Data-Driven Penalty Calibration: Slope Heuristics and the CAPUSHE Package

Jean-Patrick Baudry

Université Paris-Sud 11
Project SELECT (INRIA)
UPMC

Joint work with C. Maugis & B. Michel

September 3, 2010

Table of Contents

Contrast Minimization General Settings

- $X_1, \ldots, X_n, \ldots$ i.i.d. from an unknown distribution.
- s a function of interest linked to this distribution.
- The contrast minimization approach relies on an empirical contrast γ_n, depending on $X_1, \ldots, X_n$ and such that

$$t \mapsto \mathbb{E}[\gamma_n(t)]$$

  reaches a minimum value at $t = s$.
- In any model S, s is estimated by the empirical contrast minimizer

$$\hat{s} \in \operatorname*{argmin}_{t \in S} \gamma_n(t).$$

- Its quality is measured by the corresponding natural loss function ℓ:

$$\forall t \in S,\ \ell(s,t) = \mathbb{E}[\gamma_n(t)] - \mathbb{E}[\gamma_n(s)].$$

Contrast Minimization: Maximum Likelihood

Contrast minimization generalizes maximum likelihood estimation:

- s is the density of the sample distribution itself.
- For any density t, $\gamma_n(t) = -\frac{1}{n}\sum_{i=1}^{n} \log t(X_i)$.
- In any model S, the minimum contrast estimator is the standard MLE:

$$\hat{s} = \underset{t \in S}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log t(X_i) \right\}.$$

- The corresponding loss function is the Kullback-Leibler divergence:

$$\begin{aligned} \ell(s,t) &= \mathbb{E}\left[-\frac{1}{n} \sum_{i=1}^{n} \log t(X_i) \right] - \mathbb{E}\left[-\frac{1}{n} \sum_{i=1}^{n} \log s(X_i) \right] \\ &= d_{KL}(s,t) \geq 0, \end{aligned}$$

which is uniquely minimized at $t = s$.

Contrast Minimization: Other Contrasts

Examples of least-squares constrasts.

- Regression $\begin{cases} Y_i = s(X_i) + \varepsilon_i, \quad i = 1, \dots, n \\ \gamma_n(t) = -\frac{1}{n} \sum_{i=1}^n \big(Y_i - t(X_i)\big)^2 \\ \ell(s,t) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}\Big[(t-s)^2(X_i)\Big] \geq 0 \end{cases}$

- Gaussian white noise $\begin{cases} dY^{(n)}(x) = s(x)dx + \frac{1}{\sqrt{n}} dW(x), \\ \qquad\qquad \text{with } W \text{ a brownian motion} \\ \gamma_n(t) = \|t\|^2 - 2\int t(x) dY^{(n)}(x) \\ \ell(s,t) = \|t-s\|_2^2 \geq 0 \end{cases}$

- Density Estimation $\begin{cases} X_1, \dots, X_n \text{ i.i.d. with density } s \\ \gamma_n(t) = \|t\|^2 - \frac{2}{n} \sum_{i=1}^n t(X_i) \\ \ell(s,t) = \|t-s\|_2^2 \geq 0 \end{cases}$

Model Selection

Model chosen based on the data:

- $(S_m)_{m\in\mathcal{M}}$ a model collection.
- Estimator in model S_m: $\hat{s}_m \in \operatorname{argmin}_{t\in S_m} \gamma_n(t)$.
- Ideal choice: Model $m(s)$ minimizing the risk $\mathbb{E}\left[\ell(s,\hat{s}_m)\right]$. $\hat{s}_{m(s)}$ is not an estimator of s, but is a benchmark for model selection procedures. It is the *oracle*.
- Aim: $\hat{s}_{\hat{m}}$ such that $\mathbb{E}\left[\ell(s,\hat{s}_{\hat{m}})\right]$ is as close as possible to $\mathbb{E}\left[\ell(s,\hat{s}_{m(s)})\right]$.
- A bias-variance trade-off has to be reached: Model chosen by minimizing a penalized criterion

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}} \Big\{ \underbrace{\gamma_n(\hat{s}_m) + \operatorname{pen}(m)}_{\operatorname{crit}(m)} \Big\}.$$

Model Selection: Ideal Penalty

Let us denote

$$\hat{s}_D = \underset{\{\hat{s}_m : m \in \mathcal{M}, Dim(m)=D\}}{\operatorname{argmin}} \gamma_n(\hat{s}_m).$$

$$s_D = \underset{t \in \bigcup_{Dim(m)=D} S_m}{\operatorname{argmin}} \mathbb{E}\left[\gamma_n(t)\right].$$

$$\begin{aligned} \text{crit}(D) &= \gamma_n(\hat{s}_D) + \text{pen}(D) \ \left(-\gamma_n(s)\right) \\ &= \underbrace{\gamma_n(s_D) - \gamma_n(s)}_{\approx \ell(s, s_D)} - \underbrace{\left(\gamma_n(s_D) - \gamma_n(\hat{s}_D)\right)}_{\hat{v}_D} + \text{pen}(D) \end{aligned}$$

$\rightarrow$ With the ideal penalty

$$\text{pen}_{id}(D) = \hat{v}_D + \underbrace{\ell(s, \hat{s}_D) - \ell(s, s_D)}_{\ell(s_D, \hat{s}_D)},$$

$\text{crit}_{id}(D) \approx \ell(s, \hat{s}_D)$ selects the oracle.

Model Selection: Penalized Criteria

Recall: $\text{pen}_{id}(D) = \hat{v}_D + \ell(s_D, \hat{s}_D)$.

