Statistics and Its Interface Volume 5 (2012) 263–275

# Random threshold for linear model selection, revisited

Merlin Keller* and Marc Lavielle

In [11], a random thresholding method is introduced to select the significant, or non-null, mean terms among a collection of independent random variables, and applied to the problem of recovering the significant coefficients in non-ordered model selection. We introduce a simple modification which removes the dependency of the proposed estimator on a window parameter while maintaining its asymptotic properties. A simulation study suggests that both procedures compare favorably to standard thresholding approaches, such as multiple testing or model-based clustering, in terms of the binary classification risk. An application of the method to the problem of activation detection on functional magnetic resonance imaging (fMRI) data is discussed.

Keywords and phrases: Random threshold, Non-ordered model selection, FDR, Mixture modeling, Binary risk, Oracle risk, fMRI.

## 1. INTRODUCTION

In [11], the following model is considered:

$$Y_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\mu_i$ are unknown constants, some of which are zeros, and $\varepsilon_i$ are independent, identically distributed (iid) zero-mean random variables, with known cumulative distribution function (cdf) $F_\varepsilon$. We will also note in the following $F_{|\varepsilon|}$ the cdf of their absolute values $|\varepsilon_i|$, given by: $F_{|\varepsilon|}(x) = P(|\varepsilon_1| \leq x) = (F_\varepsilon(x) - F_\varepsilon(-x))\, \mathbf{1}_{\{x\geq 0\}}$.

With this model, the problem of selecting the significant coefficients $\mu_i \neq 0$, based on the observations $Y_i$, is studied. Such a problem arises in many different application areas, such as genomics [8], or neuroimaging [7], to cite just a few.

Many methods have been proposed to perform this task. Multiple testing procedures for instance (see [8] for an overview of existing methods), have been developed to control a certain type I error rate, such as the familywise error rate (FWER), or the false discovery rate (FDR) [2] at a user-fixed level. It can be argued however that the choice of a level, which ultimately defines the subset of selected coefficients, is arbitrary, as there is no safe guideline to what an ‘optimal’ level of false detections should be.

An alternative, that allows to control both type I and type II error rates, consists in fitting a mixture model to the data, with one class for the null (zero-mean) data, and one, or more, for the non-null data. A detection threshold can then be derived, which minimizes a certain classification risk, such as the binary risk, associated to the 0 –1 loss function, resulting in a ‘naive Bayes’ classifier [12]. The main difficulty with this approach lies in the choice of a distribution for the non-null data, which may influence significantly the resulting classifier. Many authors have proposed to deal with this issue through model selection techniques (see [6, 9, 13] for instance), however it remains an open-ended problem.

In view of these difficulties, the random threshold (RT) approach introduced in [11] appears as a promising candidate, since it does not require the specification of a type-I error level, nor of a model for the non-zero mean observations. The principle of RT lies in estimating the number of significant coefficients, based on a random centering of the partial sums of the ordered observations. Because it relies on as little assumptions as possible, we expect RT to be more robust than the above-mentioned approaches.

However, to date very little is known concerning the classification performances of RT procedures; [11] essentially gives a minimal separation speed between null and non-null data for the method to attain perfect classification asymptotically. Furthermore, the algorithm described therein still depends on a window parameter, which may have some influence in presence of noisy data.

This article describes a simple modification of the RT procedure, which removes its dependency on the window parameter, while maintaining its asymptotic properties. We then study the classification performances of both techniques using numerical experiments, in comparison to the above-mentioned standard approaches.

An issue that is not addressed here is that of statistical dependency. Indeed, the methods presented in this paper are all based upon the assumption that the observations are independent. This may not always be the case in real-life situations; hence, assessing the robustness of the random threshold method to violation of this independence assumption, either theoretically or through numerical experiments, is important for certain applications.

However, before conducting such an assessment, it would be necessary to specify the form of dependency that is considered of interest. Indeed, there are infinitely many differ-

arXiv: 1010.5389

*Corresponding author.

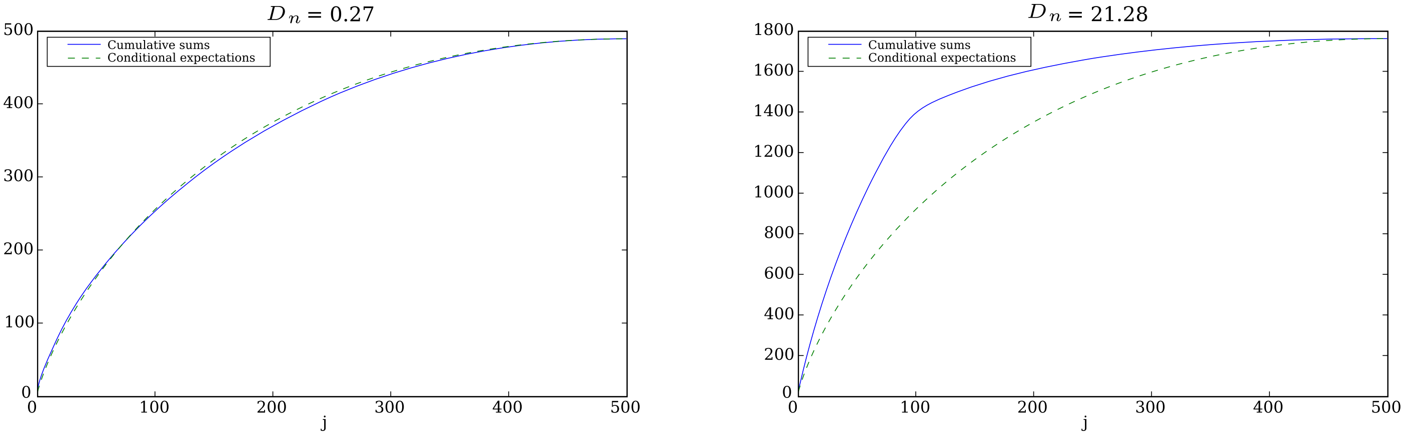

*Figure 1. Test of the presence of significant coefficients: the cumulative sums $(T_j)$ (solid line) and their conditional null expectations $(Q_j)$ (dashed lines), under the global null $\mathcal{H}_0$ (left) and in presence of significant coefficients (right).*

ent ways of introducing dependency between random variables, and each way can lead to potentially different results in terms of robustness. To cite just a few examples, in [3], the BH procedure is shown to be conservative for a particular form of dependency called *positive regression dependency*, and a modified version of the BH procedure is introduced that is conservative under very general conditions. However, this modification may be overly conservative in certain situations. Hence more specialized algorithms have been devised. For instance, in [10], a procedure is constructed specifically for data that can be divided into independent groups of correlated variables. Many other forms of dependency have also been considered, such as *subset pivotality* [8], or through hierarchical modeling, (assuming that the observations are independent conditional on a set of hidden variables), to cite just two. Hence, it seems to us that choosing a particular form of dependency and studying the ensuing properties of the random threshold approach is an open problem, which goes well beyond the scope of the present work.

The rest of this article is organized as follows. In Section 2, the original RT method is reviewed. The variable window extension is introduced in Section 3. In Section 4 the results of numerical experiments are presented, which show the good properties of RT in terms of classification. An application to fMRI data analysis is discussed in Section 5, and we conclude in Section 6 by considering the perspectives opened by this work.

# 2. ORIGINAL RANDOM THRESHOLDING PROCEDURE

## 2.1 Testing the presence of significant coefficients

We start by recalling how the presence of non-zero means is tested, that is, how the null hypothesis $\mathcal{H}_0 : \forall i \in \{1, \ldots, n\}, \mu_i = 0$ is tested in [11]. This is done by comparing the cumulative sums of the ordered observations to their conditional expectations under $\mathcal{H}_0$, according to the following steps:

1. Order the observations $|Y_{(1)}| \geq |Y_{(2)}| \geq \cdots \geq |Y_{(n)}|$.
2. For $i = 1, \ldots, n$, let $X_{(i)} = -\log(1 - F_{|\epsilon|}(|Y_{(i)}|))$.
3. Let $T_j = \sum_{i=1}^{j} X_{(i)}$ and $Q_j = \mathbb{E}_{\mathcal{H}_0}(T_j|T_n)$.
4. Define the test statistic $D_n = \max_j |T_j - Q_j|/\sqrt{n}$. The null hypothesis is rejected if $D_n > d_\alpha$, with $d_\alpha$ such that $\mathbb{P}_{\mathcal{H}_0}(D_n > d_\alpha) \leq \alpha$.

