3
Metodologia

Neste capítulo serão apresentadas as metodologias utilizadas neste trabalho. A seção 3.1 apresentará a Análise de Fatores cuja relevância é de fundamental importância para a construção dos Modelos de Índice de Difusão. Os fatores serão utilizados no preenchimento dos dados faltantes (seção 3.3), e seus escores, estimados pelo Método das Componentes Principais (seção 3.2), serão as variáveis explicativas do modelo econométrico (seção 3.4).

As seções 3.5 e 3.6 apresentarão os Testes de Diebold-Mariano e de Direction-of-Change, respectivamente. A finalidade de ambos os testes é verificar o poder preditivo do modelo proposto quando comparado aos outros modelos.

3.1
Análise de Fatores

A Análise de Fatores é um método estatístico multivariado que busca resumir um grande número de variáveis em poucas variáveis latentes não observadas, chamadas de fatores. Isto se dá pela redução da dimensionalidade da matriz de dados X através da decomposição da sua matriz de covariância Σ. Extraindo-se os autovalores e autovetores associados a esta matriz, é possível ordenar a combinação de variáveis que mais explicam a variância dos dados. Assim, se poucos fatores são necessários para explicar a maior parte da variabilidade dos dados originais, então há uma redução de dimensionalidade.

Segundo Johnson e Wichern (1998), a motivação para o uso desta análise é tentar separar, em fatores, variáveis que teoricamente deveriam explicar alguma informação comum entre elas. Ou seja, suponha que se tenha um grupo de variáveis que façam parte de certo setor da economia e que são altamente correlacionadas entre si, porém apresentam pouca correlação com um diferente grupo de variáveis que representam outro setor. Considera-se, então, que cada grupo representa um fator que é responsável pelas correlações observadas.

Logo, os fatores estimados apresentam pequena ou nenhuma correlação entre si.

O modelo de fatores mais utilizado é o Modelo de Fator Ortogonal (MFO). Este modelo considera que não há correlação (ortogonalidade) entre os mesmos. Os fatores entrarão como variáveis explicativas no modelo econométrico e a baixa ou nenhuma correlação entre eles satisfará um dos pressupostos clássicos do modelo de regressão linear que é a não ocorrência de multicolinearidade entre as variáveis explicativas.

Considere o vetor aleatório X, com *p* componentes $\left(X_1, X_2, \ldots, X_p\right)$, com média μ e matriz de covariância Σ. O modelo assume que *X* é linearmente dependente de algumas variáveis aleatórias não observáveis F_1, F_2, ..., F_m, chamadas fatores comuns, e p os erros aleatórios ε_1, ε_2, ..., ε_p, inerentes a ele. Pode-se escrever esse modelo como:

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p \end{aligned} \tag{1}$$

Ou, vetorialmente:

$$X - \mu = LF + \varepsilon \tag{2}$$

Onde $X - \mu$ é um vetor (px1), L é a matriz de pesos ou cargas (loadings) (pxm) na qual o coeficiente l_{ij} é o peso da j-ésima variável no i-ésimo fator, F são os fatores comuns (mx1), e ε é a matriz de erros (px1). Pode-se observar que $X - \mu$ é a combinação linear das *m*+p variáveis latentes F e ε.

Em relação aos vetores F e ε, cabe fazer algumas hipóteses:

$$E(F) = 0_{(mx1)},\ COV(F) = E(FF') = I_{(mxm)}.$$

$$E(\varepsilon) = 0_{(px1)},\ COV(\varepsilon) = E(\varepsilon\varepsilon') = \Psi_{(pxp)},$$

(3)

onde Ψ é uma matriz diagonal.

Considera-se, também, que F e ε são descorrelacionados,

$$COV(F, \varepsilon) = E(F\varepsilon') = 0_{(pxm)}.$$

(4)

A partir da $COV(X - \mu)$ pode-se mostrar que a matriz de covariância Σ apresenta a seguinte decomposição:

$$\Sigma = COV(X) = LL' + \Psi .$$

(5)

e que[3]

$$COV(X, F) = L .$$

(6)

E assim segue que:

$$\sigma_{ii}(x_i) = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \psi_i,\ i = 1,2,..., p$$

(7)

Assim, percebe-se que a variância de X_i é decomposta em duas partes: uma correspondente à $\sum_{j=1}^{m} l_{ij}^2$, chamada de "comunalidade", associada aos fatores, e a outra aos erros aleatórios. Com isso, pode-se dizer que a i-ésima comunalidade é a soma dos quadrados dos pesos estimados da i-ésima variável nos m fatores.

