Gaussian Mixture Regression model with logistic weights, a penalized maximum likelihood approach

E. Le Pennec L. Montuelle

LMO-Université Paris Sud
Select-INRIA Saclay

Rencontres des jeunes statisticiens
août 2013

Outline

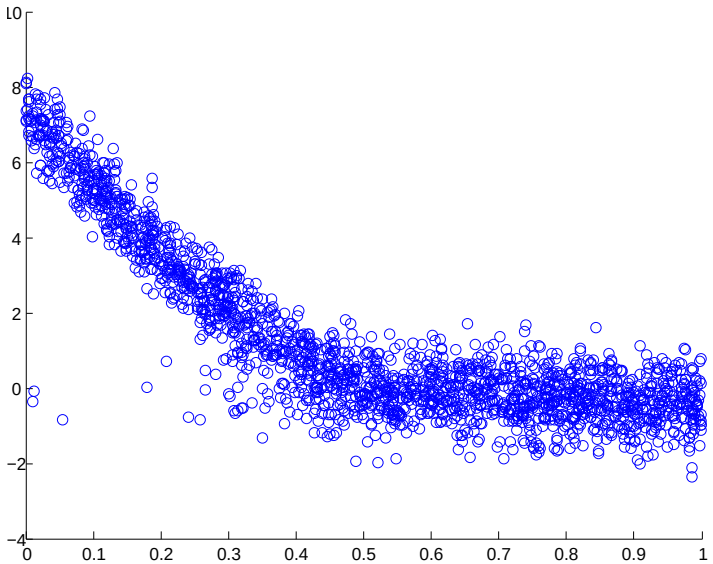
Outline

Motivation

Data: $(X_i, Y_i)_{i \leq n} \in [0;1]^d \times \mathbb{R}^p$

- $X_i \perp\!\!\!\perp X_j$
- $Y_i|(X_k)_k \perp\!\!\!\perp Y_j|(X_k)_k$
- $Y|X$ has a density s_0 w.r.t. Lebesgue measure

Aim: Estimate the conditional density $s_0(.|x)$

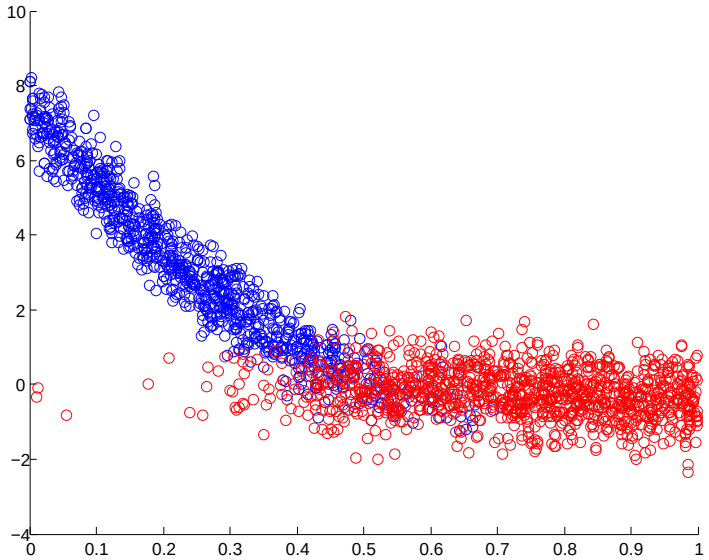


Motivation

Data: $(X_i, Y_i)_{i \leq n} \in [0;1]^d \times \mathbb{R}^p$

- $X_i \perp\!\!\!\perp X_j$
- $Y_i|(X_k)_k \perp\!\!\!\perp Y_j|(X_k)_k$
- $Y|X$ has a density s_0 w.r.t. Lebesgue measure

Aim: Estimate the conditional density $s_0(.|x)$



Framework

Family of densities:

$$s_{K,\upsilon,\Sigma,w}(y|x) = \sum_{k=1}^{K} \pi_{w(x),k} \Phi_{\upsilon_k(x),\Sigma_k}(y),$$

