**Revista Eletrônica
Paulista de Matemática**

ISSN 2316-9664
Volume 10, dez. 2017
Edição Ermac

Eduardo Rosa Godinho

UNESP - Universidade Estadual
Paulista "Júlio de Mesquita Filho"
Faculdade de Engenharia de Bauru
eduardorosagodinho@gmail.com

José Antônio Rico Coque

UNESP - Universidade Estadual
Paulista "Júlio de Mesquita Filho"
Faculdade de Engenharia de Bauru
jose.coque@hotmail.com

Giovanna Gomes Huysmans

UNESP - Universidade Estadual
Paulista "Júlio de Mesquita Filho"
Faculdade de Engenharia de Bauru
giovanna.huysmans@gmail.com

**Nair Cristina Margarido
Brondino**

UNESP - Universidade Estadual
Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
brondino@fc.unesp.br

Redução da dimensão de um conjunto de variáveis: decomposição em valores singulares versus algoritmo colônia de formigas

Reduction of the size of a set of variables: Singular Values
Decomposition versus Ant Colony Algorithm

Resumo

O presente trabalho apresenta os resultados da aplicação de dois métodos de redução de variáveis, sendo o primeiro o método de Decomposição em Valores Singulares (SVD), em conjunto com a rotação do tipo Varimax, e o segundo, o método baseado em Otimização por Colônia de Formigas (ACO). Ambos os métodos destinam-se a descartar variáveis altamente correlacionadas com outras, de forma a explicar a maior variabilidade do sistema, sem que se repita informação. Para isso, o método SVD realiza uma mudança de base, de forma que os vetores da nova base, denominados de componentes principais, são obtidos a partir da rotação dos eixos iniciais nas direções de maior variabilidade, o que permite que um número pequeno de componentes represente um grande percentual da variação total. O método baseado em ACO, por sua vez, calcula a similaridade de cossenos entre pares de variáveis, com vistas a descobrir se elas fornecem a mesma informação para o sistema. Vinte e sete variáveis econômicas e sociodemográficas brasileiras foram utilizadas nessa aplicação. Desse total, 14 variáveis foram descartadas por cada método, sendo que sete variáveis em comum foram mantidas pelos métodos.

Palavras-chave: Redução de variáveis. Decomposição por Valores Singulares. Otimização por Colônia de Formigas. Análise de correlação.

Abstract

The current work presents the results from the application of two methods of reduction of variables, being the first the method of Singular Values Decomposition (SVD), in conjunction with the Varimax-type rotation, and the second one based on Ant Colony Optimization (ACO). Both of them are concerned with choosing the variables that explain the greater variability of the system, without repeating information by two or more different variables. SVD method promotes a change of basis in which the new vectors, denominated principal components, come from the rotation of the original vectors at the directions of greater variability. The ACO based method calculates the similarity of cosines of pairs of variables, in order to find out if they provide the same information for the system. Twenty seven economic and demographic Brazilian variables were used in this application. Both SVD and ACO based methods discarded fourteen variables each one.

Keywords: Variables reduction. Singular Values Decomposition. Ant Colony Optimization. Correlation analysis.

1 Introdução

Este trabalho é parte integrante de um projeto de pesquisa maior, cujo objetivo é modelar o consumo de combustíveis e a emissão de CO₂ gerados pelo transporte rodoviário brasileiro, em função de variáveis de natureza sociodemográfica e econômica. Tendo em vista que o projeto maior visa a ajustar uma função global, por meio da aplicação de redes neurais artificiais, a dimensão do conjunto de entrada torna-se um comprometedor, uma vez que o método a ser empregado faz classificações, a partir do reconhecimento de padrões. Dessa forma, ao se usar poucas variáveis de entrada, o desempenho do modelo pode ser comprometido, em virtude de não fornecer toda a informação necessária. Por outro lado, um número excessivo de variáveis de entrada pode gerar um número muito grande de padrões a serem reconhecidos e, dessa forma, a rede apresentará dificuldades em classificar de maneira conveniente uma entrada desconhecida, ou seja, que não foi usada em nenhuma etapa do treinamento. Sendo assim, quando um número muito grande de variáveis está disponível, a utilização de métodos de redução, com vistas a manter somente aquelas que trazem alguma informação relevante ao sistema, pode aumentar o poder de classificação da rede.

