CIMPA-UCR

Optimización Combinatoria en Problemas de Regresión

- Regresión no lineal
  - Uso sobrecalentamiento simulado
  - Uso de búsqueda tabú
  - Aplicación en finanzas
- Selección de variables en regresión lineal
  - El problema
  - Algoritmo genético

Regresión

- Modelo estadístico explicativo:

$$y \leftarrow x$$

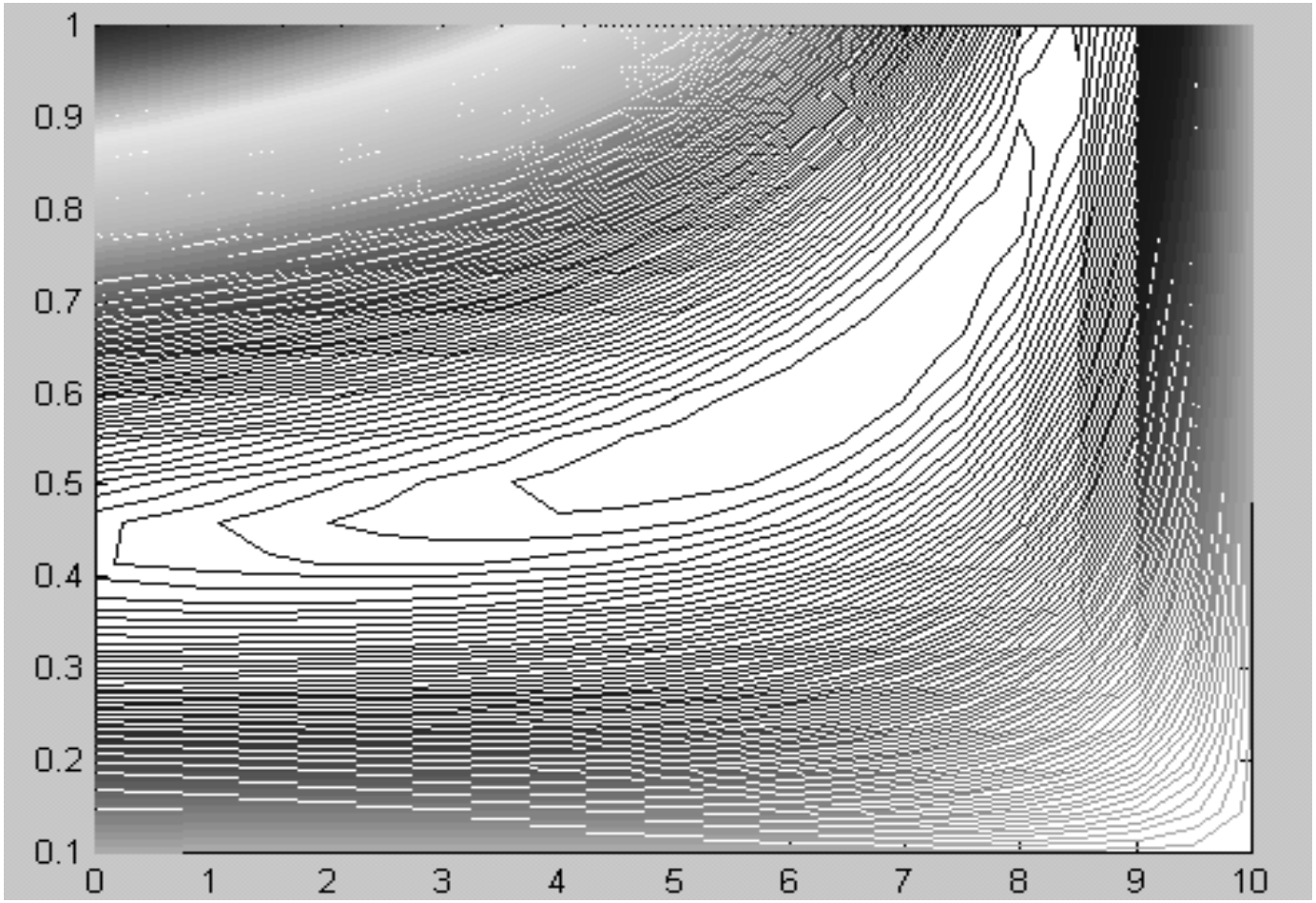
$$y = f(x) + \varepsilon$$

$$y = f_\theta(x) + \varepsilon$$

- Criterio de mínimos cuadrados:

$$S(\theta) = \sum_{i=1}^{n} [y_i - f_\theta(x_i)]^2$$

Geometría del espacio de parámetros



Gauss-Newton

- Aproximación de $f_\theta(x)$ por un polinomio de Taylor de 1er orden alrededor de θ_0

$$f_{\vec{\theta}}(x_i) = f_{\vec{\theta}^0}(x_i) + \sum_{j=1}^{p} \left[\frac{\partial f_{\vec{\theta}}(x_i)}{\partial \theta_j} \right]_{\vec{\theta}=\vec{\theta}^0} (\theta_j - \theta_j^0)$$

- Se escribe $y - f^0 = \sum_{j=1}^{p} \beta_j^0 z_j^0 + \varepsilon$
- Se usa regresión lineal múltiple

Descenso de Gradiente

- Busca la dirección de máximo descenso en cada punto de la iteración
- Se debe mover una estimación de θ en la dirección de

$$\left(-\frac{\partial E(\hat{\theta})}{\partial \theta_1}, \ldots, -\frac{\partial E(\hat{\theta})}{\partial \theta_p} \right)$$

- Teóricamente converge, pero la convergencia puede ser muy lenta

Método de Marquardt

Hace una interpolación entre las direcciones del método de Gauss-Newton y de descenso de gradiente (empíricamente se ha observado que ambas son casi ortogonales)

Observaciones

- Se itera hasta converger a un valor estable
- La convergencia no está garantizada (Draper & Smith)
- Puede converger a un óptimo local pues la búsqueda es en los contornos elipsoidales de los puntos del proceso iterativo

Mínimos Locales

- Los métodos de descenso o de búsqueda local (Gauss-Newton, descenso de gradiente, Marquardt) pueden conducir a mínimos locales del criterio
- Se puede pensar en usar metaheurísticas (sobrecalentamiento simulado, búsqueda tabú algoritmos genéticos, etc.) para evitar los mínimos locales

Un ejemplo simple

Datos:

x	-2.5	-1	1	2
y	1	1.1	-1.1	0.2

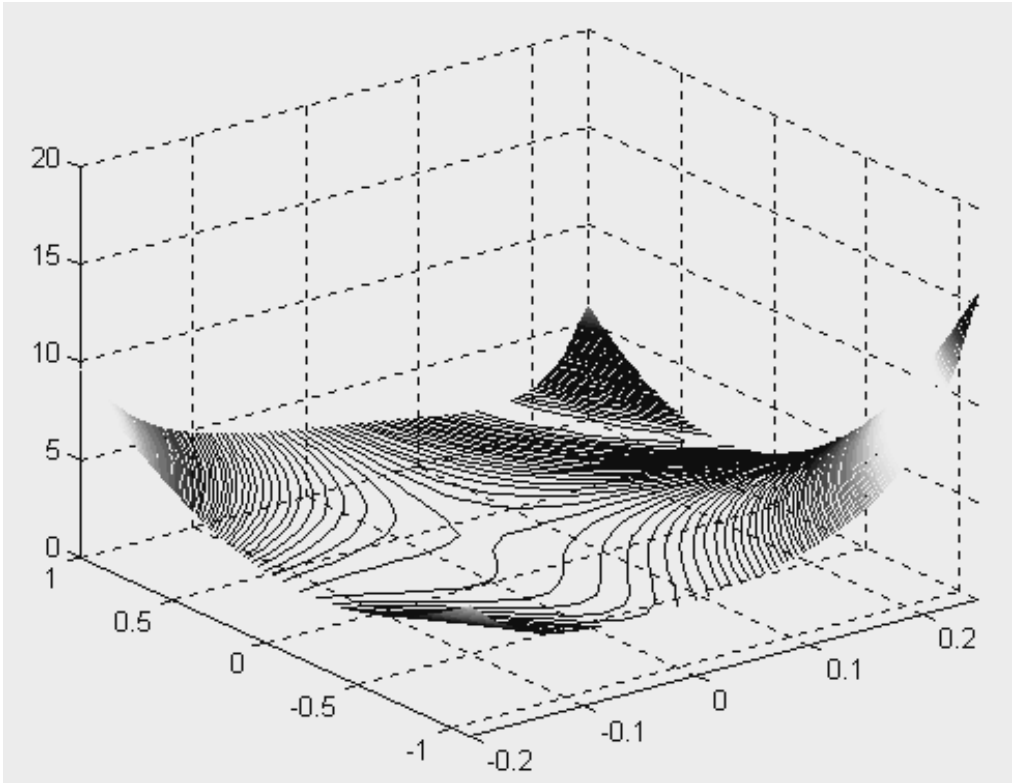
$$\hat{y} = \theta_1 e^{-\theta_2 x}$$

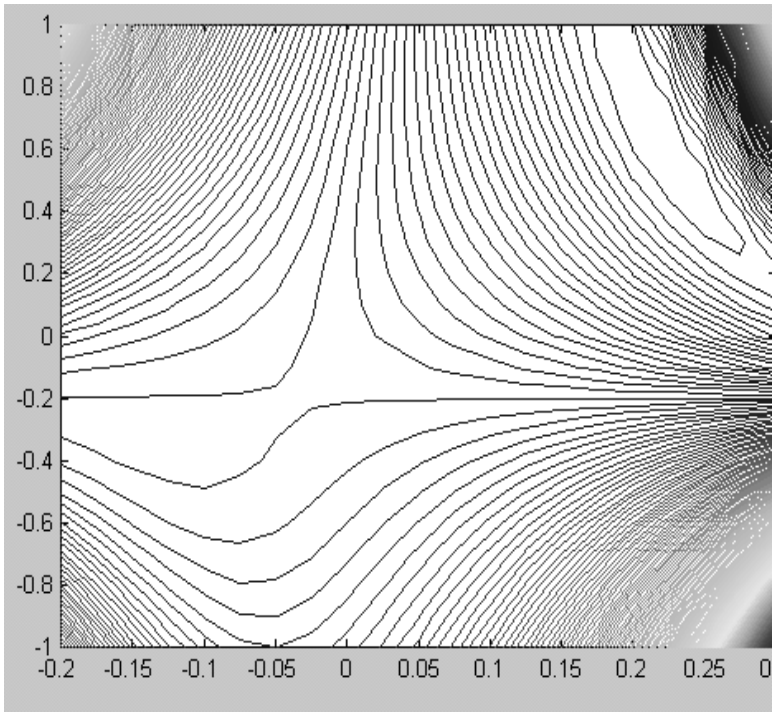
Dos óptimos locales:

$\theta^* = (0.669, 0.214)$ $\quad S(\theta^*) = 1.968$

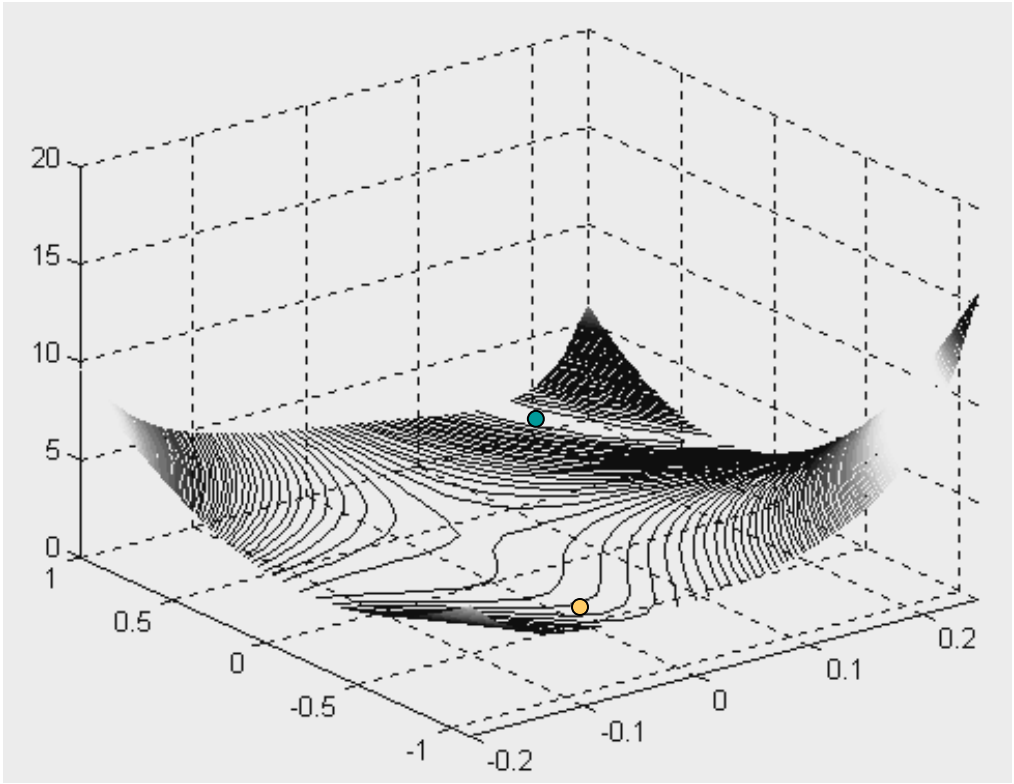
$\theta^{**} = (-0.764, -0.0298)$ $\quad S(\theta^{**}) = 3.436$

