ISSN 0101-7438

OTIMIZAÇÃO DE ENTROPIA: IMPLEMENTAÇÃO COMPUTACIONAL DOS PRINCÍPIOS MAXENT E MINXENT

Rogério Silva de Mattos *
Departamento de Análise Econômica
Universidade Federal de Juiz de Fora
Juiz de Fora – MG
E-mail: rmattos@zaz.com.br

Álvaro Veiga
Departamento de Engenharia Elétrica
PUC–Rio
Rio de Janeiro – RJ
E-mail: alvf@ele.puc-rio.br

* *Corresponding author* /autor para quem as correspondências devem ser encaminhadas

Recebido em 05/2001, aceito em 05/2002 após 1 revisão

Resumo

Os princípios de otimização de entropia MaxEnt de Jaynes (1957a,b) e MinxEnt de Kullback (1959) encontram aplicações em várias áreas de investigação científica. Ambos envolvem a otimização condicionada de medidas de entropia que são funções intrinsecamente não-lineares de probabilidades. Como constituem problemas de programação não-linear, suas soluções demandam algoritmos de busca iterativa e, além disso, as condições de não-negatividade e de soma um para as probabilidades restringem de modo particular o espaço de soluções. O artigo apresenta em detalhe (com a ajuda de dois fluxogramas) uma implementação computacional eficiente desses dois princípios no caso de restrições lineares com verificação prévia de existência de solução dos problemas de otimização. Os autores também disponibilizam rotinas de fácil uso desenvolvidas em linguagem **MatLab**®.

Palavras-chave: otimização da entropia; medida de Shannon; medida de Kullback.

Abstract

The entropy optimization principles MaxEnt of Jaynes (1957a,b) and MinxEnt of Kullback (1959) can be applied in a variety of scientific fields. Both involve the constrained optimization of entropy measures, which are intrinsically non-linear functions of probabilities. Since each is a non-linear programming problem, their solution depend on iterative search algorithms, and, in addition, the constraints that probabilities are non-negative and sum up to one restrict in a particular way the solution space. The paper presents in detail (with the aid of two flowcharts) a computer efficient implementation of those two principles in the linearly constrained case that makes a prior check for the existence of solution to the optimization problems. The authors also make available easy-to-use **MatLab**® codes.

Keywords: entropy optimization; Shannon's measure; Kullback's measure.

1. Introdução

O conceito de entropia foi introduzido na Ciência há mais de 150 anos, mas somente a partir de meados do Século XX é que difundiram-se suas aplicações por diversas áreas do conhecimento. Na raiz deste movimento, estiveram os trabalhos de Shannon (1948), que introduziu um conceito de entropia em teoria da informação e uma medida para quantificá-la, e os estudos de Jaynes (1957a,b) e Kullback (1959), que propuseram *princípios de otimização da entropia* segundo formulações distintas. Atualmente, diferentes áreas, como termodinâmica, probabilidade, estatística, pesquisa operacional, reconhecimento de padrões, economia, finanças, marketing, planejamento urbano e de transportes, dentre outras, vêm usando e desenvolvendo princípios de otimização da entropia (para diversos exemplos de aplicação em várias áreas, ver os livros de Kapur & Kesavan, 1992; Golan Judge & Miller, 1996; e Fang, Rajasekera & Tsao, 1997).

O conceito de entropia de Shannon refere-se à incerteza de uma distribuição de probabilidade e a medida que propôs destinava-se a quantificar essa incerteza. Formalmente, o princípio de Jaynes envolve a busca pela distribuição de probabilidade que maximiza a medida de Shannon, dado um conjunto de restrições lineares. Estas restrições informam características da distribuição procurada, como, por exemplo, sua média e sua variância. O princípio de Kullback, por sua vez, envolve a busca pela distribuição de probabilidade mais próxima de uma outra distribuição *a priori*, através da minimização de uma medida de divergência entre ambas, dado o mesmo conjunto de restrições. Tanto a medida de Shannon como a de Kullback são funções intrinsecamente não-lineares de probabilidades. Assim, os princípios de Jaynes e Kullback reduzem-se a problemas de programação não-linear cuja solução demanda um algoritmo de busca iterativa.

Em livro recente, Fang, Rajasekera & Tsao (1997, p. ix) apontam que, como esses (e outros) princípios de otimização de entropia foram repetidamente usados em várias áreas, muitos métodos para sua implementação (solução) foram sugeridos e utilizados. Entretanto, esses métodos carecem de uma formulação matemática mais rigorosa, que possa prover suporte aos praticantes interessados tanto no desenvolvimento de implementações computacionais eficientes como na adequada utilização dos princípios em aplicações. Por outro lado, embora esses mesmos autores apresentem um algoritmo para implementação dos princípios de Jaynes e Kullback, eles o fazem para o caso generalizado da medida de Kullback em que as distribuições são de freqüências e não de probabilidades (ver Fang, Rajasekera & Tsao (1997), pp. 51-56). No último caso, as restrições de não negatividade e de soma um para as probabilidades restringem de modo particular o conjunto de soluções viáveis, ao imporem limites inferiores e superiores para os coeficientes lineares das equações de restrição.

Este artigo objetiva apresentar em detalhe, com o auxílio de dois *fluxogramas*, a construção de um algoritmo para implementação computacional daqueles dois princípios de otimização da entropia no caso de distribuições de probabilidade discretas. O algoritmo aqui apresentado foi desenvolvido seguindo abordagem semelhante à proposta por Agmon, Alhassid & Levine (1979). No entanto, as descrições de ambos os princípios, do algoritmo e a caracterização das condições para existência de solução são feitas aqui de forma mais clara e direta, facilitando a implementação da metodologia através de diferentes linguagens computacionais. Como um subproduto do artigo, os autores disponibilizam (mediante solicitação por *e-mail*) funções escritas por eles em linguagem **MatLab**® que implementam os dois princípios e são de fácil uso. Pretende-se, com isso, estimular a utilização de técnicas de otimização da entropia no contexto brasileiro. O público-alvo são pesquisadores e profissionais que utilizam

intensivamente métodos quantitativos em suas áreas de atuação, em particular aqueles que trabalham com metodologias de P.O.

O artigo está organizado da seguinte forma. Na seção 2, o conceito de entropia na teoria da informação é revisto e breves considerações são tecidas sobre sua extensão a outras áreas onde o interesse se centra sobre distribuições de proporção e não de probabilidade. Nas seções 3 a 5, são discutidas as medidas de entropia de Shanon e de Kullback, bem como o formalismo dos princípios de Jaynes e Kullback. Na seção 6, a implementação computacional propriamente dita é descrita em detalhe e também são discutidas as condições para existência de solução factível dos problemas. Na Seção 7, um algoritmo de busca iterativa baseado no método de Newton é sugerido. Na Seção 8, são apresentadas as funções em linguagem **MatLab**® desenvolvidas pelos autores para implementar o algoritmo proposto. Na seção 9, é apresentado um exemplo de aplicação do algoritmo em um problema de determinação do número de viagens, típico de planejamento de transportes. Finalmente, na seção 10, são tecidos alguns comentários conclusivos. Há também dois apêndices: o primeiro contém um fluxograma de todo o processamento do algoritmo proposto para implementar os dois princípios de otimização da entropia; o segundo descreve, com a ajuda de outro fluxograma, a implementação da Fase 1 do Método Simplex, usada para testar a existência de solução factível dos problemas de otimização considerados.

2. Entropia

Originária de estudos de termodinâmica, onde foi introduzida para caracterizar o grau de desordem de um sistema, a noção de entropia já foi objeto de muitas controvérsias e distintas formulações. O conceito de entropia adotado por Shannon (1948) foi responsável por aplicações de relevo em diversos campos de investigação científica, embora seu trabalho tenha se destacado mais pela medida de quantificação de entropia que propôs, cujas propriedades despertaram o interesse em outras áreas. Nesta seção, é feita uma breve apresentação do conceito de entropia na teoria da informação, visando preparar o terreno para, nas duas próximas seções, falar-se de medidas para sua quantificação.

Segundo Kapur & Kesavan (1992), o conceito de Shannon poderia ser chamado de *entropia na teoria da informação* e refere-se à incerteza de uma distribuição de probabilidade. Na verdade, o conceito de incerteza é mais geral, podendo-se falar, basicamente, em três tipos de incerteza: a *incerteza determinística*, em que não são conhecidos os estados que um sistema pode assumir; a *incerteza entrópica*, em que são conhecidos os estados possíveis, mas não as chances de ocorrência de cada um deles; e a *incerteza probabilística*, em que são conhecidos não só os estados possíveis mas também a distribuição de probabilidade para eles (todavia, não se pode determinar qual irá ocorrer com certeza).

A entropia na teoria da informação corresponde à *incerteza probabilística* associada a uma distribuição de probabilidade. Cada distribuição reflete um certo grau de incerteza e diferentes graus de incerteza estão associados a diferentes distribuições (embora diferentes distribuições possam refletir o mesmo grau de incerteza). De um modo geral, quanto mais "espalhada" a distribuição de probabilidade, maior incerteza ela irá refletir. Por exemplo, se alguém lança um dado de seis faces, sem saber se ele é viciado ou não, a probabilidade mais razoável a ser atribuída a cada resultado possível é 1/6, ou seja, representar a incerteza usando a distribuição uniforme. Esta atitude segue o conhecido *princípio da razão insuficiente de Laplace*, onde atribuir chances iguais aos eventos possíveis é a maneira mais

razoável de alguém refletir sua ignorância (e sua incerteza) quanto às chances de ocorrência de cada evento. Por outro lado, provendo-se a informação de que o dado é viciado e que ele dá números maiores (menores) que a média (=3,5, no caso uniforme) mais freqüentemente, então a pessoa naturalmente irá assumir uma distribuição alternativa à uniforme para expressar sua incerteza. A Figura 1 ilustra graficamente essa situação no caso de distribuições contínuas de probabilidade.