- $\pi_{w,k}(x) = \dfrac{e^{w_k(x)}}{\sum_{k'=1}^{K} e^{w_{k'}(x)}}$, logistic weights
- $\Phi_{\upsilon_k(x),\Sigma_k}$ density of $\mathcal{N}(\upsilon_k(x), \Sigma_k)$
- $(w_1, \ldots, w_K) \in W_K$, $(\upsilon_1, \ldots, \upsilon_K) \in \Upsilon_K$, $(\Sigma_1, \ldots, \Sigma_K) \in V_K$
  W_K, Υ_K functional spaces
  V_K covariance matrices set

Outline

Maximum likelihood approach and model selection

Model S_m defined by a choice $m = (K, W_K, \Upsilon_K, V_K)$

$$S_m = \left\{ s_{K,\upsilon,\Sigma,w} \middle| w \in W_K, \upsilon \in \Upsilon_K, \Sigma \in V_K \right\}$$

Design of the estimator:

$$\widehat{s}_m = \underset{s_{K,\upsilon,\Sigma,w} \in S_m}{\operatorname{argmax}} \sum_{i=1}^{n} \ln s_{K,\upsilon,\Sigma,w}(Y_i|X_i)$$

Model's choice:

$$\widehat{m} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \sum_{k=1}^{K} -\ln \widehat{s}_m(Y_i|X_i) + \operatorname{pen}(m).$$

$\Rightarrow$ Bias-variance trade-off

Theoretical result

Main assumption

There exist C_W and C_Υ such that, for every model S_m,
$H(\delta, W_K) \leq \dim(W_K)\left(C_W + \ln\frac{1}{\delta}\right)$, $H(\delta, \Upsilon_K) \leq \dim(\Upsilon_K)\left(C_\Upsilon + \ln\frac{1}{\delta}\right)$.

where H denotes the metric entropy.
KL Küllback-Leibler divergence
$KL^{\otimes n}$ tensorized Küllback-Leibler divergence

$$KL^{\otimes n}(s,t) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} KL(s(.|X_i), t(.|X_i))\right)$$

and $JKL_\rho^{\otimes n}$ Jensen-Küllback-Leibler divergence

$$JKL_\rho^{\otimes n}(s,t) = \frac{1}{\rho} KL^{\otimes n}(s, (1-\rho)s + \rho t)$$

Theoretical result

Main assumption

There exist C_W and C_Υ such that, for every model S_m, $H(\delta, W_K) \leq \dim(W_K)\left(C_W + \ln \frac{1}{\delta}\right)$, $H(\delta, \Upsilon_K) \leq \dim(\Upsilon_K)\left(C_\Upsilon + \ln \frac{1}{\delta}\right)$.

Oracle inequality (M. 2012)

There exists a constant C such that for any $\rho \in (0; 1)$,and any $C_1 > 1$, there are two constants κ et C_2, depending only on ρ and C_1, such that

$$\mathbb{E}\left[JKL_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{m \in \mathcal{M}} \left(\inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\text{pen}(m)}{n} \right) + \frac{C_2}{n},$$

with $\text{pen}(m) = \kappa(C + \ln n)\dim(S_m)$.

Outline

Model collection for the choice of K

Let $X \in [0;1]$ and $Y \in \mathbb{R}$,

$$S_K = S_m = \left\{ (x,y) \mapsto \sum_{k=1}^{K} \pi_{w(x),k} \Phi_{\upsilon_k(x),\Sigma_k}(y) \middle| w \in W_K, \upsilon \in \Upsilon_K, \Sigma \in V_K \right\},$$