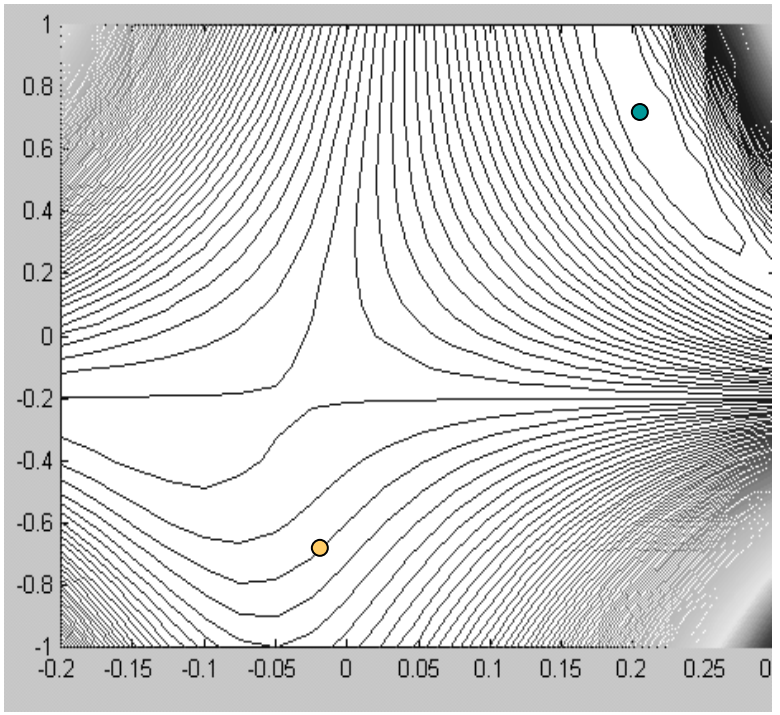
Ilustración





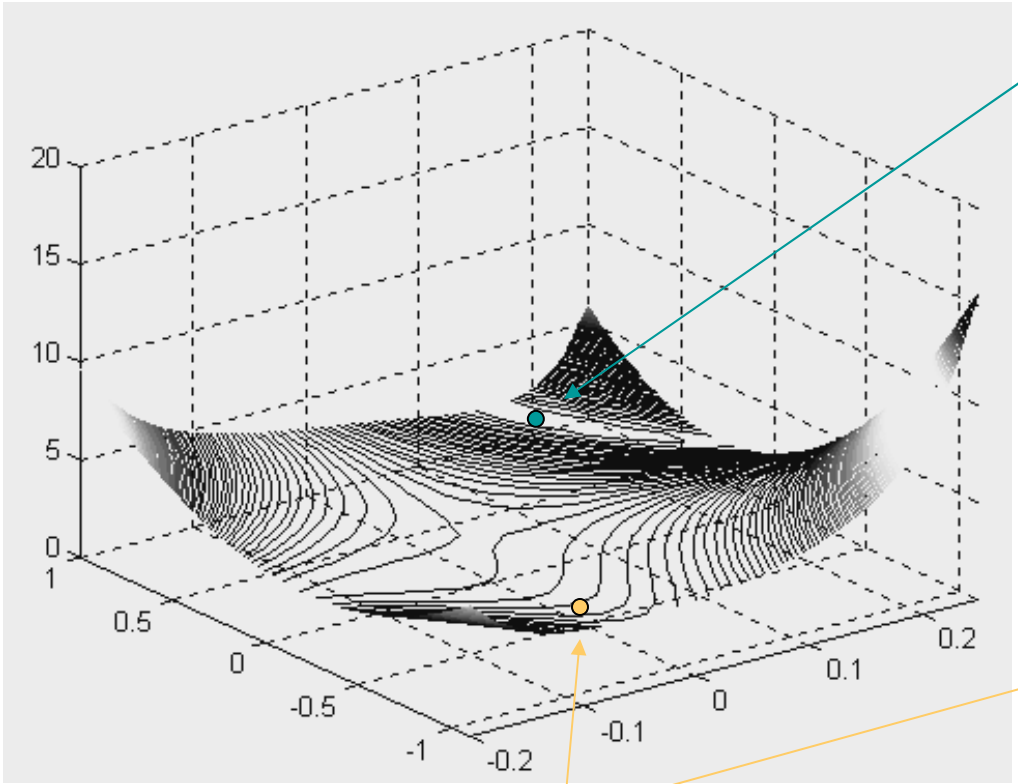
Ilustración

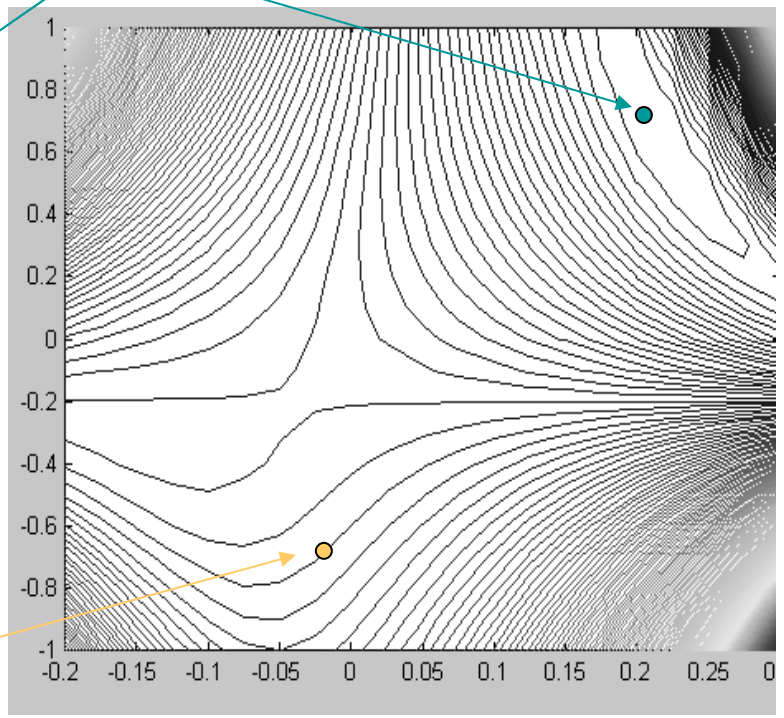




Ilustración

Optimo global: 1.968

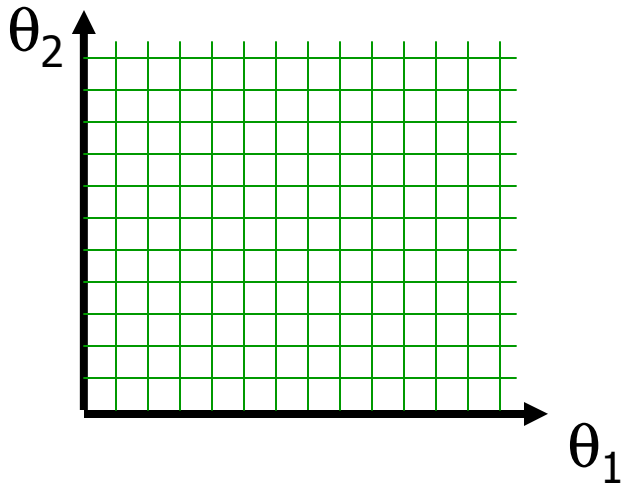




Optimo local: 3.436

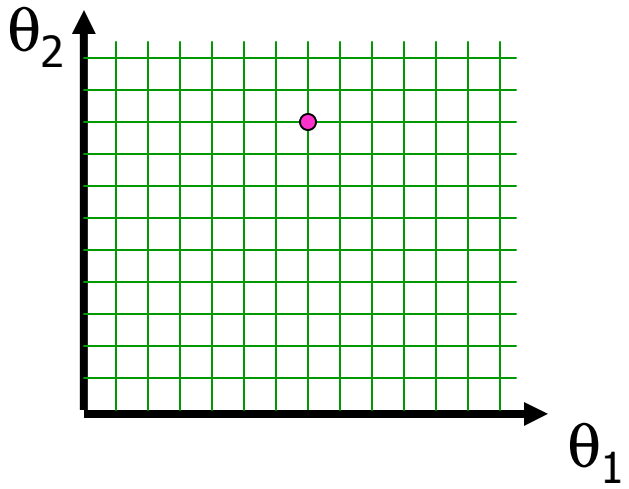
Aplicación de SS & BT

- Hacer un mallado



Aplicación de SS & BT

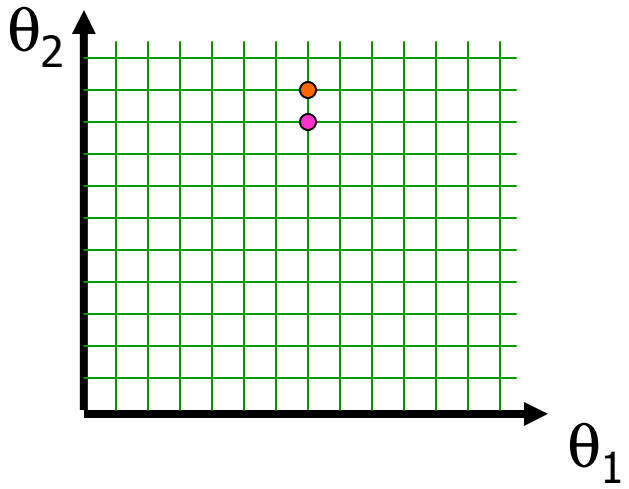
- Hacer un mallado



Un punto en la malla: una solución del problema

Aplicación de SS & BT

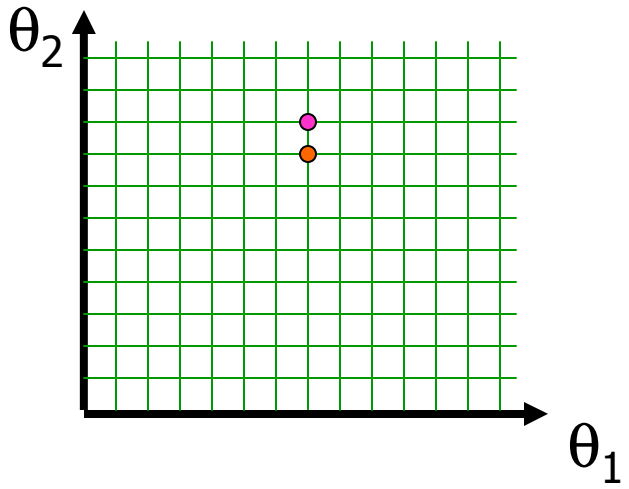
- Hacer un mallado



Mover al norte

Aplicación de SS & BT

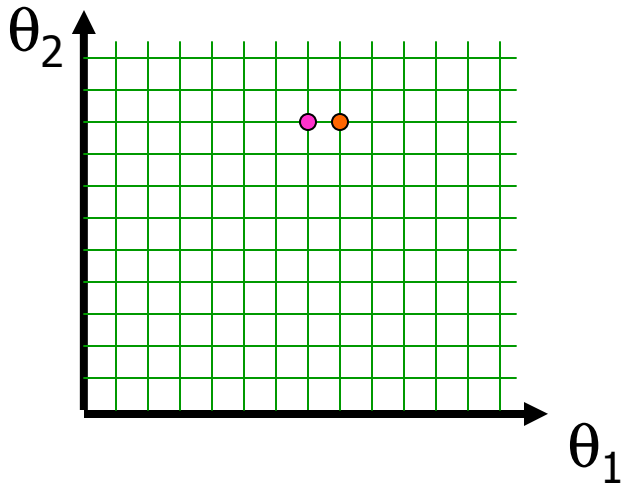
- Hacer un mallado



Mover al sur

Aplicación de SS & BT

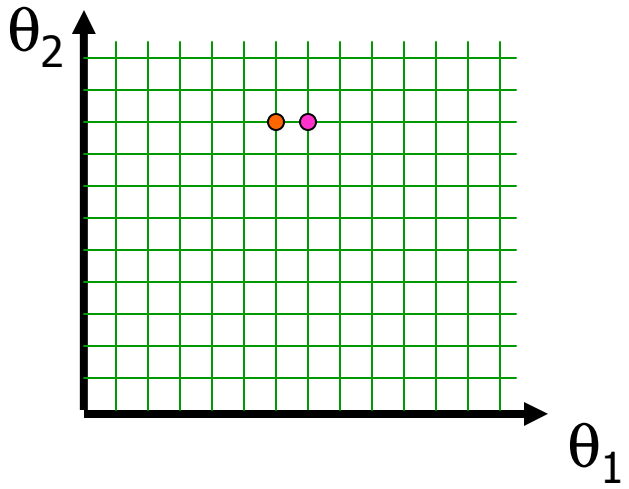
- Hacer un mallado



Mover al este

Aplicación de SS & BT

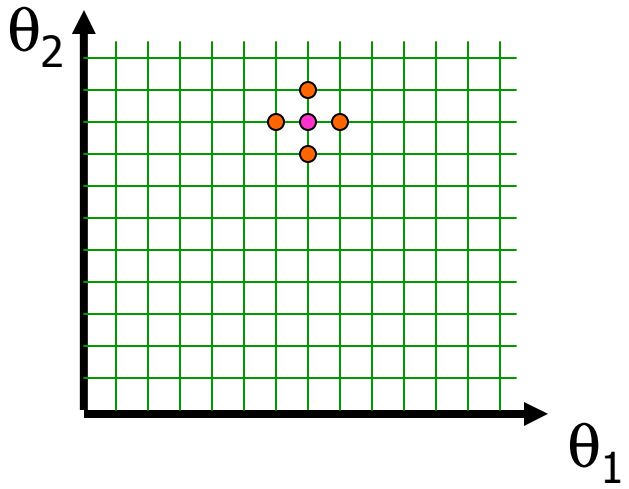
- Hacer un mallado



Mover al oeste

Aplicación de SS & BT

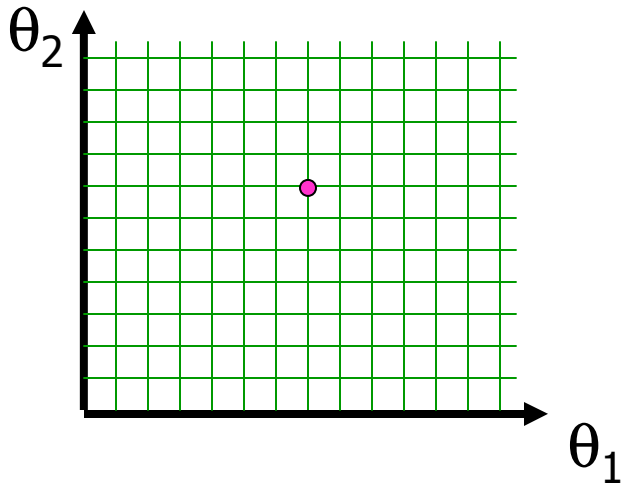
- Hacer un mallado



Vecindario de tamaño 4 (=2*p*)

Geometría del espacio de parámetros

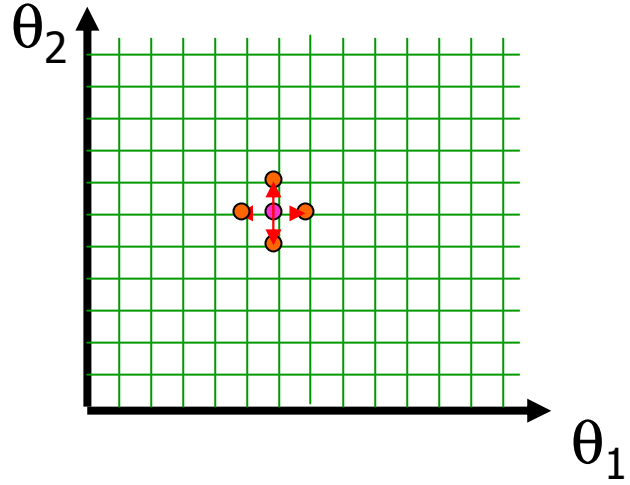
- Espacio con tantas dimensiones como parámetros tenga el modelo



Un punto del espacio: una solución del problema

Geometría del espacio de parámetros

- Movimiento

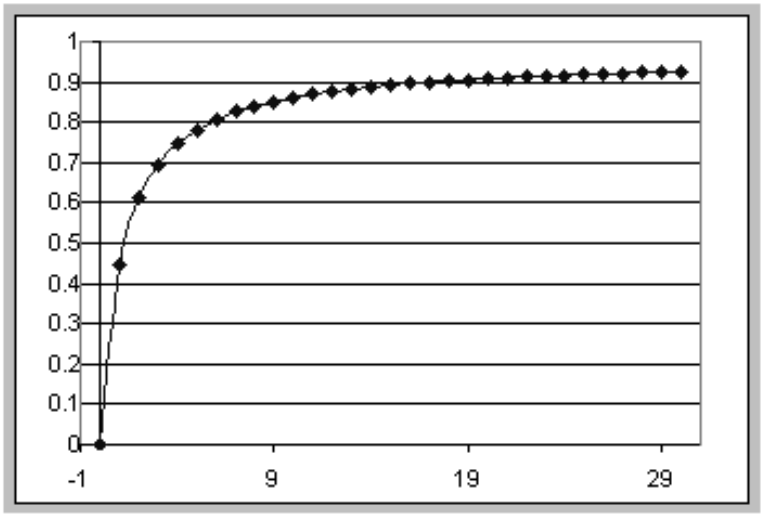


- Uso de la regla de Metropolis o una lista tabú de movimientos

Modelo Michaelis-Menten

x	0.3	0.5	1	2	3	4	5	10
y	0.17	0.27	0.43	0.65	0.73	0.78	0.79	0.83

$$\hat{y} = \frac{\theta_1 x}{x + \theta_2}$$

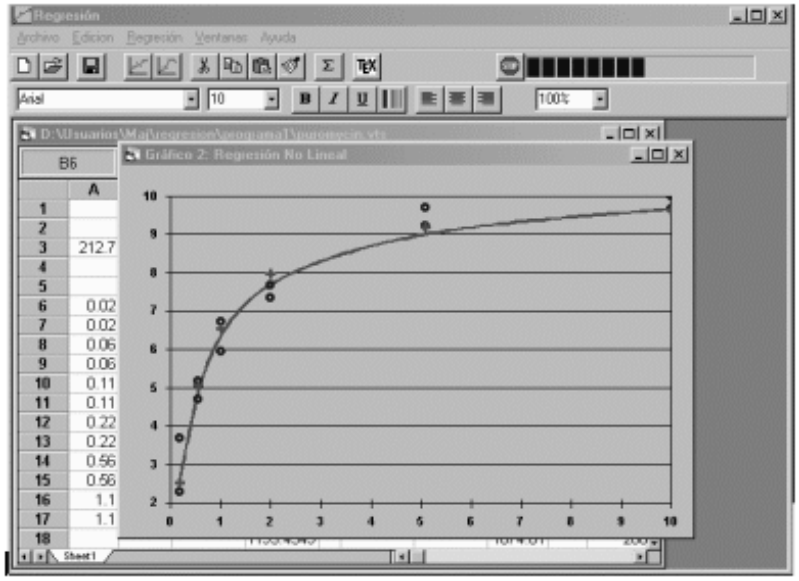


Metodo	θ_1	θ_2	S(θ)
G-N	0.96	1.14	0.008256
SS	0.96	1.14	0.008256
BT	0.96	1.12	0.008325
$1/y = g(1/x)$	0.92	1.58	0.066007
$x/y = g(x)$	0.92	1.00	0.010154
$y = g(y/x)$	1.01	1.37	0.010952
Semi-parm.	1.00	1.34	0.010419

Datos Puromycin

x	.02	.06	.11	.22	.56	1.1
y	76 47	97 107	123 139	159 152	191 201	207 200

$$\hat{y} = \frac{\theta_1 x}{x + \theta_2}$$



Metodo	θ_1	θ_2	S(θ)
G-N	212.7	0.0641	1195.455
SS	212.69	0.0641	1195.448
BT	212.76	0.0638	1196.097
Linealiz.			1920.640

Nuestros resultados en RNL

- No obtenemos mejores resultados que el método de Gauss-Newton sino prácticamente iguales, pero sí mejoramos a las linealizaciones.
- Esto es cierto tanto en regresión **simple** como **múltiple**.
- Pero, no es necesario conocer ni calcular **derivadas**
- Se puede pasar a cualquier **norma** fácilmente

Aplicación de la RNL

- Estimación de la curva soberana o vector de precios en el mercado de valores

La Curva Soberana

Construir e implementar una curva de rendimientos diaria cero cupón para el mercado financiero de Costa Rica (vector de precios), con las siguientes características:

a. La metodología tiene que ser transparente y objetiva: esto es, fácil de verificar y fácil de implementar, y además replicable.

