Statistica Sinica **28** (2018), 229-253
doi:https://doi.org/10.5705/ss.202016.0011

ALMOST THE BEST OF THREE WORLDS: RISK, CONSISTENCY AND OPTIONAL STOPPING FOR THE SWITCH CRITERION IN NESTED MODEL SELECTION

Stéphanie Van Der Pas and Peter Grünwald

Leiden University and CWI

Abstract: We study the switch distribution, introduced by van Erven, Grünwald and De Rooij (2012), applied to model selection and subsequent estimation. While switching was known to be strongly consistent, here we show that it achieves minimax optimal parametric risk rates up to a $\log\log n$ factor when comparing two nested exponential families, partially confirming a conjecture by Lauritzen (2012) and Cavanaugh (2012) that switching behaves asymptotically like the Hannan-Quinn criterion. Moreover, like Bayes factor model selection, but unlike standard significance testing, when one of the models represents a simple hypothesis, the switch criterion defines a *robust* null hypothesis test, meaning that its Type-I error probability can be bounded irrespective of the stopping rule. Hence, switching is consistent, insensitive to optional stopping and almost minimax risk optimal, showing that, Yang's (2005) impossibility result notwithstanding, it is possible to 'almost' combine the strengths of AIC and Bayes factor model selection.

Key words and phrases: AIC-BIC dilemma, consistency, exponential family, model selection, optional stopping, post model selection estimation, switch distribution, worst-case risk.

1. Introduction

We consider the following standard model selection problem, where we have i.i.d. observations $X_1, \ldots, X_n$ and wish to select between two nested parametric models,

$$\mathcal{M}_0 = \{p_\mu \mid \mu \in M_0\} \quad \text{and} \quad \mathcal{M}_1 = \{p_\mu \mid \mu \in M_1\}. \tag{1.1}$$

Here the X_i are random vectors taking values in some set $\mathcal{X}$, $M_1 \subseteq \mathbb{R}^{m_1}$ for some $m_1 > 0$ and $\mathcal{M}_0 = \{p_\mu : \mu \in M_0\} \subset \mathcal{M}_1$ represents an m_0-dimensional submodel of $\mathcal{M}_1$, where $0 \leq m_0 < m_1$. We may thus denote $\mathcal{M}_0$ as the 'simple' and $\mathcal{M}_1$ as the 'complex' model. We assume that $\mathcal{M}_1$ is an exponential family, represented as a set of densities on $\mathcal{X}$ with respect to some fixed underlying measure, so that p_μ represents the density of the observations, and we take it to be given in

its mean-value parameterization. As the notation indicates, we require, without loss of generality, that the parameterizations of $\mathcal{M}_0$ and $\mathcal{M}_1$ coincide, that is $M_0 \subset M_1$ is itself a set of m_1-dimensional vectors, the final $m_1 - m_0$ components of which are fixed to known values. We restrict ourselves to the case in which both M_1 and the restriction of M_0 to its first m_0 components are products of open intervals.

Most model selection methods output not just a decision $\delta(X^n) \in \{0, 1\}$, but also an indication $r(X^n) \in \mathbb{R}$ of the strength of evidence, such as a p-value or a Bayes factor. As a result, such procedures can often be interpreted as methods for hypothesis testing, where $\mathcal{M}_0$ represents the *null* model and $\mathcal{M}_1$ the alternative; a very simple example of our setting is when the X_i consist of two components $X_i \equiv (X_{i1}, X_{i2})$, which according to $\mathcal{M}_1$ are independent Gaussians, whereas under $\mathcal{M}_2$ they can have an arbitrary bivariate Gaussian distribution and hence can be dependent. Since we allow $\mathcal{M}_0$ to be a singleton, this setting also includes some simple, classical yet important settings such as testing whether a coin is biased ($\mathcal{M}_0$ is the fair coin model, $\mathcal{M}_1$ contains all Bernoulli distributions).

We consider three desirable properties of model selection methods: (a) optimal worst-case risk rate of post-model selection estimation (with risk measured in terms of squared error loss, squared Hellinger distance, Rényi or Kullback-Leibler divergence); (b) consistency, and (c), for procedures which also output a strength of evidence $r(X^n)$, whether the validity of the evidence is insensitive to optional stopping under the null model. We evaluate the recently introduced model selection criterion δ_{sw} based on the switch distribution (van Erven, Grünwald and De Rooij (2012)) on properties (a), (b) and (c).

The switch distribution, introduced by[1] van Erven, Grünwald and de Rooij (2007), was originally designed to address the *catch-up phenomenon*, which occurs when the best predicting model is not the same across sample sizes. The switch distribution can be interpreted as a modification of the Bayesian predictive distribution. It also has an MDL interpretation: if one corrects standard MDL approaches (Grünwald (2007)) to take into account that the best predicting method changes over time, one naturally arrives at the switch distribution. Lhéritier and Cazals (2015) describe a practical application for two-sample sequential testing related to the developments in this paper, but in a nonparametric context. We briefly give the definitions relevant to our setting in Section 2; for

[1]Matlab code for implementing model selection, averaging and prediction by the switch distribution is available at `http://www.blackwellpublishing.com/rss`. In general run times are comparable to those of the corresponding Bayesian methods.

all further details we refer to van Erven, Grünwald and De Rooij (2012) and S5 in the Supplementary Materials.

When evaluating any model selection method, there is a well-known tension between properties (a) and (b): the popular AIC method (Akaike (1973)) achieves the minimax optimal parametric rate of order $1/n$ in the above problem, but is inconsistent; the same holds for the many popular model selection methods that asymptotically tend to behave like AIC, such as k-fold (for fixed k) and leave-one-out-cross-validation, the bootstrap and Mallow's C_p in linear regression (Efron (1986); Shao (1997); Stone (1977)). On the other hand, BIC (Schwarz (1978)) is consistent in the sense that, for large enough n, it selects the smallest model containing the 'true' μ, but it misses the minimax parametric rate by a factor of $\log n$. The same holds for traditional Minimum Description Length (MDL) approaches (Grünwald (2007)) and Bayes factor model selection (BFMS) (Kass and Raftery (1995)), of which BIC is an approximation. This might lead one to wonder if there exists a single method that is optimal in both respects. A key result by Yang (2005) shows that this is impossible: any consistent method misses the minimax optimal rate by a factor $g(n)$ with $\lim_{n\to\infty} g(n) = \infty$.

In Section 4.2 we show that, Yang's result notwithstanding, the switch distribution allows us to get very close to satisfying properties (a) and (b) at the same time, at least in the above problem (Yang's result was shown in a nested linear regression rather than our exponential family context, but it does hold in our exponential family setting as well; see the discussion at the end of Section 3.3). We prove that in our setting, the switch model selection criterion δ_{sw} misses the minimax optimal rate only by a factor of $g_{\mathrm{sw}}(n) \asymp \log\log n$ (Theorem 1). Property (b), strong consistency, was shown by van Erven, Grünwald and De Rooij (2012). The factor $g_{\mathrm{sw}}(n) \asymp \log\log n$ is an improvement over the factor resulting from Bayes factor model selection, $g_{\mathrm{BFMS}}(n) \asymp \log n$. Indeed, as discussed in the introduction of van Erven, Grünwald and De Rooij (2012), the catch-up phenomenon that the switch distribution addresses is intimately related to the rate-suboptimality of Bayesian inference. van Erven, Grünwald and De Rooij (2012) show that, while model selection based on switching is consistent, sequential prediction based on model averaging with the switching method achieves minimax optimal *cumulative* risk rates in general parametric and nonparametric settings, where the cumulative risk at sample size n is obtained by summing the standard, instantaneous risk from 1 to n. In contrast, in nonparametric settings, standard Bayesian model averaging typically has a cumulative risk rate that is larger by a $\log n$ factor. Using the cumulative risk is natural in sequential pre-

diction settings, but van Erven, Grünwald and De Rooij (2012) left open the question of how switching would behave for the more standard, instantaneous risk. In contrast to the cumulative setting, we cannot expect to achieve the optimal rate here by Yang's (2005) result, but it is interesting to see that switching gets so close.