with $V_K = V^K$, where V is the set of all covariance matrices,

$$W_K = \Upsilon_K = \mathcal{D}^K,$$

$$\mathcal{D} = \left\{ x \mapsto ax + b | (a,b) \in \mathbb{R}^2 \right\}.$$

Numerical experiments

❶ $s_0 \in (S_m)_{m \in \mathcal{M}}$

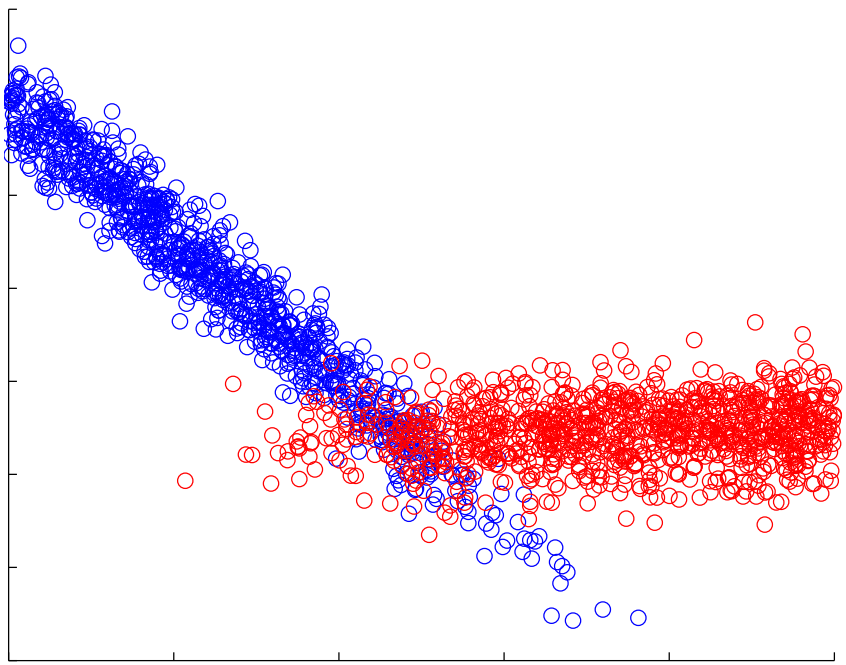


Figure: 2000 points; affine means

❷ $s_0 \notin (S_m)_{m \in \mathcal{M}}$

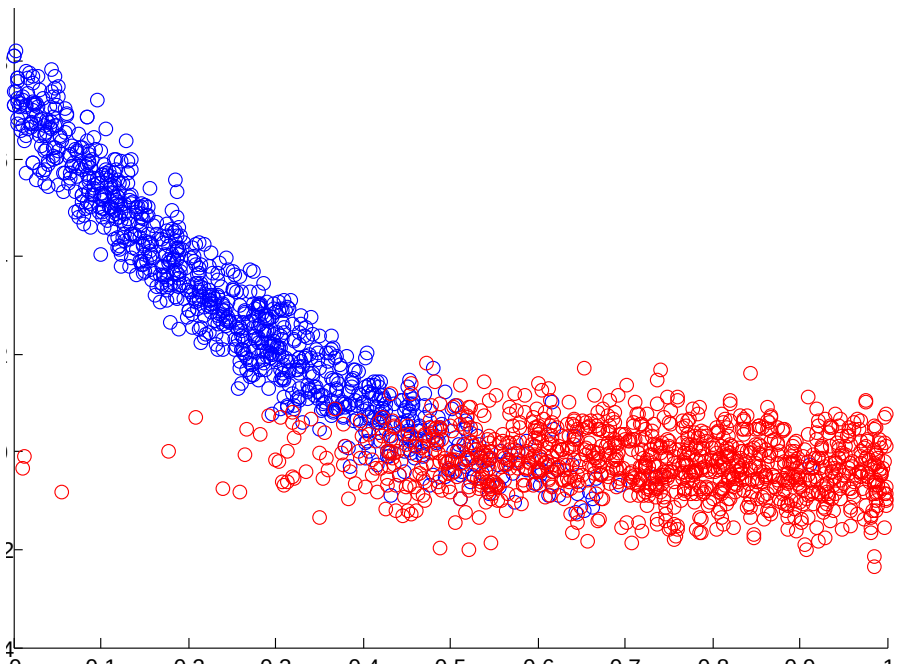


Figure: 2000 points; parabolic means

Description of Newton-EM algorithm

Newton-EM

Parameters given by initialization
Iterate until convergence:

