AutoMS: Automatic Model Selection for Novelty Detection with Error Rate Control

Yifan Zhang[1*] **Haiyan Jiang**[2,3†] **Haojie Ren**[4†] **Changliang Zou**[1] **Dejing Dou**[3]
[1] School of Statistics and Data Sciences, LPMC, KLMDASR and LEBPS, Nankai University, Tianjin, China
[2] Machine Learning Department, MBZUAI, Abu Dhabi, UAE
[3] Baidu Research, Baidu Inc., Beijing, China
[4] School of Mathematical Science, Shanghai Jiao Tong University, Shanghai, China
yfzhang_stat@mail.nankai.edu.cn, haiyan.jiang@mbzuai.ac.ae
haojieren@sjtu.edu.cn, zoucl@nankai.edu.cn, doudejing@baidu.com

Abstract

Given an unsupervised novelty detection task on a new dataset, how can we automatically select a "best" detection model while simultaneously controlling the error rate of the best model? For novelty detection analysis, numerous detectors have been proposed to detect outliers on a new unseen dataset based on a score function trained on available clean data. However, due to the absence of labeled data for model evaluation and comparison, there is a lack of systematic approaches that are able to select the "best" model/detector (i.e., the algorithm as well as its hyperparameters) and achieve certain error rate control simultaneously. In this paper, we introduce a unified data-driven procedure to address this issue. The key idea is to maximize the number of detected outliers while controlling the false discovery rate (FDR) with the help of Jackknife prediction. We establish non-asymptotic bounds for the false discovery proportions and show that the proposed procedure yields valid FDR control under some mild conditions. Numerical experiments on both synthetic and real data validate the theoretical results and demonstrate the effectiveness of our proposed AutoMS method. The code is available at: https://github.com/ZhangYifan1996/AutoMS.

1 Introduction

With the development of big data, novelty detection has become critical for many machine learning applications, such as network intrusion detection [1], fraud detection in finance [2, 3], event detection in the earth sciences [4], and so on. For these applications, there are a very large number of "normal" instances (i.e., inlier) but a lack of "abnormal" data (i.e., novelty, or outlier), because collecting abnormal data is much more expensive or difficult than collecting normal data [5]. This leads us to a situation where the available historical dataset contains only normal observations, and the basic goal is to detect outliers on a new unseen test dataset. Novelty detection has achieved remarkable success in many application domains with numerous methods/algorithms being developed in machine learning literature [6–8], however, two problems need to be addressed in the current novelty detection models, (1) lack of error rate control in detection algorithms, and (2) ad-hoc model selection of algorithms/hyperparameters.

The most common way to solve the novelty detection problem is to use a detection algorithm to learn a plausible score function that assigns a score to each point in the test dataset. Generally, data with large scores are more likely to be detected as outliers. A decision threshold is given to determine

[*] Y. Zhang did this work during her internship at Baidu Research.
[†] Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

whether a data point is an outlier or not. There is an obvious trade-off between precision and recall: a small threshold would detect more outliers but make more false discoveries as well. Therefore, a more viable way is to reformulate the novelty detection problem as a multiple hypothesis testing problem, and adopt some error rate control procedure as suitable criteria to choose the threshold [9].

Model selection (including algorithm selection or hyperparameter tuning) is a well-known feature of these novelty detection algorithms and arises naturally since none of these methods can universally outperform others on *all* tasks as noted in Wolpert and Macready [10], Zhao et al. [11]. Due to the special setting of novelty detection, model evaluation and model comparison are not feasible when only "clean" data are provided as training dataset. Those strategies are usually ad-hoc [11, 12], which may greatly hamper their applicability to novelty detection in general cases. Furthermore, these methods only target the recall through model selection without considering the false discoveries. Therefore, it is desirable to develop a unified framework to select the "best" detector with certain error rate control.

We see that designing a data-driven procedure to select the "best" model/detector while simultaneously achieving some error rate control remains an open challenge. In this paper, we propose a data-driven procedure, AutoMS, to implement automatic model selection for novelty detection with error rate control. We apply the Jackknife technique [13] to fully explore the clean data and avoid the randomness caused by data-splitting. With the help of Jackknife prediction, we select from a pool of diversified models/detectors the "best" detector that maximizes the number of detected outliers while controlling the false discovery rate. The theoretical guarantees for the FDR control of the selected "best" detector requires further investigation because the error distribution of the "best" detector is different from that of a given detector. We address this issue by carefully studying the bound of false discovery proportion (FDP) and FDR which is calculated by taking expectation of FDP.

To our best knowledge, this is the first work to systematically select the "best" model while achieving certain error rate control. The main contributions of this work are summarized as follows.
(1) The proposed AutoMS can select the best model and simultaneously control the error rate of the best model. Notably, AutoMS is a unified data-driven framework, and it does not rely on any labeled anomalous data for model selection.
(2) To our best knowledge, it is the first effort to select the "best" model/detector with theoretical guarantees in the view of FDR control. We establish non-asymptotic bounds for the FDP and show that the proposed AutoMS yields valid FDR control.
(3) The AutoMS can be easily coupled with commonly used novelty detection algorithms. Extensive numerical experiments indicate that AutoMS outperforms other methods significantly, with respect to both error rate control and detection power.

2 Related Work

FDR Control. Reformulating the outlier detection problem as a multiple testing problem, a commonly-used error rate in this setting is the false discovery rate (FDR): the expected proportion of true inliers among the test points reported as outliers. The FDR control [14–16] has been fully studied in the statistical literature and is a particularly useful tool to maintain the ability to reliably detect true alternatives without excessive false discovery results when large-scale hypotheses are simultaneously tested.

Along this line, Bates et al. [17] pioneered a method to split the clean dataset into two independent parts, one for training the score function and one for validating the empirical cumulative distribution function of the score function which will be used for constructing p-values for the new observations. They showed that the FDR control can be achieved with theoretical finite sample guarantees. It involves, however, a one-time random split of the clean dataset, which leads to a "p-value lottery" problem. That is, those p-values based on sample-splitting are sensitive and difficult to be reproduced in practice. Moreover, the procedure in Bates et al. [17] can only control the FDR for a given novelty detection model, and no model efficiency is considered in their paper.

Model Selection in Novelty Detection. As described in Sect. 1, various detection algorithms have been developed for different problems or different types of data, such as LOF [18], one-class SVM [19], iForest [20] and so on. More discussions can be found in Pimentel et al. [6], Aggarwal [21]. However, no universal learning model performs well on all problem instances Wolpert and Macready [10]. Therefore, it raises the problem of model selection, including algorithm selection

and hyperparameter tuning. For example, AutoOD [12] proposed an automated outlier detection framework focusing on automatic neural architecture search. Recent work METAOD [11] built a measure for task similarity and use meta-learning to select the most similar task compared to the historical ones. Most existing work is usually based on sophisticated black-box machine learning algorithms to learn the space of the normal data. While generally effective in practice, they rarely offer a clear guarantee about the output.

We note that some statistical criteria can be used for model selection in novelty detection; e.g. using mass-volume and excess-mass [22] or some clustering validation metrics [23] to measure the quality of a score function, and using IREOS [24] to assess the separability. However, Ma et al. [25] indicated that those criteria are not useful in practice and cannot lead to a universal detector for general cases. Our proposed AutoMS is not limited to one specific algorithm or model but can select the "best" one among any set of models/detectors.

3 Methodology: Automatic Model Selection

In novelty detection problems, we are given a *clean* training dataset of inliers $\mathcal{D} = \{\tilde{Z}_j\}_{j=1}^m$ containing m independently and identically distributed (i.i.d.) observations $\tilde{Z}_j \sim \mathbb{P}_0$. The goal of the novelty detection problem is to identify the outlier set $\mathcal{O} = \{Z_i \in \mathcal{U} : Z_i \text{ is } \textit{not} \text{ drawn from } \mathbb{P}_0\}$ among a new unseen test dataset $\mathcal{U} = \{Z_i\}_{i=1}^n$.

One of the most common approaches in addressing novelty detection is to employ a detection algorithm $\mathcal{M}$ to learn a plausible score function $S_{\mathcal{M}}(\cdot)$ with the clean dataset $\mathcal{D}$ and use the score $S_{\mathcal{M}}(Z_i)$ to assess whether $Z_i \in \mathcal{U}$ is an outlier. Generally, it is assumed that the larger the score $S_{\mathcal{M}}(Z_i)$, the more likely Z_i is an outlier. Hence, the decision set (detected outlier set) is selected as: $\widehat{\mathcal{O}}_{\mathcal{M}} = \{Z_i \in \mathcal{U} : S_{\mathcal{M}}(Z_i) \geq L_{\mathcal{M}}\}$, where $L_{\mathcal{M}}$ is a threshold associated with detector $\mathcal{M}$. Denote the subset of $\mathcal{U}$ containing all inliers as $\mathcal{O}^c = \{Z_i \in \mathcal{U} : Z_i \text{ is drawn from } \mathbb{P}_0\}$ with size n_0, and $\mathcal{O} \bigcup \mathcal{O}^c = \mathcal{U}$. Let $|\mathcal{A}|$ denote the number of the elements in set $\mathcal{A}$. Denote $a \vee b = \max\{a, b\}$.

Taking a multiple testing perspective, the null hypothesis is that a new observation is an inlier, and the alternative asserts it is an outlier. That is, the novelty detection problem is translated into a multiple testing problem as follows,

$$\mathbb{H}_{0i} : Z_i \sim \mathbb{P}_0 \text{ v.s. } \mathbb{H}_{1i} : Z_i \nsim \mathbb{P}_0 \quad i = 1, \cdots, n. \tag{1}$$

3.1 Problem Definition

Problem (Model Selection of Unsupervised Novelty Detection) The model selection problem for unsupervised novelty detection problem can be stated as follows.

Given a training dataset of inliers $\mathcal{D} = \{\tilde{Z}_j\}_{j=1}^m$, a new unlabeled dataset $\mathcal{U} = \{Z_i\}_{i=1}^n$ which may contain outliers, and all models $\mathcal{G} = \bigcup \mathcal{M}$;

Select the best model $\mathcal{M}^* \in \mathcal{G}$, *such that* the FDR of $\mathcal{M}^*$ is controlled under the target level α.

The false discovery proportion (FDP) and true discovery proportion (TDP) are defined as follows

$$\text{FDP} = \frac{|\widehat{\mathcal{O}}_{\mathcal{M}} \bigcap \mathcal{O}^c|}{1 \vee |\widehat{\mathcal{O}}_{\mathcal{M}}|}, \quad \text{TDP} = \frac{|\widehat{\mathcal{O}}_{\mathcal{M}} \bigcap \mathcal{O}|}{|\mathcal{O}|}. \tag{2}$$

The false discovery rate (FDR) is the expectation of FDP, i.e., $\text{FDR} = \mathbb{E}(\text{FDP})$. The true detection rate (TDR) is defined as the expectation of TDP, i.e., $\text{TDR} = \mathbb{E}(\text{TDP})$. A detector is superior to another if the former has a larger TDR for the same FDR level.

3.2 Construction of AutoMS

We begin with the FDR control for a given detector $\mathcal{M}$ and then give the AutoMS of how to select the "best" detector from $\mathcal{G}$. For a detector $\mathcal{M}$, we learn a score function $S_{\mathcal{M}}(\cdot)$ based on the historical clean dataset $\mathcal{D}$ and evaluate the score $S_{\mathcal{M}}(Z_i)$ at the new data point $Z_i \in \mathcal{U}$. The canonical approach to FDR control in multiple testing problem is to apply the Benjamini-Hochberg (BH) procedure [14–16] to the p-values of scores $S_{\mathcal{M}}(Z_i)$ for any $Z_i \in \mathcal{U}$. That is, the data-dependent threshold L is

selected as follows,

$$L = \inf\left\{t > 0 : \frac{nG_{\mathcal{M}}(t)}{|\{Z_i \in \mathcal{U} : S_{\mathcal{M}}(Z_i) \geq t\}| \vee 1} \leq \alpha\right\},$$

where $G_{\mathcal{M}}(t) = \mathbb{P}_{\mathbb{H}_0}(S_{\mathcal{M}}(Z_i) \geq t|\mathcal{D})$ is the p-value of $S_{\mathcal{M}}(Z_i)$ for $Z_i \in \mathcal{U}$ when $\mathbb{H}_0$ holds.

However, it is quite a challenge to estimate reliable p-values $G_{\mathcal{M}}(t)$ since $\mathbb{P}_0$ is unknown and the score function $S_{\mathcal{M}}(\cdot)$ may be derived from a sophisticated black-box machine learning algorithm. Bates et al. [17] suggested randomly splitting $\mathcal{D}$ into two disjoint parts with equal sample size, one for training the score function $S_{\mathcal{M}}(\cdot)$ and the other for estimating $G_{\mathcal{M}}(t)$ with the help of conformal inference. This single-random-splitting based method suffers from, (1) the estimation of the p-values depends on the random split; (2) different split ratios also lead to different results.