Uma importante característica da entropia na teoria da informação, ou incerteza probabilística, é que ela está diretamente associada ao grau de similaridade entre as probabilidades de uma distribuição. Segundo Kapur & Kesavan (1992), este aspecto confere uma importante versatilidade à essa noção de entropia que lhe permite ser estendida e adaptada, enquanto conceito, à várias outras disciplinas. Entretanto, esta extensão/adaptação já foi questionada na literatura (Georgescu-Roegen, 1971) por não estar em consonância com a noção original de entropia em termodinâmica e nem com a própria noção de entropia na teoria da informação.

Sem pretender aprofundar essa discussão, o fato é que a medida introduzida por Shannon para quantificar entropia em teoria da informação também se presta a quantificar diversos conceitos de interesse em outras disciplinas. Se, ao invés de distribuição de probabilidades, trata-se de distribuição de *proporções*, como a distribuição intersetorial do produto industrial ou a distribuição espacial da ocupação residencial, é possível utilizar-se de modo interessante as medidas de entropia desenvolvidas em teoria da informação. Sob esta perspectiva, elas servem para medir igualdade, espalhamento, similaridade, diversidade, complexidade de sistemas e outros conceitos que aparecem em diversas áreas do conhecimento, ainda que tais conceitos não tenham uma relação direta com alguma noção clássica de entropia.

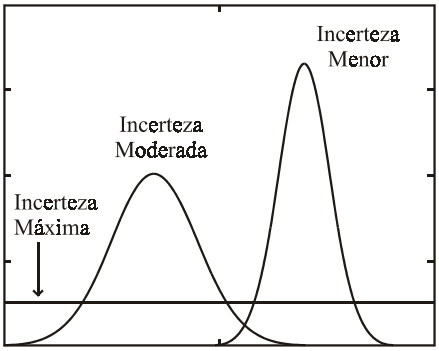


Figura 1 – Representação de incerteza com distribuições de probabilidade.

3. Medida de Entropia

Shannon (1948) derivou uma medida para quantificar o grau de incerteza de uma distribuição de probabilidade. Denominando S a medida de entropia de Shannon, sua expressão formal para distribuições discretas de probabilidade é dada por:

$$S(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln p_i \qquad (1)$$

onde $\mathbf{p}^T = [p_1,\ldots, p_N]$ é a distribuição de probabilidade (o sobrescrito "T" representa transposição matricial). Esta medida é sempre não-negativa e assume seu valor máximo $S(\mathbf{p}) = \ln N$ quando $\mathbf{p}^T = [1/N,\ldots, 1/N]$ (i.e., a distribuição uniforme). Qualquer outra distribuição faz S ser menor do que $\ln N$. Seu mínimo ocorre em $S(\mathbf{p}) = 0$, situação onde há ausência de incerteza, quando então $\mathbf{p}$ é degenerada em uma das p_is (i.e., uma $p_i = 1$ e as demais iguais a zero).

Usando um método axiomático, Shannon derivou essa medida de modo que ela refletisse certas características desejadas. Posteriormente, outros matemáticos demonstraram que ela atende a outras propriedades de interesse adicional (Kapur & Kesavan, 1992, pp. 23-35). As propriedades de S mais relevantes para os fins deste trabalho são:

S1. $S(p_1, p_2,\ldots, p_n)$ é uma função duas vezes diferenciável de $p_1, p_2,\ldots, p_N$.

S2. $S(p_1, p_2,\ldots, p_n)$ é simétrica em relação à permutação de $p_1, p_2,\ldots, p_N$.

S3. $S(1/N, 1/N,\ldots, 1/N)$ é uma função monotonamente crescente de N.

S4. S é uma função estritamente côncava de $p_1, p_2,\ldots, p_N$.

S1 é importante por permitir a aplicação de técnicas para maximização de funções diferenciáveis. S2 significa que as p_is podem ter sua ordem invertida no cômputo de S que esta não se altera. S3 significa que a entropia da distribuição uniforme (máxima entropia possível) cresce quanto maior for o número de resultados possíveis N. Por último, S4 é de especial relevância, como se verá adiante, pois garante que S tenha um único máximo (global), mesmo quando sujeita a restrições lineares.

As propriedades de S permitem que ela também seja aplicada em diversos outros contextos. Quando, ao invés de probabilidades, as p_is representarem proporções, isto é: $p_i = v_i / \Sigma_{i=1}^{N} v_i$, onde v_i = valor da i-ésima parcela não negativa de uma soma, a medida de Shannon também pode ser aplicada, o que viabiliza aplicações em outras disciplinas. Por exemplo, S pode ser usada para medir o grau de igualdade (ou desigualdade) da distribuição de renda entre várias classes sociais, ou o grau de espalhamento das ocupações residenciais dentro de uma cidade.

4. Medida de Entropia Cruzada

Kullback (1959) introduziu outra importante medida em teoria da informação. A *medida de entropia cruzada* de Kullback é um caso particular de medidas de divergência direcionada e serve para medir a diferença entre duas distribuições de probabilidade. Sejam $\mathbf{p}^T = [p_1,\ldots, p_N]$ e $\mathbf{q}^T = [q_1,\ldots, q_N]$ duas distribuições quaisquer, e seja K a medida de Kullback. No caso de distribuições discretas de probabilidade, K é definida como:

$$K(\mathbf{p} : \mathbf{q}) = \sum_{i=1}^{N} p_i \ln \frac{p_i}{q_i}. \qquad (2)$$

Embora K não seja uma medida de incerteza de uma distribuição de probabilidade, ela serve aos mesmos propósitos que a medida de Shannon. A expressão (2) indica que K é uma medida de divergência ou diferença entre $\mathbf{p}$ e $\mathbf{q}$. É fácil verificar que $K(\mathbf{p}:\mathbf{q}) \neq K(\mathbf{q}:\mathbf{p})$, daí ela ser uma medida de divergência *direcionada*. Quanto maior a diferença/divergência entre $\mathbf{p}$ e $\mathbf{q}$, maior será o valor de K. Dado o valor de N, seu máximo é atingido quando $\mathbf{p}$ é degenerada e $\mathbf{q}$ é a distribuição uniforme, situação em que $K = \ln N$ (ver expressão (3) abaixo). Quanto mais parecidas forem $\mathbf{p}$ e $\mathbf{q}$, menor será K; no limite, se $\mathbf{p} = \mathbf{q}$, então $K = 0$.

Quando **q** é a distribuição uniforme, isto é, $\mathbf{q}^T = \mathbf{u}^T = [1/N,\ldots, 1/N]$, então K também pode ser usada para medir incerteza ou entropia, pois neste caso:

$$K(\mathbf{p} : \mathbf{u}) = \ln N - (-\sum_{i=1}^{N} p_i \ln p_i) = \ln N - S(\mathbf{p}) . \qquad (3)$$

Sendo $\ln N$ a entropia da distribuição uniforme (constante para um dado N), os graus de entropia de diferentes distribuições podem ser medidos e comparados entre si com base em suas divergências (medidas com a expressão (3)), em relação à distribuição uniforme. Adicionalmente, K também serve para indicar o grau de similaridade entre as entropias de duas distribuições, ainda que nenhuma delas seja a uniforme. Neste último caso, a medida de Kullback se presta a comparar diferentes distribuições com uma distribuição fixa qualquer.

K também apresenta várias propriedades atraentes. Dentre elas, destacam-se:

K1. $K(\mathbf{p}:\mathbf{q})$ é uma função duas vezes diferenciável de $p_1, p_2,\ldots, p_N$;

K2. $K(\mathbf{p}:\mathbf{q})$ é simétrica em relação à permutação dos pares $(p_1, q_1),\ldots, (p_N, q_N)$;

K3. K é uma função estritamente convexa de $p_1, p_2,\ldots, p_N$.

K4. $K(\mathbf{p}:\mathbf{q}) \geq 0$ (não negatividade);

K5. $K(\mathbf{p}:\mathbf{q}) = 0$ se e apenas se $\mathbf{p} = \mathbf{q}$;

As propriedades K1, K2 e K3 possuem implicações para K análogas às que S1, S2 e S4 apresentaram, respectivamente, para a medida de Shannon. As propriedades K4 e K5 são duas características de distâncias métricas (no entanto, K não é uma métrica pois não atende às propriedades de simetria e de desigualdade do triângulo que toda medida de distância tem de apresentar para ser uma métrica). Da mesma forma que a medida de Shannon, a medida de Kullback se presta a estudos de entropia em que as distribuições se refiram a proporções e não a probabilidades.

5. Otimização da Entropia

Em teoria da informação, maximizar a entropia significa determinar a distribuição de probabilidade que represente o máximo de incerteza, dadas certas restrições. Ou seja, significa determinar a distribuição com maior grau de similaridade entre suas probabilidades, ou que seja mais parecida com a uniforme e diferindo dela apenas devido às restrições. Estas, por sua vez, refletem algum tipo de informação prévia sobre o fenômeno probabilístico de interesse, como, por exemplo, a média e a variância da distribuição que se quer determinar.

O princípio de maximizar a entropia (MaxEnt) através da medida de Shannon, dado um conjunto de restrições, foi introduzido por Jaynes (1957a,b). Posteriormente, Kullback (1959) introduziu o princípio de minimização da entropia cruzada (MinxEnt), através do qual se procura minimizar a medida K, de divergência direcionada entre duas distribuições **p** e **q**, também sujeito a um conjunto de restrições. O princípio MinxEnt de Kullback generaliza o MaxEnt de Jaynes, pois permite que se incorpore, através de **q**, alguma informação *a priori* sobre a forma da distribuição de probabilidade procurada ao se otimizar a entropia. Quando **q** é a distribuição uniforme, o princípio MinxEnt se reduz ao princípio MaxEnt. Quando **q** é uma outra distribuição qualquer, o princípio MinxEnt envolve encontrar a distribuição **p** mais parecida com *a priori* **q**, ou a distribuição cuja entropia é a mais próxima da de **q**.

Nas subseções 5.1 e 5.2, o formalismo característico dos princípios MaxEnt e MinxEnt são introduzidos, buscando-se salientar as implicações para a implementação computacional de ambos que será apresentada na seção 6.