Newton algorithm over weights if the likelihood increases, up to 5 iterations

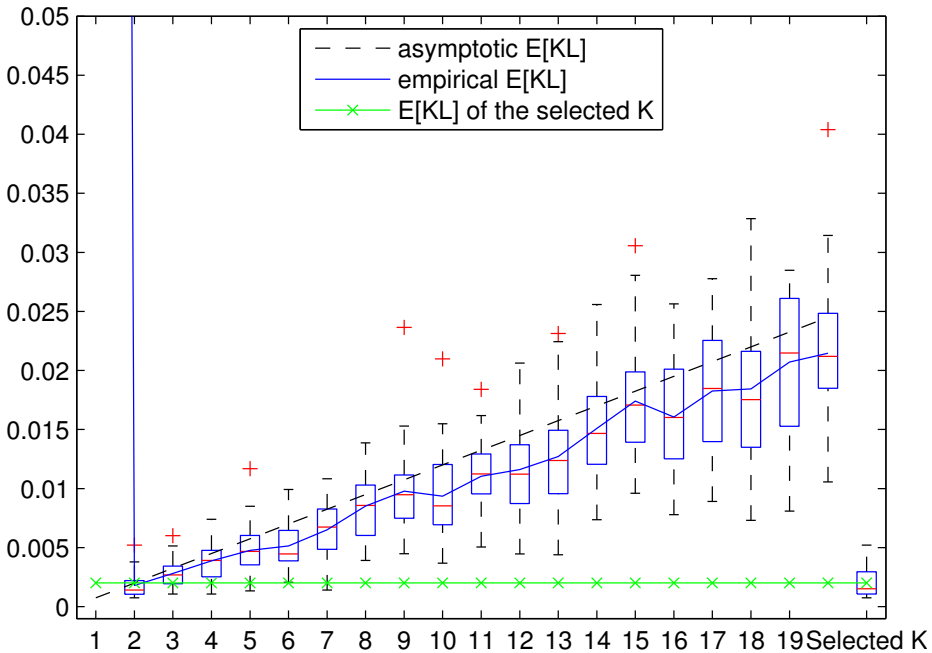
Linear regression to update mean and variance in each class

Newton-EM initialization

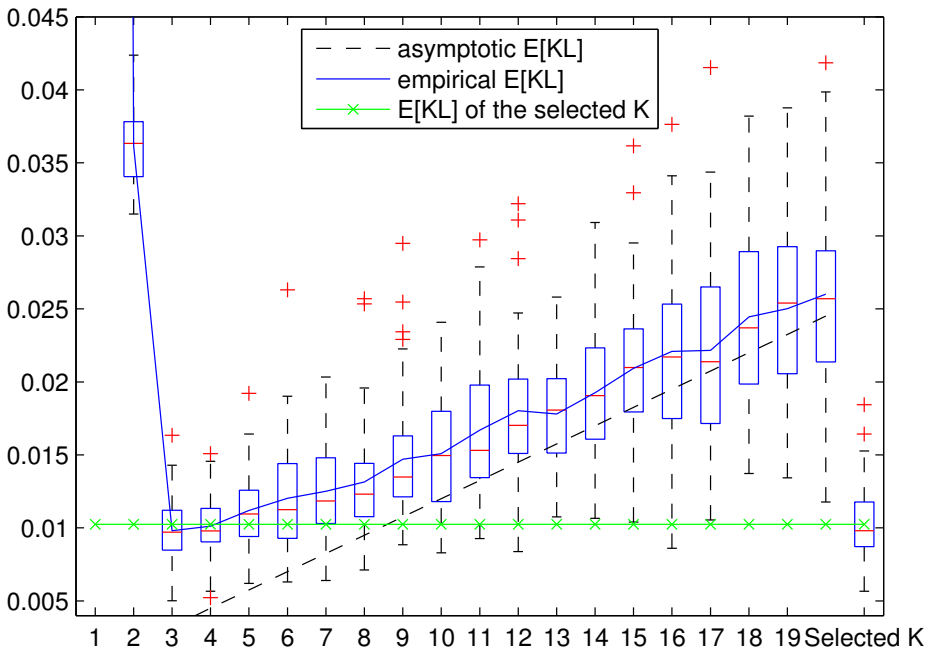
- Draw K couples of points (X_i, Y_i) uniformly among the data, defining K lines
- Classify the data: $k = \text{argmin}_l |Y_i - \upsilon_l(X_i)|$
- 3 iterations of Newton-EM
- Repeat 50 times the previous steps

Chosen initialization: the one among the 50 with the greatest likelihood

Evolution of the risk according to the number of mixture components - 2 000 data points