Em geral, nesse processo de seleção, utiliza-se métodos mais tradicionais como a Análise de Componentes Principais (PCA) e a Decomposição em Valores Singulares (SVD). Porém, métodos baseados em sistemas biológicos têm aparecido na literatura como substitutos eficientes para diversos tipos de problemas, em especial, aqueles que visam a ajustar funções. Dentro desse contexto, visando ao melhor resultado possível ao final do processo de seleção e posterior ajuste da função global, este trabalho apresenta os resultados fornecidos pela aplicação de dois métodos de seleção, a saber: o método estatístico Decomposição em Valores Singulares (SVD) e um método heurístico baseado em Otimização por Colônia de Formigas (ACO).

No caso da Decomposição em Valores Singulares, todas as variáveis são tratadas de maneira simultânea e a covariância entre as mesmas é levada em consideração no processo de seleção. Esse tipo de técnica, em geral, pode ser utilizada para gerar novas entradas para a rede neural, a partir de combinações lineares das variáveis disponíveis, de maneira que as novas entradas não sejam correlacionadas entre si e que poucas variáveis geradas expliquem a maior parte da variabilidade do sistema. Nessa aplicação, o método não foi utilizado com esse fim, uma vez que o objetivo não era gerar novas variáveis para o treinamento, mas sim descartar as que traziam pouca informação adicional.

No método baseado em Algoritmo Colônia de Formigas, por sua vez, as variáveis entram no sistema sem estar descorrelacionadas e são tratadas em agrupamentos gerados de forma aleatória. Sendo assim, no início do processo, quando as formigas partem de variáveis diferentes, mas que são altamente correlacionadas, as demais variáveis a serem selecionadas pelo método podem diferir, pelo fato de que as duas variáveis de partida podem apresentar relações distintas com as outras variáveis do sistema. Como consequência, esse método pode estar mais propenso a escolher variáveis altamente correlacionadas ao final do processo.

O método SVD foi aplicado por Godinho e Brondino (2017), que trabalharam com a redução de um conjunto de 19 séries temporais de variáveis econômicas e demográficas, a partir da utilização de duas componentes principais. O mesmo problema foi abordado por Coque e Brondino (2017) a partir de um método baseado em Algoritmo Colônia de Formigas. No presente trabalho, além da inclusão de 12 variáveis de natureza socioeconômica e demográfica no conjunto, tais como número médio de anos de estudo, taxa de desemprego, taxa de urbanização e matrículas nos ensinos básico e superior, será utilizado um número maior de componentes principais com o objetivo de melhorar o percentual de variância explicada pelas

componentes. Além dessas modificações, esse trabalho também contempla uma comparação dos resultados da aplicação dos dois métodos na seleção.

Sendo assim, tendo em vista que essas duas técnicas apresentam algumas diferenças na forma de seleção, decidiu-se por aplicá-las ao mesmo problema, com vistas a comparar os produtos finais, em termos de coerência nas escolhas. Análise semelhante não foi encontrada na literatura pesquisada e, nesse contexto, encontra-se o diferencial desse trabalho.

Cabe ressaltar que, em uma etapa posterior, os dois conjuntos selecionados serão adicionados a uma rede neural e o desempenho dos mesmos será confrontado a partir dos erros observados.

Além dessa Introdução e da descrição dos dados apresentada na seção 2, este artigo divide-se em mais três seções. A terceira apresenta a teoria acerca de cada método de redução. Na quarta seção, os resultados obtidos por cada método, assim como uma comparação entre os mesmos, serão apresentados. A última seção é dedicada às considerações finais.

