Métodos Numéricos de Otimização e NLLS

Bibliografia: Greene

Claudio Lucinda

FEA/USP

Overview

Métodos Numéricos

- Como vimos na aula passada, estimadores de extremo são uma classe que abrange todos os métodos que iremos estudar neste curso.
- E acho que deve ser intuitivo neste momento imaginar que um aspecto prático destes estimadores é determinar **como** se alcança aquele extremo da função.
- Neste começo da aula, iremos discutir métodos numéricos aplicados à otimização.

Soluções de forma fechada e caso geral

- Suponha a seguinte função, com $\mathbf{C}$ sendo uma matriz positiva semidefinida:

$$F(\theta) = a + \mathbf{b}'\theta - \frac{1}{2}\theta'\mathbf{C}\theta$$

- As CPO e solução são:

$$\frac{\partial F}{\partial \theta} = \mathbf{b} - \mathbf{C}\theta = \mathbf{0} \rightarrow \theta = \mathbf{C}^{-1}\mathbf{b}$$

Esta é uma solução de forma fechada ou solução analítica. Sabendo $\mathbf{C}$ e $\mathbf{b}$, sabemos θ

Otimização Numérica

- Agora iremos nos focar no caso mais geral de se maximizar uma função de várias variáveis, $\max_\theta F(\theta)$
- Evidentemente, minimizar $F(\theta)$ equivale a maximizar $-F(\theta)$
- Uma forma óbvia de se buscar pelo θ que maximiza a função é por tentativa e erro. Se θ estiver em um intervalo pequeno e for fácil calcular a função em cada ponto, uma estratégia como essa (também conhecida como **grid search**) funciona.
- Existem também métodos mais sofisticados que fazem esta tentativa e erro de forma mais organizada.
  - Um método bastante robusto é o Nelder Mead - a melhor explicação está neste vídeo aqui: **link**.
  - Outros são o Simulated Annealing, Differential Evolution, Genetic Algorithm
- Um exemplo gráfico do desempenho relativo dos métodos tá nesse **link** aqui
- Outro caminho é a partir do fato que a solução passa por uma derivada, e a utilização desta derivada pode ser útil.

Algoritmos

- Podemos resumir o que veremos mais adiante na seguinte forma geral do algoritmo:
- Começando com o valor inicial θ_0, em uma iteração genérica t, se o θ_t não for o valor ótimo para $F(\theta)$, calcule um vetor de direção Δ_t, e um tamanho do passo λ_t, e a partir daí;
- $\theta_{t+1} = \theta_t + \lambda_t \Delta_t$
- Diferentes métodos nos fornecem soluções para Δ_t e λ_t. Em alguns casos, o λ_t não é obtido junto com o Δ, e um passo subsidiário (**line search**) é feito para se encontrar o valor ideal para este parâmetro.
- Agora vamos focar nos métodos baseados no gradiente, que são aqueles em que Δ é calculado a partir do gradiente da função.

Métodos Baseados no Gradiente

Nessa família de métodos, $\mathbf{\Delta_t} = \mathbf{W_t g_t}$, em que $\mathbf{W_t}$ é uma matriz positiva semidefinida e $\mathbf{g_t} = \mathbf{g}(\theta_\mathbf{t}) = {}^{\partial F(\theta_t)}/_{\partial \theta_\mathbf{t}}$ é o gradiente da função.
Isso é motivado pelo seguinte. Considere uma expansão de Taylor para $F\left(\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t\right)$ em torno de $\lambda_t = 0$

$$F\left(\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t\right) \simeq F\left(\boldsymbol{\theta}_t\right) + \lambda_t \mathbf{g}\left(\boldsymbol{\theta}_t\right)' \boldsymbol{\Delta}_t$$

Seja $F\left(\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t\right) = F_{t+1}$. Então,

$$F_{t+1} - F_t \simeq \lambda_t \mathbf{g}_t' \boldsymbol{\Delta}_t$$

Se $\Delta_t = \mathbf{W}_t \mathbf{g}_t$, então

$$F_{t+1} - F_t \simeq \lambda_t \mathbf{g}_t' \mathbf{W}_t \mathbf{g}_t$$

Se $\mathbf{g}_t$ não é $\mathbf{0}$ e λ_t é pequeno o suficiente, então $F_{t+1} - F_t$ é positivo.
Portanto, se $F(\boldsymbol{\theta})$ não estiver no máximo, sempre podemos encontrar um "step" tal que uma iteração como essa leva a um aumento na função.