We now turn to robustness to optional stopping. While consistency here is an asymptotic and even somewhat controversial notion (see Section 6),there exists a nonasymptotic property closely related to consistency that, while arguably more important in practice, has received relatively little attention in the literature. This is the insensitivity to optional stopping. In statistics, the issue was thoroughly discussed, yet never completely resolved, in the 1960s; nowadays, it is viewed as a highly desirable feature of testing methods by, for example, psychologists; see Wagenmakers (2007); Sanborn and Hills (2014). In particular, it is often argued (Wagenmakers (2007)) that the fixed stopping rule required by the classical Neyman-Pearson paradigm severely and unnecessarily restricts the application domain of hypothesis testing, invalidating much of the p-values reported in the psychological literature. Some 55% of psychologists admitted in a survey to deciding whether to collect more data after looking at their results to see if they were significant (John, Loewenstein and Prelec (2012)). We analyze property (c) in terms of *robust null hypothesis tests*, formally defined in Section 5. A method defines a robust null hypothesis test if (1) it outputs evidence $r(X^n)$ that does not depend on the stopping rule used to determine n, and (2) (some function of) $r(X^n)$ gives a bound on the Type-I error that is valid no matter what the stopping rule. Standard (Neyman-Pearson) null hypothesis testing and tests derived from AIC-type methods are not robust in this sense. For example, such tests cannot be used if the stopping rule is simply unknown, as is often the case when analyzing externally provided data — but this is just the tip of an iceberg of problems with nonrobust tests. For an exhaustive review of such problems we refer to Wagenmakers (2007) who builds on, amongst others, Berger and Wolpert (1988) and Pratt (1962).

Now, as first noted by Edwards, Lindman and Savage (1963), in simple versus composite testing, the output of BFMS, the Bayes factor, does provide a robust null hypothesis test. This is one of the main reasons why for example, in psychology, Bayesian testing is becoming more and more popular (Dienes (2011); Andrews and Baguley (2012)), even among 'frequentist' researchers (Sanborn and Hills (2014)). In Section 5 we show that the same holds for the switch criterion: if $\mathcal{M}_0$ is a singleton, so (1.1) reduces to a simple versus composite hypothesis

test, then the evidence $r(X^n)$ associated with the switching criterion has the desired robustness property as well, and thus in this sense behaves like the Bayes factor method. The advantage, from a frequentist point of view, of switching as compared to Bayes is then that switching is more sensitive: our risk rate results directly imply that the Type II error (1 − power) of the switch criterion goes to 0 as soon as, at sample size n, the distance between the 'true' distribution μ_1 and the null model, $\inf_{\mu \in M_0} \|\mu - \mu_1\|_2^2$, is of order $(\log \log n)/n$; for Bayes factor testing, this distance must be of order $(\log n)/n$ (this was informally recognized by Lhéritier and Cazals (2015), who reported substantially larger power of switching as compared to the Bayes factor method in a sequential two-sample test setting).

Thus, for singleton $\mathcal{M}_0$, switching gives us minimax rate optimality up to a $\log \log n$ factor (in contrast to BFMS), consistency (in contrast to AIC-type methods), and nonasymptotic insensitivity to optional stopping (in contrast to standard Neyman-Pearson testing), in combination with a small Type-II error. For composite $\mathcal{M}_0$, we show that nonasymptotic robustness to optional stopping still holds, albeit only in a much weaker sense — thus pointing towards an obvious goal for future work: the modification of the switch distribution to get full optional stopping robustness for composite $\mathcal{M}_0$.

Organization This paper is organized as follows. The switch criterion is introduced in Section 2. In Section 3, we provide some preliminaries: we list the loss/risk functions for which our result holds, describe the sets in which the truth is assumed to lie, and discuss the tension between consistency and rate-optimality. Suitable post-model-selection estimators to be used in combination with the switch criterion are introduced in Section 4, after which our main result on the worst-case risk of the switch criterion is stated. We also go into the relationship between the switch criterion and the Hannan-Quinn criterion in that section. In Section 5 we define robust null hypothesis tests, give some examples, and show that testing by switching has the desired nonasymptotic robustness to optional stopping; in constrast, AIC does not satisfy such a property at all and the Hannan-Quinn criterion only satisfies an asymptotic analogue. We also provide some simulations that illustrate our results. Section 6 provides some additional discussion and ideas for future work. Proofs are given in the Supplementary Materials.

Notation and Conventions We use $x^n = x_1, \ldots, x_n$ to denote n observations, each taking values in a sample space $\mathcal{X}$. For a set of parameters M, $\mu \in M$, and $x \in \mathcal{X}$, $p_\mu(x)$ denotes the density or mass function of x under the distribution

$\mathbb{P}_\mu$ of random variable X, taking values in $\mathcal{X}$. This is extended to n outcomes by independence, so that $p_\mu(x^n) := \prod_{i=1}^n p_\mu(x_i)$ and $\mathbb{P}_\mu(X^n \in A_n)$, abbreviated to $\mathbb{P}_\mu(A_n)$, denotes the probability that $X^n \in A_n$ for $X^n = X_1, \ldots, X_n$ i.i.d. $\sim \mathbb{P}_\mu$. Similarly, $\mathbb{E}_\mu$ denotes expectation under $\mathbb{P}_\mu$. We write $a_n \asymp b_n$ to denote $0 < \lim_{n\to\infty} \inf a_n/b_n \leq \lim_{n\to\infty} \sup a_n/b_n < \infty$. When we refer to a sample size n, $n \geq 3$.

When we refer to standard properties of exponential families they can be found, in precise form, in (Barndorff-Nielsen (1978)) and, on a less formal level, in (Grünwald (2007, Chapters 18,19)).

2. Model Selection by Switching

The *switch distribution* (van Erven, Grünwald and de Rooij (2007); van Erven, Grünwald and De Rooij (2012)) is a modification of the Bayesian predictive distribution, inspired by Dawid (1984) 'prequential' approach to statistics and the related *Minimum Description Length* (MDL) Principle (Barron, Rissanen and Yu (1998); Grünwald (2007)). The corresponding *switch criterion* can be thought of as Bayes factor model selection with a prior on meta-models, where each meta-model consists of a sequence of basic models and associated starting times: until time t_1, follow model k_1, from time t_1 to t_2, follow model k_2, and so on. The fact that we only need to select between two nested parametric models allows us to considerably simplify the set-up of van Erven, Grünwald and De Rooij (2012), who dealt with countably infinite sets of arbitrary models.

It is convenient to directly introduce the switch criterion as a modification of the Bayes factor model selection (BFMS). Assuming equal prior $1/2$ on each of the models $\mathcal{M}_0$ and $\mathcal{M}_1$, BFMS associates each model $\mathcal{M}_k$, $k \in \{0, 1\}$, with a *marginal distribution* $p_{B,k}$ with

$$p_{B,k}(x^n) := \int_{\mu \in M_k} \omega_k(\mu) p_\mu(x^n) d\mu, \tag{2.1}$$

where ω_k is a prior density on M_k. It then selects model $\mathcal{M}_1$ if and only if $p_{B,1}(x^n) > p_{B,0}(x^n)$.

The basic idea behind MDL model selection is to generalize this in the sense that each model $\mathcal{M}_k$ is associated with *some* 'universal' distribution $p_{U,k}$; one then picks the k for which $p_{U,k}(x^n)$ is largest. $p_{U,k}$ may be set to the Bayesian marginal distribution, but other choices may be preferable in some situations. Switching is an instance of this; in our simplified setting, it amounts to associating $\mathcal{M}_0$ with a Bayes marginal distributon $p_{B,0}$ as before. $p_{U,1}$ however is set to

the *switch distribution* $p_{\mathrm{sw},1}$. This distribution corresponds to a switch between models $\mathcal{M}_0$ and $\mathcal{M}_1$ at some sample point s, which is itself uncertain; before point s, the data are modelled as coming from $\mathcal{M}_0$, using $p_{B,0}$; after point s, they are modelled as coming from $\mathcal{M}_1$, using $p_{B,1}$. Formally, we denote the strategy that switches from the simple to the complex model after t observations by $\bar{p}_t$; $p_{\mathrm{sw},1}$ is then defined as the marginal distribution by averaging $\bar{p}_t$ over t, with some probability mass function π (analogous to a Bayesian prior) over $t \in \{1, 2, \ldots\}$:

$$\bar{p}_t(x^n) = p_{B,0}(x^{t-1}) \cdot p_{B,1}(x_t, \ldots, x_n \mid x^{t-1}),$$
$$p_{\mathrm{sw},1}(x^n) = \sum_{t=1}^{\infty} \pi(t) \bar{p}_t(x^n),$$

where switching at $t = 1$ corresponds to predicting with $p_{B,1}$ at each data point, and switching at any $t > n$ to predicting with $p_{B,0}$. Even for i.i.d. models, $p_{B,1}(x_t, \ldots, x_n \mid x^{t-1})$ usually depends on x^{t-1} — the Bayes predictive distribution learns from data. The model selection criterion δ_{sw} mapping sequences of arbitrary length to $k \in \{0, 1\}$ is then defined, for each n, as:

$$\delta_{\mathrm{sw}}(x^n) = \begin{cases} 0 & \text{if } \dfrac{p_{\mathrm{sw},1}(x^n)}{p_{\mathrm{B},\,0}(x^n)} \leq 1, \\ 1 & \text{if } \dfrac{p_{\mathrm{sw},1}(x^n)}{p_{\mathrm{B},\,0}(x^n)} > 1. \end{cases} \tag{2.2}$$

When defining $p_{\mathrm{sw},1}$ it is sufficient to consider switching times that are a power of two. Thus, we restrict attention to 'priors' π on switching time with support on $2^0, 2^1, 2^2, \ldots$. For our subsequent results to hold, π should be such that $\pi(2^i)$ decays like $i^{-\kappa}$ for some $\kappa > 1$. An example of such a prior with $\kappa = 2$ is $\pi(2^i) = 1/((i+1)(i+2))$, $\pi(j) = 0$ for any j that is not a power of 2.

