhecendo-se a transformação que produziu os Componentes Principais como aquela que gera um novo conjunto de j's ou coordenadas que sejam perpendiculares entre si, o cosseno do ângulo formado por dois quaisquer dos eixos deve ser zero, ou seja, eles são ortogonais. Estas condições podem ser expressas assim:

$$\sum_{j=1}^{p} \alpha_{(jk)} \; \alpha_{(jm)} = 1, \quad com \;\; k \neq m \qquad k, m = 1, ..., p \tag{4}$$

$$\sum_{j=1}^{p} \alpha_{(jk)} \; \alpha_{(jm)} = 1, \quad com \;\; k = m$$

o qual em álgebra vetorial se denomina "delta Kronecker".

É possível expressar a condição anterior em forma matricial definindo-se uma matriz α com o arranjo das (p x p) constantes $\alpha_{(jk)}$.

$$\alpha = \begin{bmatrix} \alpha_{(11)} & \cdots & \cdots & \cdots & \alpha_{(1p)} \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \alpha_{(jk)} & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \alpha_{(p1)} & \cdots & \cdots & \cdots & \alpha_{(pp)} \end{bmatrix} \tag{5}$$

Para satisfazer a condição de ortogonalidade é preciso que:

$$\alpha \,.\, \alpha' = \alpha' \,.\, \alpha = \alpha^{-1} \,.\, \alpha = I$$

e diz-se, então, que α é uma matriz ortogonal. Pode-se expressar a transformação linear de Componentes Principais através da seguinte matriz:

$$Y_{(nxp)} = X_{(nxp)} \; . \; \alpha_{(pxp)} \tag{6}$$

Na equação (6) foi representada a matriz original completa. Fazendo-se uma matriz de dados como a da definição (1), a qual se aplica a transformação ortogonal α, obtém-se uma nova matriz de dimensão igual a matriz original (nxp).

Para cada um dos indivíduos (n total), se calculam novos valores correspondentes as variáveis não correlacionadas. A nova matriz Y terá também uma matriz variância-covariância que será diagonal, se as variáveis (combinações lineares originadas pela transformação ortogonal α) não estiverem correlacionadas.

A condição da equação (3) também pode ser expressa em termos matriciais como:

$$E\left\langle Y^{'} Y\right\rangle = E\left\langle (X.\alpha)^{'} (X.\alpha)\right\rangle$$

aplicando as propriedades das operações matriciais, tem-se:

$$E\left\langle Y^{'} Y\right\rangle = E\left\langle \alpha^{'} X^{'} X \; \alpha\right\rangle$$

e introduzindo-se um operador esperança matemática, sendo α uma matriz de constantes

$$E\left\langle Y^{'} Y\right\rangle = \alpha^{'} E\left\langle \; X^{'} X \; \alpha\right\rangle$$

onde a esperança de ($X^{'} X$) é uma matriz de covariância dos dados originais representada por S, sendo que seu estimador amostral deve satisfazer a condição:

$$E\left\langle Y^{'} Y\right\rangle = \alpha^{'} S \; \alpha = \wedge \tag{7}$$

onde $\wedge$ é uma matriz diagonal, se as covariâncias amostrais das variáveis forem nulas. Esta matriz $\wedge$ terá na diagonal principal os valores das variâncias das novas variáveis ou Componentes Principais.

Multiplicando-se ambos os membros da equação (7) por α e recordando-se que $\alpha \,.\, \alpha' = I$, obtém-se:

$$(\alpha . \wedge) = (S . \alpha) \tag{8}$$

A matriz $\wedge$ pode ser expressa como:

$$\wedge = \begin{bmatrix} \lambda_{(1)} & 0 & 0 & \ldots & 0 \\ 0 & \lambda_{(2)} & : & : & : \\ : & : & \lambda_{(k)} & : & : \\ : & : & : & \lambda_{(p-1)} & : \\ 0 & 0 & 0 & 0 & \lambda_{(p)} \end{bmatrix}$$

A expressão matricial da equação (8) pode ser:

$$\begin{bmatrix} \alpha_{(11)} & \ldots & \ldots & \ldots & \alpha_{(1p)} \\ : & \ldots & \ldots & \ldots & : \\ : & \ldots & \alpha_{(jk)} & \ldots & : \\ : & \ldots & \ldots & \ldots & : \\ \alpha_{(p1)} & \ldots & \ldots & \ldots & \alpha_{(pp)} \end{bmatrix} \begin{bmatrix} \lambda_{(1)} & 0 & 0 & \ldots & 0 \\ 0 & \lambda_{(2)} & : & : & : \\ : & : & \lambda_{(k)} & : & : \\ : & : & : & \lambda_{(p-1)} & : \\ 0 & 0 & 0 & 0 & \lambda_{(p)} \end{bmatrix} =$$

$$\begin{bmatrix} S_{(11)} & \ldots & \ldots & \ldots & S_{(1p)} \\ : & \ldots & \ldots & \ldots & : \\ : & \ldots & S_{(jk)} & \ldots & : \\ : & \ldots & \ldots & \ldots & : \\ S_{(p1)} & \ldots & \ldots & \ldots & S_{(pp)} \end{bmatrix} \begin{bmatrix} \alpha_{(11)} & \ldots & \ldots & \ldots & \alpha_{(1p)} \\ : & \ldots & \ldots & \ldots & : \\ : & \ldots & \alpha_{(jk)} & \ldots & : \\ : & \ldots & \ldots & \ldots & : \\ \alpha_{(p1)} & \ldots & \ldots & \ldots & \alpha_{(pp)} \end{bmatrix}$$

Multiplicando-se a 1ª linha da 1ª matriz pela 1ª coluna da 2ª matriz obtém-se:

$$\alpha_{(1\,1)} \,.\, \lambda_{(1)}$$

sendo que o resto dos termos se anulam. Efetuando-se a mesma operação, com as matrizes a direita do sinal de igual, obtém-se:

$$S_{(11)}.\alpha_{(1\,1)} + S_{(12)}.\alpha_{(2\,1)} + ... + S_{(1\,p)}\ .\ \alpha_{(p\,1)}.$$

Para manter a igualdade, dos elementos homólogos das matrizes, deve-se igualá-los:

$$\alpha_{(11)}\ \lambda_{(1)} = S_{(11)}.\alpha_{(1\,1)} + ... + S_{(1\,p)}\ .\ \alpha_{(p\,1)}.$$

O segundo elemento da primeira coluna da matriz resultante se obterá multiplicando 2ª linha da primeira matriz pela 1ª coluna da segunda matriz:

$$\alpha_{(21)}\ \lambda_{(1)} = S_{(21)}.\alpha_{(1\,1)} + ... + S_{(2\,p)}\ .\ \alpha_{(p\,1)}.$$

O p-ésimo elemento da primeira coluna da matriz resultante se obterá multiplicando a última linha da primeira matriz pela primeira linha da segunda matriz:

$$\alpha_{(p1)}\ \lambda_{(1)} = S_{(p1)}.\alpha_{(1\,1)} + ... + S_{(p\,p)}\ .\ \alpha_{(p\,1)}.$$

Este sistema de p equações pode-se reordenar colocando todos os termos em um mesmo membro, igualando cada equação a zero e tirando fatores comuns $\alpha(k1)$ entre os componentes (k=1, ..., p).

$$\begin{array}{c}
0=\left\langle S_{(11)}-\lambda_{(1)}\right\rangle\alpha_{(11)}+S_{(12)}\ \alpha_{(21)}+...+S_{(1p)}\ \alpha_{(p1)}\\
0=S_{(11)}\ \alpha_{(11)}+\left\langle S_{(22)}-\lambda_{(1)}\right\rangle\alpha_{(21)}+...+S_{(2p)}\ \alpha_{(p1)}\\
\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
0=S_{(p1)}\ \alpha_{(11)}+...+\left\langle S_{(p\,p)}-\lambda_{(1)}\right\rangle\ \alpha_{(p1)}
\end{array}$$

Esse mesmo sistema de equações pode ser expresso de forma matricial:

$$\begin{bmatrix} S_{(11)} & \dots & \dots & \dots & S_{(1p)} \\ : & \dots & \dots & \dots & : \\ : & \dots & S_{(jk)} & \dots & : \\ : & \dots & \dots & \dots & : \\ S_{(p1)} & \dots & \dots & \dots & S_{(pp)} \end{bmatrix} \begin{bmatrix} \lambda_{(1)} & 0 & 0 & \dots & 0 \\ 0 & \lambda_{(1)} & 0: & \dots & 0 \\ 0 & 0: & \lambda_{(1)} & \dots & : \\ : & \dots & \dots & \lambda_{(1)} & : \\ 0 & 0 & 0 & 0 & \lambda_{(1)} \end{bmatrix} x$$

$$\begin{bmatrix} \alpha_{(11)} \\ \alpha_{(21)} \\ : \\ : \\ : \\ \alpha_{(p1)} \end{bmatrix} = 0 \qquad (9)$$

Nota-se que esta equação só interfere o 1º elemento da diagonal principal da matriz α, ou seja $\lambda_{(1)}$. Comparando-se esta equação com a transformação da definição (8), vê-se que é igual a matriz B da definição, a matriz de covariância S. Como mostrou-se, esta equação tem tantas soluções quantas dimensões tiver a matriz S. Neste caso a equação terá p soluções.

Se este mesmo procedimento que foi detalhado para a 1ª coluna da matriz resultante da equação (8) aplica-se também para outras colunas restantes, obtendo-se uma equação idêntica, no que implica das p soluções encontradas igualando-se a zero o determinante da matriz de covariância menos λI,

$$|S - \lambda I| = 0$$

que se conhece como polinômio característico da matriz S. Na equação (9) existem (p + 1) incógnitas (p valores de $\alpha_{(j1)}$ e $\lambda_{(1)}$) e só p equações, pelo qual o sistema tem solução única.

Considerando-se os p possíveis sistemas gerados pela equação (8), se terá (p x p) incógnitas correspondentes aos p elementos (dos $\alpha_{(jk)}$) e p incógnitas dos $\lambda_{(j)}$, pois só (p x p) equações existem. O restante das equações se encontram a partir da condição de normalização, tornando o sistema com solução única, devendo-se observar que:

$$\sum_{j=1}^{p} \alpha(j1)^2 = 1 \qquad para\ todo\ k.$$

Este sistema de equações dará lugar a p vetores que satisfaçam a equação (9), cada um destes vetores se denomina de vetor próprio. É importante mostrar que a matriz α, formada por p vetores próprios, não é simétrica, e que cada uma das colunas identifica uma nova variável, (Componente Principal), a qual é uma combinação linear de todas as variáveis originais.

Conhecida a matriz α é possível após multiplicar-se a matriz original de observações X por α, e obter uma nova matriz de dados Y, tal como foi definida na equação (6).

3.3.2 – MÁXIMA VARIABILIDADE

Pela forma como são gerados os Componentes Principais, também satisfazem a condição de sintetizar em forma decrescente a variância do conjunto original de dados.

Desejando-se achar a combinação linear que sintetize a máxima variância, deverá encontrar-se o Maximo da expressão da variância da equação (1) que define um Componente Principal.

$$Y = \sum_{j=1}^{p} \alpha_{(j)}\ X_{(j)}\ ,$$

do qual só se eliminam os sub-índices que identifica a nova variável.

$$Var\ (Y) = Var\left[\sum_{j=1}^{p} \alpha_{(j)}\ X_{(j)}\right] \tag{10}$$

Para calcular a variância da expressão entre colchetes deve-se recordar que as variáveis $X_{(j)}$ deverão somar-se as covariâncias. Recordando que a covariância $S_{(j\ h)}$ é igual a covariância $S_{(h\ j)}$, a expressão fica do seguinte modo:

$$Var(Y) = \sum_{j=1}^{p} \alpha_{(1)}^2\ S_{(j\,j)}\ + 2\sum_{\substack{j=1 \\ h=2}}^{p} \alpha_{(j)}\ \alpha_{(h)}\ S_{(j\,h)},\ j<h \tag{11}$$

recordando que $\alpha_{(j)}$ é uma constante, e que a variância do produto de uma variável aleatória por uma constante é igual ao produto da variância dessa variável aleatória pelo quadrado da constante. O somatório do segundo termo se faz para todos os valores de j e h diferentes, sendo j < h. A equação (11) pode ser escrita como:

$$Var(Y) = \sum_{j=1}^{p} \alpha_{(1)}^2 \, S_{(j\,j)} + \sum_{\substack{j=1 \\ h=2}}^{p} \alpha_{(j)} \, \alpha_{(h)} \, S_{(j\,h)}, \; j \neq h \tag{12}$$

da equação (11) só foi eliminado o 2 e fez-se a soma para todas as possíveis combinações de j e h diferentes, ainda é possível sintetizar mais equações anteriores agrupando-se os termos em:

$$Var(Y) = \sum_{\substack{j=1 \\ h=2}}^{p} \alpha_{(j)} \, \alpha_{(h)} \, \alpha_{(j\,h)} \tag{13}$$

onde j e h assumem todas as combinações de valores possíveis. Na equação anterior observa-se que a determinação da variância e das covariâncias originais (ou seja dos valores de $S_{(j\,h)}$) entre todas as variáveis originais. As constantes $\alpha_{(j)}$ e $\alpha_{(h)}$ que interferem na equação são os valores das p constantes que formam o vetor associado com essa nova variável (Componente Principal). E tem-se que são os elementos do vetor da equação (9) ou os elementos de uma coluna da matriz α definida em (5).

É necessário encontrar o máximo da equação (13) com restrições sujeitas às expressas na equação (4). Neste caso, se trata dos elementos de um mesmo vetor e, portanto, seguindo a nomenclatura da equação (13) temos:

$$\sum_{j=h=1}^{p} \alpha_{(j)} \, \alpha_{(h)} = 1, \tag{14}$$

ou seja, a soma dos quadrados dos valores deve ser igual a 1.

Derivando-se a equação (13) em relação aos valores de $\alpha_{(j)}$, considerando-se a restrição da equação (14) através do uso do multiplicador de Lagrange, tem-se:

$$\partial < \sum_{j=h=1}^{p} \alpha_{(j)}\, \alpha_{(h)}\, S_{(jh)} - g\left(\sum_{j=1}^{p} \alpha_{(j)}^{2} - 1 \right) > / \partial \alpha_{(j)} = 0 \qquad (15)$$

Esta equação pode ser derivada em relação a p possíveis valores de $\alpha_{(j)}$. Supondo-se que j=1, obtém-se:

$$\sum_{h=1}^{p} \alpha_{(h)}\, S_{(1h)} + \sum_{h=1}^{p} \alpha_{(h)}\, S_{(h1)} - g\,(2\alpha_{(1)}) = 0 \qquad (16)$$

como viu-se ao deduzir a equação (13). O duplo somatório inclui todas as combinações possíveis. Ao examinar a equação (16) levando em conta as propriedades de simetria da matriz de covariância pode-se observar que:

$$2\sum_{h=1}^{p} \alpha_{(h)}\, S_{(1h)} - 2\, g\, \alpha_{(1)} = 0, \qquad (17)$$

de onde podemos eliminar sem alterar a equação o número 2 em ambos os membros. A partir da equação (15), e derivando-se em relação a $\alpha_{(2)}$, ..., $\alpha_{(p)}$ pode-se encontrar equações similares a (17) que em conjunto, formam um sistema onde a incógnita é o valor g. este sistema será:

$$\sum_{h=1}^{p} \alpha_{(h)}\, S_{(1h)} - g\, \alpha_{(1)} = 0$$

$$\sum_{h=1}^{p} \alpha_{(h)}\, S_{(2h)} - g\, \alpha_{(2)} = 0$$

.....

$$\sum_{h=1}^{p} \alpha_{(h)}\, S_{(ph)} - g\, \alpha_{(p)} = 0$$

a qual pode-se expressar sem alterá-la:

$$\sum_{h=1}^{p} \alpha_{(h)} S_{(1h)} = g\ 1_{(1)}$$

$$\sum_{h=1}^{p} \alpha_{(h)} S_{(1h)} = g\ 1_{(2)}$$

.....

$$\sum_{h=1}^{p} \alpha_{(h)} S_{(ph)} = g\ 1_{(p)}$$

Cada uma dessas equações desse sistema pode expressar-se como o produto de vetores, dos quais pode-se organizar em forma matricial. Para a primeira equação será:

$$< \alpha_{(1)}\ \alpha_{(2)} ... \alpha_{(p)} > \begin{bmatrix} S_{(11)} \\ S_{(21)} \\ : \\ : \\ S_{(p1)} \end{bmatrix} = 1\ (1)\ g \tag{18}$$

que poderá repetir-se para cada uma das restantes. Estes vetores podem ser dispostos em uma matriz.

$$< \alpha_{(1)}\ \alpha_{(2)} ... \alpha_{(p)} > \begin{bmatrix} S_{(11)} & S_{(12)} & ..... & S_{(1p)} \\ S_{(21)} & S_{(22)} & ..... & S_{(2p)} \\ : & ..... & ..... & : \\ : & ..... & ..... & : \\ S_{(p1)} & ..... & ..... & S_{(pp)} \end{bmatrix} =$$

$$= < \alpha_{(1)}\ \alpha_{(2)} ... \alpha_{(p)} > . g \tag{19}$$

É possível expressar o segundo termo da equação (19) em forma matricial, obtém-se:

$$< \alpha_{(1)} \ \alpha_{(2)} ... \alpha_{(p)} > \begin{bmatrix} S_{(11)} & S_{(12)} & ..... & S_{(1p)} \\ S_{(21)} & S_{(22)} & ..... & S_{(2p)} \\ : & ..... & ..... & : \\ : & ..... & ..... & : \\ S_{(p1)} & ..... & ..... & S_{(pp)} \end{bmatrix} =$$

$$= < \alpha_{(1)} \ \alpha_{(2)} ... \alpha_{(p)} > \begin{bmatrix} g & 0 & ..... & 0 \\ 0 & g & ..... & 0 \\ 0 & 0 & ..... & : \\ 0 & 0 & ..... & g \end{bmatrix} \qquad (20)$$

Reordenando a equação (20), igualando-se a zero e tirando-se o vetor das constantes como fator comum à esquerda, obtém-se:

$$< \alpha_{(1)} \ \alpha_{(2)} ... \alpha_{(p)} > \left[\begin{bmatrix} S_{(11)} & S_{(12)} & ..... & S_{(1p)} \\ S_{(21)} & S_{(22)} & ..... & S_{(2p)} \\ : & ..... & ..... & : \\ : & ..... & ..... & : \\ S_{(p1)} & ..... & ..... & S_{(pp)} \end{bmatrix} + \right.$$

$$\left. - \begin{bmatrix} g & 0 & ..... & 0 \\ 0 & g & ..... & 0 \\ 0 & 0 & ..... & : \\ 0 & 0 & ..... & g \end{bmatrix} \right] = 0$$

onde:

$$1 x \, p \, (p \, x \, p \, - \, p \, x \, p) = 1 x \, p$$

À esquerda do "=" se encontra um vetor para o qual existem p incógnitas, uma matriz cujo valores são conhecidos que representam as variâncias e covariâncias das variáveis originais, e uma segunda matriz com uma incógnita, g. Existem p equações e (p + 1) incógnitas. Em conseqüência, para que o sistema tenha solução única, deve-se encontrar outra equação. Esta é a condição que se

tem imposto para os $\alpha_{(j)}$, cuja a soma quadrática deve ser 1, conforme a equação (4), fazendo-se com que o sistema tenha solução única.

É possível reordenar este sistema e obter a transposta de ambos os membros. É preciso transpor linhas por colunas, e se obter então, um vetor (p x 1), ou seja, um vetor coluna em vez de vetor linha:

$$\left[\begin{bmatrix} S_{(11)} & S_{(12)} & ..... & S_{(1p)} \\ S_{(21)} & S_{(22)} & ..... & S_{(2p)} \\ : & ..... & ..... & : \\ : & ..... & ..... & : \\ S_{(p1)} & ..... & ..... & S_{(pp)} \end{bmatrix} + \begin{bmatrix} g & 0 & & ... & 0 \\ 0 & g & 0 & ... & 0 \\ 0 & 0 & g & ... & 0 \\ : & : & : & : & : \\ 0 & 0 & 0 & ... & g \end{bmatrix}\right] x$$

$$x \begin{bmatrix} \alpha_{(1)} \\ \alpha_{(2)} \\ : \\ : \\ \alpha_{(p)} \end{bmatrix} = 0 \qquad (21)$$

onde:

$$(p\,x\,p - p\,x\,p).\ p\,x1 = p\,x1$$

A diferença entre as matrizes não se altera sendo que são simétricas. A matriz e o vetor trocam de posição afim de viabilizar as operações (esta é uma propriedade das operações matriciais).

