

# MAT-468: Sesión 5, Ecuaciones no lineales y Optimización

Felipe Osorio

<http://fosorios.mat.utfsm.cl>

Departamento de Matemática, UTFSM



El objetivo de esta sección es resolver problemas del tipo:

- ▶ **Ecuaciones no lineales:** Suponga que estamos interesados en **obtener una raíz**  $\hat{\mathbf{x}}$  para el sistema de  $k$  ecuaciones en  $k$  incógnitas:

$$\mathbf{g}(\mathbf{x}) = \mathbf{0},$$

donde  $\mathbf{g} : \mathbb{R}^k \rightarrow \mathbb{R}^k$ .

- ▶ **Optimización no restringida:** Considere  $\phi : S \rightarrow \mathbb{R}$ , donde  $S \subset \mathbb{R}^k$ . Entonces, deseamos obtener  $\hat{\mathbf{x}}$ , **solución del problema:**

$$\min_{\mathbf{x} \in S} \phi(\mathbf{x}) \quad \text{o} \quad \max_{\mathbf{x} \in S} \phi(\mathbf{x})$$



Sea  $\phi : S \rightarrow \mathbb{R}$  una función real-valuada definida en  $S$  y sea  $\mathbf{c} \in S \subset \mathbb{R}^k$ . Decimos que

- ▶  $\phi$  tiene un **mínimo local** en  $\mathbf{c}$  si existe una bola<sup>1</sup>  $B(\mathbf{c})$  tal que

$$\phi(\mathbf{x}) \geq \phi(\mathbf{c}), \quad \forall \mathbf{x} \in S \cap B(\mathbf{c}).$$

- ▶  $\phi$  tiene un **mínimo local estricto** en  $\mathbf{c}$  si podemos escoger  $B(\mathbf{c})$  tal que

$$\phi(\mathbf{x}) > \phi(\mathbf{c}), \quad \forall \mathbf{x} \in S \cap B(\mathbf{c}), \mathbf{x} \neq \mathbf{c}.$$

- ▶  $\phi$  tiene un **mínimo absoluto** en  $\mathbf{c}$  si

$$\phi(\mathbf{x}) \geq \phi(\mathbf{c}), \quad \forall \mathbf{x} \in S.$$

- ▶  $\phi$  tiene un **mínimo absoluto estricto** en  $\mathbf{c}$  si

$$\phi(\mathbf{x}) > \phi(\mathbf{c}), \quad \forall \mathbf{x} \in S, \mathbf{x} \neq \mathbf{c}.$$

---

<sup>1</sup> $B(\mathbf{c}) = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^k, \|\mathbf{x} - \mathbf{c}\| < r\}$



Sea  $\phi : S \rightarrow \mathbb{R}$  una función real-valuada definida en  $S$  y sea  $\mathbf{c} \in S \subset \mathbb{R}^k$ . Decimos que

- ▶  $\phi$  tiene un **mínimo local** en  $\mathbf{c}$  si existe una bola<sup>1</sup>  $B(\mathbf{c})$  tal que

$$\phi(\mathbf{x}) \geq \phi(\mathbf{c}), \quad \forall \mathbf{x} \in S \cap B(\mathbf{c}).$$

- ▶  $\phi$  tiene un **mínimo local estricto** en  $\mathbf{c}$  si podemos escoger  $B(\mathbf{c})$  tal que

$$\phi(\mathbf{x}) > \phi(\mathbf{c}), \quad \forall \mathbf{x} \in S \cap B(\mathbf{c}), \mathbf{x} \neq \mathbf{c}.$$

- ▶  $\phi$  tiene un **mínimo absoluto** en  $\mathbf{c}$  si

$$\phi(\mathbf{x}) \geq \phi(\mathbf{c}), \quad \forall \mathbf{x} \in S.$$

- ▶  $\phi$  tiene un **mínimo absoluto estricto** en  $\mathbf{c}$  si

$$\phi(\mathbf{x}) > \phi(\mathbf{c}), \quad \forall \mathbf{x} \in S, \mathbf{x} \neq \mathbf{c}.$$