To prepare for Theorem 1, we instantiate the switch criterion to the problem (1.1). We define $p_{B,1}$ as any distribution of the form (2.1) where ω_1 is a continuous prior density on M_1 that is strictly positive on all $\mu \in M_1$. Because we parameterized $\mathcal{M}_0$ in terms of an M_0 that has a fixed value on its final $m_1 - m_0$ components, it is an m_0-dimensional family with an m_1-dimensional parameterization, so one cannot easily express a prior on $\mathcal{M}_0$ as a density on M_0. Thus, when M_0 has a single element ν, we define $p_{B,0} = p_\nu$, when $m_0 > 0$, we define $\Pi'_0 : M_0 \to \mathbb{R}^{m_0}$ as the projection of $\mu \in M_0$ on its first m_0 components, and $\Pi'_0(M_0) := \{\Pi'_0(\mu) : \mu \in M_0\}$. For $\mu \in M_0$, we define $p_{\Pi'_0(\mu)} = p_\mu$, and we then let ω_0 be a continuous strictly positive prior density on $\Pi'_0(M_0)$, and we define $p_{B,0}(x^n) := \int_{\mu' \in \Pi'_0(M_0)} \omega_0(\mu') p_{\mu'}(x^n) d\mu'$.

That we associate $\mathcal{M}_1$ with a distribution incorporating a 'switch' from $\mathcal{M}_0$

to $\mathcal{M}_1$ *does not mean* that we really believe that data were sampled, until some point t, according to $\mathcal{M}_0$, then according to $\mathcal{M}_1$; rather, it is suggested by prequential and MDL considerations, that suggest that one should pick the model that performs best in sequentially predicting data. If the data are sampled from a distribution in $\mathcal{M}_1$ that is not in $\mathcal{M}_0$, but quite close to it in KL divergence, then $p_{\mathrm{B},1}$ is suboptimal for sequential prediction, and can be substantially outperformed by $p_{\mathrm{sw},1}$, see van Erven, Grünwald and De Rooij (2012).

The criterion (2.2) is not equivalent to the special case of the construction of van Erven, Grünwald and De Rooij (2012), specialized to two models, but rather a simplification thereof. We do this for ease of exposition; the impact of simplifying the original switch criterion to (2.2) is minimal. Varying the exponent κ in the prior $\pi(2^i) \propto i^{-\kappa}$ defined above — which is a free parameter of the switch distribution — has a stronger effect on the switch criterion than switching between the two versions of the switch distribution. This is explained in the Supplementary Materials, where we also explain why all our results continue to hold if we were to follow the original construction; conversely, the strong consistency result for the construction of van Erven, Grünwald and De Rooij (2012) trivially continues to hold for the criterion (2.2).

3. Rate-Optimality of Post-Model Selection Estimators

This section contains some background to our main result, Theorem 1. In Section 3.1, we first list the loss functions for which our main result holds, and define the CINECSI sets in which the truth is assumed to lie. We then discuss the minimax parametric risk for our model selection problem in Section 3.2. This section ends with a discussion on the generality of the impossibility result of Yang (2005) in Section 3.3.

3.1. Loss functions and CINECSI sets

Let $\mathcal{M} = \{p_\mu \mid \mu \in M\}$ be an exponential family given in its mean-value parameterization with $M \subset \mathbb{R}^m$ a product of m open, possibly but not necessarily unbounded intervals for some $m > 0$. We do not require the family to be 'full'; for example, the Bernoulli model with success probability $\mu \in M_1 = (0.2, 0.4)$ counts as an exponential family in our (standard) definition.

Suppose that we measure the quality of a density $p_{\mu'}$ as an approximation to p_μ by a loss function $L : M \times M \to \mathbb{R}$. The standard definition of the (instantaneous) *risk* of estimator $\breve{\mu} : \bigcup_{i>0} \mathcal{X}^i \to M$ at sample size n, as defined

relative to loss L, is given by its expected loss,

$$R(\mu, \breve{\mu}, n) = \mathbb{E}_\mu \left[L(\mu, \breve{\mu}(X^n)) \right],$$

where $\mathbb{E}_\mu$ denotes expectation over $X_1, \ldots, X_n$ i.i.d. $\sim \mathbb{P}_\mu$. Popular loss functions are: *squared error loss*: $d_{SQ}(\mu', \mu) = \|\mu' - \mu\|_2^2$; *standardized squared error loss*:

$$d_{ST}(\mu' \| \mu) := (\mu - \mu')^T I(\mu')(\mu - \mu'), \tag{3.1}$$

where T denotes transpose, $I(\cdot)$ is the Fisher information matrix, and we view μ and μ' as column vectors; Rényi divergence of order 1/2: $d_R(\mu', \mu) = -2 \log \mathbb{E}_{\mu'}[(p_\mu(X)/p_{\mu'}(X))^{1/2}]$; squared Hellinger distance $d_{H^2}(\mu', \mu) = 2(1 - \mathbb{E}_{\mu'}[(p_\mu(X)/p_{\mu'}(X))^{1/2}])$; and Kullback-Leibler divergence $D(p_{\mu'} \| p_\mu)$, henceforth abbreviated to $D(\mu' \| \mu)$.

There is a direct relationship between the Rényi divergence and squared Hellinger distance:

$$d_{H^2}(\mu', \mu) = 2 \left(1 - e^{-d_R(\mu', \mu)/2} \right). \tag{3.2}$$

We show below that these loss functions are all equivalent (equal up to universal constants) on CINECSI sets, defined as follows.

Definition 1 (CINECSI)**.** *A* CINECSI (Connected, Interior-Non-Empty-Compact-Subset-of-Interior) subset *of a set M is a connected subset of the interior of M that is itself compact and has nonempty interior.*

The following proposition is proved in the Supplementary Materials.

Proposition 1. *Let M be the mean-value parameter space of an exponential family, and let M' be a CINECSI subset of M. Then there exist positive constants $c_1, c_2, \ldots, c_6$ such that for all $\mu, \mu' \in M'$,*

$$c_1 \|\mu' - \mu\|_2^2 \leq c_2 \cdot d_{ST}(\mu' \| \mu) \leq d_{H^2}(\mu', \mu) \leq d_R(\mu', \mu) \leq D(\mu' \| \mu) \leq c_3 \|\mu' - \mu\|_2^2. \tag{3.3}$$

and for all $\mu' \in M', \mu \in M$,

$$d_{H^2}(\mu', \mu) \leq c_4 \|\mu' - \mu\|_2^2 \leq c_5 \cdot d_{ST}(\mu' \| \mu) \leq c_6 \|\mu' - \mu\|_2^2. \tag{3.4}$$

CINECSI subsets are a variation on the INECCSI sets of Grünwald (2007). Our main result holds for all the above loss functions, and for general 'sufficiently efficient' estimators. The equivalence of the losses on CINECSI sets helps in the proofs, but we never require these estimators to be restricted to CINECSI subsets of M — although, since we require M to be open, every 'true' $\mu \in M$ will lie in *some* CINECSI subset M' of M, albeit unknown.

3.2. Minimax parametric risk

We say that a quantity f_n converges at rate g_n if $f_n \asymp g_n$. We say that an estimator $\breve{\mu}$ is *minimax-rate optimal* relative to a model $\mathcal{M} = \{p_\mu \mid \mu \in M\}$ *restricted to a subset* $M' \subset M$ if

$$\sup_{\mu \in M'} R(\mu, \breve{\mu}, n) \asymp \inf_{\dot{\mu}} \sup_{\mu \in M'} R(\mu, \dot{\mu}, n),$$

where $\dot{\mu}$ ranges over all estimators of μ at sample size n.