Asymptotic approach:

- Akaike's AIC (73), density estimation with log-likelihood contrast:

$$\ell(s_D, \hat{s}_D) \approx \hat{v}_D \approx \frac{D}{2n}.$$

$$\text{pen}_{AIC}(D) = \frac{1}{n} D$$

- Mallows' Cp (73), least-square regression:

$$\text{pen}_{Cp}(D) = 2\frac{\sigma^2}{n} D$$

The models dimension and the number of models are bounded, and $n \to \infty$.

Model Selection: Penalized Criteria

Recall: $\text{pen}_{id}(D) = \hat{v}_D + \ell(s_D, \hat{s}_D)$.

Nonasymptotic approach:

- Models, models dimension and number of models may depend on n.
- Example: Change Points Detection. Regression by piecewise constant functions with endpoints to be chosen on the grid $\left\{\frac{j}{n} : 0 \leq j \leq n\right\}$.
- Birgé & Massart (97, 01, 07) and others get nonasymptotic bounds on $\hat{v}_D + \ell(s_D, \hat{s}_D)$ through concentration results on empirical processes.
- They derive "optimal" penalties typically such that

$$\text{pen}(D) = \kappa D$$

or $$\text{pen}(D) = \kappa D\left(2 + \log\left(\frac{n}{D}\right)\right) \qquad \text{(Lebarbier, 05)}$$

- But κ is unknown and may depend on s, the sample distribution...

Slope Heuristics: Data-driven Penalty Calibration

Recall:
$$\text{crit}(D) \approx \underbrace{\gamma_n(\hat{s}_D) - \gamma_n(s_D)}_{\ell(s,s_D) - \hat{v}_D} + \text{pen}(D)$$

$\rightarrow$ With the penalty ($\alpha > 0$)

$$\begin{aligned} \text{pen}_\alpha(D) &= \alpha \hat{v}_D \\ &= \alpha \kappa_{min} D, \end{aligned} \qquad \text{(for example)}$$

$\text{crit}_\alpha(D) = \ell(s, s_D) + (\alpha - 1)\hat{v}_D$ selects models among the most complex iff $\alpha < 1$.

"$\text{pen}_{min} = \hat{v}_D$ is a minimal penalty".

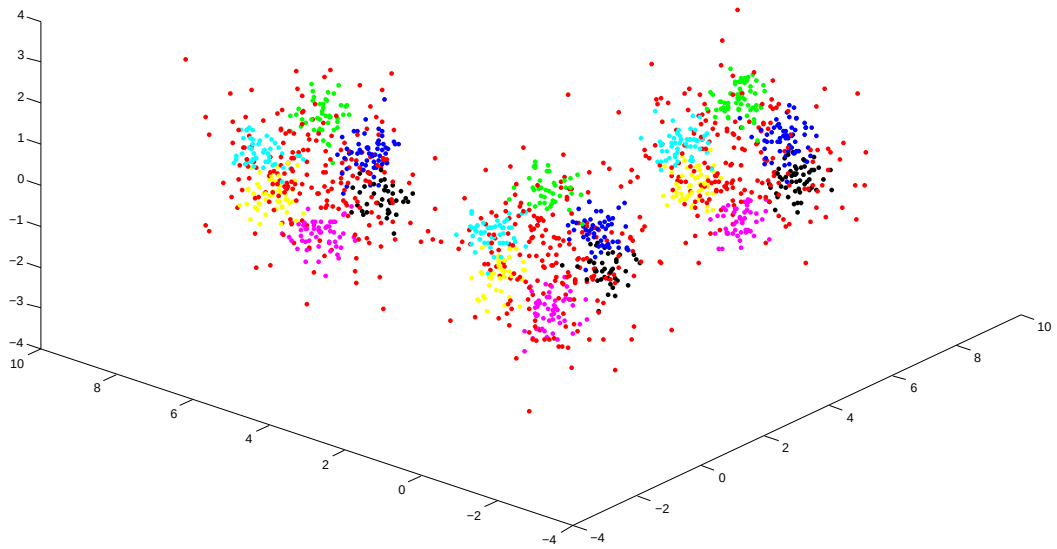
$\rightarrow$ Deduce the optimal penalty ($\hat{v}_D \approx \ell(s_D, \hat{s}_D)$)

$$\begin{aligned} \text{pen}_{opt} &= 2 \times \text{pen}_{min} \\ &= 2\alpha \kappa_{min} D. \end{aligned} \qquad \text{(for example)}$$

(Birgé & Massart, 07 ; Arlot & Massart, 09)