2 Dados utilizados

Para essa aplicação, considerou-se 27 séries temporais obtidas junto a sites de órgãos oficiais e medidas mensalmente entre julho de 2001 e julho de 2016, a saber (AGÊNCIA NACIONAL DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS, 2016; ASSOCIAÇÃO NACIONAL DOS FABRICANTES DE VEÍCULOS AUTOMOTORES, 2016; BANCO CENTRAL DO BRASIL, 201-?; INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA, 2015; INSTITUTO DE PESQUISA ECONÔMICA APLICADA, 2016?; KNOEMA, 2016?):

- consumo em $\text{m}^3/\text{mês}$ de álcool anidro (CAA), álcool hidratado (CAH), diesel (CD) e gasolina;
- PIB em US\$;
- número de matrículas no ensino básico (MB); número de matrículas no ensino superior (MS);
- número de admissões de trabalhadores (AT), de demissões de trabalhadores (DT);
- salário mínimo real em R\$ (SM); exportações em US\$ (E); importações em US\$ (I);
- média do número de anos de estudo (MAE);
- preços em US\$ do litro do álcool (PA), do diesel (PD) e da gasolina (PG);
- produção de automóveis (PAu), de caminhões (PC), de comerciais leves (PCL), e de ônibus (PO);
- número de licenciamentos de automóveis (LA), de caminhões (LC), de comerciais leves (LCL), e de ônibus (LO);
- taxa de desemprego (TD); densidade populacional em habitantes por km^2 ; população urbana em relação à total (UR).

3 Descrição dos métodos

3.1 A decomposição em valores singulares

O método de Decomposição em Valores Singulares (SVD) promove a decomposição de uma matriz X por meio da Equação 1.

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \quad (1)$$

onde $\mathbf{X}$ é uma matriz $m \times n$ de posto r , $\mathbf{U}$ e $\mathbf{V}$ são ortogonais e $\mathbf{\Sigma}$ é uma matriz diagonal.

A matriz $\mathbf{X}^T\mathbf{X}$ é proporcional à covariância das linhas de $\mathbf{X}$, mantendo as colunas como dimensão amostral. A matriz $\mathbf{U}$ é ortogonal e, portanto, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, em que $\mathbf{U}^T$ é a inversa à esquerda de $\mathbf{U}$. Além disso, pelo fato de que $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$ é matriz diagonal, tem-se $\mathbf{\Sigma} = \mathbf{\Sigma}^T$. Usando essas propriedades, a matriz $\mathbf{X}^T\mathbf{X}$ pode ser reescrita conforme a Equação 2.

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T \quad (2)$$

A matriz $\mathbf{X}^T\mathbf{X}$ é simétrica e positiva semi-definida. Pelo Teorema Espectral (HOFFMAN; KUNZE, 1971), a mesma é diagonalizável numa base de autovetores ortonormais e seus autovalores são reais positivos. Pode-se observar na Equação 2 que a matriz $\mathbf{V}$ diagonaliza $\mathbf{X}^T\mathbf{X}$ e, portanto, suas colunas são os autovetores ortonormais (*componentes*) de $\mathbf{X}^T\mathbf{X}$ associados aos autovalores $\lambda_i = \sigma_i^2$.

Fazendo a mesma análise para $\mathbf{X}\mathbf{X}^T$, que é proporcional à variância das colunas de $\mathbf{X}$, constata-se que as colunas de $\mathbf{U}$ são os autovetores normalizados de $\mathbf{X}\mathbf{X}^T$. Os elementos da diagonal de $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$ são denominados *valores singulares*. Esses autovalores são arranjados em ordem decrescente e são tais que $\sigma_i > 0$ para $1 \leq i \leq r$ e $\sigma_i = 0$ para $r + 1 \leq i \leq n$. Os r valores singulares da diagonal de $\mathbf{\Sigma}$ são as raízes quadradas dos autovalores não nulos de $\mathbf{X}\mathbf{X}^T$ e também de $\mathbf{X}^T\mathbf{X}$ e são tais que os σ_i^2 são proporcionais às variâncias das componentes principais.