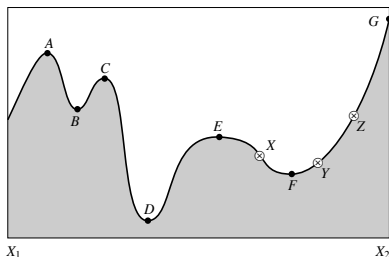
---

<sup>1</sup> $B(\mathbf{c}) = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^k, \|\mathbf{x} - \mathbf{c}\| < r\}$



- ▶ Si  $\phi$  tiene un **mínimo** en  $c$ , entonces  $\psi := -\phi$  tiene un **máximo** en  $c$
- ▶ Si  $c$  está en el interior de  $S$  y  $\phi$  es diferenciable en  $c$ , entonces  $c$  es un **punto crítico** de  $\phi$  si

$$\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}} \Big|_{\mathbf{x}=\mathbf{c}} = \mathbf{0}.$$



- ▶ La **existencia de extremos** de  $\phi$  esta asegurada por el **Teorema de Weierstrass**.
- ▶ **Condición necesaria para un mínimo local**: Sea  $\phi : S \rightarrow \mathbb{R}$  función real-valuada definida en  $S \subset \mathbb{R}^k$  y asuma que  $\phi$  tiene un **mínimo local** en un punto interior  $c \in S$ . Si  $\phi$  es diferenciable en  $c$ , entonces

$$\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}} \Big|_{x=c} = \mathbf{0}.$$

Además, si  $\phi$  es **dos veces diferenciable** en  $c$ , entonces

$$H \phi(c) = \frac{\partial^2 \phi(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} \Big|_{x=c}$$

es matriz **definida positiva**.

**Observación:** Si  $\phi$  tiene un **máximo local**, la condición anterior es substituída por  $H \phi(c) \leq \mathbf{0}$ .



## Ejemplo de ecuaciones no lineales: Método de momentos

El **método de momentos** es basado en resolver el sistema de ecuaciones:

$$E(X^r) = \frac{1}{n} \sum_{i=1}^n x_i^r, \quad r = 1, \dots, k.$$

Por ejemplo, podemos considerar obtener estimadores para  $\alpha$  y  $\beta$  como solución del **sistema  $2 \times 2$** :

$$\begin{aligned}\alpha\beta &= \frac{1}{n} \sum_{i=1}^n x_i, \\ \alpha\beta^2(1 - \alpha) &= \frac{1}{n} \sum_{i=1}^n x_i^2.\end{aligned}$$

Que puede ser planteado como obtener una **raíz de la ecuación** (no lineal)  $g(\theta) = 0$ .



## Método de Newton: Ecuaciones no lineales

Suponga que se desea resolver  $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ . De este modo, suponiendo que  $\mathbf{g}$  es **función suficientemente suave**,

$$\mathbf{g}(\mathbf{x}) = \mathbf{g}(\mathbf{x}^{(0)}) + \dot{\mathbf{g}}(\mathbf{x}^{(0)})(\mathbf{x} - \mathbf{x}^{(0)}) + \mathbf{r}(\mathbf{x} - \mathbf{x}^{(0)}), \quad (1)$$

donde  $\dot{\mathbf{g}}(\mathbf{x}^{(0)}) = \partial \mathbf{g}(\mathbf{x}) / \partial \mathbf{x}^\top |_{\mathbf{x}=\mathbf{x}^{(0)}}$  y  $\mathbf{r}(\mathbf{u})$  debe satisfacer,

$$\lim_{\mathbf{u} \rightarrow \mathbf{0}} \frac{\mathbf{r}(\mathbf{u})}{\|\mathbf{u}\|} = \mathbf{0}.$$