5.1 MaxEnt

A aplicação do princípio MaxEnt pressupõe expressá-lo formalmente como um problema de otimização (doravante problema MaxEnt), da seguinte forma:

$$\begin{aligned} \underset{\mathbf{p}}{Max} \quad & S = -\sum_{i=1}^{N} p_i \ln p_i \\ s.a. \quad & \begin{cases} \sum_{i=1}^{N} p_i = 1 \\ \sum_{i=1}^{N} p_i g_{ri}(x_i) = a_r \quad r = 1,...,M \\ p_i \geq 0 \end{cases} \end{aligned} \tag{4}$$

(onde *s.a.*. significa "sujeito a"). As funções $g_{ri}(x_i)$, $r = 1,\ldots, M$, são funções dos resultados possíveis x_i, $i = 1,\ldots, N$. Note-se que o conjunto de restrições é formado por $M + 1$ restrições lineares e N restrições de não-negatividade, constituindo um típico problema de programação não-linear. Porém, a presença do termo $\ln p_i$ em S implica que esta medida não está definida para valores negativos das p_is, de modo que as N restrições de não-negatividade são não operantes (embora $ln0$ não seja definido, a medida S, no entanto, está definida para valores nulos das p_is porque, quando $x \to 0$, lim $(x.lnx) = 0$). Isto simplifica o problema, permitindo que se aplique diretamente o método dos multiplicadores de Lagrange para otimização de funções não lineares com restrições de igualdade apenas.

A primeira das restrições lineares $\Sigma_{i=1}^{N} p_i = 1$ é chamada de *restrição natural*, porque reflete a necessidade de que toda distribuição de probabilidade some um. As M restrições $\Sigma_{i=1}^{N} p_i g_{ri}(x_i) = a_r$ são denominadas de *restrições de consistência*. Nas aplicações em probabilidade, cada a_r geralmente representa o momento de ordem r (o que implica fazer $g_{ri}(x_i) = x_i^r$ ou $g_{ri}(x_i) = (x_i - \mu)^r$, com μ representando a média da distribuição) ou então um momento *característico* da distribuição de probabilidade (sobre momentos característicos, ver por exemplo, Kapur & Kesavan, 1992, p. 359). Em várias aplicações onde as p_is são tratadas como *proporções*, a_r, x_i e g_{ri} representam outro tipo de informação conhecida sobre o fenômeno de interesse (ver o exemplo da Seção 9, e também diversos outros nos livros de Kapur & Kesavan, 1992, e Fang & Tsao, 1997).

Usando-se o método do multiplicador de Lagrange, o problema MaxEnt (4) pode ser posto na seguinte forma irrestrita:

$$\underset{\mathbf{p},\lambda_0,\mathbf{z}}{Max}\, L_s = -\sum_{i=1}^{N} p_i \ln p_i + (\lambda_0 - 1)\left(\sum_{i=1}^{N} p_i - 1\right) + \sum_{r=1}^{M} \lambda_r \left(\sum_{i=1}^{N} p_i g_{ri}(x_i) - a_r\right) \tag{5}$$

onde $(\lambda_0 – 1)$ e o vetor $\mathbf{z}^{\mathrm{T}} = [\lambda_1,\ldots,\lambda_M]$ representam os $M + 1$ multiplicadores de Lagrange associados às $M + 1$ restrições. O multiplicador $\lambda_0 - 1$, ao invés de simplesmente λ_0, foi usado com a primeira restrição por conveniência matemática (Kapur & Kesavan, 1992, pp. 43-44),

uma vez que permite simplificar as expressões apresentadas adiante. Aplicando a condição de primeira ordem para um extremo local, dada por $\nabla L_s(\mathbf{p},\lambda_0,\mathbf{z}) = \mathbf{0}$ (i.e., gradiente nulo ou conjunto das derivadas parciais iguais a zero), e manipulando algebricamente o sistema resultante, são obtidas as seguintes expressões:

$$p_i = \frac{\exp\left(-\sum_{r=1}^{M}\lambda_r g_{ri}(x_i)\right)}{\sum_{i=1}^{N}\exp\left(-\sum_{r=1}^{M}\lambda_r g_{ri}(x_i)\right)} \qquad i = 1,\ldots, N \qquad (6)$$

$$a_r = \frac{\sum_{i=1}^{N} g_{ri}(x_i)\exp\left(-\sum_{r=1}^{M}\lambda_r g_{ri}(x_i)\right)}{\sum_{i=1}^{N}\exp\left(-\sum_{r=1}^{M}\lambda_r g_{ri}(x_i)\right)} \qquad r = 1,\ldots, M\,. \qquad (7)$$

A expressão (6) caracteriza a chamada *distribuição de probabilidade MaxEnt*. O sistema de $M + N$ equações em $M + N$ incógnitas formado por (6) e (7) apresenta uma relação intrinsecamente não-linear entre as probabilidades p_i e os multiplicadores de Lagrange λ_r, de modo que não é possível derivar uma solução analítica para p_i e λ_r, simultaneamente, em função apenas dos elementos conhecidos a_r e g_{ri}. Logo, a solução do sistema tem de ser obtida usando-se um algoritmo de busca iterativa. Note-se que um dos multiplicadores de Lagrange, λ_0, foi eliminado na manipulação algébrica (e, logo, do sistema de equações (6) e (7)), mas é simples verificar que ele pode ser obtido a partir dos demais multiplicadores segundo $\lambda_0 = \ln[\Sigma_{i=1}^{N}\exp(-\Sigma_{r=1}^{M}\lambda_r g_{ri}(x_i))]$.

Além disso, o problema (5) pode ser colocado ainda de uma outra forma. Substituindo os p_is do Lagrangeano L_s pelas expressões em (6) e realizando uma pequena manipulação algébrica, obtém-se:

$$L_s^*(\mathbf{z}) = \ln\left[\sum_{i=1}^{N}\exp\left(-\sum_{r=1}^{M}\lambda_r g_{ri}(x_i)\right)\right] + \sum_{r=1}^{M}\lambda_r a_r\,. \qquad (8)$$

Esta nova expressão apresenta como argumentos apenas os multiplicadores de Lagrange em $\mathbf{z}^T = [\lambda_1,\ldots,\lambda_M]$, embora seja intrinsecamente não-linear em relação a eles. Isto permite uma formulação dual irrestrita para o problema MaxEnt que consiste em *Min* $L_s^*(\mathbf{z})$. É possível mostrar que L_s^* é uma função estritamente convexa dos multiplicadores de Lagrange $\mathbf{z}$ (Golan, Judge & Miller, 1996, pp. 25-26), o que assegura que o problema dual apresenta uma única solução (se houver solução). Posto deste modo, o problema tem sua dimensão reduzida, pois agora o sistema a ser resolvido é composto apenas por (7). Ao invés de se procurar iterativamente as $N + M$ variáveis em $(\mathbf{p}^T,\mathbf{z}^T)$, basta a procura dos M multiplicadores de Lagrange em $\mathbf{z}$. Uma vez determinada a solução ótima para estes, automaticamente ficam determinados os N valores para as p_is, através da relação (6).

5.2 MinxEnt

Analogamente, o princípio MinxEnt tem de ser formalizado como um problema de otimização (problema MinxEnt), da seguinte forma:

$$\begin{aligned} \underset{\mathbf{p}}{Min} \quad & K = \sum_{i=1}^{N} p_i \ln \frac{p_i}{q_i} \\ s.a. \quad & \begin{cases} \sum_{i=1}^{N} p_i = 1 \\ \sum_{i=1}^{N} p_i g_{ri}(x_i) = a_r \quad r = 1,...,M \\ p_i \geq 0. \end{cases} \end{aligned} \tag{9}$$

O que muda em relação ao problema anterior é que agora busca-se minimizar a Medida de Kullback. Como mencionado antes, esta mede a distância entre a distribuição definida pelas p_is em relação a uma distribuição *a priori* definida pelas q_is. O sistema de restrições lineares é o mesmo e tem o mesmo papel que no problema MaxEnt. Escrevendo $K = \Sigma_{i=1}^{N} p_i(\ln p_i - \ln q_i)$, verifica-se que esta medida também não está definida para valores negativos das p_is (e das q_is), de modo que a restrição $p_i \geq 0$ também é não operante aqui. Se $q_i = 1/N$ (distribuição uniforme), (9) reduz-se ao problema MaxEnt, conforme se pode perceber pela expressão para K dada em (3).

Usando-se o método dos multiplicadores de Lagrange, o problema MinxEnt também pode ser posto na seguinte forma irrestrita:

$$\begin{aligned} \underset{\mathbf{p},\lambda_0,\mathbf{z}}{Max} \; L_k = & \sum_{i=1}^{N} p_i \ln \frac{p_i}{q_i} + (\lambda_0 - 1)\left(\sum_{i=1}^{N} p_i - 1\right) \\ & + \sum_{r=1}^{M} \lambda_r \left(\sum_{i=1}^{N} p_i g_{ri}(x_i) - a_r\right) \end{aligned} \tag{10}$$

onde, novamente, $\lambda_0 - 1$ e $\mathbf{z}$ representam multiplicadores de Lagrange. De forma análoga, aplicando a condição de primeira ordem para um extremo local $\nabla L_s(\mathbf{p},\lambda_0,\mathbf{z}) = \mathbf{0}$ (gradiente nulo ou conjunto das derivadas parciais iguais a zero), e manipulando algebricamente o sistema resultante, são obtidas as seguintes expressões:

$$p_i = \frac{q_i \exp\left(-\sum_{r=1}^{M} \lambda_r g_{ri}(x_i)\right)}{\sum_{i=1}^{N} q_i \exp\left(-\sum_{r=1}^{M} \lambda_r g_{ri}(x_i)\right)} \tag{11}$$

$$a_r = \frac{\sum_{i=1}^{N} g_{ri}(x_i)\, q_i \exp\left(-\sum_{r=1}^{M} \lambda_r g_{ri}(x_i)\right)}{\sum_{i=1}^{N} q_i \exp\left(-\sum_{r=1}^{M} \lambda_r g_{ri}(x_i)\right)} \tag{12}$$