[3] Johnson e Wichern (1998, página 398).

Apesar da semelhança do objetivo da Análise de Fatores em estimar L e Ψ (ver equação (1)) com o do Modelo de Regressão Linear, é importante verificar que aquele não utiliza variáveis observadas. Ou seja, na Análise de Fatores, para estimar os coeficientes e o termo de erro aleatório, utiliza-se da matriz de covariância Σ. Para isto, é preciso impor restrições nesta para que suas estimações sejam únicas e adequadas para os dados do estudo (ver Johnson e Wichern (1998, pág. 396)).

Com relação aos métodos de estimação dos parâmetros do modelo, destacam-se o Método das Componentes Principais e o da Máxima Verossimilhança. Ambos são os mais utilizados, pois a solução deles pode ser rotacionada para simplificar a interpretação dos fatores.

No presente trabalho será apresentado apenas o primeiro método, pois segundo Stock e Watson (2002 a e b), este é o indicado para a estimação dos fatores que serão utilizados como variáveis explicativas no Modelo de Índice de Difusão. Para detalhes do Método de Máxima Verossimilhança, ver Johnson e Wichern (1998, página 411).

3.2
Método das Componentes Principais

O Método das Componentes Principais tem como vantagens em relação aos outros métodos de estimação sua simplicidade, fácil implementação e adequação para grandes bases de dados. Além disso, pode ser utilizado para lidar com valores faltantes na base de dados. Apesar dos fatores não serem diretamente interpretados, o uso das regressões lineares ajuda a identificar os fatores extraídos.

O objetivo deste método é estimar $\hat{L}$ (pxm) e $\hat{\Psi}$ (pxp). Sendo Σ uma matriz diagonal, o uso do Teorema da Decomposição Spectral é interessante na demonstração da redução do conjunto de dados pelo método. Assim, seja Σ com seus pares ordenados autovalor, autovetor (λ_i, e_i), com $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Então:

$$\Sigma = \lambda_1 e_1 e'_1 + \lambda_2 e_2 e'_2 + \cdots + \lambda_p e_p e'_p$$

$$= \left[\sqrt{\lambda_1}e_1 \middle| \sqrt{\lambda_2}e_2 \middle| \cdots \middle| \sqrt{\lambda_p}e_p\right] * \begin{bmatrix} \sqrt{\lambda_1}e'_1 \\ \sqrt{\lambda_2}e'_2 \\ \vdots \\ \sqrt{\lambda_p}e_p \end{bmatrix}$$

(9)

Esta é a decomposição da matriz de covariância para a análise de fatores, tendo (m = p) e ψ_i = 0 para qualquer i. Assim, de (7) e (8), tem-se:

$$\Sigma = LL' + 0 = LL'$$

(10)

Nesta decomposição todas as matrizes apresentam dimensão (pxp). Com isso, percebe-se que os pesos (*loadings*) do j-ésimo fator são os coeficientes da j-ésima componente principal.

Como se quer a variância explicada em poucos fatores (m), ignora-se a contribuição dos p-m autovalores que menos explicam a variabilidade dos dados. Assim:

$$\left[\sqrt{\lambda_1}e_1 \middle| \sqrt{\lambda_2}e_2 \middle| \cdots \middle| \sqrt{\lambda_m}e_m\right] * \begin{bmatrix} \sqrt{\lambda_1}e'_1 \\ \sqrt{\lambda_2}e'_2 \\ \vdots \\ \sqrt{\lambda_m}e_m \end{bmatrix} + \psi_{(pxp)}$$

(11)

Onde $\psi_{(pxp)}$ é a matriz diagonal de covariância dos erros aleatórios. Ou seja,

$$\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2 \ ,$$

(12)

para $i = 1,2,\cdots, p$. Essa solução é conhecida como Solução das Componentes Principais.

A estimação impõe restrições na matriz de covariância dos dados Σ. Sabe-se, pela inferência estatística, que a matriz de covariância amostral S é o estimador consistente, não tendencioso e eficiente de Σ.