To yield more stable estimates of p-values in practice, we propose to learn the score function $S_{\mathcal{M}}(\cdot)$ on the whole clean dataset $\mathcal{D}$, where our key idea stems from the **Jackknife estimate**. Denote $\mathcal{D}^{[-j]} = \mathcal{D}\backslash\{\tilde{Z}_j\}$ as the subset of $\mathcal{D}$ with the j-th observation removed. Define $S_{\mathcal{M}}^{[-j]}(\cdot)$ as the score function trained on $\mathcal{D}^{[-j]}$. Then we can evaluate m scores $\{S_{\mathcal{M}}^{[-j]}(\tilde{Z}_j)\}_{j=1}^m$ on the whole clean dataset $\mathcal{D}$. The p-value function can be estimated by

$$G_{m,\mathcal{M}}(t) = \frac{|\{\tilde{Z}_j \in \mathcal{D} : S_{\mathcal{M}}^{[-j]}(\tilde{Z}_j) \geq t\}|}{m}. \tag{3}$$

The Jackknife technique provides an approximation of the score, $S_{\mathcal{M}}^{[-j]}(\tilde{Z}_j)$, for any $\tilde{Z}_j \in \mathcal{D}$. Under the null $\mathbb{H}_0$ that $Z_i \in \mathcal{U}$ is from $\mathbb{P}_0$, $S_{\mathcal{M}}^{[-j]}(\tilde{Z}_j)$ can be used for estimating the empirical distribution of $S_{\mathcal{M}}(Z_i)$. And the corresponding threshold $L_{\mathcal{M}}$ is determined via

$$L_{\mathcal{M}} = \inf\left\{0 < t < \bar{t}_{m,\mathcal{M}} : \frac{nG_{m,\mathcal{M}}(t)}{|\{Z_i \in \mathcal{U} : S_{\mathcal{M}}(Z_i) \geq t\}| \vee 1} \leq \alpha\right\}, \tag{4}$$

where $\bar{t}_{m,\mathcal{M}}$ is the upper bound for t and its choice will be discussed in Sect. 4 in detail, and $S_{\mathcal{M}}(\cdot)$ is the same score function trained on $\mathcal{D}$. Then, the detected set is identified as

$$\widehat{\mathcal{O}}_{\mathcal{M}} = \{Z_i \in \mathcal{U} : S_{\mathcal{M}}(Z_i) \geq L_{\mathcal{M}}\}. \tag{5}$$

Further, we select the "best" detector $\mathcal{M}^*$ via

$$\mathcal{M}^* = \arg\max_{\mathcal{M}\in\mathcal{G}} |\widehat{\mathcal{O}}_{\mathcal{M}}|. \tag{6}$$

Algorithm 1 AutoMS: Automatic Model Selection

1: **Input:** New dataset $\mathcal{U}$, clean dataset $\mathcal{D}$, target FDR level α and a pool of detectors $\mathcal{G}$.
2: **for** $\mathcal{M} \in \mathcal{G}$ **do**
3: **for** $j = 1, \cdots, m$ **do**
4: Train score function $S_{\mathcal{M}}^{[-j]}(\cdot)$ based on $\mathcal{D}^{[-j]}$ (i.e. $\mathcal{D}^{[-j]} = \mathcal{D}\backslash\{\tilde{Z}_j\}$);
5: Compute the score at $\tilde{Z}_j \in \mathcal{D}$, i.e., $S_{\mathcal{M}}^{[-j]}(\tilde{Z}_j)$;
6: **end for**
7: Learn score function $S_{\mathcal{M}}(\cdot)$ based on $\mathcal{D}$;
8: Compute the scores $S_{\mathcal{M}}(Z_i)$ for any $Z_i \in \mathcal{U}$, and find a threshold $L_{\mathcal{M}}$ by Eq. (4);
9: The detected set is $\widehat{\mathcal{O}}_{\mathcal{M}} = \{Z_i \in \mathcal{U} : S_{\mathcal{M}}(Z_i) \geq L_{\mathcal{M}}\}$.
10: **end for**
11: Select the "best" detector $\mathcal{M}^*$ from $\mathcal{G}$ via Eq. (6).
12: **Input:** $\widehat{\mathcal{O}}_{\mathcal{M}^*} = \{Z_i \in \mathcal{U} : S_{\mathcal{M}^*}(Z_i) \geq L_{\mathcal{M}^*}\}$.

We refer to our procedure as AutoMS short for *Automatic Model Selection* and it is summarized in Algorithm 1. In fact, our AutoMS consists of two phases: offline training in lines 2-7 and online prediction in lines 8-9 of pseudo-code in Algorithm 1. Note that the running time of offline training is not critical. Meanwhile, the additional running time is the detector selection in line 11, which should incur a small run-time overhead due to only maximizing a finite set.

Note that the pool of detectors $\mathcal{G} = \bigcup \mathcal{M}$ can be a set of different algorithms or a set of hyperparameters for one typical algorithm, or both, which fits the definition of model selection including

algorithm selection and hyperparameter tuning. Note that the size of $\mathcal{G}$, $|\mathcal{G}|$, can be infinite, and $|\mathcal{G}|$ is allowed to increase as the size of $\mathcal{D}$. More details will be discussed in Sect. 4.

Rationale. Here is the intuition why the selected detector obtained by AutoMS is the "best" one. The general idea is as follows. A good detector needs to achieve a large TDR while controlling the false discoveries under the target FDR level α. **With FDR control**, the selected detector which has the largest number of discoveries is roughly the one with the largest TDR. Otherwise, simply selecting the largest number of discoveries **without FDR control**, is not even a correct criterion, because in this case, most detected discoveries may be inliers and true outliers may not be detected. For every detector $\mathcal{M} \in \mathcal{G}$, with FDR control, its FDP can be approximately controlled around the target FDR level α, as the proposed AutoMS resets the corresponding novelty threshold for every $\mathcal{M}$. Therefore, there are roughly $|\widehat{\mathcal{O}}_{\mathcal{M}}|(1-\alpha)$ true outliers in the detected set $\widehat{\mathcal{O}}_{\mathcal{M}}$, with the TDR being approximated by $\frac{|\widehat{\mathcal{O}}_{\mathcal{M}}|(1-\alpha)}{|\mathcal{O}|}$. Since the number of true outliers $|\mathcal{O}|$ and the target FDR level α are fixed in a given novelty detection task, selecting the largest $\frac{|\widehat{\mathcal{O}}_{\mathcal{M}^*}|(1-\alpha)}{|\mathcal{O}|}$, namely the largest $|\widehat{\mathcal{O}}_{\mathcal{M}^*}|$, leads to the "best" detector $\mathcal{M}^*$.

Why Jackknife not Jackknife+. Barber et al. [26] introduced the Jackknife+ as a new method for constructing predictive confidence intervals. They provide non-asymptotic coverage guarantees under no assumptions beyond the training and test dataset being exchangeable and also study the theoretical guarantees for the original Jackknife with stability assumptions (see Section 5 of their paper). Jackknife+ needs to calculate m leave-one-out predictions for one test point while the original Jackknife only calculates one predicted value [26]. The model evaluation cost (counts the number of times evaluating a fitted model on a new point) of Jackknife+ is mn whereas the original Jackknife is of cost n, indicating that Jackknife+ has a much higher computation complexity compared to the original. Thus, since most novelty detection algorithms satisfy the stability assumption, we are more inclined to apply the original Jackknife technique rather than Jackknife+ to detectors that meet stability assumption, and we also give theoretical guarantees to achieve the non-asymptotic bounds of FDP and maintain FDR control.

4 Statistical Performance Guarantees

In this section, we will theoretically establish NON-asymptotic bounds for FDP of the selected detector $\mathcal{M}^*$ by AutoMS and further show that it can control the FDR under the following mild assumptions. Define the number of detectors $\varpi_m = |\mathcal{G}|$ and let $A_n = o(n)$ be a pre-specified sequence.

Assumption 1 (Density)**.** For any detector $\mathcal{M} \in \mathcal{G}$ and sufficiently large m, the conditional density of $S_{\mathcal{M}}(Z)$ given training dataset $\mathcal{D}$ is uniformly bounded above by a universal constant $c_f > 0$, where $Z \in \mathcal{U}$ is a new point.

Assumption 2 (Learning stability)**.** For any fixed $\mathcal{M}$, the score $S_{\mathcal{M}}$ satisfies: for large enough m,

$$\mathbb{P}(|S_{\mathcal{M}}(Z) - \tilde{\mu}_{\mathcal{M}}(Z)| \leq \kappa_m) \geq 1 - \zeta_m,$$

with some sequences satisfying $\kappa_m = o(1)$ and $\zeta_m = o(1)$ as $m \to \infty$, and some function $\tilde{\mu}_{\mathcal{M}}$.

Assumption 3 (Rates)**.** Denote $B_m = 2c_f\kappa_m + 3\zeta_m$. Suppose that $B_m = o(1)$, $n\varpi_m B_m = o(1)$ and $m\varpi_m/A_n^{2/3} = o(1)$.

Our main theoretical result on the validity of the AutoMS method for both FDP and FDR control is given by the next theorem.

Theorem 1 (FDR control)**.** *Suppose Assumptions 1–3 hold. Let* $0 < \delta < 1$ *and* $0 < \alpha < 1$. *There exist universal constants* $C_1 > 0$ *and* $C_2 > 0$ *so that the FDP of the proposed method satisfies*

$$\mathrm{FDP}(L_{\mathcal{M}^*}) \leq \alpha \left[1 + \frac{4nB_m}{A_n} + C_1\sqrt{\frac{\varpi_m}{\delta A_n^{2/3}}} + C_2\sqrt{\frac{\varpi_m W_{mn}}{\delta}} \right], \tag{7}$$

with probability at least $1 - 2\delta$, *and*

$$\limsup_{(m,n)\to\infty} \mathrm{FDR}(L_{\mathcal{M}^*}) \leq \alpha, \tag{8}$$

where $W_{mn} = \left(\frac{n}{mA_n^{2/3}} + \frac{n}{A_n^{4/3}}\right)(1 + 15B_m + 2m^2B_m)$.

It is worth mentioning that even if all detectors control the FDR, we cannot assert without theoretical proof that the optimal detector $\mathcal{M}^*$ can still guarantee the FDR control. And Bates et al. [17] only gives **the expectation version** that FDR can be controlled for a given detector. Briefly, we first prove that the FDP of each detector can be bounded *uniformly* through the uniform convergence of the Jackknife estimator. Hence, the selected "best" detector $\mathcal{M}^*$ with the largest number of discoveries will have the same non-asymptotic bounds for FDP. Then we give the asymptotic FDR control for the optimal detector. Detailed proofs for theoretical results are provided in **Appendix E**.

In Eq. (4), we specify a data-driven upper bound $\bar{t}_{m,\mathcal{M}}$ when determining $L_{\mathcal{M}}$. We define $\bar{t}_{m,\mathcal{M}} = G^-_{m,\mathcal{M}}(A_n/n)$ where $G_{m,\mathcal{M}}(t)$ is defined in Eq. (3) and $G^-_{m,\mathcal{M}}(y) = \inf\{t \geq 0 : G_{m,\mathcal{M}}(t) \leq y\}$ with $0 \leq y \leq 1$. The upper bound $\bar{t}_{m,\mathcal{M}}$ for threshold $L_{\mathcal{M}}$ can facilitate the derivation of non-asymptotic bounds for FDP but is actually not required in practice. In other words, this makes an implicit assumption that a non-trivial (not ∞) threshold can be attained within the interval $(0, \bar{t}_{m,\mathcal{M}}]$. By Theorem 1, we can see that a large A_n would yield faster convergence of FDP. However, too large A_n may lead to a trivial threshold. In fact, how large A_n could be depends on the number of identifiable outliers in $\mathcal{U}$ [27], which is certainly unknown to us. Alternatively, by examining the requirements in Assumption 3, we can set $A_n \sim \min\{(m\varpi_m)^2, n^{1/2}\}$ for simplicity.

Remark 1 (Stability). Since we use the Jackknife to estimate FDP, certain stability conditions seem to be necessary. In Assumption 2, $\tilde{\mu}$ needs not to be close to an optimal population detector. We only need the $S_{\mathcal{M}}$ to concentrate around $\tilde{\mu}$. This is just a stability assumption rather than a consistency assumption. When the sampling stability fails to hold, the $S^{[-j]}_{j,\mathcal{M}}$ may have a substantially different distribution than $G_{n,\mathcal{M}}(t)$. In that case, we may resort to the sample-splitting procedure instead.

Remark 2 (Rates). In general, Assumption 3 is used to ensure the uniform consistency of the estimated p-values, which performs an important role in the convergence of the FDR control. Specifically, $A_n = o(n)$ is a pre-specified sequence which gives a non-trivial lower bound of the true p-value function $G_{\mathcal{M}}(t)$ with $G_{\mathcal{M}}(t) \geq \frac{A_n}{n} = o(1)$. The B_m is a quantity used to measure the difference of two p-values estimated in different ways, as in Lemma C.3 and Lemma C.2. And $B_m = o(1)$ indicates the two p-values should be close enough which is a very mild condition, and it can be expected that the p-values will be close if the training sample sizes of inliers m and test sample sizes of inliers n_0 are large enough in practice. In the proposed AutoMS, the number of detectors $\varpi_m = |\mathcal{G}|$ can go to infinity with some certain rate related to the training sample size m and test sample size n. The assumptions $n\varpi_m B_m = o(1)$ and $m\varpi_m/A_n^{2/3} = o(1)$ indicate the required rate of ϖ_m to ensure the uniform consistency among detectors.

Remark 3 (FDP bound). Although the FDP bound might be slightly larger than the target FDR level α in finite samples, the FDP bound converges asymptotically to the target FDR level α. Actually, in novelty detection tasks, it is not necessary to obtain procedures of finite FDR control with additional assumptions, instead, a powerful and stable procedure which controls FDR asymptotically is good enough for most cases.

5 Experiments and Evaluation

5.1 Experiment Setting

In this section, we evaluate the performance of the proposed AutoMS in terms of FDR control and detection power (TDP) through extensive numerical studies on both synthetic and real data. We report the results based on 100 replication experiments throughout this paper. We construct a detector set $\mathcal{G}$ with 83 candidate models by combining 6 novelty detection algorithms and their corresponding hyperparameters (See Table S1 in **Appendix A** for the complete list). These models/detectors are available in the Python "**PyOD**" [28] package. Note that the size of detector set $\mathcal{G}$ can be infinite and our AutoMS can still be used.

