

# Feature selection and volatility modeling of European options

Katya Krymova  
Vadim Strijov

Russian Academy of Sciences  
Computing Center

EURO XXIV  
July 11-14  
Lisbon

## European option

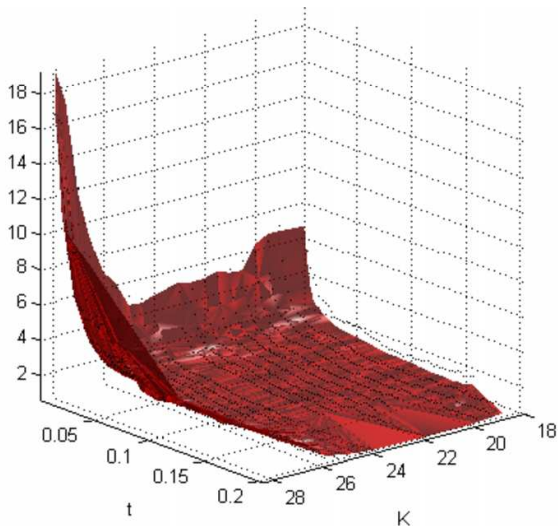
- The option is an instrument that conveys the right, but not the obligation, to engage in a future transaction on some underlying security.
- European option is an option that may only be exercised on expiration

Implied volatility:

$$\sigma = \mathit{arg} \min_{\sigma \in [0, 1.5]} (C_{K,t} - C(\sigma, P_t, B, K, t)).$$

- $\mathbf{K}$  is strike price,
- $t$  is the set of time ticks,
- $B$  is risk-free rate,
- $C_{K,t}$  is the historical price,
- $P_t$  is the historical security price.

## Volatility smile model



## Background

- The set of strike prices, time ticks and volatility values is given.
- Feature generation is necessary.
- The number of the features exceeds the number of the objects.
- Feature selection is a must.



## Feature generation

Let

- $\Xi = \{\xi^u\}_{u=1}^U$  be the set of free variables;
- $G = \cup\{g_v\}_{v=2}^V$  be the finite set of primitives given by experts.

For example  $G = \{\frac{1}{x}, \sqrt{x}, \ln(x), \tanh(x)\}$ .

Put  $a_\iota = g_v(\xi^u)$  determined by  $\iota = (v-1)U + u$ .

Put  $x_j = \prod a_{i_1} a_{i_2} \dots a_{i_s}$ ,

$i_1, i_2, \dots, i_s \in \{1, 2, \dots, UV\}$  and  $s = 1, 2, \dots, R$ .

$$\xi_u \xrightarrow{g_v} g_v(\xi_u) \stackrel{\text{def}}{=} a_\iota \xrightarrow{\prod} x_j$$

## Problem statement

The sample:

$$\{(\mathbf{x}^i, y^i) | i = 1, \dots, m\}, \quad \mathbf{x}^i \in \mathbb{R}^n, y^i \in \mathbb{R}^1, \quad n = |N|, N \subset \mathbb{N}.$$

The design matrix:  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ .

The set of indices  $\mathcal{A} \subset \mathcal{Z} = \{1, 2, \dots, n\}$  corresponding the optimal model is required.

The optimal model:

$$y^i = \sum_{j \in \mathcal{A}} x_j^i w_j,$$
$$S = \sum_{i=1}^m \left( \sum_{j \in \mathcal{A}} x_j^i w_j - y^i \right)^2 \rightarrow \min.$$

## Bayesian model selection

- 1 Estimation of model parameters.
- 2 Model evidence is  $p(D|f_i)$ .

The likelihood function  $p(f_i|D)$  is given by Bayes' formula:

$$p(f_i|D) = \frac{\mathbf{p}(\mathbf{D}|\mathbf{f}_i)p(f_i)}{p(D)}.$$

For all  $i, j$ ,  $p(f_i) = p(f_j)$ .

Comparison of models:

$$\frac{p(f_i|D)}{p(f_j|D)} = \frac{p(D|f_i)}{p(D|f_j)}.$$

## Evidence calculation

The likelihood function

$$p(D|\mathbf{w}, f, \beta) = \prod_{i=1}^m \mathcal{N}(y^i | f(\bar{\mathbf{x}}^i, \mathbf{w}), \beta^{-1}).$$

Let  $\mathbf{w} \sim \mathcal{N}(0, \alpha I)$ .