b. La curva de rendimientos sería adoptada por la Bolsa Nacional de Valores de Costa Rica, como un instrumento oficial de referencia para todas las entidades financieras del país.

Bonos Considerados

1. En Colones: instrumentos cero cupón emitidos por el Ministerio de Hacienda (G) y el Banco Central de Costa Rica (BCCR)

- G-tp0 - títulos de propiedad con cero-cupón
- BCCR-bem0 - bonos de estabilización monetaria cero cupón.
- G-tp - títulos de propiedad tasa fija (bullet).
- BCCR-bem - bonos de estabilización monetaria tasa fija (bullet).

2. En Dólares: idem:

- BCCR-cd$ - certificados de depósito tasa fija en dólares (bullet).
- G-tp$ - titulos de propiedad tasa fija en dólares (bullet).

Notación

Si el instrumento *k* paga *r* en cupones en los instantes

$$0 \leq t_1 < \ldots < t_n$$

y el principal principal *p* es pagado en el instante t_n, entonces el valor teórico (precio) del instrumento es

$$V_k = r \sum_{i=1}^{n-1} e^{z(t_i)t_i} + (r+p)e^{z(t_n)t_n}$$

donde $z(t)$ es la curva de rendimientos cero cupón que debemos estimar.

Objetivo General

Encontrar la curva de rendimientos $z(t)$ que hace que los valores teóricos V_k estén lo más cerca posible de los valores observados en el mercado financiero,

$$\widehat{V}_k$$

La relación entre $z(t)$ y la función de la tasa *forward* instantánea $f^i(t)$ es:

$$z(t) = \frac{\int_0^t f^i(\tau)\, d\tau}{t}$$

Modelos

(Estudio actual)

- Nelson-Siegel
- Svensson

(Trabajo futuro en otros proyectos)

- Merrill Lynch Exponential Splines (MLES)
- ACP estimation
- Hull & White Trinomial tree
- otros

Uso de los Modelos

Nelson & Siegel	Svensson	Splines
Bélgica	Alemania	Colombia
Colombia	Bélgica	Estados Unidos
España	Canadá	Japón
Estados Unidos	España	Noruega
Francia	Estados Unidos	Reino Unido
Italia	Finlandia	Banco Mundial
Reino Unido	Francia	
	Noruega	
	Suecia	
	Suiza	
	Reino Unido	

Fuente: Filipovic (1999), Vázquez y Melo, Banco Mundial.

Modelo de Nelson-Siegel

Tasa instantánea forward:

$$f^i(t) = \beta_0 + \beta_1(e^{-\lambda_1 t}) + \beta_2(\lambda_1 t\, e^{-\lambda_1 t})$$

donde:

- t : tiempo para maduración, $t = \tau - s$
- 4 parámetros: $\beta_0,\ \beta_1,\ \beta_2,\ \lambda_1$

Esto implica que

$$z(t) = \beta_0 + \beta_1\left(\frac{1 - e^{-\lambda_1 t}}{\lambda_1 t}\right) + \beta_2\left(\frac{1 - e^{-\lambda_1 t}}{\lambda_1 t} - e^{-\lambda_1 t}\right)$$

Modelo de Svensson

Forward instant rate:

$$f^i(t) = \beta_0 + \beta_1(e^{-\lambda_1 t}) + \beta_2(\lambda_1 t\, e^{-\lambda_1 t}) + \beta_3(\lambda_2 t\, e^{-\lambda_2 t})$$

donde: $\lambda_1, \lambda_2 > 0$, y

- t : tiempa para la maduración, $t = \tau - s$
- 6 parameters: $\beta_0, \beta_1, \beta_2, \beta_3, \lambda_1, \lambda_2$