Performance Measures. Since we focus on both error rate control and detection power, we use $\mathrm{FDR} = \mathbb{E}(\mathrm{FDP})$ and $\mathrm{TDR} = \mathbb{E}(\mathrm{TDP})$ as performance measures for each model in this paper. According to the definitions in Eq. (2) of Sect. 3, we have $\mathrm{FDP} = 1 - \mathrm{Precision}$, and $\mathrm{TDP} = \mathrm{Recall}$. Correspondingly, the empirical FDR and empirical TDR are calculated by averaging of FDPs and TDPs over 100 replications, respectively. We say a detection procedure is superior to its competitor if

it has a larger TDR for the same target FDR level α, that is, a superior detector has a larger empirical TDR whose empirical FDR is below or around the target FDR level α. The preset target FDR level α should be an acceptable level, which means all detectors that control the FDR at this level α should be accepted. As long as the FDR of the selected detector is below a preset level α, we have achieved our goal on FDR control and we should focus on improving TDR.

Baseline methods. To evaluate the performance of the proposed **AutoMS** method, we compare the proposed AutoMS with two different types of baselines: (i) model selection methods (METAOD [11]) and (ii) error-rate-controlled methods (SRS-based methods [17]). And we also consider two versions of the proposed AutoMS: **AutoMS-JK** and **AutoMS-SRS**, where AutoMS-JK uses Jackknife technique and AutoMS-SRS replaces the Jackknife technique with single-random-splitting (SRS) technique for estimating score. We explore the performances of the four methods with respect to FDR and TDR. It is worth mentioning that **SRS** is guaranteed with FDR control for any given detector, but there is no model selection considered in SRS-based methods. **METAOD** is a model selection method using meta-learning without considering error rate control, while the **AutoMS** can do both model selection and error rate control, as shown in Table 1. For more details on the baseline methods, Refer to **Appendix B**.

Table 1: Model-selection and error-rate-control methods.

Method	Model Selection	Error Rate Control
SRS [17]	✗	✓
METAOD [11]	✓	✗
AutoMS-SRS	✓	✓
AutoMS-JK	✓	✓

5.2 Evaluation of the Jackknife

Firstly, We would like to show the advantage of Jackknife prediction with a toy example using synthetic data. We generate a training dataset $\mathcal{D}$ and a test dataset $\mathcal{U}$ from a Gaussian mixture model, where their sample sizes are $m = n = 2000$ respectively and $\mathcal{U}$ contains 10% outliers. More details of the setting can be found in Sect. 5.3. In Fig. 1, we compare the Jackknife prediction to the single-random-splitting (SRS) method in Bates et al. [17] with different split ratios when the detector $\mathcal{M}$ is the OCSVM with its default hyperparameter in the PyOD package [28]. Focusing on one training set $\mathcal{D}$, we randomly split $\mathcal{D}$ 100 times for each split ratio of SRS method while Jackknife technique only has one prediction result. We can see from Fig. 1 that the Jackknife provides a quite tight FDP control and achieves a high TDP while finding more true outliers. As discussed above, the SRS-based methods depend on the random split and the split ratio and are usually unstable due to the randomness of the segmentation, making it hard to reproduce.

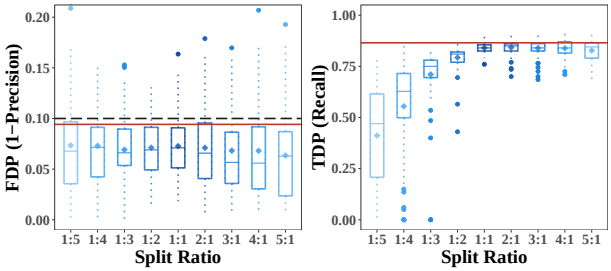

Figure 1: Performance of SRS-OCSVM [17] under different split ratios. Each box-plot shows the distribution of FDP and TDP under 100 simulations. The red line shows the corresponding results from Jackknife prediction. The black dashed line is the target FDR level $\alpha = 0.1$.

The Jackknife estimate helps improve the stability and the accuracy of the scores due to the usage of more samples to train the score function. Meanwhile, the Jackknife procedure will require a bit more computation compared to one-time data splitting, if m is not too small. As the Jackknife procedure for $S_{\mathcal{M}}^{[-j]}(\cdot)$ is trained offline, one can always replace the Jackknife with cross-validation to reduce the computing burden to a certain extent. Instead of omitting one observation, we can randomly partition $\mathcal{D}$ into k equal-sized folds and remove one fold from $\mathcal{D}$ as the subsample to train the score function. In the following of this section, we use $k = 100$ folds in the AutoMS-JK procedure.

5.3 Synthetic Data

Data Description. We consider the same synthetic setting as in Bates et al. [17]. We will compare the performance of different methods (the proposed AutoMS-JK, AutoMS-SRS, SRS-based methods and METAOD) when the target FDR level is set to $\alpha = 0.1$. Then we will show the results under

different target FDR levels. The data are generated by sampling each data point $Z_i \in \mathbb{R}^{50}$ from a multivariate Gaussian mixture model, $Z_i = W_i + \sqrt{a}V_i$, where W_i is one of 50 cluster centers sampled independently from the uniform distribution on $[-3, 3]^{50}$ and $V_i \in \mathbb{R}^{50}$ are i.i.d from $N(0, 1)$. The difference between inliers and outliers is the constant a with $\mathbb{H}_0 : a = 1$ v.s. $\mathbb{H}_1 : a = 2$. We generate a clean dataset $\mathcal{D}$ with $m = 2000$ observations and a test dataset $\mathcal{U}$ with $n = 2000$ observations in which 90% are sampled as inliers. We aim to identify the remaining 10% of outliers in $\mathcal{U}$.

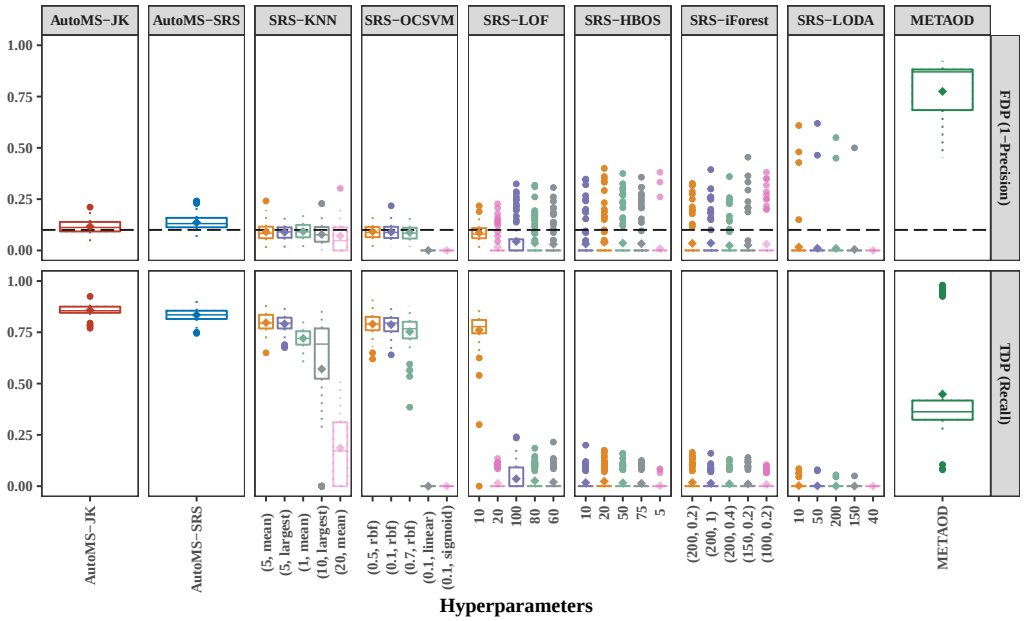

Figure 2: Comparisons of FDP (more stable around the dashed line) and TDP (the larger the better) among AutoMS-JK, AutoMS-SRS, SRS-based methods (with different hyperparameters) and METAOD. The dashed line is the target FDR level α which is set to $\alpha = 0.1$.

Results. With the given target FDR level $\alpha = 0.1$, the AutoMS (either AutoMS-JK or AutoMS-SRS) outperforms all baseline methods as shown in Fig. 2. (a) All the SRS-based methods can perfectly control the FDR with FDPs scattering around the given target FDR level α. But in terms of detection power, the TDP performance of SRS-based methods vary greatly with different hyperparameters and different algorithms. Some SRS-based methods, such as SRS-LODA, SRS-HOBOS, SRS-iForest, even lose detection power with almost zero TDP/recall. This necessitates model selection among numerous novelty detection methods. (b) On the other hand, METAOD gives large FDPs, which means that most of the reported points are not true outliers. This phenomenon can be explained as METAOD does not employ any FDR control procedure. And the TDPs of METAOD do not perform well because the model selection of METAOD is based on historical data which may not be suitable for the data we use. (c) AutoMS-JK and AutoMS-SRS are superior to all SRS-based methods and METAOD with respect to the average TDR, with the average FDR slightly above the target FDR level α, as it is more difficult to control the FDR of the selected detector than given detector. Additionally, AutoMS-JK gives a higher TDR than AutoMS-SRS, which indicates that using the Jackknife method instead of SRS enhances detection power. Although both AutoMS and SRS-based methods can control the FDR, AutoMS has better detection power with a larger TDR compared to SRS-based methods. (d) The reason why the FDR control of AutoMS is more difficult than SRS-based methods is that the FDP distribution of our selected detector is different from a given detector. Take a simple example: there are 10 standard Gaussian variables, each with a mean 0. The largest variable among those 10 does not have a mean 0. Therefore, the selected detector is more difficult to control FDR.

Fig. 3 shows the performance of AutoMS-JK, AutoMS-SRS and SRS-based-best methods under different target FDR levels. In particular, SRS-LOF-best is the best SRS-LOF detector with the largest TDR among all its candidate hyperparameters. Although all SRS-based-best methods can control the FDR under any target FDR level, the average TDPs of AutoMS-JK/AutoMS-SRS are higher than all the SRS-based-best methods at every target FDR level. The average FDP of AutoMS-JK/AutoMS-SRS is slightly larger than the given target FDR level α, but their difference $\mathrm{FDP} - \alpha$ does not

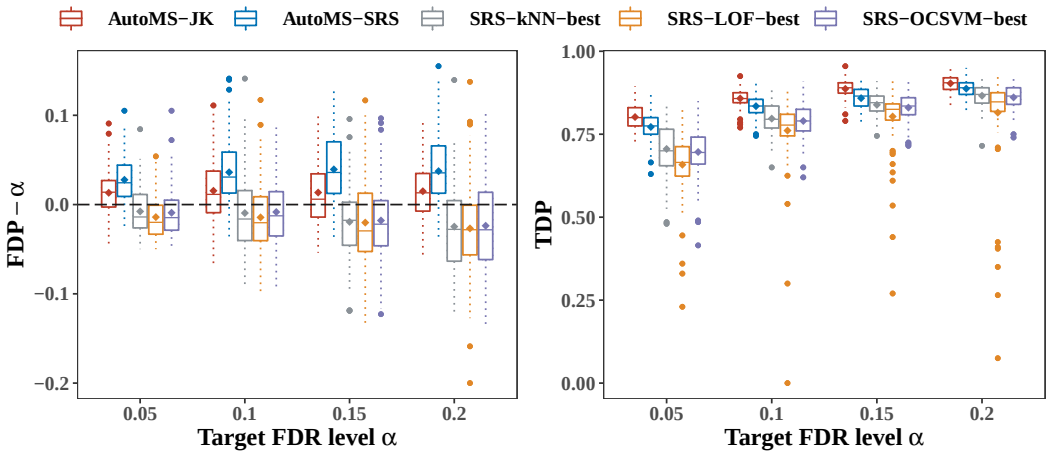

Figure 3: Performance of AutoMS-JK, AutoMS-SRS and SRS-based-best methods with different target FDR level α. The difference between the FDP and the target FDR level α shows the deviation of FDP from α.

increase as α becomes larger. The FDR of AutoMS-JK has a smaller deviation above the target α than AutoMS-SRS. Actually, in novelty detection tasks, it is not necessary to obtain procedures with finite FDR control, instead, a powerful and stable procedure that controls FDR asymptotically around the target FDR level α is good enough for most cases. That is what AutoMS does. AutoMS can uniformly offer much higher TDPs than SRS-based-best methods under different target FDR levels.

5.4 Real Data

We compare the proposed AutoMS with two different types of baselines: (i) model selection methods (METAOD [11]) and (ii) error-rate-controlled methods (SRS-based methods [17]), as in Table 1.

Datasets. Four real-world datasets for outlier detection are considered, and the details of the datasets are summarized in Table 2. We use empirical TDR (Recall) and empirical FDR (1-Precision) as performance measures. For real data implementation, we construct the training dataset $\mathcal{D}$ by randomly sampling $m = \min\{20000, N_0/3\}$ inlier samples from the total N_0 inliers. Then we focus on a test dataset generating mechanism under which 90% of the $n = \min\{2000, (N_0 + N_1)/6\}$ observations in each test dataset $\mathcal{U}$ are sampled from N_0 inliers, and we seek to identify the remaining 10% of outliers (i.e., 10% outlier ratio). For results on 11 more real-world datasets [29] as used in Zhao et al. [11], refer to **Appendix D**.

Table 2: Summary of real-world datasets for novelty detection.