Steepest Ascent/Descent

O algoritmo mais simples é o de maior subida ("Steepest Ascent"), que usa

$$\mathbf{W} = \mathbf{I} \rightarrow \Delta = \mathbf{g}$$

Como seu nome implica, a direção é a de maior aumento de $F(.)$.
Outra virtude é que o passo de "line search" tem uma solução óbvia; pelo menos perto do máximo, o λ ótimo é

$$\lambda = \frac{-\mathbf{g}'\mathbf{g}}{\mathbf{g}'\mathbf{H}\mathbf{g}}$$

Em que

$$\mathbf{H} = \frac{\partial^2 F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

A iteração fica sendo

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\mathbf{g}_t'\mathbf{g}_t}{\mathbf{g}_t'\mathbf{H}_t\mathbf{g}_t}\mathbf{g}_t$$

Calculando Derivadas e Hessianos

- Nesse método, temos problemas potenciais com o **H**.
- Podemos ter que a própria matriz seja difícil de calcular.
- Além disso, especialmente quando o θ_t está longe do máximo, não dá para garantir que o **H** seja negativa definida – o que faz com que o algoritmo divirja ao invés de convergir.
- Usualmente este algoritmo converge muito devagar.

Método de Newton/BHHH

A base para o método de Newton é uma expansão de Taylor de primeira ordem a partir das condições de primeira ordem, em torno de um ponto qualquer $\boldsymbol{\theta}^0$

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \simeq \mathbf{g}^0 + \mathbf{H}^0 \left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right) = \mathbf{0}$$

Em que o sobrescrito $\boldsymbol{\theta}^0$ indica que foi avaliado em θ_0. Resolvendo para $\boldsymbol{\theta}$ e igualando $\boldsymbol{\theta}$ a $\boldsymbol{\theta}_{t+1}$ e $\boldsymbol{\theta}^0$ a $\boldsymbol{\theta}_t$, temos

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{H}_t^{-1}\mathbf{g}_t$$

Portanto, para o Método de Newton, temos:

$$\mathbf{W} = -\mathbf{H}^{-1}, \quad \boldsymbol{\Delta} = -\mathbf{H}^{-1}\mathbf{g}, \quad \lambda = 1.$$

O método BHHH utilizado em muitos pacotes econométricos envolve substituir o Hessiano $\mathbf{H}$ pelo produto externo dos gradientes. Ele é formalmente válido no contexto de funções verossimilhança e se baseia na igualdade da matriz de informação (a ser vista ainda).

Quadratic Hill Climbing

- Uma extensão do método de Newton para lidar com o problema que o Hessiano pode não ser negativo definido longe do ótimo é o "quadratic hill climbing".
- Em cada iteração, se o **H** não for negativa definida, ela seria substituída por

$$\mathbf{H}_\alpha = \mathbf{H} - \alpha\mathbf{I}$$

Em que α é um número positivo grande o suficiente para assegurar que $\mathbf{H}_\alpha$ seja negativa definida.

Métodos Quasi-Newton

Uma classe de algoritmos é chamada de "Quasi-Newton" porque, ainda que usem as ideias do Método de Newton, não usam os hessianos diretamente e possuem propriedades muito boas de convergência. Eles tem a forma de:

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{E}_t$$

Em que $\mathbf{E}_t$ é uma matriz positiva definida
Enquanto $\mathbf{W}_0$ for positiva definida (matriz identidade costuma ser usada), $\mathbf{W}_t$ será Positiva Definida em toda iteração.
No método Davidon-Fletcher-Powell (DFP), após um número suficiente de iterações, $\mathbf{W}_{t+1}$ será uma aproximação de $-\mathbf{H}^{-1}$. Definindo

$$\boldsymbol{\delta}_t = \lambda_t \boldsymbol{\Delta}_t \quad \text{and} \quad \gamma_t = \mathbf{g}(\boldsymbol{\theta}_{t+1}) - \mathbf{g}(\boldsymbol{\theta}_t)$$