The evidence is given by:

$$p(D|f, w, \alpha, \beta) = \int p(D|\mathbf{w}, f, \beta) p(\mathbf{w}|f, \alpha) d\mathbf{w}.$$

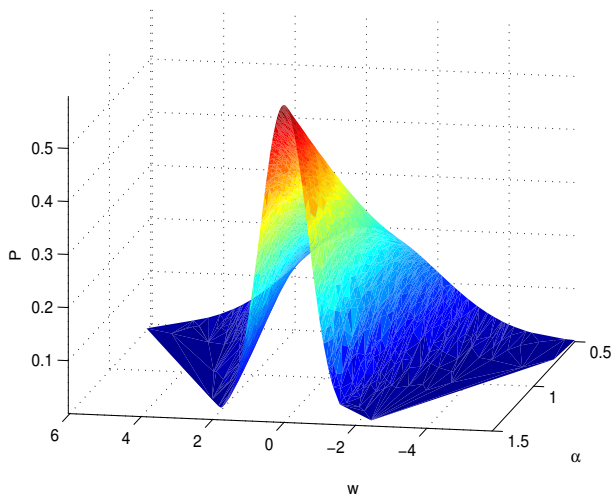
Calculation:

$$\begin{aligned} \ln p(D|f, w, \alpha, \beta) &= -\frac{w}{2} \|\mathbf{y} - X\mathbf{w}\|^2 - \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} - \frac{1}{2} \ln H + \\ &\quad + \frac{m}{2} \ln \beta + \frac{l}{2} \ln \alpha - \frac{m}{2} \ln 2\pi, \end{aligned}$$

$l$  is the number of parameters in the model.



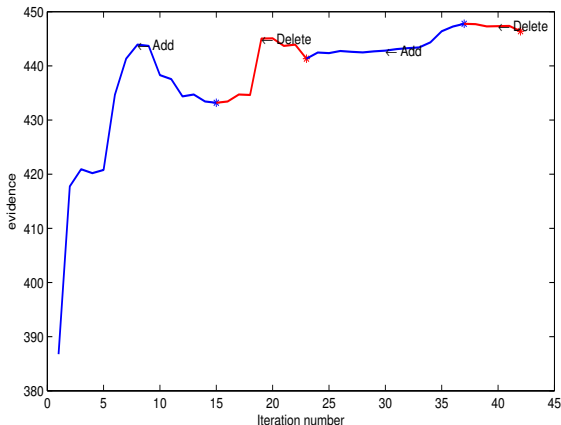
## An illustration of parameter distribution depending on $\alpha$



## Selection of the most evident model

The evidence from Coherent Bayesian Inference is maximizing:

1. **Add** 2. **Delete** features, while the evidence value is increasing and some steps while decreasing.



## Algorithm of evidence maximization

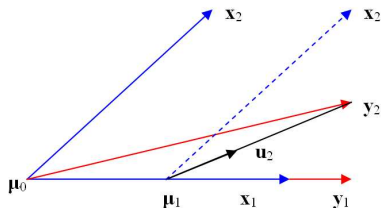
Set of all features indexes  $\mathcal{Z} = \{1, 2, \dots, n\}$ .

The current set  $\mathcal{A}_k$ .  $\mathcal{A}_0 = \emptyset$ .

Consider  $k$ -th algorithm step.

- 1 Adding:  $\mathcal{A}_k = \mathcal{A}_{k-1}$ ; features are added from  $\mathcal{Z} \setminus \mathcal{A}_{k-1}$  into  $\mathcal{A}_k$  by criterion  $\mathcal{C}_L$ , while criterion  $\mathcal{C}_E$  is completed.
  - 2 Deleting: features are deleted by criterion  $\mathcal{C}_D$ , while criterion  $\mathcal{C}_E$  is completed.
- $\mathcal{C}_L$  : LARS step.
  - $\mathcal{C}_D$  : Belsley method.
  - $\mathcal{C}_E$  : evidence is not less than  $\mathcal{E}_{min}$ .

# 1. Add a feature: Least angle regression (LARS)



Put  $\mu = X\mathbf{w}$ .

