Comparison of methods for estimating the non zero components of a Gaussian vector.

Sylvie HUET[1]

Rapport technique 2005-4, 21 pp.

Unit Mathmatiques et Informatique Appliques
INRA
Domaine de Vilvert
F-78352 Jouy-en-Josas Cedex

[1] sylvie.huet@jouy.inra.fr

Abstract. We propose a method based on a penalised likelihood criterion, for estimating the number of non-zero components of the mean of a Gaussian vector. Following the work of Birgé and Massart in Gaussian model selection, we choose the penalty function such that the resulting estimator minimises the Kullback risk. A simulation study compares the proposed method to others based on penalised likelihood criterion and on thresholding. Two applications are shown: the first one concerns the differential analysis of macro-array data, and the second one the analysis of un-replicated factorial designs.

Résumé. Un critère de choix de modèle basé sur la vraisemblance pénalisée est proposé afin d'estimer le nombre de composantes non nulles de l'espérance d'un vecteur gaussien. Comme proposé par Birgé et Massart pour le problème de sélection de variables dans le modèle gaussien, la fonction de pénalité est calculée afin de minimiser le risque Kullback de l'estimateur. La méthode est comparée, à l'aide de simulations, aux méthodes basées sur des critères de vraisemblance pénalisée ainsi qu'aux méthodes de seuillage. La méthode est comparée, à l'aide de simulations, aux méthodes basées sur des critères de vraisemblance pénalisée ainsi qu'aux méthodes de seuillage. Elle est appliquée à l'analyse différentielle des gènes et à l'estimation dans un plan factoriel.

Subjclass. 62G05, 62G09

Keywords. Kullback risk, Model selection, Penalised likelihood criteria, differential analysis, macro-array, factorial designs

Introduction

The following regression model is considered:

$$\boldsymbol{X} = \boldsymbol{m} + \tau\boldsymbol{\varepsilon}, \;\; \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, I_n),$$

where $\boldsymbol{X} = (X_1, \dots X_n)^T$ is the vector of observations. The expectation of $\boldsymbol{X}$, say $\boldsymbol{m} = (m_1, \dots, m_n)^T$, and the variance τ^2 are unknown. Assuming that some of the components of $\boldsymbol{m}$ are equal to zero, our objective is to estimate the number of zero components as well as their positions.

We denote by J a subset of $J_n = \{1, 2, \dots, n\}$ with dimension k_J and by J^c its complement in J_n. We consider the collection $\mathcal{J}$ of all subsets of J_n with dimension less than k_n for some k_n less than n:

$$\mathcal{J} = \{J \subset J_n, k_J \leq k_n\}.$$

Let $\boldsymbol{x} = (x_1, \dots, x_n)^T$, then for each subset $J \in \mathcal{J}$ we denote by $\boldsymbol{x}_J$ the vector in $\mathbb{R}^n$ whose component i equals x_i if i belongs to J and 0 if not.

For each subset J in the collection $\mathcal{J}$, assuming that $\boldsymbol{m} = \boldsymbol{m}_J$, the maximum likelihood estimators of the parameters $(\boldsymbol{m}, \tau^2)$ are $(\boldsymbol{X}_J, \widehat{\sigma}_J^2)$, where

$$\widehat{\sigma}_J^2 = \frac{1}{n} \sum_{i \in J^c} X_j^2,$$

and the maximum of the log-likelihood equals $-(n/2) \log(\widehat{\sigma}_J^2)$. The problem is to choose an estimator of $(\boldsymbol{m}, \tau^2)$ by choosing the best J in $\mathcal{J}$, say $\widehat{J}$, and taking $(\widehat{\boldsymbol{m}}, \widehat{\tau}^2) = (\boldsymbol{X}_{\widehat{J}}, \widehat{\sigma}_{\widehat{J}}^2)$. We associate to each estimator in the collection a risk defined as

$$R(J) = \mathrm{E}\left\{\mathcal{K}_{(\boldsymbol{m},\tau^2)}\left(\boldsymbol{X}_J, \widehat{\sigma}_J^2\right)\right\}, \tag{1}$$

where for all $\boldsymbol{g} \in \mathbb{R}^n$ and σ positive, $\mathcal{K}_{(\boldsymbol{m},\tau^2)}(\boldsymbol{g}, \sigma^2)$ denotes the Kullback-Leibler divergence:

$$\mathcal{K}_{(\boldsymbol{m},\tau^2)}\left(\boldsymbol{g}, \sigma^2\right) = \frac{n}{2}\left\{\log\left(\frac{\sigma^2}{\tau^2}\right) - 1 + \frac{\tau^2 + \sum_{i=1}^n (m_i - g_i)^2/n}{\sigma^2}\right\}.$$

The ideal subset J^*, defined as the minimiser of the risk over all the subsets in the collection,

$$R(J^*) = \inf_{J \in \mathcal{J}} R(J),$$

is estimated by a model selection procedure. Namely, J^* is estimated by $\widehat{J}$ that minimises a penalised likelihood criterion defined as follows:

$$\mathrm{crit}(J, \mathrm{pen}) = \frac{n}{2} \log\left(\widehat{\sigma}_J^2\right) + \mathrm{pen}(k_J), \tag{2}$$

where pen is a penalty function that depends on k_J.

Several criteria have been proposed in the literature, the most famous ones being the Akaike and the Schwarz criteria.

The Akaike criterion [1] with

$$\mathrm{pen}_{AIC}(k) = k \tag{3}$$

is based on an asymptotic expansion of the Kullback-Leibler divergence calculated in the maximum likelihood estimator of $(\boldsymbol{m}, \tau^2)$ on one subset J with dimension k. It can be shown that $\mathrm{crit}(J, \mathrm{pen}_{AIC})$ is an asymptotically unbiased estimator of $\mathrm{E}\left[\mathcal{K}_{(m,\tau^2)}(\widehat{\boldsymbol{X}}_J, \widehat{\sigma}_J^2)\right]$ (up to some terms that do not depend on k).

The SIC criterion,

$$\mathrm{pen}_{SIC}(k) = \frac{1}{2} k \log(n), \tag{4}$$

was proposed by Schwartz [19] and Akaike [2]. Schwartz derived its penalty function from Bayesian arguments and asymptotic expansions. It is shown by Nishii [14] that if the penalty function is written as

$\text{pen}(k) = c_n k$ such that $c_n \to \infty$ and $c_n/n \to 0$, then $\widehat{J}_{c_n} = \widehat{J}(\text{pen})$ converges to J_0 in probability, where J_0 is the set of indices on which the components of $\boldsymbol{m}$ are non zero.

The *AMDL* (for approximate minimum description length) criterion proposed by Rissanen [17]

$$\text{pen}_{AMDL}(k) = \frac{3}{2} k \log(n) \tag{5}$$

was studied by Antoniadis et al. [3] for determining the number of nonzero coefficients in the vector of wavelet coefficients.