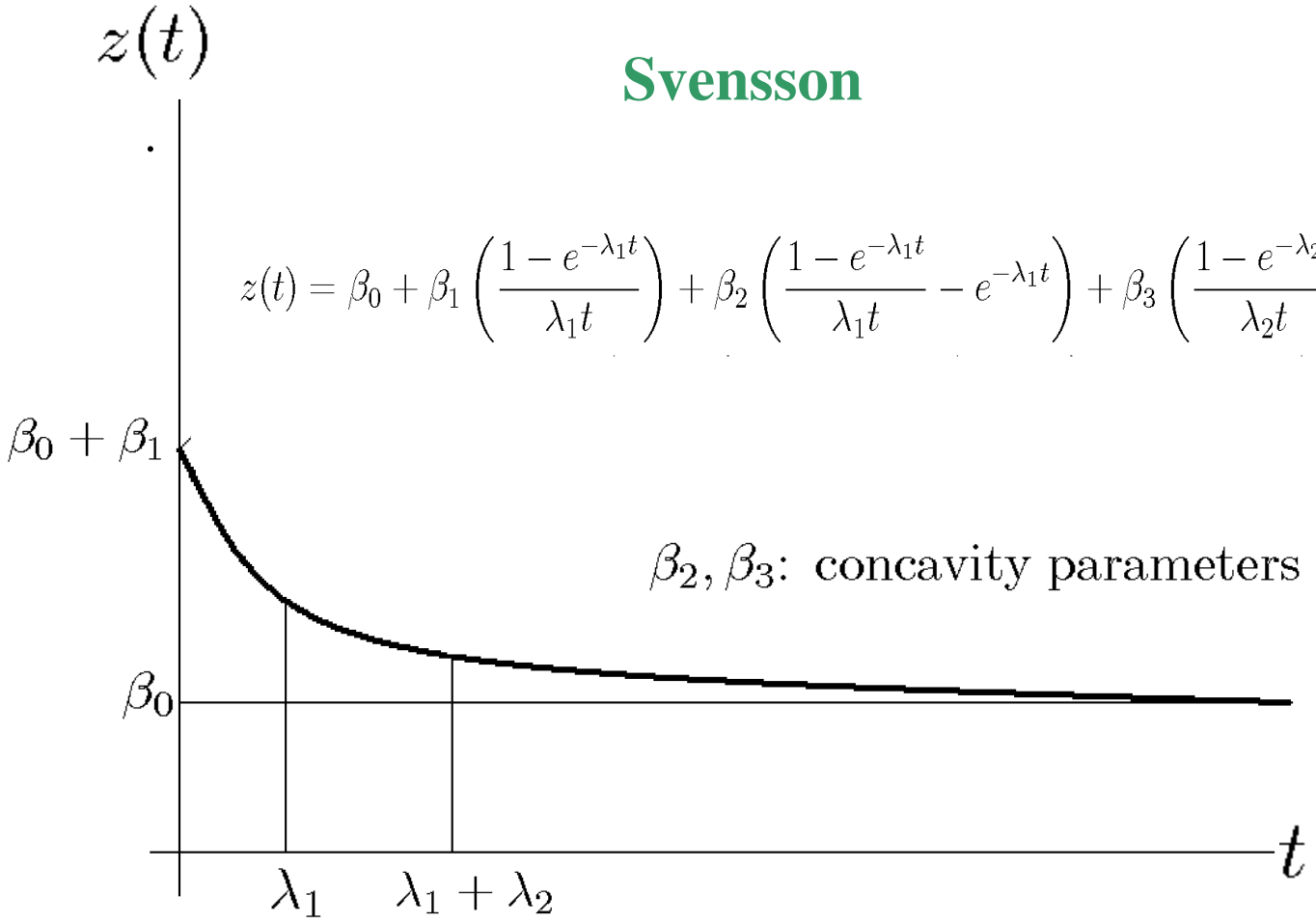
- Esto implica:

$$z(t) = \beta_0 + \beta_1\left(\frac{1-e^{-\lambda_1 t}}{\lambda_1 t}\right) + \beta_2\left(\frac{1-e^{-\lambda_1 t}}{\lambda_1 t} - e^{-\lambda_1 t}\right) + \beta_3\left(\frac{1-e^{-\lambda_2 t}}{\lambda_2 t} - e^{-\lambda_2 t}\right)$$

Interpretación de Parámetros

Svensson

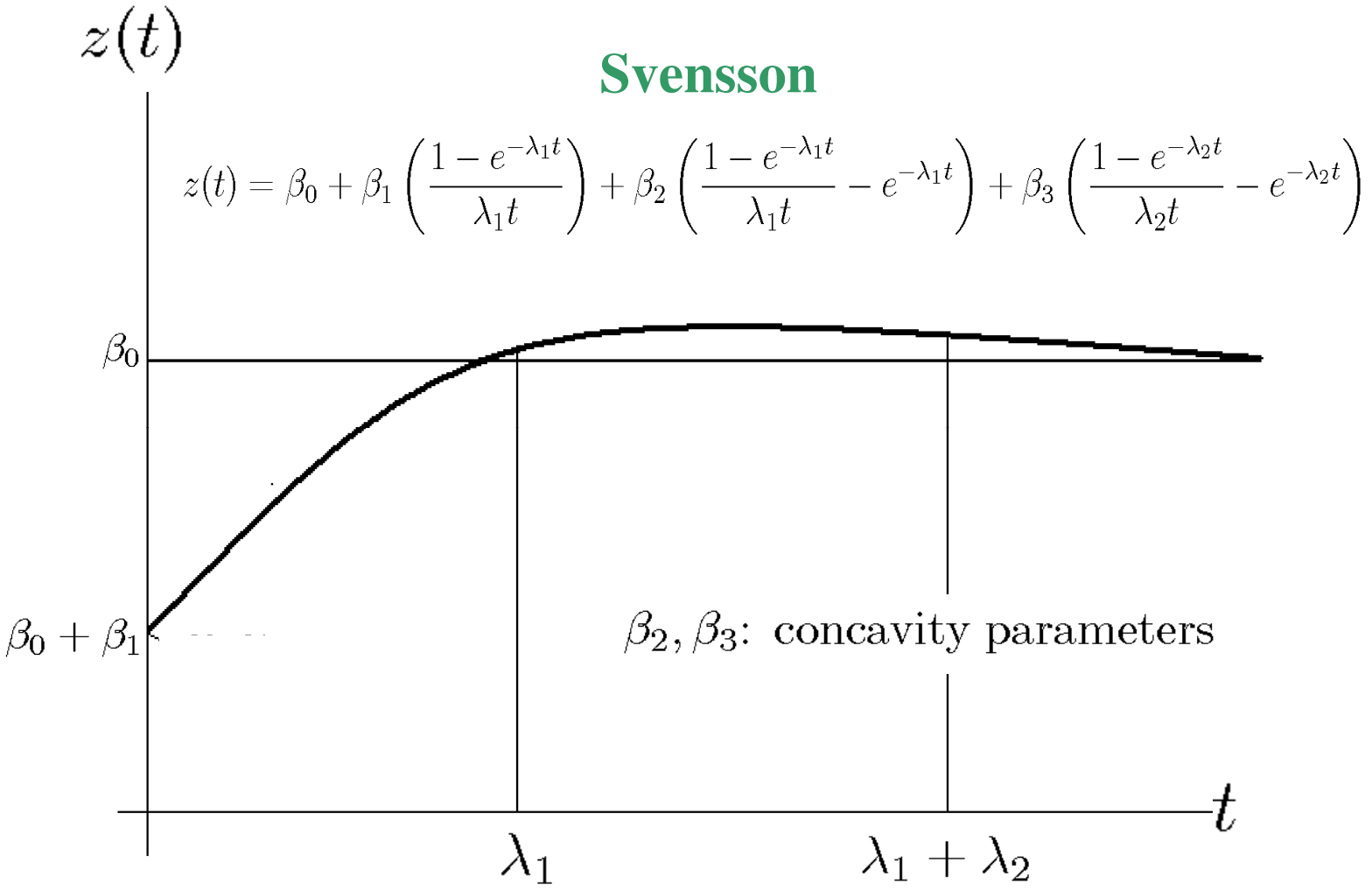
$$z(t) = \beta_0 + \beta_1 \left(\frac{1 - e^{-\lambda_1 t}}{\lambda_1 t} \right) + \beta_2 \left(\frac{1 - e^{-\lambda_1 t}}{\lambda_1 t} - e^{-\lambda_1 t} \right) + \beta_3 \left(\frac{1 - e^{-\lambda_2 t}}{\lambda_2 t} - e^{-\lambda_2 t} \right)$$



Interpretación de Parámetros

Svensson

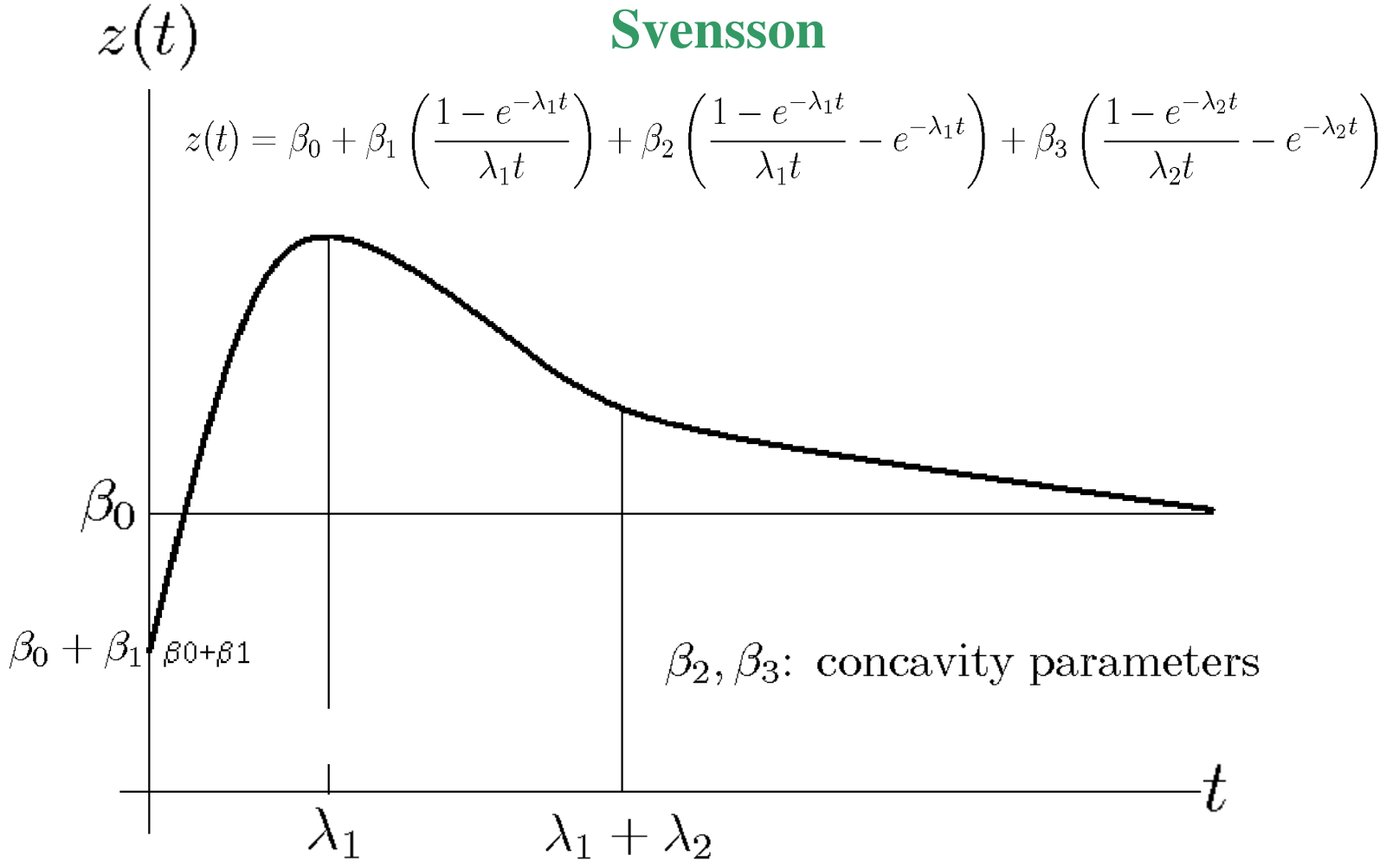
$$z(t) = \beta_0 + \beta_1 \left(\frac{1 - e^{-\lambda_1 t}}{\lambda_1 t} \right) + \beta_2 \left(\frac{1 - e^{-\lambda_1 t}}{\lambda_1 t} - e^{-\lambda_1 t} \right) + \beta_3 \left(\frac{1 - e^{-\lambda_2 t}}{\lambda_2 t} - e^{-\lambda_2 t} \right)$$



Interpretación de Parámetros

Svensson

$$z(t) = \beta_0 + \beta_1 \left(\frac{1 - e^{-\lambda_1 t}}{\lambda_1 t} \right) + \beta_2 \left(\frac{1 - e^{-\lambda_1 t}}{\lambda_1 t} - e^{-\lambda_1 t} \right) + \beta_3 \left(\frac{1 - e^{-\lambda_2 t}}{\lambda_2 t} - e^{-\lambda_2 t} \right)$$



Criterios de Optimización

- Mínimos cuadrados:

$$E^2(V,\hat{V}) = \sum_{k=1}^{K}(V_k - \hat{V}_k)^2.$$

- Mínimos cuadrados ponderados

$$E^2(V,\hat{V}) = \sum_{k=1}^{K}\frac{(V_k - \hat{V}_k)^2}{B_k\,(1+T(k))}$$

$$,\quad E^2(V_k,\widehat{V}_k) = \sum_{k=1}^{K}\frac{(V_k - \widehat{V}_k)^2}{D_k}$$