Note that the cumulative sums are not computed directly from the ordered observations, but from the transforms $X_{(1)}, \ldots, X_{(n)}$ which, under $\mathcal{H}_0$, are an ordered series of $\mathcal{E}(1)$ random variables, where $\mathcal{E}(\lambda)$ stands for the exponential distribution with rate parameter $\lambda$. The conditional expectations $Q_j = \mathbb{E}_{\mathcal{H}_0}(T_j|T_n)$ can then be computed using the following result (see [11] for details).

**Theorem 1** (Conditional expectations of ordered exponential variables)**.** *Under $\mathcal{H}_0$, the $X_{(i)}$ are an ordered series of $\mathcal{E}(1)$ random variables. It follows that:*

i) $\mathbb{E}_{\mathcal{H}_0}(X_{(i)}) = \sum_{\ell=i}^{n} \frac{1}{\ell}$.
ii) $\mathbb{E}_{\mathcal{H}_0}(T_j) = j + j \sum_{\ell=j}^{n} \frac{1}{\ell}$.
iii) $\mathbb{E}_{\mathcal{H}_0}(T_j|T_n) = \frac{\mathbb{E}_{\mathcal{H}_0}(T_j)}{\mathbb{E}_{\mathcal{H}_0}(T_n)} T_n$.

Furthermore, for $n \geq 100$, it is shown in [11] using Monte-Carlo simulations that:

$$\mathbb{P}_{\mathcal{H}_0}(D_n > 0.65) \approx 0.05, \tag{2}$$

which provides an approximate calibration for the above test. It is illustrated in Figure 1, on a dataset of $n = 500$ observations $Y_i$ simulated according to model (1), with noise terms $\varepsilon_i$ sampled from the $\mathcal{N}(0, 1)$ distribution.

Under the global null $\mathcal{H}_0$, that is, when all $\mu_i$ are equal to zero, (Figure 1, left), the gap between the curve representing the cumulative sums $(T_j)$ and that of their conditional expectations $(Q_j)$ remains small. Consequently, the resultant test statistic value $D_n = 0.27$, does not exceed the critical value 0.65 given by (2).

In contrast, when we add $n_1 = 100$ non-zero means $\mu_1 = \cdots = \mu_{100} = 5$, (see Figure 1, right), the $T_j$'s become substantially larger than their expected values $Q_j$ under $\mathcal{H}_0$, resulting in a gap between the corresponding curves. Note that this gap is most significant around $j = 100$, because, for all $j$, $T_j$ is the sum of the $j$ largest observations (after transformation), containing mostly non-zero means for $j \leq 100$. Consequently, the ensuing test statistic $D_n = 21.28$ is far above the critical value 0.65.

## 2.2 Selecting the significant coefficients

Upon rejection of the null hypothesis, the following task consists in selecting the significant coefficients. The procedure for doing so can be interpreted in a data-dependent 'multiple hypothesis testing' setting, as described hereafter. Consider the null hypothesis $\mathcal{H}_0$, as defined in Section 2.1, and the set of alternative hypotheses:

$$\begin{aligned} \mathcal{H}_{1,k} : \forall i \in \{1, \dots, k\},\ \mu_{(i)} > 0, \text{ and} \\ \mu_{(k+1)} = \cdots = \mu_{(n)} = 0. \end{aligned}$$

In other terms, $\mathcal{H}_{1,k}$ corresponds to the hypothesis that the $k$ largest observations only have non-zero means. Even though in real-life datasets null and non-null data are never perfectly separated, in general one cannot expect more than to discriminate between such hypotheses in non-ordered model selection. Note that this is equivalent to choosing a certain detection threshold to separate null from non-null data.

The RT procedure first computes the $X_{(i)}$'s using the same steps 1 and 2 as in Section 2.1, then adds the following steps:

3. Let $K_n$ be some positive integer. For $1 \leq k \leq n - K_n$ and $1 \leq j \leq K_n$, compute:

$$\begin{aligned} T_{k,j} &= \sum_{i=k+1}^{k+j} X_{(i)} \\ Q_{k,j} &= \mathbb{E}_{\mathcal{H}_{1,k}}(T_{k,j} | T_{k,K_n}) \\ \eta_k &= \max_{1 \leq j \leq K_n} |T_{k,j} - Q_{k,j}| / \sqrt{n}. \end{aligned}$$

4. Let $\widehat{k}_n = \operatorname{argmin}_{1 \leq k \leq n - K_n} \eta_k$.

As in the previous section, for each $k$, $X_{(k+1)}, \dots, X_{(k+n-k)}$ is an ordered series of $\mathcal{E}(1)$ variables under $\mathcal{H}_{1,k}$, so $Q_{k,j}$ can easily be computed using Theorem 1. Heuristically, the partial cumulative sums $T_{k,j}$ are compared to their conditional expected values $Q_{k,j}$ under the hypotheses $\mathcal{H}_{1,k}$, for $k = 1, \dots, n - K_n$. The number of significant coefficients is estimated as the index $\widehat{k}_n$ corresponding to the minimal gap between $T_{k,j}$ and $Q_{k,j}$, as evaluated by $\eta_k$.

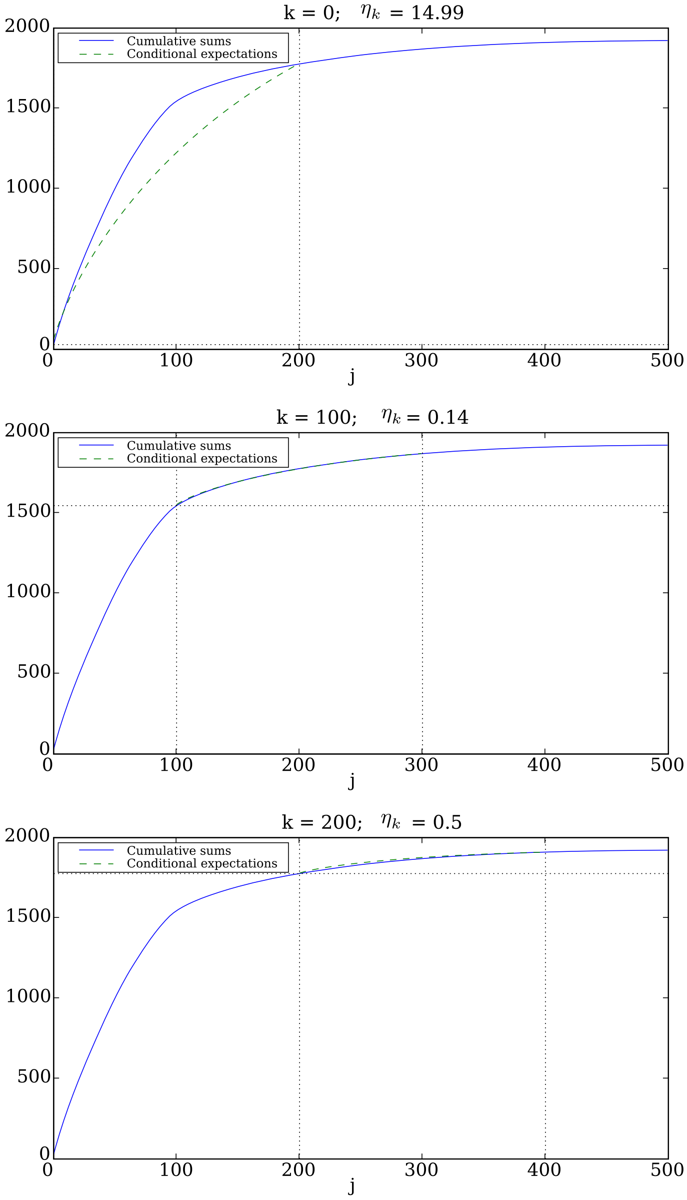

*Figure 2. Random threshold procedure: partial cumulative sums $(T_{k,j})_j$ (solid lines) and their conditional null expectations $(Q_{k,j})_j$ (dashed lines) for $k = 0$ (top), $k = 100$ (middle), and $k = 200$ (bottom), with window width $K_n = 200$.*

This is illustrated in Figure 2, using a dataset simulated exactly as in Section 2.1, that is, with $n_1 = 100$ significant coefficients. Loosely speaking, the procedure uses a sliding window with width $K_n$, and compares the cumulative sums $T_j$ within this window to their conditional expectations under the hypothesis that the window contains the $K_n$ largest null terms. For $k = 1$, the window contains in fact mainly significant terms, so that the $T_{1,j}$'s are well above their ex-