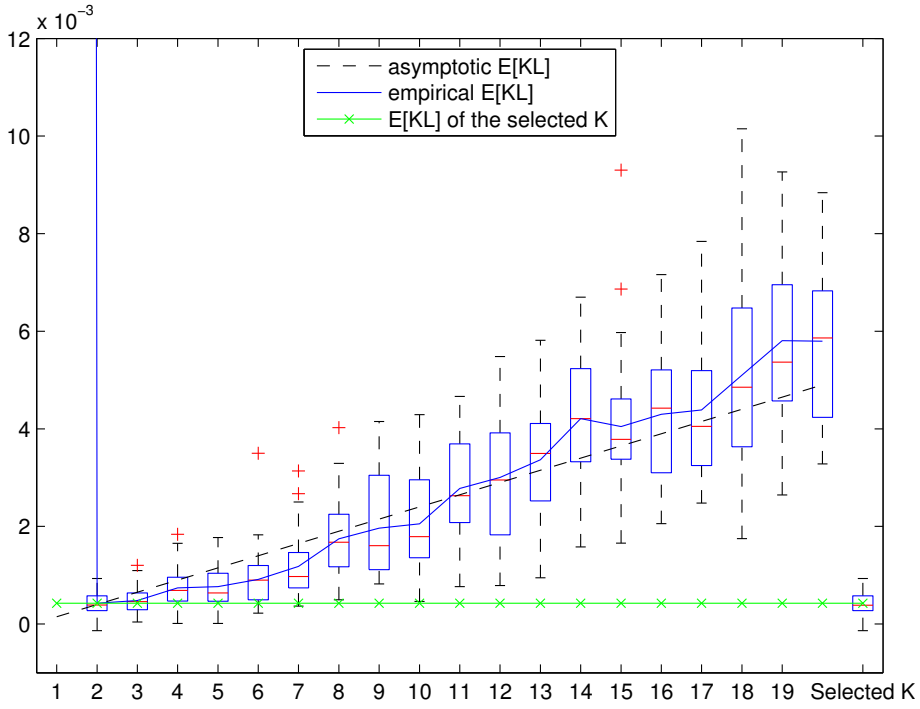


❶ $s_0 \in (S_m)_{m \in \mathcal{M}}$

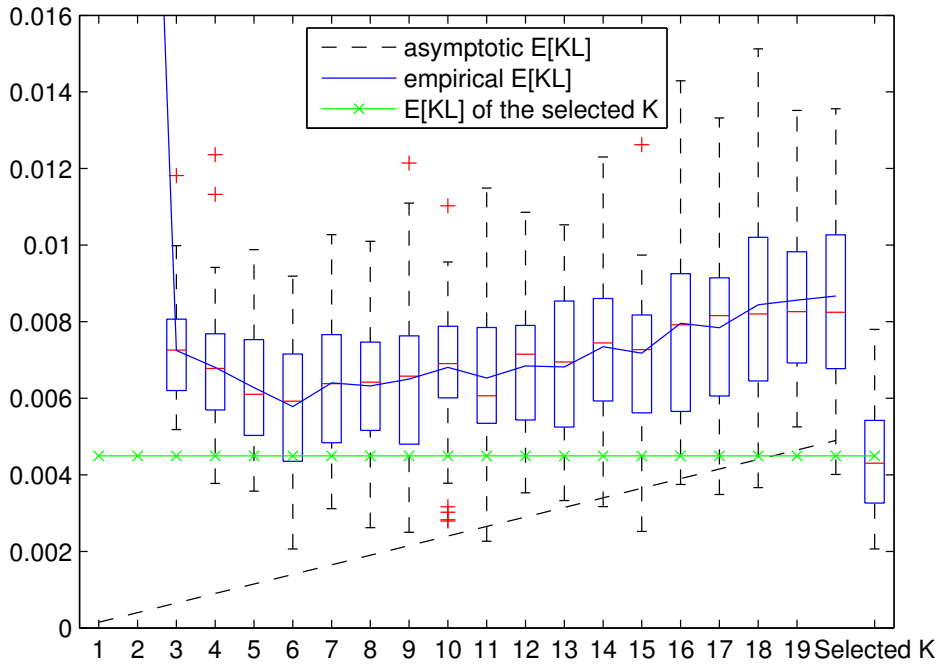


❷ $s_0 \notin (S_m)_{m \in \mathcal{M}}$

Evolution of the risk according to the number of mixture components - 10 000 data points