Resultados

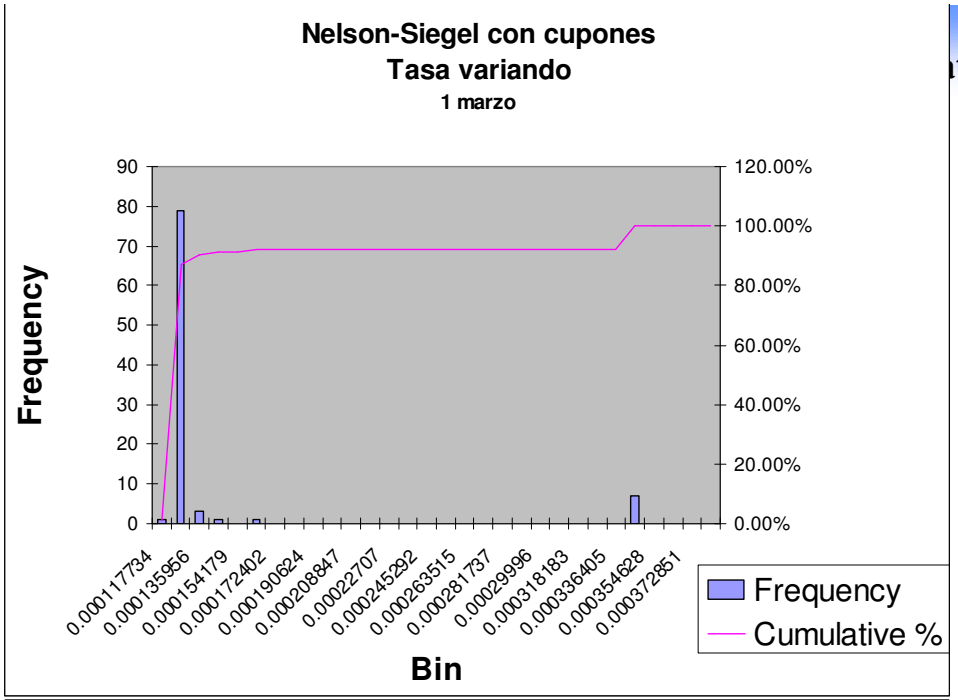
1 de marzo

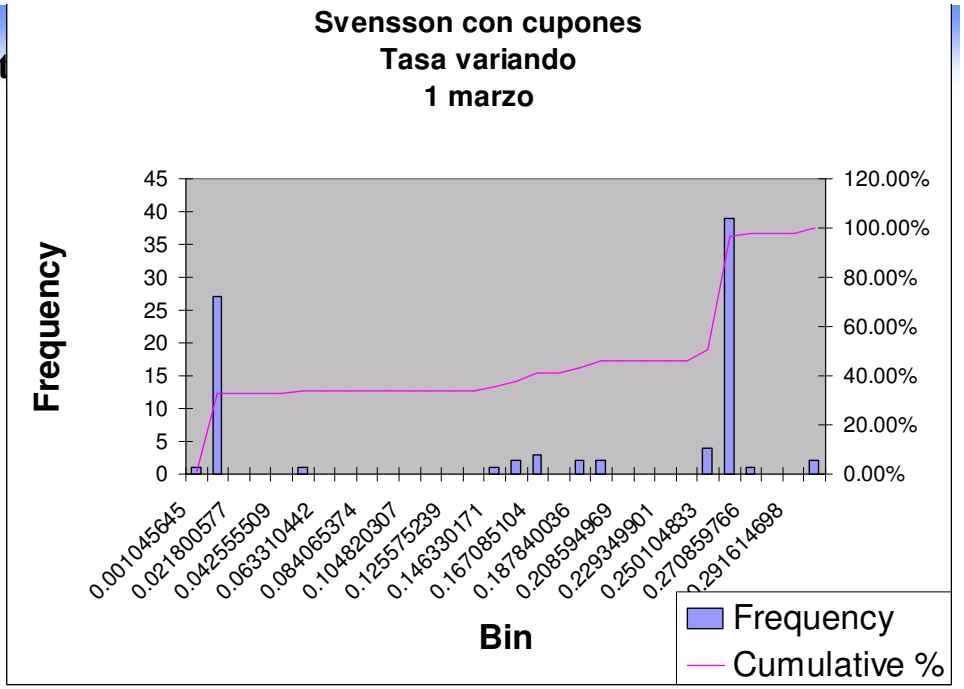
Tasa fija	NS		NSS	
	Con cupón	Sin cupón	Con cupón	Sin cupón
Error del mínimo	0.0001310	0.0005668	0.0017126	0.0049300
Número de aciertos	95	88	92	94
Percentil 1	15	19	1	31
Percentil 5	58	50	1	35
Tiempo total	ND	ND	ND	ND
Tasa variando	**NS**		**NSS**	
	Con cupón	Sin cupón	Con cupón	Sin cupón
Error del mínimo	0.0001177	0.0004223	0.0010456	0.0040077
Número de aciertos	92	91	85	74
Percentil 1	34	17	1	1
Percentil 5	79	81	2	1
Tiempo total	ND	ND	ND	ND

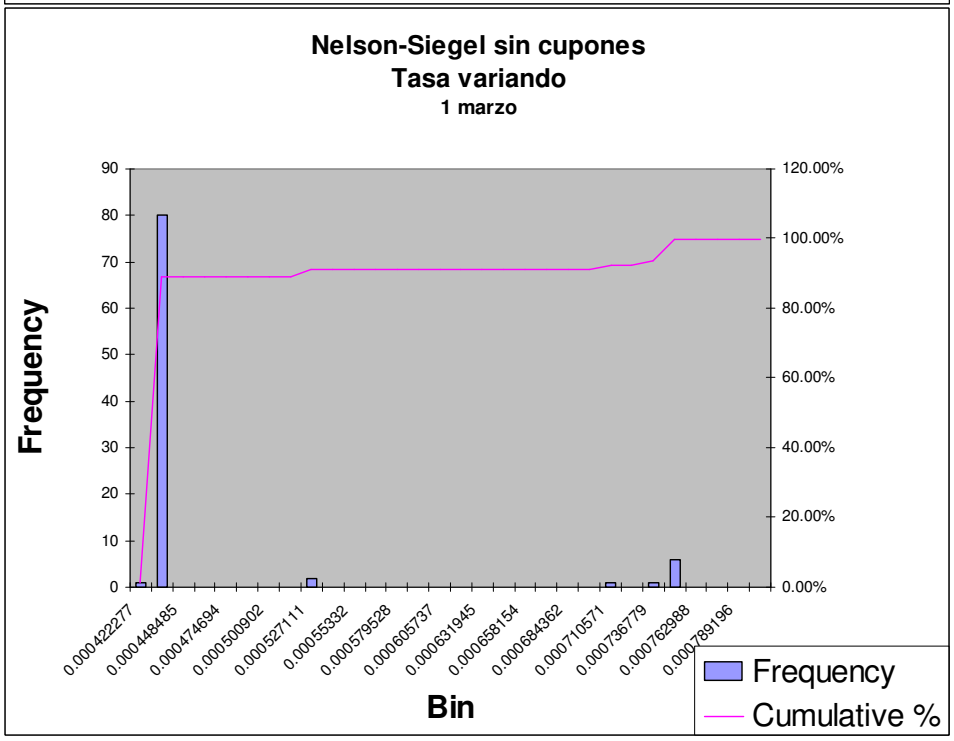
Resultados

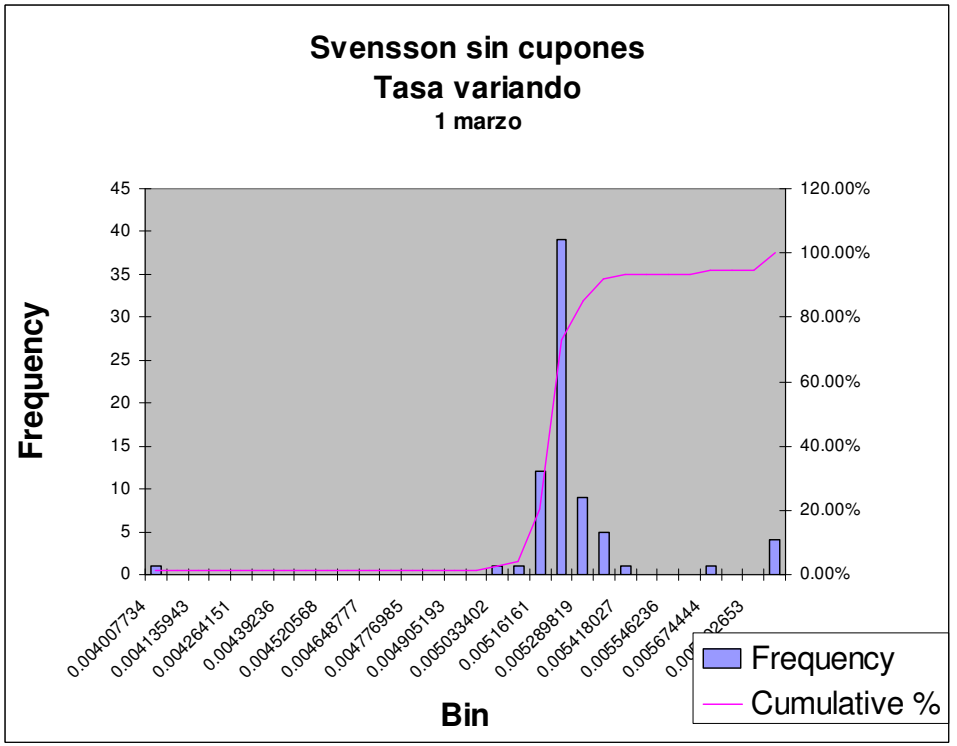
2 de marzo

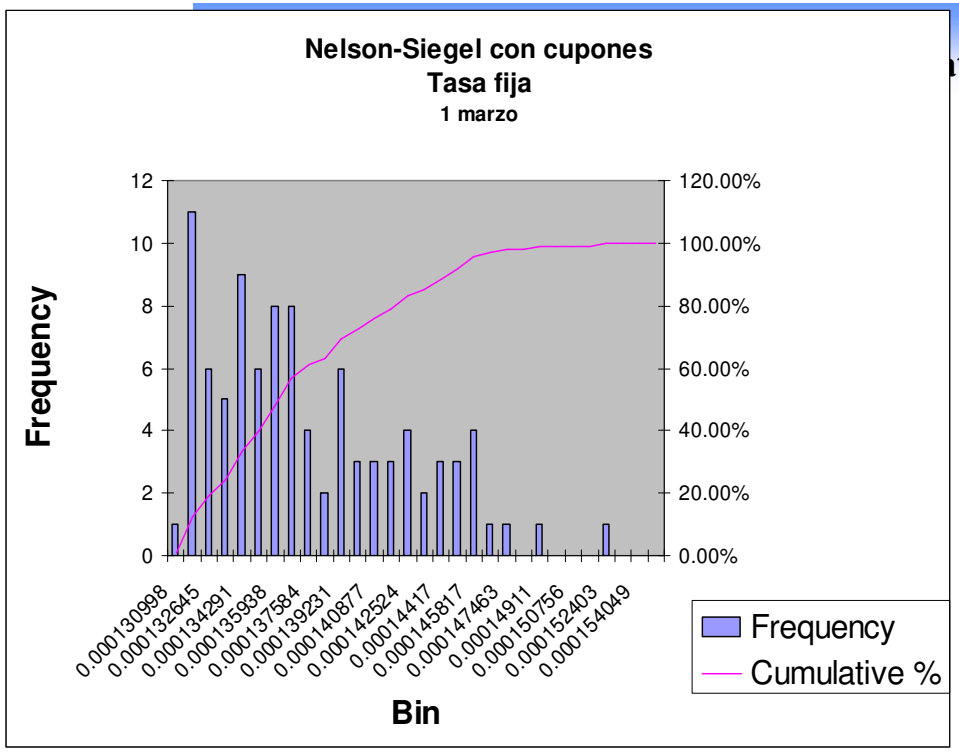
Tasa fija	NS		NSS	
	Con cupón	Sin cupón	Con cupón	Sin cupón
Error del mínimo	0.0000494	0.0000156	0.0077321	0.0018325
Número de aciertos	74	95	92	96
Percentil 1	2	5	3	1
Percentil 5	3	6	92	1
Tiempo total	4′ 8″	3′ 55″	5′ 20″	5′. 13″
Tasa variando	NS		NSS	
	Con cupón	Sin cupón	Con cupón	Sin cupón
Error del mínimo	0.0000516	0.0000257	0.0075450	0.0012706
Número de aciertos	88	88	89	93
Percentil 1	3	3	10	1
Percentil 5	12	3	65	5
Tiempo total	4′ 42″	4′ 48″	5′ 58″	6′ 16″