O algoritmo DFP atualiza W da seguinte forma:

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \frac{\delta_t \delta_t'}{\delta_t' \gamma_t} + \frac{\mathbf{W}_t \gamma_t \gamma_t' \mathbf{W}_t}{\gamma_t' \mathbf{W}_t \gamma_t}$$

BFGS

O método Broyden-Fletcher-Goldfarb-Shanno (BFGS) method envolve a subtração de $v\mathbf{d}\mathbf{d}'$ da atualização do DFP, em que $v = (\gamma_t' \mathbf{W}_t \gamma_t)$ e

$$\boldsymbol{d}_t = \left(\frac{1}{\delta_t' \gamma_t} \right) \delta_t - \left(\frac{1}{\gamma_t' \mathbf{W}_t \gamma_t} \right) \mathbf{W}_t \gamma_t$$

Existe alguma evidência que este método é mais eficiente que o DFP. Qualquer método que tenha a forma:

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{Q}\mathbf{Q}'$$

Vai preservar a "definidade" de $\mathbf{W}$, independentemente do número de colunas em $\mathbf{Q}$.

Gradientes

- Como vocês viram anteriormente, o cálculo dos gradientes de uma função é algo fundamental para a implementação dos algoritmos discutidos anteriormente.
- Existem duas formas de se calcular estes gradientes:
  - Analiticamente: A função objetivo é tal que é possível avaliar não apenas a função mas os seus gradientes em um ponto θ_0 qualquer
  - Numericamente: Podemos aproximar o gradiente no ponto θ_0 por meio de diferenças finitas. Um exemplo é dado pelas diferenças finitas **centradas**, para o i-ésimo elemento de θ

$$\frac{\partial F(\theta)}{\partial \theta_i} = \frac{F(\cdots, \theta_i + \epsilon, \cdots) - F(\cdots, \theta_i - \epsilon, \cdots)}{2\epsilon}$$

Gradientes Numéricos – Problemas

- Problemas com derivadas numéricas:
  - Aumenta o custo computacional: você tem que fazer $2K + 1$ cálculos de funções
  - A determinação de ϵ pode causar problemas (muito grande, a aproximação pode ser ruim, muito pequena pode ter pouca variação para calcular a derivada).
  - Erros de aproximação podem se acumular.
  - Observação: Os Hessianos podem ser calculados numericamente também, mas sofrem dos mesmos problemas nem podemos nos assegurar que ele será negativo definido.
- Valores Iniciais
  - São Importantes para garantir que tenhamos convergência para um ótimo global
  - Importantes em termos de velocidade
  - Recomendação (a) - Tentar vários conjuntos de valores iniciais e (b) Tentar descobrir valores iniciais que façam um sentido inicial

NLLS

- O NLLS é um caso especial de estimador de extremo baseado na seguinte função objetivo:

$$m(\beta, \mathbf{Y}, \mathbf{X}) = \frac{1}{n}[\mathbf{Y} - \mathbf{g}(\beta, \mathbf{X})]'[\mathbf{Y} - \mathbf{g}(\beta, \mathbf{X})]$$

$$\hat{\beta} = \arg\min_{\beta} \left[\frac{1}{n}[\mathbf{Y} - \mathbf{g}(\beta, \mathbf{X})]'[\mathbf{Y} - \mathbf{g}(\beta, \mathbf{X})]\right]$$

- Uma questão importante em econometria é a de **identificação** – definida aqui como os parâmetros do modelo poderem ser determinados de forma única a partir da população observável que gerou os dados.
- Isso não é uma questão da **amostra** que temos, e sim do **modelo** e da **população** que temos.
- Identificação é um passo anterior à análise.