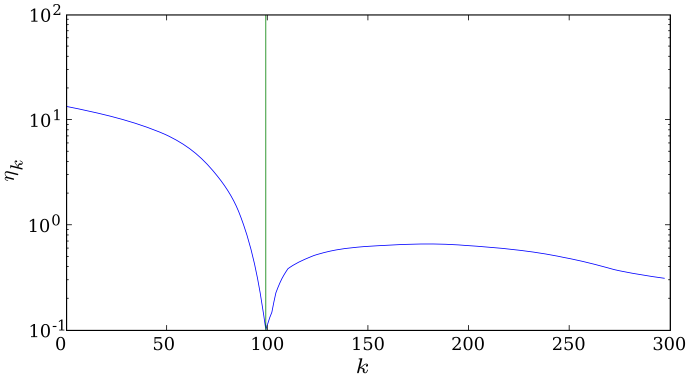

*Figure 3. Random threshold procedure: the sequence $\eta_k$ on a logarithmic scale; its minimum is attained for $\widehat{k}_n = 99$.*

pected values, yielding a normalized gap of $\eta_1 = 14.99$. For $k = 100$, the window indeed contains mostly the $K_n$ largest null terms, so $T_{100,j}$ and $Q_{100,j}$ are of the same order, yielding a much smaller gap ($\eta_{100} = 0.14$). Finally, for $k = 200$, the window contains null terms, but not the $K_n$ largest, so the cumulative sums $(T_{k,j})$ become lower than their expected values. Consequently, the gap increases ($\eta_{200} = 0.5$). Figure 3 shows the complete sequence of $\eta_k$ values, with a clear minimum at $\widehat{k}_n = 99$, close to the true number of significant coefficients.

# 3. EXTENSIONS AND ASYMPTOTIC PROPERTIES

## 3.1 Unknown distribution extension

The RT method recalled in the previous section may be difficult to apply to real-life problems, where the noise distribution $F_\epsilon$ is in general unknown. In [11], an extension is proposed to the case where $F_\epsilon$ is a parametric distribution $F_\epsilon(\cdot\,;\theta)$, with $\theta$ unknown. Quite naturally, this consists in estimating $\theta$ under each hypothesis $\mathcal{H}_{1,k}$ from the null data $Y_{(k+1)}, Y_{(k+2)}, \ldots, Y_{(n)}$, then using this estimate to derive the transforms $X_{(i)}$, for $i = k+1, \ldots, k+K_n$. More precisely, having chosen some positive integer $K_n$, the extension consists in performing for $1 \le k \le n - K_n$ the following steps:

1. Let $\widehat{\theta}_k = \widehat{\theta}(Y_{k+1}, \ldots, Y_n)$ be an estimate of $\theta$.
2. For $i = 1, \ldots, n$, let $X_{(i)}(\widehat{\theta}_k) = -\log(1 - F_{|\epsilon|}(|Y_{(i)}|\,;\widehat{\theta}_k))$.
3. For $1 \le j \le K_n$, compute:
   - $T_{k,j}(\widehat{\theta}_k) = \sum_{i=k+1}^{k+j} X_{(i)}(\widehat{\theta}_k)$,
   - $Q_{k,j}(\widehat{\theta}_k) = \mathbb{E}_{\mathcal{H}_{1,k}}(T_{k,j}(\widehat{\theta}_k)|T_{k,K_n}(\widehat{\theta}_k))$.
4. Compute $\eta_k(\widehat{\theta}_k) = \max_{1\le j\le K_n} |T_{k,j}(\widehat{\theta}_k) - Q_{k,j}(\widehat{\theta}_k)|/\sqrt{n}$.

Finally, the estimated number of components is given as before by

$$\widehat{k}_n = \mathrm{argmin}_{1\le k\le n-K_n} \eta_k(\widehat{\theta}_k).$$

This simple extension is much more computationally intensive than the original procedure, since the $X_{(i)}$'s for $i = k+1, \ldots, k+K_n$ must be re-computed for each $k$, instead of once and for all.

To illustrate this, we applied the unknown distribution extension to the same simulated dataset used to illustrate the original procedure. We defined the null distribution as the Gaussian $\mathcal{N}(0;\sigma^2)$, and the unknown variance $\sigma^2$ was estimated by the usual mean squares: $\widehat{\sigma}^2_k = \frac{1}{n-k}\sum_{i=k+1}^{n} Y^2_{(i)}$. The estimated number of significant coefficient, $\widehat{k}_n = 105$, was still found close to its true value, and so was the corresponding standard error estimate $\widehat{\sigma}^2_{\widehat{k}_n} = 0.93$.

## 3.2 Varying window extension

As we have seen previously, the RT procedure depends on a parameter $K_n$ which can be interpreted as a window width, since $\eta_k$ is a function of $X_{(k+1)}, \ldots, X_{(k+K_n)}$. $K_n$ must be smaller than the number of null coefficients, but at the same time not too small, or $\eta_k$ would become unstable. Hence choosing an appropriate value for $K_n$ may be a hindrance in practice, especially since we want the RT method to be adaptive and depend as little as possible on any form of tuning.

This issue can be avoided by re-defining $\eta_k$ as a function of $X_{(k+1)}, \ldots, X_{(n)}$, thus replacing the fixed width $K_n$ by a varying width $n-k$, which requires no prior tuning. We define the following procedure, starting with the same steps 1 and 2 as in Section 2.1, and adding the following steps:

3. Let $\kappa_n$ be a lower bound on the number of null coefficients. For $1 \le k \le n - \kappa_n$ and $1 \le j \le n-k$, compute:

$$\begin{aligned} T_{k,j} &= \textstyle\sum_{i=k+1}^{k+j} X_{(i)} \\ Q_{k,j} &= \mathbb{E}_{\mathcal{H}_{1,k}}(T_{k,j}|T_{k,n-k}) \\ \eta_k &= \max_{1\le j\le n-k} |T_{k,j} - Q_{k,j}|/\sqrt{n-k}. \end{aligned} \tag{3}$$

4. Let $\widehat{k}_n = \mathrm{argmin}_{1\le k\le n-\kappa_n} \eta_k$.

In other terms, $\eta_k$ would be strictly equal to the test statistic $D_n$ defined in Section 2.1, if the sequence $(X_{(i)})_{1\le i\le n}$ were replaced by the subsequence $(X_{(i)})_{k+1\le i\le n}$, i.e., the set of null terms under $\mathcal{H}_{1,k}$.

Notice that $\widehat{k}_n$ is independent from $\kappa_n$, as long as $\eta$ reaches its global minimum on $\{1, \ldots, n-\kappa_n\}$ (see below for a discussion on the choice of $\kappa_n$). The varying window extension presented here can of course be combined with the unknown distribution extension in Section 3.1.

Figure 4 illustrates the varying window extension on the same simulated dataset as previously. Intuitively, the ‘sliding’ window is replaced by a ‘shrinking’ window, which at first encloses all observations, than progressively reduces

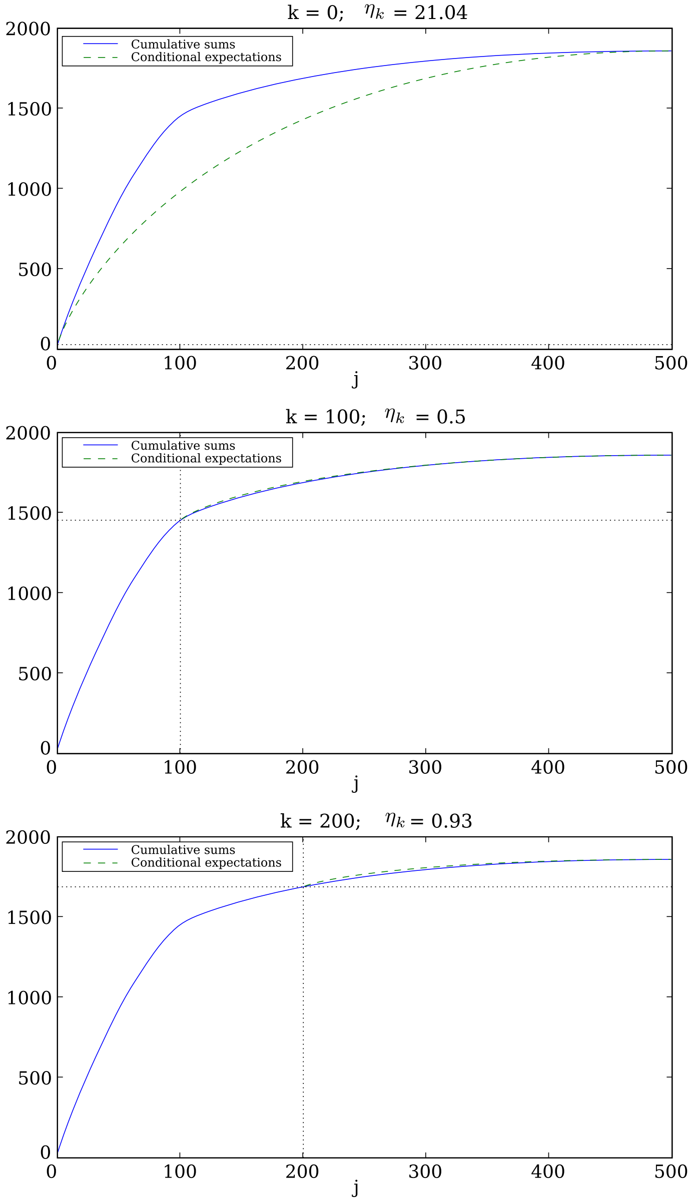

*Figure 4. Random threshold with a varying window width: partial cumulative sums $(T_{k,j})_j$ (solid lines) and their conditional null expectations $(Q_{k,j})_j$ (dashed lines) for $k = 0$ (top), $k = 100$ (middle), and $k = 200$ (bottom).*