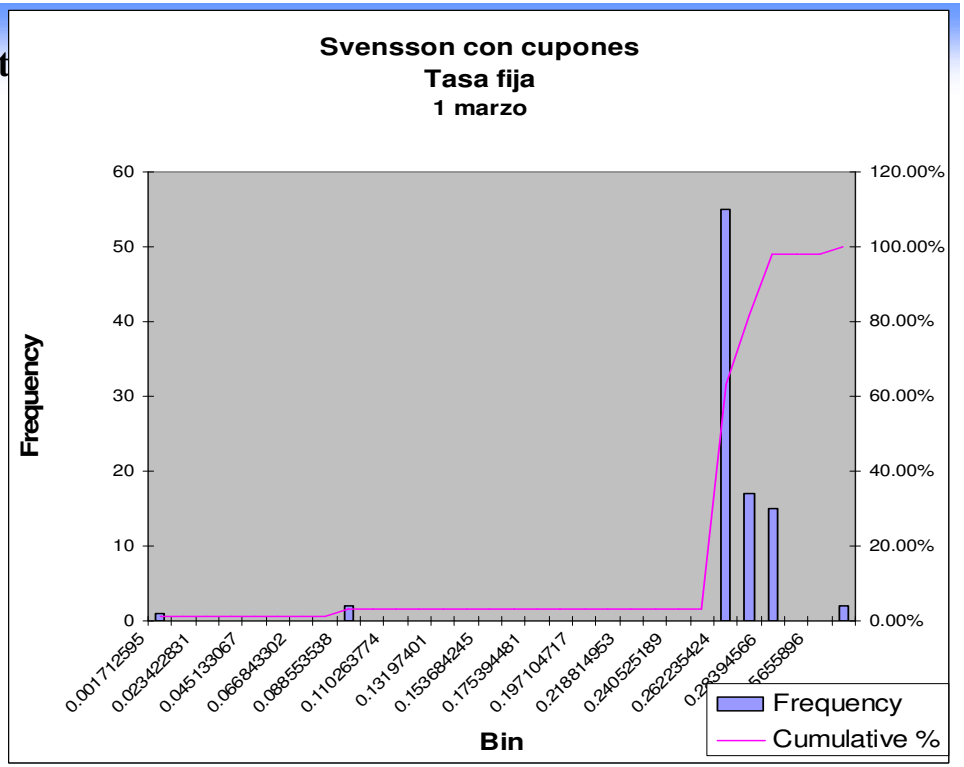

Nelson-Siegel con cupones
Tasa variando
1 marzo
Frequency
Bin
Frequency
Cumulative %

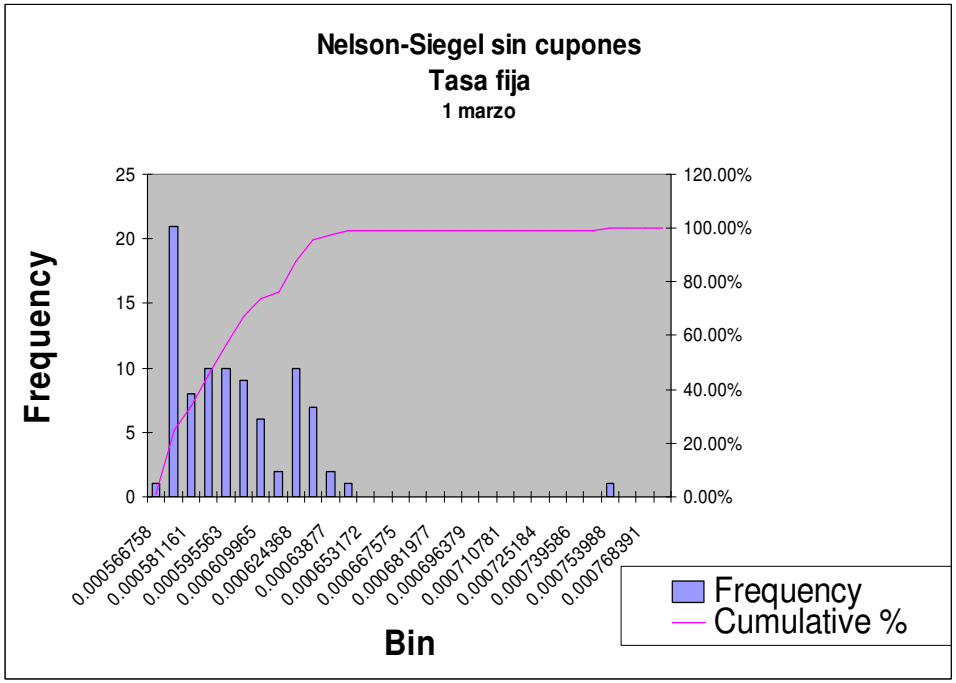

Svensson con cupones
Tasa variando
1 marzo
Frequency
Bin
Frequency
Cumulative %

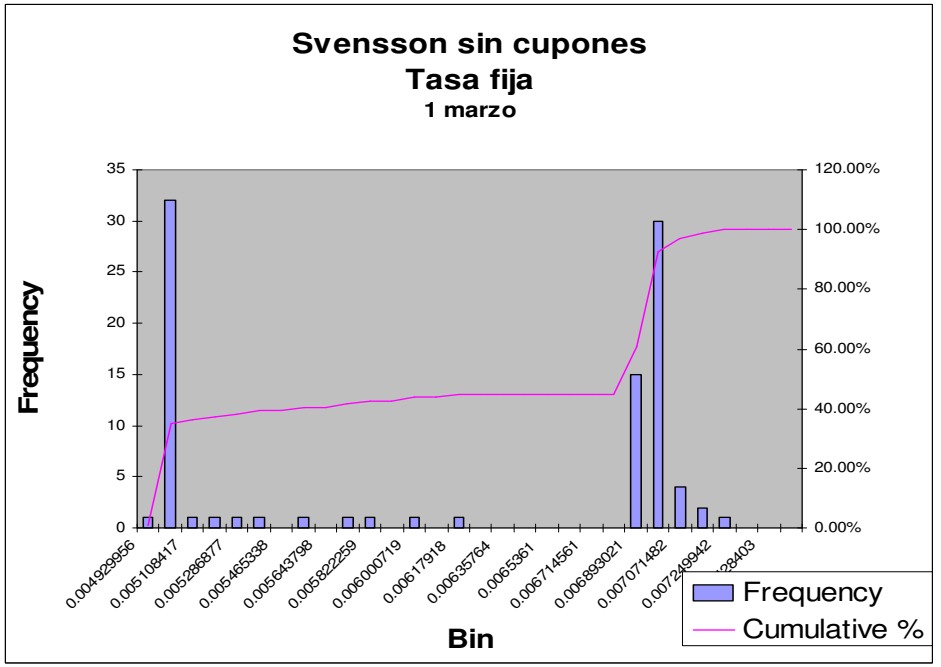

Nelson-Siegel sin cupones
Tasa variando
1 marzo
Frequency
Bin
Frequency
Cumulative %


Svensson sin cupones
Tasa variando
1 marzo
Frequency
Bin
Frequency
Cumulative %


Nelson-Siegel con cupones
Tasa fija
1 marzo
Frequency
Bin
Frequency
Cumulative %


Svensson con cupones
Tasa fija
1 marzo
Frequency
Bin
Frequency
Cumulative %


Nelson-Siegel sin cupones
Tasa fija
1 marzo
Frequency
Bin
Frequency
Cumulative %


Svensson sin cupones
Tasa fija
1 marzo
Frequency
Bin
Frequency
Cumulative %

Comentarios

- El valor de la función de error es más bajo, generalmente, para el modelo NS que para el modelo NSS. Esto quiere decir que si se ejecuta varias veces y se toma la mejor solución, en general el modelo NS dará un mejor ajuste.
- El número de aciertos para ambos modelos es comparable; esto quiere decir que ambos modelos pueden diverger un porcentaje de veces similar. Para el número de aciertos no hay diferencia significativa entre usar cupón o no, o entre tasa fija y tasa variando.
- Los percentiles indican qué tantas soluciones suboptimales están cercanas al mejor óptimo obtenido. En algunas ocasiones NS fue superior, pero en otras fue NSS el mejor. Para cualquiera de los dos modelos, se puede apreciar que no siemprese obtienen las soluciones en un intervalo cercano a la mejor solución encontrada, lo cual es reflejo de que la función encuentra muy a menudo óptimos locales. Esto se puede apreciar también en los histogramas que muestran la frecuencia de soluciones obtenidas en cada intervalo (ver en el Anexo B los histogramas para los datos del 1 de marzo).
- El modelo NS es más rápido que el modelo NSS, pero no hay diferencias en tiempos entre usar cupones o no, o entre tasa fija y tasa variando.

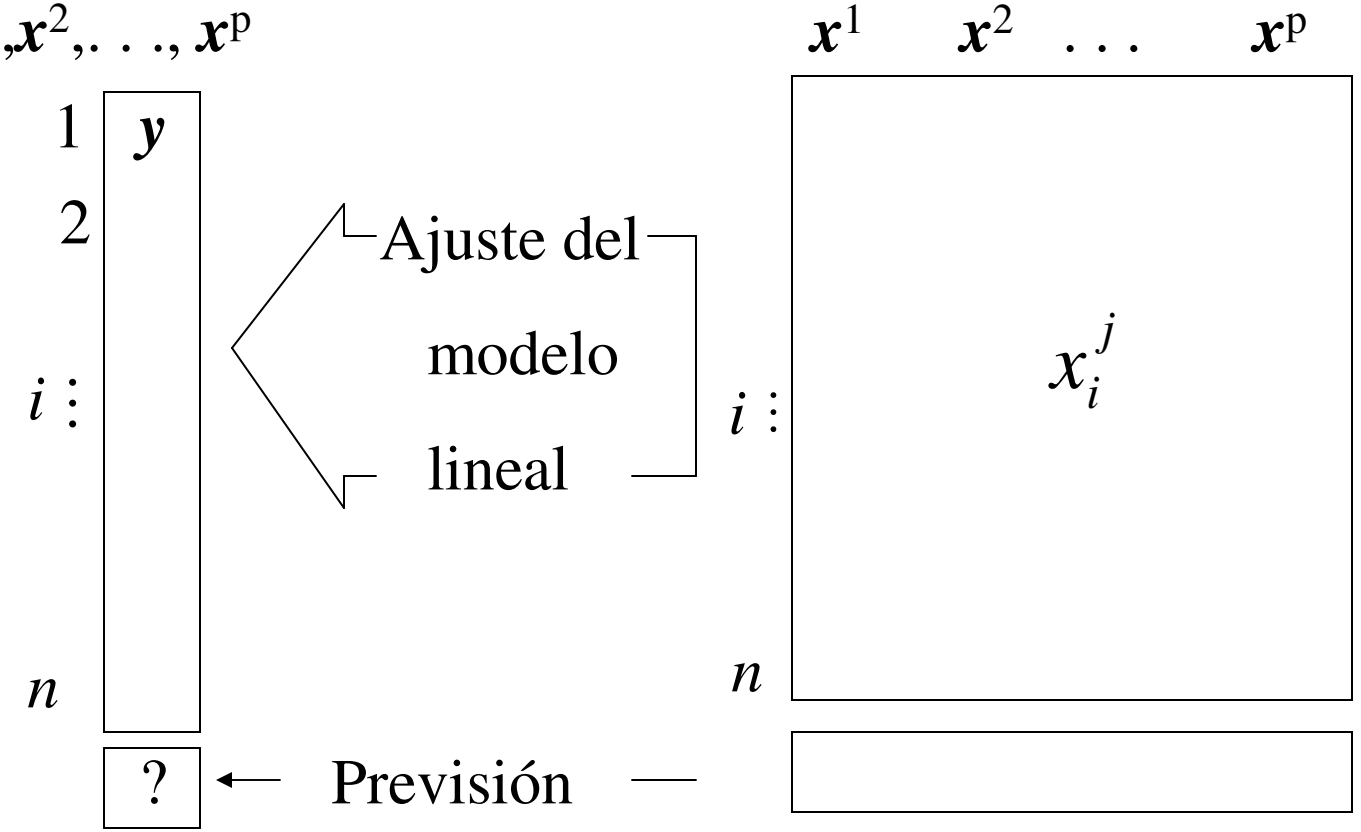
Conclusiones

- El modelo de Nelson-Siegel tiene un mejor rendimiento que el de Svensson en nuestro estudio experimental, a pesar de que el modelo de Svensson es una generalización del de Nelson-Siegel.
- **Razón:** la optimización no lineal del problema con 6 parámetros (modelo de Svensson) es considerablemente más complicada que la correspondiente con sólo 4 parámetros (modelo de Nelson-Siegel), con el software actual.