For parametric models, (3.2) is typically of order $1/n$ when R is defined relative to any of the loss measures defined in Section 3.1 and M' is an arbitrary CINECSI subset of M (van der Vaart (1998)) — models for which this holds include e.g. most location families and all curved exponential families, which include as a special case all standard exponential families. For this reason, we refer to $1/n$ as the *minimax parametric rate*. The restriction $\mu \in M'$ is imposed only on the data-generating distribution, not on the estimators and, since we require models with open parameter sets M such that for every $\delta > 0$, there is a CINECSI subset M'_δ of M with $\sup_{\mu \in M} \inf_{\mu' \in M'_\delta} \|\mu - \mu'\|_2^2 < \delta$, every possible $\mu \in M$ will also lie in some CINECSI subset M'_δ that 'nearly' covers M_δ. This makes the restriction to CINECSI M' a mild one. Still a restriction is necessary; at least for squared error loss, for most exponential families, we have $\inf_{\dot{\mu}} \sup_{\mu \in M'_\delta} R(\mu, \dot{\mu}, n) = C_\delta / n$ for some constant $C_\delta > 0$, which can grow arbitrarily large as $\delta \to 0$.

Now consider a model selection criterion $\delta : \bigcup_{i>0} \mathcal{X}^i \to \{0, 1, \ldots, K-1\}$ that selects, for given data x^n of arbitrary length n, one of a finite number K of parametric models $\mathcal{M}_0, \ldots, \mathcal{M}_{K-1}$ with respective parameter sets $M_0, \ldots, M_{K-1}$. One way to evaluate the quality of δ is to consider the risk attained after first selecting a model and then estimating the parameter vector μ using an estimator $\breve{\mu}_k$ associated with each model $\mathcal{M}_k$. This *post-model selection estimator* (Leeb and Pötscher (2005)) is denoted by $\breve{\mu}_{\breve{k}}(x^n)$, where $\breve{k}$ is the index of the model selected by δ. The risk of a model selection criterion δ is thus $R(\mu, \delta, n) = \mathbb{E}_\mu \left[L(\mu, \breve{\mu}_{\breve{k}}(X^n)) \right]$, where L is a given loss function, and its worst-case risk relative to μ restricted to $M'_k \subset M_k$ is given by

$$\sup_{\mu \in M'_k} R(\mu, \delta, n) = \sup_{\mu \in M'_k} \mathbb{E}_\mu \left[L(\mu, \breve{\mu}_{\breve{k}}(X^n)) \right]. \tag{3.5}$$

Definition 2. *A model selection criterion* δ achieves the minimax parametric rate *if there exist estimators* $\breve{\mu}_k$*, one for each* $\mathcal{M}_k$ *under consideration, such that, for* every *CINECSI subset* M'_k *of* M*,*

$$\sup_{\mu \in M_k'} R(\mu, \delta, n) \asymp \frac{1}{n}.$$

The restriction $\mu \in M_k'$ is imposed only on the data-generating distribution, not on the estimators.

3.3. The result of Yang (2005) transplanted to our setting

We specialize the above setting to problem (1.1) where we select between two nested exponential families, given in their mean-value parameterization. Thus $\mathcal{M}_1$ contains distributions from an exponential family parametrized by an m_1-dimensional mean vector μ, and the 'simple' model $\mathcal{M}_0$ contains distributions with the same parametrization, where the final $m_1 - m_0$ components are fixed to values $\nu_{m_0+1}, \ldots, \nu_{m_1}$. We require that M_1 and M_0 are of the form

$$\begin{aligned} M_1 &= (\zeta_{1,1}, \eta_{1,1}) \times \cdots \times (\zeta_{1,m_1}, \eta_{1,m_1}), \\ M_0 &= (\zeta_{0,1}, \eta_{0,1}) \times \cdots \times (\zeta_{0,m_0}, \eta_{0,m_0}) \times \{\nu_{m_0+1}\} \times \cdots \times \{\nu_{m_1}\} \end{aligned} \tag{3.6}$$

where, for $j = 1, \ldots, m_0$, we have $-\infty \le \zeta_{1,j} \le \zeta_{0,j} < \eta_{0,j} \le \eta_{1,j} \le \infty$; and for $j = m_0 + 1, \ldots, m_1$, we have $-\infty \le \zeta_{1,j} < \nu_j < \eta_{1,j} \le \infty$.

For example, $\mathcal{M}_1$ could contain all normal distributions with mean μ and variance σ^2, with mean value parameters $\mu_1 = \mu^2 + \sigma^2$ and $\mu_2 = \mu$, and $M_1 = (0, \infty) \times (-\infty, \infty)$, while $\mathcal{M}_0$ could contain all normal distributions with mean zero and unknown variance σ^2, so $M_0 = (0, \infty) \times \{0\}$.

Yang (2005) showed in a linear regression context that a model selection criterion cannot both achieve the minimax optimal parametric rate and be consistent. Our (3.8) provides some insight into why this can occur. A similar inequality in Yang's paper, in a linear regression context, remains valid in our exponential family setting, and the derivations are essentially equivalent.

For $\mu_1 = (\mu_{1,1}, \ldots, \mu_{1,m_1})^T \in M_1$, take

$$\Pi_0(\mu_1) := (\mu_{1,1}, \ldots, \mu_{1,m_0}, \nu_{m_0+1}, \ldots, \nu_{m_1})^T \tag{3.7}$$

to be the *projection* of μ_1 on M_0. Here Π_0 is a function from $\mathbb{R}^{m_1}$ to $\mathbb{R}^{m_1}$, whereas Π_0' in Section 2 is a function from $\mathbb{R}^{m_1}$ to $\mathbb{R}^{m_0}$; $\Pi_0(\mu_1)$ and $\Pi_0'(\mu_1)$ agree in the first m_0 components. Thus $\Pi_0(\mu_1)$ minimizes, among all $\mu \in M_0$, the squared Euclidean distance $\|\mu - \mu_1\|_2^2$ to p_{μ_1}, and it also minimizes, among $\mu \in M_0$, the KL divergence $D(p_{\mu_1} \| p_\mu)$ (Grünwald (2007, Chap. 19)); we think of it as the 'best' approximation of the 'true' μ_1 within M_0, and usually abbreviate $\Pi_0(\mu_1)$ to μ_0.

Let A_n be the event that the complex model is selected at sample size n.

Since $\mathcal{M}_1$ is an exponential family, the MLE $\widehat{\mu}_1$ is unbiased and $\widehat{\mu}_0$ coincides with $\widehat{\mu}_1$ in the first m_0 components, so that $\mathbb{E}_{\mu_1}\left[\mu_0 - \widehat{\mu}_0(X^n)\right] = 0$. Hence we can rewrite, for any $\mu_1 \in M_1$, the squared error risk as

$$\begin{aligned} R(\mu_1, \delta, n) &= \mathbb{E}_{\mu_1}\left[\mathbf{1}_{A_n}\|\mu_1 - \widehat{\mu}_1(X^n)\|_2^2 + \mathbf{1}_{A_n^c}\|\mu_1 - \widehat{\mu}_0(X^n)\|_2^2\right] \\ &= \mathbb{E}_{\mu_1}\left[\mathbf{1}_{A_n}\|\mu_1 - \widehat{\mu}_1(X^n)\|_2^2 + \mathbf{1}_{A_n^c}\|\mu_0 - \widehat{\mu}_0(X^n)\|_2^2 + 1_{A_n^c}\|\mu_1 - \mu_0\|_2^2\right] \\ &\leq \mathbb{E}_{\mu_1}\left[\|\mu_1 - \widehat{\mu}_1(X^n)\|_2^2 + \|\mu_0 - \widehat{\mu}_0(X^n)\|_2^2\right] + \mathbb{P}(A_n^c)\|\mu_1 - \mu_0\|_2^2 \\ &\leq 2R(\mu_1, \widehat{\mu}_1, n) + \mathbb{P}(A_n^c)\|\mu_1 - \mu_0\|_2^2. \end{aligned} \tag{3.8}$$

The first term on the right of (3.8) is of order $1/n$. The second term depends on the probability of selecting the simple model when it is not true. A low worst-case risk is attained if this probability is small, even if the true parameter is close to μ_0. This leaves the possibility for a risk-optimal model selection criterion to incorrectly select the complex model with high probability, or, a risk-optimal model selection method may not be consistent if the simple model is correct. The theorem by Yang (2005) shows that it cannot be. It seems likely that his result holds in much more general settings: a procedure attains a low worst-case risk by selecting the complex model with high probability, which leads to inconsistency if the simple model is correct. The same holds in our exponential family problem (1.1) as long as $\mathcal{M}_0 = \{\nu\}$ is a singleton (van der Pas (2013)). As the switch criterion is strongly consistent (van Erven, Grünwald and De Rooij (2012)), the worst-case risk rate of the switch criterion cannot be of the order $1/n$ in general.