Identificação

- Para que um modelo tenha parâmetros identificados, isso significa que se os valores deles fossem diferentes nós teríamos diferentes distribuições dos dados observáveis.
- Se isso não ocorre (a mesma distribuição dos dados observáveis pode acontecer com diferentes distribuições dos parâmetros), não temos como achar um estimador pra esses parâmetros.
- No contexto de NLLS:
  - Como o termo erro é aditivo, com média zero e variância σ^2, as distribuições dependem somente de $g(\beta, \mathbf{X})$.
  - No entanto, a matriz $\mathbf{X}$ ter posto cheio não é mais suficiente para garantir identificação do modelo.

Propriedades do estimador

- O estimador NLLS é não viesado:

$$\hat{\beta} = \beta_0 + [\mathbf{x}(\beta_\mathbf{0})'\mathbf{x}(\beta_\mathbf{0})]^{-1}\mathbf{x}(\beta_\mathbf{0})'\varepsilon$$

$$\mathbf{x}(\beta_\mathbf{0}) = \left.\frac{\partial \mathbf{g}(\mathbf{x}, \beta_\mathbf{0})}{\partial \beta'}\right|_{\theta_0}$$

E normalmente distribuído:

$$\hat{\beta} \overset{p}{\sim} \mathbb{N}(\beta_\mathbf{0}, \sigma^\mathbf{2}[\mathbf{x}(\beta_\mathbf{0})'\mathbf{x}(\beta_\mathbf{0})]^{-\mathbf{1}})$$

Estimando $Cov(\beta)$

- Para estimarmos a matriz VC do $\hat{\beta}$, precisamos dos resíduos (se a função média condicional é consistente tá OK).
- Uma estimativa consistente de σ^2 é dada por $S^2 = {\hat{\varepsilon}'\hat{\varepsilon}}/{(n-K)}$
- A estimativa consistente de $Cov(\beta)$ é dada então por:

$$\hat{cov}(\hat{\beta}) = S^2 \left[\left. \frac{\partial \mathbf{g}(\mathbf{x}, \beta_{\mathbf{0}})}{\partial \beta'} \right|_{\hat{\beta}} \left. \frac{\partial \mathbf{g}(\mathbf{x}, \beta_{\mathbf{0}})}{\partial \beta'} \right|_{\hat{\beta}}' \right]^{-1}$$

Os termos dentro do colchete são os gradientes da função objetivo em torno do ótimo. Eles saem naturalmente depois que você utilizou um otimizador numérico aqui. Se não utilizou um otimizador numérico, aí você precisa derivar a função na mão e substituir.

Testes de Hipóteses e o Método Delta

- Vamos imaginar que tenhamos um vetor de estimativas de parâmetros, denominado $\hat{\beta}$, e gostaríamos de obter um intervalo de confiança para uma função – não linear – de β, que denominaremos $\gamma = g(\beta)$.
- O valor estimado de γ seria

$$\hat{\gamma} = g(\hat{\beta})$$

Fazendo uma expansão de Taylor em volta do valor "verdadeiro' de β, que denominaremos de β_0, temos:

$$\hat{\gamma} \simeq g(\beta_0) + \bigtriangledown_\beta g(\beta_0)(\hat{\beta} - \beta_0)$$

Uma vez que $g(\beta_0) = \gamma_0$, podemos reorganizar a equação acima:

$$\hat{\gamma} - \gamma_0 = \bigtriangledown_\beta g(\beta_0)(\hat{\beta} - \beta_0)$$

Método Delta (II)

Multiplicando os dois lados por $N^{1/2}$, para assegurar uma velocidade de convergência assintótica adequada, temos que a variância de $\hat{\gamma} - \gamma_0$ é igual à seguinte forma quadrática:

$$(\bigtriangledown_\beta g(\beta_0))V(\beta)(\bigtriangledown_\beta g(\beta_0))^T$$

Esta forma quadrática nos dá a seguinte expressão para a função não linear dos parâmetros:

$$Var(\hat{\gamma}) = (\bigtriangledown_\beta g(\hat{\beta}))V(\beta)(\bigtriangledown_\beta g(\hat{\beta}))^T$$