Comparando-se a equação (21), que reflete a condição que deve cumprir uma transformação linear para sintetizar a máxima variabilidade, com a equação (9), que expressa a condição de que deve satisfazer uma transformação linear para que as variáveis resultantes não estejam correlacionadas observa-se que são idênticas, exceto no caso em que os valores de uma das incógnitas se tem chamado $\lambda_{(1)}$, e no outro g. Em ambos os casos, e como se demonstrou na equação (9), existem p soluções que são os valores próprios gerados por S, cada um dos quais origina um conjunto de valores para o vetor de constantes $\alpha_{(j)}$.

A transformação linear que sintetiza a máxima variabilidade corresponderá, pois, à gerada pelo valor de $\lambda_{(j)}$ que seja maior. Convencionalmente esta solução máxima se tem denominado $\lambda_{(1)}$ é a notação que se utiliza de maneira que se cumpra:

$$\lambda_{(1)} \geq \lambda_{(2)} \geq ... \geq \lambda_{(p)} \tag{22}$$

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

3.4 – NOVA EXPRESSÃO DOS DADOS

Conhecendo-se os valores próprios gerados pela matriz de covariância de um conjunto de dados, é possível calcular todas as constantes que formam a matriz α de transformação, definida na equação (6). Uma vez encontrada esta matriz, é possível após multiplicar a matriz original dos dados X e obter uma nova matriz de dados Y.

Esta matriz de dados transformada, terá as características que mencionou-se antes.

i) Para cada observação o indivíduo terá p valores que correspondem a cada um dos Componentes Principais ou novas variáveis;

ii) A matriz de covariância deste conjunto de dados será diagonal (se as novas variáveis não estiverem correlacionadas) os valores da variância de cada variável serão os valores próprios encontrados ao resolver o polinômio característico da matriz de covariância e dos dados originais;

iii) A variância do primeiro Componente Principal será a maior, e cada um dos seguintes componentes terá uma variância menor, visto que o último componente será o que possui a menor variância;

iv) O vetor centróide da nova matriz também será submetido a mesma transformação linear.

$$\overline{Y}_{1xp} = \overline{X}'_{1xp} . \alpha_{pxp}$$

Utiliza-se a transposta dos vetores centróides, como na definição (2). $\overline{X}$ é um vetor coluna.

Nesta nova expressão de dados tem-se uma série de propriedades que pode-se sintetizar em:

a) $E<Y_{(k)}>=E<X'>\alpha_{(k)}, onde\, \alpha_{(k)}\ é\ \ o\ \ k-ésimo\ \ vetor\ próprio$;

b) $Var<Y_{(k)}>=\lambda_{(k)}, onde\, \lambda_{(k)}\ é\ \ o\ \ k-ésimo\ \ valor\ próprio$;

c) $Cov<Y_{(k)}, Y_{(m)}>=0,\ \ para\, k \neq m$;

d) $Var\ [Y_{(1)}] \geq Var\ [Y_{(2)}] \geq \ ... \ \geq Var\ [Y_{(p)}] \geq 0$;

e) $\sum_{k=1}^{p} Var<Y_{(k)}> = tr\ S$

f) $\prod_{k=1}^{p} Var<Y_{(k)}> = |S|$.

Deseja-se que a média das novas variáveis geradas sejam zero, assim, deve-se efetuar uma translação, e deixar a matriz Y dos novos dados. O valor das médias das novas variáveis como se definiu em a), calcula-se a partir da equação (23). Assim, uma matriz de dados transformados e com média zero será:

$$\begin{aligned} Y^* &= Y - \overline{X}'_{\alpha} = X_{\alpha} - \overline{X}'_{\alpha} \\ Y^* &= [X - 1.\overline{X}'].\alpha \end{aligned} \qquad (23)$$

ao reescrever Y por sua expressão em função dos dados originais, onde 1 representa um vetor coluna de números 1 que, ao multiplicar-se pelo vetor linha $\overline{X}'$, obtém-se uma matriz que permite centrar as variáveis originais. Logo, isola-se à direita a matriz de transformação com fator comum α e obtém-se a equação (23).

3.5 – USO DA MATRIZ DE CORRELAÇÃO

Utilizou-se até agora, a matriz de covariância S dos dados originais, mostrada na definição (3). É possível calcular os valores e os vetores próprios e, portanto, a matriz de transformação α utilizando-se os dados padronizados, em cujo caso a matriz de covariância será a matriz de correlação da definição (4).

Os valores da diagonal principal de r (a matriz de correlação) são os números 1, se as novas variáveis padronizadas possuem variância unitária. Isto significa que o conjunto de dados a partir do qual se geram os Componentes Principais têm a mesma importância para todas as variáveis observadas. Esta situação pode ou não ser desejável, mas importa destacar que o uso da matriz de correlação implica em uma ponderação das variáveis originais, dando a cada uma a mesma importância, independente dos valores relativos de sua variância.

Quando utiliza-se a matriz r, trocam-se algumas das propriedades dos valores e vetores próprios gerados. Das propriedades mencionadas anteriormente cabe destacar que as correspondentes a e) e f) devem expressar-se em termos de r:

e) $\sum_{k=1}^{p} Var\langle Y_{(k)} \rangle = tr\ r = p;$

a matriz r tem números 1 na diagonal principal e dimensão p;

f) $\prod_{k=1}^{p} Var\langle Y_{(k)} \rangle = |r|;$

a matriz de transformação α assim gerada será diferente a partir de S. Esta característica dos valores e vetores próprios determinam que a Análise de Componentes Principais é sensível à troca de escalas. Por esse motivo, é muito importante examinar cuidadosamente os dados originais em suas médias, variâncias e covariâncias a fim de decidir que tipo de matriz convém utilizar. A troca de escala pode-se introduzir antes de encontrar S ou r e que a interpretação deverá dar-se aos componentes encontrados.

Utilizando-se a matriz de correlação para gerar os Componentes Principais deve-se usar a matriz X padronizada (variáveis com média zero e variância unitária) para aplicar a equação (6). O

método de Componentes Principais é sensível à troca de escalas e será impossível obter novas variáveis não correlacionadas (Componentes Principais) usando-se expressões diferentes para o cálculo da matriz α e da matriz Y.

Utilizando os dados padronizados é necessário corrigir-se segundo a equação (23), pois a média de qualquer das variáveis será zero.

O método de Análise de Componentes Principais é utilizado para estudar a estrutura de dependência e consiste em determinar os coeficientes αij; i, j = 1, ..., p encontrando p combinações lineares. Assim,

$$Y1=\sum_{j=1}^{p}\alpha 1\, j\ X_j, \ldots, YP=\sum_{j=1}^{p}\alpha\, p\, j\ X_j \quad \text{onde:}$$

$$Cov(Y_i,Y_j)=0 \ \ para\ i,j=1,\ldots,p; i\neq j$$

$$V(Y_1)\geq V(Y_2)\geq \ldots \geq V(Y_P) \quad \text{e,}$$

$$\sum_{i=1}^{p}V(Y_i)=\sum_{i=1}^{p}\sigma_{i\,i}$$

Destas equações vê-se que as novas variáveis Y1, ..., YP são não correlacionadas e são ordenadas por suas variâncias. Além disso a variância total

$$V=\sum_{i=1}^{p}\sigma_{i\,i}$$

continua igual após a transformação. Desse modo um subconjunto dos primeiros Y_i's pode explicar grande parte da variância total e, portanto, produzir uma descrição parcimoniosa da estrutura de dependência entre as variáveis originais.

Discutiremos primeiro os detalhes do método em termos de parâmetros populacionais e depois em termos de estimação amostrais. Neste tipo de análise não é necessário pressupor uma distribuicão normal multivariada[1], contudo esta distribuição é conveniente desde que as

1 – Ver anexo -03

combinações lineares das variáveis distribuídas normalmente também sejam normais, e desde que sejam completamente determinadas por μ *e* $\sum$.

Pode-se pressupor que $\mu=(0, ..., 0)'$ e explicando a estrutura de dependência dada por $\sum$, iremos explicar completamente a distribuição de X_1, ..., X_p.

Supondo-se que $\sum$ seja conhecida, admite-se que:

$$Y_1 = \alpha_{1\,1}\, X_1 + \ ... + \alpha_{1\,p}\ X_p$$

Deseja-se encontrar $\alpha_{1\,1}$, ... , $\alpha_{1\,p}$, de modo que:

$$V_{(Y\,1)} = \sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_{1\,i}\ \alpha_{1\,j}\ \sigma_{i\,j}$$

seja maximizada sujeita a condição de que

$$\sum_{j=1}^{p} (\alpha_{1\,j})^2 = 1,$$

esta condição garante a singularidade da solução.

A solução de $\alpha_1 = (\alpha_{11}, ... , \alpha_{1\,p})$ é chamada de primeiro autovetor e é associado com o maior autovalor da matriz $\sum$. Este valor é igual para a variância $V_{(Y1)}$. A combinação linear

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

é a primeira Componente Principal de X_1, ..., X_p e explica

$\dfrac{100\ .\ V_{(Y1)}}{V}$ por cento da variância total.

A seguir admite-se que

$$Y2 = \alpha_{21} X_1 + ... + \alpha_{2p} X_p$$

e deseja-se encontrar α_{21}, ... , α_{2p} de modo que

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

seja maximizada sujeita às mesmas condições de que

$$\sum_{j=1}^{p} (\alpha_{2j})^2 = 1 \text{ e,}$$

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

A primeira condição garante uma singularidade da solução e a segunda condição garante que Y_1 e Y_2 sejam não correlacionados.

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

$$Y2 = \alpha_{21} X_1 + ... + \alpha_{2p} X_p$$

é a segunda Componente Principal de X_1, ..., X_p e explica

$\frac{100 \, . \, V_{(Y2)}}{V}$ por cento da variância total.

Os dois componentes explicam a variância total em termos percentuais por:

$$\frac{100\ [\ V_{(Y1)}\ +\ V_{(Y2)}\ ]}{V}$$

De maneira análoga deve-se encontrar os Y_i's componentes restantes.

3.6 – INTERPRETAÇÃO DOS COMPONENTES PRINCIPAIS

Em Análise de Componentes Principais é necessário calcular e interpretar tanto os valores próprios gerados como os vetores próprios. Deve-se decidir quantos valores próprios serão considerados, se desejarmos reduzir a dimensão original de p variáveis a m (sendo m < p). Tem-se que tomar cuidado ao interpretar os vetores próprios, sendo que o método não é independente da escala de medição das variáveis originais.

Do ponto de vista geométrico e espacial é possível conceituar a matriz de dados multivariados de duas maneiras:

i) como um conjunto de n indivíduos (elementos) em um espaço definido por p variáveis e;

ii) como um conjunto de p variáveis definidas em um espaço de n dimensões.

No primeiro caso, as observações serão pontos que representam um indivíduo (elemento) em um espaço definido pelas variáveis (cada j será uma variável); e no 2º cada ponto representará uma variável definida em um espaço cujos j's serão cada um de n indivíduos.

No primeiro caso compara-se elementos considerados em função de suas características, e os vetores linha $X_{(i)}$. Este procedimento, denominado técnica Q, se utiliza em análise discriminante, em análise por conglomerados (que também é possível haver conglomerados de variáveis) e em análise de ordenamento multidimensional.

No segundo caso, ao contrário, se comparam colunas, para obter-se informações a respeito da relação entre características consideradas em função dos elementos, e compara-se vetores $X_{(j)}$ em um espaço de dimensão n. Esta técnica chama-se técnica R, pois a matriz de correlação r deve ser calculada para poder-se iniciar as análises, através da Análise de Componentes Principais, análise fatorial e análise de correlação canônica.

A interpretação geométrica dos Componentes Principais é dada por cada variável X_1, ..., X_p que é representada por uma coordenada no eixo origem $\mu=(\mu_1, ..., \mu_p)'$. Estes p eixos formam um espaço p-dimensional com cada realização $X = (X_1, ..., X_P)'$ representada por um ponto cujas coordenadas são $X_1 = X_1, ..., X_P = X_P$.

Na Análise de Componentes Principais, procura-se uma rotação de eixos de modo que a variável Y_1 representada pelo primeiro novo eixo tenha variância máxima. A variável Y_2 representada pelo segundo novo eixo é não correlacionada com Y_1 e tem variância máxima sob esta restrição. Similarmente, a variável Y_q representada pelo q-ésimo novo eixo é não correlacionada com Y_1, Y_2, ..., Y_{q-1} e tem variância máxima sob estas restrições com q = 3, ..., p.

Admitindo-se que f(x) seja a função densidade normal[2] do vetor aleatório $X = (X_1, ..., X_P)'$, então a desigualdade $f(x) \leq c$, onde c é uma constante, define-se uma região de espaço p-dimensional chamada elipsóide de concentração. Pode-se mostrar que estes Componentes Principais estão na direção dos eixos principais da elipsóide de concentração.

A Figura 1 mostra um espaço bidimensional definido por X_1 e X_2 com a origem em $\mu_1 \; e \; \mu_2$. A elipsóide de concentração é simplesmente uma elipse.

O primeiro Componente Principal $Y_1=\alpha_{11} X_1+\alpha_{12} X_2$ está na direção do maior eixo da elipse, o segundo Componente Principal $Y_2=\alpha_{21} X_1+\alpha_{22} X_2$ está na direção do menor eixo da elipse.

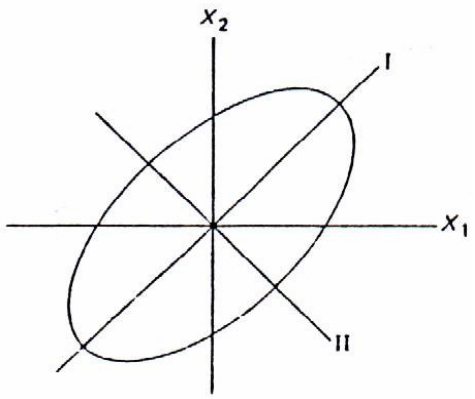

FIGURA 01 – Elipse representando dois componentes principais em um espaço bidimensional.

2 – Ver anexo -03

3.7 – SELEÇÃO DO NÚMERO DE COMPONENTES

Anteriormente mostrou-se que a soma das variâncias e das variáveis originais, e a transformação de S, é igual à soma dos valores próprios da matriz S. Por sua vez, a variância de cada Componente Principal (λ_i) é o valor próprio (α_i) que lhe deu origem. Assim:

$$\sum_{v=1}^{p} S_{(j\ j)} = \sum_{k=1}^{p} \lambda_{(k)} \quad j,\ k = 1,\ ...,\ p$$

do que se reduz das propriedades b) e e) mencionadas no item 3.4.

Cada Componente Principal explicará uma proporção da variabilidade total e essa proporção pode ser calculada, mediante quociente entre o valor próprio e a transformação de S. Este quociente denomina-se proporção da variabilidade total explicada pelo k-ésimo componente e se calcula assim:

$$Variação\ \exp licada = \frac{\lambda_{(k)}}{tr\ S} = \frac{V_{(Y\ k)}}{V} \tag{24}$$

Como os valores próprios se ordenam em forma crescente, é possível selecionar os primeiros m valores próprios (sendo m < p) e a eficiência do ajuste dos dados originais pelos novos m componentes principais será dada pela proporção da variação total explicada pela soma dos m primeiros valores próprios, ou seja:

$$Porcentagem\ \ da\ Variação\ Total = \frac{\sum_{k=1}^{m} \lambda_{(k)}}{tr\ \ S} . 100 \tag{25}$$

Tanto a equação (14) como a (25) podem ser expressas em forma de proporção ou de porcentagem. No 1º caso, a variação total do conjunto original de dados será a unidade e no 2º, será 100.

Aplicando a propriedade e) (item 3.4) das equações anteriores é possível expressar a variação explicada por cada componente ou pelos m primeiros componentes em função da soma total dos valores próprios. Assim, obtém-se:

$$Porcentagem\ de\ Variação\ \text{explicada}\ pelo\ k-ésimo\ componente = \frac{\lambda_{(k)}}{\sum_{k=1}^{p}\lambda_{(k)}}$$

$$Porcentagem\ de\ Variação\ \text{explicada}\ pelos\ m\ componentes = \frac{\sum_{k=1}^{m}\lambda_{(k)}}{\sum_{k=1}^{p}\lambda_{(k)}}$$

Quando se consideram todos os componentes, fazendo-se m = p, a proporção da variação explicada é 1 e a porcentagem é 100%.

Ao decidir quantos componentes deve-se manter em uma situação particular, deve-se examinar quantos componentes são necessários incluir para que a porcentagem de variação explicada seja satisfatória. Não é possível aplicar um teste de hipótese que tenha validade para toda situação e que permita decidir quando se tem alcançado o "nível satisfatório". Em princípio, 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, este critério é devido a CATTEL (1966) e tende a incluir um número alto de componentes.

Outro critério para selecionar o número de componentes, consiste em incluir só aqueles cujos valores próprios sejam superiores à média. Se utilizarmos a matriz r, incluir-se-á os componentes cujos valores próprios sejam maiores que 1, este é o critério de KAISER (citado por MARDIA et alli), o qual tende incluir muito poucos componentes quando o número original de variáveis é inferior a vinte.

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

3.8 – CORRELAÇÃO ENTRE VARIÁVEIS ORIGINAIS E COMPONENTES PRINCIPAIS

No caso bivariado, por exemplo ao estudar um modelo linear de relação entre as variáveis x e y, como na regressão linear simples, o coeficiente de correlação é dado por:

$$r = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)\ .\ \mathrm{var}(y)} \tag{26}$$

A expressão (26) elevada ao quadrado, nos fornece o coeficiente de determinação, que constitui uma medida de associação entre duas variáveis. Esta medida poderá ser comparada com um teste de hipótese para valores próximos de zero quando se supõe binormalidade, efetuando-se um teste para valores absolutos elevados do coeficiente de correlação, (que varia entre -1 e 1). Para aplicar-se um teste de hipótese deve-se dispor de uma amostra grande e aplicar-se a conhecida distribuição assintótica da estatística Z de Fischer calculada a partir de r.

Para estudar-se a correlação entre as variáveis originais e os Componentes Principais tem-se que calcular todas as correlações entre cada variável original e cada nova variável.

Seja X o vetor de dimensão (px1) das p variáveis originais e Y um vetor, também (px1) das nova variáveis (combinações lineares das originais) deve-se encontrar uma expressão da esperança do produto dos vetores, e reescrever o vetor por sua expressão em função da transformação linear que seguem a equação (6).

$$E < X, Y > = E < X\ .\ X'_{\alpha} >$$

Aplicando-se a equação (8) e relembrando que o estimador de $\sum$ é S, e que $\sum$ é a matriz de covariância das variáveis originais, tem-se:

$$E < X, Y' > = E < X, X' > \alpha = \sum \alpha = \alpha \alpha' \sum \alpha = \alpha r$$

Assim, a covariância entre x(j) e y(k) será o elemento da matriz $\alpha . \lambda$ colocado na posição (jk). A matriz α possui na posição (jk) o j-ésimo elemento do k-ésimo vetor próprio, ou seja α(jk). A matriz λ é diagonal e os elementos não são os valores próprios da matriz S, pelo qual:

$$\mathrm{cov} < x(j), y(k) > = \alpha(jk)\,\lambda(k) \quad j, k = 1, ..., p \tag{27}$$

Para calcular-se a correlação deve-se dividir a covariância pela raiz das variâncias das variáveis, ou seja:

$$r(jk) = \frac{\alpha(jk)\ .\ \lambda(k)}{(S(jj)\ .\ \lambda(k))^{1/2}}$$

que pode ser expressa por:

$$r(jr) = \alpha(jk) < \lambda(k) / S(jj) >^{1/2} \tag{28}$$

Utilizando-se a matriz de correlação dos dados originais R, as variâncias S(jj) serão unitárias e a fórmula (28) se reduzirá.

$$r(jr) = \alpha(jk) < \lambda(k) >^{1/2} \tag{29}$$

As equações (28) e (29) representam a correlação entre as variáveis originais x(j) e o k-ésimo Componente Principal. O quadrado do coeficiente de correlação é uma medida de associação entre eles e uma maneira de quantificar a proporção de variação total de uma variável original explicada pelo componente k. Observa-se que o denominador da equação (28) é a raiz da variância da variável x(j). Assim, ao elevar-se quadrado se obtém uma ponderação da variação explicada pela k-ésima combinação linear.