Slope Heuristics: Data-driven Penalty Calibration

Illustration: the "Bubbles" Dataset

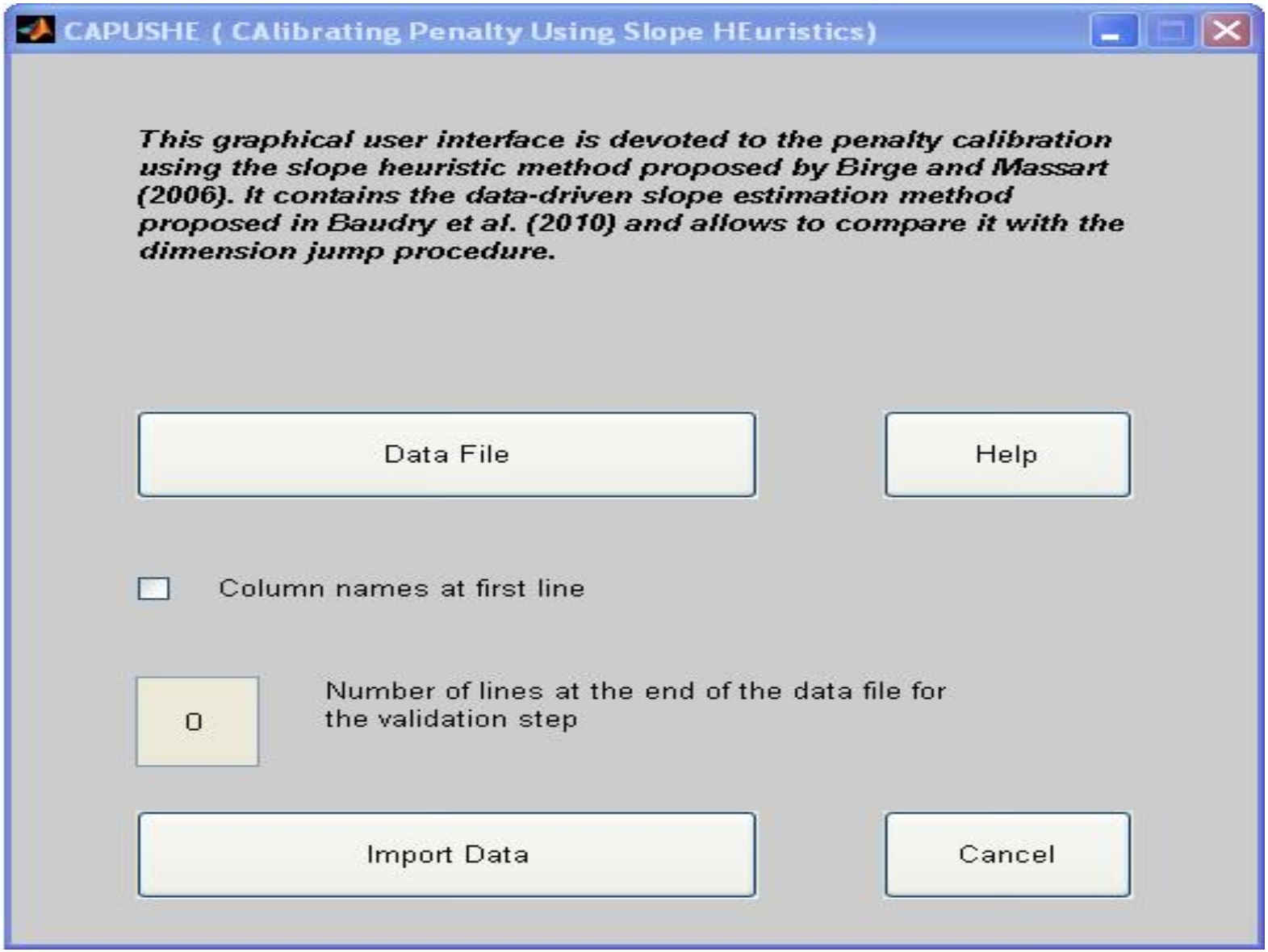


1000-sample of a 21-component Gaussian mixture.

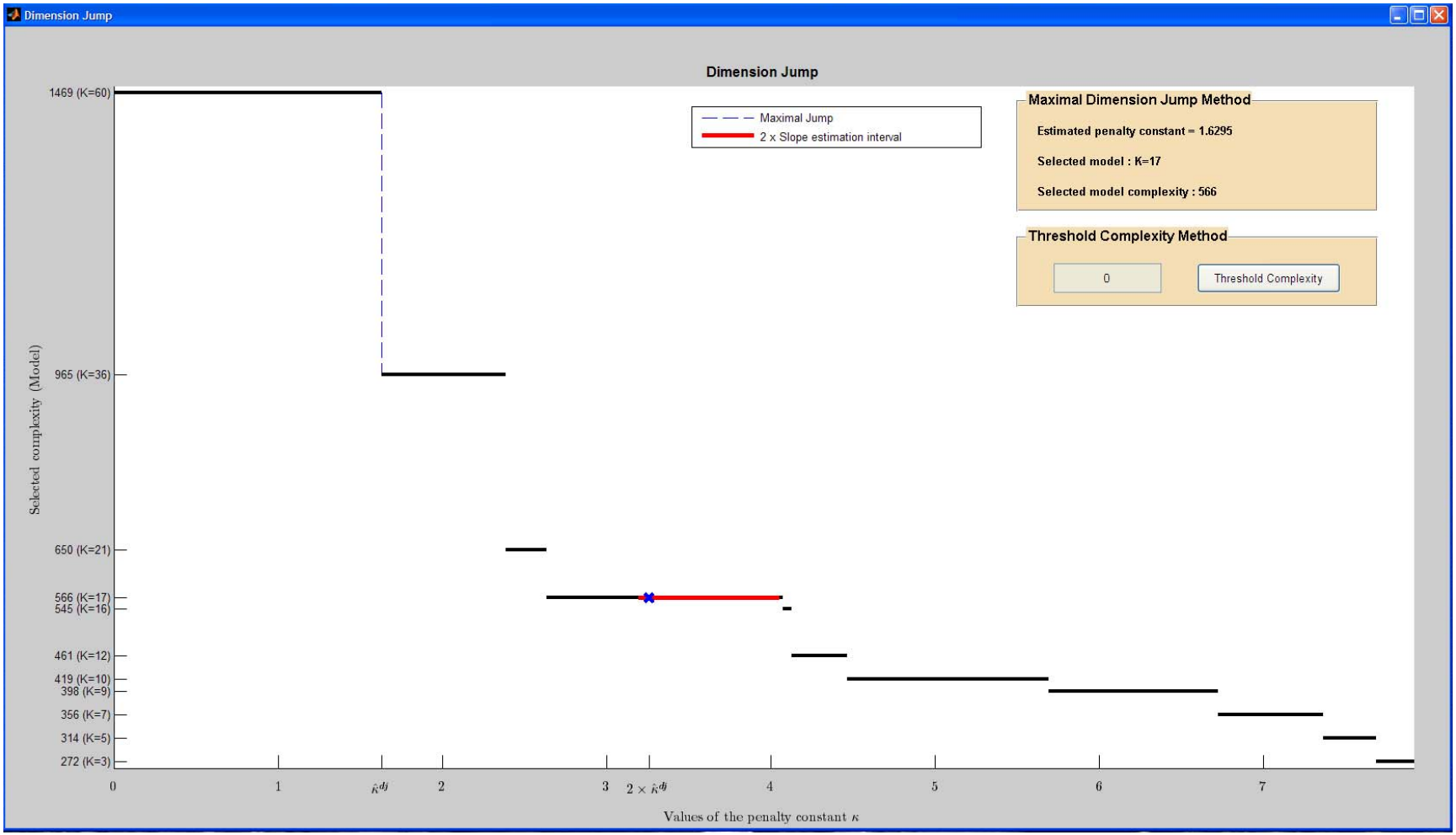
- Model S_m: (spherical) Gaussian mixture model with m components.
- $\forall t \in S_m, \gamma_n(t) = -\sum_{i=1}^{n} \log t(X_i)$.
- Penalty shape: $\text{pen}(D) = \kappa D$ (Maugis & Michel, 09).