The calculation of the penalty function of the criterion we propose is based on the following equality:

$$R(J) = \mathcal{K}_{(\boldsymbol{m},\tau^2)}\left(\boldsymbol{m}_J, \mathrm{V}_J(\boldsymbol{X})\right) + \mathrm{E}\left\{\mathcal{K}_{(\boldsymbol{m}_J,\mathrm{V}_J(\boldsymbol{X}))}\left(\boldsymbol{X}_J, \widehat{\sigma}_J^2\right)\right\}, \tag{6}$$

where $\mathrm{V}_J(\boldsymbol{X})$ is defined as follows:

$$\mathrm{V}_J(\boldsymbol{X}) = \frac{1}{n}\mathrm{E}\left\{\sum_{i=1}^{n}(X_i - m_i \mathbb{I}_{i\in J})^2\right\} = \frac{k}{n}\tau^2 + \frac{1}{n}\sum_{i\in J^c} m_i^2.$$

The quantity $\mathcal{K}_{(\boldsymbol{m},\tau^2)}\left(\boldsymbol{m}_J, \mathrm{V}_J(\boldsymbol{X})\right)$ is analogous to a bias term: it represents the distance between the expectation and the variance of $\boldsymbol{X}$ under the model J and the parameters $(\boldsymbol{m}, \tau^2)$. It is equal, up to some terms that do not depend on J, to $(n/2)\log(\mathrm{V}_J(\boldsymbol{X}))$. If $\mathrm{V}_J(\boldsymbol{X})$ is estimated by $\widehat{\sigma}_J^2$, the second term on the right hand side of Equation (6) is analogous to a *variance term*. The penalty function is calculated so that it compensates both this *variance term* and the bias due to the estimation of $\log(\mathrm{V}_J(\boldsymbol{X}))$ by $\log(\widehat{\sigma}_J^2)$.

In [13] it is shown that if the penalty function is written as follows:

$$\text{pen}(k) = n\left\{c_1 \log\left(\frac{n}{k}\right) + c_2\right\}\frac{k}{n-k} \tag{7}$$

for some constants c_1, c_2, then

$$\mathrm{E}\left[\mathcal{K}_{(\boldsymbol{m},\tau^2)}\left(\widehat{\boldsymbol{m}}, \widehat{\tau}^2\right)\mathbb{I}_\Omega\right] \leq C_1\left(\inf_{J\in\mathcal{J}}\left\{\mathcal{K}_{(\boldsymbol{m},\tau^2)}\left(\boldsymbol{m}_J, \mathrm{V}_J(\boldsymbol{X})\right) + \text{pen}(k_J)\right\}\right) + C_2$$

where C_1 and C_2 are some constants and Ω is a set of probability greater than $1 - \kappa/n$.

In this paper, we calculate the constant (c_1, c_2) in the penalty function by a simulation study. We compare our procedure to procedures based on generalisations of the AIC criterion, and to procedures based on the penalisation of the residual sum of squares : the method proposed by Birgé and Massart [6] and the threshold methods. We present

an extensive simulation study to evaluate the behaviour of the methods in different situations, considering the cases where the number of non-zero components in $\boldsymbol{m}$ equals 0, is small and large.

The method is applied to the differential analysis of macro-array data where the purpose is to determine significant variation of gene expression under different treatments and to the analysis of un-replicated factorial and fractional factorial designs such as those in screening experiments.

1. The method

Let $\widehat{J}(\text{pen})$ be the subset in $\mathcal{J}$ that minimises $\text{crit}(J, \text{pen})$ defined in Equation (2),

$$\widehat{J}(\text{pen}) = \arg\min_{J\in\mathcal{J}} \text{crit}(J, \text{pen}). \tag{8}$$

The components of $\boldsymbol{m}$ that are not estimated by 0, correspond to the greatest absolute values of the components of $\boldsymbol{X}$. Let $\{\ell_1, \ldots, \ell_n\}$ be the order statistics, such that the $X^2_{\ell_i}$ are sortered in the descending order: $X^2_{\ell_1} \geq \ldots \geq X^2_{\ell_n}$, and let J_k be the subset of $\mathcal{J}$ corresponding to the k first order statistics, $J_k = \{\ell_1, \ldots, \ell_k\}$. Then for all subset J in $\mathcal{J}$ with dimension $k_J = k$, we have

$$\frac{n}{2}\log\left(\widehat{\sigma}^2_{J_k}\right) + \text{pen}(k) \leq \frac{n}{2}\log\left(\widehat{\sigma}^2_{J}\right) + \text{pen}(k_J),$$

and the problem reduces to choose k, less than n, that minimises $\text{crit}(J_k, \text{pen})$, say $\widehat{k}$, and to take $\widehat{J} = J_{\widehat{k}}$.

2. Comparison with other criteria.

2.1. **The penalised residual sum of squared approach.** Birgé and Massart [6] provided a general approach to model selection via penalisation for Gaussian regression with known variance. For each $J \in \mathcal{J}$, the Kullback risk for the maximum likelihood estimator of $\boldsymbol{m}$ when $\boldsymbol{m} = \boldsymbol{m}_J$ is the quadratic risk defined as follows:

$$Q(J) = \frac{1}{2}\text{E}\left\{\sum_{i=1}^{n}(X_i\mathbb{1}_{i\in J} - m_i)^2\right\} \tag{9}$$

and the likelihood penalised estimator is defined by $\widehat{J}(\underline{\text{pen}}) = \arg\min_{J\in\mathcal{J}} \underline{\text{crit}}(J, \underline{\text{pen}})$, where

$$\underline{\text{crit}}(J, \underline{\text{pen}}) = \frac{n}{2}\widehat{\sigma}^2_J + \underline{\text{pen}}(k_J). \tag{10}$$

We call this estimator a RSS-penalised estimator and we denote the penalty function by $\underline{\text{pen}}$ to highlight that the residual sum-of-squares

is penalised, not its logarithm. They propose to choose the penalty function as follows:

$$\underline{\text{pen}}(k) = \tau^2 \left\{ c_1 \log\left(\frac{n}{k}\right) + c_2 \right\},$$

for some constants c_1, c_2. For practical issues τ^2 has to be estimated and the penalty function calibrated. This last point is discussed in section 3.

2.2. **Threshold estimators.** Another important class of estimators is the class of threshold estimators. The estimator of $\boldsymbol{m}$ equals $\boldsymbol{X}_{\widehat{J}}$ where $\widehat{J}$ is defined as the set of indices i in $\{1, \ldots, n\}$ such that the absolute value of X_i is greater than a threshold t. The method consists in choosing a decreasing sequence $t(k)$ of positive numbers and comparing the order statistics $X^2_{\ell_1}, \ldots, X^2_{\ell_n}$ to $t^2(1), \ldots, t^2(n)$. Then define

$$\left.\begin{array}{rcl} \widehat{k} & = & 0 \text{ if } X^2_{\ell_k} < t^2(k) \ \ \forall k \geq 1 \\ \widehat{k} & = & \max_k \left\{ X^2_{\ell_k} \geq t^2(k) \right\} \text{ if not} \end{array}\right\}, \tag{11}$$

and choose $\widehat{t} = t(\widehat{k})$. The link between threshold and penalised estimators is done as follows: $\widehat{k}$ is the location of the right most local minimum of the quantity $\underline{\text{crit}}(J_k, \underline{\text{pen}})$ defined at Equation (10) by taking the following penalty function:

$$\begin{array}{rcl} \underline{\text{pen}}(0) & = & 0 \\ \underline{\text{pen}}(k) & = & \dfrac{1}{2} \displaystyle\sum_{l=1}^{k} t^2(l), \text{ if } k \geq 1. \end{array}$$

The link between the threshold estimator and a logRSS-penalised is done in the same way: the threshold estimator defined a logRSS-penalised estimator by setting

$$\begin{array}{rcl} \text{pen}(0) & = & 0 \\ \text{pen}(k) & = & \dfrac{n}{2} \displaystyle\sum_{l=1}^{k} \log\left(1 + \frac{t^2(l)}{n\widehat{\sigma}^2_{J_l}}\right), \text{ if } k \geq 1 \end{array}$$

in $\text{crit}(J_k, \text{pen})$ defined at Equation (2).