	Credit Card [30]	Covertype [31]	Satellite [32]	Shuttle [33]
# Features d	30	10	36	9
# Inliers N_0	284,807	283,301	5,732	45,586
# Outliers N_1	492	2747	703	12,414

Results. Fig. 4 shows the performance of different methods on real-world datasets, and Table 3 gives the empirical FDR and empirical TDR of different methods, where SRS-best is the best SRS detector with the largest TDR among all detectors. From Fig. 4 and Table 3, we can observe that AutoMS can maintain the FDPs well around the target FDR level $\alpha = 0.1$, while METAOD has inflated the empirical FDP to values around or above 0.5 on all real-world datasets, indicating that in the detected set $\widehat{\mathcal{O}}$ obtained by METAOD, around 50% of the detected outliers are not true outliers. It is worth mentioning that, all SRS-based methods can maintain the FDR level around 0.1 as well as expected from their theoretical guarantees. As for true detection rate (TDR), AutoMS has reasonable empirical TDR on all datasets, while SRS-based methods sometimes perform too conservatively, have no detected outliers and their performance varies greatly from dataset to dataset.

In addition, the performances of SRS-LODA-best and SRS-LOF-best are rather different, particularly on the "Covertype" dataset and "Credit Card" dataset, which means SRS-based-best methods can not give a consistent good result on different tasks and its performance can be severely affected by the base detector the SRS used. On the other hand, AutoMS keeps accurate and stable FDR levels on all datasets and has reasonable TDR for different tasks. Overall, AutoMS can uniformly offer much

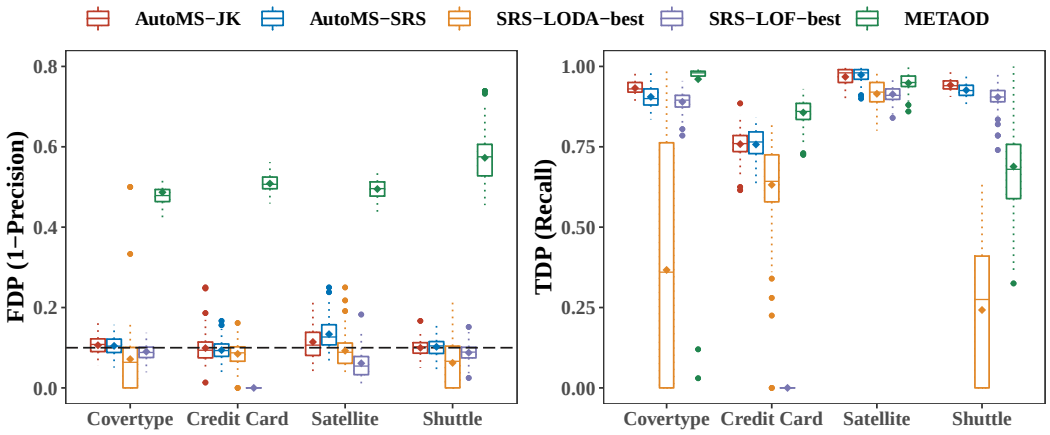

Figure 4: Performance on four different real-world datasets. Each box-plot shows the distribution of FDP and TDP. The dashed line is the target FDR level $\alpha = 0.1$.

higher TDRs than SRS-based-best methods across all settings and AutoMS can maintain the FDR level with a comparable TDR compared with METAOD. Table 4 shows results under different target FDR α levels (refer to **Appendix C** for more results and details).

Table 3: Empirical FDR and TDR of different methods when the target FDR level is set $\alpha = 0.1$.

Dataset	AutoMS-JK		AutoMS-SRS		SRS-best		METAOD	
	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$
Covertype	0.106	0.933	0.105	0.905	0.090	0.890	0.487	0.960
Credit Card	0.099	0.758	0.094	0.757	0.086	0.674	0.508	0.856
Satellite	0.114	0.968	0.134	0.974	0.069	0.963	0.495	0.948
Shuttle	0.100	0.942	0.102	0.926	0.088	0.914	0.573	0.689

Table 4: Empirical FDR and TDR under different target FDR levels α on "Shuttle" dataset.

Shuttle	$\alpha = 0.05$		$\alpha = 0.10$		$\alpha = 0.15$		$\alpha = 0.20$	
Method	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$
AutoMS-JK	0.042	**0.883**	0.100	**0.942**	0.151	**0.951**	0.209	**0.953**
AutoMS-SRS	0.051	0.659	0.102	**0.926**	0.159	**0.952**	0.216	**0.956**
SRS-HBOS-best	0.024	0.400	0.083	0.554	0.131	0.733	0.177	0.748
SRS-OCSVM-best	0.010	0.139	0.088	0.914	0.133	0.942	0.177	0.946
METAOD	0.573	**0.689**	0.573	0.689	0.573	0.689	0.573	0.689

6 Conclusion

The detector selection problem in novelty detection analysis is a challenging problem and it has not been thoroughly investigated in the literature, especially in terms of theoretical guarantees. In this paper, we propose an effective data-driven procedure, named AutoMS, that can select the "best" detector which offers high TDR and controls FDR simultaneously. The AutoMS is fully automatic without any supervision and is easily implemented. We give theoretical results showing that the proposed AutoMS can achieve the non-asymptotic bounds of FDP and maintain FDR control under mild conditions. Extensive experiments on synthetic and real-world datasets show AutoMS significantly improves the detection power and yields accurate FDR control.

Acknowledgments and Funding Disclosure

We thank anonymous area chair and reviewers for their helpful comments. Changliang Zou's research was supported by the National Natural Science Foundation of China Grants (Nos. 11925106, 12231011, 11931001 and 11971247). Haojie Ren's research was supported by the National Natural Science Foundation of China Grants (No.12101398) and the Shanghai Sailing Program. This work was partially supported by the Technology Innovation Institute and MBZUAI joint project (NO. TII/ARRC/2073/2021): Energy-based Probing for Spiking Neural Networks.

References

[1] Hung-Jen Liao, Chun-Hung Richard Lin, Ying-Chih Lin, and Kuang-Yuan Tung. Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, 36 (1):16–24, 2013.

[2] Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, and J Christopher Westland. Data mining for credit card fraud: A comparative study. *Decision support systems*, 50(3):602–613, 2011.

[3] Tzu-Hsuan Lin and Jehn-Ruey Jiang. Credit card fraud detection with autoencoder and probabilistic random forest. *Mathematics*, 9(21):2683, 2021.

[4] Jan Verbesselt, Achim Zeileis, and Martin Herold. Near real-time disturbance detection using satellite image time series. *Remote Sensing of Environment*, 123:98–108, 2012.

[5] Rémi Domingues. Probabilistic modeling for novelty detection with applications to fraud identification. *arXiv preprint arXiv:1903.01730*, 2019.

[6] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[7] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021.

[8] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021.

[9] Quentin Rebjock, Baris Kurt, Tim Januschowski, and Laurent Callot. Online false discovery rate control for anomaly detection in time series. *Advances in Neural Information Processing Systems*, 34, 2021.

[10] David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[11] Yue Zhao, Ryan Rossi, and Leman Akoglu. Automatic unsupervised outlier model selection. *Advances in Neural Information Processing Systems*, 34, 2021.

[12] Yuening Li, Zhengzhang Chen, Daochen Zha, Kaixiong Zhou, Haifeng Jin, Haifeng Chen, and Xia Hu. Autood: Neural architecture search for outlier detection. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 2117–2122. IEEE, 2021.

[13] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[14] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, 57(1):289–300, 1995.

[15] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[16] John D Storey, Jonathan E Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

[17] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *arXiv preprint arXiv:2104.08279*, 2021.

[18] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[19] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[20] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

[21] Charu C Aggarwal. *Outlier Analysis*. Springer, 2017.

[22] Nicolas Goix. How to evaluate the quality of unsupervised anomaly detection algorithms? *arXiv preprint arXiv:1607.01152*, 2016.

[23] Thanh Trung Nguyen, Uy Quang Nguyen, et al. An evaluation method for unsupervised anomaly detection algorithms. *Journal of Computer Science and Cybernetics*, 32(3):259–272, 2016.

[24] Henrique O Marques, Ricardo JGB Campello, Jörg Sander, and Arthur Zimek. Internal evaluation of unsupervised outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(4):1–42, 2020.

[25] Martin Q Ma, Yue Zhao, Xiaorong Zhang, and Leman Akoglu. A large-scale study on unsupervised outlier model selection: Do internal strategies suffice? *arXiv preprint arXiv:2104.01422*, 2021.

[26] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

[27] Lilun Du, Xu Guo, Wenguang Sun, and Changliang Zou. False discovery rate control under general dependence by symmetrized data aggregation. *Journal of the American Statistical Association*, just-accepted:1–34, 2021.

[28] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019. URL http://jmlr.org/papers/v20/19-011.html.

[29] Andrew Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng Keen Wong. Anomaly detection meta-analysis benchmarks. *Oregon State University*, 2016. URL https://ir.library.oregonstate.edu/concern/datasets/47429f155.

[30] Credit Card. Credit card fraud detection, 2016. URL https://www.kaggle.com/mlg-ulb/creditcardfraud.

[31] Covertype. Forest covertype, 1998. URL http://odds.cs.stonybrook.edu/forestcovercovertype-dataset.

[32] Satellite. Satellite image, 1993. URL http://odds.cs.stonybrook.edu/satimage-2-dataset.

[33] Shuttle. Statlog (shuttle), 1993. URL https://archive.ics.uci.edu/ml/datasets/Statlog+(Shuttle).

[34] Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*, pages 59–63, 2012.

[35] Tomáš Pevnỳ. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 102(2):275–304, 2016.

[36] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 427–438, 2000.

[37] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 11:2973–3009, 2010.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See at the end of Sect. 1, we summarize our main contributions. And the experimental results in Sect. 5 validate the theoretical results in Sect. 4.
   (b) Did you describe the limitations of your work? [Yes] See Remark 1 in Sect. 4, we discuss the stability limitation (Assumption 2) of detector algorithm.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is foundational research and not tied to particular applications.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We read the ethics review guidelines and ensured that our paper conforms to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumption 1-Assumption 3 in Sect. 4.
   (b) Did you include complete proofs of all theoretical results? [Yes] See **Appendix E** in the Supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Algorithm 1 in Sect. 3, we give the pseudo-code of our AutoMS procedure, and release the code on github `https://github.com/ZhangYifan1996/AutoMS`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We describe the generation of our synthetic data in Sect. 5.3 and give the candidate detectors we use in **Appendix A** of the Supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Our experimental results are all shown with boxplots.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use the Python "**PyOD**" package and have cited the creator. The real data we use is open source and the URLs are cited.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Material for "AutoMS: Automatic Model Selection for Novelty Detection with Error Rate Control"

Here we list all the terms we use in the paper.

- FDP, false discovery proportion (FDP), is the proportion of the normal samples that are erroneously detected as outliers.

$$\text{FDP} = \frac{|\widehat{\mathcal{O}}_{\mathcal{M}} \bigcap \mathcal{O}^c|}{1 \vee |\widehat{\mathcal{O}}_{\mathcal{M}}|},$$

  where $\mathcal{O} \subset \mathcal{U}$ is the outlier set, $\mathcal{O}^c \subset \mathcal{U}$ is the inlier set, $\mathcal{U} = \mathcal{O} \cup \mathcal{O}^c$, $\widehat{\mathcal{O}}_{\mathcal{M}}$ is the decision set containing samples detected as outliers.
- TDP, true discovery proportion (TDP), is the proportion of the outliers that are correctly detected,

$$\text{TDP} = \frac{|\widehat{\mathcal{O}}_{\mathcal{M}} \bigcap \mathcal{O}|}{|\mathcal{O}|}.$$

- FDR, false discovery rate (FDR), is the expectation of FDP,

$$\text{FDR} = \mathbb{E}(\text{FDP}).$$

- TDR, true discovery rate (TDR), is the expectation of TDP,

$$\text{TDR} = \mathbb{E}(\text{TDP}).$$

- Precision, is the fraction of inliers that are correctly detected among the samples are detected. And FDP=1-Precision.
- Recall/Power, is the fraction of outliers that are correctly detected among the true outliers. And TDP= Recall = Power. Power = TDP.
- $\widehat{\text{FDR}}$ is the empirical FDR, which is the average FDP over 100 repetitions.
- $\widehat{\text{TDR}}$ is the empirical TDR, which is the average TDP over 100 repetitions.

A AutoMS Detector Set

We give an example of the detector set $\mathcal{G}$. We use 6 well-known novelty detection methods: Histogram-Based Outlier Score (HBOS), isolation Forest (iForest), K Nearest Neighbors (KNN), Lightweight On-line Detector of Anomalies (LODA), Local Outlier Factor (LOF) and One-Class Support Vector Machine (OCSVM), combined with their corresponding hyperparameters as a set of detectors $\mathcal{G}$ with 83 candidate models in our experiments. Those detectors are available in the Python "**PyOD**" [28] package. A complete list of the detector set $\mathcal{G}$ is shown in Table S1.

Note that the size of detector set $\mathcal{G}$ can be infinite, and our AutoMS procedure can still be used.

Table S1: An example of the detector set $\mathcal{G}$: detection methods with their hyperparameters (with default hyperparameters in bold).