A SVD pode ser utilizada para promover a redução da dimensionalidade do conjunto de dados, a partir da manutenção de k valores singulares no conjunto, de forma que esses expliquem a maior proporção de variância (PV) possível, conforme Equação 3.

$$PV = \frac{\sigma_1^2 + \dots + \sigma_k^2}{\sum_{i=1}^r \sigma_i^2}, \quad (3)$$

onde σ_1 e σ_r são o maior e o menor valor singular, respectivamente.

A matriz $\mathbf{U}$ é organizada de tal forma que o primeiro eixo é o mais significativo, uma vez que corresponde ao maior autovalor e consequentemente a variância dos pontos ao longo desse eixo é a maior. Usualmente, a aplicação do método permite escolher a dimensão da matriz $\mathbf{X}$ que contém a estrutura de interesse e aquela que contém variabilidade amostral.

Após a obtenção das componentes principais, pode-se aplicar às mesmas uma rotação, de forma que cada novo vetor apresente um número pequeno de cargas (*loadings*) altas e um número grande de cargas pequenas. Com esse arranjo, cada vetor desse novo conjunto representará um número pequeno de variáveis, o que torna mais fácil identificar os agrupamentos. Dentre os métodos disponíveis, o mais comumente empregado é o método *Varimax* (KRUSKAL, 1983). Esse método consiste em encontrar uma rotação que maximize a variância das cargas, ou seja, que maximize V na Equação (4).

$$V = \sum (q_{il}^2 - \overline{q_{il}^2})^2, \quad (4)$$

onde q_{il}^2 é o quadrado da carga da i -ésima variável no fator l e $\overline{q_{il}^2}$ é a média dos quadrados dessas cargas.

Nesta aplicação, a matriz X , de ordem 181×27 , foi organizada de tal forma que as variáveis relativas a cada uma das séries temporais foram normalizadas e dispostas por colunas. A Decomposição por Valores Singulares e as representações gráficas associadas foram obtidas a partir de um *script* construído no software MATLAB®.

3.2 A otimização por colônia de formigas

Os métodos que utilizam Algoritmo por Colônia de Formigas (ACO - Ant Colony Optimization) são baseados no comportamento biológico de uma colônia de formigas. O seu princípio é o modo como as formigas procuram por comida e voltam à colônia, depositando feromônio pelo caminho que passam, o qual serve como meio de comunicação entre esta e as outras formigas para guiá-las até o alimento. As formigas da colônia tendem a escolher o caminho com maior concentração de feromônio, pois este é o caminho pelo qual a maioria das formigas passou mais vezes em menos tempo, ou seja, é o caminho mais curto entre a colônia e a fonte de comida (DORIGO; STÜTZLE, 2004).

O algoritmo utilizado nesse trabalho foi implementado em MATLAB® e é denominado UFSACO (TABAKHI; MORADI; AKHLAGHIAN, 2014). Se $A = (a_1, a_2, \dots, a_n)$ e $B = (b_1, b_2, \dots, b_n)$ são os vetores que representam as variáveis A e B , uma medida de similaridade entre as mesmas é dada pelo módulo do cosseno que os vetores A e B formam e é dada pela Equação 5. Pode-se observar que se A e B forem aproximadamente paralelos, ou seja, representarem praticamente a mesma informação para o sistema, tem-se $\text{sim}(A, B) \cong 1$ e se os dois forem ortogonais, teremos $\text{sim}(A, B) \cong 0$.