CAPUSHE

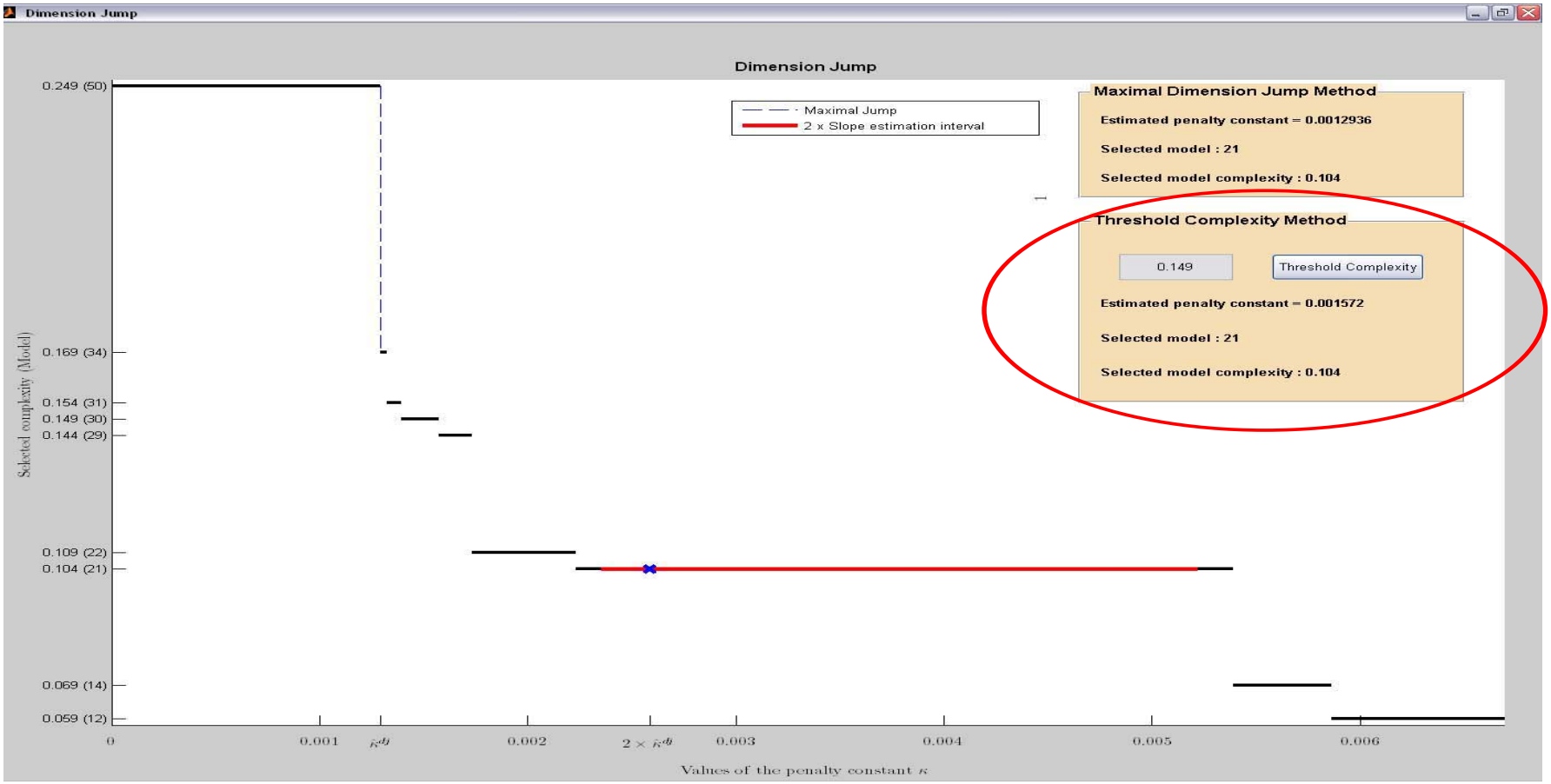
(Baudry, Maugis, Michel, 10)