A expressão (11) caracteriza a *distribuição de probabilidade MinxEnt*. O sistema (11)–(12) difere do sistema (6)–(7) apenas pela presença do termo q_i multiplicando o expoente nos numeradores e denominadores das expressões para p_i e a_r. Aqui, da mesma forma, cai-se em um sistema de equações intrinsecamente não-lineares em p_i e λ_r, de modo que é impossível derivar soluções analíticas para estas incógnitas em função dos elementos conhecidos a_r e g_{ri}. Substituindo a expressão para p_i em (11) no Lagrangeano L_k em (10) e manipulando, obtém-se:

$$L_k^*(\mathbf{z}) = -\ln\left[\sum_{i=1}^{N} q_i \exp\left(-\sum_{r=1}^{M} \lambda_r g_{ri}(x_i)\right)\right] - \sum_{r=1}^{M} \lambda_r a_r \ . \qquad (13)$$

Assim, o problema (9) admite uma formulação dual irrestrita, que consiste em *Max* L_k*(**z**). É possível mostrar que L_k* é uma função estritamente côncava dos multiplicadores de Lagrange **z** (Kapur & Kesavan, 1992, pp. 167-168), o que garante que o problema MinxEnt dual também tem uma única solução (se houver solução). Aqui também fica simplificado o problema ao se reduzir a dimensão $M + N$ do espaço de busca para M. O sistema de equações não lineares a ser resolvido no problema dual é formado apenas por (12), que envolve apenas M equações nos M multiplicadores de Lagrange **z**. Achando-se estes, automaticamente determinam-se as N probabilidades p_is por (11).

Os problemas duais para MaxEnt e MinxEnt admitem formulações análogas quando se redefine o último de modo que se torne também um problema de minimização, isto é:

MinxEnt Dual: $\underset{\mathbf{z}}{Min} \quad -L_k^*(\mathbf{z})$.

A vantagem disso é que é possível desenvolver um único tipo de programa, voltado para um problema de minimização, que pode ser usado em ambos os casos MaxEnt e MinxEnt. O restante deste artigo apresenta a implementação computacional de um algoritmo para encontrar as soluções de ambos os problemas.

6. Implementação Computacional

Conforme visto na seção anterior, encontrar a distribuição MaxEnt envolve resolver um sistema de M equações intrinsecamente não-lineares em M incógnitas. Para se achar a solução desse sistema, é necessário algum método de procura iterativa pela solução, o que requer implementação computacional mesmo para pequenas dimensões do problema. Antes, porém, de se discutir a montagem de um tal algoritmo, é preciso examinar as condições para existência de soluções factíveis dos problemas MaxEnt e MinxEnt. Embora os sistemas alvos sejam (7) e (12), que se apresentam não-lineares em **z**, o ponto de partida será o sistema de restrições lineares nas p_is, porque estas é que são as incógnitas de interesse e porque é mais fácil derivar as condições factíveis para um sistema linear.

6.1 O Sistema de Restrições

Considere-se, inicialmente, o sistema de restrições comum aos princípios MaxEnt e MinxEnt, apresentado no Quadro 6.1 Ele é composto por uma restrição natural de soma unitária das probabilidades e M restrições de consistência, formando um conjunto de $M + 1$ equações lineares. Existe um total de N incógnitas, constituídas pelas N p_is. As restrições de não negatividade $p_1 \geq 0$, $p_2 \geq 0$, …, $p_N \geq 0$ não aparecem no Quadro 6.1, pois são não operantes, como discutido anteriormente.

Conforme a teoria de sistemas lineares, o sistema do Quadro 6.1 admite as seguintes possibilidades:

- se $M + 1 \geq N$, então: *ou* existe um número infinito de soluções, *ou* existe uma única solução, *ou* não existe solução factível;
- se $M + 1 < N$, então: *ou* existe um número infinito de soluções, *ou* não existe solução factível.

Quadro 6.1 – O sistema de restrições operantes dos princípios MaxEnt e MinxEnt

	N incógnitas	
Restrição natural:	$p_1 + p_2 + \ldots + p_N = 1$	$M + 1$ equações
Restrições de consistência	$p_1 g_{11} + p_2 g_{12} + \ldots + p_N g_{1N} = a_1$ . . . $p_1 g_{M1} + p_2 g_{M2} + \ldots + p_N g_{MN} = a_1$	

Se não existir solução factível, não é possível aplicar o princípio de Jaynes nem o de Kullback e não existe uma distribuição MaxEnt/MinxEnt. Se, por outro lado, existir solução factível, então pode-se afirmar que:

a) quando $M + 1 \geq N$, com pelo menos N restrições linearmente independentes, maximizar a entropia é sinônimo de resolver o sistema de equações lineares formado pelas restrições (se $M + 1 = N$) ou por um subgrupo das restrições (se $M + 1 > N$);
b) quando $M + 1 < N$, maximizar a entropia envolve escolher uma única solução dentre um número infinito de soluções.

A afirmação a) significa que não é necessário maximizar a medida de Shannon ou minimizar a de Kullback se o sistema de restrições lineares admite no máximo uma única solução factível. Neste caso, podem ser ignoradas S ou K e, simplesmente, obter-se a distribuição MaxEnt/MinxEnt usando-se procedimentos padronizados para resolução de sistemas de equações lineares.

A afirmação b), por sua vez, significa que encontrar a distribuição MaxEnt/MinxEnt envolve escolher uma distribuição de probabilidade entre infinitas distribuições alternativas que atendem ao sistema de restrições. Otimizar a medida de Shannon ou a de Kullback, portanto, só faz sentido neste último caso.

6.2 Condições para existência de solução factível

Como otimizar a entropia só faz sentido se $M + 1 < N$, este é o caso relevante a ser considerado a partir daqui. Entretanto, esta situação admite também a possibilidade de não existir solução factível para o sistema de restrições. Se for este o caso, não faz sentido aplicar um algoritmo de busca. É importante, então, estabelecer quais as condições precisas para a existência de solução do sistema de restrições lineares quando $M + 1 < N$.

Consideremos a seguinte representação matricial do sistema de restrições:

$$\begin{bmatrix} \mathbf{1}^{\mathrm{T}} \\ \mathbf{G} \end{bmatrix} \mathbf{p} = \begin{bmatrix} 1 \\ \mathbf{a} \end{bmatrix} \qquad (14)$$

onde $\mathbf{l}$ = vetor de dimensão N com todos os elementos iguais a 1, $\mathbf{G} = \{g_{ri}(x_i)\}$ é a matriz de coeficientes do subsistema de restrições de consistência, $\mathbf{p}^{\mathrm{T}} = [p_1,\ldots,p_N]$ e $\mathbf{a}^{\mathrm{T}} = [a_1,\ldots,a_M]$. As condições para existência de solução factível para o sistema (14) quando $M + 1 < N$ são as seguintes:

a) a matriz $[\mathbf{1}\ \mathbf{G}^{\mathrm{T}}]^{\mathrm{T}}$ tem de apresentar posto cheio ($= M + 1$), o que implica que as restrições têm de ser linearmente independentes entre si;

b) Se $M = 1$, isto é, se há uma única restrição de consistência, de modo que $\mathbf{G} = [g_{11},\ldots, g_{1N}]$, então:

$$min\{\mathbf{G}\} < a_1 < max\{\mathbf{G}\}. \qquad (15)$$

A condição a) é óbvia. A condição b) tem uma justificativa intuitiva. Como o coeficiente a_1 é uma média ponderada (pelas probabilidades p_is) das g_{1i}s, ele não pode ser inferior à menor nem superior à maior delas.

Os sinais de desigualdade estrita em (15) são relevantes porque:

- se $a_1 \to min\{\mathbf{G}\}$, então $\lambda_1 \to \infty$ e a distribuição MaxEnt/MinxEnt degenera-se em $\mathbf{p}^{\mathrm{T}} = (1,0,\ldots,0)$ com $S = 0$ (não há incerteza com um único resultado possível);
- se $a_1 \to max\{\mathbf{G}\}$, então $\lambda_1 \to -\infty$ e a distribuição MaxEnt/MinxEnt degenera-se em $\mathbf{p}^{\mathrm{T}} = (0,\ldots,1)$ com $S = 0$;
- se $a_1 = min\{\mathbf{G}\} = max\{\mathbf{G}\}$, então $\mathbf{G}$ é linearmente dependente de $\mathbf{1}^{\mathrm{T}}$, ou seja, a única restrição de consistência é linearmente dependente da restrição natural. Logo, a condição b) inclui a condição a).

c) Se $M > 1$, então a condição (15) tem de ser generalizada, o que é obtido pré-multiplicando-se ambos os lados do subsistema de restrições de consistência em (14) por um vetor de constantes $\mathbf{c}^{\mathrm{T}} = [c_1,\ldots,c_M]$, tal que $\mathbf{c}^{\mathrm{T}}\mathbf{G}\mathbf{p} = \mathbf{c}^{\mathrm{T}}\mathbf{a}$. Usando-se o mesmo argumento intuitivo, tem-se a condição generalizada:

$$min\{\mathbf{c}^{\mathrm{T}}\mathbf{G}\} < \mathbf{c}^{\mathrm{T}}\mathbf{a} < max\{\mathbf{c}^{\mathrm{T}}\mathbf{G}\},\ \forall\ \mathbf{c} \neq 0 \qquad (16)$$

onde $\mathbf{c}^{\mathrm{T}}\mathbf{G}$ é um vetor-linha ($1\times N$) e $\mathbf{c}^{\mathrm{T}}\mathbf{a}$ é um escalar (para uma prova formal de (16), ver Alhassid, Agmon & Levine, 1978).