in width as the largest observations are left out. Otherwise the same observations as in the fixed-width case hold: when $k = 1$, the cumulative sums $(T_{k,j})_j$ are larger than their conditional expectations $(Q_{k,j})_j$, resulting in a large gap ($\eta_0 = 21.04$). This gap is considerably reduced when $k = 100$ as the $T_{k,j}$'s and the $Q_{k,j}$'s become of the same order ($\eta_{100} = 0.5$), then the gap increases again for $k = 200$ as the cumulative sums become smaller than their expected values ($\eta_{200} = 0.93$). Though not shown here, the sequence $\eta_k$ attains its minimum in $\widehat{k}_n = 99$, satisfyingly close to the true value.

*Choosing $K_n$ and $\kappa_n$* Both $K_n$ and $\kappa_n$ must satisfy the constraint of being lower bounds on the number of null data. Since in most datasets null data outnumber non-null data, tuning both parameters to half the number of observations seems reasonable.

However, in the original random threshold method, $K_n$ is a key parameter as it determines the number of points used to calculate the criterion $\eta_k$ for all $k = 1, \ldots, (n - K_n)$, which in turn serves to select the number of non-null observations. Hence, changing the value of $K_n$ mechanically changes the set of observations detected as non-null.

In contrast, in the varying window extension, the criterion $\eta_k$ is calculated using $(n - k)$ observations for $k = 1, \ldots, (n - \kappa_n)$, and $\kappa_n$ simply specifies the minimal number of points required to compute the criterion, in order to ensure that it is not too noisy. Hence it is essentially included as a safeguard to prevent erratic values at very high threshold values; the result of the selection procedure does not depend on the value of this parameter, so long as it remains in a reasonable range of values. In this respect the varying window extension is truly adaptive, since it requires no parameter tuning.

## 3.3 Asymptotic properties

The estimator of the number of significant coefficients presented in Section 2.2 is consistent. This is the main result in [11], and it can be extended to the varying window setting. We start by recalling the following asymptotic framework:

**AF1** There exists $t^\star \in (0, 1)$ and a subset $I_{k_n^\star}$ of $\{1, \ldots, n\}$, with $k_n^\star = [t^\star n]$ and $|I_{k_n^\star}| = k_n^\star$, such that $\mu_i \neq 0$ if $i \in I_{k_n^\star}$. For all other index, $\mu_i = 0$.

**AF2** For any $i \in I_{k_n^\star}$, $|\mu_i| \geq \alpha_n$, where $\alpha_n \to \infty$ according to the distribution of the $(\epsilon_i)$. More specifically, let $\boldsymbol{\Phi}_{(1)}$ be the distribution of $\max_{1 \leq i \leq n} |\epsilon_i|$ and $(a_n, b_n)$ such that $\Phi_{(1)}(a_n + b_n x) \to \boldsymbol{W}(x)$ for some fixed distribution $\boldsymbol{W}$. Then $(\alpha_n)$ satisfies

$$\frac{\alpha_n - 2a_n}{b_n} \to \infty.$$

**AF3** $\kappa_n / n \to c$ such that $0 < c < 1 - t^\star$.

We then have the following result:

**Theorem 2** (Consistency of the random threshold)**.** *Let $\widehat{k}_n$ stand for the estimator defined in Section 2.2. Under assumptions* **AF1**, **AF2**, **AF3**, *$\widehat{k}_n$ is consistent in the sense that*

$$P\left(\left|\frac{\widehat{k}_n}{n} - t^\star\right| > u_n\right) \to 0, \tag{4}$$

*for any positive decreasing sequence $(u_n)$ such that $\sqrt{n} u_n \to \infty$.*

This result is refined in [11] by deriving an upper bound, which we do not detail here, on the convergence rate of the probability in Equation (4), for a particular choice of sequence $(u_n)$. Consistency also holds in the unknown distribution case, under a different set of assumptions which we do not recall here, and under the varying window extension, as shown in Appendix A.

This theorem is interesting in that it gives a convergence rate for $\widehat{k}_n$, provided that a minimal signal-to-noise (SNR) ratio is attained, represented by a lower bound $\alpha_n$ on the absolute values $|\mu_i|$ of the non-zero means (assumption **AF2**). Note that, in order for the random threshold (or any other threshold for that matter) to asymptotically separate perfectly null from non-null data, the SNR must necessarily become arbitrarily large as the sample size increases.

However, this theorem provides no clue to what happens when the SNR remains bounded, as we expect to be the case in real-life applications. In the remainder of this paper, our goal is to explore the behavior of the RT approach in such cases.

## 4. SIMULATION STUDY

In order to assess empirically the classification properties of RT, we designed several numerical experiments. Our goal was to compare the binary classification risk of the RT procedure (with both fixed and varying window width), to those obtained by model-based clustering and FDR control techniques. Specifically, we used a mixture-model, estimated via an expectation-maximization (EM) algorithm [5] to approximate the risk minimizing detection threshold, and the Benjamini-Hochberg (BH) procedure [2] to derive a threshold controlling the FDR at a certain level. We dismissed FWER control techniques as they essentially yield constant thresholds at a given level, and are therefore of little interest when compared to adaptive approaches.

We considered two cases, depending on whether the null distribution $F_\epsilon$ was considered as known or not. Note that the BH procedure is based on the p-values $p_i = 1 - F_{|\epsilon|}(|Y_i|)$, hence it requires that $F_\epsilon$ be known, whereas this same distribution can be estimated using the EM algorithm. So in order to compare methods on a fair basis, we compared RT to the BH procedure when the null distribution was known, and to mixture model fit otherwise.

### 4.1 Results with known null distribution

We chose in this first experiment to directly simulate the $X_i$, rather than the $Y_i$. Datasets of $n = 10{,}000$ observations were generated, containing each $n_1 = 1{,}000$ signicant terms. These were sampled from the Gamma distribution $\mathcal{G}(\alpha, \beta)$, where $\beta$ is a scale parameter (the chosen values for these parameters are given below). The remaining 9,000 null terms were sampled from the $\mathcal{E}(1)$ distribution.

Note that the data simulated in this fashion does not have the additive structure specified by (1), but rather follows a mixture of two distributions (an exponential and a gamma). Hence only the null distribution (the exponential) is available to the algorithm, the alternative (the gamma) being misspecified with respect to the above formulation. This is important as it means that the random threshold procedure does not benefit here from any additional information as compared to the BH procedure, and ensures that the comparison is fair in this respect.

Table 1 shows the average classification risks obtained by the different methods over 100 simulated datasets and for different choices of the Gamma distribution parameters. More precisely, we chose to compute the ratio of each attained risk to the lowest achievable (oracle) risk, which makes more sense since perfect classification is in general unattainable.