For analysing un-replicated factorial and fractional factorial designs, several authors proposed threshold estimators. See for example Box and Meyer [11], Lenth [18], and Haaland and O'Connell [15]. The idea is to choose a threshold that should provide a powerful testing procedure for identifying non-null effects. Lenth [18] proposed a threshold

estimator based on constant $t(k)$'s. He proposed to estimate τ as follows:

$$\widehat{\tau} = 1.5 \times \text{median}\{|X_i| \text{ for } |X_i| < 2.5 s_0\}, \tag{12}$$

and he defined a simultaneous margin of error with approximately 95% confidence by taking

$$t_{SME} = \widehat{\tau}\, T^{-1}\left(\gamma_n, \frac{n}{3}\right), \tag{13}$$

where $\gamma_n = (1 + 0.95^{1/n})/2$ and $T^{-1}(\gamma, d)$ denotes the γ-quantile of a student variable with d degrees of freedom. The choice $d = n/3$ comes from the comparison of the empirical distribution of $\widehat{\tau}^2$ to chi-squared distribution.

For the problem of testing simultaneously several hypotheses, Benjamini and Hochberg [5] proposed a procedure that controls the false discovery rate. Precisely, the procedure seeks to ensure that at most a fraction q of the rejected null hypotheses corresponds to false rejections. It corresponds to a threshold estimator with

$$t(k) = \widehat{\tau}\Phi^{-1}\left(1 - \frac{qk}{2n}\right) \text{ and } t_{FDR} = t(\widehat{k}), \tag{14}$$

where $\widehat{k}$ is defined by Equation (11). It can be shown that the penalty function $\underline{\text{pen}}_{FDR}(k)$ is of order $k \log(n/k)$.

Foster and Stine [12] compared the performances of adaptive variable selection to that obtained by Bayes expert and proposed an approximate empirical Bayes estimator defined as a threshold estimator with

$$t(k) = \widehat{\tau}\sqrt{2\log\left(\frac{n}{k}\right)} \text{ and } t_{FS} = t(\widehat{k}), \tag{15}$$

where $\widehat{k}$ is defined by Equation (11).

3. Choice of the constants in the penalty function

3.1. **The minimum Kullback risk estimator.** From a practical point of view, we want to have in hand a penalty function such that the risk associated with the corresponding estimator is as close as possible to $R(J^*)$, the minimum of the risks associated with the sets J in the collection $\mathcal{J}$, denoted in the following $\mathcal{R}_n^*(\boldsymbol{m}, \tau^2)$. The penalised estimator defined by Equations (2) and (8) is slightly modified to take into account the cases where the function $\text{crit}(J_k, \text{pen})$ is not a convex function of k. In these cases $\widehat{k}$ is the location of the right most local minimum of $\text{crit}(J_k, \text{pen})$.

Let us denote by $\mathcal{R}_n(\boldsymbol{m}, \tau^2, \text{pen})$ the risk associated with $\widehat{J}(\text{pen})$, namely

$$\mathcal{R}_n(\boldsymbol{m}, \tau^2, \text{pen}) = \mathrm{E}\left\{\mathcal{K}_{(\boldsymbol{m},\tau^2)}\left(\boldsymbol{X}_{\widehat{J}(\text{pen})}, \widehat{\sigma}^2_{\widehat{J}(\text{pen})}\right)\right\}.$$

We are looking for a penalty function that minimises, uniformly for all $(\boldsymbol{m}, \tau^2)$, the risk ratio:

$$r_n(\text{pen}) = \sup_{\boldsymbol{m}} \frac{\mathcal{R}_n(\boldsymbol{m}, \tau^2, \text{pen})}{\mathcal{R}^*_n(\boldsymbol{m}, \tau^2)}.$$

Note that τ^2 has been removed from the supremum because $\mathrm{E}\left\{\mathcal{K}_{(\boldsymbol{m},\tau^2)}\left(\boldsymbol{X}_J, \widehat{\sigma}^2_J\right)\right\}$ depends on the pair $(\boldsymbol{m}, \tau^2)$ only through the ratio $\boldsymbol{m}/\tau$.

We are looking for (c_1, c_2) such that the penalty function defined by Equation (7) minimises the risk ratio $r_n(\text{pen})$, denoted in the following by $r_n(c_1, c_2)$.

The risk ratio is estimated by simulations. We consider several values of n, say $n = 2^j$ for j equals 4 to 10. For each value of n, we take $k_n = n/2$, and we consider several values of k_0, namely $k_0 \in \{0\} \cup K$ where K is a subset of integers greater than 1 and smaller than $n/3$. For each n and each $k_0 \in K$, we consider six vectors $\boldsymbol{m}^\ell(k_0)$, for $\ell = 1, \ldots, 6$ as follows: for $i = 1, \ldots, k_0$, $m^1_i(k_0) = 5$, $m^2_i(k_0) = 5i^{-0.25}$, $m^3_i(k_0) = 5i^{-0.5}$, $m^4_i(k_0) = 5i^{-1}$, $m^5_i(k_0) = 10i^{-0.5}$, $m^6_i(k_0) = 10i^{-0.25}$; for $i > k_0$, $m^\ell_i(k_0) = 0$.

Taking $\tau = 1$, for each n and each $\boldsymbol{m}^\ell(k_0)$ we estimate $\mathcal{R}^*_n(\boldsymbol{m}^\ell(k_0), \tau^2)$ as the empirical mean of values obtained on 500 simulations. Then, for several values of (c_1, c_2), we estimate the quantities $\mathcal{R}_n(\boldsymbol{m}^\ell(k_0), \tau^2, \text{pen}(c_1, c_2))$ on the basis of 500 simulations. Finally we calculate the supremum over ℓ and k_0 of the ratios of these two estimators to get an estimation of $r_n(c_1, c_2)$. The variations of $r_n(c_1, c_2)$ are shown in Figure 1. When $c_2 = 0$, it appears that the value of c_1 that minimises the risk ratio decreases with n. When c_2 increases, the minimum value of the risk ratio decreases. When $c_2 = 4$, and n large enough, the minimum value is close to 2 and is attained for a value of c_1 close to 2. When $n = 16$ and $n = 32$, the minimum value of the risk ratio is larger than 2 (around 6) and the minimum is attained for c_1 around 4 when $n = 32$ and around 8 when $n = 16$. This suggests that for small values of n, the penalty function should be modified. For example, adding a term in $\sqrt{\log(n/k)}$ in the penalty function, might improve the behaviour of the risk ratio. We chose not to pursue in this way and to define the minimum Kullback risk estimator with the following penalty function:

$$\text{pen}_{MKR}(k) = n\left\{\log\left(\frac{n}{k}\right) + 2\right\}\frac{k}{n-k}. \tag{16}$$