Note-se aqui que, fazendo $\mathbf{G}(r)$ a r-ésima linha de $\mathbf{G}$, a generalização (16) não envolve simplesmente estipular que $min\{\mathbf{G}(r)\} < a_r < max\{\mathbf{G}(r)\}$ para todo $r = 1,\ldots,M$, porque embora esta condição seja necessária, ela não é suficiente. Uma razão para isso pode ser pensada quando se tem, por exemplo, duas restrições de momento em que $a_1 = \Sigma_{i=1}^{N} x_i / N$ (média) e $a_2 = \Sigma_{i=1}^{N} x_i^2 / N$ (2º momento não centrado). Se o problema for definido de modo que a_1 esteja próximo de $\min_i\{g_{1i}\}$ e a_2, em contraposição, esteja próximo de $\max_i\{g_{2i}\}$, então é de se esperar que o sistema não seja factível. Nesta situação, tende a haver uma inconsistência, pois a média é muito baixa e o 2º momento muito alto. Conseqüentemente, a faixa admissível de valores para os a_rs tem de ser mais estreita que a sugerida por (15).

Da mesma forma que antes, os sinais de desigualdade estrita em (16) são relevantes porque:

- se $\mathbf{c}^{\mathrm{T}}\mathbf{a} \to min\{\mathbf{c}^{\mathrm{T}}\mathbf{G}\}$, então $\lambda_r \to \infty$ ($r = 0,1,\ldots,M$), a distribuição MaxEnt/MinxEnt é degenerada em p = (1,0,…,0) com $S = 0$;
- se $\mathbf{c}^{\mathrm{T}}\mathbf{a} \to max\{\mathbf{c}^{\mathrm{T}}\mathbf{G}\}$, então $\lambda_r \to -\infty$ ($r = 0,1,\ldots,M$), a distribuição MaxEnt/MinxEnt é degenerada em $\mathbf{p}^{\mathrm{T}} = (0,\ldots,1)$ com $S = 0$;
- se $\mathbf{c}^{\mathrm{T}}\mathbf{a} = min\{\mathbf{c}^{\mathrm{T}}\mathbf{G}\} = max\{\mathbf{c}^{\mathrm{T}}\mathbf{G}\}$, então $\mathbf{G}$ é linearmente dependente de $\mathbf{1}^{\mathrm{T}}$, ou seja, as restrições de consistência são linearmente dependentes da restrição natural. A condição (16), portanto, inclui as duas anteriores.

A questão agora é: como verificar se o sistema (14) satisfaz (16)? Agmon, Alhassid & Levine (1979) sugerem que se divida o problema em 2 partes:

i) verificar se $[1\ \mathbf{G}^{\mathrm{T}}]^{\mathrm{T}}$ tem posto cheio ($= M + 1$) usando-se o procedimento Gram-Schmidt de ortogonalização de matrizes e;

ii) verificar se (14) tem solução factível usando-se o algoritmo da Fase 1 do Método Simplex de programação linear.

Na implementação computacional aqui desenvolvida em linguagem **MatLab**®, i) pode ser facilmente verificada usando-se a função RANK, que indica o posto de uma matriz. Para verificar ii), entretanto, foi necessário o desenvolvimento de uma função específica e esse é o assunto da próxima subseção.

6.3 A Fase 1 do Método Simplex

O objetivo do Método Simplex é encontrar uma solução ótima para o seguinte problema padrão de programação linear:

$$\text{(PLP)} \begin{cases} \underset{\mathbf{x}}{Min} & \mathbf{d}^{\mathrm{T}}\mathbf{x} \\ s.a. & \begin{cases} \mathbf{A}\mathbf{x} - \mathbf{I}\mathbf{s} = \mathbf{b} \\ \mathbf{x}, \mathbf{s} \geq \mathbf{0} \end{cases} \end{cases}$$

onde:

$\mathbf{d}$ = vetor de coeficientes lineares da função objetivo;

$\mathbf{x}$ = vetor das N variáveis de interesse;

$\mathbf{s}$ = vetor das M variáveis de folga;

$\mathbf{A}$ = matriz ($M \times N$) dos coeficientes de x nas M restrições e

$\mathbf{I}$ = matriz ($M \times M$) de identidade.

Assume-se no problema padrão que $\mathbf{b} \geq 0$. O sistema de restrições $\mathbf{A}\mathbf{x} - \mathbf{I}\mathbf{s} = \mathbf{b}$ possui, portanto, M restrições e $N + M$ incógnitas e representa o conjunto de soluções factíveis para o PLP.

Para ser inicializado, o algoritmo do Método Simplex precisa dispor de uma solução básica inicial (SBI) do sistema de restrições. Uma SBI é um determinado valor de $(\mathbf{x}^{\mathrm{T}},\mathbf{s}^{\mathrm{T}})$ no qual existem N elementos nulos. A chamada *Fase 1 do Método Simplex* corresponde a essa etapa de busca de uma SBI para se inicializar o algoritmo. Uma importante característica da Fase 1 é que ela ou acha uma solução básica inicial ou *detecta que o PLP não tem solução factível* (Solow, 1984). Daí ser útil para indicar se há ou não solução para os problemas MaxEnt e MinxEnt.

Operacionalmente, a Fase 1 usa o mesmo Método Simplex, mas para resolver um PLP modificado (PLPM), no qual são usadas variáveis artificiais $\mathbf{y}$, como segue:

$$\text{(PLPM)} \begin{cases} \underset{\mathbf{y}}{Min} & \mathbf{1}^{\mathrm{T}}\mathbf{y} \\ s.a. & \begin{cases} \mathbf{A}\mathbf{x} - \mathbf{I}\mathbf{s} + \mathbf{I}\mathbf{y} = \mathbf{b} \\ \mathbf{x}, \mathbf{s}, \mathbf{y} \geq \mathbf{0}. \end{cases} \end{cases}$$

Na prática, as restrições lineares do problema MaxEnt e MinxEnt são de igualdade, o que implica que $\mathbf{s} = \mathbf{0}$. Então, definindo-se $(\bar{\mathbf{x}}^{\mathrm{T}}, \bar{\mathbf{y}}^{\mathrm{T}})$ como a solução ótima do PLPM, isto é, da Fase 1, podem haver duas situações: a) $\bar{\mathbf{y}} = \mathbf{0}$, logo existe uma SBI para o PLP original ou b) $\bar{\mathbf{y}} \neq 0$, logo não existe solução factível para o PLP original ou para o seu sistema de restrições.

Uma vez detectado que não há solução factível para o sistema de restrições, o programa de computador tem de ser abortado. O algoritmo detalhado para implementação da Fase 1 do Método Simplex está apresentado no Apêndice II. Na próxima seção, será apresentado o algoritmo desenvolvido para buscar iterativamente a solução dos problemas MaxEnt e MinxEnt.

7. Método de Newton

A obtenção de uma solução para os problemas MaxEnt e MinxEnt duais demanda a utilização de um algoritmo de procura iterativa. Ambos são problemas de otimização não-linear com restrições de igualdade apenas, para os quais existem vários métodos de procura disponíveis. Na presente implementação, foi adotado o conhecido Método de Newton, devido às suas vantagens de convergência rápida e certa (quando a função-objetivo é estritamente convexa), além de ser de fácil programação.

Na implementação deste método, é realizado o cálculo do gradiente, do Hessiano e da inversa do Hessiano da função objetivo a cada iteração. O cálculo da inversa constitui uma desvantagem em problemas maiores, mas como a dimensão do Hessiano nos problemas MaxEnt e MinxEnt é determinada pelo número de restrições *M*, raramente este número será grande. Obviamente, melhoras podem ser introduzidas com outros enfoques, como por exemplo os métodos de quase-Newton (Bertsekas, 1995).

Considere-se, inicialmente, as representações primal e dual para o problema MaxEnt, escritas em forma vetorial e matricial no Quadro 7.1:

Quadro 7.1 – Representações primal e dual de MaxEnt

PRIMAL	**DUAL**
$\underset{\mathbf{p}}{Max} \quad S = \mathbf{p}^{\mathrm{T}} \ln \mathbf{p}$ $s.a. \quad \begin{bmatrix} \mathbf{1}^{\mathrm{T}} \\ \mathbf{G} \end{bmatrix} \mathbf{p} = \begin{bmatrix} 1 \\ \mathbf{a} \end{bmatrix}$	$\underset{\mathbf{z}}{Min} \quad L_s^* = \lambda_0(\mathbf{z}) + \mathbf{a}^{\mathrm{T}}\mathbf{z}$ (*irrestrito*)

onde $\lambda_0(\mathbf{z}) = \ln\left(\mathbf{1}^{\mathrm{T}} \exp(\mathbf{G}^{\mathrm{T}}\mathbf{z})\right)$ e $\mathbf{z}^{\mathrm{T}} = [\lambda_1, \ldots, \lambda_M]$. A solução do primal é dada por:

$$\mathbf{p} = \frac{\exp[-\mathbf{G}^{\mathrm{T}}\mathbf{z}]}{\exp(\lambda_0)} = \frac{\exp[-\mathbf{G}^{\mathrm{T}}\mathbf{z}]}{\mathbf{1}^{\mathrm{T}} \exp[-\mathbf{G}^{\mathrm{T}}\mathbf{z}]} . \tag{19}$$

A implementação do Método de Newton para o problema dual envolve a busca iterativa de $\mathbf{z}^*$ = solução ótima do problema através do seguinte algoritmo:

$$\begin{aligned} \mathbf{z}_k &= \mathbf{z}_{k-1} + \mathbf{d}_k \\ \mathbf{d}_k &= -\mathbf{H}_k^{-1}\mathbf{h}_k \end{aligned} \qquad (20)$$

onde k = índice de iteração, $\mathbf{d}_k$ = direção de Newton, $\mathbf{h}_k = \nabla L_k^*$ = gradiente de L_k^* e $\mathbf{H}_k = \nabla^2 L_k^*$ = matriz hessiana de L_k^*. A direção de Newton $\mathbf{d}_k$ usa informações referentes às primeiras e segundas derivadas em relação ao vetor de multiplicadores de Lagrange $\mathbf{z}$, que estão embutidas no gradiente e na matriz hessiana. Como ambos são conhecidos analiticamente, podem ser calculados diretamente, isto é, sem o uso de aproximações, da seguinte forma:

$$\begin{aligned} \mathbf{h}_k &= \mathbf{a} - \mathbf{G}\mathbf{p}_k \\ \mathbf{P}_k &= \mathbf{diag}(\mathbf{p_k}) \\ \mathbf{H}_k &= \mathbf{G}[\mathbf{P}_k - \mathbf{p}_k\mathbf{p}_k^{\mathrm{T}}]\mathbf{G}^{\mathrm{T}} \end{aligned} \qquad (21)$$

onde $\mathbf{p}_k$ é calculado a cada iteração segundo (19). Note-se que a matriz hessiana $\mathbf{H}_k$ nada mais é do que a matriz $M \times M$ de variância-covariância das colunas de $\mathbf{G}$ calculada sob a distribuição $\mathbf{p}_k$, sendo, portanto, definida positiva sempre.