Efetuando-se um somatório em k, e fazendo a soma da variância explicada pelos p Componentes Principais para a variabilidade original x(j), obter-se-á o valor 1. Pode-se efetuar o somatório para os m primeiros componentes incluídos, logo após a seleção destinada a reduzir a dimensão de um conjunto de dados e determinar qual é a proporção da variância de cada variável original, considerando o novo subconjunto. Isto é possível, se os Componentes Principais não estão correlacionados entre si. Assim, em termos de matriz de covariâncias tem-se que:

$$r^2(jk) = \frac{\lambda(k)\,\alpha^2(jk)}{S(jj)} \qquad j\ , k = 1, ..., p \tag{30}$$

para o k-ésimo componente, e somando-se para os primeiros m's:

$$\sum_{k=1}^{m} r^2(jk) = \frac{1}{S(jj)} \sum_{k=1}^{m} \lambda(k) \, . \, \alpha^2(jk)$$

Da análise detalhada destas proporções e dos elementos de cada vetor próprio podem-se tirar as conclusões necessárias para explicar a estruturação de um conjunto de dados multivariados.

3.9 – TESTE DE HIPÓTESE PARA OS VALORES PRÓPRIOS

Ao selecionar os Componentes Principais e considerar uma descrição pode ser útil dispor de um teste para afirmar que os últimos (p – m) componentes são zero. Isto implica que $\wedge$, é na realidade de grau m (existem p – m variáveis originais que são combinações lineares das m restantes). Sem problemas, nesse caso também S, a matriz de covariância estimada a partir de n observações, seria de grau m com probabilidade 1 e um teste de que $\alpha_{(p)}$, (menor valor estimado) é igual a zero seria trivial (MARDIA et alli, 1979).

Quando pode-se supor multinormalidade na população da qual se extrai a amostra é possível verificar os tipos de hipóteses ao calcular-se os valores próprios da matriz S.

Assim,

a) a proporção da variação explicada pelos últimos (p – m) componentes é menor que um certo valor. Isto seria:

$$H_0 : \frac{\sum_{k=1}^{m} \lambda_{(k)}}{\sum_{k=1}^{p} \lambda_{(k)}} = W$$

$$m < p$$

$$H_1 : \frac{\sum_{k=1}^{m} \lambda_{(k)}}{\sum_{k=1}^{p} \lambda_{(k)}} < W$$

O estimador amostral de W é dado por w, e seguirá uma distribuição normal com média W e variância:

$$Var(W) = T^2 = \frac{2\ tr\ \Sigma}{(n-1)\ (tr\ \Sigma)^2}\ (w^2 - 2aw + a^2)\ ,$$

onde:

$$a : \frac{\sum_{k=1}^{m} \lambda^2_{(k)}}{\sum_{k=1}^{p} \lambda^2_{(k)}} \qquad m < p\ .$$

É possível estimar T^2 utilizando a matriz de covariância S e os valores próprios dessa matriz. A matriz quadrada do valor estimado chama-se t e pode ser usada para construir uma estatística de teste

$$Z = \frac{w - W}{t}$$

que tem distribuição normal típica, usando-se com média zero e variância 1, para o qual existem tabelas de probabilidade que permitem tomar decisões em relação à hipótese formulada e construir intervalos de confiança.

b) Testar a hipótese de que os últimos (p – m) valores próprios são iguais.

$$H_0 : \lambda_{(p)} = \lambda_{(p-1)} = ... = \lambda_{(m+1)}$$

Este teste, também conhecido como teste de isotropia implica que as ultimas (p – m) dimensões dos dados estão dispersos em uma hiperesfera e, portanto, para incluir um dos componentes na análise deveria implicar a inclusão de todos os restantes.

A estatística utilizada para testar esta hipótese se obtém pelo método da razão de verossimilhança e da função $-2 \log \alpha^* = n\ .\ p(a\ 1 \log\ g)$ que se distribui aproximadamente como um Qui-quadrado. Sendo L^* a razão de verossimilhança quando se supõe multinormalidade. Onde:

n é o tamanho da amostra;

p é o número total de variáveis observadas;

a é a média aritmética dos valores próprios e;

g a média geométrica dos valores próprios.

Após um arranjo algébrico é possível utilizar-se:

$$a'' = \frac{\sum_{k=m+1}^{p} \alpha_{(k)}}{(p-m)} ,$$

onde:

$\alpha_{(k)}$ é uma estimativa de λk,

a'' é a média aritmética dos últimos (p – m) valores próprios estimados que correspondem aos incluídos na hipótese nula, e

$g'' = < 1\ (m+1)\ .\ 1\ (m+2)\ ...\ 1(p) > (1/(p-m))$ é a média geométrica dos últimos p – m valores próprios estimados, e a estatística pode escrever-se como:

$$-2 \log \alpha^{*} = n\ .\ p(a'' - 1 - \log\ g'')$$

Barlett propõe uma aproximação (citado por MARDIA at alli, (1979)) de tal forma que a equação será:

$$X^{2} = \left[n - \frac{2p + 11}{6} \right] (p-m)\ .\ \log \frac{a''}{g''}$$

distribuindo-se assintoticamente como Qui-quadrado com $1/2(p-m+2)\,(p-m-1)$ graus de liberdade.

4 – MODELOS UNIVARIADOS DE BOX & JENKINS

Neste item será apresentado um resumo sobre a metodologia de Box & Jenkins para modelos univariados. Para maiores detalhes ver Box & Jenkins, 1970; Souza & Camargo, 1991 e Camargo, 1992.

Teoricamente Box & Jenkins assumem que a série temporal Y_t é uma realização particular de um processo estocástico gerado pela passagem sucessiva de um processo ruído branco a_t a uma seqüência de dois filtros lineares, um estável e outro instável, conforme apresentado no diagrama da Figura 02.

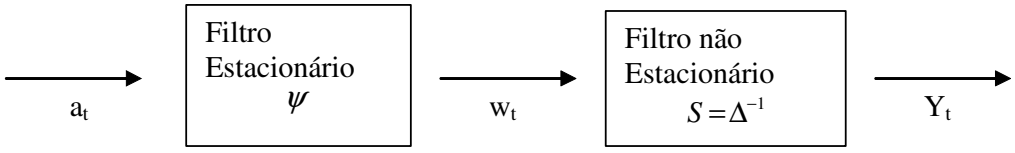

FIGURA 02 – Filtros lineares estacionário e não estacionário

A modelagem de Box & Jenkins, conforme originalmente proposta pelos autores em 1970, 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:

i) Identificação da estrutura do modelo, realizada com base na análise da função de autocorrelação parcial, etc...

ii) Estimação dos parâmetros do modelo, isto é, $\phi's, \theta's \; e \; \sigma_a^2$ (ou $\phi's, \theta's$ no caso sazonal) através da maximização da função de verossimilhança condicional, equivalente à estimação por mínimos quadrados assumindo normalidade de a_t.

iii) Verificação do modelo ajustado: através de uma análise de resíduos para se verificar a adequabilidade do modelo para previsões.

iv) Previsão dos valores futuros $Y_T(1), ..., Y_T(\ell)$ através do modelo devidamente testado.

4.1 – CRITÉRIOS DE AJUSTAMENTO

ENGLE e BROWN (1986) fazem uma descrição dos critérios aqui apresentados e analisam empiricamente os erros de previsão para os modelos selecionados por diferentes critérios tais como:

a) Critério de informação de AKAIKE (AIC), é dado por:

AIC = -2 log (máxima verossimilhança) + 2 (número de parâmetros)

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

O BIC é dado por:

$$BIC = \log(\hat{\sigma}_a^2) + \frac{\log(N)K}{N} \tag{31}$$

onde: K é o número de parâmetros

N é o numero de resíduos

$\hat{\sigma}_a^2$ é o estimador de máxima verossimilhança da variância residual

c) Coeficiente de determinação (R^2)

i) Regressão

Define-se o coeficiente de determinação "R^2" como sendo a Soma dos Quadrados dos Resíduos estimados (SQR), comparando-a com a correspondente Soma dos Quadrados Totais (SQT).

Assim, R^2 pode ser interpretado como a fração da variação total (SQT) que é explicada pela reta de regressão de mínimos quadrados, ou seja, mede o percentual de explicação do modelo com relação a variação da série original e a variância residual ($\hat{\sigma}_a^2$), dado por:

$$R^2 = 1 - \frac{SQR}{SQT} \tag{32}$$

ii) Série temporal univariada não sazonal

Verifica-se a adequabilidade do modelo de séries temporais, comparando-se os resíduos estimados pelo modelo escolhido, com aqueles obtidos pelo modelo univariado mais simples, ou seja, passeio aleatório com fator de crescimento constante β, dado por:

$$Y_t = Y_{t-1} + \beta + \varepsilon_t$$

O coeficiente de determinação R^2D, é dado por:

$$R^2 D = 1 - [SQR / \sum_{t=2}^{N} (\Delta Y_t - \overline{\Delta} \overline{Y}_t)^2] \qquad (33)$$

onde $\overline{\Delta} \overline{Y}$ é a média da série da 1ª diferença simples ($Y_t - Y_{t-1}$).

iii) série temporal univariada sazonal

Seja s o período sazonal e considerando-se a versão sazonal do modelo simples, ou seja o passeio aleatório dado por:

$$\Delta Y_t = \sum \gamma_j \, Y_{tj} + \varepsilon_t; \qquad t = 2, 3, \ldots, T$$

onde: $Y_{tj} = 0$ para todo t diferente do mês j

γ_j, j = 1, 2, ..., s são os fatores sazonais correspondentes

Aplicando o método dos mínimos quadrados a equação (* *) e seja SQRO a soma dos quadrados dos resíduos correspondentes. Assim, se SQR é a soma dos quadrados dos resíduos do modelo, o coeficiente de determinação pode ser expresso por:

$$R^2 S = 1 - \frac{SQR}{SQRO} \qquad (34)$$

4.2 – Critérios de Previsão

Para avaliar-se o erro de previsão nesta dissertação será usado o Erro Médio Percentual de Previsão (EMPP)

$$e_t(h) = Y_{t+h} - \hat{Y}_t(h). \tag{36}$$

onde e_t (h) é o erro de previsão h passos-à-frente.

5 – APLICAÇÕES E DISCUSSÕES

Neste item apresentaremos a análise empírica das variáveis macroeconômicas, através das metodologias de componentes principais e séries temporais.

5.1 – INTRODUÇÃO

Os dados usados na análise empírica foram coletados na revista Conjuntura Econômica, os quais representam a política monetária, política de preços, atividade econômica, agregados de crédito e balança comercial.

5.2 – VARIÁVEIS UTILIZADAS NA ANÁLISE EMPÍRICA

As variáveis analisadas estão compreendidas no período amostral de janeiro de 1985 a dezembro de 1991, sendo as seguintes:

1) Política monetária, composta pelos meios de pagamento, nos conceitos (M1, M2, M3, M4) e Base Monetária.

O conceito das séries representativas de moeda no Brasil, são:

M1: Papel moeda em poder do público + depósitos à vista no setor bancário.

M2: M1 + depósitos a prazo de títulos federais (exclui a carteira própria do banco central e das instituições financeiras).

M3: M2 + depósitos de poupança.

M4: M3 + saldo dos títulos públicos federais em circulação (exclui a carteira do banco central).

Base Monetária: É o somatório do saldo em papel-moeda mais reservas em moeda dos bancos comerciais, bancos do Brasil e caixas econômicas abrangendo somente as contas do banco central do Brasil. Sendo a razão entre os meios de pagamentos e o multiplicador bancário.

2) Política de preços: Composta pelos seguintes índices: IGP-DI (Índice Geral de Preços – Disponibilidade Interna), IGP-OG (Índice Geral de Preços Oferta Global), IPA-DI (Índice de Preços por Atacado – Disponibilidade Interna), IPA-OG (Índice de Preços por Atacado Oferta Global), IPA-

> OG-PI (Índice de Preços por Atacado Oferta Global de Produtos Industrializados), IPA-OG-PA (Índice de Preços por Atacado Oferta Global de Produtos Agrícola).

Onde:

O IGP representa três atividades no país que são as operações em geral, preços de varejo e a construção civil;

O IPA representa cada tipo de operação na formação da despesa interna bruta que são: a produção, transporte e comercialização a grosso modo de bens de consumo e de produção.

O IGP é formado pelo IGP-DI e IGP-OG, sendo que o IGP-DI é formado em 60% pelo IPA, 30% pelo IPC-RJ mais 10% do INCC, o somatório de todos estes índices do IGP-DI formam o IGP-OG, obtendo-se assim um total de 100%.

O IPA é formado também pelo IPA-DI e IPA-OG, sendo assim as suas composições:

O IPA-DI é formado por bens de consumos, que são os duráveis (utilidades domésticas e outros) e não-duráveis (gêneros alimentícios e outros), com uma participação de 55.8450% e por bens de produção que é constituído de matéria-prima, materiais de construção, máquinas, veículos, equipamentos e outros com uma contribuição de 44.1550%. O somatório de bens de consumo e bens de produção formam o IPA-DI totalizando 100%.

O IPA-OG é formado pelo IPA-OG-PA e IPA-OG-PI, sendo que o IPA-OG-PI contribui para a formação do índice com 30.6292% e o IPA-OG-PI que é a indústria de transformação e a extrativa mineral sendo responsável por 69.37708% para a composição do índice, o somatório destes dois índices formam o IPA-OG com um total de 100%.

3) Atividade Econômica composta pelas variáveis: Produção Industrial (base fixa), Indicador do Nível de Atividade FIESP (Federação da Industrias de São Paulo) e Uso da Capacidade Instalada.

4) Agregados de Crédito, composto pelos empréstimos concedidos em milhões de cruzeiros ao Setor Privado que são: Créditos concedidos ao Banco do Brasil e aos Bancos Comerciais e o total de empréstimos.

5) Balança Comercial que é composta pelo: Saldo da Balança Comercial, Exportações (Básicos, Industrializados, Semimanufaturados, Manufaturados e Transações Especiais), Importações (Petróleo Bruto e Demais).

Para o desenvolvimento deste estudo, foi utilizado o seguinte procedimento:

- cálculo da média, desvio-padrão e do coeficiente de variação de cada variável;

- cálculo da matriz de covariância, utilizando os dados originais;

- cálculo de R a partir de S;

- cálculo dos valores e vetores próprios com a sua respectiva proporção de variação total explicada, para cada um dos componentes;

- cálculo do coeficiente de correlação entre as variáveis e os componentes principais;

- cálculo da proporção da variação original explicada para cada componente principal;

- seleção dos componentes principais a serem utilizados na combinação linear;

- cálculo da correlação das variáveis originais com os componentes selecionados.

Nesta análise, todas as variáveis foram consideradas com variação percentual mensal e a matriz de correlação utilizada foi a padronizada.

5.3 – ANÁLISE DA POLÍTICA MONETÁRIA

Para a política monetária, composta pelas seguintes variáveis: MEIOS DE PAGAMENTOS nos conceitos M1 (X1), M2 (X2), M3 (X3), M4 (X4) e BASE MONETÁRIA (X5), inicialmente foi feita uma análise univariada, cujas estatísticas estão apresentadas na Tabela 01.

TABELA 01 – Estatísticas univariadas

Variáveis	Média	Desvio-Padrão	Coef. De Variação
X1	19.2902	28.8446	149.5297
X2	16.6750	14.7777	88.6219
X3	17.3059	14.3727	83.0511
X4	17.4869	13.4615	76.9807
X5	18.7103	24.4396	130.6213

Pode-se observar que os altos valores do coeficiente de variação, indicam uma grande heterogeneidade dentro de cada variável, sendo que a variável que apresentou menor variabilidade no período analisado foi a M4.

A seguir foi realizada a análise de componentes principais, determinando-se o dos autovalores e autovetores para cada componente, que estão apresentados nas Tabelas 02 e 03 respectivamente.

$$Autovalores = \frac{\text{var. explicada pelo CP } x \text{ var. total}}{100}$$

TABELA 02 – Relação dos autovalores da política monetária

componentes	1	2	3	4	5
autovalores	3.1181	1.5228	0.1792	0.1316	0.0294

TABELA 03 – Coeficientes dos cinco componentes principais

	Autovetores				
Variáveis	α_1	α_2	α_3	α_4	α_5
X1	0.3294	-0.6091	0.6652	0.2754	-0.0460
X2	0.2952	-0.6485	-0.5659	-0.4074	0.0776
X3	0.5265	0.1514	-0.4233	0.7201	-0.0451
X4	0.5161	0.2920	0.1317	-0.4041	-0.6839
X5	0.5106	0.3167	0.2015	-0.2762	0.7225

A variância total é igual a 5, pois temos 5 variáveis, como pode ser encontrado, fazendo-se o somatório dos autovalores.

Para encontrarmos o que cada componente explica da variação total, basta fazermos:

$$\frac{100 \,.\, (3.1181)}{5} = 62.3620\%$$

Assim, o primeiro componente explica 62.3620% da variância total, e pode ser escrito na forma de combinação linear da seguinte maneira:

$$Y1 = 0.3294\ X1 + 0.2952\ X2 + 0.5265\ X3 + 0.5161\ X4 + 0.5106\ X5$$

Pode-se verificar que a condição de singularidade foi mantida, ou seja, o somatório de todos os autovetores elevado ao quadrado somam a unidade.

O segundo componente explica:

$$\frac{100\ .\ (1.5228)}{5} = 30.4560\%$$

Vimos que o segundo componente explica 30.4560% da variância total, e pode ser escrito em combinação linear da seguinte maneira:

$$Y2 = -0.6091\ X1 - 0.6485\ X2 + 0.1514\ X3 + 0.2920\ X4 + 0.3167\ X5$$

A condição de singularidade também foi mantida, ou seja, o somatório de todos os autovetores elevado ao quadrado somam a unidade.

A variância total explicada por estes dois componentes é de:

$$\frac{100\ .\ [V(Y1) + V(Y2)]}{V} = 92.8180\%$$

A política monetária foi explicada em quase 100% em apenas dois componentes principais, reduzindo a dimensionalidade do problema de cinco para duas variáveis escritas em combinação linear.

Similarmente os outros componentes são escritos da mesma maneira.

Na Tabela 04 mostra-se o percentual de explicação da variância total pelos seis componentes encontrados para a política monetária e o percentual de variância acumulada.

TABELA 04 – Percentual da variância explicada por cada componente e o percentual acumulado.

Número de componentes	Percentual da variância	Percentual acumulada da variância
C1	62.33625	62.34
C2	30.85808	93.19
C3	3.58518	96.78
C4	2.63232	99.41
C5	0.58818	100.00

Para se determinar quais as variáveis que mais contribuem para cada componente principal, pode-se analisar a relação $\left|\alpha_{ji}\ [V(Yj)]^{1/2}\right|$, com i = j = 1, ..., 5, fornecendo assim o valor da correlação absoluta ou utilizando-se a matriz de correlação, fornecendo direto a contribuição de cada variável para a formação da nova variável.

$$[V(Y1)]^{1/2} = (3.1181)^{1/2} = 1.7658$$
$$[V(Y2)]^{1/2} = (1.5228)^{1/2} = 1.2340$$
$$[V(Y3)]^{1/2} = (0.1792)^{1/2} = 0.4233$$
$$[V(Y4)]^{1/2} = (0.1316)^{1/2} = 0.3628$$
$$[V(Y5)]^{1/2} = (0.0294)^{1/2} = 0.1715$$

Para se determinar a correlação absoluta entre os autovetores e os componentes principais deve-se fazer:

$$X1 \rightarrow |0.3294\ x\ 1.7658| = 0.5816$$
$$X2 \rightarrow |0.2952\ x\ 1.2340| = 0.3643$$
$$X3 \rightarrow |0.5265\ x\ 0.4233| = 0.2229$$
$$X4 \rightarrow |0.5161\ x\ 0.3628| = 0.1872$$
$$X5 \rightarrow |0.5106\ x\ 0.1715| = 0.0876$$

A variável que maior contribuição ofereceu para a primeira combinação linear foi a variável X1, seguida das variáveis X2, X3, X4 e X5, então pode-se dizer que Y1 possui características na sua grande maioria da variável MEIOS DE PAGAMENTOS – M1.