Method	Article	hyperparameters
HBOS	Goldstein and Dengel [34]	n_histograms:{5,**10**,20,30,40,50,75,100}
iForest	Liu et al. [20]	n_estimators:{50,**100**,150,200}
		max_feature:{0.2,0.4,0.6,0.8,**1.0**}
KNN	Ramaswamy et al. [36]	n_neighbors:{1,**5**,10,15,20,50,75,100}
		method:{**"largest"**,"mean"}
LODA	Pevnỳ [35]	n_bins: {**10**,20,30,40,50,75,100,150,200}
LOF	Breunig et al. [18]	n_neighbors:{10,**20**,30,40,50,60,70,80,90,100}
OCSVM	Schölkopf et al. [19]	ν:{0.1,0.3,**0.5**,0.7,0.9}
		kernel:{**"rbf"**,"linear","poly","sigmoid"}

B Experiment Setting with Details

To evaluate the performance of the proposed **AutoMS** method, we compare the proposed AutoMS with two different types of baselines: (i) model selection methods (METAOD [11]) and (ii) error-rate-controlled methods (SRS-based methods [17]), as in Table 1. (i) For model selection methods, we use METAOD [11] as a representative, since it deals with automatic unsupervised outlier model selection which uses the performances of historical outlier detection benchmark datasets as prior experience to automatically do model selection via meta-learning. (ii) For error-rate-controlled methods, we choose SRS-based methods [17] as typical methods to give statistical guarantees for the error rate control, but there is no model selection considered in SRS-based methods.

We consider two versions of the proposed AutoMS: **AutoMS-JK** and **AutoMS-SRS**, as shown in Table 1, where AutoMS-JK uses Jackknife technique and AutoMS-SRS replaces the Jackknife technique with single-random-splitting (SRS) technique. We explore the performances of the four methods with respect to FDR and TDR.

We give the detailed introduction for the four methods: (1) **SRS** [17], **METAOD** [11], (3) **AutoMS-JK** and (4) **AutoMS-SRS**.

(1) **SRS** [17] employs the single-random-splitting (SRS) based method in conjunction with a fixed detector, and constructs conformal p-values to control the FDR. We combine SRS with different detectors from the same $\mathcal{G}$ we use for AutoMS. In particular, SRS-LOF means that we use LOF as the base detector in combination with SRS, while SRS-LOF-best is the best SRS-LOF detector with the largest TDR among all its candidate hyperparameters. **SRS-based** methods [17] are representative error-rate-controlled methods which combine SRS with different base detectors from $\mathcal{G}$ as used for AutoMS. It is worth mentioning that SRS is guaranteed with FDR control for any given detector. But there is no model selection considered in SRS-based methods.

(2) **METAOD** [11] is a model selection algorithm which uses meta-learning to learn the performance of a large body of detection models on historical outlier detection benchmark datasets, and carries over this prior experience to automatically select an effective model to be employed on a new dataset. **METAOD** is a representative model selection method using meta-learning without considering error rate control, while the **AutoMS** can do both model selection and error rate control.

(3) **AutoMS-JK** is our proposed model selection procedure using Jackknife technique to select the "best" detector among a bunch of candidate detectors while controlling the FDR.

(4) **AutoMS-SRS** is a variant of AutoMS that replaces the Jackknife technique with single-random-splitting (SRS) to incorporate our model selection procedure.
As the single-random-splitting (SRS) can be combined with our model selection procedure, hereafter called AutoMS-SRS. AutoMS-SRS can be regarded as a special case of AutoMS and also has the theoretical guarantees that the selected model yields **asymptotically** valid FDR control. But SRS does not fully explore the clean data and can cause randomness by data-splitting. So we use the Jackknife method instead of SRS to improve the accuracy and stability of the estimated p-values and enhance detection power. Specially, when the sample size is too large, we use AutoMS-SRS to reduce the computation complexity.

C Detailed Experiment Results on 4 Real-world Datasets in the Main Paper

In this part, we provide detailed experiment results on the four real-world datasets, and we provide additional experiment results on 11 more real-world datasets in the next part.

We show the empirical FDR and TDR as performance measures of different methods on four real-world datasets with different target FDR levels α in Table S2, as a detailed extension of Table 3 and Table 4 in Sect. 5 of the main paper. As introduced in Sect. 5, the empirical FDR and TDR are calculated as the expectation of FDP and TDP, respectively. Note the performance of METAOD is same under different α since METAOD does not employ any FDR control procedure.

Table S2: Empirical FDR and TDR of AutoMS-JK, AutoMS-SRS, SRS-based methods and METAOD on four real-world datasets under different target FDR levels α (extension of Table 3 and Table 4).

Method	$\alpha = 0.05$		$\alpha = 0.10$		$\alpha = 0.15$		$\alpha = 0.20$	
	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$
Covertype								
AutoMS-JK	0.056	**0.905**	0.106	**0.933**	0.158	**0.943**	0.209	**0.948**
AutoMS-SRS	0.055	0.861	0.105	0.905	0.164	0.930	0.210	0.932
SRS-HBOS-best	0.000	0.000	0.062	0.001	0.138	0.007	0.221	0.011
SRS-iForest-best	0.033	0.011	0.090	0.022	0.102	0.045	0.165	0.042
SRS-KNN-best	0.046	0.802	0.092	0.871	0.135	0.902	0.181	0.922
SRS-LODA-best	0.012	0.146	0.072	0.367	0.116	0.593	0.165	0.759
SRS-LOF-best	0.046	0.848	0.090	0.890	0.135	0.912	0.177	0.926
SRS-OCSVM-best	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
METAOD	0.487	**0.960**	0.487	**0.960**	0.487	**0.960**	0.487	**0.960**
Credit Card								
AutoMS-JK	0.047	**0.608**	0.099	**0.758**	0.158	0.805	0.207	**0.830**
AutoMS-SRS	0.050	0.595	0.094	0.757	0.146	**0.807**	0.198	**0.830**
SRS-HBOS-best	0.041	0.469	0.086	0.674	0.132	0.762	0.175	0.797
SRS-iForest-best	0.040	0.226	0.087	0.460	0.131	0.606	0.180	0.704
SRS-KNN-best	0.000	0.000	0.006	0.002	0.050	0.001	0.243	0.006
SRS-LODA-best	0.035	0.345	0.085	0.632	0.129	0.746	0.178	0.797
SRS-LOF-best	0.000	0.000	0.000	0.000	0.000	0.000	0.150	0.000
SRS-OCSVM-best	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
METAOD	0.508	**0.856**	0.508	**0.856**	0.508	**0.856**	0.508	**0.856**
Satellite								
AutoMS-JK	0.067	**0.951**	0.114	**0.968**	0.165	**0.969**	0.218	**0.970**
AutoMS-SRS	0.068	**0.968**	0.134	**0.974**	0.183	**0.975**	0.236	**0.974**
SRS-HBOS-best	0.046	0.720	0.099	0.761	0.149	0.787	0.212	0.823
SRS-iForest-best	0.039	0.866	0.086	0.898	0.134	0.920	0.178	0.932
SRS-KNN-best	0.041	0.941	0.069	0.963	0.105	0.969	0.152	0.971
SRS-LODA-best	0.050	0.876	0.092	0.915	0.134	0.935	0.182	0.950
SRS-LOF-best	0.024	0.883	0.061	0.913	0.106	0.931	0.164	0.947
SRS-OCSVM-best	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
METAOD	0.495	0.948	0.495	0.948	0.495	0.948	0.495	0.948
Shuttle								
AutoMS-JK	0.042	**0.883**	0.100	**0.942**	0.151	**0.951**	0.209	**0.953**
AutoMS-SRS	0.051	0.659	0.102	**0.926**	0.159	**0.952**	0.216	**0.956**
SRS-HBOS-best	0.024	0.400	0.083	0.554	0.131	0.733	0.177	0.748
SRS-iForest-best	0.044	0.417	0.092	0.469	0.138	0.504	0.178	0.528
SRS-KNN-best	0.015	0.195	0.086	0.906	0.130	0.942	0.171	0.944
SRS-LODA-best	0.022	0.102	0.062	0.242	0.117	0.363	0.169	0.498
SRS-LOF-best	0.010	0.124	0.088	0.904	0.137	0.933	0.184	0.944
SRS-OCSVM-best	0.010	0.139	0.088	0.914	0.133	0.942	0.177	0.946
METAOD	0.573	**0.689**	0.573	0.689	0.573	0.689	0.573	0.689

From Table S2, we see that the empirical FDR values of AutoMS (either AutoMS-JK or AutoMS-SRS) and SRS-based methods are all around the target FDR levels for all datasets. In finite samples, the empirical FDR of AutoMS is slightly larger than the given target FDR level α, but their difference does not increase as α becomes larger. We also list the Standard Deviation for FDP and TDP of AutoMS-JK, AutoMS-SRS, SRS-based methods and METAOD on four real-world datasets under different target FDR levels α in Table S3.

Moreover, with similar target FDR level α, AutoMS can uniformly offer much higher TDRs than SRS-based methods across all settings, as the former automatically selects the most suitable model

Table S3: Standard deviation for empirical FDR and TDR of AutoMS-JK, AutoMS-SRS, SRS-based methods and METAOD on four real-world datasets under different target FDR levels α .

	$\alpha = 0.05$		$\alpha = 0.10$		$\alpha = 0.15$		$\alpha = 0.20$	
Method	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$
Covertype								
AutoMS-JK	0.017	0.022	0.022	0.021	0.026	0.021	0.027	0.019
AutoMS-SRS	0.018	0.036	0.023	0.034	0.026	0.029	0.027	0.020
SRS-HBOS-best	0.000	0.000	**0.228**	0.003	**0.285**	0.016	**0.365**	0.022
SRS-iForest-best	0.075	0.019	0.137	0.030	0.131	0.051	0.233	0.048
SRS-KNN-best	0.018	0.045	0.024	0.033	0.028	0.022	0.036	0.021
SRS-LODA-best	0.024	0.269	0.127	**0.380**	0.109	**0.366**	0.062	**0.279**
SRS-LOF-best	0.019	0.042	0.021	0.034	0.027	0.025	0.029	0.024
SRS-OCSVM-best	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
METAOD	**0.066**	**0.128**	0.066	0.128	0.066	0.128	0.066	0.128
Credit Card								
AutoMS-JK	0.021	0.076	**0.036**	0.046	0.118	0.037	0.181	0.053
AutoMS-SRS	0.021	0.089	0.026	0.049	0.031	0.028	0.034	0.024
SRS-HBOS-best	0.020	0.114	0.028	0.071	0.029	0.049	0.032	0.032
SRS-iForest-best	**0.030**	0.114	0.029	0.095	0.035	**0.081**	0.039	**0.063**
SRS-KNN-best	0.000	0.000	0.028	0.008	**0.224**	0.004	**0.371**	0.012
SRS-LODA-best	0.028	**0.240**	0.029	**0.145**	0.035	0.066	0.036	0.046
SRS-LOF-best	0.000	0.000	0.000	0.000	0.000	0.000	0.366	0.000
SRS-OCSVM-best	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
METAOD	0.021	0.039	0.021	0.039	0.021	0.039	0.021	0.039
Satellite								
AutoMS-JK	**0.039**	0.031	0.043	0.027	0.053	0.029	0.043	0.027
AutoMS-SRS	0.030	0.024	0.040	0.024	0.045	0.023	0.039	0.023
SRS-HBOS-best	0.032	**0.056**	**0.045**	**0.045**	**0.062**	**0.049**	**0.055**	**0.040**
SRS-iForest-best	0.022	0.033	0.040	0.036	0.047	0.026	0.052	0.024
SRS-KNN-best	0.018	0.026	0.031	0.012	0.039	0.008	0.049	0.007
SRS-LODA-best	0.029	0.054	0.042	0.038	0.049	0.033	0.046	0.026
SRS-LOF-best	0.016	0.035	0.034	0.026	0.042	0.026	0.042	0.024
SRS-OCSVM-best	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
METAOD	0.022	0.027	0.022	0.027	0.022	0.027	0.022	0.027
Shuttle								
AutoMS-JK	0.013	0.030	0.022	0.017	0.027	0.019	0.030	0.020
AutoMS-SRS	0.021	0.149	0.021	0.024	0.024	0.024	0.027	0.028
SRS-HBOS-best	0.026	0.162	0.037	**0.260**	0.031	**0.203**	0.033	**0.204**
SRS-iForest-best	0.023	0.054	0.030	0.050	0.038	0.046	0.045	0.053
SRS-KNN-best	0.026	**0.328**	0.021	0.026	0.023	0.017	0.027	0.016
SRS-LODA-best	0.037	0.145	0.054	0.201	0.065	0.197	0.054	0.138
SRS-LOF-best	0.023	0.267	0.024	0.036	0.028	0.023	0.035	0.019
SRS-OCSVM-best	0.022	0.308	0.021	0.027	0.027	0.017	0.028	0.017
METAOD	**0.067**	0.169	**0.067**	0.169	**0.067**	0.169	**0.067**	0.169

while maintaining the FDR control. On the Satellite dataset, the SRS-KNN-best method achieves the same high empirical TDP as AutoMS. It is because the optimal algorithm and hyperparameter selected by AutoMS are almost the same as those used by SRS-KNN-best.

On the other hand, the METAOD algorithm has a higher empirical TDR (i.e. higher recall) than AutoMS in some settings (such as "Covertype" dataset and "Credit Card" dataset), while the empirical FDR (i.e. 1 - Precision) of METAOD can be as large as 50%, indicating that around 50% of outliers detected by METAOD are not true outliers. That is, the recall of METAOD can be larger than AutoMS in some applications, while its precision can only be maintained around or smaller than 50%, which

results from the fact that there is no error control in METAOD. In conclution, AutoMS is a more stable novelty detection method containing model selection and FDR control.

We can see that the larger the target FDR level α, the larger the empirical TDR, which means the more relaxation of the target FDR level, the higher improvement of empirical TDR, indicating the trade-off between precision and recall of detection methods. Adjusting the target FDR level leaves us the trade-off that with a larger target FDR level α the detection method can detect more outliers but produce more false discoveries.