Para aplicar o método na matriz S a partir dos dados originais, deve-se aplicar a variância ($\sum (Xi - \overline{X})^2$) na forma matricial. Logo:

$$x_j - \overline{x} = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{pj} \end{bmatrix} - \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix} = \begin{bmatrix} x_{1j} - \overline{x}_1 \\ x_{2j} - \overline{x}_2 \\ \vdots \\ x_{pj} - \overline{x}_p \end{bmatrix}$$

(13)

Assim, a solução do Método da Componente Principal para a Análise de Fatores é simplesmente especificar os pares aleatórios $(\hat{\lambda}_1, \hat{e}_1)$, $(\hat{\lambda}_2, \hat{e}_2)$, ..., $(\hat{\lambda}_p, \hat{e}_p)$, onde $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Fazendo m ‹ p ser o número de fatores comuns, a matriz de pesos (*loadings*) $\{\hat{l}_{ij}\}$ é dada por:

$$L = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1 \middle| \sqrt{\hat{\lambda}_2}\hat{e}_2 \middle| \cdots \middle| \sqrt{\hat{\lambda}_m}\hat{e}_m \right]$$

(14)

Analogamente a (12), os erros aleatórios são estimados como:

$$\hat{\psi}_i = s_{ii} - \sum_{j=1}^{m} \hat{l}_{ij}^2$$

(15)

É fácil verificar por (11) e (14) que a contribuição total do primeiro fator equivale a $\hat{\lambda}_1$, assim como a contribuição do segundo fator é igual a $\hat{\lambda}_2$, e assim por diante (assumindo-se que o autovetor tenha tamanho unitário).

A proporção do total da variância (Θ) explicada pode assim ser definida:

$$\Theta = \begin{cases} \dfrac{\hat{\lambda}_j}{s_{11} + s_{22} + \cdots + s_{pp}} \\ \dfrac{\hat{\lambda}_j}{\varphi} \end{cases}$$

(16)

Onde s_{ii} equivale à variância amostral e φ equivale ao número de fatores estimados. Cabe afirmar que a primeira razão deve-se quando da utilização da matriz S (matriz de covariância amostral), e a segunda, da matriz R (matriz de correlação amostral). Esta última aparece como uma alternativa interessante, já que a correlação informa a relação linear entre as variáveis, uma a uma, levando em consideração a variância de cada uma delas.

Com a obtenção da matriz de pesos L, pode-se recuperar a matriz de covariância Σ por (10). Assim, quando se aplica a propriedade da esperança matemática, o termo de erro aleatório não é utilizado.

Os escores serão utilizados como variáveis explicativas no modelo econométrico. Eles são estimativas dos fatores não observados da matriz de dados X. Espera-se encontrar fatores que representem grupos que caracterizam setores da economia.

Uma informação pertinente a respeito do Método das Componentes Principais está na presença de valores faltantes na matriz de dados. Ele não pode ser implementado na sua forma tradicional quando há informações incompletas na amostra.

A literatura para a solução do problema de dados faltantes (*missing values*) é extensa e bastante explorada. Existem diversas maneiras de solucionar tais problemas, que vão desde soluções simples, como interpolação e utilização da média amostral para seus preenchimentos, a técnicas mais avançadas como o uso do Filtro de Kalman e do Algoritmo EM (*Expectation-Maximization Algorithm*).

Devido à presença de *missing values* e séries com diferentes frequências, foi feito um estudo bastante cauteloso utilizando-se o algoritmo EM para o preenchimento desses dados faltantes. Na seção a seguir será apresentada uma explicação minuciosa deste algoritmo.

3.3
O Algoritmo EM para Análise de Componentes Principais

Utilizado para preenchimento de dados faltantes na matriz de dados, o algoritmo *Expectation-Maximization* (EM) foi originalmente formulado por Dempster et al.(1977) e é utilizado para satisfazer a limitação decorrida do processo de estimação das Componentes Principais. Na base de dados do presente estudo, o uso deste algoritmo foi de fundamental importância, pois havia diversas séries com periodicidades incompletas e frequências diferentes. Para solucionar este problema, foi utilizada uma versão do mesmo, denominada Algoritmo EM para Análise de Componentes Principais (ACP), elaborado por Rowies (1997) e adotada por Stock e Watson (2002 a e b).