O componente C1, que mais agregou explicação ao desconhecimento inicial a cerca da variabilidade dos dados, apresentou correlação mais alta com meios de pagamentos.

Para a segunda componente principal temos:

$$X1 \rightarrow |-0.6091\ x\ 1.7658| = 1.0755$$
$$X2 \rightarrow |-0.6485\ x\ 1.2340| = 0.8002$$
$$X3 \rightarrow |0.1514\ x\ 0.4233| = 0.0641$$
$$X4 \rightarrow |0.2920\ x\ 0.3628| = 0.1059$$
$$X5 \rightarrow |0.3167\ x\ 0.1715| = 0.0543$$

A variável que maior contribuição ofereceu para a segunda combinação linear foi a variável X1, seguida das variáveis X2, X4, X3 e X5, então podemos dizer que Y2 possui características na sua grande maioria da variável MEIOS DE PAGAMENTOS no conceito M1.

5.4 – ANÁLISE DA POLÍTICA DE PREÇOS

Para a política de preços, composta pelas seguintes variáveis: IGP-DI (X1), IGP-OG (X2), IPA-DI (X3), IPA-OG (X4), IPA-OG-PA (X5), IPA-OG-PI (X6), inicialmente foi feita uma análise univariada, cujas estatísticas estão apresentadas na Tabela 05.

TABELA 05 – Estatísticas univariadas

Variáveis	Média	Desvio-Padrão	Coef. De Variação
X1	17.7670	15.2964	86.0944
X2	28.5420	114.2405	400.2535
X3	17.7967	15.6728	88.0659
X4	17.6848	15.6133	88.2862
X5	17.9318	15.5276	86.5926
X6	18.8753	17.3590	91.9667

Pode-se observar que os altos valores do coeficiente de variação, indicam uma grande heterogeneidade dentro de variável, sendo que a variável que apresentou menor variabilidade no período analisado foi o IGP-DI.

A seguir foi realizada a análise de componentes principais, determinando-se o valor dos autovalores e autovetores para cada componente, que estão apresentados nas Tabelas 06 e 07 respectivamente.

$$Autovalores = \frac{\text{var. exp}licada\ \ pelo\ \ CP\ \ x\ \ \text{var. }total}{100}$$

TABELA 06 – Relação dos autovalores da variável política de preços

componentes	1	2	3	4	5	6
autovalores	4.6637	0.9878	0.2650	0.07694	0.0049	0.0016

Na Tabela 07 apresenta-se os autovetores das seis componentes principais.

TABELA 07 – Coeficientes das seis componentes principais

	Autovetores					
Variáveis	α_1	α_2	α_3	α_4	α_5	α_6
X1	0.4593	-0.0238	-0.1056	-0.3472	0.0778	0.2265
X2	0.0586	0.9980	0.0213	0.0101	-0.0033	0.0004
X3	0.4599	-0.0203	-0.1053	-0.3376	0.2079	-0.7872
X4	0.4603	-0.0205	-0.0688	-0.3219	0.5922	0.5731
X5	0.4143	-0.0458	0.8443	0.3352	-0.0240	-0.0198
X6	0.4364	-0.0223	-0.5096	0.7411	-0.0016	0.0054

A variância total é igual a 6, pois temos 6 variáveis, como pode ser encontrado fazendo-se o somatório dos autovalores.

Para encontrar-se a explicação de cada componente na variação total, basta fazermos:

$$\frac{100\ .\ (4.6637)}{6} = 77.7385\%$$

Assim, o primeiro componente explica 77.7285% da variância total e pode ser escrito na forma de combinação linear da seguinte maneira:

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

Pode-se verificar que a condição de singularidade foi mantida, ou seja, o somatório de todos os autovetores elevado ao quadrado somam a unidade.

O segundo componente explica:

$$\frac{100\ .\ (0.9878)}{6} = 16.46374\%$$

Vimos que o segundo componente explica 16.46364% da variância total, e pode ser escrito em combinação linear da seguinte maneira:

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

A condição de singularidade também foi mantida, ou seja, o somatório de todos os autovetores elevado ao quadrado somam a unidade.

A variância total explicada por estes dois componentes é de:

$$\frac{100\ .\ [V(Y1) + V(Y2)]}{V} = 94.19\,\%$$

A política de preços foi explicada em quase 100% com apenas dois componentes principais, reduzindo a dimensionalidade do problema de seis para duas variáveis escritas em combinação linear.

Similarmente os outros componentes são escritos da mesma maneira.

Na Tabela 08 mostra-se o percentual de explicação da variância total pelos seis componentes encontrados para a política de preços e o percentual de variância acumulada.

TABELA 08 – Percentual da variância explicada para cada componente e o percentual acumulado.

Número de componentes	**Percentual da variância**	**Percentual acumulada da variância**
C1	77.72854	77.73
C2	16.46374	94.19
C3	4.41759	98.61
C4	1.28235	99.89
C5	0.08109	99.97
C6	0.02670	100.00

Para se determinar quais as variáveis que mais contribuem para cada componente principal, pode-se analisar a relação $\left|\alpha_{ji}\ [V(Yj)]^{1/2}\right|$, com i = j = 1, ..., 6, fornecendo assim o valor da correlação absoluta ou utilizando-se a matriz de correlação, fornecendo direto a contribuição de cada variável para a formação da nova variável.

$$[V(Y1)]^{1/2} = (4.6637)^{1/2} = 2.1596$$
$$[V(Y2)]^{1/2} = (0.9878)^{1/2} = 0.9939$$
$$[V(Y3)]^{1/2} = (0.2650)^{1/2} = 0.5148$$
$$[V(Y4)]^{1/2} = (0.0769)^{1/2} = 0.2774$$
$$[V(Y5)]^{1/2} = (0.0049)^{1/2} = 0.0697$$
$$[V(Y6)]^{1/2} = (0.0016)^{1/2} = 0.0402$$

Para se determinar a correlação absoluta entre os autovetores e os componentes principais devemos fazer:

$$X1 \rightarrow |0.4593\ x\ 2.1596| = 0.9919$$
$$X2 \rightarrow |0.0586\ x\ 0.9936| = 0.0585$$
$$X3 \rightarrow |0.4599\ x\ 0.5148| = 0.2368$$
$$X4 \rightarrow |0.4603\ x\ 0.2774| = 0.1277$$
$$X5 \rightarrow |0.4143\ x\ 0.0697| = 0.0289$$
$$X6 \rightarrow |0.4364\ x\ 0.0400| = 0.0175$$

A variável que maior contribuição ofereceu para a primeira combinação linear foi a variável X1, seguida das variáveis X3, X4, X2, X5 e X6, então pode-se dizer que Y1 possui características na sua grande maioria da variável IGP-DI.

O componente C1, que mais agregou explicação ao desconhecimento inicial a cerca da variabilidade dos dados, apresentando a correlação mais alta com os índices de preços por atacado.

Para a segunda componente principal temos:

$$X1 \rightarrow |-0.0238\ x\ 2.1596| = 0.0514$$
$$X2 \rightarrow |0.9980\ x\ 0.9936| = 0.9916$$
$$X3 \rightarrow |-0.0203\ x\ 0.5148| = 0.0105$$
$$X4 \rightarrow |-0.0205\ x\ 0.2774| = 0.0057$$
$$X5 \rightarrow |-0.0458\ x\ 0.0697| = 0.0032$$
$$X6 \rightarrow |-0.0223\ x\ 0.0400| = 0.0009$$

A variável que maior contribuição ofereceu para a segunda combinação linear foi a variável X2, seguida das variáveis X1, X3, X4, X5 e X6, então podemos dizer que Y2 possui características na sua grande maioria da variável IGP-OG.

5.5 – ATIVIDADE ECONÔMICA

Nesta análise considerou-se a atividade econômica, composta pelas seguintes variáveis: PRODUÇÃO INDUSTRIAL – BASE FIXA (X1), INDICADOR DO NÍVEL DE ATIVIDADE (X2) e USO DA CAPACIDADE INSTALADA (X3), na Tabela 09 apresenta-se as estatísticas univariadas para cada uma das variáveis que representam a atividade econômica.

TABELA 09 – Estatísticas univariadas

Variáveis	**Média**	**Desvio-Padrão**	**Coef. De Variação**
X1	0.3536	9.3975	2657.6640
X2	0.1883	5.8721	3118.4811
X3	0.6800	3.7003	544.1617

Pode-se observar que todas as variáveis apresentam grande heterogeneidade, e a variável menos heterogênea no período analisado foi o uso da capacidade instalada.

A seguir foi realizada a análise por componentes principais, determinando-se o valor dos autovalores e autovetores para cada componente, que estão relacionados nas Tabelas 10 e 11 respectivamente.

$$Autovalores = \frac{\text{var. explicada pelo CP } x \text{ var. total}}{100}$$

TABELA 10 – Relação dos autovalores da atividade econômica

componentes	1	2	3
autovalores	2.0703	0.8245	0.1052

Na Tabela 11 apresenta-se os autovetores das três componentes principais.

TABELA 11 – Coeficientes das três componentes principais

Variáveis	*Autovetores*		
	α_1	α_2	α_3
X1	0.6589	-0.2406	-0.7127
X2	0.6514	-0.2914	0.7006
X3	0.3763	0.9258	0.0353

A variância total é igual a 3, pois temos 3 variáveis, que pode ser encontrada fazendo-se o somatório dos autovalores.

Para encontrarmos a explicação de cada componente na variação total, basta fazermos:

$$\frac{100\ .\ (2.0703)}{3} = 69.0100\,\%$$

Pode-se notar que o primeiro componente explica 69.01% da variância total, e pode ser escrito na forma de combinação linear da seguinte forma:

$$Y1 = 0.6589\ X1 + 0.6513\ X2 + 0.3762\ X3$$

A condição de singularidade foi mantida, ou seja, o somatório de todos os autovetores elevado ao quadrado somam a unidade.

O segundo componente explica:

$$\frac{100\ .\ (0.8245)}{3} = 36.3857\,\%$$

Vimos que o segundo componente explica 36.3857% da variância total, podendo ser escrito em combinação linear da seguinte forma:

$$Y2 = -0.2406\ X1 - 0.2914\ X2 + 0.9258\ X3$$

A condição de singularidade também foi mantida, ou seja, o somatório de todos os autovetores elevado ao quadrado somam a unidade.

A variância total explicada por estes dois componentes é de:

$$\frac{100\ .\ [V(Y1) + V(Y2)]}{V} = 96.4933\%$$

Pode-se verificar que a variável atividade econômica foi explicada em 96.4933% com apenas dois componentes principais, reduzindo-se a dimensionalidade do problema de três para duas variáveis escritas em combinação linear.

Similarmente o outro componente é escrito da mesma maneira.

Na Tabela 12 mostra-se o percentual de explicação da variância total pelos três componentes encontrados para a atividade econômica e o percentual de variância acumulada.

TABELA 12 – Percentual da variância explicada por cada componente e o percentual acumulado.

Número de componentes	**Percentual da variância**	**Percentual acumulada da variância**
C1	69.00976	69.00978
C2	27.48357	96.49333
C3	3.50667	100.00000

Para se determinar quais as variáveis que mais contribuem para a formação de cada componente principal, pode-se analisar a relação $\left|\alpha_{ji}\ [V(Yj)]^{1/2}\right|$, com i = j = 1, ..., 3, fornecendo assim o valor da correlação absoluta ou utilizando-se a matriz de correlação, fornecendo direto a contribuição de cada variável para a formação da nova variável.

$$[V(Y1)]^{1/2} = (2.0703)^{1/2} = 1.4388$$
$$[V(Y2)]^{1/2} = (0.8245)^{1/2} = 0.9080$$
$$[V(Y3)]^{1/2} = (0.1052)^{1/2} = 0.3243$$

Para se determinar a correlação absoluta entre os autovetores e os componentes principais devemos fazer:

$$X1 \rightarrow |0.6589\ x\ 1.4388| = 0.9480$$
$$X2 \rightarrow |0.6513\ x\ 0.9080| = 0.5913$$
$$X3 \rightarrow |0.3762\ x\ 0.3243| = 0.1220$$

A variável que maior contribuição ofereceu para a primeira combinação linear foi a variável X1, seguida das variáveis X2 e X3, então podemos dizer que Y1 possui características na sua grande maioria da variável PRODUÇÃO INDUSTRIAL.

O componente C1, que mais agregou explicação ao desconhecimento inicial a cerca da variabilidade dos dados, apresentou correlação mais alta com produção industrial.

Para a segunda componente principal temos:

$$X1 \rightarrow |-0.2185 \; x \; 1.4388| = 0.3143$$
$$X2 \rightarrow |-0.2646 \; x \; 0.9080| = 0.2402$$
$$X3 \rightarrow |0.9258 \; x \; 0.3243| = 0.3002$$

A variável que maior contribuição ofereceu para a segunda combinação linear foi a variável X1, seguida das variáveis X2 e X3, então pode-se dizer que Y2 possui características na sua grande maioria da variável PRODUÇÃO INDUSTRIAL.

O componente C2, que mais agregou explicação ao desconhecimento inicial a cerca da variabilidade dos dados, apresentou correlação mais alta com uso da capacidade instalada.

5.6 – ANÁLISE DOS AGREGADOS

Para os agregados de créditos, compostos pelas seguintes variáveis: CRÉDITO TOTAL (X1), BANCOS COMERCIAIS (X2) e BANCO DO BRASIL (X3), inicialmente foi feita uma análise univariada cujas estatísticas estão apresentados na Tabela 13.

TABELA 13 – Estatísticas univariadas

Variáveis	Média	Desvio-Padrão	Coef. De Variação
X1	16.7350	13.8876	82.9853
X2	15.4674	101464	65.5990
X3	17.7910	13.1564	73.9496

Pode-se observar que os altos valores para o coeficiente de variação, indicam uma grande heterogeneidade dentro de cada variável, sendo que a variável que apresentou menor variabilidade no período analisado foi a variável bancos comerciais.

A seguir foi realizada a análise de componentes principais, determinando-se o valor dos autovalores e autovetores para cada componente, que estão apresentados nas Tabelas 14 e 15.

$$Autovalores = \frac{\text{var. exp}licada\ \ pelo\ \ CP\ \ x\ \ \text{var. }total}{100}$$

TABELA 14 – Relação dos autovalores da balança comercial

componentes	1	2	3
autovalores	2.2601	0.5118	0.2280

TABELA 15 – Coeficientes das três componentes principais

	Autovetores		
Variáveis	α_1	α_2	α_3
X1	0.6155	0.0835	0.7836
X2	0.5467	-0.7614	-0.3483
X3	0.5675	0.6428	-0.5144

A variância total é igual a 3, pois temos 3 variáveis, ou simplesmente, fazendo-se o somatório dos autovalores.

Para descobrir-se a explicação de cada componente na variação total, basta fazermos:

$$\frac{100\ .\ (2.2601)}{3} = 75.3366\%$$

Assim, o primeiro componente explica 75.3366% da variância total, e pode ser escrito na forma de combinação linear da seguinte maneira:

$$Y1 = 0.6155\ X1 + 0.5467\ X2 + 0.5675\ X3$$

Pode-se verificar que a condição de singularidade foi mantida, ou seja, o somatório de todos os autovetores elevado ao quadrado somam a unidade.

O segundo componente explica:

$$\frac{100\ .\ (0.5118)}{3} = 17.6000\%$$

Vimos que o segundo componente explica 17.6000% da variância total, e pode ser escrito em combinação linear da seguinte maneira:

$$Y2 = 0.0835\ X1 - 0.7614\ X2 + 0.6428\ X3$$

Pode-se verificar que a condição de singularidade foi mantida, ou seja, o somatório de todos os autovetores elevado ao quadrado somam a unidade.

A variância total explicada por estes dois componentes é de:

$$\frac{100\ .\ [V(Y1) + V(Y2)]}{V} = 92.9366\%$$

A variável investimentos foi explicada em quase 100% em apenas dois componentes principais, reduzindo a dimensionalidade do problema de três para duas variáveis escritas em combinação linear.

Similarmente os outros componentes são escritos da mesma maneira.

Na Tabela 16 mostra-se o percentual de explicação da variância total pelos três componentes encontrados para a variável política monetária e o percentual de variância acumulada.

TABELA 16 – Percentual da variância explicada por cada componente e o percentual acumulado.

Número de componentes	**Percentual da variância**	**Percentual acumulada da variância**
C1	75.33853	75.33853
C2	17.06249	92.40101
C3	7.59899	100.00000

Para se determinar quais as variáveis contribuem mais para cada componente principal, podemos analisar a relação $\left|\alpha_{ji}\ [V(Yj)]^{1/2}\right|$, com i = j = 1, ..., 3, fornecendo assim o valor da correlação absoluta ou utilizando-se a matriz de correlação, fornecendo direto a contribuição de cada variável para a formação da nova variável.

$$[V(Y1)]^{1/2} = (2.2601)^{1/2} = 1.5034$$
$$[V(Y2)]^{1/2} = (0.5118)^{1/2} = 0.7154$$
$$[V(Y3)]^{1/2} = (0.2280)^{1/2} = 0.4775$$

Para determinarmos a correlação absoluta entre os autovetores e os componentes principais devemos fazer:

$$X1 \rightarrow |0.6155 \; x \; 1.5034| = 0.9253$$
$$X2 \rightarrow |0.5467 \; x \; 0.7154| = 0.3911$$
$$X3 \rightarrow |0.5675 \; x \; 0.4775| = 0.2710$$

A variável que maior contribuição ofereceu para a primeira combinação linear foi a variável X1, seguida das variáveis X2 e X3, então podemos dizer que Y1 possui características na sua grande maioria da variável CRÉDITO TOTAL.

O componente C1, que mais agregou explicação ao desconhecimento inicial a cerca da variabilidade dos dados, apresentou correlação mais alta com o crédito total.

Para a segunda componente principal temos:

$$X1 \rightarrow |0.0835 \; x \; 1.5034| = 0.1255$$
$$X2 \rightarrow |-0.7614 \; x \; 0.7154| = 0.5447$$
$$X3 \rightarrow |0.6428 \; x \; 0.4775| = 0.3069$$

A variável que maior contribuição ofereceu para a segunda combinação linear foi a variável X2, seguida das variáveis X3 e X1, então pode-se dizer que Y2 possui características na sua grande maioria da variável BANCOS COMERCIAIS.

5.7 – ANÁLISE DA BALANÇA COMERCIAL

Para a balança comercial, composta pelas seguintes variáveis: SALDO DA BALANÇA COMERCIAL (X1), IMPORTAÇÕES DE COMBUSTÍVEIS E LUBRIFICANTES (X2) e EXPORTAÇÕES DE PRODUTOS INDUSTRIALIZADOS (X3), inicialmente foi feita uma análise univariada cujas estatísticas estão apresentados na Tabela 17.

TABELA 17 – Estatísticas univariadas

Variáveis	Média	Desvio-Padrão	Coef. De Variação
X1	0.3354	5.8983	1758.1016
X2	0.0364	0.3572	981.3186
X3	0.0129	0.1518	1176.7442

Pode-se observar que os altos valores para o coeficiente de variação, indicam uma grande heterogeneidade dentro de cada variável, sendo que a variável que apresentou menor variabilidade no período analisado foi a importação de combustíveis e lubrificantes.

A seguir foi realizada a análise de componentes principais, determinando-se o valor dos autovalores e autovetores para cada componente, que estão apresentados nas Tabelas 18 e 19, respectivamente.

$$Autovalores = \frac{\text{var. explicada pelo CP } x \text{ var. total}}{100}$$

TABELA 18 – Relação dos autovalores da balança comercial

componentes	1	2	3
autovalores	1.3175	1.0820	0.6004

Na Tabela 19 apresenta-se os autovetores das duas componentes principais.

TABELA 19 – Coeficientes das duas componentes principais

Variáveis	*Autovetores*		
	α_1	α_2	α_3
X1	0.6811	-0.3725	-0.6302
X2	0.1054	0.9018	-0.4192
X3	0.7245	0.2191	0.6535

A variância total é igual a 3, pois temos 3 variáveis, ou simplesmente, fazendo-se o somatório dos autovalores.