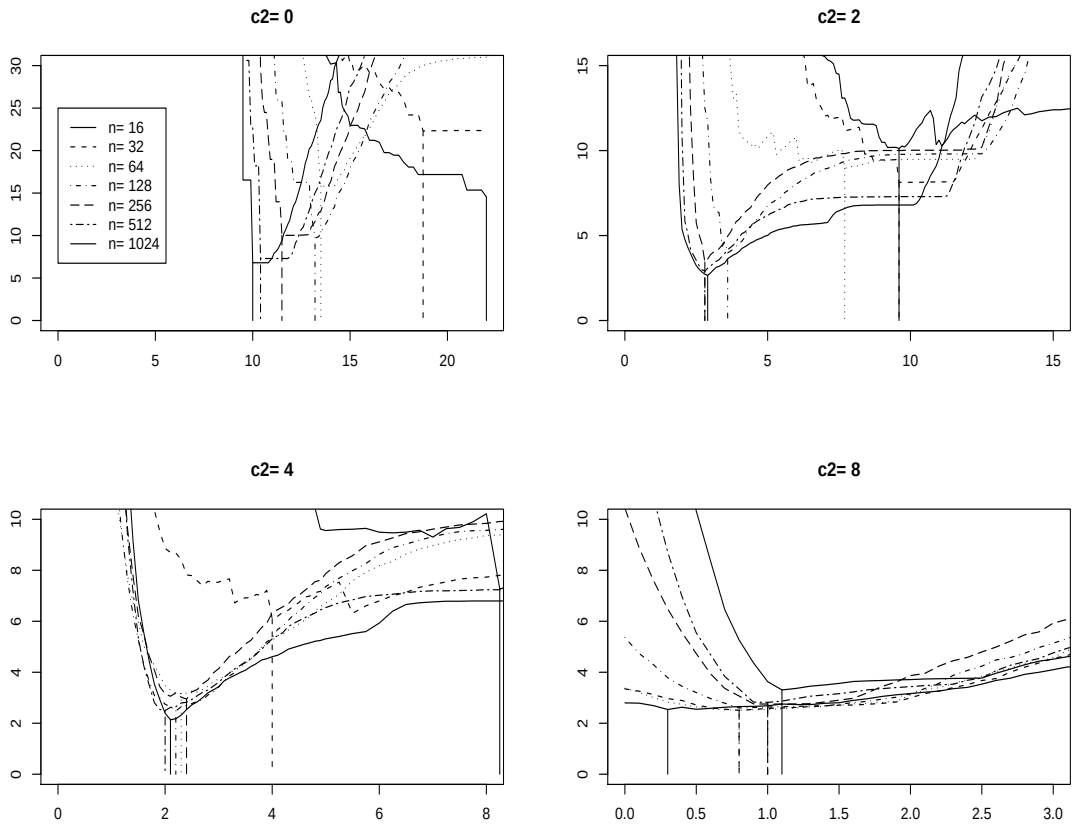

FIGURE 1. For each value of n, variations of $r(c_1, c_2)$ as a function of c_1, for different values of c_2.

The simulation study reported in section 4 evaluates its performances both for small and large values of n.

3.2. **The estimator proposed by Birgé and Massart.** The approach is similar to the preceding one. We are looking for a penalty function such that the associated quadratic risk is as close as possible to $\underline{\mathcal{R}^*}_n(\boldsymbol{m}, \tau^2) = \inf_{J \in \mathcal{J}}\{Q(J)\}$ where $Q(J)$ is the quadratic risk defined at Equation (9). We are thus looking for a penalty function that minimises, uniformly for all $(\boldsymbol{m}, \tau^2)$ the risk ratio defined as follows:

$$\underline{r}_n(\underline{\text{pen}}) = \sup_{\boldsymbol{m} \in \mathbb{R}^{n,*}} \frac{\underline{\mathcal{R}}_n(\boldsymbol{m}, \tau^2, \underline{\text{pen}})}{\underline{\mathcal{R}^*}_n(\boldsymbol{m}, \tau^2)},$$

where $\underline{\mathcal{R}}_n(\boldsymbol{m}, \tau^2, \underline{\text{pen}})$ is the risk associated with $\widehat{J}(\underline{\text{pen}})$ defined by Equation (10), and $\mathbb{R}^{n,*}$ denotes the set of $\mathbb{R}^n$ vectors whose components are not all equal to zero. Taking $\tau^2 = 1$, we are looking for the best constants c_1, c_2 in the penalty function. For each n and each $m^\ell(k_0), k_0 \in K$ defined previously, we calculate $\underline{\mathcal{R}^*}_n(\boldsymbol{m}^\ell(k_0), \tau^2)$. Then for several values of (c_1, c_2) we estimate the quantities $\underline{\mathcal{R}}_n(\boldsymbol{m}^\ell(k_0), \tau^2, \underline{\text{pen}}(c_1, c_2))$ on the basis of 500 simulations. The value of (c_1, c_2) that minimises the estimated risk ratio are found to be $c_1 = 1$ and $c_2 = 2$. Therefore the estimator of Birgé and Massart will be defined with the following penalty function:

$$\underline{\text{pen}}_{BM}(k) = \widehat{\tau}^2 \left\{ \log\left(\frac{n}{k}\right) + 2 \right\} k, \tag{17}$$

where $\widehat{\tau}^2$ is a suitable variance estimate. For the sake of comparison with the FS threshold estimators, let us note that $(1/k)\sum_{l=1}^{k} \log(k/l)$ is smaller than 1. Therefore we get the following inequality

$$\underline{\text{pen}}_{BM}(k) \geq \underline{\text{pen}}_{FS}(k) + \widehat{\tau}^2 k.$$

It shows that the large values of k are less likely to be found using the BM criteria than the FS criteria.

4. Comparison with other criteria via simulation

In this section we compare the performances of our criterion with others. We consider the following criteria :

- Criteria based on penalised logarithm of the residual sum of squares
  - The SIC criterion defined at Equation (4).
  - The AMDL criterion defined at Equation (5)
  - The MKR criterion, that aims at minimising the Kullback risk, defined at Equation (16).
- Threshold estimators or criteria based on penalised residual sum of squares. For these estimators we need to choose a variance estimator. Following the results given by Haaland and O'Connell [15] we chose the estimator given by Lenth [18], see Equation (12), that should generally perform well for moderate to large numbers of non-null effects.
  - The SME estimator defined at Equation (13).
  - The FDR estimator defined at Equation (14) with $q = 0.05$.
  - The FS estimator defined at Equation (15).
  - The estimator proposed by Birgé and Massart using the penalty function defined at Equation (17).

We carry out a simulation study considering several models defined as follows: we choose

- 3 values of n, $n = 20, 100, 5000$,
- for each value of $n = 20$ and $n = 100$, we consider 3 values for k_0, $k_0 = 0, n/20, n/4$, For $n = 5000$ we consider $k_0 = 0, 10, 1250$.
- when k_0 is non null, we set for $i = 1, \ldots, k_0$, $m_i = \mu$ with $\mu = 5$.

Proceeding as in Section 3, taking $\tau = 1$, we calculate $\underline{\mathcal{R}^*}_n(k_0)$ and for each $J \in \mathcal{J}$ we estimate $R(J)$ as the empirical mean based on 1000 simulations. The calculation of $\mathcal{R}^*_n(k_0) = \inf_{J\in\mathcal{J}} \{R(J)\}$ follows. Moreover, for each $\widehat{k}$ defined above and each simulation, we calculate the quadratic and the Kullback distances respectively denoted by $\underline{\mathcal{K}}_n(k_0)$ and $\mathcal{K}_n(k_0)$, and the false discovery ratio defined when $\widehat{k}$ is positive as

$$\begin{aligned} \rho_n(k_0) &= 0 \text{ if } \widehat{k} = 0 \\ \rho_n(k_0) &= \frac{\sum_{i\in J_{\widehat{k}}} \mathbf{I}_{i>k_0}}{\widehat{k}} \text{ if } \widehat{k} > 0. \end{aligned}$$

On the basis of 1000 simulations, we compare the methods by comparing the following quantities:

(1) the empirical means and the medians of the quadratic and Kullback distances denoted respectively $\underline{\mathcal{K}}_n^{\text{mean}}(k_0, \mu, \ell)$, $\underline{\mathcal{K}}_n^{\text{median}}(k_0, \mu, \ell)$, $\mathcal{K}_n^{\text{mean}}(k_0, \mu, \ell)$, $\mathcal{K}_n^{\text{median}}(k_0, \mu, \ell)$.
(2) the empirical means of the false discovery ratios
(3) the distributions of $\widehat{k}$

Results for the case $k_0 = 0$. The results given at Tables 1, 2 and 3 suggest the following comments:

- The Kullback risk is very sensitive to overestimation. Let us note that the difference between the mean and the median of the Kullback risk ratios is very large. This is a consequence of the convergence of the Kullback-Leibler divergence $\mathcal{K}_{(\boldsymbol{m},\tau^2)}(\boldsymbol{g}, \sigma^2)$ towards infinity when σ^2 tends to zero. Because $\widehat{\sigma}^2_{J_k}$ decreases when k increases, the risk ratio explodes when the criterion leads to over-fitting.
- The best methods are SME, MKR, FDR and BM whatever the value of n, and AMDL when $n = 100$ or $n = 5000$.
- When $n = 20$, all the methods, except SME, tend to overestimate k_0, see Table 3. When $n = 100$, MKR and AMDL give the same results than SME. When $n = 5000$, the BM method corrects this tendency while the probability for $\widehat{k}$ to be strictly positive is estimated by 14.3% for the FS method.

Table (a)

n	SIC	AMDL	MKR	SME	FDR	FS	BM	$\mathcal{R}_n^*(0)$
20	389	80	15	1.4	8.9	48	13	0.657
100	2000	1.2	1.8	1.4	2.8	17	3.3	0.590
5000	246	1.1	1.2	1.4	1.5	6.6	1.3	0.468

Table (b)

n	SIC	AMDL	MKR	SME	FDR	FS	BM
20	318	0.66	0.46	0.40	0.56	0.97	0.55
100	1930	0.37	0.39	0.39	0.47	1.10	0.44
5000	245	0.46	0.47	0.48	0.48	0.73	0.47

TABLE 1. Kullback risk ratios when $k_0 = 0$: Table (a) gives the ratios $\mathcal{K}_n^{\text{mean}}(0)/\mathcal{R}_n^*(0)$; the last column gives the values of the minimum Kullback risk. Table (b) gives the ratios $\mathcal{K}_n^{\text{median}}(0)/\mathcal{R}_n^*(0)$.

n	SIC	AMDL	MKR	SME	FDR	FS	BM
20	9.3	1.6	0.55	0.13	0.86	2.02	0.91
100	46	0.17	0.36	0.26	0.83	4.03	0.91
5000	107	0	0.08	0.17	0.19	2.57	0.14

TABLE 2. Quadratic risks when $k_0 = 0$: $\underline{\mathcal{K}}_n^{\text{mean}}(0)$. The median of the quadratic risks equal 0 for all methods except SIC.

n	SIC	AMDL	MKR	SME	FDR	FS	BM
20	100%	19.7%	8.5%	3%	15.5%	29.9%	14.8%
100	100%	2.2%	4%	3.8%	10.3%	32.5%	8.6%
5000	100%	0%	0.7%	1.6%	1.7%	14.3%	0.9%

TABLE 3. Results when $k_0 = 0$: estimated probabilities for $\widehat{k}$ to be positive.

- The SIC method does not penalised enough the high dimensions: when $n = 20$ the SIC criteria decreases from 0 to k_n. As a consequence, k_0 is estimated by $k_n = 10$. When $n = 100$, k_0 is estimated by $k_n = 50$ in 997 cases (over 1000 simulations). When $n = 5000$, the range of values for $\widehat{k}$ equals $[5, 43]$.
- The AMDL method improves the SIC method, especially when $n = 100$ and $n = 5000$. When $n = 20$, the probability for $\widehat{k}$ to be positive equals 19.7% and the number of $\widehat{k}$ equal to 10 is 62 (over 1000 simulations).

	SIC	AMDL	MKR	SME	FDR	FS	BM
$\widehat{k}=0$	0	5.1	9.4	45.5	6.6	2.2	7.9
$\widehat{k}=1$	0	71.9	73.4	52.5	72.4	46.8	68.1
$\widehat{k}>1$	100	23	17	20	21	51	24
r^{mean}	171	37	15	3.8	10	35	12
r^{median}	134	0.8	0.7	3.5	0.8	4	0.9
$\underline{r}^{\text{mean}}$	18	5.5	5.1	12	4.5	7.5	5.3
$\underline{r}^{\text{median}}$	18	0.8	0.7	3.7	0.9	5.1	1.0
ρ	90	17	11	1.2	13	36	15

TABLE 4. Case $n = 20$, $k_0 = 1$, $\mu = 5$, $\ell = 0$. The three first lines give the percentage of simulations for which $\widehat{k} = 0$, $\widehat{k} = 1$ and $\widehat{k} > 1$. The four following lines give respectively $r^{\text{mean}} = \mathcal{K}_n^{\text{mean}}(k_0, \mu, \ell)/\mathcal{R}_n^*(k_0, \mu, \ell)$, $r^{\text{median}} = \mathcal{K}_n^{\text{median}}(k_0, \mu, \ell)/\mathcal{R}_n^*(k_0, \mu, \ell)$, $\underline{r}^{\text{mean}} = \underline{\mathcal{K}}_n^{\text{mean}}(k_0, \mu, \ell)/\underline{\mathcal{R}^*}_n(k_0, \mu, \ell)$, and $\underline{r}^{\text{median}} = \underline{\mathcal{K}}_n^{\text{median}}(k_0, \mu, \ell)/\underline{\mathcal{R}^*}_n(k_0, \mu, \ell)$. The last line ρ gives the mean of the false discovery ratios.

Results when k_0 is positive and small. Let us discuss the results when $n = 20$ and $k_0 = 1$ given at Table 4, $n = 100$ and $k_0 = 5$ given at Table 5, and $n = 5000$ and $k_0 = 10$ given at Table 6. In all cases the values of k that minimise the quadratic and the Kullback risk ratios equal k_0.

- The best methods are MKR, FDR and BM. They tend to over estimate k_0 when $n = 20$ and $n = 100$ and to underestimate k_0 when $n = 5000$ and $k_0 = 10$.
- The behaviour of the AMDL method depends strongly on n. When $n = 20$ it tends to overestimate k_0, while when $n = 100$ and $n = 5000$, k_0 is underestimated. When $n = 5000$, the method is the worse (after SIC) both from the point of view of the Kullback and quadratic risks.
- The SME method always underestimates k_0, while the FS methods always overestimates k_0.
- The SIC method fails to work: the range of the values of $\widehat{k}$ is $[8, 10]$ when $n = 20$ and $k_0 = 1$, $[39, 50]$ when $n = 100$ and $k_0 = 5$ and $[18, 51]$ when $n = 5000$ and $k_0 = 10$.