So we can adjust our target FDR level α to obtain the satisfied FDR and TDR. However, the performance of METAOD is same under different target FDR levels α since METAOD does not employ any FDR control procedure. And AutoMS can uniformly offer much higher TDRs than SRS-based-best methods across all FDR levels. Therefore, our AutoMS is a practical model selection method, which takes into account FDR and TDR.

D Additional Experiment Results on More Real-world Datasets

As for the experiment scale of our AutoMS method, we are here trying to make it clear that why there is no need for our AutoMS method to go through hundreds of datasets. As MetaOD uses meta-learning, METAOD requires a large number of datasets as the historical benchmarks to measure the similarity between the test dataset and benchmark datasets. And the different test-bechmark-dataset-similarity will effect the results of METAOD, while our AutoMS approach has no special requirements for datasets. We have shown experiments on 4 datasets in the main paper, and the proposed AutoMS method works on all other datasets, of course not only on these 4 datasets shown in the main paper, which can be obtained from Theorem 1 (FDR control) in our paper. We have conducted additional experiments on 11 more real-world datasets. The conclusion still remains the same that AutoMS can always control FDR while METAOD cannot. To our best knowledge, is the first effort of model selection for novelty detection with theoretical guarantees in the view of FDR control. Our proposed AutoMS method can select the best model and simultaneously control the error rate of the best model.

Table S4: Summary of 11 more real-world datasets.

Dataset	d	N_0	N_1	Outlier Ratio (p)
abalone	7	2096	2081	49.7%
comm.and.crime	101	1001	993	49.7%
imgseg	18	1320	990	42.9%
letter.rec	16	10022	9978	49.9%
magic.gamma	10	12332	6688	35.2%
opt.digits	62	3357	2263	40.2%
pageb	10	4913	560	10.2%
skin	3	194198	50859	20.8%
spambase	57	2788	1813	39.3%
synthetic	10	10000	10000	50.0%
wave	21	3343	1657	33.1%

We provide additional experiment results on 11 more real-world datasets which are pre-processed data from Emmott et al. [29]. And the details of 11 real-world datasets are summarized in Table S4. The outlier ratio of the original dataset is calculated by $p = \frac{N_1}{N_0+N_1}$. We construct the training dataset $\mathcal{D}$ by random sampling $m = \min\{20000, N_0/3\}$ from the total N_0 inliers. As for the test dataset $\mathcal{U}$, we need $n = \min\{2000, (N_0 + N_1)/6\}$ samples in total to generate the test dataset. We fix the outlier ratio as p, then randomly sample $n \times (1 - p)$ inliers from the remaining $N_0 - m$ inliers, and randomly sample $n \times p$ outliers from N_1 outliers. When constructing the test dataset $\mathcal{U}$, the outlier ratio for each dataset is the same as the original dataset as shown in Table S4.

Note that there are only inliers in the training dataset $\mathcal{D}$, and there are p percent of outliers in the test dataset $\mathcal{U}$. For example, for the "abalone" dataset, there are 49.7% outliers in its test dataset $\mathcal{U}$, and the goal of novelty detection is to detect the 49.7% outliers, using only clean data $\mathcal{D}$. The goal of model selection for unsupervised outlier detection is to select the best outlier detection model. Our

goal here is to select the best model and simultaneously to control the error rate of the best model. We propose AutoMS method to achieve the goal of model selection for unsupervised outlier detection with error rate control.

Table S5: Empirical FDR and TDR of AutoMS-JK, AutoMS-SRS, SRS-best and METAOD on 11 more real-world datasets when the target FDR level $\alpha = 0.1$, where SRS-best is the best detector of SRS-based methods with the largest TDR among all detectors.

Dataset	Outlier	AutoMS-JK		AutoMS-SRS		SRS-best		METAOD	
	Ratio (p)	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$	$\widehat{\text{FDR}}$	$\widehat{\text{TDR}}$
abalone	49.7%	0.077	0.277	0.096	0.335	0.062	0.284	0.192	0.436
comm.and.crime	49.7%	0.064	0.540	0.091	0.607	0.056	0.497	0.197	0.427
imgseg	42.9%	0.046	0.510	0.122	0.544	0.059	0.496	0.230	0.448
letter.rec	49.9%	0.052	0.419	0.056	0.269	0.048	0.243	0.273	0.339
magic.gamma	35.2%	0.071	0.352	0.083	0.354	0.065	0.337	0.244	0.581
opt.digits	40.2%	0.065	0.767	0.096	0.743	0.059	0.718	0.219	0.592
pageb	10.2%	0.138	0.374	0.182	0.477	0.110	0.353	0.546	0.720
skin	20.8%	0.128	0.943	0.086	0.976	0.079	0.969	0.595	0.373
spambase	39.3%	0.067	0.673	0.089	0.717	0.064	0.646	0.390	0.239
synthetic	50.0%	0.059	0.994	0.064	0.994	0.050	0.991	0.225	0.460
wave	33.1%	0.121	0.148	0.156	0.167	0.070	0.092	0.358	0.374

Table S5 shows the results of empirical FDR and empirical TDR of different methods (including AutoMS-JK, AutoMS-SRS, SRS-best and METAOD) on 11 more real-world datasets when the target FDR level $\alpha = 0.1$. As introduced above, these 11 real-world datasets have different outlier ratios which are shown in Table S4. The outlier ratio ranges from 10% to 50%. Unsupervised novelty detection becomes more difficult when the outlier ratio goes higher, because only inliers are used in the cleaning training dataset $\mathcal{D}$ and more outliers need to be detected in the test dataset $\mathcal{U}$. The outlier ratio of the 11 real-world datasets ranges from 10% to 50%. The experiment results in Table S5 show that at different outlier ratio, our AutoMS approach outperforms SRS-best and MATAOD for all the 11 real-world datasets.

Table S5 shows that the empirical TDR of AutoMS is higher than the empirical TDR of SRS-best since SRS-based methods only consider FDR control without model selection, while AutoMS selects the best detector with largest TDR while the empirical FDR is still controlled asymptotically around the FDR level.

The effectiveness of our AutoMS can be found and supported by the theories of our paper, which theoretically guarantees that AutoMS can control the FDR. It is worth noting that METAOD requires a large number of datasets as the historical benchmark to measure the similarity between the test dataset and benchmark datasets as it uses meta-learning.

All the empirical FDRs of all datasets using METAOD are as large as 0.2 or even larger, indicating a very high false discovery rate, which means METAOD can not control the FDR. However, our proposed AutoMS can improve the TDR while controlling the FDR.

For the two versions of AutoMS, sometimes AutoMS-SRS has higher empirical TDR than AutoMS-JK, but the empirical FDR of AutoMS-SRS has a larger deviation above target FDR level α than AutoMS-JK. Table S6 gives the standard deviation of empirical FDR and empirical TDR of different methods on 11 more real-world datasets when the target FDR level $\alpha = 0.1$.

Table S6: Standard deviation of empirical FDR and empirical TDR with different methods (including AutoMS-JK, AutoMS-SRS, SRS-best and METAOD) on 11 more real-world datasets when the target FDR level $\alpha = 0.1$, where SRS-best is the best detector of SRS-based methods with the largest TDR among all detectors.

Dataset	Outlier Ratio (p)	AutoMS-JK $\widehat{\text{FDR}}$	AutoMS-JK $\widehat{\text{TDR}}$	AutoMS-SRS $\widehat{\text{FDR}}$	AutoMS-SRS $\widehat{\text{TDR}}$	SRS-best $\widehat{\text{FDR}}$	SRS-best $\widehat{\text{TDR}}$	METAOD $\widehat{\text{FDR}}$	METAOD $\widehat{\text{TDR}}$
abalone	49.7%	0.040	**0.110**	**0.050**	0.087	0.033	0.090	0.031	0.051
comm.and.crime	49.7%	0.026	0.072	0.029	0.055	0.031	**0.158**	**0.044**	0.083
imgseg	42.9%	0.033	0.056	**0.125**	**0.097**	0.033	0.076	0.041	0.053
letter.rec	49.9%	0.013	0.025	0.015	0.041	0.016	**0.057**	**0.047**	0.045
magic.gamma	35.2%	0.020	0.026	0.023	0.028	0.021	0.032	**0.029**	**0.064**
opt.digits	40.2%	0.023	0.039	**0.117**	0.067	0.020	0.099	0.075	**0.122**
pageb	10.2%	**0.086**	0.100	0.077	0.095	0.068	**0.118**	0.039	0.071
skin	20.8%	0.186	0.181	0.014	0.011	0.014	0.020	**0.231**	**0.194**
spambase	39.3%	0.020	0.069	**0.077**	0.060	0.024	**0.105**	0.044	0.035
synthetic	50.0%	0.007	0.003	0.008	0.003	0.008	0.002	**0.016**	**0.015**
wave	33.1%	**0.086**	0.066	0.056	0.062	0.056	0.048	0.061	**0.072**

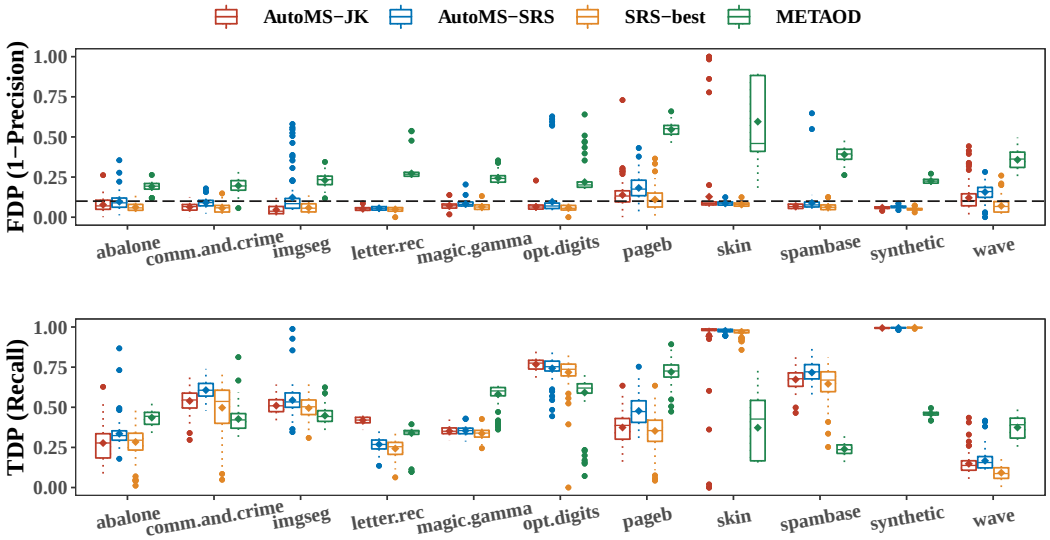

Figure S1: Performance on eleven different real-world datasets. Each box-plot shows the distribution of FDP and TDP. The dashed line is the target FDR level $\alpha = 0.1$.

Fig. S1 shows the performance of AutoMS-JK, AutoMS-SRS, SRS-best and METAOD on different real-world datasets when the target FDR level $\alpha = 0.1$. The empirical FDR of AutoMS-JK and AutoMS-SRS is all controlled around the given target FDR level $\alpha = 0.1$, The empirical FDR of AutoMS-JK has a smaller deviation above the target FDR α than AutoMS-SRS on "pageb" and "wave" datasets. The empirical TDR of AutoMS-JK and AutoMS-SRS is higher than the SRS-best methods at every outlier detections. Although the empirical TDR of METAOD is much higher than AutoMS, the empirical FDR is the worst as on error rate control has been taken into count. Therefore, our AutoMS procedure can be more practical in real-world applications, which takes into account FDR and TDR and selects the best detector.

E Proofs

E.1 Proof of Theorem

We present the proof of the main Theorem 1 in the Appendix. Before starting, we make clear the notations first. Let $\mathcal{D}$ and $\mathcal{U}$ be the training dataset and test dataset, respectively. Let $\mathcal{O}^c \subset \mathcal{U}$ be the subset of $\mathcal{U}$ containing all the inliers with size n_0, $\mathcal{O} \subset \mathcal{U}$ be the subset of all the outliers, and $\mathcal{O} \bigcup \mathcal{O}^c = \mathcal{U}$. When there is no risk of confusion, we also denote $\mathcal{U}$ the indices of the test dataset, $\mathcal{O}$ the indices of outliers, and $\mathcal{O}^c$ the indices of inliers. Let $S_{i,\mathcal{M}}$ be the short of $S_{\mathcal{M}}(Z_i)$ for $Z_i \in \mathcal{U}$ (i.e., $i \in \mathcal{U}$).

Note $S_{i,\mathcal{M}}$ and $S_{k,\mathcal{M}}$ are i.i.d. for each pair $(i,k) \in \mathcal{U}$ conditional on $\mathcal{D}$ and detector $\mathcal{M}$. We define $G_{\mathcal{M}}(t) = \mathbb{P}(S_{\mathcal{M}}(Z) > t \mid \mathcal{D})$, $G^{-}_{\mathcal{M}}(y) = \inf\{t \geq 0 : G_{\mathcal{M}}(t) \leq y\}$ with $0 \leq y \leq 1$.