Para encontrarmos o que cada componente explica da variação total, basta fazermos:

$$\frac{100 \,.\, (1.3175)}{3} = 43.9179\,\%$$

Assim, o primeiro componente explica 43.9179% da variância total, e pode ser escrito na forma de combinação linear da seguinte maneira:

$$Y1 = 0.6811\ X1 + 0.1054\ X2 + 0.7245\ X3$$

Pode-se verificar que a condição de singularidade foi mantida, ou seja, o somatório de todos os autovetores elevado ao quadrado somam a unidade.

O segundo componente explica:

$$\frac{100\ .\ (1.0820)}{3} = 36.0067\,\%$$

Vimos que o segundo componente explica 36.0067% da variância total, e pode ser escrito em combinação linear da seguinte maneira:

$$Y2 = -0.9018\ X1 + 0.0918\ X2 + 0.2191\ X3$$

A condição de singularidade também foi mantida, ou seja, o somatório de todos os autovetores elevado ao quadrado somam a unidade.

A variância total explicada por estes dois componentes é de:

$$\frac{100\ .\ [V(Y1) + V(Y2)]}{V} = 79.9846\,\%$$

A explicação da variável balança comercial foi explicada em quase 80% em apenas dois componentes principais, reduzindo-se a dimensionalidade do problema de três para duas variáveis escritas em combinação linear.

Na Tabela 20 mostra-se o percentual de explicação da variância total pelos dois componentes encontrados para a variável balança comercial e o percentual de variância acumulada.

TABELA 20 – Percentual da variância explicada por cada componente e o percentual acumulado.

Número de componentes	Percentual da variância	Percentual acumulada da variância
C1	43.9180	43.9180
C2	36.0669	79.9849
C3	20.0151	100.0000

Para se determinar quais as variáveis contribuem mais para cada componente principal, podemos analisar a relação $\left|\alpha_{ji}\ [V(Yj)]^{1/2}\right|$, com i = j = 1, ..., 3, fornecendo assim o valor da correlação absoluta ou utilizando-se a matriz de correlação, fornecendo direto a contribuição de cada variável para a formação da nova variável.

$$[V(Y1)]^{1/2} = (1.3175)^{1/2} = 1.1478$$
$$[V(Y2)]^{1/2} = (1.0820)^{1/2} = 1.0402$$
$$[V(Y3)]^{1/2} = (0.6004)^{1/2} = 0.7749$$

Para determinarmos a correlação absoluta entre os autovetores e os componentes principais devemos fazer:

$$X1 \rightarrow |0.6811\ x\ 1.1478| = 0.7818$$
$$X2 \rightarrow |0.1054\ x\ 1.0402| = 0.1096$$
$$X3 \rightarrow |0.7245\ x\ 0.7749| = 0.5614$$

A variável que maior contribuição ofereceu para a primeira combinação linear foi a variável X1, seguida das variáveis X3 e X2, então podemos dizer que Y1 possui características na sua grande maioria da variável SALDO DA BALANÇA COMERCIAL.

O componente C1, que mais agregou explicação ao desconhecimento inicial a cerca da variabilidade dos dados, apresentou correlação mais alta com importação de combustíveis e lubrificantes.

Para a segunda componente principal temos:

$$X1 \rightarrow |-0.3726\ x\ 1.1478| = 0.4277$$
$$X2 \rightarrow |0.9018\ x\ 1.0402| = 0.9381$$
$$X3 \rightarrow |0.2191\ x\ 0.7749| = 0.1698$$

A variável que maior contribuição ofereceu para a segunda combinação linear foi a variável X2, seguida das variáveis X1 e X3, então pode-se dizer que Y2 possui características na sua grande maioria da variável IMPORTAÇÃO DE COMBUSTÍVEIS E LUBRIFICANTES.

5.8 – AJUSTAMENTO E PREVISÃO

Pode-se observar através das Figuras 03 a 12 (Anexo 1) que todas as séries apresentam heterocedasticidade na variância. Também pode-se observar alguns valores aberrantes. A maioria das variáveis foram transformadas, a fim de garantir a homocedasticidade da variância.

Mostra-se nas Fifuras 13 a 17 (anexo 2) a previsão para a variável referência e da variável de maior representatividade, juntamente com o valor observado da variável de maior representatividade, no período compreendido de janeiro a dezembro de 1992.

Os critérios de ajustamento utilizados foram os critérios AIC, BIC, coeficiente de explicação ajustado e variância residual e para a previsão o Erro Médio Percentual de Previsão (EMPP).

5.8.1 – AJUSTAMENTO

i) Política Monetária

i.1) Meios de Pagamento M1

O modelo com intervenção encontrado foi um AR(1), sendo descrito da seguinte maneira:

$$Z_t = \underset{(9.07)}{31.562} + \underset{(2.50)}{0.28563}\ Z_{t-1} + a_t + \underset{(12.10)}{168.83}\ X_{63} +$$

$$+ \underset{(6.14)}{62.405}\ X_{62},\ldots + \underset{(6.47)}{88.562}\ X_{16} + \underset{(5.57)}{30.282}\ X_{12},\ldots +$$

$$+ \underset{(2.53)}{11.196}\ X_{34}$$

As estatísticas de ajuste do modelo são:

$R^2 = 77.890\%$ AIC = 0.53700 BIC = 0.55740

Observando-se as intervenções feitas, pode-se verificar que as estimativas dos coeficientes das variáveis de intervenção apresentam os sinais esperados.

A intervenção ocorrida no instante 62 apresenta um aumento para o instante 63, onde este aumento apresenta um crescimento de mais de 100% do mês de fevereiro de 1990 para o mês de março deste mesmo ano, logo após há uma redução na oferta de moeda em poder do publico, estas intervenções são reflexos do Plano Collor I. Havendo um enxugamento de papel-moeda circulante. Após março de 1990 esta intervenção apresenta-se com características sazonais.

O crescimento ocorrido em dezembro de 1985, no instante 12 é caracterizado pelo efeito sazonal, pois a partir deste período este efeito repete-se, sendo este crescimento da circulação de moeda é devido ao mês de dezembro, onde há uma maior circulação de moeda devido as festas de fim de ano.

O aumento caracterizado pelo modelo de intervenção em abril de 1986, no instante 16 é devido ao Plano Cruzado que impôs congelamento de preços em março de 1986, fazendo com que houvesse maior circulação de papel-moeda, a duração deste plano foi até dezembro de 1986.

O acréscimo do volume de M1 no mês de dezembro de 1987, no instante 36, é caracterizado pelo efeito sazonal, visto que este aumento se repete em todos os anos subseqüentes do período analisado, este aumento de circulação de moeda é devido as festas de final de ano.

i.2) Primeira Componente Principal (Variável de referência)

O modelo com intervenção encontrado foi um AR(1), descrito da seguinte maneira:

$$Z_t = \underset{(11.62)}{28.978} + \underset{(5.32)}{0.52301}\ Z_{t-1} + a_t + \underset{(10.69)}{102.18}\ X_{63} +$$

$$+ \underset{(6.19)}{27.876}\ X_{36},\ldots - \underset{(-3.59)}{14.946}\ X_{13},\ldots + \underset{(5.08)}{35.627}\ X_{62},\ldots +$$

$$+ \underset{(3.72)}{33.767}\ X_{16}$$

As estatísticas de ajuste do modelo são:

$R^2 = 78.034\%$ AIC = 0.47345 BIC = 0.49385

Como pode-se verificar as intervenções ocorridas na variável referência apresentam os sinais esperados dos coeficientes das variáveis de intervenção e também apresentam as intervenções no mesmo período e do mesmo tipo das que ocorreram na variável de maior representatividade (M1), o que já se esperava pois a variável M1 além de ser a que mais contribuiu para a formação da variável referência, é a que compõe os agregados M2, M3 e M4.

ii) Política de Preços

ii.1) IGP-DI (transformação raiz quadrada)

O modelo encontrado com intervenção foi um AR(2), descrito da seguinte maneira:

$$Z_t = \underset{(11.36)}{4.0806} + \underset{(13.30)}{1.3399}\ Z_{t-1} - \underset{(-5.05)}{0.30748}\ Z_{t-2} + a_t +$$

$$+ \underset{(10.84)}{4.0598}\ X_{63} + \underset{(4.33)}{1.3621}\ X_{49} - \underset{(-4.94)}{1.5615}\ X_{15},\ldots +$$

$$+ \underset{(5.58)}{1.9578}\ X_{16}, \ldots + \underset{(4.11)}{1.2802}\ X_{61}, \ldots$$

A estatística de ajuste do modelo é:

$R^2 = 90.90\%$

Observa-se que as estimativas dos coeficientes das variáveis de intervenção, apresentam os sinais esperados. Estas intervenções levam a concluir que a intervenção ocorrida em março de 1986, no instante 15, apresentam um decrescimento no índice geral de preços, onde no mesmo período foi implantado o Plano Cruzado, como uma tentativa de estabilização da economia brasileira. No mês de janeiro de 1989, no instante 49, há um aumento do índice, sendo no período seguinte reduzido de 36.56% para 11.8%, esta redução, deve-se a implantação do Plano Cruzado II. No inicio do ano de 1990, há um crescimento expressivo no índice geral de preços, este crescimento sinaliza um crescimento da inflação, mostrado este crescimento através das intervenções ocorridas em janeiro, fevereiro e março de 1990, nos instantes 61, 62 e 63, sendo que o pico da inflação ocorreu em março de 1990, período este que mais um plano de estabilização foi implantado, o Plano Collor I, cujo efeito foi refletido na redução da taxa de inflação, pois o IGP foi reduzido de 81.31% em março de 1990 para 11.33% em abril de 1990.

ii.2) Primeira Componente Principal (Variável de referência, transformação logarítmica)

O modelo com intervenção encontrado foi um AR(1), descrito da seguinte maneira:

$$\underset{\quad}{Z_t} = \underset{(7.30)}{2.2511} + \underset{(13.42)}{0.84033}\ Z_{t-1} + a_t - \underset{(-4.26)}{1.2573}\ X_{16} +$$

$$- \underset{(-3.92)}{1.1623}\ X_{64} + \underset{(2.89)}{0.96529}\ X_{52}$$

A estatística de ajuste do modelo é:

$R^2 = 76.735\%$

As estimativas dos coeficientes das variáveis de intervenção, apresentam os sinais esperados. A intervenção ocorrida em abril de 1986, no instante 16, apresenta uma redução, sendo esta devida a reflexos do Plano Cruzado, implantado no período anterior. A intervenção de abril de 1989, no instante 52, mostra um aumento da variável referência, sinalizando assim que o Plano

Verão de fevereiro de 1989 já não atinge os seus objetivos que são de baixar a taxa inflacionária do país. Em abril de 1990, a intervenção ocorrida no instante 64, vimos a implantação de mais um plano econômico, o Plano Brasil Novo, implantado em março de 1990, refletindo neste instante uma redução nos índices de preços, assim como o da inflação.

iii) Atividade Econômica

iii.1) Produção Industrial (transformação raiz quadrada)

O modelo com intervenção encontrado foi um AR(1), descrito da seguinte maneira:

$$Z_t = \underset{(25.76)}{5.2302} + \underset{(8.11)}{0.71048}\ Z_{t-12} + a_t - \underset{(8.58)}{4.1426}\ X_{64} +$$

$$+ \underset{(5.29)}{2.0682}\ X_{65} + \underset{(2.08)}{1.23925}\ X_{76} - \underset{(-3.13)}{1.4377}\ X_{60},\ldots$$

As estatísticas de ajuste do modelo são:

$R^2 = 78.814\%$ AIC = 3.1739 BIC = 3.2719

Observa-se que as estimativas dos coeficientes das variáveis de intervenção tem os sinais esperados. Estas intervenções levam a concluir que a intervenção ocorrida em abril de 1990, no instante 64, mostra uma redução, significando os reflexos do plano de estabilização Brasil Novo, seguida de intervenção ocorrida em maio de 1990, no instante 65, onde já ocorre um aumento da produção industrial, devido ao efeito do plano dos empresários foram forçados a reestudar os seus investimentos e as modificações que deveriam ser feitas, motivo pelo qual o crescimento da produção ocorreu um mês depois. Anterior a abril de 1990 havia uma instabilidade na economia, não facilitando o aumento da produção, como pode-se verificar pela intervenção ocorrida em dezembro de 1989, no instante 60, tendo esta um efeito sazonal, mostrando que dezembro é um mês de baixa para o setor produtivo. Em abril de 1991, no instante 76, foi possível verificar também um aumento de produtividade no setor.

iii.2) Primeira Componente Principal (variável de referência, transformação raiz quadrada)

O modelo com intervenção encontrado foi um AR(2), descrito da seguinte maneira:

$$Z_t = \underset{(23.81)}{36.484} - \underset{(-3.68)}{0.39840}\ Z_{t-4} + \underset{(6.98)}{0.67604}\ Z_{t-12} + a_t +$$

$$\underset{(-9.39)}{-36.885}\ X_{64} + \underset{(7.11)}{27.94}\ X_{65} \underset{(-2.67)}{-11.611}\ X_{36}, ... +$$

$$\underset{(-4.43)}{-17.405}\ X_{63} + \underset{(3.14)}{13.532}\ X_{39}, ...$$

As estatísticas de ajuste do modelo são:

$R^2 = 81.906\%$ AIC = 3.3651 BIC = 3.6262

Observa-se que as estimativas dos coeficientes das variáveis de intervenção apresentam os sinais esperados. Houve duas intervenções sazonais em dezembro de 1987 e março de 1988. A partir de dezembro de 1987, no instante 36, houve sempre uma redução no incremento percentual da atividade econômica, pois dezembro é um mês caracterizado pela venda de produtos acabados e de serviços acabados do que pela produção. A intervenção de março de 1988, no instante 39, caracteriza um aumento de produção nesses meses, pois há uma retomada total a produção. As intervenções ocorridas em março, abril e maio de 1990, são reflexos do plano de estabilização Brasil Novo.

iv) Agregados de Crédito

iv.1) Crédito Total (transformação raiz quadrada)

O modelo com intervenção encontrado foi um AR(1), descrito da seguinte maneira:

$$Z_t = \underset{(25.0)}{2.8379} + \underset{(7.76)}{0.67219}\ Z_{t-1} + a_t - \underset{(-15.16)}{3.0485}\ X_{65} +$$

$$+ \underset{(4.35)}{1.0085}\ X_{63} + \underset{(3.73)}{0.53292}\ X_{41} + \underset{(2.68)}{0.62426}\ X_{62}, \ldots +$$

$$+ \underset{(2.83)}{0.57068}\ X_{31}$$

As estatísticas de ajuste do modelo são:

$R^2 = 82.841\%$ AIC = 4.5178 BIC = 4.5761

Observa-se que as estimativas dos coeficientes das variáveis de intervenção tem os sinais esperados. A intervenção ocorrida no instante 62, em fevereiro de 1990, mostra um aumento de créditos concedidos, aumento este que continua até março de 1990, no instante 63, onde ocorre o Plano Collor I, sendo este refletido na intervenção ocorrida em maio de 1990, no instante 65, que apresenta uma redução de créditos concedidos ao setor privado, efeitos do plano de estabilização.

iv.2) Primeira Componente Principal (variável de referência, transformação raiz quadrada)

O modelo com intervenção encontrado foi um AR(1), descrito da seguinte maneira:

$$Z_t = \underset{(5.49)}{0.35325} + \underset{(8.78)}{0.74774}\ Z_{t-1} + a_t - \underset{(-4.37)}{0.27709}\ X_{65} +$$

$$+ \underset{(4.56)}{0.29448}\ X_{55} - \underset{(-4.41)}{0.2997}\ X_{64} - \underset{(-3.41)}{0.21619}\ X_{51} +$$

$$+ \underset{(2.66)}{0.18359}\ X_{21}$$

As estatísticas de ajuste do modelo são:

$R^2 = 78.278\%$ AIC = 5.9337 BIC = 6.0211

As estimativas dos coeficientes das variáveis de intervenção apresentam os sinais esperados. A intervenção ocorrida em setembro de 1986, no instante 21, apresenta um aumento nos empréstimos concedidos ao setor privado. Em março de 1989, no instante 51 ocorre uma redução de empréstimos, sendo que em julho de 1989, no instante 55 há um aumento de 28.88 pontos percentuais do instante anterior. Em abril de 1990 no instante 64 e em maio de 1990, no instante 65 há uma redução de empréstimos, este enxugamento da concessão de empréstimos, deve-se ao efeito de Plano Brasil Novo, ocorrido em março de 1990.

v) Balança Comercial

v.1) Saldo da Balança Comercial

O modelo com intervenção encontrado foi um modelo constante descrito da seguinte maneira:

$$Z_t = \underset{(541.11)}{25.184} + a_t + \underset{(358.0)}{149.03}\ X_6 + \underset{(114.51)}{47.666}\ X_{31} +$$

$$\underset{(-58.89)}{-24.184}\ X_{70} + \underset{(12.28)}{5.1133}\ X_{25} + \underset{(3.30)}{1.3753}\ X_{73}$$

As estatísticas de ajuste do modelo são:

$R^2 = 99.946\%$ AIC = -1.6966 BIC = -1.5230

As intervenções ocorridas em junho de 1985, no instante 6; janeiro de 1987, no instante 25 e julho de 1987, no instante 31 apresentam um crescimento no saldo da balança comercial, este saldo positivo representa um superavit nas exportações brasileiras, nos períodos apresentados. Em outubro de 1990, no instante 70, há um decrescimento do saldo da balança comercial, mas em janeiro de 1991, no instante 73, este quadro é revertido, mostrando um aumento no saldo, onde as exportações superaram as exportações.

v.2) Primeira Componente Principal (variável de referência)

O modelo com intervenção encontrado foi um modelo constante descrito da seguinte maneira:

$$Z_t = \underset{(465.5)}{17.587} + a_t + \underset{(96.76)}{32.694}\ X_{30} - \underset{(-49.1)}{16.587}\ X_{69} + \underset{(10.16)}{3.4336}\ X_{24}$$

$$+ \underset{(3.13)}{1.0576}\ X_{67} + \underset{(2.97)}{1.0019}\ X_{72}$$

As estatísticas de ajuste do modelo são:

R^2 = 99.354% AIC = -2.1139 BIC = -1.9402

Verifica-se que todas as estimativas apresentam sinais esperados das variáveis de intervenção, sendo que a intervenção ocorrida em dezembro de 1986, no instante 24, apresenta um crescimento da variável referencia, a intervenção ocorrida em junho de 1987, no instante 30, apresenta também um aumento. A intervenção ocorrida em julho de 1990, no instante 67, apresenta um aumento e logo após na intervenção ocorrida em setembro de 1990, no instante 69, há um decrescimento de -0.66% para -16.609% no mês de outubro do mesmo ano, sendo que a intervenção ocorrida em dezembro de 1990, no instante 72, apresenta um crescimento de 0.29% do mês anterior para 0.98%.

5.8.2. PREVISÃO PARA O PERÍODO DE JANEIRO A DEZEMBRO DE 1992

i) Política Monetária

Na Tabela 21 apresenta-se os valores observados e previstos, da variável de maior representatividade, bem como os seus valores previstos da variável referência.

TABELA 21 – Valores previstos e observados para a variável M1 e os valores previstos para a Variável Referência em 1992.

Meses	M1	V. Referência	V. Observado
Janeiro	-4.1	-4.8	-5.1
Fevereiro	32.6	31.8	34.7
Março	2.6	2.9	1.3
Abril	24.8	25.5	26.6
Maio	16.4	16.1	15.3
Junho	27.8	28.4	30.4
Julho	13.5	18.0	12.2
Agosto	26.9	27.9	29.4
Setembro	21.0	22.0	20.5
Outubro	23.5	22.8	22.5
Novembro	42.8	40.9	41.1
Dezembro	37.4	36.8	36.0

EMPP (M1) = -15.83%

EMPP (Variável Referência) = 28.33%

ii) Política de preços

Na Tabela 22 apresenta-se os valores observados e previstos, da variável de maior representatividade, bem como os seus valores previstos da variável referência.

TABELA 22 – Valores previstos e observados para a variável IGP-DI e os valores previstos para a Variável Referência em 1992.