De este modo, basándose en la **aproximación de primer orden** en (1), tenemos

$$\begin{aligned} \mathbf{0} &= \mathbf{g}(\mathbf{x}) \\ &\approx \mathbf{g}(\mathbf{x}^{(0)}) + \dot{\mathbf{g}}(\mathbf{x}^{(0)})(\mathbf{x} - \mathbf{x}^{(0)}), \end{aligned}$$

Es decir,  $\dot{\mathbf{g}}(\mathbf{x}^{(0)})(\mathbf{x} - \mathbf{x}^{(0)}) = -\mathbf{g}(\mathbf{x}^{(0)})$  y, si  $\dot{\mathbf{g}}(\mathbf{x}^{(0)})$  tiene inversa. Entonces,

$$\mathbf{x} = \mathbf{x}^{(0)} - \{\dot{\mathbf{g}}(\mathbf{x}^{(0)})\}^{-1} \mathbf{g}(\mathbf{x}^{(0)}). \quad (2)$$





## Método de Newton: Ecuaciones no lineales

Suponga que se desea resolver  $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ . De este modo, suponiendo que  $\mathbf{g}$  es **función suficientemente suave**,

$$\mathbf{g}(\mathbf{x}) = \mathbf{g}(\mathbf{x}^{(0)}) + \dot{\mathbf{g}}(\mathbf{x}^{(0)})(\mathbf{x} - \mathbf{x}^{(0)}) + \mathbf{r}(\mathbf{x} - \mathbf{x}^{(0)}), \quad (1)$$

donde  $\dot{\mathbf{g}}(\mathbf{x}^{(0)}) = \partial \mathbf{g}(\mathbf{x}) / \partial \mathbf{x}^\top |_{\mathbf{x}=\mathbf{x}^{(0)}}$  y  $\mathbf{r}(\mathbf{u})$  debe satisfacer,

$$\lim_{\mathbf{u} \rightarrow \mathbf{0}} \frac{\mathbf{r}(\mathbf{u})}{\|\mathbf{u}\|} = \mathbf{0}.$$

De este modo, basándose en la **aproximación de primer orden** en (1), tenemos

$$\begin{aligned} \mathbf{0} &= \mathbf{g}(\mathbf{x}) \\ &\approx \mathbf{g}(\mathbf{x}^{(0)}) + \dot{\mathbf{g}}(\mathbf{x}^{(0)})(\mathbf{x} - \mathbf{x}^{(0)}), \end{aligned}$$

Es decir,  $\dot{\mathbf{g}}(\mathbf{x}^{(0)})(\mathbf{x} - \mathbf{x}^{(0)}) = -\mathbf{g}(\mathbf{x}^{(0)})$  y, si  $\dot{\mathbf{g}}(\mathbf{x}^{(0)})$  tiene inversa. Entonces,

$$\mathbf{x} = \mathbf{x}^{(0)} - \{\dot{\mathbf{g}}(\mathbf{x}^{(0)})\}^{-1} \mathbf{g}(\mathbf{x}^{(0)}). \quad (2)$$



## Método de Newton: Ecuaciones no lineales

La ecuación anterior motiva el esquema iterativo:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \{-\dot{\mathbf{g}}(\mathbf{x}^{(k)})\}^{-1} \mathbf{g}(\mathbf{x}^{(k)}). \quad (3)$$

que define el **Método de Newton**.

La convergencia del algoritmo es **local** (ver, por ejemplo, Dennis y Schnabel, 1996). Esto significa que la estimación inicial  $\mathbf{x}^{(0)}$  es cercana a la solución.

La convergencia local requiere de los siguientes supuestos:

- ▶ La ecuación  $\mathbf{g}(\mathbf{x}) = \mathbf{0}$  tiene una solución  $\hat{\mathbf{x}}$ .
- ▶ La función  $\dot{\mathbf{g}}(\cdot)$  es Lipschitz continua cerca de  $\hat{\mathbf{x}}$ .
- ▶  $\dot{\mathbf{g}}(\hat{\mathbf{x}})$  es no singular.