Define

$$G_{n,\mathcal{M}}(t) = n_0^{-1} \sum_{i \in \mathcal{O}^c} \mathbb{I}(S_{i,\mathcal{M}} > t),$$

where $n_0 = |\mathcal{O}^c|$ and recall

$$\mathcal{M}^* = \arg\max_{\mathcal{M} \in \mathcal{G}} |\widehat{O}_{\mathcal{M}}| = \arg\max_{\mathcal{M} \in \mathcal{G}} \sum_{i=1}^{n} \mathbb{I}(S_{i,\mathcal{M}} > L_{\mathcal{M}}),$$

$$L_{\mathcal{M}} = \inf\left\{0 < t \leq \bar{t}_{m,\mathcal{M}} : \frac{nG_{m,\mathcal{M}}(t)}{1 \vee \sum_{i=1}^{n} \mathbb{I}(S_{i,\mathcal{M}} > t)} \leq \alpha\right\}, \tag{C.1}$$

where $\bar{t}_{m,\mathcal{M}} = G^{-}_{m,\mathcal{M}}(A_n/n)$, $G_{m,\mathcal{M}}(t) = m^{-1}\sum_{j=1}^{m} \mathbb{I}(S^{[-j]}_{j,\mathcal{M}} > t)$ and $A_n = o(n)$ is a pre-specified sequence.

In the proof part, we use c and C denote strictly positive constants that can be changed from place to place, and they do not depend on n and m. Let

$$\tilde{G}_{\mathcal{M}}(t) = \mathbb{E}\{G_{\mathcal{M}}(t)\}.$$

Proof of Theorem 1. Briefly, we first prove that the FDP of each detector can be bounded **uniformly** according to the uniform convergence properties Lemma C.1-C.3. Hence, the selected detector with the largest number of discoveries will have the same non-asymptotic bounds for FDP. Then we give the asymptotic FDR control for the optimal detector.

By Lemma C.3-(ii) we have

$$\sup_{0 \leq t \leq \bar{t}_{\mathcal{M}}} \left| \frac{\tilde{G}_{\mathcal{M}}(t)}{G_{\mathcal{M}}(t)} - 1 \right| \leq 4nB_m/A_n. \tag{C.2}$$

By Lemmas C.1 and C.2, we have

$$\sup_{0 \leq t \leq \bar{t}_{\mathcal{M}}} \left| \frac{G_{n,\mathcal{M}}(t)}{G_{\mathcal{M}}(t)} - 1 \right| \leq C_1 \delta^{-1/2} \sqrt{\frac{\varpi_m}{A_n^{2/3}}}, \tag{C.3}$$

$$\sup_{0 \leq t \leq \tilde{t}_{\mathcal{M}}} \left| \frac{G_{m,\mathcal{M}}(t)}{\tilde{G}_{\mathcal{M}}(t)} - 1 \right| \leq C_2 \delta^{-1/2} \left\{ \varpi_m \left(\frac{n}{mA_n^{2/3}} + \frac{n}{A_n^{4/3}} \right) (1 + 15B_m + 2m^2 B_m) \right\}^{1/2} \tag{C.4}$$

hold simultaneously with probability at least $1 - 2\delta$.

Under the event that Eqs. (C.3) and (C.4) hold, we have

$$G_{m,\mathcal{M}}(\bar{t}_{\mathcal{M}}) \leq (1 + g_{m,n}) G_{\mathcal{M}}(\bar{t}_{\mathcal{M}}) \leq (1 + g_{m,n}) A_n/2,$$

where $g_{m,n} = o(1)$ under Assumption 3. Hence, with probability at least $1 - 2\delta$, for sufficiently large m and n,

$$G_{m,\mathcal{M}}(\bar{t}_{\mathcal{M}}) \leq G_{m,\mathcal{M}}(\bar{t}_{m,\mathcal{M}}) = A_n.$$

Since $G_{m,\mathcal{M}}(t)$ is a deceasing function of t, we have $\bar{t}_{m,\mathcal{M}} \leq \bar{t}_{\mathcal{M}}$ with high probability, and hence $\bar{t}_{m,\mathcal{M}^*} \leq \bar{t}_{\mathcal{M}^*}$. Similarly, $\bar{t}_{m,\mathcal{M}^*} \leq \tilde{t}_{\mathcal{M}^*}$ with at least $1 - 2\delta$.

Write

$$\begin{aligned}
\mathrm{FDP}(L_{\mathcal{M}^*}) &= \frac{\sum_{i\in\mathcal{O}^c}\mathbb{I}(S_{i,\mathcal{M}^*} > L_{\mathcal{M}^*})}{1\vee\sum_{i=1}^n \mathbb{I}(S_{i,\mathcal{M}^*} > L_{\mathcal{M}^*})} \\
&= \frac{n_0}{n}\times\frac{nG_{m,\mathcal{M}^*}(L_{\mathcal{M}^*})}{1\vee\sum_{i=1}^n \mathbb{I}(S_{i,\mathcal{M}^*} > L_{\mathcal{M}^*})}\times\frac{G_{\mathcal{M}^*}(L_{\mathcal{M}^*})}{\tilde{G}_{\mathcal{M}^*}(L_{\mathcal{M}^*})}\times\frac{G_{n,\mathcal{M}^*}(L_{\mathcal{M}^*})}{G_{\mathcal{M}^*}(L_{\mathcal{M}^*})}\times\frac{\tilde{G}_{\mathcal{M}^*}(L_{\mathcal{M}^*})}{G_{m,\mathcal{M}^*}(L_{\mathcal{M}^*})} \\
&\le \alpha\times E_1\times E_2\times E_3 := \alpha R(L_{\mathcal{M}^*}).
\end{aligned}$$

In the case that $L_{\mathcal{M}^*} = \infty$, the result holds by definition. Thus, we only consider the case that $L_{\mathcal{M}^*} \le \bar{t}_{m,\mathcal{M}^*}$. Using Eqs.(C.2)-(C.4) to get bounds for the terms E_1, E_2 and E_3, respectively, Eq.(7) is proved.

Then, for any $\epsilon > 0$,

$$\mathrm{FDR} \le (1+\epsilon)\alpha R(L_{\mathcal{M}^*}) + \mathbb{P}\left(\mathrm{FDP} \ge (1+\epsilon)\alpha R(L_{\mathcal{M}^*})\right),$$

which proves the second part of this theorem by using Assumption 3. □

E.2 Some Lemmas

We need the following lemmas to prove Theorem 1. The first lemma establishes the uniform convergence of $G_{n,\mathcal{M}}(t)$.

Lemma C.1 (Uniform consistency of the estimator with $\mathcal{U}$)**.** *Suppose Assumption 1-Assumption 3 hold. Denote $\bar{t}_{\mathcal{M}} = G^-_{\mathcal{M}}(A_n/(2n))$. For $\delta > 0$, we have with probability at least $1-\delta$,*

$$\sup_{\mathcal{M}\in\mathcal{G}}\sup_{0\le t\le\bar{t}_{\mathcal{M}}}\left|\frac{G_{n,\mathcal{M}}(t)}{G_{\mathcal{M}}(t)} - 1\right| \le C\sqrt{\frac{\varpi_m}{\delta A_n^{2/3}}}.$$

Proof. For any given $\mathcal{M}\in\mathcal{G}$, note that $G_{\mathcal{M}}(t)$ is a deceasing and continuous function of t. Let $a_n = A_n/2$, $z_0 \le z_1 \cdots \le z_{h_n} \le 1$, and $t_k = G^-_{\mathcal{M}}(z_k)$, where $z_0 = a_n/n$, $z_k = a_n/n + b_n\exp(k^\tau)/n$, $h_n = \{\log((n-a_n)/b_n)\}^{1/\tau}$ with $b_n/a_n \to 0$ and $0 < \tau < 1$. Note that $|G_{\mathcal{M}}(t_k)/G_{\mathcal{M}}(t_{k+1}) - 1| \le Cb_n/a_n$ uniformly holds in k. It is therefore enough to obtain the convergence of

$$D_{n,\mathcal{M}} = \sup_{0\le k\le h_n}\left|\frac{\sum_{i\in\mathcal{O}^c}\{\mathbb{I}(S_{i,\mathcal{M}} > t_k) - n_0G_{\mathcal{M}}(t_k)\}}{n_0G_{\mathcal{M}}(t_k)}\right|.$$

We calculate the second moment

$$\begin{aligned}
D_{\mathcal{M}}(t) :&= \mathbb{E}\left(\left[\sum_{i\in\mathcal{O}^c}\{\mathbb{I}(S_{i,\mathcal{M}} > t) - G_{\mathcal{M}}(t)\}\right]^2 \mid \mathcal{D}\right) \\
&= \sum_{i\in\mathcal{O}^c}\mathbb{E}\left[\{\mathbb{I}(S_{i,\mathcal{M}} > t) - G_{\mathcal{M}}(t)\}^2 \mid \mathcal{D}\right] \\
&= n_0G_{\mathcal{M}}(t)\{1 - G_{\mathcal{M}}(t)\} \\
&\le n_0G_{\mathcal{M}}(t).
\end{aligned} \tag{C.5}$$

From result (C.5) and Chebyshev's inequality, for $\epsilon > 0$ we have

$$\begin{aligned}
\mathbb{P}(D_{n,\mathcal{M}} > \epsilon \mid \mathcal{D}) &\le \sum_{k=0}^{h_n}\mathbb{P}\left(\left|\frac{\sum_{i\in\mathcal{O}^c}\{\mathbb{I}(S_{i,\mathcal{M}} > t_k) - G_{\mathcal{M}}(t_k)\}}{n_0G_{\mathcal{M}}(t_k)}\right| \ge \epsilon \mid \mathcal{D}\right) \\
&\le \frac{1}{\epsilon^2}\sum_{k=0}^{h_n}\frac{D_{\mathcal{M}}(t_k)}{\{n_0G_{\mathcal{M}}(t_k)\}^2} \\
&\le \frac{1}{\epsilon^2}\left\{\sum_{k=0}^{h_n}\frac{1}{n_0G_{\mathcal{M}}(t_k)}\right\}.
\end{aligned}$$

Moreover, observe that

$$\sum_{k=0}^{h_n} \frac{1}{n_0 G_{\mathcal{M}}(t_k)} = \frac{n}{n_0}\left(\frac{1}{a_n} + \sum_{k=1}^{h_n} \frac{1}{a_n + b_n e^{k^\tau}}\right) \le C b_n^{-1}.$$

By Assumption 3,

$$\mathbb{P}(\sup_{\mathcal{M}\in\mathcal{G}} D_{n,\mathcal{M}} > \epsilon) \le \sum_{\mathcal{M}\in\mathcal{G}} \mathbb{P}(D_{n,\mathcal{M}} > \epsilon) \le C\varpi_m/(\epsilon^2 b_n).$$

Hence, conditional on $\mathcal{D}$, we have with probability at least $1-\delta$,

$$\begin{aligned}\sup_{\mathcal{M}\in\mathcal{G}} \sup_{0\le t\le \bar{t}_{\mathcal{M}}} \left|\frac{G_{n,\mathcal{M}}(t)}{G_{\mathcal{M}}(t)} - 1\right| &\le C \max\left\{\sqrt{\varpi_m/(\delta b_n)}, b_n/a_n\right\} \\ &\le C\sqrt{\frac{\varpi_m}{\delta a_n^{2/3}}},\end{aligned}$$

where we take $b_n = a_n^{2/3}$. Finally, note that this holds uniformly in $\mathcal{D}$, and thus we complete the proof. □

Lemma C.2 (Uniform consistency of the Jackknife estimator)**.** *Suppose Assumption 1-Assumption 3 hold. Denote $\tilde{t}_{\mathcal{M}} = \tilde{G}^{-}_{\mathcal{M}}(A_n/2n)$. For $\delta > 0$, we have with probability at least $1-\delta$,*

$$\begin{aligned}&\sup_{\mathcal{M}\in\mathcal{G}} \sup_{0\le t\le \tilde{t}_{\mathcal{M}}} \left|\frac{\sum_i \mathbb{I}(S_{j,\mathcal{M}}^{[-j]} > t)}{m\tilde{G}_{\mathcal{M}}(t)} - 1\right| \\ &\quad \le C\delta^{-1}\left\{\varpi_m\left(\frac{n}{mA_n^{2/3}} + \frac{n}{A_n^{4/3}}\right)(1 + 15B_m + 2m^2 B_m)\right\}^{1/2}. \qquad \text{(C.6)}\end{aligned}$$

Proof. Define $G_{\mathcal{M}}^{[-j]}(t) = \mathbb{P}(S_{\mathcal{M}}^{[-j]}(Z) > t \mid \mathcal{D}^{[-j]})$, where $\mathcal{D}^{[-j]} = \mathcal{D} \setminus \tilde{X}_j$. Using the same partition technique in Lemma C.1, we need to obtain a bound for

$$D_{m,\mathcal{M}} = \sup_{0\le k\le h_n} \left|\frac{\sum_j \left\{\mathbb{I}(S_{j,\mathcal{M}}^{[-j]} > t_k) - \tilde{G}_{\mathcal{M}}(t_k)\right\}}{m\tilde{G}_{\mathcal{M}}(t_k)}\right|.$$

Again, the main step is to bound

$$D_{\mathcal{M}}(t) = \mathbb{E}\left[\sum_j \left\{\mathbb{I}(S_{j,\mathcal{M}}^{[-j]} > t) - \tilde{G}_{\mathcal{M}}(t)\right\}\right]^2.$$

By the construction of Jackknife estimators on $\mathcal{D}$, $\{1 \le j \le m : S_{j,\mathcal{M}}^{[-j]}\}$ are not independent with each other. Observe for each pair (j,k)

$$\begin{aligned}&\mathbb{E}\left\{\mathbb{I}(S_{j,\mathcal{M}}^{[-j]} > t)\mathbb{I}(S_{k,\mathcal{M}}^{[-k]} > t)\right\} \\ &= \mathbb{E}\left[\mathbb{E}\left\{\mathbb{I}(S_{j,\mathcal{M}}^{[-j]} > t)\mathbb{I}(S_{k,\mathcal{M}}^{[-k]} > t) \mid \mathcal{D}^{[-j,-k]}\right\}\right] \\ &= \mathbb{E}\left\{\mathbb{P}\left(S_{j,\mathcal{M}}^{[-j]} > t, S_{k,\mathcal{M}}^{[-k]} > t \mid \mathcal{D}^{[-j,-k]}\right)\right\} \\ &\le \mathbb{E}\left\{\left|\mathbb{P}\left(S_{j,\mathcal{M}}^{[-j]} > t, S_{k,\mathcal{M}}^{[-k]} > t \mid \mathcal{D}^{[-j,-k]}\right) - \mathbb{P}\left(S_{j,\mathcal{M}}^{[-j,-k]} > t, S_{k,\mathcal{M}}^{[-j,-k]} > t \mid \mathcal{D}^{[-j,-k]}\right)\right|\right\} \\ &\quad + \mathbb{E}\left\{\mathbb{P}\left(S_{j,\mathcal{M}}^{[-j,-k]} > t, S_{k,\mathcal{M}}^{[-j,-k]} > t \mid \mathcal{D}^{[-j,-k]}\right)\right\} \\ &:= Q_1 + Q_2.\end{aligned}$$

By the fact that $S_{j,\mathcal{M}}^{[-j,-k]}$ and $S_{k,\mathcal{M}}^{[-j,-k]}$ are independent conditional on $\mathcal{D}^{[-j,-k]}$, we have $Q_2 = \mathbb{E}\{G_{\mathcal{M}}^{[-j,-k]}(t)\}^2$. In Lemma C.5, we set $Y_1 = S_{j,\mathcal{M}}^{[-j]}$, $Y_2 = S_{k,\mathcal{M}}^{[-k]}$, $Y_3 = S_{j,\mathcal{M}}^{[-j,-k]}$, and $Y_4 = S_{k,\mathcal{M}}^{[-j,-k]}$. Under Assumption 2, we have

$$Q_1 \leq 4c_f\kappa_m + 6\zeta_m,$$

by applying Lemma C.5.