Regresión lineal Múltiple

Situación: $x^1, x^2, \ldots, x^p$, y variables cuantitativas

Objetivo: explicar y a partir de una combinación lineal de $x^1, x^2, \ldots, x^p$



Regresión Lineal

- Dados $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_p,\mathbf{y}$, se buscan coeficientes $b_1,b_2,\ldots,b_p$ tales que

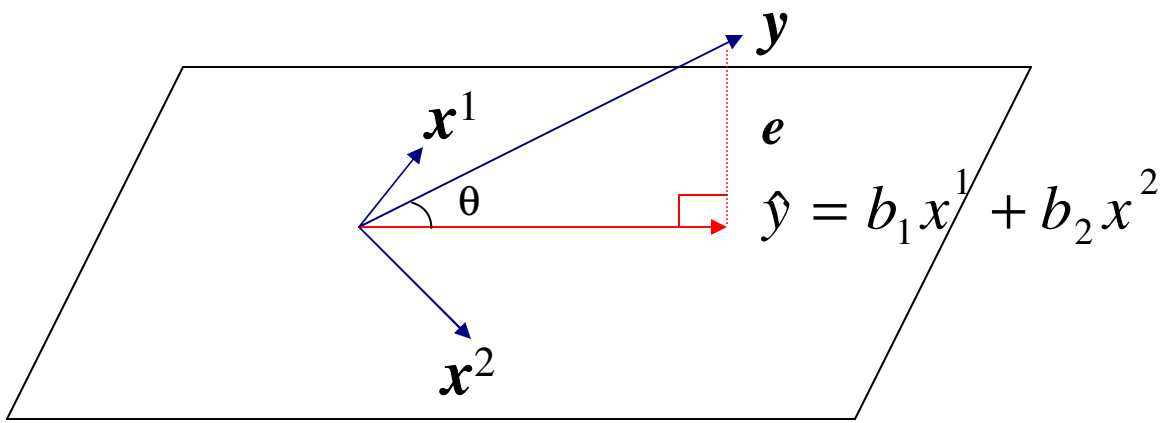
$$\left\| \mathbf{y} - \sum_{j=1}^{p} b_j \mathbf{x}_j \right\|^2 \rightarrow \min_{(b_1,\ldots,b_p)}$$

- $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_p$: predictores o vars. explicativas
- $\mathbf{y}$: variable a explicar (var. dependiente)

Formulación Geométrica

En R^n:

(centradas)



Por teorema de Pitágoras $\|y\|^2 = \|\hat{y}\|^2 + \|e\|^2$

$$\mathrm{var}(y) = \mathrm{var}(\hat{y}) + \mathrm{var}(y - \hat{y})$$

Varianza explicada Varianza residual

Solución conocida

- Si $X^t X$ es invertible entonces la solución es:

$$b = (X^t X)^{-1} X^t y$$

- Calidad de la regresión: coeficiente de determinación

$$R^2 = \cos\theta = r(y, \hat{y}) = \frac{\text{var } \hat{y}}{\text{var } y}$$

Selección de Variables en Regr.

- Se debe balancear dos objetivos en conflicto objectives:
  1. Inlcuir todas las variables con poder predictivo legítimo
  2. Excluir cualquier variables redundante
- Determinar el mejor subconjunto de predictores para incluir en el modelo
  1. No existe una única definición de "mejor"
  2. Diferentes algoritmos pueden producir diferentes soluciones
  3. Los problemas se magnifican si hay mucha correlation entre los predictores

Criterio

- Si se incluyen más variables en el modelo, el R^2 solo puede crecer
- No se quiere incluir muchas variables independientes que parezca que no contribuyen mucho al modelo
- Se debe selecccionar un criterio que crezca únicamente si las nuevas variables que se incluyan añaden informacion significativa al modelo

R cuadrado ajustado

$$\overline{R}_q^2 = 1 - \left(\frac{p-1}{p-q}\right)(1-R_q^2)$$

$$= R_q^2 - \left(\frac{q-1}{p-q}\right)(1-R_q^2) \rightarrow \max$$

q es el número de variables incluidas en el modelo

El índice puede bajar en valor si la contribución de una nueva variable incluida es menor que el impacto que tiene en el numero de g.l.

Métodos & Algoritmos

- Todos los posibles subconjuntos (problema combinatorial)
- Métodos paso a paso, hacia adelante (greedy: óptimo local)
  - Iniciar sin predictores
  - Incluir el predictor con la mayor correlación con *y*
  - Anadir predicto con la mayor correlacion parcia con *y*, considerando los predictores ya seleccionados
  - Stop when a numerical criterion holds
- Métodos paso a paso, hacia atrás (greedy: óptimo local)

AG para Selección de Variables

- Cromosomas: indicadoras de la variables (presence=1, absence=0, de una variable en el modelo)
- Fitness: adjustado $R^2 + \alpha$
- Selección: proporcional al fitness
- Operadores:
  - Mutación: con probabilidad p_m
  - Cruzamiento: con probabilidad p_c
- Criterio de parada: si [max_iter] ó [var(fitness) < ε]

Parámetros

- Maximo número de iteraciones
- Tamaño de la población
- Probabilidad de cruzamiento
- Probabilidad de mutación
- Tolerancia al número de condición
- α
- ε

Algunos Resultados

- Experimentos: 100 corridas en cada tabla de datos con población inicial al azar
- Factores: p_c, p_m, tamaño de la población
- Max_iter = 200
- $\alpha = 0.1$
- Tolcond = 100 000 000
- $\varepsilon =$

Longley Data

Number of objects: 16
Number of explanatory variables: 6

The labor statistics data set of Longley (1967) is noted for being ill conditioned

The Longley (JASA, 1967, p.819-841) regression coefficients have for many years been a reliable benchmark for testing regression algorithms and statistical packages.

Longley Data

- The data set contains one dependent variable, Employment (total derived employment)
- Six independent variables:
  - **Prices** (GNP implicit price deflator with year 1954 = 100),
  - **GNP** (gross national product),
  - **Jobless** (unemployment),
  - **Military** (size of armed forces),
  - **PopSize** (non-institutional population aged 14 and over), and
  - **Year** (year).

Longley Data Table

Prices	GNP	Jobless	Military	PopSize	Year	**Employment**
83	234289	2356	1590	107608	1947	**60323**
88.5	259426	2325	1456	108632	1948	**61122**
88.2	258054	3682	1616	109773	1949	**60171**
89.5	284599	3351	1650	110929	1950	**61187**
96.2	328975	2099	3099	112075	1951	**63221**
98.1	346999	1932	3594	113270	1952	**63639**
99	365385	1870	3547	115094	1953	**64989**
100	363112	3578	3350	116219	1954	**63761**
101.2	397469	2904	3048	117388	1955	**66019**
104.6	419180	2822	2857	118734	1956	**67857**
108.4	442769	2936	2798	120445	1957	**68169**
110.8	444546	4681	2637	121950	1958	**66513**
112.6	482704	3813	2552	123366	1959	**68655**
114.2	502601	3931	2514	125368	1960	**69564**
115.7	518173	4806	2572	127852	1961	**69331**
116.9	554894	4007	2827	130081	1962	**70551**

Longley Data: Preliminary results

Results from the ordinary regression algorithm

Coeficients	
-3482258.6346	constant
15.061872271	X1 - DEFL
-0.035819179293	X2 - GNP
-2.0202298038	X3 - UNEM
-1.0332268672	X4 - MIL
-0.051104105653	X5 - POP
1829.1514646	X6 - TIME

Longley – AGSelVar Output

The GASelVar algorithm has been applied with the following factors and levels:

- three population sizes (20, 50 and 100)
- three values of mutation probability (0.01, 0.1 and 0.2)
- three values of crossover probability (0.3, 0.5 and 0.7)

Longley – AGSelVar Output

Best fitness obtained: 1.099767

This fitness was associated to the following sets of variable selection:

1	0	0	0	0	0
1	0	0	0	1	0
0	0	0	1	0	0
1	0	1	1	0	0
1	0	0	0	1	1
1	1	0	0	0	1

Longley – AGSelVar Output

The attraction rate (AR) is a measure that gives us an idea how appropiate are the parameters.

Population size = 20:

Pm=	0.01	0.01	0.01	0.1	0.1	0.1	0.2	0.2	0.2	
Pc=	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	
AR=	**42**	**32**	**34**	**38**	**33**	**50**	**28**	**26**	**30**	**mean: 34.8**

Population size = 50:

Pm=	0.01	0.01	0.01	0.1	0.1	0.1	0.2	0.2	0.2	
Pc=	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	
AR=	**63**	**60**	**64**	**53**	**56**	**58**	**53**	**62**	**57**	**mean: 58.4**

Population size = 100:

Pm=	0.01	0.01	0.01	0.1	0.1	0.1	0.2	0.2	0.2	
Pc=	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	
AR=	**87**	**93**	**83**	**83**	**82**	**79**	**84**	**86**	**86**	**mean: 84.8**

ANOVA: F = 190.5, the null hypothesis is rejected (p value < 0.0001)

Longley – AGSelVar Output

Mean Values of Fitness

Best fitness obtained: 1.099767

PopSize = 20

Mean Value of Fitness	%Approximation to Best Fitness
0.55974	50.89623529
0.479682	43.61669335
0.479759	43.62369484
0.529709	48.16556598
0.499639	45.43135046
0.739316	67.22478489
0.489489	44.50842769
0.479453	43.59587076
0.654932	59.5518869

PopSize = 50

Mean Value of Fitness	%Approximation to Best Fitness
0.749772	68.17553173
0.809414	73.59868045
0.849405	77.23499614
0.75935	69.04644347
0.799302	72.67921296
0.898973	81.74213265
0.839025	76.29115985
0.849329	77.22808559
0.859097	78.11627372

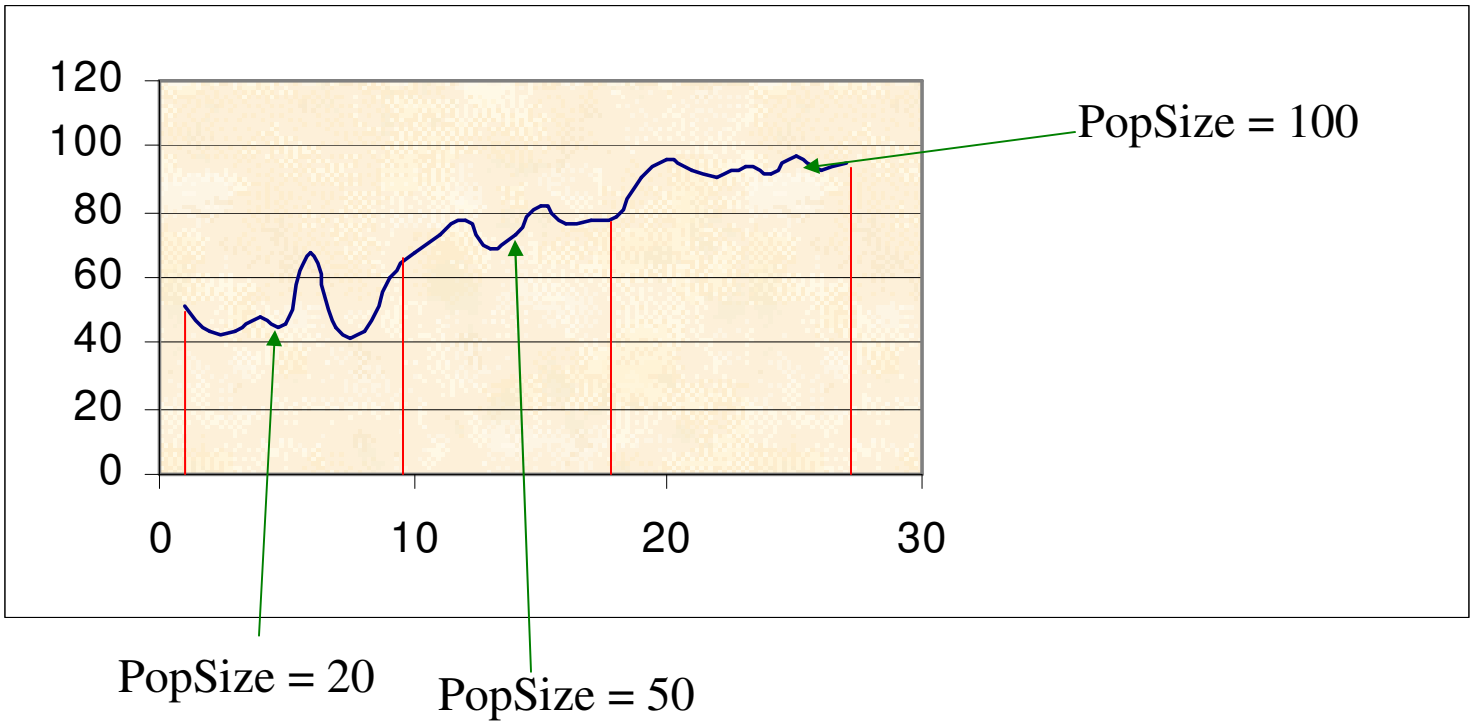
PopSize = 100

Mean Value of Fitness	%Approximation to Best Fitness
0.999676	90.8988904
1.059662	96.35331848
1.019442	92.69618019
0.999523	90.88497836
1.029363	93.59828036
1.00933	91.7767127
1.069278	97.2276855
1.019557	92.70663695
1.049435	95.42339423

ANOVA: F = 137.1, the null hypothesis is rejected (p value < 0.0001)

Longley – AGSelVar Output

Behaviour of the mean value of fitness, depending on the Population Size



Machine Data

- Number of objects: 209
- Number of explanatory variables: 6
- Maximum nb. of iterations: 200
- Total number of runs: 100

Machine – AGSelVar Output

The GASelVar algorithm has been applied with the following factors and levels:

- three population sizes (20, 50 and 100)
- three values of mutation probability (0.01, 0.1 and 0.2)
- three values of crossover probability (0.3, 0.6 and 0.7)

Machine – AGSelVar Output

Best fitness obtained: 0.917388

This fitness was associated to the following sets of variable selection:

1	1	0	1	1	1	(21 times)
1	1	1	1	0	0	(1 time)
0	1	1	1	1	1	(1 time)
1	1	1	0	1	1	(1 time)
1	1	1	0	0	0	(1 time)
0	0	1	1	0	1	(1 time)
1	0	1	1	0	1	(1 time)

CIMPA-UCR

Machine – AGSelVar Output

The attraction rate (AR) is a measure that gives us an idea how appropiate are the parameters.