Slope Heuristics: Dimension Jump



Slope Heuristics: Dimension Jump

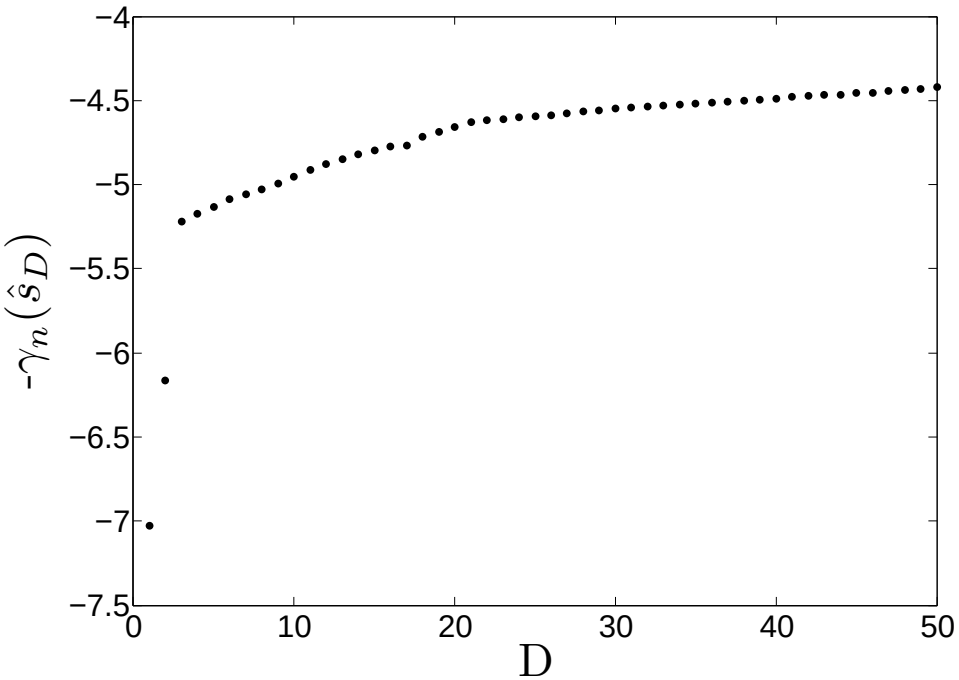


Thresholding the jump
(Arlot & Massart, 09)

Slope Heuristics: Data-driven Slope Estimation

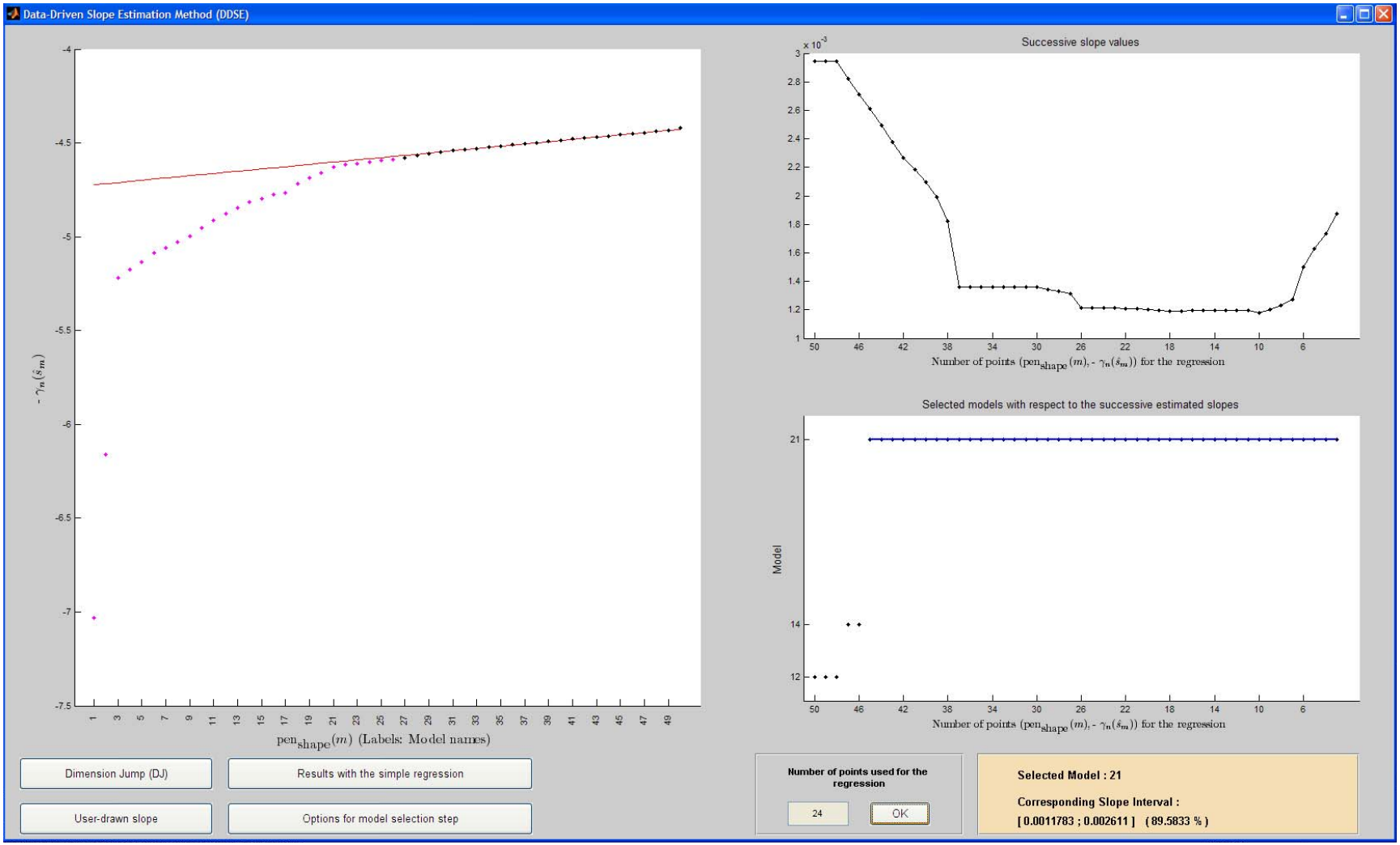
$$\begin{aligned}\gamma_n(\hat{s}_D) - \gamma_n(s) &= \gamma_n(\hat{s}_D) - \gamma_n(s_D) + \quad \gamma_n(s_D) - \gamma_n(s) \\ &\approx \qquad -\hat{v}_D \qquad\quad + \mathbb{E}\big[\gamma_n(s_D) - \gamma_n(s)\big].\end{aligned}$$