**Zero step:**  $\mu_0 = \mathbf{0}$ , the residual vector  $\varepsilon_0 = \mathbf{y} - \mu_0$ .

**First step:**  $\text{corr}(\mathbf{y}, \mathbf{x}_1) > \text{corr}(\mathbf{y}, \mathbf{x}_2)$ , then  $\mu_1 = \mu_0 + w_1 \mathbf{x}_1$ , where  $w_1$  provides  $\mathbf{y} - \mu_1$  to lie on the bisecting line between  $\mathbf{x}_1, \mathbf{x}_2$ .

**Second step:** the parameter  $w_2$  satisfies

$$\mu_2 = \mu_1 + w_2 \mathbf{u}_2 = \mathbf{y}$$

for  $m = 2$ , where  $\mathbf{u}_2$  is the unit vector.

## 2. Delete a feature: Belsley method

Singular value decomposition of the correlation matrix  $X^T X$ :  $X^T X = U \Lambda V^T$ ,

where  $\Lambda$  is the diagonal matrix with singular values  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ .

Condition indices :  $\eta_j = \frac{\lambda_{max}}{\lambda_j}$ .

The variance of  $w_i$  :  $\sigma^{-2} \mathcal{V}(w_i) = (q_{i1} + q_{i2} + \dots + q_{in}) \sum_{j=1}^n \frac{v_{ij}^2}{\lambda_{ij}^2}$ ,  
 $q_{ij}$  is the contribution of corresponding summand.

$$\hat{j} = \arg \max \eta_j$$

Condition indices	$\mathcal{V}(w_1)$	$\mathcal{V}(w_2)$	...	$\mathcal{V}(w_n)$
$\eta_{\hat{j}}$	$q_{1\hat{j}}$	$q_{2\hat{j}}$	...	$q_{n\hat{j}}$

Correlated features correspond large  $q_{k\hat{j}}$ .

## Experiment: historical data of European options

- $\mathbf{K} = \{K_1, K_2, \dots, K_9\} = \{1400, 1425, \dots, 1575, 1600\}$  is the set of strike prices,
- $t = \{t_1, t_2, \dots, t_{36}\}$  is the set of time ticks (Maturity),
- $C_{K,t}$  is the historical option price,
- $P_t$  is the historical security price.

The sample set for regression analysis

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^m = \{(\langle K_i, t_i \rangle, \sigma_i)\}_{i=1}^m.$$

Set of primitives  $G = \{\frac{1}{x}, \sqrt{x}, \ln(x), \tanh(x)\}$ .

## Index mapping

Maturity	$C_{K_1,t}$	$C_{K_2,t}$	...	$C_{K_8,t}$	$C_{K_9,t}$	Price
-129	129.70	109.00	...	17.60	9.10	1495.42
-128	129.70	109.10	...	18.00	10.10	1494.25
...	...	...	...	...	...	...
-1	90	64.3	...	0.7	0.25	1473.99

Implied volatility

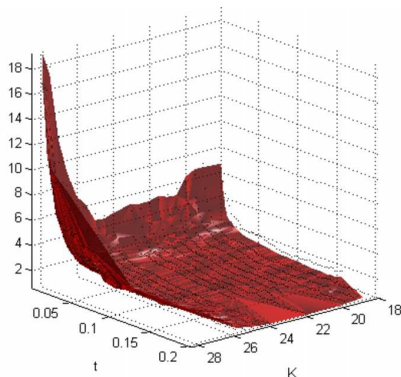
$$\sigma_i = \arg \min_{\sigma \in [0, 1.5]} (C_{K_i, t_i} - C(\sigma, P_{t_i}, B, K_i, t_i)),$$