Meses	IGP-DI	V. Referência	V. Observado
Janeiro	30.75	17.85	26.84
Fevereiro	42.36	18.59	24.79
Março	20.42	19.24	20.70
Abril	18.94	19.80	18.54
Maio	17.82	19.80	22.44
Junho	17.07	20.29	21.41
Julho	18.63	20.70	21.69
Agosto	18.41	21.06	25.54
Setembro	18.40	21.36	27.37
Outubro	18.35	21.62	24.93
Novembro	18.41	21.84	24.22
Dezembro	18.48	22.02	23.69

EMPP (IGP-DI) = -159.75%

EMPP (Variável Referência) = 308.25%

iii) Atividade Econômica

Na Tabela 23 apresenta-se os valores observados e previstos, da variável de maior representatividade, bem como os seus valores previstos da variável referência.

TABELA 23 – Valores previstos e observados para a variável PRODUÇÃO INDUSTRIAL (PI) e os valores previstos para a Variável Referência em 1992.

Meses	**PI**	**V. Referência**	**V. Observado**
Janeiro	1.27	1.42	1.61
Fevereiro	3.37	3.34	3.35
Março	3.61	3.59	3.58
Abril	-1.72	-1.82	-1.97
Maio	4.93	4.84	4.92
Junho	7.78	7.82	7.93
Julho	5.89	5.78	6.06
Agosto	-1.32	-1.30	-1.25
Setembro	-0.42	-0.38	-0.34
Outubro	2.52	2.58	2.56
Novembro	3.70	3.72	3.75
Dezembro	-12.06	-13.10	-13.00

EMPP (PI) = -2.91%

EMPP (Variável Referência) = 5.16%

iv) Agregados de Crédito

Na Tabela 24 apresenta-se os valores observados e previstos, da variável de maior representatividade, bem como os seus valores previstos da variável referência.

TABELA 24 – Valores previstos e observados para a variável CRÉDITO e os valores previstos para a Variável Referência em 1992.

Meses	**CRÉDITO**	**V. Referência**	**V. Observado**
Janeiro	27.84	27.48	29.84
Fevereiro	25.50	25.98	26.30
Março	22.65	22.06	21.59

EMPP (CRÉDITO) = 58.00%

EMPP (Variável Referência) = 61.66%

v) Balança Comercial

Na Tabela 25 apresenta-se os valores observados e previstos, da variável de maior representatividade, bem como os seus valores previstos da variável referência.

TABELA 25 – Valores previstos e observados para a variável SALDO DA BALANÇA COMERCIAL (SBC) e os valores previstos para a Variável Referência em 1992.

Meses	**SBC**	**V. Referência**	**V. Observado**
Janeiro	-0.028	-0.0158	0.288
Fevereiro	-0.028	-0.0158	-0.052
Março	-0.028	-0.0158	0.668
Abril	-0.028	-0.0158	-0.137
Maio	-0.028	-0.0158	0.117
Junho	-0.028	-0.0158	-0.055
Julho	-0.028	-0.0158	0.171
Agosto	-0.028	-0.0158	-0.106
Setembro	-0.028	-0.0158	0.020

EMPP (SBC) = 13.17%

EMPP (Variável Referência) = 11.70%

Observando-se o Erro Médio Percentual de Previsão, pode-se afirmar que a Metodologia de Componentes Principais apresentou um excelente desempenho na redução do número de variáveis, mantendo o mesmo grau de informação do conjunto de variáveis originais, bem como proporcionou que se identificasse qual a variável mais representativa em cada conjunto de variáveis econômicas, tais como: a Política Monetária (M1), a Política de Preços (IGP-DI), Atividade Econômica (Produção Industrial), Agregados de Crédito (Total de créditos) Balança Comercial (Saldo da Balança Comercial).

6 – CONCLUSÃO

Em geral, os modelos estatísticos não competem entre si, mas sim possuem características diferentes, sendo que cada situação pode se beneficiar com o uso de um modelo e não de outro.

Neste trabalho foram testados dois modelos, o modelo obtido por COMPONENTES PRINCIPAIS e o modelo ARIMA de Box & Jenkins, existindo vantagens e desvantagens na utilização de cada um deles.

Assim, neste trabalho encontramos através da metodologia de componentes principais a variável referência, para o conjunto de variáveis componentes da Política Monetária, Política de Preços, Atividade Econômica, Agregados de Crédito e Balança Comercial, bem como a variável que apresenta maior contribuição na formação da variável referência.

Tanto a variável referência, como a variável de maior representatividade apresentaram as mesma estrutura de modelagem.

Os resultados obtidos levam na direção de mais estudos e pesquisas da utilização da metodologia de componentes principais aplicado ao estudo de séries temporais.

Como continuidade deste trabalho podemos citar:

- determinação de índices microeconômicos para empresas da região;

- utilização de componentes principais na determinação de indicadores de qualidade;

- determinação de um índice macroeconômico, que possa representar os ciclos econômicos brasileiros.

REFERÊNCIAS BIBLIOGRÁFICAS

AFIFI, A. A. & AZEN, S. P. Statistical Analysis. A computer oriented approach. Los Angeles, Califórnia. 2ª ed. 1971.

ANDERSON, T.W. An introduction to multivariate analysis, Wiley, New York, N. Y, 2ª ed. 1984.

ARNOLS, S. F. The teory of linear models and multivariate analysis, Wiley, New York, N. Y. 1981.

BOX, G. E. & JENKINS, G. M. Time series analysis, forecasting and control, San Francisco. Holden Day.

CAMARGO, M. E. Modelagem Clássica e Bayesiana: uma evidência empírica do processo inflacionário brasileiro. Tese de Doutorado. Programa de Pós-Graduação. UFSC. 1992.

CHATTERJEE, S. Y. & PRICE, B. Regression analysis by examples, Wiley, New York, N. Y. 228 págs. 1977.

CROOKES, J. G., CROSTON, K. & SYPSAS, P. Process Components for Multivariate Time Series Analysis. University of Lancaster. Journal of the Operational Research Society, 31(4): 325-330

FLURY, B. & RIEDWYL, H. Multivariate statistics a pratical approach. Ed. Chapman and Hall, New York. 1988.

GIRI, N. C. Multivariate Statistical Inference. Department of Mathmatics University of Montreal, Montreal, Quebec, Canadá. 1977.

GNANADESIKAN, R. & KETTERING, J. R. Robust estimates, residuals and outlier detection with multiresponse data, Biometrics, 28(1): 81 – 124. 1972.

GREENBERG, E. Minimum variance properties of principal component regression, JASA, 70(349): 194 – 197. 1975.

GUNST, R. F., WEBSTER, J. T. & MASON, R. L. A comparison of least squares and latent root regression estimator, Technometrics, 18(1): 75 – 83. 1976.

HARVEY, A. C. Forecasting, Strutural Time Series Models and the Kalman Filter. Cambridge University Press. 1990.

HAWKINS, D. M. The detection of errors in multivariate data using principal components, JASA, 69(346): 340 344. 1974.

HENRIQUEZ, W. Análisis de calificaciones por componentes principales. Tesis de Maestria, Posgrado em Estadística, Universdidad Central de Venezuela, Maracay, Venezuela. 1985.

HOCKING, R. R. The analysis and selection of variables in linear regression. Biometrics, 32(1): 1-49. 1976.

HOTELLING, H. Analysis of a complex of statistical variables into principal components, J. Educ. Psychol., 24: 417 – 441 e 498 – 520. 1933.

JACKSON, E. J. & MUDHOLKAR, G. S. Control Procedures for Residuals Associated With Principal Component Analysis. Technometrics, 21(3) 341 – 349. 1979.

JOHNSON, R. A. & WICHEM, D. W. Applied Multivariate Statistical Analysis: Prentice Hall, Inc., 1982.

KENDALL, M. G. A course in multivariate analysis, Griffin, Londres, 152 págs. 1957.

KENDALL, M. G. A course in multivariate analysis, 2ª ed. New York. 1980.

KLOEK, T. & MENNES, L. B. M. Simulateneous Equations Estimations base don Principal Components of Predetermined Variables. Econometrica, 28 (1) 45 – 61. 1960.

KUBRUSLY, L. S. Uma estratégia de análise multivariada. Laboratório nacional de Computação Científica/CNPq, Série de C&T, Vol. 1, nº 4(1982). Rio de Janeiro. 1982.

KUBRUSLY, L. S. & GOUVÊA, V. H. C. Análise de estabilidade no tempo das matrizes do Balanço Energético Nacional (1976-1980). Revista de Econometria. 8(2): 93 – 108, novembro. 1988.

MARDIA, K. V., KENT, J. T. & BIBBY, J. M. Multivariate analysis, Academic, Londres, 521 págs. 1979.

MARQUARDAT, D. W. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation, Technometrics, 12(3): 591 – 612. 1970.

MATTEUCCI, S., COLMA, A. & PLA, L. Análisis regional de la vegetación y el ambiente del Estado del Falcón: La vegetación del Departamento de Investigación, IUTC, Coro, Venezuela, 292 págs. 1982.

MATTEUCCI, S. & COLMA, A. Metodologia para el estúdio de la vegetación, monografia científica nº 22, série de biologia, Secretaría General de la Organizacion de los Estados Americanos, Washington, D. C., 168 págs. 1982.

MORRISON, D. F. Multivariate statistical methods, McGraw, New York, 2ª. Ed. 515 págs. 1976.

PINTO, L. A. M. S. Componentes Principais: Um teste para a eficiência de Previsão a Curto Prazo. Dissertação apresentada ao Departamento de Engenharia Industrial da PUC/RJ. Rio de Janeiro. 1981.

REVISTA CONJUNTURA ECONÔMICA, Instituto Brasileiro de Economia Fundação Getúlio Vargas, Rio de Janeiro. 1980 a 1993.

SOUZA, A. M. Aplicação e performance da análise de intervenção em séries macroeconômicas brasileiras, Monografia de Especialização apresentada ao CPGEMQ – UFSM, Santa Maria. 1991.

SOUZA, R. C. & CAMARGO, M. E. Trabalho sobre a análise de séries temporais. Rio de Janeiro. PUC. 1990.

SOUZA, J. Análise em Componentes Principais e suas Aplicações. Métodos Estatísticos nas Ciências Psicossociais, Vol II, Ed. Thesaurus. 1988.

VELAZQUEZ, G. & PLA, L. Diagnóstico lechero de los distritos Mauroa, Federación Y Zamora del Estado Falcón, Informe final, 3 vols, 234 págs. 1984.

WEISBERG, S. Applied linear regression, Wiley, New York, N. Y. 283 págs. 1980.