The binary risk and oracle threshold can be computed as follows. Consider a given dataset $(X_i, Z_i)_i$, where $Z_i$ is a binary variable, equal to 0 if $X_i$ is a null term (sampled from the $\mathcal{E}(1)$ distribution), and 1 if $X_i$ is a non-null term (sampled from the Gamma distribution). Then the overall classification error associated with a given detection threshold $t$ is given by:

$$c(t) = \sum_{Z_i=0} \mathbf{1}_{\{X_i > t\}} + \sum_{Z_i=1} \mathbf{1}_{\{X_i \le t\}},$$

that is, the sum of type I (false detections) and type II (false non-detections) errors. The oracle threshold $t^*$ is then chosen to minimize this classification error:

$$t^* = \arg\min_t c(t).$$

It can be seen that both RT approaches perform in general better than FDR control through the BH procedure,

*Table 1. Ratio of binary classification risks with respect to the lowest attainable (oracle) risk for FDR control at different levels and for the RT procedure with fixed ($K_n = 5{,}000$) and variable ($\kappa_n = 5{,}000$) window width, averaged over 100 simulated datasets. For each method, the highest attained risk ratio is shown in bold*

| | | | $\beta$ | | | $\beta$ | | | $\beta$ | | | $\beta$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1.0 | 2.0 | 3.0 | 1.0 | 2.0 | 3.0 | 1.0 | 2.0 | 3.0 | 1.0 | 2.0 | 3.0 | 1.0 | 2.0 | 3.0 |
| | 5.0 | 1.88 | 1.85 | 1.36 | 1.6 | 1.08 | 1.12 | 1.31 | 1.06 | 1.65 | **1.31** | 1.15 | 1.11 | 1.24 | 1.13 | 1.10 |
| $\alpha$ | 6.0 | 2.4 | 1.66 | 1.14 | 1.59 | 1.04 | 1.61 | 1.19 | 1.33 | 3.00 | 1.30 | 1.14 | 1.14 | **1.25** | 1.12 | 1.14 |
| | 7.0 | **2.7** | 1.37 | 1.09 | 1.42 | 1.21 | **2.91** | 1.07 | 2.01 | **6.02** | 1.27 | 1.13 | 1.16 | 1.23 | 1.12 | 1.17 |
| | | | FDR 0.01 | | | FDR 0.05 | | | FDR 0.1 | | | fix. RT | | | var. RT | |

*Table 2. Ratio of binary classification risks over oracle risk for model-based clustering (left) and for the RT procedure with fixed ($K_n = 500$) and variable ($\kappa_n = 500$) window width (middle and right), averaged over 100 simulated datasets. For each method, the highest attained risk ratio is shown in bold*

| | | $\sigma$ | | | $\sigma$ | | | $\sigma$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1.0 | 2.0 | 3.0 | 1.0 | 2.0 | 3.0 | 1.0 | 2.0 | 3.0 |
| | 1.0 | 1.03 | 1.03 | 1.08 | 1.03 | 1.06 | 1.02 | 1.03 | 1.06 | 1.03 |
| $\mu$ | 2.0 | 1.06 | 1.03 | 1.04 | 1.32 | 1.13 | 1.05 | 1.30 | 1.12 | 1.05 |
| | 3.0 | **1.11** | 1.06 | 1.04 | **1.60** | 1.19 | 1.08 | **1.55** | 1.18 | 1.08 |
| | | GMM fit | | | fix. RT | | | var. RT | | |

with a slight advantage to the varying window extension. Most importantly, the classification risks they attain is never more than 1.31 and 1.25 times the oracle risk for the fixed and varying window versions, respectively. In contrast to these near optimal performances, whatever the chosen level of FDR control, the BH procedure always performs poorly for at least one model, with an average classification risk that rises as high as 6.02 times the optimal one in the worst case.

Of course, it can be argued that comparing the RT and BH methods in terms of the binary classification risk is unfair, since the BH aims to control the FDR, which is essentially a type I error rate. Our goal here is to compare RT with a standard thresholding method. Such methods are usually multiple-testing based, and so aim at maximizing the test power while maintaining a type I error rate under a prespecified level. What we show here is that an equally simple procedure can balance type I and type II error rates without requiring any pre-specified level, but rather by adapting to the SNR in the dataset at hand.

The above results suggest that the RT approach, due to its adaptive nature, is more stable than error rate control techniques, that depend on the choice of a false detection level, as we had anticipated. Moreover, the excellent performance of the RT methods, which attained near optimal risks on all the simulated datasets, is very encouraging for this approach.

## 4.2 Results with unknown null distribution

To illustrate the unknown distribution case, we simulated $n = 1{,}000$ observations $Y_i$, among which $n_1 = 100$ where sampled from the $\mathcal{N}(\mu, \sigma^2)$ distribution with $\mu > 0$ and represented the significant terms, and $n - n_1 = 900$ where sampled from the $\mathcal{N}(0, 1)$ distribution and represented the null terms. We used less observations than in the known distribution case because the unknown null distribution extension is much more computer-intensive.

We implemented an EM algorithm to estimate a two-class Gaussian mixture model (GMM) from the data, with one zero-mean class to model the null data. As is often the case with iterative algorithms, providing initial values for the model parameters was the main problem we encountered. We found an efficient strategy for doing so, taking advantage of the fact that the negative data contained mostly null terms, and could provide a good initial guess for the null distribution variance and mixture weight. Details of the algorithm are given in Appendix B. Table 2, shows the average ratios of the classification risks obtained by the different methods with respect to the oracle risk, over 100 simulated datasets and for different choices of the Gaussian distribution parameters for the significant terms.

*Table 3. Results for bimodal non-null data (averaged over 100 simulated datasets)*

| **4.01** | **2.03** | **1.89** |
|---|---|---|
| GMM fit | fix. RT | var. RT |

All methods performed satisfyingly, yielding close to optimal risks, with model-based clustering performing slightly better than the RT methods. This comes as little surprise, since in this case the former approach has several advantages: it is based on a parametric model for the significant terms that is precisely the one used to simulate the data, and it explicitly minimizes the binary classification risk. In contrast, the RT approach is not based on the probability model used to generate the data, and does not explicitly minimize a classification risk. Nevertheless, it gave good results.

Furthermore, performances of model-based methods can deteriorate when based on the wrong assumptions. To illustrate this, we simulated $n = 5{,}000$ observations $Y_i$ among which $n_1 = 950$ where sampled from the $\mathcal{N}(3, 1)$ distribution, $n_2 = 50$ from the $\mathcal{N}(20, 1)$ distribution, and the remaining 4,000 from the $\mathcal{N}(0, 1)$ distribution and represented the null data. Consequently, the significant terms had a bimodal distribution. Most of these terms were next to the null mode, and a small number where next to a more distant mode.

This way, we hoped to trick the mixture model, which assumed a unimodal distribution for the significant terms, into detecting only the distant mode, while merging the other mode with the null distribution. This is exactly what happened, as can be seen in Table 3: the mixture-model fit performs significantly worse than the RT approach in this case, the latter maintaining a reasonable, though also degraded, classification risk.

Of course it can be argued that such a dataset does not represent a realistic situation; our point here is simply to illustrate the increased robustness of RT due to the fact that it requires no assumptions other than a noise model. It can also be discussed that an alternative to the simplistic two-class GMM used here would be to allow a variable number of classes, combined with a model selection framework [6, 9, 13]. However, implementing such complex strategies would be non-trivial, especially concerning the algorithm's initialization. This last issue could be addressed for instance by using stochastic extensions of the EM, such as the stochastic averaging EM (SAEM) [4], in order to reduce dependency to initial values. In contrast to such sophisticated strategies, the simplicity of the RT approach, which requires minimal implementation and virtually no tuning, appears as a key advantage in practice, especially in view of the good performances suggested by this study.

## 5. APPLICATION TO FMRI DATA ANALYSIS

We now apply the random threshold approach to functional magnetic resonance imaging (fMRI) data analysis. fMRI is a modality of *in vivo* brain imaging that allows to measure the variations of cerebral blood oxygen levels induced by the neural activity of a subject lying inside a MRI scanner and submitted to a series of stimuli. A sequence of three-dimensional (3D) images of the brain is thus acquired, measuring over time a vascular effect of neural activity known as the blood oxygenation level dependent (BOLD) effect. From the time series recorded in each voxel, and the occurrence times for each stimulus, one may compute an estimate of the BOLD effect of the subject in response to any given stimulus, and more generally to any difference or combination of stimuli (contrast) [7, 18].

Thus, the fMRI data for one subject generally consists in a spatial map of $z$-scores $(Y_1, \ldots, Y_p)$, where $p$ is the number of voxel in the search volume (which can be as high as 100,000), and $Y_i$ the estimated BOLD effect at voxel $i$. This map of measures of cerebral activity, also termed *activation map*, is plagued by several sources of uncertainty: the natural variability of brain activity, and the estimation noise induced by the MRI scanner. Thus, model (1) provides a potentially good representation of the activation map $(Y_1, \ldots, Y_p)$, with significant terms corresponding to voxels in brain regions involved in the task under study. More precisely, in the classical fMRI data analysis terminology, voxel $i$ is said to be *inactivated* if $\mu_i = 0$, *activated* if $\mu_i > 0$ and *deactivated* if $\mu_i < 0$, using the notations of model (1).

Of course, it is only reasonable to model the activation map $(Y_1, \ldots, Y_p)$ according to (1) if the $Y_i$ can be considered as statistically independent, meaning that the BOLD effect measurement errors in neighboring voxels are independent. It turns out that this independence assumption is standard in the fMRI literature (see [7, 18] for instance). This is justified by the fact that the data goes through a series of pre-processing steps that includes a whitening step that renders the voxels spatially independent. Typically, a certain amount of spatial dependency is then re-introduced in a controlled way through smoothing (whose primary goal is to increase signal to noise ratio and inter-subject matching of homologous brain areas), but standard detection procedures remain voxelwise and ignore this dependency (see [14] for a review of such procedures). The data considered here is smoothed using a 5 mm full-width at half maximum (FWHM) Gaussian filter, hence the amount of spatial dependency is very small, given that the voxel width is 3 mm.