O algoritmo completo de busca pela solução dual de MaxEnt é formado pelas expressões em (20) e (21). Sua implementação demanda ainda resolver a questão da inicialização e a definição do critério de parada. Em ambos os casos, existem procedimentos alternativos que podem ser utilizados. Na inicialização, é preciso definir-se o vetor inicial de multiplicadores de Lagrange $\mathbf{z}_0^{\mathrm{T}} = [\lambda_{10},\ldots, \lambda_{M0}]$. Um cuidado a ser tomado é o de que, se $\mathbf{z_0}$ for definido na região assintótica, então isto pode provocar uma quase singularidade da matriz H_1. Um procedimento mais seguro pode ser definir-se $\mathbf{z}_0 = \mathbf{0}$. De fato, nos diversos testes realizados com a função MAXENT escrita em **MatLab**®, este valor de inicialização $\mathbf{z}_0$ foi o que apresentou os resultados corretos com maior freqüência. Por sua vez, o critério de parada também pode ser definido de diversas formas. A que foi adotada na função MAXENT é baseada na norma do gradiente e em um número máximo de iterações, de modo que o algoritmo pára quando $\| \mathbf{h}_k \| < 10^{-5}$ ou quando o número de iterações exceder 30.

Considere-se, agora, as representações primal e dual para o problema MinxEnt, que estão apresentadas no Quadro 7.2.

Quadro 7.2 – Representações primal e dual de MinxEnt

PRIMAL	**DUAL**
$\underset{\mathbf{p}}{Min} \quad K = \mathbf{p}^{\mathrm{T}}(\ln\mathbf{p} - \ln\mathbf{q})$ $s.a. \quad \begin{bmatrix}\mathbf{1}^{\mathrm{T}} \\ \mathbf{G}\end{bmatrix}\mathbf{p} = \begin{bmatrix}1 \\ \mathbf{a}\end{bmatrix}$	$\underset{\mathbf{z}}{Min} \quad -L_k^* = \lambda_0(\mathbf{z}) + \mathbf{a}^{\mathrm{T}}\mathbf{z}$ $(irrestrito)$

onde $\lambda_0(\mathbf{z}) = \ln\left(\mathbf{q}^{\mathrm{T}} \exp[-\mathbf{G}^{\mathrm{T}}\mathbf{z}]\right)$ e $\mathbf{z}^{\mathrm{T}} = [\lambda_1,\ldots, \lambda_M]$. A solução do primal para $\mathbf{p}$ é dada por:

$$\mathbf{p} = \frac{\mathbf{q} \circ \exp[-\mathbf{G}^{\mathrm{T}}\mathbf{z}]}{\exp(\lambda_0)} = \frac{\mathbf{q} \circ \exp[-\mathbf{G}^{\mathrm{T}}\mathbf{z}]}{\mathbf{q}^{\mathrm{T}} \exp[-\mathbf{G}^{\mathrm{T}}\mathbf{z}]} \tag{22}$$

onde " $\circ$ " representa o produto elemento a elemento de Hadamard.

Dado que a representação dual de MinxEnt no Quadro 7.2 está escrita como um problema de minização, a solução ótima pode ser procurada usando o mesmo algoritmo composto por (20) e (21). A única diferença reside no cômputo de $\mathbf{p}_k$, que no caso MinxEnt é feito de acordo com (22), onde se tem a presença da distribuição *a priori* **q** no numerador e no denominador da expressão. As recomendações acerca dos procedimentos de inicialização e em relação ao critério de parada são as mesmas.

8. Funções em Linguagem MatLab

A linguagem de programação **MatLab®**, desenvolvida pela empresa *The Math Works Inc*., foi utilizada para se escrever um programa para solucionar os problemas MaxEnt e MinxEnt, conforme a metodologia apresentada nas seções anteriores. Na verdade, foram escritas três funções: MAXENT, MINXENT e PHASE1. Um detalhamento dos procedimentos realizados por ambas as funções MAXENT e MINXENT encontra-se no fluxograma apresentado no Apêndice 1. Para a função PHASE1, seu detalhamento encontra-se no fluxograma apresentado no Apêndice 2.

A única diferença, como se verá abaixo, entre as funções MAXENT e MINXENT é que na segunda é permitido ao usuário entrar uma distribuição *a priori* **q** e o cálculo de **p** a cada iteração é feito segundo (22). Nos demais aspectos, o algoritmo é o mesmo. Ambas as funções realizam testes, antes de implementar o método de Newton, para verificar:

a) se $[1\ \mathrm{G}^{\mathrm{T}}]^{\mathrm{T}}$ tem posto cheio (= M + 1) e;

b) se existe solução factível para o sistema de restrições.

O teste em a) é feito através da função RANK disponível na linguagem **MatLab®** e o teste em b) é realizado usando-se a função PHASE1, que implementa o algoritmo da Fase 1 do Método Simplex (conforme explicado na subseção 6.3 e no Apêndice 1), e que foi aqui desenvolvida especialmente para isso.

8.1 Função MAXENT

Implementa a solução do problema de maximização da medida de entropia de Shannon sujeita a restrições lineares. A sintaxe da função é:

[**p**,**z**] = MAXENT(**G**,**a**,*op*).

Na entrada, o usuário fornece a matriz de coeficientes das restrições de consistência **G** e o vetor **a**. Este último tem de ser entrado como um vetor-coluna. Na saída, a função fornece os valores da distribuição ótima (distribuição MaxEnt) **p** e do vetor ótimo **z** de multiplicadores de Lagrange. Se o usuário desejar que na saída a função apresente gráficos da evolução de L_s* a cada iteração e da distribuição MaxEnt, ele tem que definir na entrada *op* = 1. Se desejar apenas o gráfico da distribuição MaxEnt, tem que definir *op* = 2 e se não quiser gráfico nenhum, tem de fazer *op* = outro número.

8.2 Função MINXENT

Implementa a solução do problema de minimização da medida de entropia cruzada de Kullback sujeita a restrições lineares. A sintaxe da função é:

[**p**,**z**] = MINXENT(*G*,**a**,**q**,*op*).

Na entrada, o usuário fornece a matriz de coeficientes das restrições de consistência G, o vetor **a** e a distribuição a priori **q**. Estes dois últimos vetores têm de ser entrados como vetores-coluna. Na saída, a função fornece os valores da distribuição ótima (distribuição MinxEnt) **p** e do vetor ótimo **z** de multiplicadores de Lagrange. As demais características são as mesmas que para a função MAXENT.

9. Um exemplo de determinação do número de viagens

Para ilustrar a aplicação dos dois princípios de otimização da entropia, bem como das funções MAXENT e MINXENT desenvolvidas com essa finalidade, esta seção apresenta um pequeno exemplo onde eles são usados para resolver um problema típico de planejamento de transportes. Este problema refere-se à determinação do número de viagens entre regiões dentro de uma localidade hipotética (e.g., Novaes, 1981).

O dados disponíveis foram criados artificialmente e estão apresentados na Tabela 9.1. O lado esquerdo da tabela apresenta as freqüências absolutas de viagens originadas das regiões O_1, O_2 e O_3, com destino às regiões D_1, D_2 e D_3. Por sua vez, o lado direito apresenta esses dados como proporções do total de viagens. Por exemplo, o número de viagens de O_1 para D_1 é 30 na matriz de fluxos e está associado à proporção 0,0952 (=30/315) na matriz de proporções.

Tabela 9.1 – Dados artificiais de número de viagens

	Matriz de Fluxos				**Matriz de Proporções**			
	***D*₁**	***D*₂**	***D*₃**					
***O*₁**	30	52	10	92	0,0952	0,1651	0,0317	0,2921
***O*₂**	45	65	35	145	0,1429	0,2063	0,1111	0,4603
***O*₃**	9	42	27	78	0,0286	0,1333	0,0857	0,2476
	84	159	72	315	0,2667	0,5048	0,2286	1

Em geral, no problema de determinação do número de viagens, são conhecidos os valores totais das colunas e das linhas das matrizes, mas não o conteúdo das células. O objetivo é determinar as proporções de viagens de cada região para as demais, que se assume sejam desconhecidas (e, determinando-se as proporções, é fácil calcular as freqüências correspondentes). Em outras palavras, o objetivo é determinar, ou "estimar", a matriz de proporções $\mathbf{P} = \{p_{ij}\}$ apresentada na Tabela 9.1. O fato de que esta seja conhecida no âmbito do exemplo é útil porque permite avaliar a capacidade de recuperação das proporções desagregadas com os métodos MaxEnt e MinxEnt, bem como compará-los entre si.

O problema de determinação de viagens pode ser formulado como um problema MaxEnt ou MinxEnt, como segue:

$$\underset{\mathbf{P}}{Max} \quad \sum_{i=1}^{3}\sum_{j=1}^{3} F(p_{ij})$$

$$s.a. \quad \begin{cases} \sum_{i=1}^{3}\sum_{j=1}^{3} p_{ij} = 1 \\ \sum_{i=1}^{3} p_{ij} = p_{\bullet j} \quad j = 1,...,3 \\ \sum_{j=1}^{3} p_{ij} = p_{i\bullet} \quad i = 1,...,3. \end{cases}$$

Se $F(p_{ij}) = -p_{ij}\ln p_{ij}$, então o problema é do tipo MaxEnt; se $F(p_{ij}) = p_{ij}\ln(p_{ij}/q_{ij})$, então é do tipo MinxEnt. Cada variável p_{ij} pode representar a probabilidade de um indivíduo viajar da região-origem O_i para a região-destino D_j. Alternativamente, pode-se falar em p_{ij} como a *proporção* do total de viagens realizadas de O_i para D_j, e essa será a terminologia usada aqui. Existem nove proporções a se determinar através da solução do problema.