Segundo Rowies (1997), o preenchimento dos dados faltantes por seu algoritmo apresenta diversas vantagens. Primeiro, permite um pequeno custo computacional ao calcular os autovalores e seus respectivos autovetores, mesmo na presença de matrizes de dados de grandes dimensões. Segundo, permite suas estimações, mesmo tendo presente dados faltantes. Por fim, aproveita todos os benefícios do algoritmo EM tradicional em termos de estimar as informações faltantes pelo método da máxima verossimilhança a cada iteração.

Chamando de y a amostra “completa” do espaço amostral Y, e x a “incompleta” do espaço amostral X, pode-se referir x como uma matriz “completa” na qual somente é vista desse jeito via y.

Seja a densidade amostral $f(x|\phi)$, onde ϕ é o vetor de parâmetros, pode-se derivar sua família de densidade amostral $g(y|\phi)$ e escrever sua relação da seguinte maneira:

$$g(y|\phi) = \int_{X(y)} f(x, y|\phi)dx$$

(17)

O algoritmo EM encontra o valor ϕ que maximiza $g(y|\phi)$ dado o valor obervado y, usando a densidade amostral associada $f(x, y|\phi)$.

Na p-ésima iteração, o procedimento é realizado em duas etapas:

- Expectation (E): são estimados os estados não observados $t(x)$ a partir dos dados completos.

$$t^{(p)} = E\left[t(x)\middle|y, \phi^{(p)}\right]$$

(18)

- Maximization (M): estima-se o parâmetro do vetor $\phi^{(p+1)}$ como solução das equações,

$$E\left[t(x)|\phi\right] = t^{(p)}$$

(19)

definindo o estimador de máxima verossimilhança de ϕ.

3.4
Modelo de Índice de Difusão Linear

O modelo elaborado por Stock e Watson (2002 a e b) considera um pequeno número de fatores latentes estimados como representantes das variáveis de cada setor da economia. Com isto, utilizando esses fatores como preditores, pode-se dizer que representam o resumo do movimento das séries a partir de sua matriz de correlação. O uso da Análise de Fatores, neste sentido, é uma solução sofisticada que satisfaz todos os pressupostos estatísticos.

Conforme discutido na seção 2.1, diversos foram os obstáculos encontrados para se construir modelos que satisfaçam a teoria e possibilitem fazer uma previsão confiável. O uso do grande número de candidatos a preditores (N) afetaria pressupostos importantes do modelo clássico de regressão linear. Dentre esses problemas, destaca-se o aumento nos erros de estimação, tendo como consequência uma diminuição na confiabilidade da previsão.

No artigo original, os autores mostraram que quando N é grande, as estimativas dos fatores pelo Método das Componentes Principais são consistentes e tendem ao verdadeiro fator com erros idiossincráticos correlacionados[4]. Este resultado confirma-se mesmo quando N > T, onde N são os candidatos a previsores e T o tamanho da amostra para estimação. Mostraram, também, que a previsão $\hat{y}_{t+h}$ dos fatores multiplicados por seus coeficientes (ambos estimados) converge para os verdadeiros valores y_{t+h}. Além disso, sabendo do fato da instabilidade temporal dos modelos de previsão de ciclos de negócios, eles comprovaram que a consistência dos resultados mantém-se quando ela é pequena e fracamente dependente no espaço.

Para se fazer a previsão do modelo, deve-se usar dois procedimentos. O primeiro consiste em estimar séries temporais dos fatores (escores) a partir dos N preditores. O segundo consiste em estimar por Mínimos Quadrados Ordinários a relação entre a variável a ser prevista e os fatores.

Assim, sejam X_t os N candidatos a previsores, T o tamanho da amostra para estimação e y_t a variável a ser estimada, assume-se que o vetor (X_t, y_{t+h}) admita ser representado por r fatores latentes comuns através do modelo de fatores F_t. A estimação do modelo dá-se algebricamente da seguinte maneira:

$$X_t = \Lambda F_t + e_t \tag{20}$$

[4] Caso os distúrbios e_t em (20) fossem independentes, esta equação se equivaleria ao Modelo Clássico de Análise Fatorial. Para mais detalhes, ver Johnson e Wichern (1998).

e

$$\hat{y}_{t+h} = \hat{\beta}_F^{'} \hat{F}_t + \hat{\beta}_w^{'} w_t + \hat{\varepsilon}_{t+h}$$

(21)

onde h é o horizonte de previsão, os vetores $\hat{F}_T$ e w_t são os vetores com os fatores estimados e os lags de y_t, respectivamente. Além disso, ε_{t+h} é o erro de previsão e e_t é o vetor de distúrbios idiossincráticos.