In a typical fMRI study however, not one but several subjects are recruited from a population of interest, and scanned while submitted to the same series of stimuli. Activation maps associated with a given contrast are obtained for each subject, as described above, and used as input data for inference at the between-subject level, where the goal is to evidence a general brain activity pattern. Mass univariate, or voxel-based, detection [7] is to date the most widely used approach to address this question. It starts with normalizing individual images onto a common brain template using nonrigid image registration. Next, a $t$-statistic is computed for each voxel to locally assess mean group effects.

In both single-subject and multi-subject fMRI data analysis, the problem of activation detection can be formulated statistically as that of detecting non-zero means among a collection of observations. The most common approach consists in thresholding a statistical map of brain activity [7]. Multiple-testing techniques are widely used [14, 15], as well as mixture models. The Gamma-Gaussian mixture model (GGM) is most often used in this context [1]. It uses a Gaussian distribution for null, or inactivated, data, a Gamma distribution for activated data, and a negative Gamma distribution for deactivated data.

These methods suffer from certain limitations, as discussed in the previous sections, hence the RT appears as an appealing alternative in this context. Thus, we decided to compare the regions detected by the different approaches, to see if RT succeeded in recovering regions known to be involved in certain well-studied cognitive tasks.

### 5.1 Data acquisition and preprocessing

We used a real fMRI dataset from the Localizer database [16], involving a cohort of 38 right-handed subjects, and acquired as follows. The participants were presented with a series of stimuli or were engaged in tasks such as passive viewing of horizontal or vertical checkerboards, left or right click after audio or video instructions, computation (subtraction) after video or audio instructions, sentence listening and reading. Events occurred randomly in time (mean inter stimulus interval: 3s), with ten occurrences per event type, and ten event types in total.

Functional images were acquired on a General Electric Signa 1.5T scanner using an Echo Planar Imaging sequence. Each volume consisted of 34 64 × 64 3 mm-thick axial contiguous slices. A session comprised 130 scans. Anatomical T1

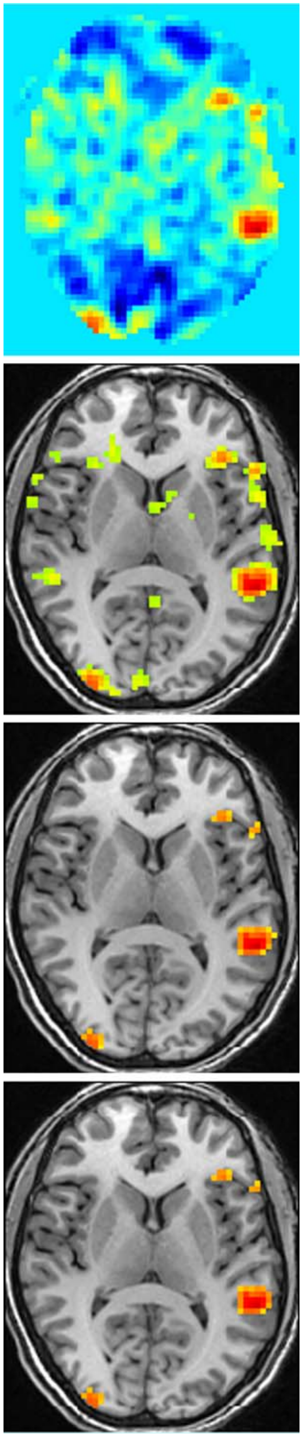

*Figure 5. Axial slice from a $z$-score map for the "sentence minus checkerboard" contrast, using a temperature palette (not shown here) for the $z$-score values. From top to bottom: Unthresholded, thresholded by GGM model fit, varying-window and fixed-window ($K_n = 15{,}000$) random thresholding. Detected activations are shown against the subject's anatomical image.*

weighted images were acquired on the same scanner, with a spatial resolution of $1 \times 1 \times 1.2$ mm$^3$. Finally, the cognitive performance of the subjects was checked using a battery of syntactic and computation tasks.

Single-subject analyzes were conducted using SPM5 (`http://www.fil.ion.ucl.ac.uk`). Data were submitted successively to motion correction, slice timing, normalization to the MNI template and spatial smoothing using an isotropic 5 mm-FWHM Gaussian filter. For each subject, BOLD contrast images were obtained from a fixed-effect analysis on all sessions. Group analyzes were restricted to the intersection of all subjects' whole-brain masks, comprising 43,367 voxels.

We considered the $t$-score maps computed for different contrasts of experimental conditions. These were first converted to $z$-score maps, to obtain approximatively Gaussian statistics in inactivated voxels. Using these maps as input data, we then compared the detection thresholds obtained by Gamma-Gaussian mixture modeling (GGM), fixed-window random thresholding and the varying-window extension, also using the unknown variance extension in both cases (see Section 3.1). For simplicity, we only present here the results obtained for a fixed window equal to $K_n = 15{,}000$.

## 5.2 Individual subject activation map

Our first illustration concerns the activation map of a single subject, for the "sentence minus checkerboard" contrast. This contrast subtracts the effect of viewing horizontal and vertical checkerboards from that of reading video instructions, thus allowing to detect brain regions specifically implicated in the reading task.

Figure 5, left, shows an axial slice from the $z$-score map before thresholding. Activations are clearly seen in Wernicke's and Broca's areas (right and upper right), which are known to be involved in language processing (see [17], for instance). The detection threshold found by GGM fit for the $z$-score map (2.03) was much lower than those found by the random threshold procedure, both with a varying window (3.19) and a fixed window (3.33).

The random thresholds with fixed and variable windows yield very similar activation maps in this case, which seem to capture the activated regions seen in the raw map. In contrast, the much lower threshold found by mixture modeling detects several smaller clusters, some of which may be false positives.

## 5.3 Group activation map

In this second example, we consider a group activation map, specifically a map of $t$-statistics computed from the individual contrast maps of 15 subjects, thus enabling to infer regions of positive mean effects in the parent population. Our choice of limiting the number of subjects, rather than using the whole cohort, was driven by the fact that many fMRI studies are conducted on groups of less than 20 subjects.

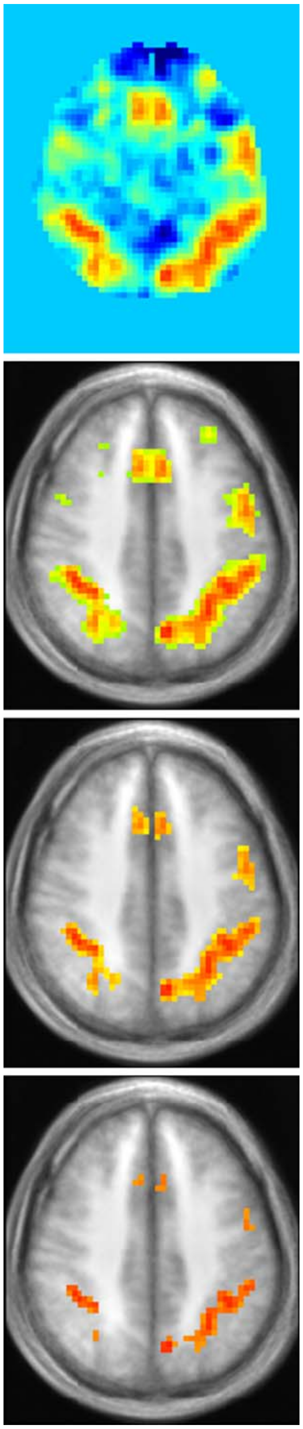

*Figure 6. Axial slice from the group activation $z$-score map for the "calculation minus sentences" contrast, using a temperature palette (not shown here) for the $z$-score values. From top to bottom: Unthresholded, thresholded by GGM model fit, varying-window and fixed-window ($K_n = 15{,}000$) random thresholding. Detected activations are shown against the mean anatomical image of all subjects.*

We report results for the "calculation minus sentences" contrast, which subtracts activations due to reading or hearing instructions from the overall activations detected during the mental calculation tasks. This contrast may thus reveal regions that are specifically involved in the processing of numbers.

Figure 6, left, shows an axial slice from the activation map before thresholding, with clear activations in the bilateral anterior cingulate (upper middle), bilateral parietal (lower left and right) and right precentral (upper right) regions, all known to be involved in number processing [16].