Population size = 20:

Pm=	0.01	0.01	0.01	0.1	0.1	0.1	0.2	0.2	0.2	
Pc=	0.3	0.6	0.7	0.3	0.6	0.7	0.3	0.6	0.7	
AR=	**24**	**31**	**32**	**34**	**36**	**44**	**32**	**40**	**41**	**mean: 34.9**

Population size = 50:

Pm=	0.01	0.01	0.01	0.1	0.1	0.1	0.2	0.2	0.2	
Pc=	0.3	0.6	0.7	0.3	0.6	0.7	0.3	0.6	0.7	
AR=	**50**	**59**	**58**	**57**	**69**	**66**	**54**	**80**	**76**	**mean: 63.2**

Population size = 100:

Pm=	0.01	0.01	0.01	0.1	0.1	0.1	0.2	0.2	0.2	
Pc=	0.3	0.6	0.7	0.3	0.6	0.7	0.3	0.6	0.7	
AR=	**70**	**74**	**83**	**76**	**86**	**89**	**84**	**95**	**95**	**mean: 83.6**

ANOVA: $F = 73.37$, the null hypothesis is rejected (p value < 0.0001)

Machine – AGSelVar Output

Mean Values of Fitness Best fitness obtained: 0.917388

PopSize = 20

Mean Value of Fitness	%Approximation to Best Fitness
0.904765	98.6240282
0.912744	99.4937802
0.910924	99.2953908
0.910712	99.2722817
0.912658	99.4844057
0.910765	99.278059
0.912844	99.5046807
0.913876	99.617174
0.910587	99.2586561

PopSize = 50

Mean Value of Fitness	%Approximation to Best Fitness
0.914758	99.7133165
0.915919	99.8398715
0.915977	99.8461938
0.916225	99.873227
0.913904	99.6202261
0.916644	99.9189002
0.913778	99.6064915
0.917134	99.9723127
0.917104	99.9690425

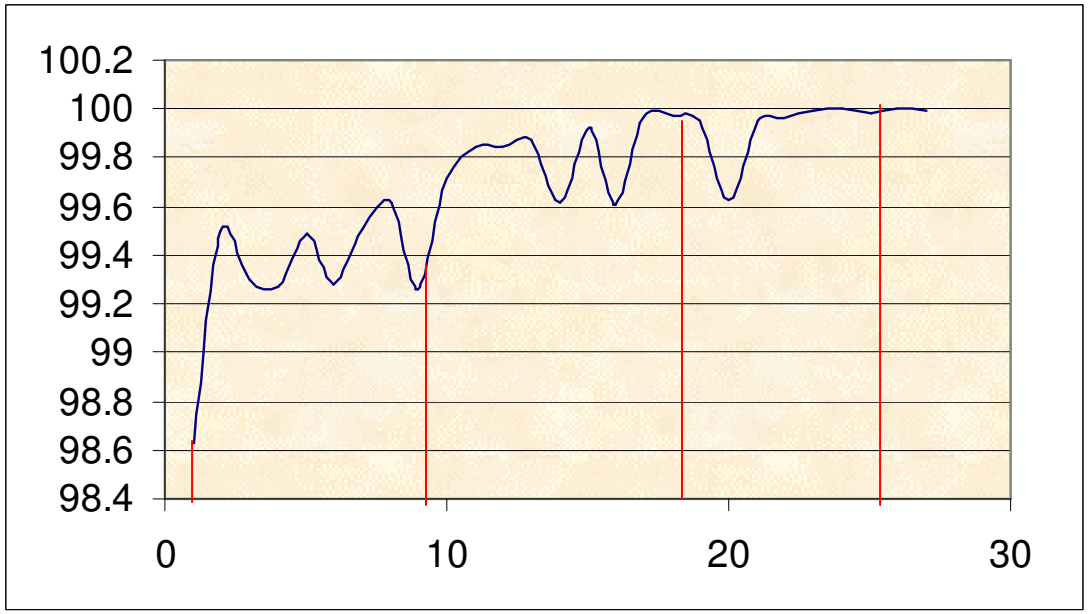
PopSize = 100

Mean Value of Fitness	%Approximation to Best Fitness
0.916982	99.9557439
0.913963	99.6266574
0.916932	99.9502937
0.917024	99.9603221
0.917318	99.9923696
0.917374	99.9984739
0.917245	99.9844123
0.917388	100
0.91736	99.9969479

ANOVA: $F = 25.25$, the null hypothesis is rejected (p value < 0.0001)

Machine – AGSelVar Output

Behaviour of the mean value of fitness, depending on the Population Size



Triazines Data

- Number of objects: 186
- Number of explanatory variables: 60
- Maximum nb. of iterations: 200
- Total number of runs: 100

Triazines – AGSelVar Output

The GASelVar algorithm has been applied with the following factors and levels:

- three population sizes (100, 300 and 500)
- three values of mutation probability (0.01, 0.1 and 0.2)
- three values of crossover probability (0.3, 0.5 and 0.7)

Triazines – AGSelVar Output

Population size: **100**

(Pm,Pc):	(0.01,0.3),	**(0.01,0.5),**	(0.01,0.7),
	(0.1,0.3),	(0.1,0.5),	(0.1,0.7),
	(0.2,0.3),	(0.2,0.5),	(0.2,0.7),
Best fitness:	1.052185	**1.052581**	1.051929
	1.052182	1.052017	1.052003
	1.051725	1.052160	1.052228

Variable selection

01110111111100101001010010000001110110111100001110010110001

Triazines – AGSelVar Output

Population size: **300**

(Pm,Pc):	(0.01,0.3),	(0.01,0.5),	(0.01,0.7),
	(0.1,0.3),	(0.1,0.5),	(0.1,0.7),
	(0.2,0.3),	(0.2,0.5),	(0.2,0.7),
Best fitness:	1.052861	1.052637	1.052846
	1.053086	1.052634	1.052513
	1.052256	1.052415	1.052584

Variable selection:

111111101100100111101100110001011010111101000111111100010110

Triazines – AGSelVar Output

Population size: **500**

(Pm,Pc):	(0.01,0.3),	(0.01,0.5),	(0.01,0.7),
	(0.1,0.3),	(0.1,0.5),	(0.1,0.7),
	(0.2,0.3),	**(0.2,0.5)**,	(0.2,0.7),
Best fitness:			
	1.052495	1.052664	1.052557
	1.052956	1.052410	1.052797
	1.052318	**1.053029**	1.052514

Variable selection:

11111111110010111100101010101011011111011100110110000101011 1

Triazines – AGSelVar Output

Mean Values of Fitness

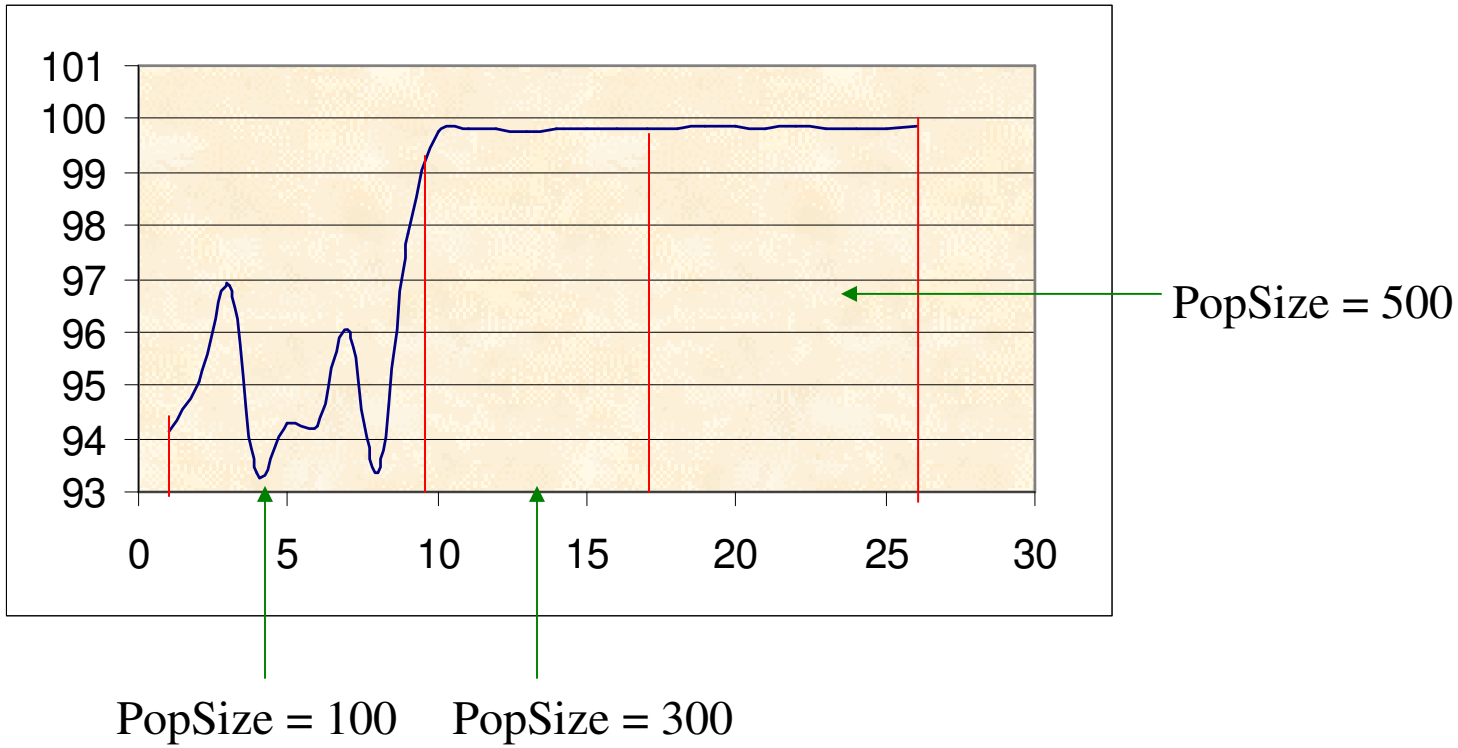
Mean Value of Fitness	Best Fitness	%Approximation to Best Fitness
0.990619	1.052185	94.14874761
1.00058	1.052581	95.05966762
1.01927	1.051929	96.89532278
0.982212	1.052182	93.35000979
0.991699	1.052017	94.26644246
0.991358	1.052003	94.2352826
1.01003	1.051725	96.03556063
0.982247	1.05216	93.35528817
1.02985	1.052228	97.87327461

Mean Value of Fitness	Best Fitness	%Approximation to Best Fitness
1.05026	1.052861	99.7529588
1.05062	1.052637	99.808386
1.05069	1.052846	99.7952217
1.05047	1.053086	99.7515872
1.0504	1.052634	99.7877705
1.05059	1.052513	99.8172944
1.05014	1.052256	99.7989083
1.05034	1.052415	99.8028344

Mean Value of Fitness	Best Fitness	%Approximation to Best Fitness
1.05041	1.052495	99.8018993
1.05101	1.052664	99.8428748
1.0511	1.052557	99.8615752
1.05072	1.052956	99.7876454
1.05089	1.05241	99.8555696
1.05107	1.052797	99.8359608
1.05063	1.052618	99.8111376
1.05082	1.053029	99.7902242
1.05089	1.052514	99.8457028

ANOVA: $F = 78.70$, the null hypothesis is rejected (p value < 0.0001)

Triazines – AGSelVar Output



Present applications

- Banana data:

  Data collected in 4 countries in the Caribbean basin, in 40 farms, 8 samples in each one, 60 variables. *Aim*: to select a minimum variable set for describing soil quality and health

- CableTV data:

  32 variables that describe market and demographic behavior of districts. *Aim*: to forecast the sales with a reduced set of variables

Final Comments

- It seems that the Population Size is important to the quality of the results
- Further comparisons should be made:
  - Tuning of the parameters
  - Compare to alternative methods (stepwise, ...)