$$\begin{aligned}\gamma_n(\hat{s}_D) &\approx \gamma_n(s) + \ell(s, s_D) - \hat{v}_D \\ &\approx \gamma_n(s) + \ell(s, s_D) - \kappa_{min} D. \qquad \text{(for example)}\end{aligned}$$

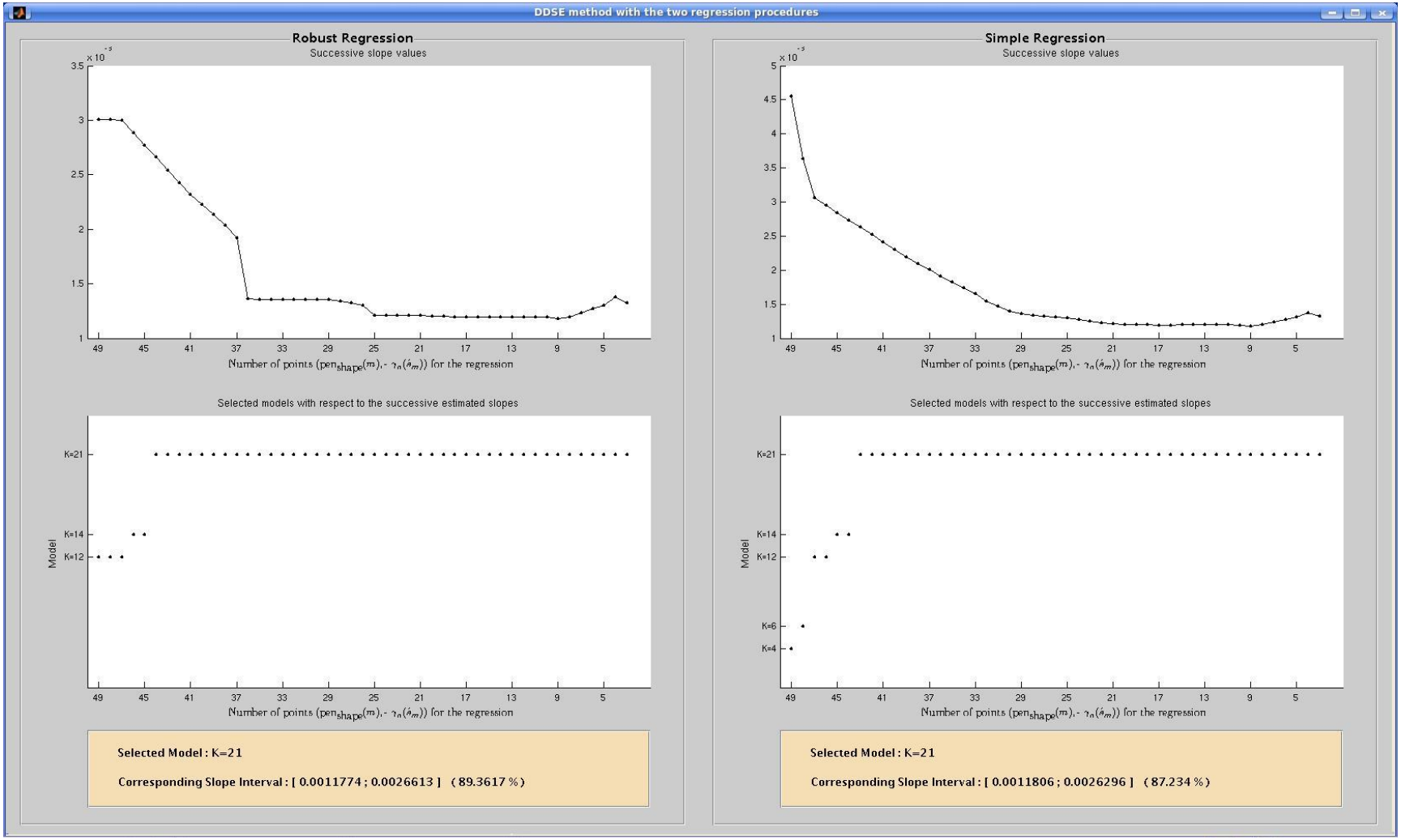


Slope Heuristics: Data-driven Slope Estimation

Recall: $\gamma_n(\hat{s}_D) \approx \gamma_n(s) + \ell(s, s_D) - \kappa_{min} D.$