where

$$(K_k.t_\tau) = (K_i, t_i) \in \mathbf{K} \times t,$$

$$i = \tau + k(|T| - 1).$$

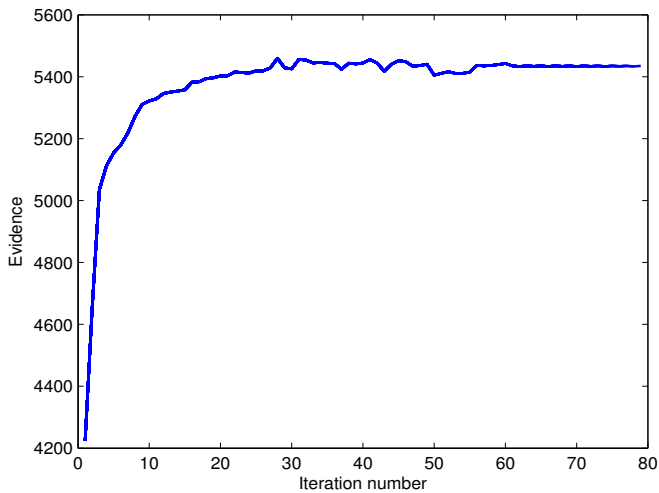
## Example of the resulting model



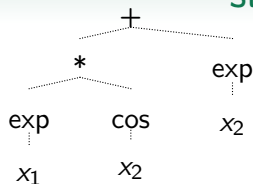
$$\begin{aligned}
 &w_1 + w_2 t^{\frac{1}{2}} \ln K + w_3 \ln K \ln t + w_4 K^{\frac{1}{2}} \ln^2 K + w_5 K^{-\frac{1}{2}} t^{-2} + \\
 &+ w_6 \ln^2 K \tanh K + w_7 K^{-2} \ln K + w_8 t \ln K + w_9 t^{\frac{1}{2}} \ln K \ln t + \\
 &+ w_{10} \ln K \ln^2(t) + w_{11} \ln K \tanh^2 t + w_{12} K^{-3} + w_{13} K^{-1} t^{-1} \tanh K.
 \end{aligned}$$



## Evidence convergence

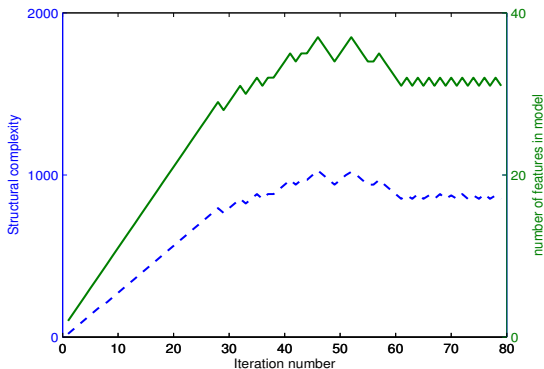


## Structural complexity



The model is represented as a tree.

The structural complexity is a sum of the number of nodes in all subtrees of given tree.



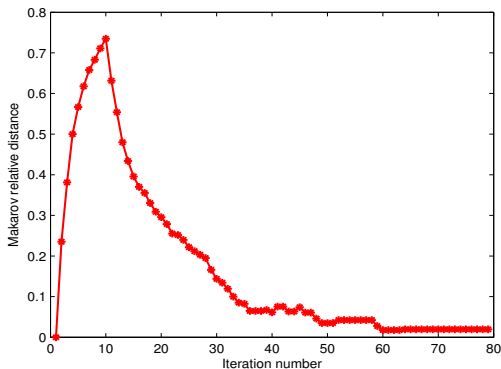
## Distance between models

Distance measure is determined by number of nodes  $p_{12}$  in maximum common subgraph of two trees  $T^1$  and  $T^2$ .

Let  $p_1$  and  $p_2$  — numbers of nodes in trees  $T^1$  and  $T^2$ .

Distance measure:

$$r_{12} = (p_1 + p_2 - 2p_{12}) / (p_1 + p_2).$$



## Results of comparison

Algorithm	CV	AIC	BIC	$C_p$	$\lg \kappa$	$k$
Genetics	0,107	-1152	-1072	337	13	26
GMDH	0,194	-1076	-1045	745	6	10
Stepwise	0,154	-1092	-1055	644	7	12
Ridge	0,146	-819	-330	832	33	160
Lasso	0,147	-1089	-1034	611	5	18
Stagewise	0,096	-1157	-1077	324	9	26
FOS	0,135	-1105	-1044	527	7	20
LARS	0,095	-1102	-1017	492	7	28
Proposed	0,123	-1118	-1054	469	2	21

$$AIC = m \ln(S/m) + 2k, \quad BIC = m \ln(S/m) + k \ln m,$$

$$C_p = S/\sigma^2 - 2k + m.$$

## Conclusion

- The feature generation method was proposed.
- The proposed algorithm construct the model in the maximum evidence neighborhood.
- This algorithm allows to get more stable models in comparison with well-known algorithms.

The proposed algorithm is similar to stepwise regression.

But instead of common criterion (Mallow's  $C_p$ ) the algorithm uses the Coherent Bayesian Inference. This allows to obtain the model plausible given the data.