4. Main Result

We perform model selection by using the switch criterion; after the model selection, we estimate the underlying parameter μ. We discuss post-model selection estimators suitable to our problem in Section 4.1. We present our main result in Section 4.2: the worst-case risk for the switch criterion under the loss functions listed in Section 3.1 attains the minimax parametric rate up to a $\log\log n$ factor.

4.1. Post-model selection: sufficiently efficient estimators

Our goal is to determine the worst-case rate for the switch criterion applied to two nested exponential families, which we combine with an estimator as follows: if the simple model is selected, μ is estimated by an estimator $\breve{\mu}_0$ with range M_0; if the complex model is selected, μ is estimated by another estimator $\breve{\mu}_1$ with range M_1. Our result holds for all estimators $\breve{\mu}_0$ and $\breve{\mu}_1$ that are *sufficiently efficient*:

Definition 3 (sufficiently efficient)**.** *The estimators* $\{\breve{\mu}_k \to M_k \mid k \in \{0,1\}\}$ *are* sufficiently efficient *with respect to a divergence measure* $d_{\mathtt{gen}}(\cdot\|\cdot)$ *if, with* $\mu_0 = \Pi_0(\mu_1)$*, for every CINECSI subset* M_1' *of* M_1 *there exists a constant* $C > 0$ *such that for all* n*,*

$$\sup_{\mu_1 \in M_1'} \mathbb{E}_{\mu_1}[d_{\mathtt{gen}}(\mu_0 \| \breve{\mu}_0)] \leq C \cdot \sup_{\mu_1 \in M_1'} \mathbb{E}_{\mu_1}[d_{\mathtt{gen}}(\mu_1 \| \breve{\mu}_1)] \leq \frac{C}{n}. \tag{4.1}$$

This is a stronger requirement than just rate-optimality: we additionally require that, if the estimate $\breve{\mu}_0$ is used on data sampled from $\mu_1 \in M_1$ ('misspecification'), then $\breve{\mu}_0$ converges to μ_0, the best approximation of μ_1 within M_0 at rate $O(1/n)$. In the Supplementary Materials we provide a detailed discussion of sufficiently efficient estimators by means of several examples.

4.2. Main result: risk of the switch criterion

We show that for the exponential family problem under consideration, the worst-case instantaneous risk rate of δ_{sw} is of order $(\log\log n)/n$, while maintaining consistency.

The theorem holds for squared error loss, standardized squared error loss, KL divergence, Rényi divergence of order 1/2, or squared Hellinger distance, with $d_{\mathtt{gen}}$ denoting any of them. Apart from the sufficiently efficient condition on $\breve{\mu}_0$ and $\breve{\mu}_1$, we rule out the use of improper prior densities, and require that the prior probability of switching at time $t = 2^i$ be strictly decreasing and not exponentially small in i. Since these priors are user-defined, these conditions can easily be satisfied.

Theorem 1. *Let* $\mathcal{M}_0 = \{p_\mu \mid \mu \in M_0\}$ *and* $\mathcal{M}_1 = \{p_\mu \mid \mu \in M_1\}$ *be nested exponential families in their mean-value parameterization, where* $M_0 \subseteq M_1$ *are of the form* (3.6)*. If* $\breve{\mu}_0$ *and* $\breve{\mu}_1$ *are sufficiently efficient estimators relative to the chosen loss* $d_{\mathtt{gen}}$*; and if* δ_{sw} *is constructed with* $p_{B,0}$ *and* $p_{B,1}$ *defined as in Section 2 with priors* ω_k *that admit a strictly positive, continuous density; and if* $p_{\mathrm{sw},1}$ *is defined relative to a prior* π *with support on* $\{0,1,2,4,8,\ldots\}$ *and* $\pi(2^i) \propto i^{-\kappa}$ *for some* $\kappa > 1$*, then for every CINECSI subset* M_1' *of* M_1*, we have*

$$\sup_{\mu_1 \in M_1'} R(\mu_1, \delta_{sw}, n) = O\left(\frac{\log\log n}{n}\right),$$

for $R(\mu, \delta_{sw}, n)$ *the risk at sample size* n *defined relative to the chosen loss* $d_{\mathtt{gen}}$*.*

Example 1 (Switching vs. Hannan-Quinn). In their comments on van Erven, Grünwald and De Rooij (2012), Lauritzen (2012) and Cavanaugh (2012) suggested a relationship between the switch model selection criterion and the criterion due to Hannan and Quinn (1979). For the exponential family models under consideration, the Hannan-Quinn criterion with parameter c, denoted as HQ has $\delta_{\mathrm{HQ}}(x^n) = 0$ if

$$-\log p_{\widehat{\mu}_0}(x^n) < -\log p_{\widehat{\mu}_1}(x^n) + c \log\log n,$$

and is 1 otherwise. Hannan and Quinn show that this criterion is strongly consistent for $c > 1$.

As shown by Barron, Birgé and Massart (1999), under some regularity conditions, penalized maximum likelihood criteria achieve worst-case quadratic risk of the order of their penalty divided by n. One can show that this holds in our specific setting and hence, that the worst-case risk rate of HQ for our problem is of order $(\log\log n)/n$. Our main result has the same risk rate achieved by the switch distribution, thus partially confirming the conjecture of Lauritzen (2012) and Cavanaugh (2012): HQ achieves the same risk rate as the switch distribution and, for the right choice of c, is also strongly consistent. Thus the switch distribution and HQ, at least for some specific value c_0, may behave asymptotically indistinguishably. van der Pas (2013) suggests that this is indeed the case if $\mathcal{M}_0$ is a singleton and, in this sense the conjecture of Lauritzen (2012) and Cavanaugh (2012) has only been partially resolved.

To compare the two, for this parametric problem, HQ has the advantage of being simpler to analyze and implement. The criterion δ_{sw} can however, be used to define a robust hypothesis test as in Section 5 below. We show that HQ is insensitive to optional stopping in an asymptotic sense only, whereas robust tests such as the switch criterion are insensitive to optional stopping in a much stronger, nonasymptotic sense. Another advantage of switching is that it can be combined with arbitrary priors and applied more generally, for example when the constituting models are themselves nonparametric (Lhéritier and Cazals (2015)).

5. Robust Null Hypothesis Tests

Bayes factor model selection, the switch criterion, AIC, BIC, HQ, and most model selection methods used in practice are based on thresholding the output of a more informative *model comparison method*. Given data x^n, one outputs a number $r(x^n)$ between 0 and ∞ that is a deterministic function of the data x^n. Every model comparison method r and threshold t has an associated model

selection method $\delta_{r,t}$ that outputs 1 (corresponding to selecting model $\mathcal{M}_1$) if $r(x^n) \leq t$, and 0 otherwise. Such model comparison methods can often be viewed as performing a null hypothesis test with $\mathcal{M}_0$ the null hypothesis, $\mathcal{M}_1$ the alternative hypothesis, and t akin to a significance level.

Example 2 (BFMS). The output of the Bayes factor model comparison method is the posterior odds ratio $r_{\text{Bayes}}(x^n) = \mathbb{P}(\mathcal{M}_0|x^n)/\mathbb{P}(\mathcal{M}_1|x^n)$. The associated model selection method (BFMS) with threshold t selects model $\mathcal{M}_1$ if and only $r_{Bayes}(x^n) \leq t$.

Example 3 (AIC). Standard AIC selects model $\mathcal{M}_1$ if $\log(p_{\widehat{\mu}_1}(x^n)/p_{\widehat{\mu}_0}(x^n)) > m_1 - m_0$, but we consider more conservative versions of AIC that only select $\mathcal{M}_1$ if

$$\log\left(\frac{p_{\widehat{\mu}_1}(x^n)}{p_{\widehat{\mu}_0}(x^n)}\right) - (m_1 - m_0) \geq -\log t. \tag{5.1}$$

We can thus think of AIC as a model comparison method that outputs the left-hand side of (5.1), and that becomes a model selection method when supplied with a particular t.

Neyman-Pearson null hypothesis testing requires the *sampling plan*, or equivalently, the *stopping rule*, to be determined in advance to ensure the validity of the subsequent inference. Greater flexibility in choosing the sample size n is desirable (Wagenmakers (2007) provides examples and discussion). We discuss hypothesis tests that allow such flexibility in that their Type I-error probability remains bounded irrespective of the stopping rule used, and term them *robust* tests. We find that for simple vs. composite testing, both Bayes factor model selection (BFMS) and the switch distribution define such tests, whereas AIC does not and HQ does so only in an asymptotic sense.