Slope Heuristics: Data-driven Slope Estimation



Comparing robust and simple regression

"Bubbles" Experiment: Results

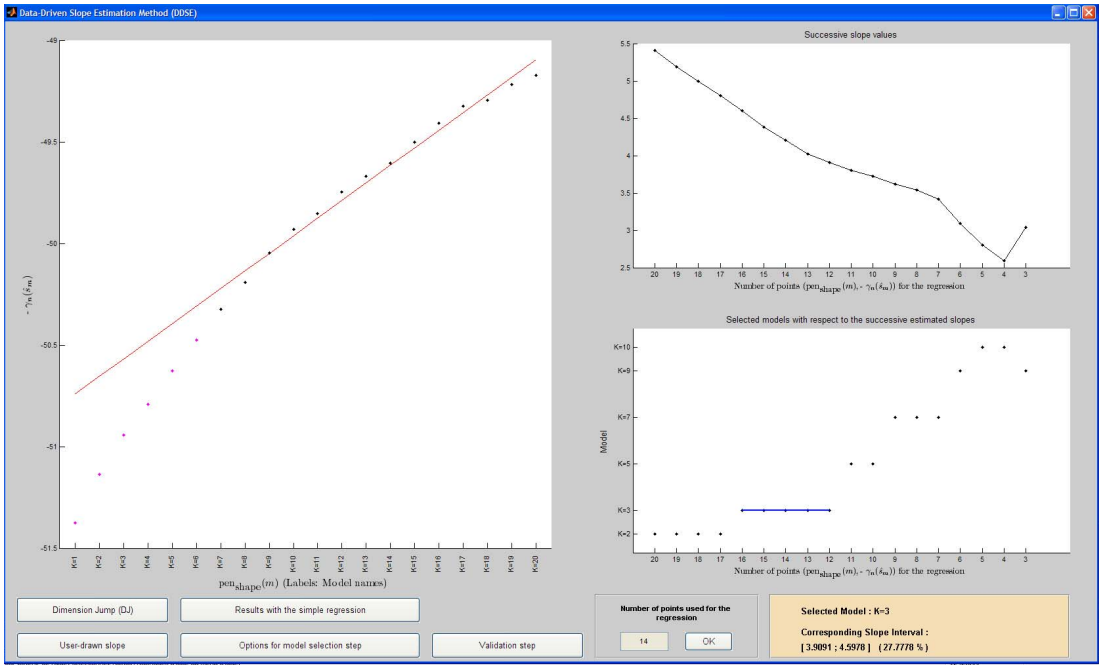
$\hat{m}$	3–4	15–18	19	20	21	22	23	24	25	≥ 35	Risk ratio
Or.				1	**76**	15	3	3	2		1
AIC										**100**	2.59
BIC		3	6	23	**57**	9	1	1			1.17
DDSE			3	7	**59**	20	6	3	2		1.06
DJ	6		3	7	**59**	18	2	3	2		1.49

Table: Number of times a model m is selected among the 100 simulations by AIC, BIC, the data-driven slope estimation method (DDSE) and the dimension jump method (DJ). The last column is the ratio between the risk of the selected estimator and the oracle risk.

Slope Heuristics: Validation Step

Transcriptome Dataset

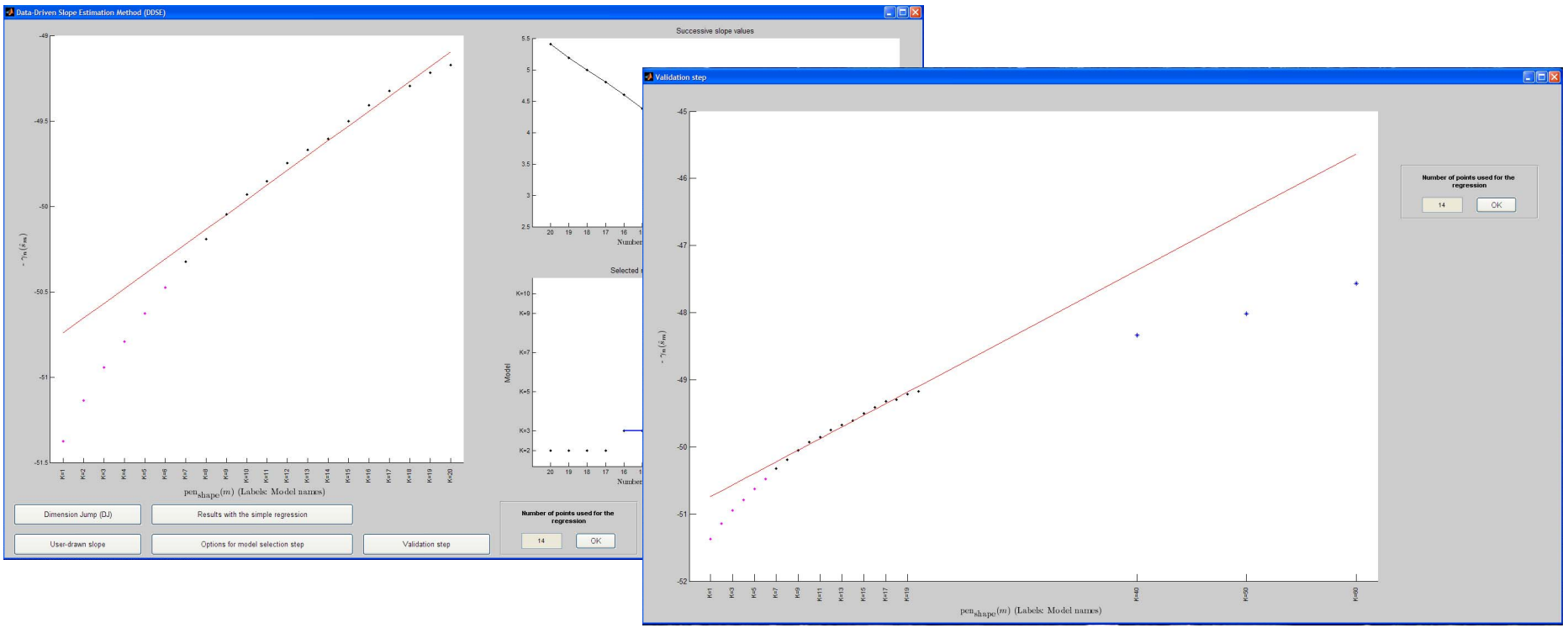
- Dataset studied by Maugis, Celeux, Martin-Magniette (07).
- $n = 1020$ genes described in 20 experiments (data in $\mathbb{R}^{20}$).
- Gaussian mixture models with equal component covariance matrices are considered.
- The model collection $(S_m)_{1 \leq m \leq 20}$ was first considered.