$$\text{sim}(A, B) = \left| \frac{\sum_{i=1}^p (a_i \cdot b_i)}{\left(\sqrt{\sum_{i=1}^p a_i^2} \right) \cdot \left(\sqrt{\sum_{i=1}^p b_i^2} \right)} \right| \quad (5)$$

O problema de seleção pode ser representado por um grafo não orientado totalmente conectado, em que as variáveis representam os nós e as similaridades representam o peso dos arcos que associam os respectivos pares de variáveis. No momento inicial, as formigas são distribuídas entre os nós, sendo que as quantidades iniciais de feromônio em cada arco, denotadas por τ , são todas constantes. A cada iteração, as formigas deslocam-se aleatoriamente para um nó diferente, a partir de duas regras possíveis: gulosa ou probabilística.

Suponhamos que i representa a variável na qual a formiga está e j representa a próxima variável a ser visitada. Na regra gulosa, a formiga que está na variável i escolhe deslocar-se para a variável j , de acordo com a Equação 6.

$$j = \text{argmax} \{ [\tau_u] [\eta(F_i, F_u)]^\beta \}, \quad (6)$$

onde $\eta(F_i, F_u) = 1/\text{sim}(A, B)$, $\beta > 0$ é um parâmetro que controla a importância da similaridade versus o feromônio, J_i^k é o conjunto de variáveis ainda não visitadas, $u \in J_i^k$ e τ_u é a quantidade de feromônio atribuída a u .

Na regra probabilística, a formiga decide deslocar-se para a cidade j , de acordo com a regra dada pela Equação 7. Para essa aplicação, assumiu-se que 30% das formigas escolhem o método guloso e as demais escolhem o método probabilístico.

$$P_k(i, j) = \begin{cases} [\tau_j] [\eta(F_i, F_j)]^\beta / \sum_{u \in J_i^k} [\tau_u] [\eta(F_i, F_u)]^\beta, & \text{se } j \in J_i^k \\ 0, & \text{caso contrário} \end{cases} \quad (7)$$

A cada iteração, a quantidade de feromônio associada à variável i é atualizada de acordo com a Equação 8.

$$\tau_i(t+1) = (1 - \varrho)\tau_i(t) + \frac{FC[i]}{\sum_{j=1}^n FC[j]}, \quad (8)$$

em que n é o número de atributos originais; $\tau_i(t)$ e $\tau_i(t+1)$ representam as quantidades de feromônio relativo à variável i nos tempos t e $t+1$, respectivamente; ϱ é o parâmetro de evaporação de feromônio; $FC[i]$ é um contador relativo ao número de visitas à variável i .

Ao final do processo, aquelas variáveis que possuírem os maiores valores de feromônio serão as que trazem mais informação ao sistema e as demais poderão ser descartadas.

4 Resultados e discussões

4.1 O resultado da seleção por SVD

Partindo-se da Equação 3, foi possível obter o gráfico apresentado na Figura 1, em que o eixo horizontal faz referência aos maiores valores singulares em ordem decrescente e o eixo vertical à fração explicada da variância por cada valor singular (colunas) e à fração cumulativa explicada pelos primeiros valores singulares (linha). Pode-se observar que a base formada pelos cinco primeiros autovetores responde por 94,18% da variância total.

Figura 1 - Variâncias relativas e cumulativas explicadas pelos valores singulares.

Por meio da Equação 3, é possível aproximar as variáveis originais utilizando a quantidade desejada de valores/vetores singulares. Para exemplificar, a Figura 2 apresenta o comportamento original da variável Exportações (em azul) e a série construída com a utilização das três primeiras componentes principais (em vermelho). Conforme pode ser observado, pelo fato de responder pela maior parcela de variância, a utilização de três vetores singulares promove uma boa aproximação da série original

Figura 2 - Exportações - série original e aproximada a partir dos 3 primeiros vetores singulares.