5.1. Bayes factors with singleton $\mathcal{M}_0$ are robust under optional stopping

In many cases, for each $0 < \alpha < 1$ there is an associated threshold $t(\alpha)$, a strictly increasing function of α, such that for every $t \leq t(\alpha)$, $\delta_{r,t}$ is a null hypothesis significance test (NHST) with type-I error probability bounded by α. In particular, $\delta_{r,t(\alpha)}$ is a standard NHST with type-I error bounded by α.

We say that model comparison method r defines a *robust null hypothesis test* for null hypothesis $\mathcal{M}_0$ if, for all $\mu_0 \in M_0$ and $0 \leq \alpha \leq 1$,

$$\mathbb{P}_{\mu_0}(\exists n : \delta_{r,t(\alpha)}(X^n) = 1) \leq \alpha. \tag{5.2}$$

Hence, a test that satisfies (5.2) is a NHST test at each fixed significance level α,

independently of the stopping rule used. If a researcher can obtain a maximum of n observations, the probability of incorrectly selecting the complex model remains bounded away from one, regardless of the number of observations made.

We may view the output of BFMS as a 'robust' variation of the p-value. This was noted by Edwards, Lindman and Savage (1963) and interpreted as a frequentist justification for BFMS.

Theorem 2 (Special Case of Eq. (2) of Shafer et al. (2011))**.** *Let $\mathcal{M}_0, \mathcal{M}_1, M_0$ and M_1 be as in Theorem 1 with common support $\mathcal{X} \subset \mathbb{R}^d$ for some $d > 0$. Let $(X_1, X_2, \ldots)$ be an infinite sequence of random vectors all with support $\mathcal{X}$, and fix distributions, $\bar{\mathbb{P}}_0$ and $\bar{\mathbb{P}}_1$ on $\mathcal{X}^\infty$ If for each n, $\bar{p}_j^{(n)}$ represents the marginal density of $(X_1, \ldots, X_n)$ for the first n outcomes under distribution $\bar{\mathbb{P}}_j$, relative to some product measure ρ^n on $(\mathbb{R}^d)^n$ then for all $\alpha \geq 0$,*

$$\bar{\mathbb{P}}_0\left(\exists n : \frac{\bar{p}_0^{(n)}(X^n)}{\bar{p}_1^{(n)}(X^n)} \leq \alpha\right) \leq \alpha.$$

We first apply this result for Bayes factor model selection, with model priors $\pi_0 = \pi_1 = 1/2$, so that $r_{\text{Bayes}}(x^n) = \mathbb{P}(\mathcal{M}_0|x^n)/\mathbb{P}(\mathcal{M}_1|x^n) = p_{B,0}(x^n)/p_{B,1}(x^n)$. We immediately see:

Corollary 1. *If $M_0 = \{\mu_0\}$ is a simple null model, then $p_{B,0} = p_{\mu_0}$ so that from (5.2), if $t(\alpha) = \alpha$, Bayes factor model selection is a robust test.*

For a composite $\mathcal{M}_0$, full robustness requires that (5.2) holds for all $\mu_0 \in M_0$. Our simulations show that this is generally not the case for Bayes factor model selection. We still have robustness in a weaker sense, robustness in prior expectation relative to prior ω_0 on M_0, in that for all $0 \leq \alpha \leq 1$,

$$\mathbb{P}_{B,0}(\exists n : \delta_{r,t(\alpha)}(X^n) = 1) \leq \alpha, \tag{5.3}$$

where $\mathbb{P}_{B,0}$ is the Bayes marginal distribution under prior ω_0. Thus, if the prior ω_0 on model $\mathcal{M}_0$ holds, then the BFMS method still gives robust p-values, independently of the stopping rule.

Remark 1. One may be interested in a significance level α_n that is a fixed function of the sample size n. Both Bayesian and switch-based model comparison may be used in this manner, and Theorem 2 still holds with α replaced by α_n.

5.2. AIC is not, and HQ is only asymptotically robust

Here, for every function $t : (0, 1) \to \mathbb{R}_{>0}$ we have, even for every *single* $0 < \alpha < 1$, that $\delta_{AIC,t(\alpha)}$ is *not* a robust test for significance level α, and AIC

cannot be transformed into a robust test in this sense. For example, compare a 0-dimensional (fixed mean μ_0) with a 1-dimensional Gaussian location family $\mathcal{M}_1$. Evaluating the left hand side of (5.1), $\delta_{AIC,t(\alpha)}$ selects the complex model if

$$\left| \sum_{i=1}^{n} \tilde{X}_i \right| \geq \frac{\sqrt{2n}}{t(\alpha)}, \tag{5.4}$$

where the $\tilde{X}_i$ are variables with mean 0 and variance 1 if $\mathcal{M}_0$ is correct. As a consequence of the Law of the Iterated Logarithm (see for example van der Vaart (1998)), with probability one infinitely many n exist such that the complex model is favored, even though it is incorrect.

In this example, the HQ criterion, in the notation of (5.4), selects the complex model if

$$\left| \sum_{i=1}^{n} \tilde{X}_i \right| \geq \sqrt{2cn \log \log n}.$$

If $c > 1$ (when HQ is strongly consistent), this inequality almost surely fails for infinitely many n, as again follows from the Law of the Iterated Logarithm. This reasoning can be extended to other exponential families, and we find that the HQ criterion with $c > 1$ is robust to optional stopping in the crude, asymptotic sense that the probability that there exist infinitely many sample sizes such that the simple model is incorrectly rejected is zero. Yet HQ does not define a robust hypothesis test in the sense above: to get the numerically precise Type I-error bound (5.2) we would need to define $t(\alpha)$ in a model-and sample-size-dependent manner, which is quite complicated in all cases except the Gaussian location families where the asymptotics hold precisely. The same type of asymptotic robustness holds for the BIC criterion as well.

5.3. Switching with singleton $\mathcal{M}_0$ is robust under optional stopping

As with BFMS, switching can be used as a robust null hypothesis test, as long as $\mathcal{M}_0$ is a singleton: we can view the switch distribution as a model comparison method that outputs odds ratio $r_{\mathrm{sw}}(x^n) = p_{B,0}(x^n)/p_{\mathrm{sw},1}(x^n)$. Until now, we used it to select model 1 if $r_{\mathrm{sw}}(x^n) \leq 1$. If instead we fix a significance level α and select model 1 if $r_{\mathrm{sw}}(x^n) \leq \alpha$, then we immediately see, by applying Theorem 2 in the same way as for the Bayes factor case, that r_{sw} constitutes a robust null hypothesis test as long as $\mathcal{M}_0$ is a singleton model (of course, if we select $\mathcal{M}_1$ as soon as r_{sw} outputs $t \leq \alpha$, then α is merely an upper bound on the Type-I error; the actual value might even be lower, as illustrated in the

simulations below). From a frequentist perspective, switching is preferable to BFMS, since it has substantially better power (Type-II error) properties. As can be seen from (3.8), there is a connection between Type-II error and the risk rate achieved by any model comparison method.

Corollary 2. *Using the same notation and conditions of Theorem 1, for any* $\alpha > 0$*, there exist constants* $C_1, C_2 > 0$ *such that, for every CINECSI subset* M_1' *of* M_1*, and every sequence* $\mu_1^{(1)}, \mu_1^{(2)}, \ldots$ *of elements of* M_1' *satisfying for all* n*,* $\inf_{\mu_0 \in M_0} \|\mu_1^{(n)} - \mu_0\|_2^2 \geq C_1(\log\log n)/n$*, we have*

$$\mathbb{P}_{\mu_1^{(n)}}\left(r_{\mathrm{sw}}(x^n) \geq \alpha\right) \leq \frac{C_2}{\log n}. \tag{5.5}$$

For a fixed significance level, the power of testing by switching goes to 1 as long as the data are sampled from a distribution $\mu_1^{(n)}$ in M_1 farther away from M_0 than $O((\log\log n)/n)$; for BFMS, the power goes to 1 if $\mu^{(n)}$ is farther away than order $O((\log n)/n)$.

Corollary 2 holds for general $\mathcal{M}_0$ including composite ones. Yet robustness to optional stopping only holds if $\mathcal{M}_0$ is a singleton; if $\mathcal{M}_0$ is composite then, using again the same argument as for the Bayes factor case, we see from Theorem 2 that the much weaker 'prior expected robustness' property (5.3) still holds. Simulations show that full robustness does fail if μ_0 is far out in the tails of the prior ω_0.