Os dados da Tabela 9.1 embutem as informações do conjunto de restrições do problema. As sete restrições formam um sistema linear com infinitas soluções possíveis a priori, logo o conjunto factível não é vazio. A primeira é a restrição natural; o grupo a seguir de 3 restrições representam o fato de que a soma das proporções na coluna j é igual à proporção agregada da coluna $p_{\cdot j}$, e o terceiro grupo de 3 restrições representam, analogamente, o fato de que a soma das proporções na linha i é igual à proporção agregada da linha $p_{i\cdot}$; assume-se também (para simplicidade de exposição) que não existem diferenciais de custo de transporte entre as regiões.

Representando-se o sistema de restrições na notação matricial adotada ao longo do artigo, tem-se $\mathbf{Gp} = \mathbf{a}$, onde $\mathbf{p} = vec(\mathbf{P})$. Então, a partir dos dados constantes da Tabela 9.1, é possível definir-se:

$$\mathbf{G} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

e:

$$\mathbf{a}^{\mathrm{T}} = (p_{\cdot 1} \quad p_{\cdot 2} \quad p_{1\cdot} \quad p_{2\cdot}) = (0{,}2921 \quad 0{,}4603 \quad 0{,}2667 \quad 0{,}5048)\,.$$

A matriz $\mathbf{G}$ só apresenta quatro linhas (e o vetor $\mathbf{a}$ só quatro elementos) porque ao representar as incógnitas do problema como proporções, duas restrições ficam redundantes. Bastaria a inclusão de apenas uma delas para fazer com que $\mathbf{G}$ apresentasse dependência linear entre suas linhas (na verdade, haveria dependência linear entre as restrição de consistência e a restrição natural).

Usando a função MAXENT, a estimativa da matriz $\mathbf{P}$ obtida seria:

$$\hat{\mathbf{P}} = \begin{bmatrix} 0{,}0779 & 0{,}1474 & 0{,}0668 \\ 0{,}1228 & 0{,}2324 & 0{,}1052 \\ 0{,}0660 & 0{,}1250 & 0{,}0566 \end{bmatrix}$$

e o grau de aderência de $\hat{\mathbf{P}}$ em relação a $\mathbf{P}$ (que é conhecida hipoteticamente) poderia ser computado como:

$$\hat{s} = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3}(p_{ij} - \hat{p}_{ij})^2} = 7{,}262\times10^{-2}\,.$$

A estimativa de **P** ainda poderia ser melhorada através do método MinxEnt. Neste caso, poderia ser usada alguma informação a priori sobre a matriz de proporções. Isto é, suponhamos que houvesse uma matriz de proporções disponível, correspondente a um período anterior, e que tivesse sido computada a partir de um *survey* ou pesquisa de campo. Por exemplo:

$$\mathbf{Q} = \begin{bmatrix} 0{,}0975 & 0{,}1656 & 0{,}0290 \\ 0{,}1430 & 0{,}1967 & 0{,}1104 \\ 0{,}0299 & 0{,}1397 & 0{,}0883 \end{bmatrix}.$$

A matriz **Q** também foi produzida artificialmente, a partir da multiplicação dos valores da matriz de fluxo da Tabela 9.1 por variáveis geradas aleatoriamente segundo uma distribuição normal com média um e desvio padrão 0,05. Definindo-se **q** = *vec*(**Q**) e usando-se agora a função MINXENT, obtém-se uma nova estimativa da matriz **P**:

$$\tilde{\mathbf{P}} = \begin{bmatrix} 0{,}0955 & 0{,}1672 & 0{,}0293 \\ 0{,}1432 & 0{,}2030 & 0{,}1141 \\ 0{,}0280 & 0{,}1346 & 0{,}0852 \end{bmatrix}.$$

Como seria de se esperar, a matriz $\tilde{\mathbf{P}}$ aproxima melhor a matriz **P** devido ao uso das informações a priori, o que se reflete no seu grau de aderência:

$$\tilde{s} = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3}(p_{ij} - \tilde{p}_{ij})^2} = 1{,}819\times10^{-2}$$

que é cerca de quatro vezes menor do que $\hat{s}$.

Vale notar aqui que o uso do princípio MinxEnt de Kullback proporciona um mecanismo alternativo à calibração dos modelos de entropia (Novaes, 1981). Se fosse adotado no exemplo aqui considerado, o processo de calibração envolveria usar o modelo MaxEnt mas ajustando-se antes os multiplicadores de Lagrange a alguma informação prévia sobre a matriz **P** e segundo algum critério de ajuste externo, como mínimos quadrados ou máxima verossimilhança. Essa informação prévia é fornecida no exemplo pela matriz **Q,** só que tal informação é usada no âmbito do próprio formalismo do princípio MinxEnt, onde **Q** é explicitamente tratada como uma distribuição a priori para a qual se busca a distribuição mais próxima segundo a medida de entropia cruzada de Kullback.

Por último é importante ressaltar que os problemas onde os princípios MaxEnt e MinxEnt são aplicáveis precisam ser adequadamente traduzidos em termos da representação matricial usada no artigo. Ou seja, é preciso definir de modo apropriado a matriz **G** e o vetores **p** e **a**, e a estratégia para fazer isso pode variar de problema para problema. Uma vantagem do dispositivo, introduzido em seções anteriores, para verificar previamente a existência de solução dos problemas é a de ajudar o analista, indicando-lhe a necessidade de revisar a formulação do problema (ou o uso adequado do formalismo) e levando-o à adoção imediata de procedimentos corretivos.

10. Comentários Finais

Este artigo apresentou uma implementação computacional eficiente dos princípios MaxEnt de Jaynes e MinxEnt de Kullback para o caso de distribuições de probabilidade discretas. Ambos os princípios podem ser formulados como problemas de otimização não-linear com restrições lineares de igualdade. A implementação eficiente dos princípios foi obtida aqui pela resolução dos problemas duais, que envolve a utilização de um algoritmo de procura dada a não-linearidade intrínseca em relação aos multiplicadores de Lagrange que são os argumentos da função dual. O algoritmo usado foi o de Newton, devido à sua boa velocidade de convergência e à sua convergência certa nos problemas MaxEnt e MinxEnt duais devido à convexidade estrita das funções objetivo.

Como existe a possibilidade de que, dependendo da forma como forem caracterizadas as restrições, os problemas de otimização não apresentem solução, foi adicionado ao algoritmo proposto um dispositivo que verifica antes a existência de soluções factíveis para os problemas. Essa verificação envolve testar se o sistema de restrições lineares comum a ambos possui solução, ou melhor, caracteriza um conjunto viável que não é vazio. O procedimento de teste foi implementado aqui através do algoritmo para a Fase 1 do Método Simplex, usado em programação linear.

Vale salientar que esse teste inicial é relevante pois evita a situação imprecisa, e desconfortável, de se aguardar um período de tempo arbitrário pela convergência do algoritmo para, caso não haja convergência, concluir pela inexistência de solução. O teste de existência, ao indicar precisamente se o problema apresenta solução, permite também identificar rapidamente deficiências de modelagem do problema aplicado e conduzir imediatamente à adoção de procedimentos para sua correção.

Outra vantagem do procedimento de verificação de solução reside no fato de que, em geral, as rotinas de otimização de propósito geral, disponíveis para uso com as principais linguagem de programação de métodos matemáticos, não incorporam ou não permitem embutir este teste inicial. Assim, rotinas desenvolvidas conforme a metodologia apresentada no artigo especialmente para os princípios MaxEnt e MinxEnt (como as escritas pelos autores em linguagem Matlab) serão de maior utilidade prática para os interessados em aplicar os princípios.

Referências Bibliográficas

(1) Alhassid, Y.; Agmon, N. & Levine, R.D. (1978). An upper bound for the entropy and its applications to the maximal entropy problem. *Chemical Physics Letters*, **53**(**1**), 22-26.

(2) Agmon, N.; Alhassid, Y. & Levine, R.D. (1979). An algorithm for finding the distribution of Maximal Entropy. *Journal of Computational Physics*, **30**, 250-258.

(3) Bertsekas, D. (1995). *Non-linear programming*. Boston: Athena Scientific.

(4) Fang, S.-C.; Rajasekera, J.R. & Tsao, H.S.J. (1997). *Entropy optimization and mathematical programming*. International Series in Operations Research & Management Sciences. Boston: Kluwer.

(5) Georgescu-Roegen, N. (1971). *The entropy law and the economic process*. Cambridge, Massachusetts: Harvard University Press.

(6) Golan, A.; Judge, G. & Miller, D. (1996). *Maximum entropy econometrics*. New York: Wiley.

(7) Jaynes, E.T. (1957a). Information theory and statistical mechanics. *Physics Review*, **106**, 620-630.

(8) Jaynes, E.T. (1957b). Information theory and statistical mechanics II. *Physics Review*, **108**, 171-190.

(9) Kapur, J.N. & Kesavan, H.K. (1992). *Entropy optimization principles with applications*. London: Academic Press.

(10) Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.