Continuidad Lipschitz cerca de  $\mathbf{x}^*$  indica que existe  $\gamma > 0$  (constante Lipschitz) tal que

$$\|\dot{\mathbf{g}}(\mathbf{x}) - \dot{\mathbf{g}}(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|,$$

para todo  $\mathbf{x}, \mathbf{y}$  suficientemente cerca de  $\mathbf{x}^*$ .



## Método de Newton: Ecuaciones no lineales

La ecuación anterior motiva el esquema iterativo:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \{-\dot{\mathbf{g}}(\mathbf{x}^{(k)})\}^{-1} \mathbf{g}(\mathbf{x}^{(k)}). \quad (3)$$

que define el **Método de Newton**.

La convergencia del algoritmo es **local** (ver, por ejemplo, Dennis y Schnabel, 1996). Esto significa que la estimación inicial  $\mathbf{x}^{(0)}$  es cercana a la solución.

La convergencia local requiere de los siguientes supuestos:

- ▶ La ecuación  $\mathbf{g}(\mathbf{x}) = \mathbf{0}$  tiene una solución  $\hat{\mathbf{x}}$ .
- ▶ La función  $\dot{\mathbf{g}}(\cdot)$  es **Lipschitz continua cerca de  $\hat{\mathbf{x}}$** .
- ▶  $\dot{\mathbf{g}}(\hat{\mathbf{x}})$  es **no singular**.

Continuidad Lipschitz cerca de  $\mathbf{x}^*$  indica que existe  $\gamma > 0$  (constante Lipschitz) tal que

$$\|\dot{\mathbf{g}}(\mathbf{x}) - \dot{\mathbf{g}}(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|,$$

para todo  $\mathbf{x}, \mathbf{y}$  suficientemente cerca de  $\mathbf{x}^*$ .



# Método de Newton: Algoritmo básico

---

**Algoritmo 1:** Método de Newton.

---

**Entrada:** Estimación inicial  $\mathbf{x}^{(0)}$  y nivel de tolerancia  $\tau$

**Salida :** Solución  $\hat{\mathbf{x}}$ .

```
1 begin
2   Hacer  $k \leftarrow 0$ 
3   Evaluar  $\dot{\mathbf{g}}(\mathbf{x}^{(k)})$  y  $\gamma = \|\mathbf{g}(\mathbf{x}^{(k)})\|^2$ 
4   while  $\gamma > \tau$  do
5     Resolver, para  $\mathbf{p}$ , el sistema de ecuaciones:
           
$$\dot{\mathbf{g}}(\mathbf{x}^{(k)})\mathbf{p} = -\mathbf{g}(\mathbf{x}^{(k)})$$

           Actualizar  $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda\mathbf{p}$ 
6      $k \leftarrow k + 1$ 
7   end
8   return  $\hat{\mathbf{x}} = \mathbf{x}^{(k+1)}$ 
9 end
```

---



- ▶ El paso Newton  $\mathbf{p}$  usualmente es calculado mediante  $\dot{\mathbf{g}}(\mathbf{x}^{(k)}) = \mathbf{LU}$  y resolver

$$\mathbf{LU}\mathbf{p} = -\mathbf{g}(\mathbf{x}^{(k)})$$

- ▶ Note que podemos **aproximar la matriz Jacobiana**  $\dot{\mathbf{g}}(\mathbf{x}) = \partial\mathbf{g}(\mathbf{x})/\partial\mathbf{x}^\top$  por **diferencias divididas**. En efecto la  $j$ -ésima columna de  $\dot{\mathbf{g}}(\mathbf{x})$  es dada por:

$$\dot{\mathbf{g}}_j(\mathbf{x}) \approx \frac{\mathbf{g}(\mathbf{x} + h\mathbf{e}_j) - \mathbf{g}(\mathbf{x})}{h},$$

para  $h$  “pequeño” y  $\mathbf{e}_j$  un vector unidad.