Though sorted in the same order as previously, the varying window random threshold (2.49) is now roughly at equal distances from the threshold found by GGM modeling (1.79) and the fixed window ($K_n = 15{,}000$) random threshold (3.06).

The three methods detected activations in the regions described above, though the fixed window random threshold seemed to miss some activations, and the GGM approach further detected smaller clusters, some of which may be false positives.

Of course one cannot conclude from these examples alone that RT is 'better' at detecting activations than GGM fit. However, the varying window extension successfully detected regions known to be involved in the two cognitive tasks considered here, while avoiding isolated peaks in other regions, which may be part of the background noise. These results suggest that the RT succeeded in capturing only the active regions, while the GGM approach seemed to detect spurious activations.

## 6. DISCUSSION

In this paper, we have introduced a simple modification to the random threshold (RT) procedure proposed in [11], to obtain an entirely unsupervised procedure for recovering non-null mean terms from a collection of independent random observations, based solely on a parametric model of the null terms. Our modification, which requires no prior tuning, conserves the consistency properties of the original procedure.

We have implemented all the different versions of the random threshold method in a Python package. This was integrated into the Neuroimaging in Python (NIPY) open-source library, freely downloadable from `http://nipy.sourceforge.net`.

We validated this approach through extensive numerical experiments, and showed that both the original procedure, based on a fixed window-width, and our extension, which uses a variable window, compare favorably to multiple testing procedures, as well as model-based clustering, in terms of the binary classification risk, with a slight advantage to our varying window extension. On the vast majority of simulated datasets, the risks achieved where close to the lowest achievable (oracle) risk, whereas each of the other approaches behaved poorly in at least one case.

The random threshold procedure does not provide explicit control over a type I error rate as a classical multiple test procedure would do, which may appear as a weakness. Rather, it selects an optimal threshold to separate null from non-null data, hence finds a compromise between type I and type II errors that is different for each dataset. It is important to understand that the aim of this approach is not to achieve a type-I error control to a user-specified level, but rather provide a completely adaptive tool that optimally classifies the data into null and non-null classes, without the need for any prior tuning, thus excluding user-chosen levels. Our simulation results show that indeed our approach performs remarkably well in terms of classification.

Thus RT appears as a very promising method for non-ordered model selection whenever no parametric assumptions are available concerning the data distribution. Such methods are needed in many application domains, as we have illustrated in the case of activation detection for fMRI data analysis.

The good classification performances of RT evidenced empirically in our simulations suggest that a promising direction for future research would be to study its properties in the mixture-model setting, and especially its large-sample behavior. An interesting question to answer would be whether the random threshold converges to a certain limit when the SNR remains constant, and if so, how does this limit compares to the oracle threshold.

Finally, as mentioned in the introduction of this article, assessing the robustness of RT, or adapting it, to cases where the independence assumption is violated, is important for many applications, hence provides another perspective for future works.

## ACKNOWLEDGEMENTS

We would like to thank Dr. Nicolas Bousquet for his careful lecture and fruitful comments. This work was partially funded by the IMAGEN European project.

## APPENDIX A. PROOF OF THE CONSISTENCY THEOREM FOR THE VARYING WINDOW EXTENSION

Following [11], we first recall some notations. Set $U_i = Y_i$ for $i \in I_{k_n^\star}$ and $v_i = Y_i$ for $i \notin I_{k_n^\star}$; notice that $(v_i)$ is a sample from the distribution $F_\epsilon$. Let $(u_{(i)})_{1\le i\le k_n^\star}$ and $(v_{(i)})_{1\le i\le n-k_n^\star}$ be the sequences $(|U_i|)$ and $(|v_i|)$ in decreasing order. Let $\Omega_n$ be the subset of $\Omega$ where $v_{(1)} < \alpha_n/2$ and $u_{(k_n^\star)} > \alpha_n/2$.

A first lemma in [11] shows that $P(\Omega_n) \to 1$, i.e., the collections $(u_{(i)})$ and $(v_{(i)})$ are stochastically in order with high probability. The proof can then be restricted to $\Omega_n$.

Now, let $\mathbb{E}_{\mathcal{H}_{1,k}}(T_{k,j})$ and $Q_{k,j}$ be defined as in Equation (3). Using Proposition 1, we have:

$$\mathbb{E}_{\mathcal{H}_{1,k}}(T_{k,j}) = j\left(1 + \sum_{i=j+1}^{n-k} 1/i\right);$$

$$\begin{aligned} Q_{k,j} &= \frac{\mathbb{E}_{\mathcal{H}_{1,k}}(T_{k,j})}{\mathbb{E}_{\mathcal{H}_{1,k}}(T_{k,n-k})} T_{k,n-k} \\ &= B_{k,j,n} T_{k,n-k}. \end{aligned}$$

Also, let $a_i = \mathbb{E}_0(Z_{(i)}) = \sum_{\ell=i}^{n} 1/\ell$. Equation (4) can be shown separately for $k > k_n^\star$ and $k < k_n^\star$. Since the two cases are treated similarly, we will restrict ourselves here to the case $k > k_n^\star$. On $\Omega_n$:

$$\begin{aligned} T_{k,j} - Q_{k,j} &= T_{k,j} - B_{k,j,n} T_{k,n-k} \\ &= \left(T_{k,j} - \mathbb{E}_{\mathcal{H}_{1,k_n^\star}}(T_{k,j})\right) \\ &\quad - B_{k,j,n}\left(T_{k,n-k} - \mathbb{E}_{\mathcal{H}_{1,k_n^\star}}(T_{k,n-k})\right) \\ &\quad + \mathbb{E}_{\mathcal{H}_{1,k_n^\star}}(T_{k,j}) - B_{k,j,n}\mathbb{E}_{\mathcal{H}_{1,k_n^\star}}(T_{k,n-k}) \\ &= R_{k,j} + S_{k,j}. \end{aligned}$$

Thus $T_{k,j} - Q_{k,j}$ is decomposed into a random part $R_{k,j}$ and a deterministic part $S_{k,j}$. Over $\Omega_n$, $R_{k,j}$ is a function of $v_{(k)}, \dots, v_{(n-k_n^\star)}$. Before going further, we now recall the following result:

Let $Z_{(1)} \ge \dots \ge Z_{(n)}$ be an ordered sequence of independent $Exp(1)$ random variables. For $1 \le j \le n$, let $T_j = \sum_{i=1}^{j} Z_{(i)}$. Introduce for $t \in [0,1]$ the random process $d_n(t) = T_{[nt]} - \mathbb{E}(T_{[nt]}|T_n)$. Then it is shown in [11] that $\frac{1}{\sqrt{n}} d_n(t)$, as a process indexed on $t \in [0,1]$, converges in distribution to a certain zero mean Gaussian process $\Delta$.

To use this result, let $k = [tn]$ and $j = [sn]$, for $0 < t < 1 - c$ and $0 < s < 1 - t^\star - t$, for $c$ in [**AF3**]. Then $\frac{1}{\sqrt{n-k}}(T_{k,j} - Q_{k,j})\mathbf{1}_{\Omega_n} = \frac{1}{\sqrt{n-k}}(T_{[tn],[sn]} - Q_{[tn],[sn]})\mathbf{1}_{\Omega_n}$, as a process indexed by $(t,s) \in (0,1)^2$, converges in distribution to the zero-mean Gaussian process:

$$\Gamma_{t,s} = \sqrt{\frac{1-t^\star}{1-t}}\left[\Delta\left(\frac{t+s-t^\star}{1-t^\star}\right) - \Delta\left(\frac{t-t^\star}{1-t^\star}\right)\right],$$

similarly, $\frac{1}{\sqrt{n-k}} B_{k,j,n}\mathbb{E}_{\mathcal{H}_{1,k_n^\star}}(T_{k,n-k})\mathbf{1}_{\Omega_n}$ converges in distribution to another zero-mean Gaussian process, and so does their sum, $\frac{1}{\sqrt{n-k}} R_{k,j}\mathbf{1}_{\Omega_n}$.