Accordingly,

$$\begin{aligned} D_{\mathcal{M}}(t) &:= \mathbb{E}\left[\sum_j \left\{\mathbb{I}(S_{j,\mathcal{M}}^{[-j]} > t) - G_{\mathcal{M}}^{[-j]}(t)\right\}\right]^2 + \mathbb{E}\left[\sum_j \left\{G_{\mathcal{M}}^{[-j]}(t) - \tilde{G}_{\mathcal{M}}(t)\right\}\right]^2 \\ &:= D_{1\mathcal{M}}(t) + D_{2\mathcal{M}}(t), \end{aligned}$$

where

$$\begin{aligned} D_{1\mathcal{M}}(t) &= \mathbb{E}\left[\sum_j G_{\mathcal{M}}^{[-j]}(t)\{1 - G_{\mathcal{M}}^{[-j]}(t)\}\right] \\ &\quad + \sum\sum_{j\neq k} \mathbb{E}\left[\left\{\mathbb{I}(S_{j,\mathcal{M}}^{[-j]} > t) - G_{\mathcal{M}}^{[-j]}(t)\right\}\left\{\mathbb{I}(S_{k,\mathcal{M}}^{[-k]} > t) - G_{\mathcal{M}}^{[-k]}(t)\right\}\right] \\ &\leq \mathbb{E}\left\{\sum_j G_{\mathcal{M}}^{[-j]}(t)\right\} + \sum\sum_{j\neq k} \mathbb{E}\left\{\mathbb{I}(S_{j,\mathcal{M}}^{[-j]} > t)\mathbb{I}(S_{k,\mathcal{M}}^{[-k]} > t) - G_{\mathcal{M}}^{[-j]}(t)G_{\mathcal{M}}^{[-k]}(t)\right\} \\ &\leq m\{\tilde{G}_{\mathcal{M}}(t) + B_m\} + \sum\sum_{j\neq k}\left[Q_1 + \mathbb{E}\{G_{\mathcal{M}}^{[-j,-k]}(t)\}^2 - \mathbb{E}\left\{G_{\mathcal{M}}^{[-j]}(t)G_{\mathcal{M}}^{[-k]}(t)\right\}\right] \\ &\leq m\{\tilde{G}_{\mathcal{M}}(t) + B_m\} + m(m-1)\left\{2B_m + 2\tilde{G}_{\mathcal{M}}(t)B_m + 3B_m^2\right\}, \end{aligned}$$

where we use Lemma C.3 to get

$$\left|\{G_{\mathcal{M}}^{[-j,-k]}(t)\}^2 - G_{\mathcal{M}}^{[-j]}(t)G_{\mathcal{M}}^{[-k]}(t)\right| \leq 2\tilde{G}_{\mathcal{M}}(t)B_m + 3B_m^2.$$

Similarly, we have

$$D_{2\mathcal{M}}(t) = 9m^2B_m^2.$$

With the above calculation and Chebyshev's inequality, for $\epsilon > 0$

$$\begin{aligned} \mathbb{P}(D_{m,\mathcal{M}} > \epsilon) &\leq \frac{1}{\epsilon^2}\sum_{k=0}^{h_n}\frac{D_{\mathcal{M}}(t_k)}{\{m\tilde{G}_{\mathcal{M}}(t_k)\}^2} \\ &\leq \frac{1}{\epsilon^2}\left\{\sum_{k=0}^{h_n}\frac{1 + 2m^2B_m}{m\tilde{G}_{\mathcal{M}}(t_k)} + \sum_{k=0}^{h_n}\frac{15B_m}{\tilde{G}_{\mathcal{M}}^2(t_k)}\right\}. \end{aligned}$$

Moreover, observe that

$$\begin{aligned} \sum_{k=0}^{h_n}\frac{1}{m\tilde{G}_{\mathcal{M}}(t_k)} &\leq Cn/(mb_n), \\ \sum_{k=0}^{h_n}\frac{1}{\tilde{G}_{\mathcal{M}}^2(t_k)} &\leq Cn/b_n^2, \end{aligned}$$

which implies that

$$\mathbb{P}(D_{m,\mathcal{M}} > \epsilon) \leq \frac{C}{\epsilon^2}\left(\frac{n}{mb_n} + \frac{n}{b_n^2}\right)(1 + 15B_m + 2m^2B_m).$$

By similar arguments in Lemmas C.1 and taking $b_n = a_n^{2/3}$, we can complete the proof. □

Lemma C.3. *Suppose Assumption 1 and Assumption 2 hold. Then we have*

(i) The following holds

$$\mathbb{E}\left\{\left|G_{\mathcal{M}}^{[-j]}(t) - G_{\mathcal{M}}(t)\right|\right\} \leq 2c_f\kappa_m + 3\zeta_m.$$

uniformly in $1 \leq j \leq m$.

(ii) It holds that $\mathbb{E}\left\{\left|G_{\mathcal{M}}(t) - \tilde{G}_{\mathcal{M}}(t)\right|\right\} \leq 4c_f\kappa_m + 6\zeta_m$.

Proof. (i) Observe

$$\begin{aligned}\left|G_{\mathcal{M}}(t) - G_{\mathcal{M}}^{[-j]}(t)\right| &= \left|\mathbb{P}(S_{\mathcal{M}}(Z) > t \mid \mathcal{D}) - \mathbb{P}(S_{\mathcal{M}}^{[-j]}(Z) > t \mid \mathcal{D}^{[-j]})\right| \\ &= \left|\mathbb{P}(S_{\mathcal{M}}(Z) > t \mid \mathcal{D}) - \mathbb{P}(S_{\mathcal{M}}^{[-i]}(Z) > t \mid \mathcal{D})\right|.\end{aligned}$$

The lemma follows immediately by using Assumption 2 and Lemma C.4 with $c_f > 0$.

(ii) By Lemma C.4, we have

$$\begin{aligned}|G_{\mathcal{M}}(t) - \tilde{G}_{\mathcal{M}}(t)| &\leq |G_{\mathcal{M}}(t) - \mathbb{P}(\tilde{\mu}_{\mathcal{M}}(Z) > t \mid \mathcal{D})| + |\mathbb{P}(\tilde{\mu}_{\mathcal{M}}(Z) > t) - \tilde{G}_{\mathcal{M}}(t)| \\ &\leq 4c_f\kappa_m + 3\zeta_m + 3\mathbb{P}(|S_{\mathcal{M}}(Z) - \tilde{\mu}_{\mathcal{M}}(Z)| > \kappa_m \mid \mathcal{D})\end{aligned}$$

and the result holds by taking expectation and using Assumption 2 again. □

Lemma C.4 (Estimation difference)**.** *Assume that two random variables* Y_1 *and* Y_2 *have density functions bounded by a constant* $c > 0$*, and satisfy* $\mathbb{P}(|Y_1 - Y_2| \leq \varepsilon) \geq 1 - \zeta$*, where* $\varepsilon > 0$ *and* $\zeta > 0$*. Then for all* $t > 0$*,* $|\mathbb{P}(Y_1 > t) - \mathbb{P}(Y_2 > t)| \leq 2c\varepsilon + 3\zeta$.

Proof. By the fact that

$$\mathbb{P}(Y_1 + \varepsilon > t, |Y_1 - Y_2| \leq \varepsilon) \leq \mathbb{P}(Y_2 > t, |Y_1 - Y_2| \leq \varepsilon) \leq \mathbb{P}(Y_1 - \varepsilon > t, |Y_1 - Y_2| \leq \varepsilon),$$

we have

$$\begin{aligned}&|\mathbb{P}(Y_1 > t) - \mathbb{P}(Y_2 > t)| \\ \leq&|\mathbb{P}(Y_1 > t) - \mathbb{P}(Y_2 > t, |Y_1 - Y_2| \leq \varepsilon)| + \mathbb{P}(|Y_1 - Y_2| > \varepsilon) \\ \leq&|\mathbb{P}(Y_1 > t) - \mathbb{P}(Y_1 > t - \varepsilon, |Y_1 - Y_2| \leq \varepsilon)| + |\mathbb{P}(Y_1 > t) - \mathbb{P}(Y_1 > t + \varepsilon, |Y_1 - Y_2| \leq \varepsilon)| \\ &+ \mathbb{P}(|Y_1 - Y_2| > \varepsilon) \\ \leq&|\mathbb{P}(Y_1 > t) - \mathbb{P}(Y_1 > t - \varepsilon)| + |\mathbb{P}(Y_1 > t) - \mathbb{P}(Y_1 > t + \varepsilon)| + 3\mathbb{P}(|Y_1 - Y_2| > \varepsilon) \\ \leq&2c\varepsilon + 3\zeta,\end{aligned}$$

where the last inequality comes from the Lipschitz condition. □

Lemma C.5. *Assume that four random variables* Y_1*,* Y_2*,* Y_3*, and* Y_4 *have density functions bounded by a constant* $c > 0$*, and satisfy* $\mathbb{P}(|Y_1 - Y_3| \leq \varepsilon) \geq 1 - \zeta$ *and* $\mathbb{P}(|Y_2 - Y_4| \leq \epsilon) \geq 1 - \varsigma$*, where* $\varepsilon > 0$*,* $\epsilon > 0$*,* $\zeta > 0$ *and* $\varsigma > 0$*. Then for all* $t > 0$*,*

$$|\mathbb{P}(Y_1 > t, Y_2 > t) - \mathbb{P}(Y_3 > t, Y_4 > t)| \leq 2c(\varepsilon + \epsilon) + 3(\zeta + \varsigma).$$

Proof. Define the events $\mathcal{E}_1 = \{|Y_1 - Y_3| \leq \varepsilon\}$, $\mathcal{E}_2 = \{|Y_2 - Y_4| \leq \epsilon\}$, and $\mathcal{E} = \mathcal{E}_1 \bigcup \mathcal{E}_2$. Note that $\mathbb{P}(Y_1 + \varepsilon > t, \mathcal{E}_1) \leq \mathbb{P}(Y_3 > t, \mathcal{E}_1) \leq \mathbb{P}(Y_1 - \varepsilon > t, \mathcal{E}_1)$ and $\mathbb{P}(Y_2 + \epsilon > t, \mathcal{E}_2) \leq \mathbb{P}(Y_4 > t, \mathcal{E}_2) \leq \mathbb{P}(Y_2 - \epsilon > t, \mathcal{E}_2)$. Then, we have

$$\begin{aligned}&\mathbb{P}(Y_1 > t, Y_2 > t) - \mathbb{P}(Y_3 > t, Y_4 > t) \\ &\leq |\mathbb{P}(Y_1 > t, Y_2 > t) - \mathbb{P}(Y_3 > t, Y_4 > t, \mathcal{E})| + \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c) \\ &\leq |\mathbb{P}(Y_1 > t, Y_2 > t) - \mathbb{P}(Y_1 - \varepsilon > t, Y_2 - \epsilon > t, \mathcal{E})| \\ &\quad + |\mathbb{P}(Y_1 > t, Y_2 > t) - \mathbb{P}(Y_1 + \varepsilon > t, Y_2 + \epsilon > t, \mathcal{E})| + \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c) \\ &\leq |\mathbb{P}(Y_1 > t, Y_2 > t) - \mathbb{P}(Y_1 - \varepsilon > t, Y_2 - \epsilon > t)| \\ &\quad + |\mathbb{P}(Y_1 > t, Y_2 > t) - \mathbb{P}(Y_1 + \varepsilon > t, Y_2 + \epsilon > t| + 3\mathbb{P}(\mathcal{E}_1^c) + 3\mathbb{P}(\mathcal{E}_2^c) \\ &\leq 2c(\varepsilon + \epsilon) + 3(\zeta + \varsigma),\end{aligned}$$

where we use the Lipschitz condition again. □