3.5
Teste de Diebold-Mariano

Apresentado em Diebold-Mariano (1995), este teste é utilizado para comparar, estatisticamente, o desempenho preditivo do modelo de interesse, tendo como *benchmark* as projeções ingênuas do modelo *random walk* (passeio aleatório). Dentre diversos testes propostos na literatura de previsão de dados no tempo, este é o mais utilizado, tendo como uma alternativa interessante o teste de Giacomi-White (2006).

Na sua forma tradicional assume-se a hipótese nula (H_0) de que as previsões do modelo a ser testado e do passeio aleatório sejam iguais. O procedimento considera a diferença do erro quadrático médio (EQM) entre ambos. Ou seja,

$$H_0 : d_t = 0$$
$$H_a : d_t \neq 0$$

(20)

onde,

$$d_t = L(y_t) - L(z_t),$$

(21)

$$L(y_t) = EQM(y_t) = \sum_{t=1}^{T} (y_t - \hat{y}_{t|t-1})^2 \ .$$

(22)

e

$$L(z_t) = EQM(z_t) = \sum_{t=1}^{T} (y_t - \hat{z}_{t|t-1})^2 \ .$$

(23)

sendo x_t, y_t e z_t os dados reais, as previsões do modelo a ser testado e as do *random walk*, respectivamente, com t = 1, ..., T.

Assim, de (22) e (23), conclui-se que o teste considera que a distância (EQM) entre as previsões do modelo e a ingênua é igual.

A estatística de teste considera que, sob H_0, não existe diferença via métrica EQM entre ambas as previsões, assim:

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}} \overset{a}{\sim} N(0,1)$$

(24)

onde $\hat{f}_d(0)$ é um estimador consistente de $f_d(0)$, sendo

$$f_d(0) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma_d(\tau) \,,$$

(25)

a densidade espectral de L na frequência zero,

$$\gamma_d(\tau) = E\left[(d_t - \mu)(d_{t-\tau} - \mu)\right]$$

(26)

é a autocovariância de L com τ lags e μ a média da população de L.

Com isso, segue-se que:

$$2\pi\hat{f}_d(0) = \sum_{\tau=-(T-1)}^{(T-1)} l\left(\frac{\tau}{S(T)}\right) * \hat{\gamma}_d(\tau) \tag{27}$$

onde

$$l\left(\frac{\tau}{S(T)}\right) \tag{28}$$

é o chamado *lag window* e, $S(T)$, o *truncation lag*. Ademais,

$$\hat{\gamma}_d(\tau) = \frac{1}{T}\sum_{t=|\tau|+1}^{T}\left(d_t - \bar{d}\right)\left(d_{t-|\tau|} - \bar{d}\right) \tag{29}$$

é a função de autocovariância.

3.6
Teste de Direction-of-Change

Considerado um caso particular do Teste de Diebold-Mariano, é também um teste que verifica o poder preditivo do modelo a ser testado. Formulado pelos mesmos autores, busca verificar se as previsões do modelo a ser testado apresentam o mesmo sinal dos dados reais. Em outras palavras, o objetivo é saber se os movimentos preditos e reais têm o mesmo movimento.

$$H_0 : \bar{d} \geq \frac{1}{2}$$
$$H_a : \bar{d} < \frac{1}{2}$$

(30)

Sendo $\bar{d}$:

$$\bar{d} = \frac{\sum_{t=1}^{n} d_t}{T}$$

(31)

onde d_t é uma variável indicadora, tal que:

d_t {
- = 1, se a previsão do modelo acerta a mudança de direção;
- = 0, caso contrário.

Assim, um valor de $\bar{d}$ maior do que 0.5 indica que o modelo consegue prever a mudança de direção. Por outro lado, se a estatística for menor do que 0.5, as previsões tendem a não seguir o movimento real da série.

A estatística de teste pode assim ser definida:

$$\frac{\left(\bar{d} - 0.5\right)}{\sqrt{0.25/T}} \overset{a}{\sim} N(0,1) .$$

(32)