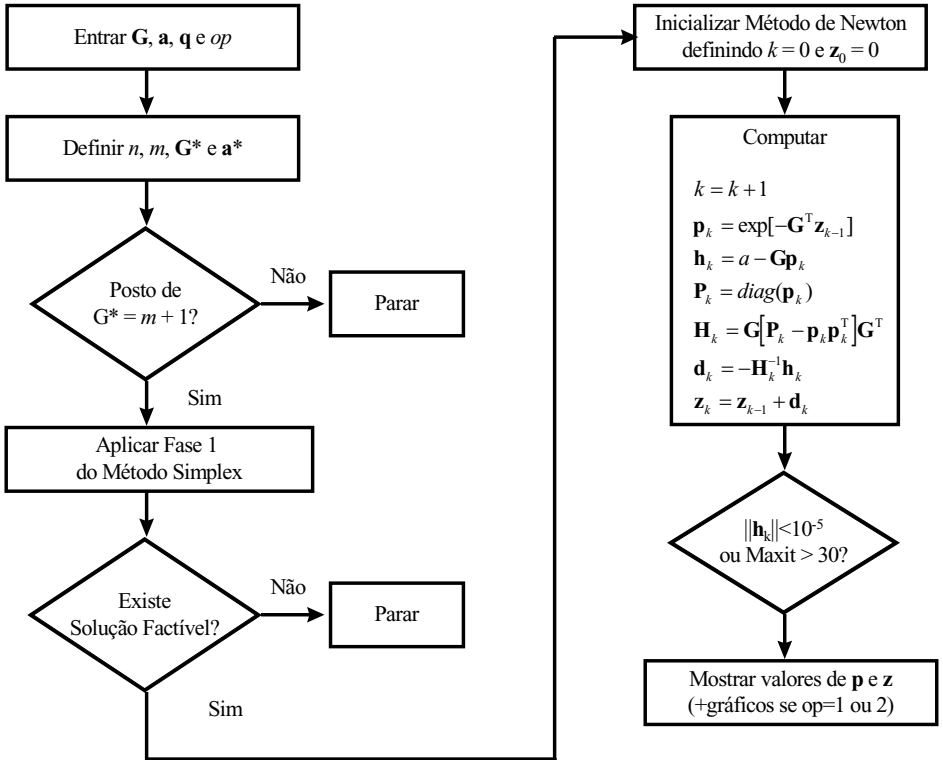
(11) Novaes, A.G. (1981). Modelos em planejamento urbano, regional e de transportes. São Paulo: Edgard Blücher.

(12) Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379-423.

(13) Solow, D. (1984). Linear algebra: an introduction to finite improvement algorithms. New York: North-Holland.

Apêndice 1

Figura A.1 – Fluxograma das Funções MAXENT e MINXENT



Nota: $G^* = [\mathbf{1}\ G^T]^T$ e $a^* = [1\ a]^T$.

Apêndice 2

Algoritmo para Teste de Existência de Solução dos Problemas MaxEnt e MinxEnt

Este apêndice descreve um algoritmo para testar a existência ou não de soluções factíveis para os problemas MaxEnt e MinxEnt e que é baseado na Fase 1 do Método Simplex. Dado um PLP padrão, a Fase 1 visa determinar, no âmbito do Método Simplex, uma SBI para a se inicializar a busca por uma solução ótima. Entretanto, ela permite também detectar se existe solução factível para o referido PLP.

Consideremos o seguinte PLP padrão com restrições sem variáveis de folga (**s** = **0**):

$$\text{(PLP)} \begin{cases} \underset{\mathbf{x}}{Min} & \mathbf{d}^{\mathrm{T}}\mathbf{x} \\ s.a. & \begin{cases} \mathbf{A}\mathbf{x} = \mathbf{b} \\ \mathbf{x} \geq \mathbf{0} \end{cases} \end{cases}$$

onde, **x** = vetor com N variáveis e **Ax = b** é um sistema com M restrições. No âmbito do PLP padrão, também assume-se que **b** ≥ 0. Usando-se M variáveis artificiais contidas em um vetor $\mathbf{y}^{\mathrm{T}} = [y_1,..,y_m]$, o PLPM da Fase 1 pode ser escrito como:

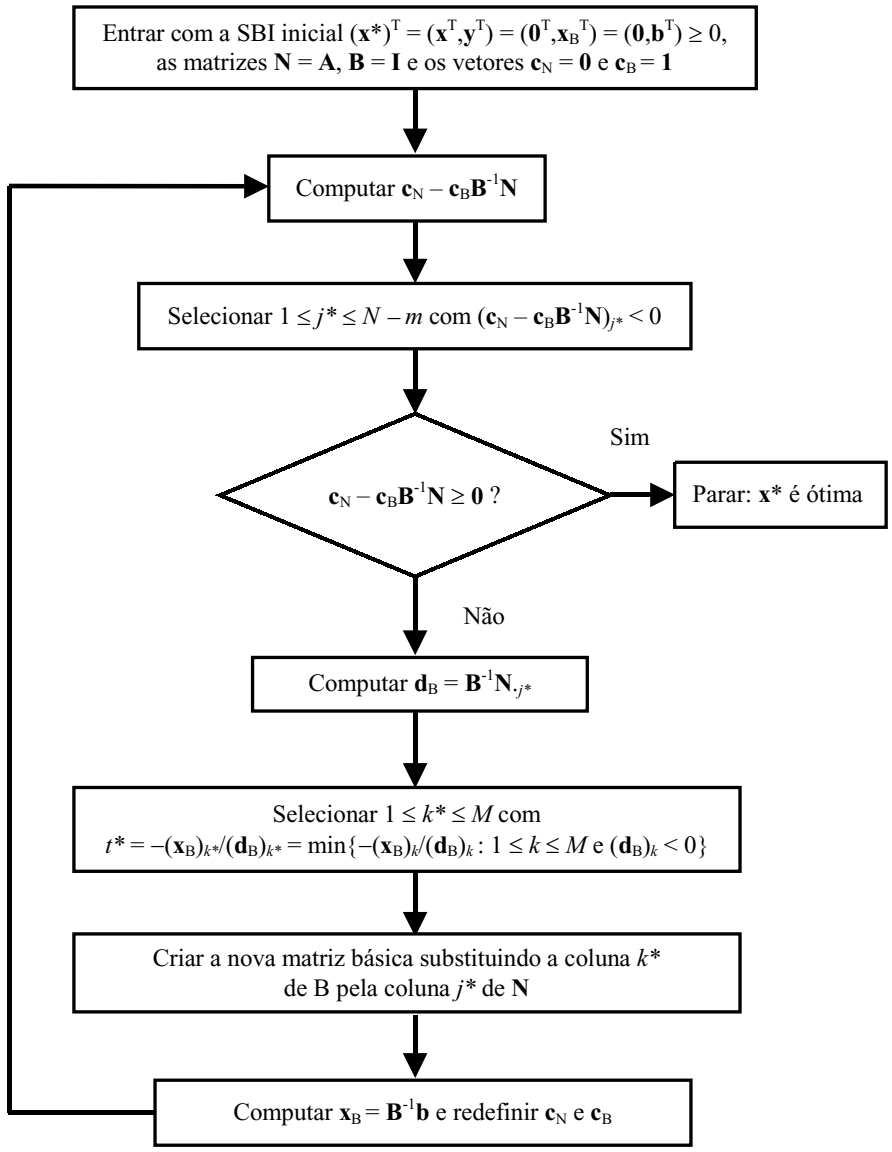
$$\text{(PLPM)} \begin{cases} \underset{\mathbf{y}}{Min} & [\mathbf{0}_n^{\mathrm{T}}\ \mathbf{1}_m^{\mathrm{T}}]\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \\ s.a. & \begin{cases} [\mathbf{A}\ \mathbf{I}]\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} = \mathbf{b} \\ \mathbf{x},\mathbf{y} \geq 0. \end{cases} \end{cases}$$

O mesmo Método Simplex para o qual se quer achar uma SBI é utilizado para encontrar uma solução para o PLPM da Fase 1. Como para qualquer PLP, encontrar uma SBI inicial para o PLPM da Fase 1 requer localizar-se uma matriz básica **B** ($m \times m$) dentro da matriz [**A I**] e computar $\mathbf{x}_{\mathrm{B}} = \mathbf{B}^{-1}\mathbf{b} \geq \mathbf{0}$, de modo que $(\mathbf{x}^{\mathrm{T}},\mathbf{y}^{\mathrm{T}}) = (\mathbf{0}^{\mathrm{T}},\mathbf{x}_{\mathrm{B}}^{\mathrm{T}})$ seja uma solução básica. É possível mostrar que o último vetor constitui uma SBI para o PLPM da Fase 1 (Solow, 1984, p. 192). Logo, não é preciso procurar uma SBI para este problema.

Considere-se agora que, antes de inicializar o algoritmo da Fase 1, sejam definidos **N** = **A** e **B** = **I**, de modo que [**A I**] = [**N B**]. Sejam também $\mathbf{c}_{\mathrm{N}}$ = vetor linha dos coeficientes das variáveis não básicas e $\mathbf{c}_{\mathrm{B}}$ = vetor linha dos coeficientes das variáveis básicas, de modo que $(\mathbf{c}_{\mathrm{N}},\mathbf{c}_{\mathrm{B}}) = (\mathbf{0}_n^{\mathrm{T}},\mathbf{1}_m^{\mathrm{T}})$. Então, o algoritmo presente na Figura A.2, que constitui uma versão modificada do Método Simplex para o PLPM da Fase 1, pode ser usado para testar a existência de solução factível do sistema **Ax** = **b** (com **x** e **b** ≥ **0**). Se a solução ótima encontrada $(\bar{\mathbf{x}}^{*})^{\mathrm{T}} = (\bar{\mathbf{x}}^{\mathrm{T}},\bar{\mathbf{y}}^{\mathrm{T}})$ apresentar $\bar{\mathbf{y}} \neq 0$, então não existe solução para **Ax = b**.

Note-se que o algoritmo da Figura A.2 opera apenas sobre o sistema de restrições do PLP original, isto é, não usa informações sobre os coeficientes (vetor **d**) da função objetivo. Assim, independentemente da função objetivo, o algoritmo para a Fase 1 na verdade testa a existência ou não de solução factível para o sistema de restrições. Em outras palavras, o algoritmo pode ser usado para fazer este teste para qualquer sistema de equações lineares (onde **x** ≥ **0** e **b** ≥ **0**). É isto que permite que ele seja usado no contexto dos problemas MaxEnt/MinxEnt, que embora não sejam um problema de programação linear, sua solução depende da factibilidade de seu sistema de restrições lineares (para o qual tem-se **x** = **p** ≥ 0).

Figura A.2 – Fluxograma* de um algoritmo para a Fase 1 do Método Simplex



Adaptado pelos autores de um fluxograma de Solow (1984, p. 175).