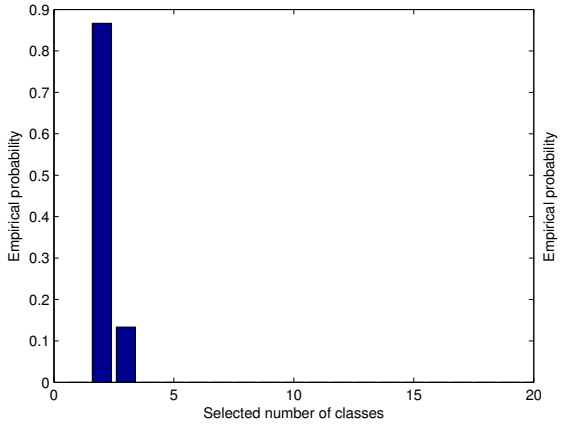


❶ $s_0 \in (S_m)_{m \in \mathcal{M}}$

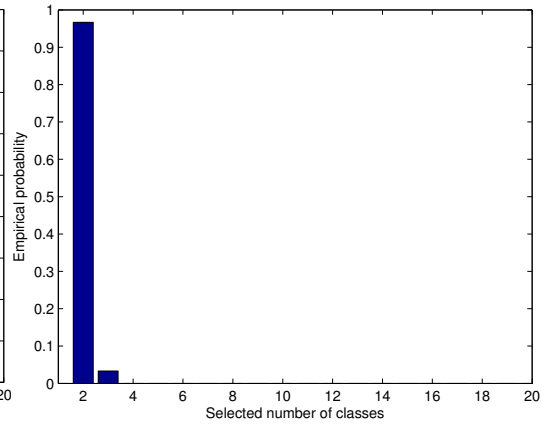


❷ $s_0 \notin (S_m)_{m \in \mathcal{M}}$

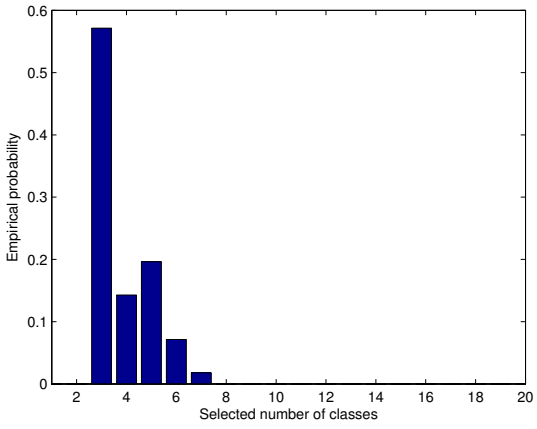
Histograms of the selected K



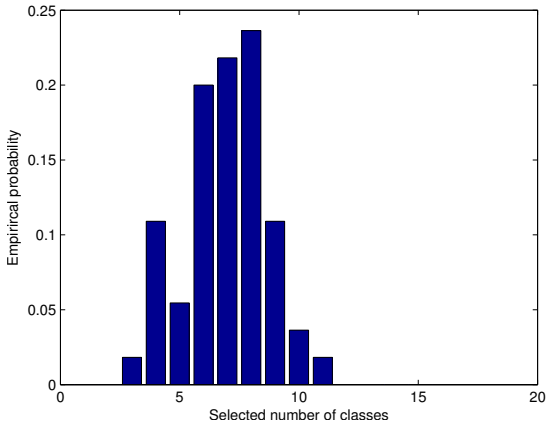
(a) $s_0 \in (S_m)_{m\in\mathcal{M}}$-2 000 points



(b) $s_0 \in (S_m)_{m\in\mathcal{M}}$-10 000 points



(c) $s_0 \notin (S_m)_{m\in\mathcal{M}}$-2 000 points



(d) $s_0 \notin (S_m)_{m\in\mathcal{M}}$-10 000 points

Conclusion

Results:

- $\text{pen}(m) = \kappa(C + \ln n)\dim(S_m)$ provides an oracle inequality
- Performant algorithm

Advantages over partitioning methods:

- Use of gradient descent
- Oblique frontier in dimension 2

→ Application to hyperspectral images segmentation

Thank you for your attention

Article:

Gaussian Mixture Regression model with logistic weights, a penalized maximum likelihood approach, `http://fr.arxiv.org/abs/1304.2696`