O *biplo*t foi inicialmente proposto por Gabriel (1971) e pode ser interpretado como uma representação gráfica no caso de dados multivariados, análoga ao diagrama de dispersão. Matrizes de posto 2 ou 3, podem ser representadas graficamente no plano e no espaço, respectivamente. Porém, quando o posto dessas matrizes é superior a 3, esse método propõe uma estratégia de representação em uma dimensão menor do que a original, geralmente em espaços de dimensão 2 ou 3. Entretanto, isso só é possível após algum processo de redução, como SVD ou Análise de Componentes Principais. A visualização gráfica fornecida pelo *biplo*t permite revelar importantes características dos dados, tais como agrupamentos, variâncias e correlações entre as variáveis.

A Figura 3 apresenta o *biplo*t com as três primeiras componentes e a Figura 4 apresenta o *biplo*t obtido a partir da rotação das mesmas pelo método *varimax*. Pode-se observar na Figura 3 que a identificação de algum agrupamento a partir das componentes 1 e 2 não é de fácil visualização. Utilizando a rotação, observa-se um melhor agrupamento das variáveis com relação às componentes 1, 2 e 3, assim como pode-se visualizar os escores de cada uma delas.

Vetores aproximadamente paralelos na Figura 4 indicam que as variáveis envolvidas contribuem com a mesma informação para o sistema e, portanto, uma pode ser descartada, sem muita perda de informação. Os ângulos θ entre pares de vetores que aparecem na Figura 4 foram calculados e assumiu-se que para aqueles pares em que $\theta = 0 \pm 0,375rad$ ou $\theta = \pi \pm 0,375rad$, uma das variáveis foi mantida, escolhendo como critério a de maior módulo. O valor $0,375rad$ foi escolhido porque $\cos(0,375) \cong 0,93$, o que indica uma alta correlação entre as variáveis e, desta forma, os vetores envolvidos são quase paralelos. Usando esse critério, as variáveis mantidas foram exportações (E), PIB, preço da gasolina (PG), preço do diesel (PD), produção de automóveis (PAu), produção de caminhões (PC), licenciamento de ônibus (LO), consumo de diesel (CD), consumo de álcool anidro (CAA), salário mínimo (SM), número de demissões de trabalhadores (DT), matrículas no ensino superior (MS), taxa de desemprego (TD).

Figura 3 - *Biplot* das componentes 1, 2 e 3.

Figura 4 - *Biplot* obtido pela rotação das componentes 1, 2 e 3.

4.2 O resultado da seleção por ACO

A Figura 5 mostra a relação de similaridade entre as variáveis, dada pela Equação 4. Quanto mais próximo de 1 (aqui representado pela cor amarelo), mais similares as variáveis são e quanto mais próximas de 0 (azul), mais diferentes. Por exemplo, existe uma baixa similaridade entre o consumo de diesel (CD) e a população urbana em relação à total (UR) e uma similaridade alta entre o preço do diesel (PD) e o preço da gasolina (PG). Na diagonal principal, todos os valores são iguais a 1 porque a similaridade de uma variável com ela mesma é sempre igual a 1, por isso a coloração amarela.

Figura 5 - Matriz de similaridades entre as variáveis.

A Figura 6 mostra o resultado final da análise. Pode-se observar que as variáveis indicadas pela cor verde são as mais importantes, ou seja, exportações (E), consumo de álcool hidratado (CAH), produção de caminhões (PC), produção de ônibus (PO), taxa de urbanização (UR), consumo de álcool anidro (CAA), consumo de diesel (CD), licenciamento de ônibus (LO), produção de automóveis (PAu), taxa de desemprego (TD), preço do álcool (PA), média do número de anos de estudo (MAE) e número de admissões de trabalhadores (AT) possuem as maiores quantidades de feromônio.

Figura 6 - Seleção de variáveis pelo método ACO.