5.4. Simulation study

We did a simulation to illustrate the differences between AIC, BIC, HQ, and the switch criterion in terms of consistency, strong consistency and robustness to optional stopping. In each setting, two of three models were compared: $\mathcal{M}_0 = \{\mathcal{N}(0,1)\}$; $\mathcal{M}_1 = \{\mathcal{N}(\mu,1), \mu \in \mathbb{R}\}$, with a normal prior with mean zero and variance equal to 100 on μ; $\mathcal{M}_2 = \{\mathcal{N}(\mu,\sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}_{>0}\}$, with a normal-inverse-gamma prior: $\mu|\sigma^2 \sim \mathcal{N}(0, C \times \sigma^2), \sigma^2 \sim IG(\alpha,\beta)$, with $C = 100, \alpha = 1, \beta = 1$.

To illustrate standard consistency, $\mathcal{M}_1$ and $\mathcal{M}_2$ were considered. In the first setting, $\mathcal{M}_1$ was true. $N = 1{,}000$ data sets of length $n = 2{,}500$ were generated from a standard normal distribution, and AIC, BIC, HQ with $c = 1.05$ and δ_{sw} were evaluated at each sample size. The average selected model index (0 for $\mathcal{M}_1$, 1 for $\mathcal{M}_2$) is given in Figure 1.

In the second setting, $\mathcal{M}_2$ was true. The data were generated from a normal distribution with mean 0 and a variance that varied. For each value of σ, $N = 1{,}000$ datasets of length $n = 2{,}500$ were generated, and the four model selection

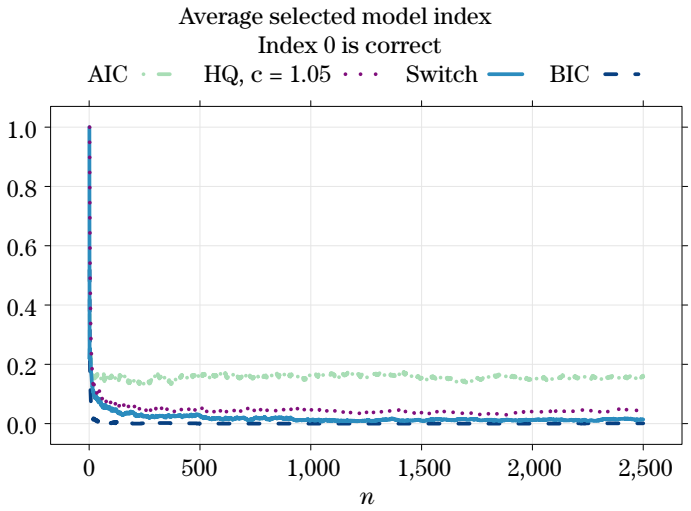

Figure 1. $N = 1{,}000$ data sets of length $n = 2{,}500$ were generated from a standard normal distribution and the criteria were evaluated at each sample size. The figure shows the average selected model index (0 for $\mathcal{M}_1$, 1 for $\mathcal{M}_2$). The true index is 0.

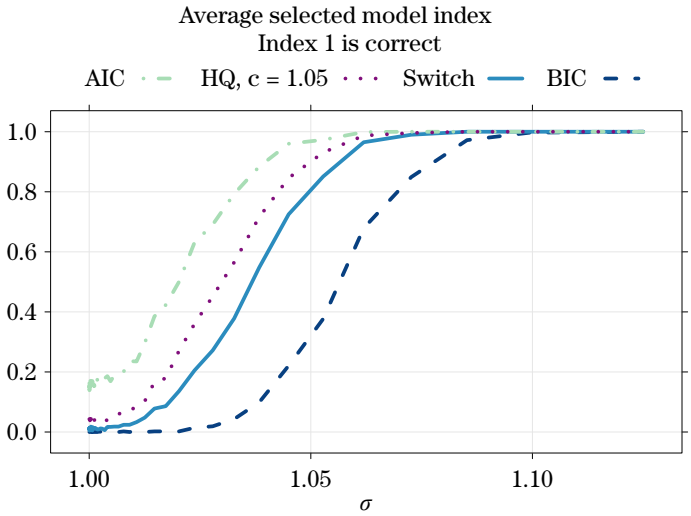

Figure 2. $N = 1{,}000$ data sets of length $n = 2{,}500$ were generated from a normal distribution with mean 0 and variance σ^2 for a range of values of σ. The criteria were evaluated at $n = 2{,}500$. The figure shows the average selected model index (0 for $\mathcal{M}_1$, 1 for $\mathcal{M}_2$). The true index is 1.

criteria were evaluated at that sample size. The average selected model index is given in Figure 2.

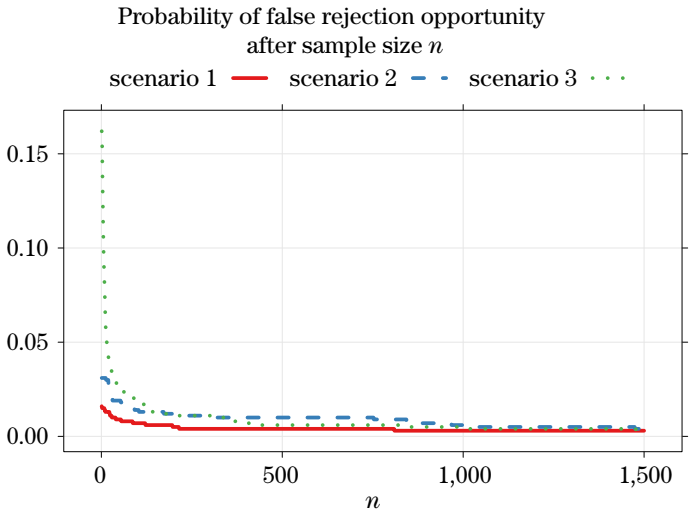

Figure 3. $N = 1{,}000$ data sets of length $n_{\max} = 10{,}000$ in each scenario, from the simple model. The complex model was selected when $\delta_{\mathrm{sw}}(x^n) > 20$. Estimated probability that there exists a model index after n at which the complex model is selected. Results shown up to $n = 1{,}500$, after which the three curves are indistinguishable.

The results are as expected. When the complex model is true, AIC is most likely to select it, at the cost of inconsistency when the simple model is true. BIC is the slowest to correctly select the complex model and the first to correctly select the simple model. HQ and δ_{sw} show intermediate behaviour, HQ being slightly more likely to select the complex model.

To illustrate strong consistency and optional stopping, three scenarios were considered: $\mathcal{M}_0$ vs $\mathcal{M}_1$, data from a standard normal distribution, "scenario 1", switching defines a test that is robust with respect to optional stopping; $\mathcal{M}_1$ vs $\mathcal{M}_2$, data from a standard normal distribution, "scenario 2"; $\mathcal{M}_1$ vs $\mathcal{M}_2$, data from a normal distribution with mean 35 and variance 1, "scenario 3".

We created $N = 1{,}000$ data sets of length $n_{\max} = 10{,}000$ in each scenario. We selected the complex model when δ_{sw} was larger than 20 corresponding to a significance level of 0.05. We estimated two probabilities at each sample size n: the probability that there is a model index after n at which the complex model will be selected (Figure 3), approximated by checking whether the complex model is selected at any sample size between n and $3n_{\max}$; the probability that there is a model index before n at which the complex model isselected (Figure 4).

Figure 3 can be interpreted as a check on strong consistency, whether the probabilities converge to 0 as $n \to \infty$. van Erven, Grünwald and de Rooij's

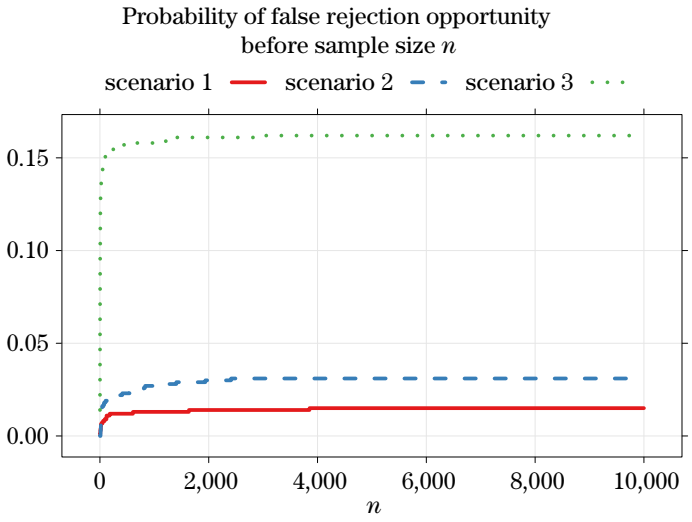

Figure 4. The setting of Figure 3. Estimated probability that there exists a model index before n at which the complex model is selected.