ANEXO 1 – GRÁFICOS

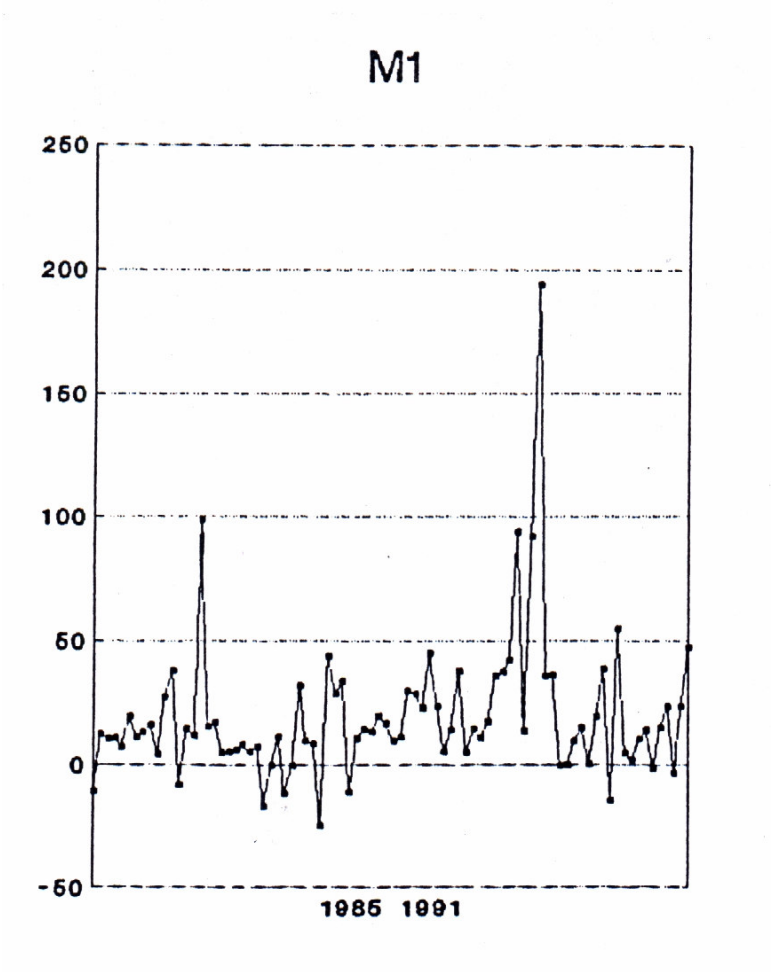

GRÁFICO 03 – Gráfico representativo da série M1

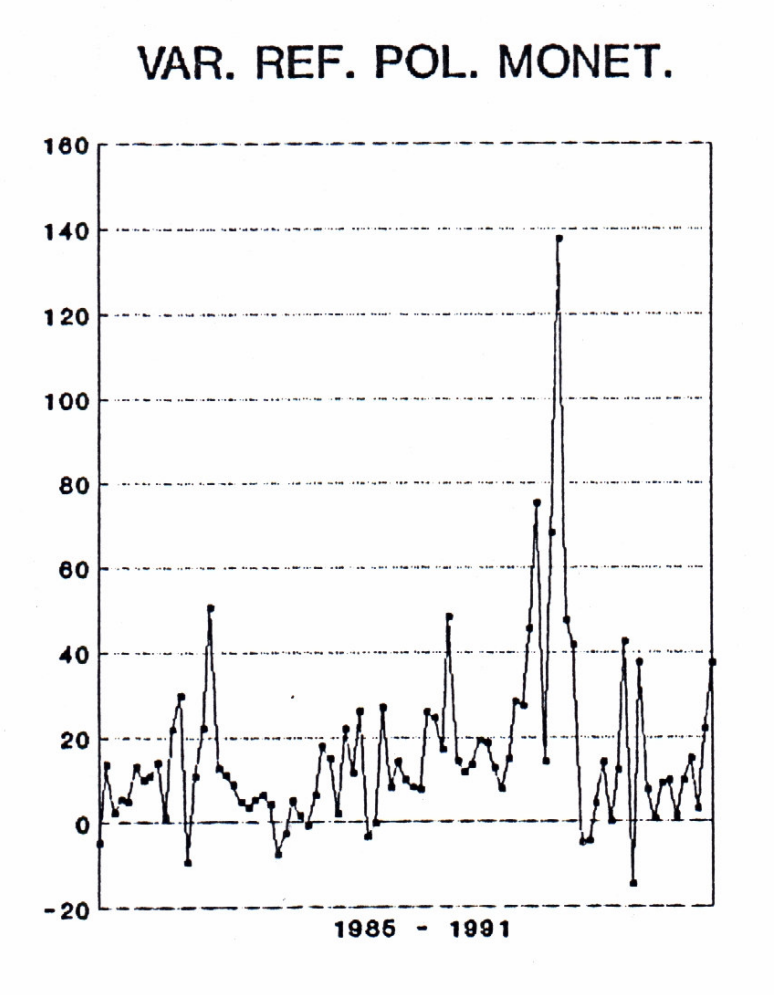

GRÁFICO 04 – Gráfico representativo da variável referência – Política Monetária

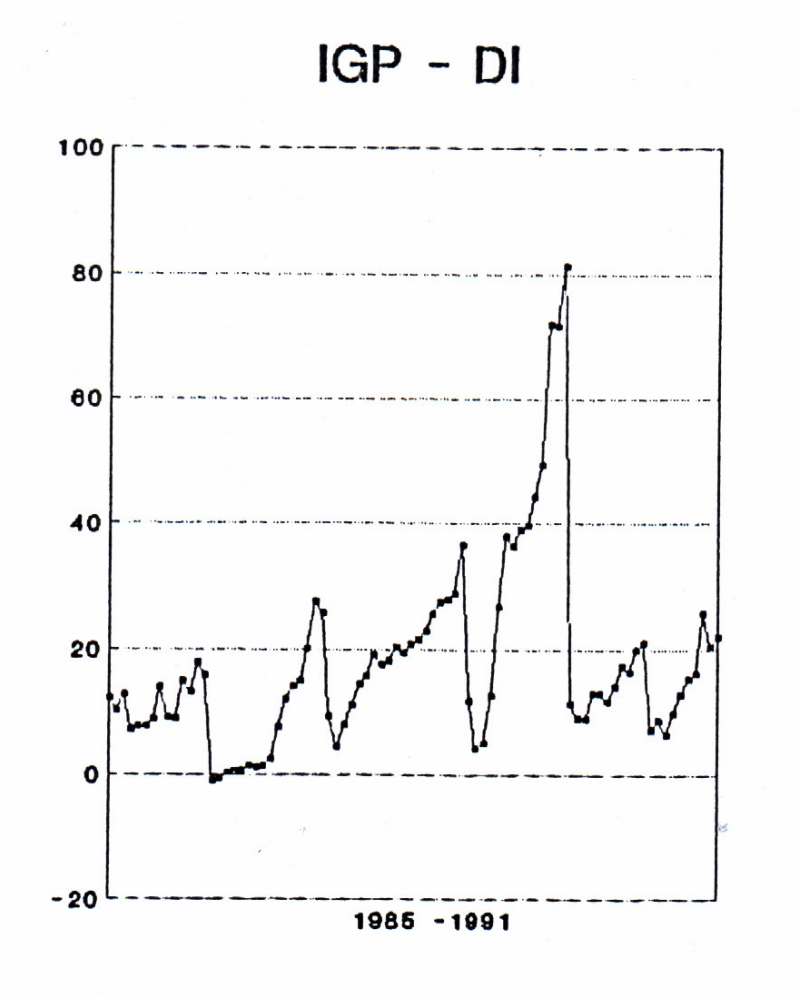

GRÁFICO 05 – Gráfico representativo da série IGP-DI

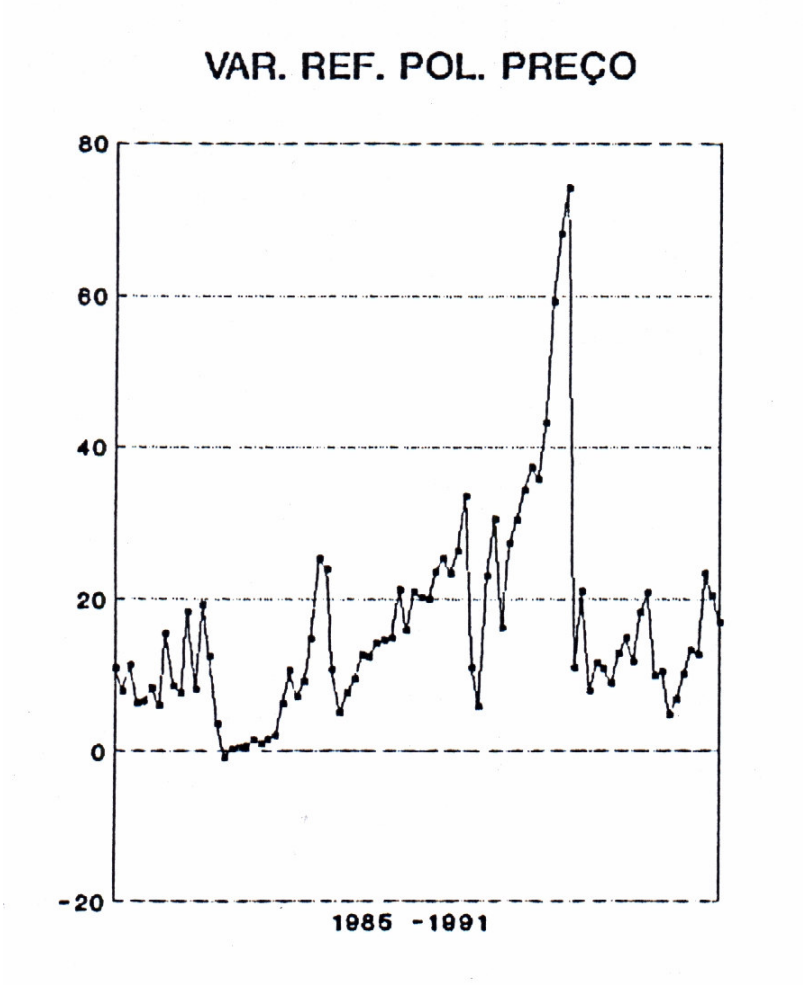

GRÁFICO 06 – Gráfico representativo da variável referência – Política de Preços

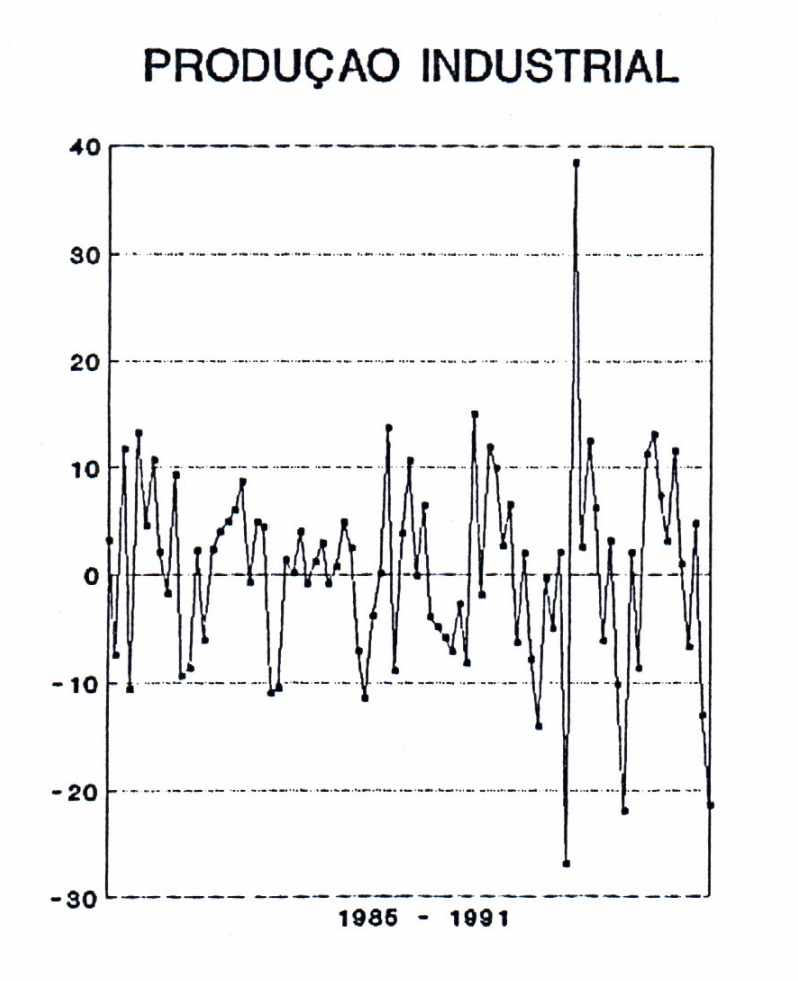

GRÁFICO 07 – Gráfico representativo da série Produção Industrial

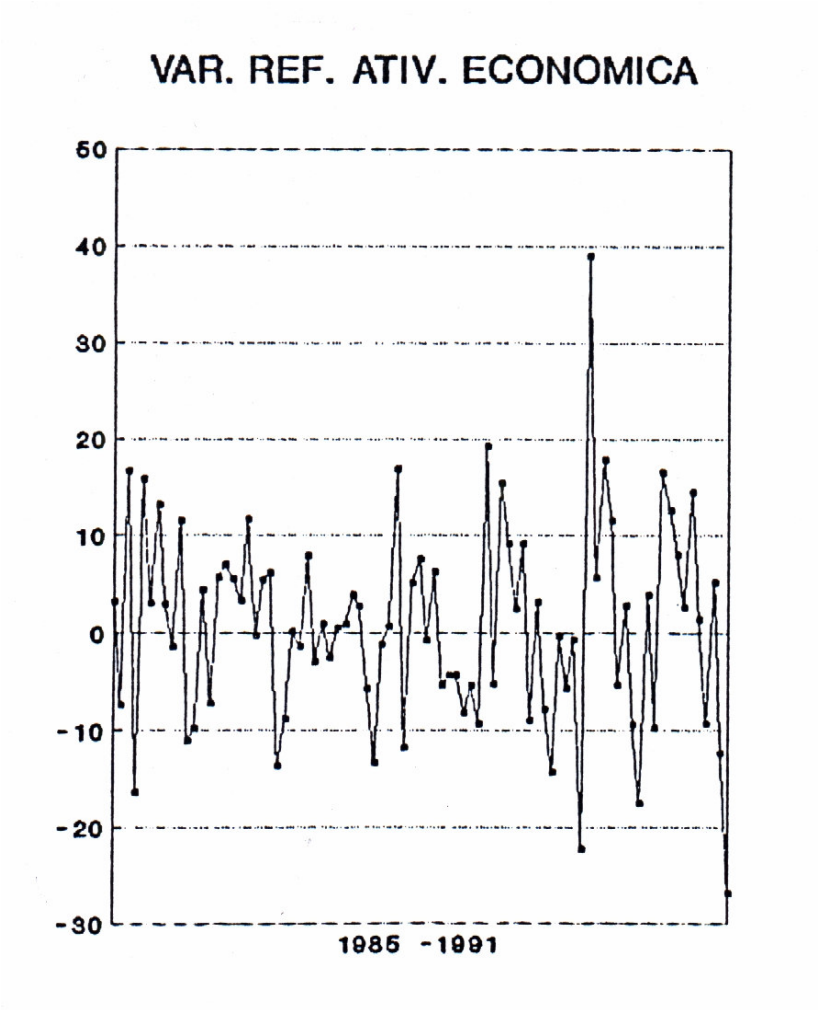

GRÁFICO 08 – Gráfico representativo da variável referência – Atividade Econômica

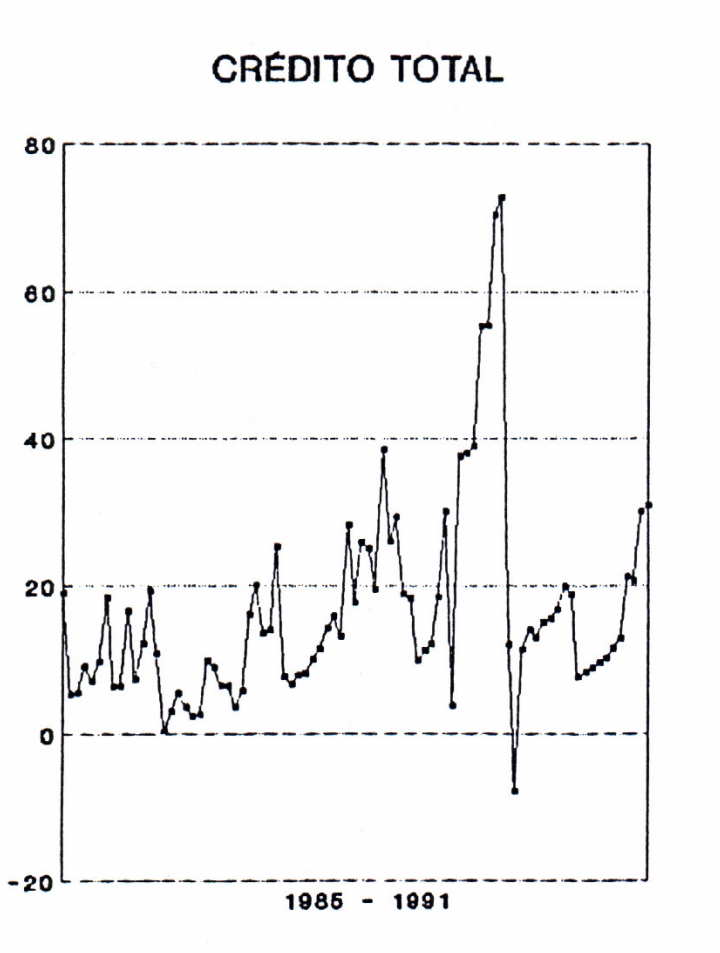

GRÁFICO 09 – Gráfico representativo da série Crédito Total

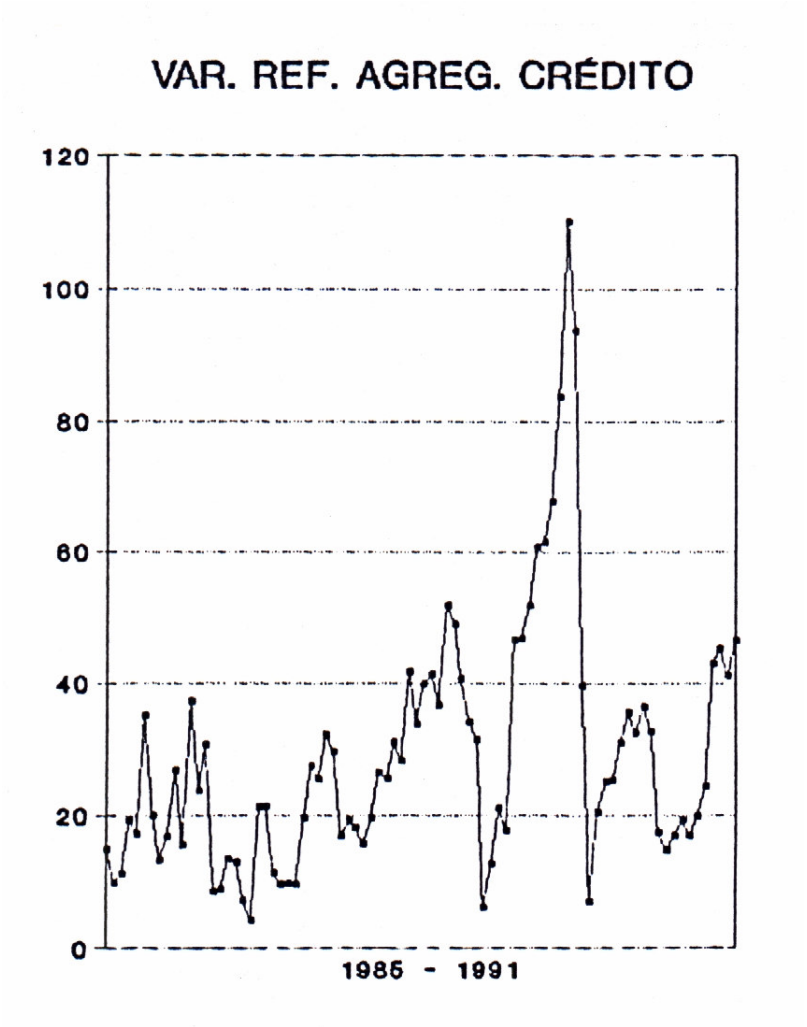

GRÁFICO 10 – Gráfico representativo da variável referência – Agregados de Crédito

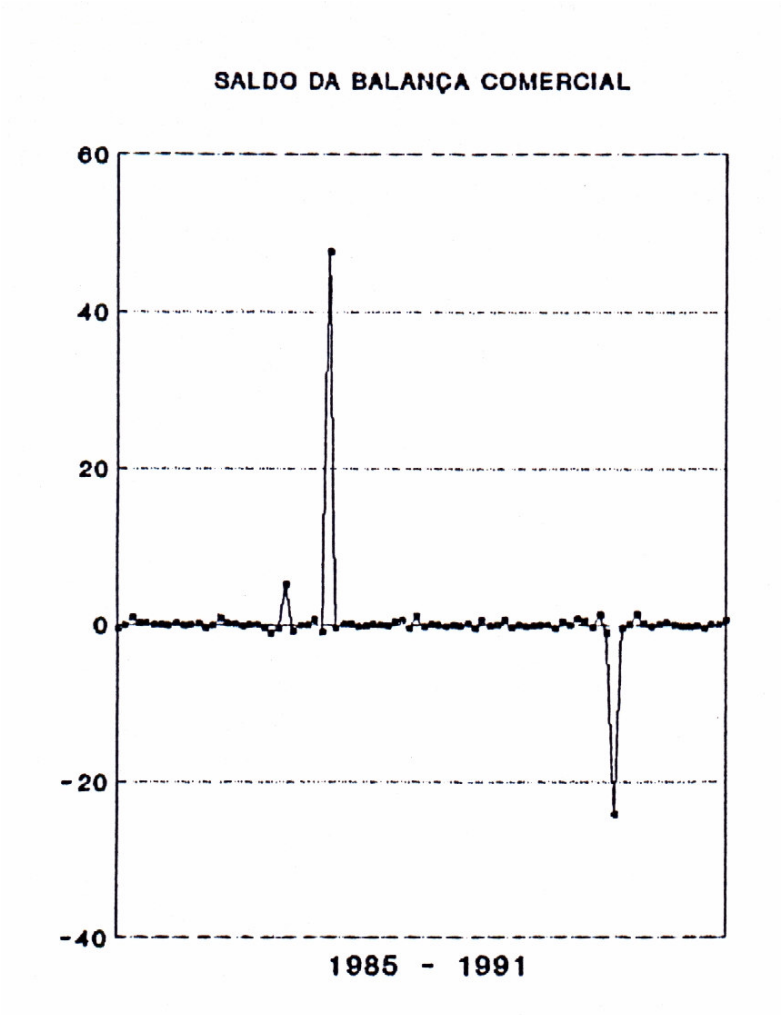

GRÁFICO 11 – Gráfico representativo da série Saldo da Balança Comercial

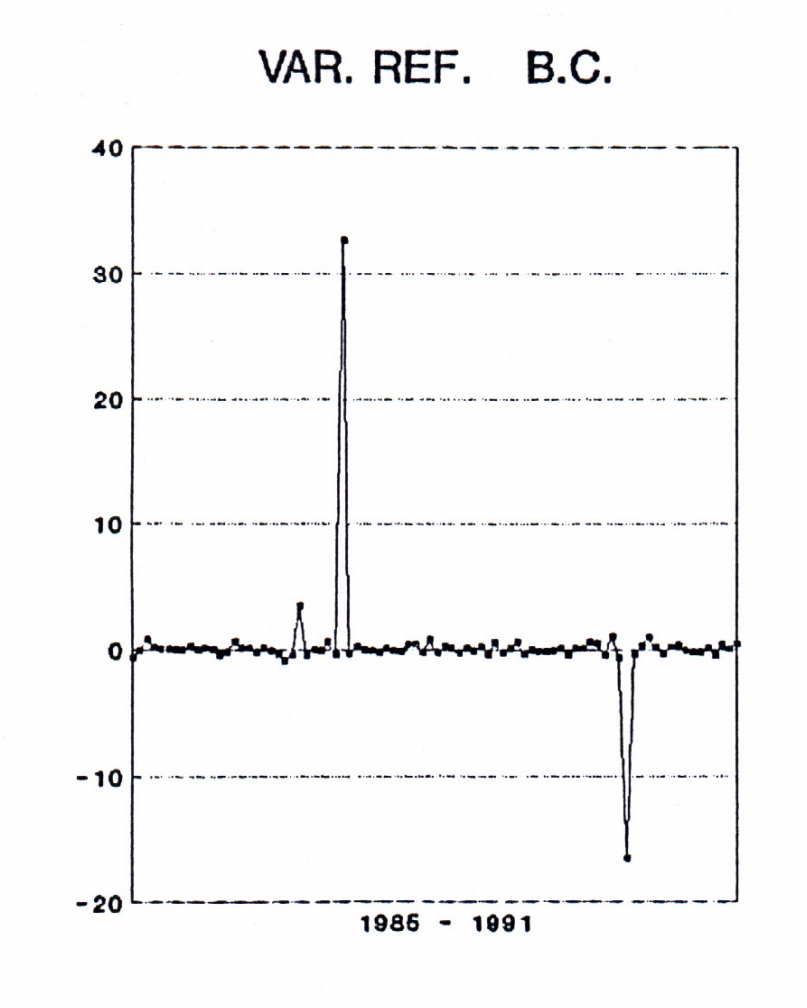

GRÁFICO 12 – Gráfico representativo da variável referência – Balança Comercial

ANEXO 2 – GRÁFICOS DAS PREVISÕES

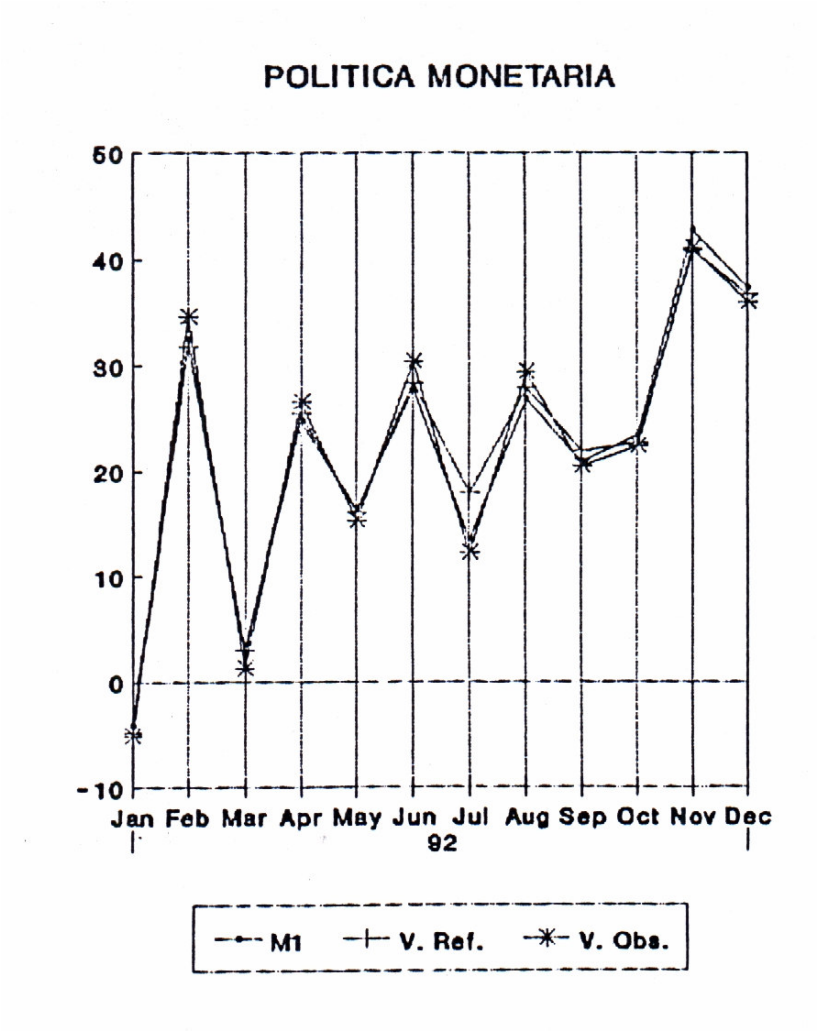

FIGURA 13 – Gráfico representativo dos valores previstos e observados para a variável M1 e os valores previstos para a Variável Referência para o ano de 1992.

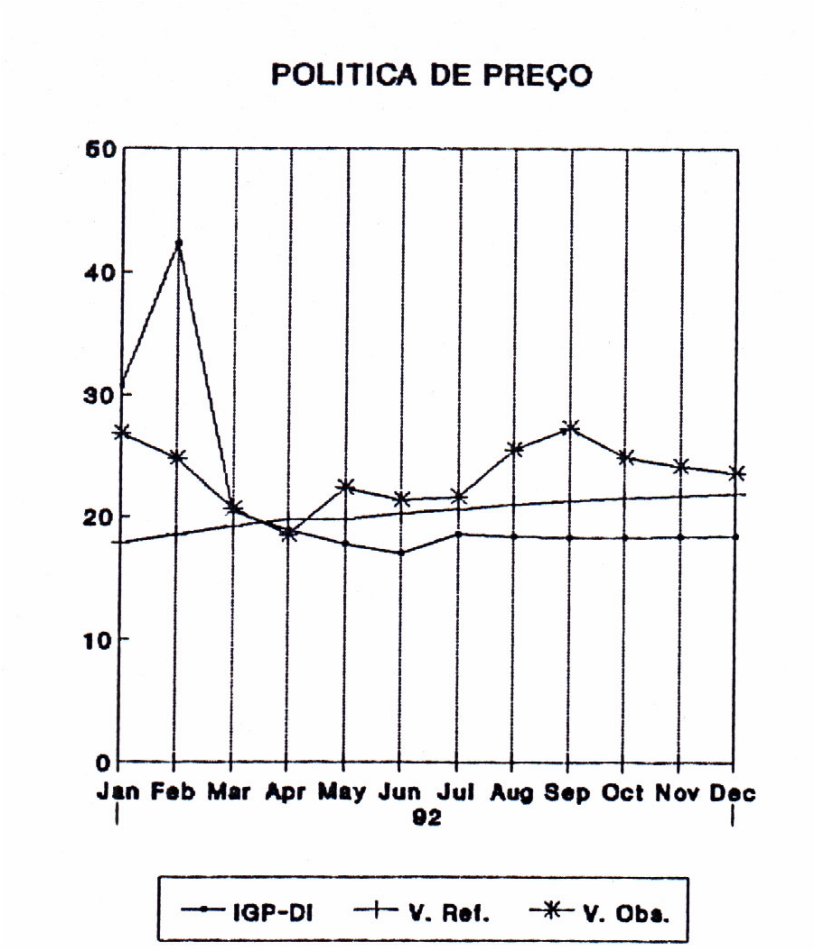

FIGURA 14 – Gráfico representativo dos valores previstos e observados para a variável IGP-DI e os valores previstos para a Variável Referência para o ano de 1992.