4.3 Comparação entre os resultados obtidos

Com relação ao esforço computacional despendido, os tempos de processamento para as duas técnicas foram medidos em um computador com Processador Intel Core I7 e memória RAM de 16 Gb. Para o método baseado em ACO, foram utilizadas 20 formigas e o processo parou após 20 iterações, apresentando duração de 0,264s até chegar ao resultado final. A SVD demandou um tempo menor, de 0,194s.

Os dois métodos descartaram sete variáveis em comum. Das variáveis restantes, sete foram mantidas por ambos, a saber: exportações (E), produção de automóveis (PAu), produção de caminhões (PC), licenciamento de ônibus (LO), consumo de diesel (CD), consumo de álcool anidro (CAA) e taxa de desemprego (TD). A Tabela 1 apresenta as variáveis mantidas por um método e que não foram mantidas pelo outro.

Uma possível explicação para as diferenças observadas pode ser devida à alta correlação (ρ) observada entre alguns pares de variáveis, tais como número de matrículas no ensino superior e média de anos de estudo ($\rho = 0,98$); preço da gasolina e preço do álcool ($\rho = 0,95$); admissão de trabalhadores e PIB ($\rho = 0,90$) e consumo de álcool hidratado e número de matrículas no ensino superior ($\rho = 0,73$).

Outra possível explicação diz respeito à utilização das covariâncias no processo de seleção pela SVD que, desta forma, faz uso de informação adicional quando comparada ao método ACO, que só leva em consideração a dependência linear.

Tabela 1 – Variáveis diferentes não descartadas pelos dois métodos

Método	Variáveis Selecionadas					
SVD	PIB	PG	SM	PD	DT	MS
ACO	AT	PA	UR	CAH	PO	MAE

5 Conclusões

Na aplicação da SVD, foi possível verificar que um espaço de dimensão três representa 89,95% da variância total, enquanto 84,01% é representada pelas duas primeiras componentes principais. Devido a isso, optou-se em utilizar os três primeiros vetores singulares de forma a captar a maior explicação possível da variância. Após rotação desses vetores, foi possível identificar não só um agrupamento de variáveis, como também indícios de dependência linear entre as mesmas. Uma comparação da série original com a obtida a partir da utilização de três componentes mostrou coerência entre as duas séries, o que mostra que carregam a carga informacional mais relevante da problemática, permitindo uma aproximação efetiva das variáveis originais com uma quantidade reduzida de dados.

O ACO, como um método heurístico, pode permitir soluções distintas, dependendo das suposições iniciais. No início da simulação, os agentes são distribuídos aleatoriamente entre as variáveis e, a partir daí, usando as regras gulosa ou probabilística, escolhem a próxima variável a ser visitada. Nessa aplicação, em particular, o número de variáveis a serem incluídas foi fixado em 13. Apesar desse comportamento, observou-se que o conjunto de variáveis mais visitadas não sofria grandes alterações entre uma simulação e outra.

A comparação entre os dois métodos mostrou que os mesmos descartaram sete variáveis em comum e retiveram também sete variáveis em comum, apresentando coerência na classificação de 14 variáveis. Das variáveis restantes apresentadas na Tabela 1, que foram mantidas por cada método, não houve concordância na retenção. Dentre as variáveis que foram

selecionadas por um método e não pelo outro, observou-se que quatro pares apresentavam correlações altas, o que indica que, enquanto a SVD manteve uma variável, o ACO escolheu outra, altamente correlacionada com a primeira. Entretanto, a eficiência das diferentes seleções poderá ser devidamente avaliada em uma etapa posterior, quando os dois conjuntos servirem de entrada para a rede neural e os erros puderem ser comparados.

Uma possível explicação para essa diferença de seleção, diz respeito à utilização das covariâncias no processo utilizado pela SVD, que usa informação adicional com relação ao método baseado em ACO, que só leva em consideração a dependência linear. Nesse sentido, uma modificação no método baseado em ACO, que incorpore algum tipo de informação acerca das correlações, já vem sendo analisada e, futuramente, essa hipótese poderá ser melhor avaliada.