Slope Heuristics: Validation Step

Transcriptome Dataset

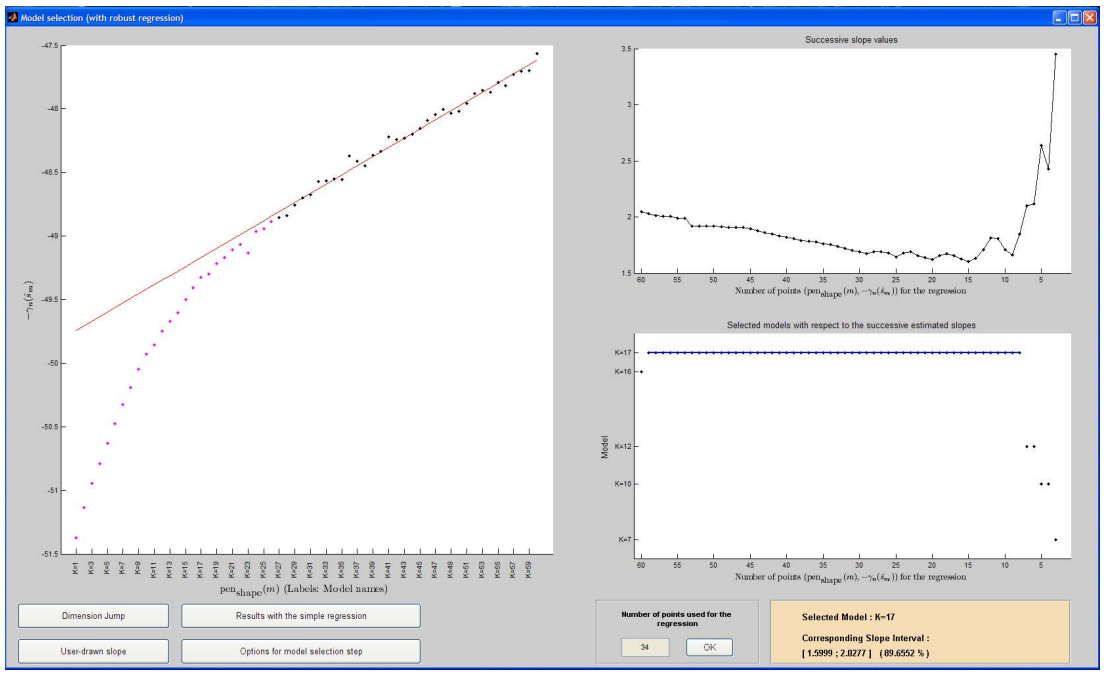
- Dataset studied by Maugis, Celeux, Martin-Magniette (07).
- $n = 1020$ genes described in 20 experiments (data in $\mathbb{R}^{20}$).
- Gaussian mixture models with equal component covariance matrices are considered.
- The model collection $(S_m)_{1 \leq m \leq 20}$ was first considered.



Slope Heuristics: Validation Step

Transcriptome Dataset

- Dataset studied by Maugis, Celeux, Martin-Magniette (07).
- $n = 1020$ genes described in 20 experiments (data in $\mathbb{R}^{20}$).
- Gaussian mixture models with equal component covariance matrices are considered.
- The model collection $(S_m)_{1 \leq m \leq 60}$ has been eventually considered.



Conclusion and Perspectives

Slope Heuristics:

- *Theoretically justified in some frameworks:* Gaussian least squares homoscedastic regression, fixed design (Birgé & Massart, 07); General heteroscedastic regression, random design, with histograms (Arlot & Massart, 09); Density estimation (Lerasle, 09); Gaussian random Markov field (Verzelen, 09)...
- *Encouraging applications in various frameworks without theoretical justification:* Estimation of oil reserves (Lepez, 02); Change point detection (Lebarbier, 05); Genomics (Villers, 07); Variable selection and clustering with Gaussian mixture models (Maugis & Michel, 10); Graph selection for computational geometry (Caillerie & Michel, 09); Model-based clustering (Baudry, 09)...
- *Theoretical and practical works are in progress:* Arlot & Bach, 09; Boucheron & Massart, 10...

Thank you for attention!

"Bubbles" Experiment: Results

d number ponents $\hat{m}$	$M_{\max}$	3	4	15–18	19	20	21	22	23	24	25	≥ 35
	40, 50					1	**76**	15	3	3	2	
	40, 50											**100**
	40, 50			3	6	23	**57**	9	1	1		
	50				3	7	**59**	20	6	3	2	
	40			1	3	7	**61**	18	4	4	2	
	50	4	2		3	7	**59**	18	2	3	2	
	40	**28**	2		2	4	**51**	10	2	1		

Table: Number of times a model m is selected among the 100 simulations by AIC, BIC, the data-driven slope estimation method (DDSE) and the dimension jump method (DJ). The last column is the ratio between the risk of the selected estimator by each method and the oracle risk.