(2007) theorem implies strong consistency in all three scenarios, and the graph confirms this. The graph also illustrates that strong consistency is an asymptotic, nonuniform version of robustness to optional stopping — from some sample size on, one does not falsely reject no matter how long one keeps sampling.

Figure 4 refers to nonasymptotic optional stopping: in scenario 1, the conditions from Theorem 2 hold, and the figure shows that the probability that the complex model is *ever* incorrectly selected even when optional stopping is used, is bounded by 0.05 (the observed bound is 0.015). In scenarios 2 and 3, the conditions from Theorem 2 do not hold. In scenario 2, the behaviour of the switch criterion is similar to that of scenario 1. In scenario 3, the probability of a false rejection opportunity before sample size n goes to 0.15, and δ_{sw} is not robust to optional stopping.

When the simplest model is not a singleton, the choice of prior on the model parameters (in scenarios 2 and 3 on μ in $\mathcal{M}_1$ and on (μ, σ^2) in $\mathcal{M}_2$) affects the results. In both scenarios δ_{sw} still satisfies the weak, prior-expected version of robustness (5.3). In scenario 2, the prior is centered at the data-generating value of zero and suggests robustness. In scenario 3, the prior is centered at zero while the data is generated with a mean of 35, 3.5 standard deviations away from the prior mean and, as the figure shows, nonasymptotic robustness is violated.

6. Discussion and Future Work

We highlight three issues which, we feel, need additional discussion: consistency; whether there is anything 'special' to the switch criterion as opposed to other possible trade-offs between risk optimality and consistency; the limitations of switching in its current form.

Consistency Following Box's maxim 'Essentially, all models are wrong, but some are useful' (Box and Draper (1987)), some consider the goal of model selection is not to select a non-existing 'true' model, but to obtain the best predictive inference or best inference about a parameter (Burnham and Anderson (2004); Forster (2000)). Another issue with consistency is that it is impossible to give a bound on the probability under $\mathbb{P}_\mu$ of selecting the wrong model at sample size n that converges to 0 uniformly for all $\mu \in M$. This nonuniformity implies that consistency is of little practical consequence for post-model selection inference (Leeb and Pötscher (2005)).

In fact there do exist situations in which a model can be correct, for example in the field of extrasensory perception (Bem (2011)), and in the area of genetic linkage (Gusella et al. (1983); Tsui et al. (1985)). While consistency is not a sufficient condition for being useful in practice, it can be desirable, for example in determining whether a certain structural relationship (e.g. dependence between variables) holds or not.

We consider studying model selection methods in terms of a finite-sample analogue. The *practical* importance of our work, is mostly that model comparison by switching defines, like Bayes, a robust null hypothesis test — providing Type-I errors irrespective of the stopping rule and with better Type-II error behaviour. We have shown robustness for singleton $\mathcal{M}_0$, however, and *the* major goal for future work is to come up with methods that are robust to optional stopping under composite $\mathcal{M}_0$.

How special is the switch distribution? Since Yang proved that in general, the conflict between consistency and risk-optimality is not resolvable, one might argue that any model selection rule just picks some position in the spectrum of behaviours of consistency vs. risk-optimality. Switching and HQ do take a special place in the consistency vs. risk-optimality spectrum as obtaining the fastest rates compatible with strong consistency, which may be viewed as asymptotic robustness to optional stopping. The switch distribution takes a special place in terms of its nonasymptotic robustness to optional stopping in that the Law of the Iterated Logarithm implies that any model comparison method that de-

fines a robust hypothesis test cannot achieve estimation rate better than order $(\log \log n)/n$. The main open question is then whether one can modify it so that robustness for composite $\mathcal{M}_0$ is achieved as well.

Future Work — Limitations of the Switch Distribution and Our Results To achieve full robustness to optional stopping with composite $\mathcal{M}_0$, some substantial changes to the switch distribution have to be made. Initial research suggests that such a modification of the switch distribution might be constructed based on techniques in Ramdas and Balsubramani (2015). This work is under development.

A limitation here is that our results are restricted to two nested exponential family models. It would be interesting to extend them to more than two models — highlighting the distinction between model selection and testing — and going beyond exponential families. It would be interesting to design an alternative, order-independent method that, like the switch distribution, is strongly consistent, near rate- and power-optimal, and is robust to optional stopping under composite $\mathcal{M}_0$.

Supplementary Materials

The online supplement contains the proofs of all theorems stated in this paper, and the relationship between the version of the switch criterion studied here, and the criterion introduced in van Erven, Grünwald and De Rooij (2012).

Acknowledgment

Our central result appeared in van der Pas (2013) for the special case $m_1 = 1$ and $m_0 = 0$, but the proof there contained an error. We are grateful to Tim van Erven for pointing this out to us. We are also thankful to the anonymous referees and to Hannes Leeb for raising the issue of whether the switch distribution has a 'special' place on the spectrum of a model selection criterion's possible risk and consistency behaviors. This research was supported by NWO VICI Project 639.073.04.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (Edited by B.N. Petrov and F.Csaki), 267–281. Akademiai Kaido, Budapest.

Andrews, M. and Baguley, B. (2012). Prior approval: the growth of Bayesian methods in psychology. *Br. J. Math. Stat. Psychol.*, **66**, 1–7.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley.

Barron, A., Rissanen, J. and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory* **44**, 2743–2760.

Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 301–413.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *J. Pers. Soc. Psychol.* **100**, 407–425.

Berger, J.O. and Wolpert, R.L. (1988). *The Likelihood Principle*. 2nd Edition. Institute of Mathematical Statistics.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Socio. Meth. Res.* **33**, 261–304.

Cavanaugh, J. E. (2012). [Catching up faster by switching sooner]: Discussion. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74**, 402–403.

Dawid, A. P. (1984). Present position and potential developments: Some personal views, statistical theory, the prequential approach. *J. R. Stat. Soc. Ser. A Stat. Soc.* **147** 278–292.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspect. Psychol. Sci.* **6**, 274–290.

Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* **70**, 193–242.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Am. Stat. Assoc.* **88**, 461–470.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *J. Math. Psychol.* **44**, 205–231.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. The MIT Press.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y., Young, A. B., Shoulson, I., Bonilla, E. and Martin, J. B. (1983). A polymorphic DNA marker genetically linked to huntington's disease. *Nature* **308**, 234–238.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **41**, 190–195.

John, L. K., Loewenstein, G. and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23**, 524–532.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795.

Lauritzen, S. (2012). [Catching up faster by switching sooner]: Discussion. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74**, 401–402.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric. Theory* **21**, 21–59.

Lhéritier, A. and Cazals, F. (2015). A sequential nonparametric two-sample test. Technical Report Research Report 8704, INRIA, Sophia Antipolis.

Pratt, J. W. (1962). On the foundations of statistical inference: Discussion. *J. Am. Stat. Assoc.*

57, 307–326.

Ramdas, A. and Balsubramani, A. (2015). Sequential nonparametric testing with the law of the iterated logarithm. Technical Report abs/1506.03488, ArXiv/CoRR.

Sanborn, A. N. and Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychon. Bull. Rev.* **21**, 283–300.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464.

Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011). Test martingales, Bayes factors and p-values. *Stat. Sci.* **26**, 84–101.

Shao, J. (1997). An asymptotic theory for linear model selection. *Stat. Sin.* **7**, 221–264.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **39**, 44–47.

Tsui, L.-C., Buchwald, M., Barker, D., Braman, J. C., Knowlton, R., Schumm, J. W., Eiberg, H., Mohr, J., Kennedy, D., Plavsic, N., Zsiga, M., Markiewicz, D., Akots, G., Brown, V., Helms, C., Gravius, T., Parker, C., Rediker, K. and Donis-Keller, H. (1985). Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* **230**, 1054–1057.

van der Pas, S. L. (2013). Almost the best of three worlds. The switch model selection criterion for single-parameter exponential families. Master's thesis, Leiden University.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van Erven, T., Grünwald, P. D. and de Rooij, S. (2007). Catching up faster in Bayesian model selection and model averaging. In *Advances in Neural Information Processing Systems 20*, 417–424. Cambridge, MA: MIT Press.

van Erven, T., Grünwald, P. D. and de Rooij, S. (2012). Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74**, 361–417.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p-values. *Psychon. Bull. Rev.* **14**, 779–804.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92**, 937–950.

Mathematical Institute, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands.
LUMC, Einthovenweg 20, 2333 ZC Leiden, The Netherlands.

E-mail: svdpas@math.leidenuniv.nl

Mathematical Institute, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands.
CWI, Science Park 123, 1098 XG, Amsterdam, The Netherlands.

E-mail: peter.grunwald@cwi.nl

(Received January 2016; accepted December 2016)