6 Agradecimentos

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo apoio financeiro ao projeto sob processo número 2016/13727-4.

7 Referências

AGÊNCIA NACIONAL DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS. **Série histórica do levantamento de preços e de margens de comercialização de combustíveis**. Brasília, DF, 2016. Disponível em: <http://www.anp.gov.br/wwwanp/precos-e-defesa-da-concorrenca/precos/levantamento-de-precos/serie-historica-do-levantamento-de-precos-e-de-margens-de-comercializacao-de-combustiveis>. Acesso em: 24 out. 2016.

ASSOCIAÇÃO NACIONAL DOS FABRICANTES DE VEÍCULOS AUTOMOTORES. **Produção, vendas e exportação de autoveículos**. São Paulo, SP, 2016. Disponível em: <http://www.anfavea.com.br/docs/SeriesTemporais.zip>. Acesso em: 25 nov. 2016.

BANCO CENTRAL DO BRASIL. SGS - **Sistema Gerenciador de Séries Temporais – v2.1**: módulo público. [S.l.], [201-?]. Disponível em: <https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries>. Acesso em: 5 nov. 2016.

COQUE, J. A. R.; BRONDINO, N. C. M. Uso de um algoritmo colônia de formigas na seleção de variáveis dependentes e independentes para futura construção de um modelo de previsão de consumo de combustíveis. In: ENCONTRO REGIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL, 4., 2017, Bauru. **Caderno de trabalhos completos e resumos**. Bauru: Unesp, Faculdade de Ciências, 2017. p. 571-573. Disponível em: <http://www.fc.unesp.br/#!/departamentos/matematica/eventos2341/ermac/cadesnos-de-trabalhos-completos-e-resumos/>. Acesso em: 5 nov. 2017.

DORIGO, M.; STÜTZLE, T. **Ant colony optimization**. Cambridge: MIT, 2004.

GABRIEL, K. R. The biplot graphic display of matrices with application to principal component analysis. **Biometrika**, v. 58, n. 3, p. 453-467, 1971.

GODINHO, E.; BRONDINO, N. C. M. Decomposição em valores singulares de um conjunto de séries temporais envolvendo variáveis econômicas e sociodemográficas brasileiras. In:

ENCONTRO REGIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL, 4., 2017, Bauru. **Caderno de trabalhos completos e resumos**. Bauru: Unesp, Faculdade de Ciências, 2017. p. 462-464. Disponível em:

<http://www.fc.unesp.br/#!/departamentos/matematica/eventos2341/ermac/cadesnos-de-trabalhos-completos-e-resumos/>. Acesso em: 5 nov. 2017.

HOFFMAN, K.; KUNZE, R. **Linear algebra**. 2. ed. New Jersey: Prentice Hall, 1971.

INSTITUTO DE PESQUISA ECONÔMICA APLICADA. **ipeadata**. [S.l.], [2016?].

Disponível em: <<http://www.ipeadata.gov.br/Default.aspx>>. Acesso em: 13 nov. 2017.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Sinopses estatísticas**. Brasília, DF, 2015. Disponível em:

<<http://portal.inep.gov.br/web/guest/sinopses-estatisticas>>. Acesso em: 5 nov. 2016.

KNOEMA. **World Data Atlas: Brazil**. [S.l.], [2016?]. Disponível em:

<<https://knoema.com/atlas/Brazil>>. Acesso em: 13 nov. 2017.

KRUSKAL, J. B. Multilinear methods. In: STATISTICAL data analysis.

Providence: American Mathematical Society, 1983. (Proceedings of Symposia in Applied Mathematics, v. 28).

TABAKHI, S.; MORADI, P.; AKHLAGHIAN, F. An unsupervised feature selection algorithm based on ant colony optimization. **Engineering Applications of Artificial Intelligence**, v. 32, p. 112-123, 2014.

Artigo recebido em jun. 2017 e aceito em nov. 2017.