## Método de Broyden: Ecuaciones no lineales

Para resolver la ecuación

$$g(x) = 0,$$

Podemos usar un algoritmo Newton, cuya derivada  $g'(x)$  en  $x^{(k)}$  es **aproximada por la diferencia**

$$b_k = \frac{g(x^{(k)}) - g(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}, \quad (4)$$

y entonces hacer la iteración

$$x^{(k+1)} = x^{(k)} - b_k^{-1} g(x^{(k)}).$$

Extensiones para **dimensiones mayores** debe imitar la aproximación en (4).



## Método de Broyden: Ecuaciones no lineales

El método de Broyden se basa en obtener una matriz  $\mathbf{B}_k$  tal que satisfaga la **ecuación secante**:

$$\mathbf{B}_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) = \mathbf{g}(\mathbf{x}^{(k)}) - \mathbf{g}(\mathbf{x}^{(k-1)}).$$

Considere  $\mathbf{x}^{(k)}$  y  $\mathbf{B}_k$  **aproximaciones** actuales de la **solución** y **matriz Jacobiana**, respectivamente, entonces

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda_k \{-\mathbf{B}_k^{-1} \mathbf{g}(\mathbf{x}^{(k)})\},$$

donde  $\lambda_k$  es un largo de paso y  $\mathbf{p}_k = -\mathbf{B}_k^{-1} \mathbf{g}(\mathbf{x}^{(k)})$  es la **dirección Broyden**.



## Método de Broyden: Ecuaciones no lineales

Luego del cálculo de  $\mathbf{x}^{(k+1)}$ ,  $\mathbf{B}_k$  es actualizada a  $\mathbf{B}_{k+1}$  usando la **actualización Broyden**

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{(\mathbf{y}_k - \mathbf{B}_k \mathbf{s}_k) \mathbf{s}_k^\top}{\mathbf{s}_k^\top \mathbf{s}_k},$$

donde

$$\mathbf{y}_k = \mathbf{g}(\mathbf{x}^{(k+1)}) - \mathbf{g}(\mathbf{x}^{(k)}), \quad \mathbf{s}_k = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}.$$

Note además que para una **implementación eficiente** del procedimiento podemos obtener  $\mathbf{B}_{k+1}^{-1}$  usando la fórmula de **Sherman-Morrison**.





- ▶ **Estimación máximo verosímil:** Considere el modelo estadístico:

$$\mathcal{P} = \{P_{\theta} : \theta \in \Theta \subset \mathbb{R}^p\},$$

tal que, tenemos la función de densidad conjunta  $f(\mathbf{y}; \theta)$ . Entonces, se define la **función de log-verosimilitud** como

$$\ell_n(\theta) = \log f(\mathbf{y}; \theta).$$

El objetivo es obtener el estimador ML como **solución del problema:**

$$\max_{\theta \in \Theta} \ell_n(\theta),$$

y denotamos a este estimador como  $\hat{\theta}$ .



Es habitual realizar los siguientes supuestos:

A1 (compacidad) Es espacio paramétrico  $\Theta \in \mathbb{R}^p$  es cerrado y acotado.

A2 (identificabilidad) El parámetro  $\theta$  es identificado, es decir para  $\theta_1 \neq \theta_2$ , se tiene

$$f(\mathbf{y}; \theta_1) \neq f(\mathbf{y}; \theta_2).$$

A3 (acotamiento) Sea  $\theta_*$  el valor verdadero para  $\theta$ . Entonces  $E_{\theta_*} \{|\log f(\mathbf{y}; \theta)|\} < \infty$  y para todo  $\theta$ ,

$$E_{\theta_*} \{(\log f(\mathbf{y}; \theta))_+\} < \infty.$$

A4 (continuidad) La densidad es continua en  $\theta$ , es decir

$$\lim_{\theta_k \rightarrow \theta} f(\mathbf{y}; \theta_k) = f(\mathbf{y}; \theta)$$



## Resultado 1 (consistencia)

Bajo los supuestos A1-A4, el MLE  $\hat{\theta}_n$  converge a  $\theta_*$ .