Results when k_0 is large. The results are given for $n = 20, 100, 5000$, $k_0 = n/4$ and $\mu = 5$ at Tables 7, 8 and 9. In any case the value of k that minimises the quadratic risk ratio equal k_0. On the other hand,

	SIC	AMDL	MKR	SME	FDR	FS	BM
$\widehat{k} < 4$	0	16.5	0.9	19.9	2.2	1	1.4
$\widehat{k} \in [4,6]$	0	83.4	87.8	79.7	87.8	56.6	80.5
$\widehat{k} > 6$	100	0.1	11.3	0.4	10	42.4	18.1
r^{mean}	220	2.1	1.75	2.3	1.9	6.2	2.2
r^{median}	212	1.2	1.	2.4	1.1	3.1	1.4
$\underline{r}^{\text{mean}}$	18	4.2	2.4	4.5	2.7	4.8	2.8
$\underline{r}^{\text{median}}$	18	1.4	1.4	4	1.4	3.9	1.9
ρ	90	0.6	4.2	0.6	7.3	30.2	10.8

TABLE 5. Case $n = 100$, $k_0 = 5$, $\mu = 5$, $\ell = 0$. The three first lines give the percentage of simulations for which $\widehat{k} < 4$, $\widehat{k} \in [4,6]$ and $\widehat{k} > 6$. The four following lines give respectively $r^{\text{mean}} = \mathcal{K}_n^{\text{mean}}(k_0,\mu,\ell)/\mathcal{R}_n^*(k_0,\mu,\ell)$, $r^{\text{median}} = \mathcal{K}_n^{\text{median}}(k_0,\mu,\ell)/\mathcal{R}_n^*(k_0,\mu,\ell)$, $\underline{r}^{\text{mean}} = \underline{\mathcal{K}}_n^{\text{mean}}(k_0,\mu,\ell)/\underline{\mathcal{R}^*}_n(k_0,\mu,\ell)$, and $\underline{r}^{\text{median}} = \underline{\mathcal{K}}_n^{\text{median}}(k_0,\mu,\ell)/\underline{\mathcal{R}^*}_n(k_0,\mu,\ell)$. The last line ρ gives the mean of the false discovery ratios.

	SIC	AMDL	MKR	SME	FDR	FS	BM
$\widehat{k} < 8$	0	96.2	11	55.1	14.4	1.4	9.9
$\widehat{k} \in [8,12]$	0	3.8	84.7	44.9	83.5	55.2	85
$\widehat{k} > 12$	100	0	4.3	0	2.1	43.4	5.1
r^{mean}	5.6	3.1	1.1	1.7	1.1	1.6	1.1
$\underline{r}^{\text{mean}}$	23	14	4.9	7.5	4.9	6.8	4.9
ρ	68.3	0	6.7	0.2	4.4	23	7.2

TABLE 6. Case $n = 5000$, $k_0 = 10$, $\mu = 5$, $\ell = 0$. The three first lines give the percentage of simulations for which $\widehat{k} < 8$, $\widehat{k} \in [8,12]$ and $\widehat{k} > 12$. The two following lines give respectively $r^{\text{mean}} = \mathcal{K}_n^{\text{mean}}(k_0,\mu,\ell)/\mathcal{R}_n^*(k_0,\mu,\ell)$ and $\underline{r}^{\text{mean}} = \underline{\mathcal{K}}_n^{\text{mean}}(k_0,\mu,\ell)/\underline{\mathcal{R}^*}_n(k_0,\mu,\ell)$. The values of r^{median} and $\underline{r}^{\text{median}}$ are not given because they are closed to r^{mean} and $\underline{r}^{\text{mean}}$. The last line ρ gives the mean of the false discovery ratios.

k^*, that minimises the Kullback risk ratio, equals k_0 when $n = 20$ and $n = 100$. For $n = 5000$, k^* is smaller than $k_0 = 1250$ and varies around 1245.

	SIC	AMDL	MKR	SME	FDR	FS	BM
$\widehat{k} < 4$	0	3.5	4.3	69.5	10.2	1.4	6.1
$\widehat{k} \in [4,6]$	0	69.2	89.4	40.2	80.4	55.2	79.1
$\widehat{k} > 6$	100	27.3	6.3	0.3	9.4	43.4	14.8
r^{mean}	21	7.5	2.8	2.6	3.2	11.1	4.2
r^{median}	16	1.9	1.2	2.8	1.6	4.5	1.8
$\underline{r}^{\text{mean}}$	3.4	4.3	2.4	4.5	2.7	4.8	2.8
$\underline{r}^{\text{median}}$	73.3	1.6	1.1	11	1.5	2.3	1.6
ρ	50	13.3	4.2	0.3	5.6	24.3	8.8

TABLE 7. Case $n = 20$, $k_0 = 5$, $\mu = 5$, $\ell = 0$. The three first lines give the percentage of simulations for which $\widehat{k} < 4$, $\widehat{k} \in [4,6]$ and $\widehat{k} > 6$. The four following lines give respectively $r^{\text{mean}} = \mathcal{K}_n^{\text{mean}}(k_0,\mu,\ell)/\mathcal{R}_n^*(k_0,\mu,\ell)$, $r^{\text{median}} = \mathcal{K}_n^{\text{median}}(k_0,\mu,\ell)/\mathcal{R}_n^*(k_0,\mu,\ell)$, $\underline{r}^{\text{mean}} = \underline{\mathcal{K}}_n^{\text{mean}}(k_0,\mu,\ell)/\underline{\mathcal{R}^*}_n(k_0,\mu,\ell)$, and $\underline{r}^{\text{median}} = \underline{\mathcal{K}}_n^{\text{median}}(k_0,\mu,\ell)/\underline{\mathcal{R}^*}_n(k_0,\mu,\ell)$. The last line ρ gives the mean of the false discovery ratios.

- The best methods are MKR, FDR and BM.
- The BM method tends to overestimate k_0 but its results are close to the MKR and FDR methods.
- The FS method overestimates k_0 while the SME method underestimates k_0.
- The behaviour of the AMDL method depends on n. When $n = 20$, k_0 is overestimated, while when $n = 100$, k_0 is underestimated: 35.8% are smaller than 21 and 34.7% are equal to 0. When $n = 5000$ the AMDL method fails to work: the penalty function increases from 0 to k_n and $\widehat{k}$ is thus estimated by 0 in all simulations.
- When $n = 20$ and $n = 100$, the SIC method fails to work: $\widehat{k}$ is estimated by $k_n = n/2$ in all simulations. When $n = 5000$ the SIC method is one of the best.

Summary. The behaviour of AMDL and SIC methods depends on k_0, and is very bad in some cases. The FS method gives good results only when $k_0 = 0$ and overestimates k_0 in other cases. This was already noticed by several authors, see for example [15]. The MKR, FDR and BM methods give similar results. We note that the BM method tends to overestimate k_0. When $n = 20$ or $n = 100$ and k_0 is small, the FDR method tends to overestimate k_0.

	SIC	AMDL	MKR	SME	FDR	FS	BM
$\widehat{k} < 21$	0	35.8	0.1	56.2	2.5	0	0.5
$\widehat{k} \in [22, 28]$	0	64.1	96.8	43.8	95.2	21.7	85.2
$\widehat{k} > 28$	100	0.1	3.1	0	2.3	78.3	14.3
r^{mean}	20	2.1	1.3	2.0	1.4	6.7	1.8
r^{median}	20	1.5	1.1	1.9	1.2	4.7	1.5
$\underline{r}^{\text{mean}}$	3.4	9.6	1.5	6.1	1.7	2.4	1.7
$\underline{r}^{\text{median}}$	3.4	2.0	1.3	4.7	1.4	2.4	1.6
ρ	50	0.3	3.2	0.1	3.5	26.5	7.4

TABLE 8. Case $n = 100$, $k_0 = 25$, $\mu = 5$, $\ell = 0$. The three first lines give the percentage of simulations for which $\widehat{k} < 21$, $\widehat{k} \in [22, 28]$ and $\widehat{k} > 28$. The four following lines give respectively $r^{\text{mean}} = \mathcal{K}_n^{\text{mean}}(k_0, \mu, \ell)/\mathcal{R}_n^*(k_0, \mu, \ell)$, $r^{\text{median}} = \mathcal{K}_n^{\text{median}}(k_0, \mu, \ell)/\mathcal{R}_n^*(k_0, \mu, \ell)$, $\underline{r}^{\text{mean}} = \underline{\mathcal{K}}_n^{\text{mean}}(k_0, \mu, \ell)/\underline{\mathcal{R}^*}_n(k_0, \mu, \ell)$, and $\underline{r}^{\text{median}} = \underline{\mathcal{K}}_n^{\text{median}}(k_0, \mu, \ell)/\underline{\mathcal{R}^*}_n(k_0, \mu, \ell)$. The last line ρ gives the mean of the false discovery ratios.