On the other hand,

$$(5) \qquad S_{k,j} = \sum_{i=1}^{k-k_n^\star} (a_{i+j} - a_i + B_{k,j,n}(a_{i+n-k} - a_i)),$$

so that there exists a constant $\gamma > 0$, which depends on $c$ in [**AF3**], such that for all $n \ge 1$, $k_n^\star < k \le n - K_n$, we have

$\sup_{1\leq j\leq n-k}|S_{k,j}| \geq \gamma(k-k_n^\star)$. Finally we use the following inequality:

$$\mathbb{P}_{k_n^\star}(\widehat{k}_n - k_n^\star > nu_n) \leq \mathbb{P}(\eta_{k_n^\star} > \inf_{k-k_n^\star > nu_n} \eta_k).$$

From Equation (5), $S_{k_n^\star,j} = 0$, hence it follows that:

$$\begin{aligned}\sqrt{n-k_n^\star}\,\eta_{k_n^\star} &= \sup_{1\leq j\leq n-k_n^\star} |R_{k_n^\star,j} + S_{k_n^\star,j}| \\ &= \sup_{1\leq j\leq n-k_n^\star} |R_{k_n^\star,j}| \\ &\leq \sup_{k\geq k_n^\star}\sup_{1\leq j\leq n-k} |R_{k,j}|.\end{aligned}$$

On the other hand,

$$\begin{aligned}\sqrt{n-k}\inf_{k-k_n^\star>nu_n}\eta_k &= \inf_{k-k_n^\star>nu_n}\sup_{1\leq j\leq n-k}|R_{k,j}+S_{k,j}| \\ &\geq \inf_{k-k_n^\star>nu_n}\sup_{1\leq j\leq n-k}|S_{k,j}| \\ &\quad - \sup_{k\geq k_n^\star}\sup_{1\leq j\leq n-k}|R_{k,j}|,\end{aligned}$$

so that we have:

$$\begin{aligned}&\mathbb{P}_{k_n^\star}(\widehat{k}_n - k_n^\star > nu_n) \\ &\quad\leq \mathbb{P}(C \sup_{\substack{k\geq k_n^\star \\ 1\leq j\leq n-k}} |R_{k,j}| \geq \inf_{k-k_n^\star>nu_n}\sup_{1\leq j\leq n-k}|S_{k,j}|) \\ &\qquad + \mathbb{P}(\Omega_n^c) \\ &\quad\leq \mathbb{P}(C \sup_{\substack{k\geq k_n^\star \\ 1\leq j\leq n-k}} |R_{k,j}| \geq \gamma n u_n) + \mathbb{P}(\Omega_n^c),\end{aligned}$$

where $C$ is a constant which depends on $c$ in [**AF3**]. This last probability vanishes as $n$ goes to infinity, due to the weak convergence of $R_{k,j}\mathbf{1}_{\Omega_n}$.

## APPENDIX B. DETAILS OF THE EM ALGORITHM FOR THE TWO-CLASS GMM WITH A ZERO-MEAN CLASS

We consider the following model:

$$\begin{aligned} Y_i|Z_i = j &\overset{iid}{\sim} \mathcal{N}(\mu_j, \sigma_j^2), \quad i=1,\ldots,n,\ j=0,1, \\ Z_i &\overset{iid}{\sim} \mathcal{B}(1,p_1), \end{aligned} \tag{6}$$

where $\mu_0 = 0$, and $p_j$ represents the proportion of data in class $j$, so that the vector of model parameters is: $\theta = (p_0, \mu_1, \sigma_0, \sigma_1)$.

Having initialized $\theta$ to $\theta^{(0)}$, the EM algorithm alternates the following steps:

**E-step.** Compute the conditional law of the indicator variable $Z_i$ at step $t$, that is, the Bernoulli defined by:

$$\begin{aligned}\mathbb{P}(Z_i = j|Y_i, \theta^{(t)}) &= \frac{f(Y_i|Z_i=j,\theta^{(t)})p_j^{(t)}}{\sum_j f(Y_i|Z_i=j,\theta^{(t)})p_j^{(t)}} \\ &:= p_{ij}^{(t)}.\end{aligned} \tag{7}$$

**M-step.** Update the estimates of model parameters by maximizing the conditional expectation of the complete log-likelihood,

$$\theta^{(t+1)} = \arg\max_\theta \mathbb{E}\left[\sum_i \log f(Y_i|Z_i,\theta)|\boldsymbol{Y}, \theta^{(t)}\right],$$

the expectation being taken with respect to the conditional distribution of the indicator variables $Z_i$ computed in the previous step. This yields:

$$\begin{aligned} p_j^{(t+1)} &= \frac{\sum_i p_{ij}^{(t)}}{n}; \\ \mu_1^{(t+1)} &= \frac{\sum_i p_{ij}^{(t)} Y_i}{\sum_i p_{ij}^{(t)}}; \\ \sigma_j^{2(t+1)} &= \frac{\sum_i p_{ij}^{(t)}(Y_i - \mu_j^{(t)})^2}{\sum_i p_{ij}^{(t)}}. \end{aligned} \tag{8}$$

Note that, throughout the iterations, $\mu_0^{(t)} \equiv 0$.

*B.0.0.1. Initialization.* An initial guess for $\sigma_0^2$ is provided by the negative data, which consists mainly of null data:

$$\sigma_0^{2(0)} = \sharp\{Y_i < 0\}^{-1}\sum_{Y_i>0} Y_i^2.$$

Then, we use a kernel estimate of the data density:

$$\widehat{f}(x) = \frac{1}{n}\sum_i \frac{K\left(\frac{x-Y_i}{h_n}\right)}{h_n}, \tag{9}$$

for a symmetric, positive kernel $K$, and a bandwidth $h_n$. In practice, we used the Gaussian kernel $K(x) = 1/\sqrt{2\pi}e^{-x^2/2}$, and $h_n = \sqrt{n}$.

By identifying the null mode of the data density kernel estimate to the null component of the mixture model, we then obtained an initial guess for the mixture weights:

$$p_0^{(0)} = \widehat{f}(0)\sqrt{2\pi\sigma_0^{2(0)}}.$$

Finally, the conditional law of the indicator variables is approached by:

$$p_{i1}^{(0)} = \min\left\{1; \frac{p_0^{(0)}\exp\{-(Y_i^2/2\sigma_0^{2(0)})\}}{\sqrt{2\pi\sigma_0^{2(0)}}}\right\}.$$

These initial guesses are used to derive initial model parameter values $\theta^{(1)}$, via the M-step described above, for $t=0$.

*Received 22 October 2010*

## REFERENCES

[1] Beckmann, C., Woolrich, M., and Smith, S. Gaussian / gamma mixture modelling of ica/glm spatial maps. In *Ninth Int. Conf. on Functional Mapping of the Human Brain*, 2003.

[2] Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995. MR1325392

[3] Benjamini, Y. and Hochberg, Y. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001. MR1869245

[4] Delyon, B., Lavielle, M., and Moulines, E. Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, 27(1):94–128, 1999. MR1701103

[5] Dempster, A. P., Laird, A. P., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977. MR0501537

[6] Efron, B., Hastie, T., Johnstone, L., and Tibshirani, R. Least angle regression. *Annals of Statistics*, 32:407–499, 2004. MR2060166

[7] Friston, K. J. *Human Brain Function*, chapter 2, pages 25–42. Academic Press, 1997.

[8] Ge, Y., Dudoit, S., and Speed, T. P. Resampling-based multiple testing for microarray data analysis. Technical report, Department of Statistics, University of California, Berkeley 2. Division of Biostatistics, University of California, Berkeley 3, 2003. MR1993286

[9] Hastie, T., Tibshirani, R., and Friedman, J., editors. *The Elements of Statistical Learning*. Springer Series in Statistics, 2001. MR1851606

[10] Hu, J. X., Zhao, H., and Zhou, H. H. False discovery rate control with groups. *Journal of the American Statistical Association*, 105:1215–1227, 2010. MR2752616

[11] Lavielle, M. and Ludeña, C. Random threshold for linear model selection. *ESAIM: Probability and Statistics*, 12:173–195, 2007. MR2374637

[12] MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 1st edition, June 2003. MR2012999

[13] Massart, S. *Concentration Inequalities and Model Selection*. Lecture Notes in Mathematics. Springer, 2003.

[14] Nichols, T. E. and Hayasaka, S. Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research*, 12(5):419–446, 2003. MR2005445

[15] Perone Pacifico, M., Genovese, C., Verdinelli, I., and Wasserman, L. False discovery control for random fields. *Journal of the American Statistical Association*, 99(468):1002–1014, 2004. MR2109490

[16] Pinel, P., Thirion, B., Mériaux, S., Jobert, A., Serres, J., Le Bihan, D., Poline, J.-B., and Dehaene, S. Fast reproducible identification and large-scale databasing of individual functional cognitive networks. *BMC Neurosci*, 8(1):91, October 2007.

[17] Price, C. J. The anatomy of language: Contributions from functional neuroimaging. *Journal of Anatomy*, 197(3):335–359, October 2000.

[18] Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., and Evans, A. C. A general statistical analysis for fMRI data. *NeuroImage*, 15(1):1–15, January 2002.

Merlin Keller
EDF R&D, Chatou
France
E-mail address: merlin.keller@edf.fr

Marc Lavielle
INRIA Saclay
France

University of Paris Sud, Orsay
France
E-mail address: Marc.Lavielle@math.u-psud.fr