## Resultado 2

Bajo condiciones apropiadas

$$\frac{1}{\sqrt{n}}\mathbf{U}(\theta_*) \xrightarrow{D} N_p(\mathbf{0}, \mathcal{F}(\theta_*))$$

donde  $\mathbf{U}(\theta) = \partial \ell_n(\theta) / \partial \theta$  y la matriz de covarianza

$$\mathcal{F}(\theta) = \text{Cov}(\mathbf{U}(\theta))$$



Considere

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(\mathbf{y}_i; \boldsymbol{\theta}).$$

y de este modo, tenemos el siguiente resultado.

## Resultado 3

Sea  $H_n(\boldsymbol{\theta}) = \ddot{\ell}_n(\boldsymbol{\theta})$  ( $= \partial^2 \ell(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top$ ). Entonces

$$\frac{1}{n} H_n(\boldsymbol{\theta}) \xrightarrow{\text{a.s.}} E(H(\boldsymbol{\theta})),$$

y  $E(-H(\boldsymbol{\theta})) = \mathcal{F}(\boldsymbol{\theta})$ , donde

$$\mathcal{F}(\boldsymbol{\theta}) = E \left\{ - \frac{\partial^2 \log f(\mathbf{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right\}.$$



## Resultado 4 (normalidad asintótica)

Bajo condiciones apropiadas, tenemos

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) \xrightarrow{D} N_p(\mathbf{0}, \mathcal{F}^{-1}(\boldsymbol{\theta}_*)).$$

## Resultado 5

Asintóticamente, para  $\boldsymbol{\theta}$  en una vecindad de  $\hat{\boldsymbol{\theta}}_n$ , tenemos que la log-verosimilitud es cuadrática.

Note que

$$\frac{1}{n} \ell_n(\boldsymbol{\theta}) = \frac{1}{n} \ell_n(\hat{\boldsymbol{\theta}}_n) + \frac{1}{n} \dot{\ell}_n^\top(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)^\top \left\{ \frac{1}{n} \ddot{\ell}_n(\boldsymbol{z}) \right\} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n),$$

para  $\boldsymbol{z}$  en la misma vecindad, y como  $\dot{\ell}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{U}(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$  el resultado sigue.



## Algoritmo Newton-Raphson

La etapa  $(k+1)$ -ésima del algoritmo es dada por

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \lambda_k \{-\ddot{\ell}(\boldsymbol{\theta}^{(k)})\}^{-1} \mathbf{U}(\boldsymbol{\theta}^{(k)}).$$

## Algoritmo Fisher-scoring

El procedimiento iterativo es definido mediante:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \lambda_k \mathcal{F}^{-1}(\boldsymbol{\theta}^{(k)}) \mathbf{U}(\boldsymbol{\theta}^{(k)}).$$



### Algoritmo BHHH

Recuerde que  $\ell_n(\boldsymbol{\theta}) = \sum_i \log f(\mathbf{y}_i; \boldsymbol{\theta})$ . De este modo,

$$\dot{\ell}_n(\boldsymbol{\theta}) = \mathbf{U}_n(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{U}_i(\boldsymbol{\theta}),$$

donde  $\mathbf{U}_i(\boldsymbol{\theta}) = \partial \log f(\mathbf{y}_i; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ . Así, podemos usar la aproximación

$$\frac{1}{n} \ddot{\ell}_n(\boldsymbol{\theta}) \approx \frac{1}{n} \sum_{i=1}^n \mathbf{U}_i(\boldsymbol{\theta}) \mathbf{U}_i^\top(\boldsymbol{\theta}).$$

Lo que lleva al algoritmo de Berndt-Hall-Hall-Hausman:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \lambda_k \left\{ \frac{1}{n} \sum_{i=1}^n \mathbf{U}_i(\boldsymbol{\theta}^{(k)}) \mathbf{U}_i^\top(\boldsymbol{\theta}^{(k)}) \right\}^{-1} \mathbf{U}_n(\boldsymbol{\theta}^{(k)}).$$