	SIC	AMDL	MKR	SME	FDR	FS	BM
$\min\{\widehat{k}\}$	1279	0	1249	627	1238	1589	1291
$q(\widehat{k}, 5\%)$	1293	0	1274	719	1258	1651	1313
$q(\widehat{k}, 50\%)$	1314	0	1280	808	1275	1727	1338
$q(\widehat{k}, 95\%)$	1336	0	1286	886	1292	1810	1367
$\max\{\widehat{k}\}$	1364	0	1264	946	1312	1859	1391
r^{mean}	1.3	3.7	1.1	2.5	1.1	4.6	1.4
$\underline{r}^{\text{mean}}$	1.5	25	1.4	9.4	1.4	2.5	1.5
ρ	5.3	0	3.1	510^{-4}	2.7	27.7	6.9

TABLE 9. Case $n = 5000$, $k_0 = 1250$, $\mu = 5$, $\ell = 0$. The five first lines give the minimum , the 0.05, 0.5 and 0.95 quantiles, and the maximum of the distribution of the $\widehat{k}$'s. The two following lines give respectively $r^{\text{mean}} = \mathcal{K}_n^{\text{mean}}(k_0, \mu, \ell)/\mathcal{R}_n^*(k_0, \mu, \ell)$, and $\underline{r}^{\text{mean}} = \underline{\mathcal{K}}_n^{\text{mean}}(k_0, \mu, \ell)/\underline{\mathcal{R}^*}_n(k_0, \mu, \ell)$. The values of r^{median} and $\underline{r}^{\text{median}}$ are not given because they are closed to r^{mean} and $\underline{r}^{\text{mean}}$. The last line ρ gives the mean of the false discovery ratios.

5. Applications

5.1. **Application to the differential analysis of macro-array data.** Statistical methods for the identification of differentially expressed genes in array experiments have been extensively studied these last years (see for example Dudoit et al. [9], Efron et al. [10], Baldy and Long [4], Delmar et al. [8], Reiner et al. [16]). Our purpose is to illustrate how the methods presented in this paper apply to such data. Clearly they apply only when the variance of the observation is the same for all genes. Because this assumption may be satisfied in some data set, possibly after some suitable transformation of the data, it is worthwhile applying these methods to such data. The difference in gene expression when *Bacillus subtilis* is grown either on methionine or on methylthioribose as sulfur source [20] is observed in a DNA array experiment. Daudin et al. [7] presented a comparison of two statistical methods for analysing these data: the first one is based on the use of mixture models with three populations (negative, non-reactive and positive), the second one is based on the use of standard analysis of variance for detecting the negative or positive expressed genes. The estimated number of differentially expressed genes was 25 using the anova method and 35 using the mixture models approach. The data are composed of 4107 genes for which we observed a quantitative response Z linked to the gene expression. Usually the logarithmic transformation is used for normalising and stabilising the variance of the observations and we are interested in the variable

$$X = \log_{10}(Z_{met}) - \log_{10}(Z_{mtr})$$

where Z_{met} (respectively Z_{mtr}), is the observed response when the bacteria is grown on methionine (respectively methylthioribose). Following the preliminary analyses of Daudin et al. [7] we suppressed 52 highly variables genes that gave incoherent results. The criteria for calculating $\widehat{k}$ using the MKR, FDR and BM methods are given at Figure 2. The observed differences X and the estimated $\boldsymbol{m}$ are given at Figure 3.

5.2. **Analysis of un-replicated factorials and fractional factorial designs.** In industrial applications, fractionated and un-replicated factorial designs are used to identify important effects (or contrasts) among a large number of factors. These designs allow no degree of freedom for the error estimation. The literature on that subject is significant, see for example a recent paper by Wang and Voss [21] who proposed to construct confidence intervals for the effects, Box and Meyer [11] who proposed to calculate the posterior probability that each contrast

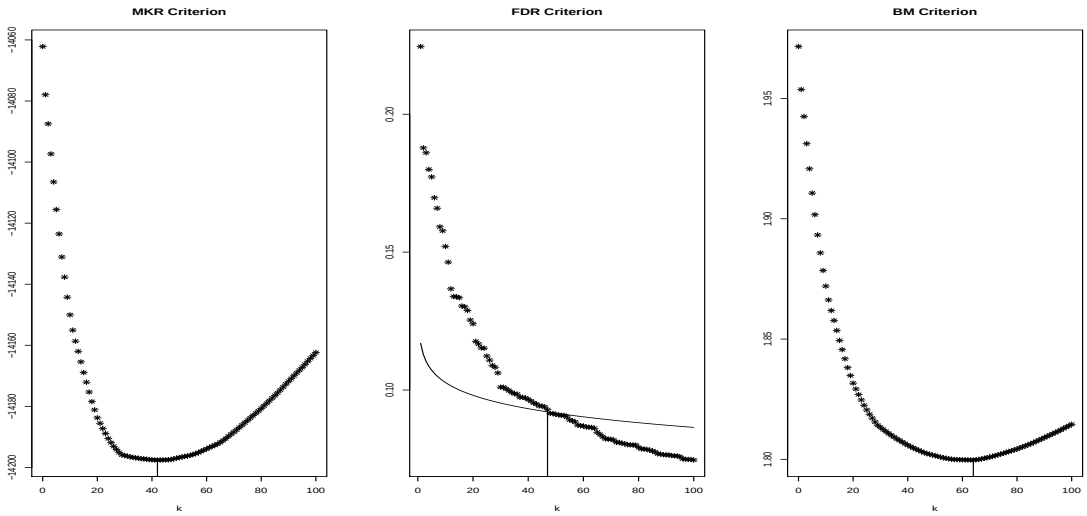

FIGURE 2. The first graph presents the variations of $\text{crit}(k, \text{pen}_{MKR})$ as a function of k. The minimum of $\text{crit}(k, \text{pen}_{MKR})$ is attained for $k = 42$. The second one presents the absolute values of X in the decreasing order and the increasing threshold values. The maximum of k for which the absolute values of X are greater than the threshold is $k = 47$. The third graph presents the variations of $\underline{\text{crit}}(k, \text{pen}_{BM})$ as a function of k. The minimum of $\underline{\text{crit}}(k, \text{pen}_{BM})$ is attained for $k = 64$.

is active assuming that the prior distribution is a mixture of Gaussian variables, Lenth [18] who proposed the SME method, Haaland and O'Connell [15] who compared the performances of several methods that generalise the procedure proposed by Lenth. They noticed that all the methods are essentially the same when they are few active contrasts. Let us see what happens for 16-run experiments with all factors at two levels, when the number of positive contrasts k_0 equals 5 among $n = 15$ contrasts. The procedures are applied with $k_n = 14$. The results are given at Table 10. The AMDL method fails to work because it does not penalise enough large values of k: the criteria decreases from 0 to k_n and in all simulations $\widehat{k}$ equal k_n. The FS method leads to overestimate k_0: $\widehat{k} = k_n$ in 7.4% of the simulations. The MKR, FDR and BM methods are nearly equivalent. We remark that in more than 10% of the simulations these methods are unable to detect the non zero means. If we compare the median of the risk ratios, the MKR method performs better than the others.