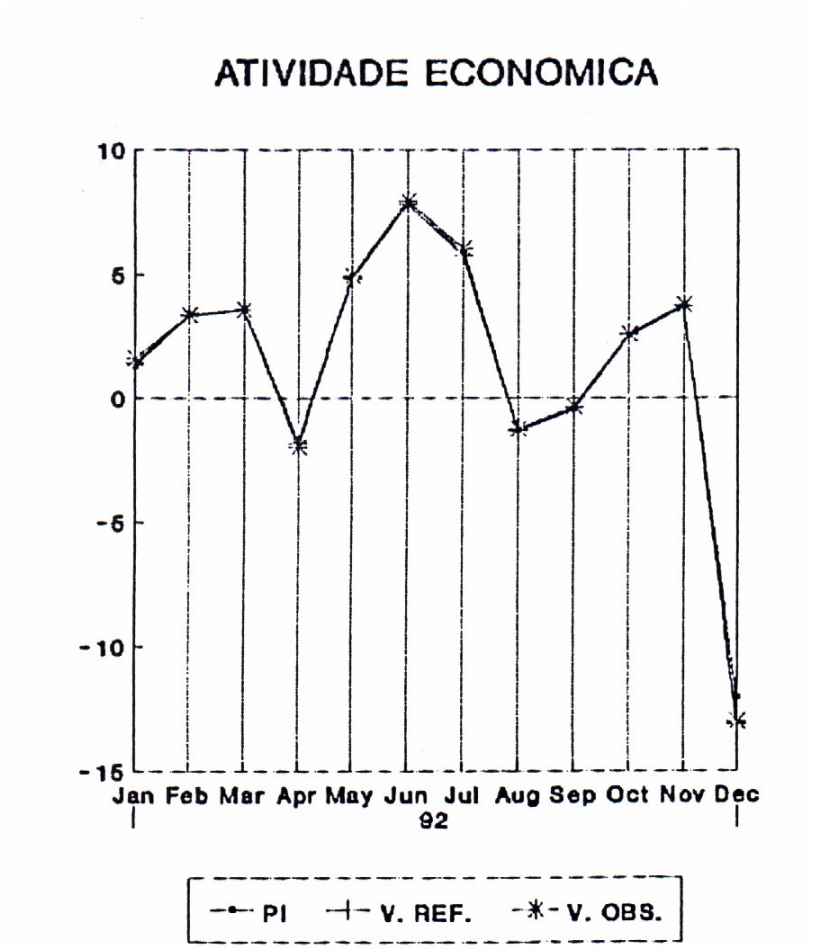

FIGURA 15 – Gráfico representativo dos valores previstos e observados para a variável PRODUÇÃO INDUSTRIAL (PI) e os valores previstos para a Variável Referência para o ano de 1992.

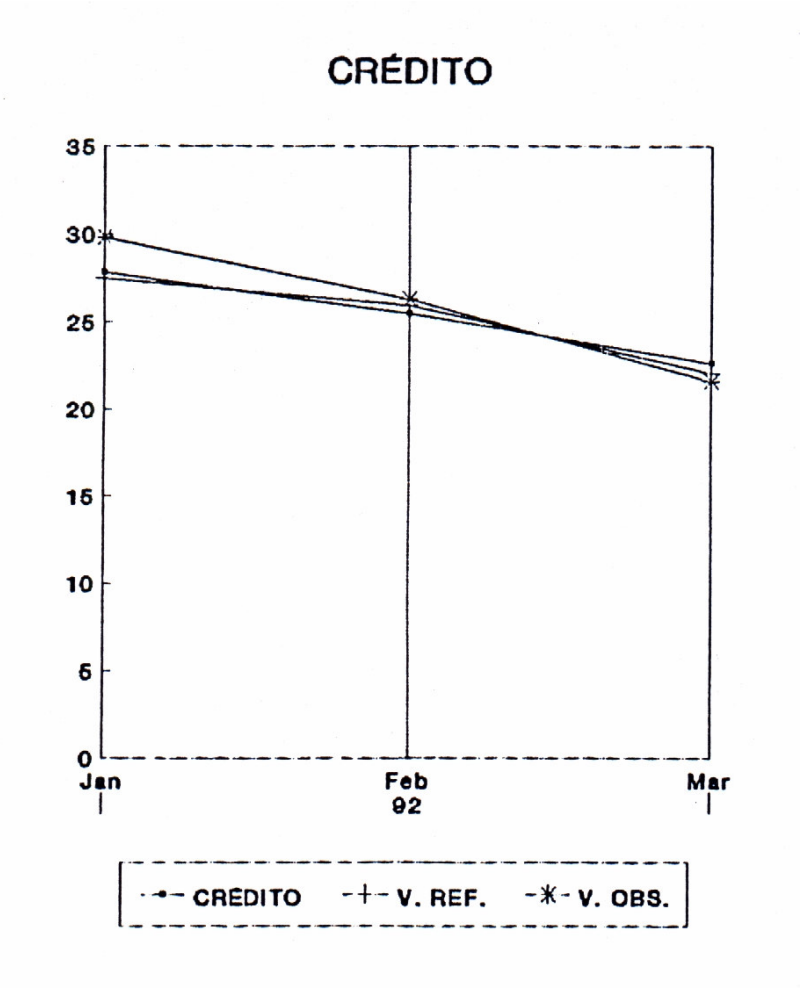

FIGURA 16 – Gráfico representativo dos valores previstos e observados para a variável CRÉDITO e os valores previstos para a Variável Referência para o ano de 1992.

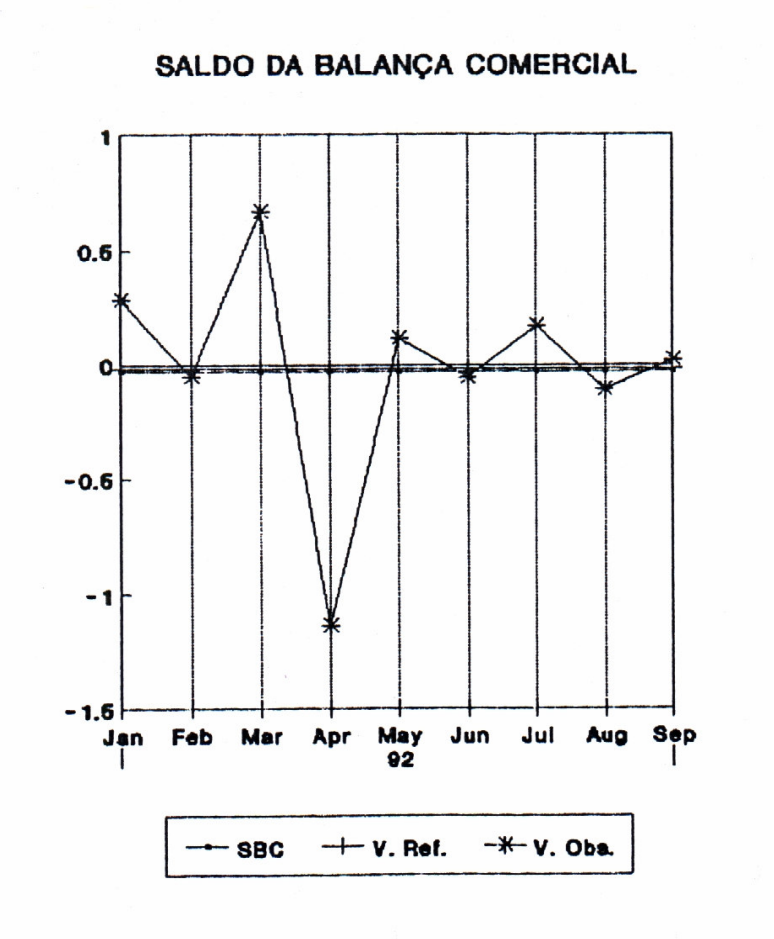

FIGURA 17 – Gráfico representativo dos valores previstos e observados para a variável SALDO DA BALANÇA COMERCIAL (SBC) e os valores previstos para a Variável Referência para o ano de 1992.

ANEXO 3 – DEFINIÇÕES BÁSICAS

Propriedades da Distribuição Normal Multivariada

Seja o par de vetores aleatórios X e Y distribuídos conjuntamente conforme uma normal multivariada tal que (X'Y')' tem média e matriz de covariância dados por:

$$\mu = \begin{bmatrix} \mu x \\ \mu y \end{bmatrix} \quad e \quad \sum = \begin{bmatrix} \sum_{xx} & \sum_{xy} \\ \sum_{yx} & \sum_{yy} \end{bmatrix}$$

Então, a distribuição de X condicional a Y também normal multivariada com média

$$\mu_{X/Y} = \mu_x + \sum_{xy} \sum_{yy}^{-1} (Y - \mu_y)$$

e a matriz de covariância

$$\sum_{X/Y} = \sum_{xx} - \sum_{xy} \sum_{yy}^{-1} \sum_{yy}$$

Observa-se que a matriz de covariância não depende de y diretamente. $\sum$ e $\sum_{yy}$ são assumidas serem não singulares, isto é, inversíveis. Embora $\sum_{yy}^{-1}$ possa ser recolocado por uma pseudo-inversa, se a inversa não existir. A prova deste resultado pode ser encontrado em vários livros da literatura estatística como por exemplo (Harvey, 1990; Johnson & Wichem, 1982).

Amostra Aleatória Multivariada

DEFINIÇÃO 1: Diz-se que um conjunto de dados constitui uma amostra aleatória multivariada se cada indivíduo tenha sido extraído aleatoriamente de uma população de indivíduos e se tenha medido e observado uma série de características.

Seja $X_{(i\,j)}$ a observação da j-ésima variável e o i-ésimo indivíduo, $X_{(i)}$ é um vetor linha que contém as observações de todas as variáveis do i-ésimo indivíduo e $x_{(j)}$ é o vetor coluna que contém todas as observações da j-ésima variável. Se define a matriz dos dados multivariados como:

$$X = (X_{(X\,i\,j)}) = \begin{bmatrix} X_{(1\,1)} & .... & .... & .... & X_{(1\,p)} \\ : & .... & .... & .... & : \\ : & .... & X_{(i\;j)} & .... & : \\ : & .... & .... & .... & : \\ X_{(1\,n)} & .... & .... & .... & X_{(n\;p)} \end{bmatrix}_{n\;X\;p}$$

de dimensão (n x p), que também pode-se expressar por:

$$X = < X_{(1)} \; ... \; X_{(p)} > = \begin{bmatrix} X_{(1)} \\ : \\ : \\ X_{(n)} \end{bmatrix}$$

A partir desta matriz que contém todas as informações estatísticas é possível calcular algumas funções, como no caso univariado, permitindo extrair conclusões dos dados.

Pode-se calcular a média para se ter uma estimação da tendência central e a variância para saber-se a respeito da dispersão dos dados ao redor da média.

Média Amostral

DEFINIÇÃO 2: Dada uma matriz de dados como a da definição (1), a média amostral da j-ésima variável será dada por:

$$\overline{X}_{(j)} = \frac{1}{n} \sum_{i=1}^{n} X_{(i\,j)}$$

e o vetor formado pelos $\overline{X}_{(j)}$ será o vetor centróide.

1. $$\overline{X} = \begin{bmatrix} \overline{X}_{(1)} \\ : \\ : \\ \overline{X}_{(p)} \end{bmatrix}$$

Variância Amostral

DEFINIÇÃO 3: Dada uma matriz de dados como a da definição (1), a variância amostral da j-ésima variavel será dada por:

$$S_{(j\,j)} = \frac{1}{n}\sum_{i=1}^{n}(X_{(i\,j)} - \overline{X}_{(j)})^2$$

e se define a covariância entre a j-ésima e a k-ésima variável por:

$$S_{(j\,k)} = \frac{1}{n}\sum_{i=1}^{n}(X_{(i\,j)} - \overline{X}_{(j)})\,(X_{(i\,k)} - \overline{X}_{(k)})$$

onde:

j, k = 1, ..., p.

A matriz formada pelo arranjo dos $S_{(j\ k)}$ e $S_{(j\ j)}$ será a matriz de variância-covariância amostral, ou simplesmente matriz de covariância amostral.

$$S = \begin{bmatrix} S_{(1\,1)} & .... & .... & .... & S_{(1\ p)} \\ : & .... & .... & .... & : \\ : & .... & S_{(j\,k)} & .... & : \\ : & .... & .... & .... & : \\ S_{(p\,1)} & .... & .... & .... & S_{(p\ p)} \end{bmatrix}$$

Coeficiente de Correlação

DEFINIÇÃO 4: A partir dos elementos da matriz S é possível calcular r, de mesma dimensão de S, cujos elementos são os coeficientes de correlação entre a j-ésima e a k-ésima variável.

$$r_{(j\,k)} = \frac{S_{(j\,k)}}{\sqrt{S_{(j\,j)}\ S_{(k\,k)}}} = \frac{S_{(j\,k)}}{S_{(j)}\ S_{(k)}}$$

Podemos organizar estes valores em uma matriz de correlação amostral, onde a diagonal principal será formada pelo número 1 e será simétrica como a matriz de covariância, pois $r_{(j\,k)} = r_{(kj)}$.

$$r = \begin{bmatrix} 1 & .... & .... & .... & r_{(1\,p)} \\ : & .... & .... & .... & : \\ : & .... & 1 & .... & : \\ : & .... & .... & .... & : \\ r_{(p\,1)} & .... & .... & .... & 1 \end{bmatrix}$$

A matriz S de covariância é uma maneira de expressar a dispersão dos dados ao redor da média. Às vezes é necessário dispor de uma escala que sintetize essa dispersão. Pode-se encontrar um número que expresse a variabilidade multivariada a partir de informação contida na mesma matriz S.

Variância Generalizada

DEFINIÇÃO 5: Dada uma matriz S tal como a mostrada na definição (3), a variância generalizada é dada pela seguinte matriz:

$$V = | S |$$

Variância Total

DEFINIÇÃO 6: Dada uma matriz S tal como a mostrada na definição (3), denomina-se variância total o traço da matriz S.

$$tr\ S = \sum_{j=1}^{p} S_{(j\,j)}$$

Tanto a variância generalizada como a variância total serão maiores, quanto maior seja a dispersão dos dados ao redor da média. Segundo Mardia e colaboradores (1979), cada medida reflete aspectos diferentes da variabilidade dos dados. A primeira desempenha um papel muito importante na geração dos estimadores de máxima verossimilhança enquanto que a segunda, a

variância total se utiliza em Análise de Componentes Principais. Deve-se ressaltar que nenhum destes conceitos tem uma equivalência em análise multivariada.

Distribuição Normal P-Variada

DEFINIÇÃO 7: Diz-se que um vetor X de dimensão (p' x' 1) tem distribuição normal p-variada (ou distribuição normal p-dimensional) com vetor média μ e matriz de covariância $\sum$ se sua função densidade é dada por:

$$f(X) = |2\,\pi \sum|^{-1/2\,\exp<-1/2\,(x-\mu)'\,\Sigma^{-1}(x-\mu)>};$$

onde:

$\sum > 0 \quad e \quad \mu$ é finito.

Isto pode ser representado por $X \cong N\,p\,(\mu, \sum)$, que deve ler-se "X tem distribuição normal p, com média μ e matriz de covariância $\sum$".

A matriz S da definição (3) é um estimador da matriz $\sum$ da definição anterior e o vetor X da definição (2) é um estimador do vetos μ da função f(x). A matriz r da definição (4) é um estimador da matriz $\sum$ quando as variáveis são padronizadas, e faz-se quando se quer transformar as variáveis originais $X_{(ij)}$ para que se tenha média zero e variância 1. Geralmente uma nova variável resultante desta transformação se simboliza por $Z_{(ij)}$. Esta transformação é feita se as p variáveis forem unidades amplamente diferentes, pois as combinações lineares delas podem ter pouca significância. Por este motivo, padronizamos as variáveis por:

$$Z = \frac{X_i - \mu}{\sigma_i} \quad ou$$

$$Z_i = \frac{X_i - X}{s_i} \qquad para \qquad i = 1, 2, ..., p$$

podendo ser mais interessante, pois Z_i é adimensional.

A matriz $\sum$ em uma distribuição multinomial ou normal multivariada pode adotar outras formas particulares. Quando as variáveis não estão correlacionadas, e tem-se quando a covariância

entre duas quaisquer delas é zero, a matriz $\sum$ será diagonal, e só terá elementos não nulos (diferentes de zero) na diagonal principal. Estes serão os valores da variância de cada variável. Se estas variáveis não correlacionadas forem padronizadas (média zero e variância 1), a matriz $\sum$ será a matriz identidade (simbolizada por I), e tem-se uma matriz com números 1 na diagonal principal e zeros fora dela. Se todas as variáveis estudadas tiverem uma variância comum, como se sucede nos modelos mostrados, a matriz de covariância poderia simbolizar-se por σ^2 I. Teríamos uma matriz diagonal em que todos os elementos seriam iguais ao valor da variância comum.

$$\begin{bmatrix} S_{(12)} & 0 & 0 & 0 & 0 \\ 0 & S_{(22)} & 0 & ... & 0 \\ 0 & 0 & S_{(33)} & ... & 0 \\ : & : & : & ... & : \\ 0 & 0 & 0 & ... & S_{(p\,p)} \end{bmatrix}$$

Esta matriz correspondente a matriz das variáveis não correlacionadas: matriz diagonal.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ : & : & : & ... & : \\ 0 & 0 & 0 & ... & 1 \end{bmatrix}$$

Matriz das variáveis não correlacionadas padronizadas: Matriz identidade.

$$\begin{bmatrix} S_{(.)} & 0 & 0 & 0 & 0 \\ 0 & S_{(.)} & 0 & ... & 0 \\ 0 & 0 & S_{(.)} & ... & 0 \\ : & : & : & ... & : \\ 0 & 0 & 0 & ... & S_{(.)} \end{bmatrix}$$

Matriz das variáveis com variância comum: Matriz diagonal.

Conforme a definição (7), para determinar a função de densidade normal multivariada é necessário calcular o determinante da matriz de covariância $\sum$, assim como a sua inversa, para o

qual deve ser não-singular, tem-se um determinante diferente de zero. Isto não ocorre quando o grau de liberdade da matriz é menor que a sua dimensão. A dimensão é dada pelo numero de variáveis que se consideram, e seja a dimensão do vetor X. O grau de liberdade corresponde ao número de vetores linearmente independentes da matriz de dados e se alguma das variáveis consideradas é uma combinação linear de outras incluídas também na matriz, o grau de liberdade será menor que a dimensão e, por fim, a matriz de covariância será singular. Neste caso, a função de densidade normal multivariada não pode ser determinada e se diz que "não existe". Para definir a distribuição é necessário recorrer a outras formas de expressões diferentes como a função geratriz de momentos.

Autovalor e Autovetor

DEFINIÇÃO 8: Seja B uma matriz quadrada de dimensão (p x p) é possível encontrar um escalar λ (lâmbda) e um vetor X de dimensão (p x 1), não nulo tal que:

$$B\,X = \lambda\,X$$

que implica em

$$B\,X - \lambda\,X = 0$$
$$(B - \lambda\,I)\,X = 0$$

tirando-se o vetor X como fator comum à direita, de modo que a operação matricial seja viável.

Para encontrar-se outra equação que permita completar o sistema se estabelece a condição de que os vetores próprios estejam normalizados. Isto equivale a dizer em termos algébricos que a soma dos quadrados dos elementos do vetor deve ser 1.

Assim, deverá cumprir-se que:

$$\alpha_1^2 + ... + \alpha_2^2 = 1$$