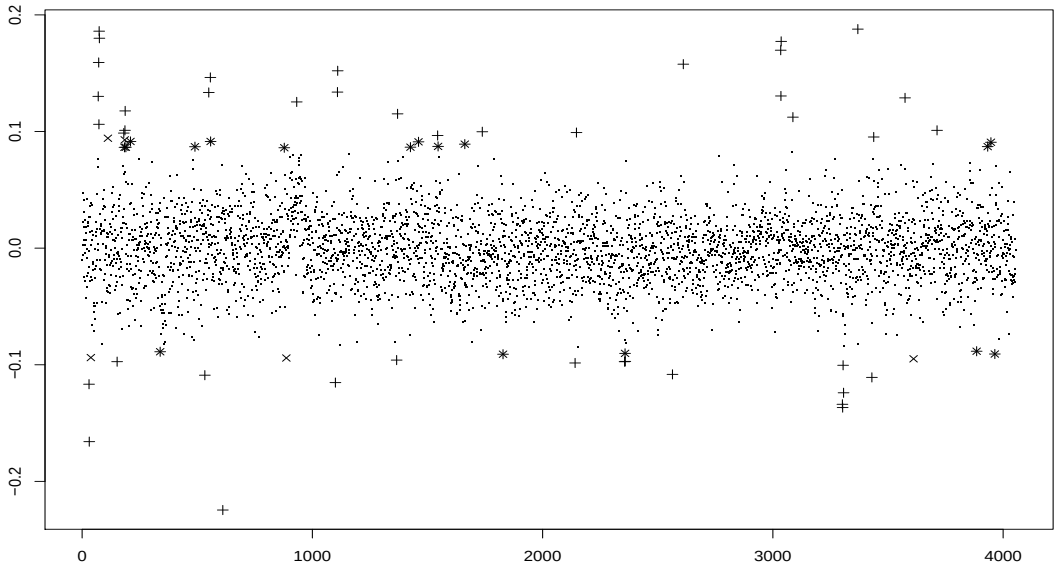

FIGURE 3. Observed values of $X_i, i = 1, \ldots, n$. The differentially expressed genes using the MKR method are identified by `+`; the additional ones using the FDR method are identified by `x`; the additional ones using the BM method are identified by `*`.

REFERENCES

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akademia Kiado.

[2] H. Akaike. A bayesian analysis of the minimum aic procedure. *Ann. Inst. Statist. Math.*, 30:9–14, 1978.

[3] A. Antoniadis, I. Gijbels, and G. Grgoire. Model selection using wavelet decomposition and applications. *Biometrika*, 84:751–763, 1997.

[4] P. Baldi and A. Long. A bayesian framework for the analysis of microarray expression data; regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519, 2001.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300, 1995.

[6] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3:203–268, 2001.

[7] J.J. Daudin, S. Ghachem, M. Descender, S. Robin, A. Hnaut, A. Sekowska, and A. Danchin. Comparison of statistical methods for differential analysis of macroarray data. Technical report, INAPG, 2001.

	SIC	AMDL	MKR	SME	FDR	FS	BM
$\widehat{k}=0$	0	0	11.6	50.2	12.1	1.9	10.9
$\widehat{k}\in]0,4[$	0	0	1.1	22.2	9.4	2	4.1
$\widehat{k}\in[4,6]$	0	0	82.9	27.4	74.2	54.4	76.1
$\widehat{k}>6$	100	100	4.4	0.2	4.3	41.7	8.9
r^{mean}	1.610^7	1.610^7	11.4	2.2	3.2	1.810^3	4.9
r^{median}	4.610^3	4.610^3	1.2	2.2	1.6	3.2	1.6
$\underline{r}^{\text{mean}}$	3	3	4.2	17	5.7	2.5	4.7
$\underline{r}^{\text{median}}$	2.8	2.8	1.1	25	1.5	1.7	1.4
ρ	64	64	2.9	0.2	3.2	20.6	5.4

TABLE 10. Case $n=15$, $k_0=5$, $\mu=5$, $\ell=0$. The three first lines give the percentage of simulations for which $\widehat{k}=0$, $\widehat{k}<4$, $\widehat{k}\in[4,6]$ and $\widehat{k}>6$. The four following lines give respectively $r^{\text{mean}}=\mathcal{K}_n^{\text{mean}}(k_0,\mu,\ell)/\mathcal{R}_n^*(k_0,\mu,\ell)$, $r^{\text{median}}=\mathcal{K}_n^{\text{median}}(k_0,\mu,\ell)/\mathcal{R}_n^*(k_0,\mu,\ell)$, $\underline{r}^{\text{mean}}=\underline{\mathcal{K}}_n^{\text{mean}}(k_0,\mu,\ell)/\underline{\mathcal{R}^*}_n(k_0,\mu,\ell)$, and $\underline{r}^{\text{median}}=\underline{\mathcal{K}}_n^{\text{median}}(k_0,\mu,\ell)/\underline{\mathcal{R}^*}_n(k_0,\mu,\ell)$. The last line ρ gives the mean of the false discovery ratios.

[8] P.R. Delmar, S. Robin, D. Tronok-Le Roux, and J.J. Daudin. Mixture model on the variance for the differential analysis of gene expression data. Technical report, INAPG, 2003.

[9] S. Dudoit, Y.H. Yang, Callow M.J, and T.P. Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Journal of the American Statistical Association*, 74:829–836, 2002.

[10] B. Efron, J. Tibshirani, J. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Asocition*, 96:1151–1160, 2001.

[11] Box E.P and R.D. Meyer. An analysis for unreplicated fractional factorials. *Technometrics*, 28(1):11–18, 1986.

[12] D.P. Foster and R.A. Stine. Adaptive variable selection competes with bayes expert. Technical report, The Wharton School of the University of Pennsylvania, Philadelphia, 2002.

[13] S. Huet. Model selection for estimating the non zero components of a gausian vector. *ESAIM*, 2005.

[14] R. Nishii. Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivariate Anal.*, 27:392–403, 1988.

[15] Haaland P.D and M.A. O'Connell. Inference for effect-saturated fractional factorials. *Technometrics*, 37(1):82–93, 1995.

[16] A. Reiner, D. Yekutieli, and Yoav Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19:368–375, 2003.

[17] J. Rissanen. Universal coding, information, prediction and estimation. *IEEE Trans. Infor. Theory*, 30:629–636, 1984.

[18] Lenth R.V. Quick and easy analysis of unreplicated factorials. *Technometrics*, 31(4):469–473, 1989.

[19] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6:461–464, 1978.

[20] S. Sekowska, S. Robin, J.J. Daudin, A. Hnaut, and A. Danchin. Extracting biological information from dna arrays: an unexpected limk between arginine and methionine metabolism in *bacillus subtilis*. *Genome Biology*, 2(6):1–12, 2001.

[21] W. Wang and D.T. Voss. On adaptive estimation in orthogonal saturated designs. *Statistica Sinica*, 13:727–737, 2003.