Variance function estimation in regression model via aggregation procedures

Ahmed Zaoui

Variance function estimation in regression model via aggregation procedures

Ahmed Zaoui[*]
LAMA, UMR-CNRS 8050,
Université Gustave Eiffel

Abstract

In the regression problem, we consider the problem of estimating the variance function by the means of aggregation methods. We focus on two particular aggregation setting: Model Selection aggregation (MS) and Convex aggregation (C) where the goal is to select the best candidate and to build the best convex combination of candidates respectively among a collection of candidates. In both cases, the construction of the estimator relies on a two-step procedure and requires two independent samples. The first step exploits the first sample to build the candidate estimators for the variance function by the residual-based method and then the second dataset is used to perform the aggregation step. We show the consistency of the proposed method with respect to the L^2-error both for MS and C aggregations. We evaluate the performance of these two methods in the heteroscedastic model and illustrate their interest in both the regression problem with reject option and the quantile regression.

Keywords: Regression, Conditional variance function, Aggregation

1 Introduction

Building efficient estimation of the level of noise is highly important for real applications and statistical analysis. In the heteroscedastic regression, which corresponds to the case where the variance of the errors depends on input variables, the heteroscedasticity must be detected and estimated. Indeed, not taking it into account in the estimation invalidates the conclusions of many statistical inference problems such as statistical tests which assume that the errors of the model all have the same variance. In addition, when using an approach that estimates the error variance as a function of input variables, the prediction intervals we obtain are likely to be more realistic than those obtained by assuming that the error variance is constant since the predictive uncertainty estimate depends on the estimate of the variance of the response variable. In general, testing and confidence intervals are two historical statistical problems where a bad calibration of the noise may lead to bad conclusions. Another important aspect of estimating the heteroskedasticity of the model is that the point estimate of the regression function is directly related to the variance function. The range of use of the variance structure in the data is even wider such as in selection the optimal kernel bandwidth (Fan, 1992), estimation correlation structure of the heteroscedastic spatial data (Opsomer et al., 1999), estimation of the quantile regression (Koenker, 2005, Shan and Yang, 2009), estimation of the signal-to-noise ratio (Verzelen and Gassiat, 2018), or choosing optimal design (Müller and Stadtmüller, 1987) and finding important applications for instance, in finance with the problems of measuring volatility or risk (Anderson and Lund, 1997, Xia et al., 2002) or long-term stock returns (Mammen et al., 2019) or time series context (Xu and Phillips, 2008). Many nonparametric economic models can be cast within the heteroscedastic regression model, for instance for application in stochastic frontier analysis (Martins-Filho and Yao, 2007) and for the analysis of causal treatment effects (Imbens and Lemieux, 2008). In our case, we highlight the interest in providing an efficient estimation of the variance function in the problem of regression with regret option where the good calibration of the rejection rule is highly dictated by the efficiency of the estimator of the noise level (Denis et al., 2020). We focus on the regression problem: we denote by $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ the couple

[*] Ahmed.Zaoui@univ-eiffel.fr

of random variables where X is the feature vector and Y is the response variable such that

$$Y = f^*(X) + W \ .$$

Here W is the noise and is such that $\mathbb{E}[W|X] = 0$ and $\mathbb{E}[W^2] < \infty$. In particular for any $x \in \mathbb{R}^d$ we write $f^*(x) = \mathbb{E}\left[Y|X = x\right]$ to denote the regression function and $\sigma^2(x) = \mathrm{Var}(Y|X = x) = \mathbb{E}\left[(Y - f^*(X))^2|X = x\right]$ to denote the conditional variance function.

Despite the popularity of the problem of estimating the noise level, there remains much to do. In particular, we study in the present paper this problem from the aggregation perspective and build estimators of the conditional variance function based on Model Selection (MS) and Convex (C) aggregations. We study their consistency properties as well as their numerical performances. Our work is motivated by recent research in regression with reject option (Denis et al., 2020). There it has been observed that the rejection rule is fully determined by the variance structure. We hope that aggregation will improve the accuracy of the method.

1.1 Related work

Our literature review consists of three related fields:

Conditional variance estimation: The problem of estimating the regression function is classical and widely studied, see for example (Biau and Devroye, 2015, Gyorfi et al., 2002, Scornet et al., 2015, Stone, 1977, Tsybakov, 2008) and references therein.

Even though the problem of estimation of the conditional variance function is less studied, it has been considered in several works that can be cast into two groups according to the nature of the design (fixed or random).

When the design is fixed, the estimation of σ^2 has been studied mainly via residual-based methods (Hall and Carroll, 1989, Härdle and Tsybakov, 1997, Ruppert et al., 1997) and difference-based methods (Brown and Levine, 2007, Müller and Stadtmüller, 1987, Wang et al., 2008). Difference-based estimators do not require the estimation of the regression function f^*. The first difference-based estimators have been developed by Müller and Stadtmüller (1987). They considered an initial variance estimates which are squared weighted sums of m observations neighbouring the fixed point where the variance function is to be estimated. The authors showed that the proposed initial variance estimates are not consistent. To solve this problem, they smoothed them with a kernel estimate. Brown and Levine (2007) presented a class of non-parametric variance estimators based on different sequences and local polynomial estimation and established asymptotic normality. Wang et al. (2008) were interested in the effect of the unknown mean on the estimation of the variance function and proved numerically that the residual-based method performs better than the first-order-difference-based estimator when the unknown regression function f^* is very smooth.

In this work, we rather focus on random design. Less methods have been proposed to estimate the conditional variance function in this case. Most classical methods are the direct and the residual-based:

1. **The direct method**: a simple decomposition the conditional variance function σ^2 is rewritten as the difference of the first two conditional moments, $\sigma^2(x) = \mathbb{E}[Y^2|X = x] - (\mathbb{E}[Y|X = x])^2$. The direct method consists in estimating the two terms in the right side separately, see for example (Devroye et al. (2018), Fan and Yao (1998), Härdle and Tsybakov (1997)). To be more specific, the direct estimator of σ^2 has the following form

$$\hat{\sigma}^2_d(x) = \hat{g}(x) - (\hat{f}(x))^2 \ ,$$

   where $\hat{g}$ and $\hat{f}$ are estimators of $\mathbb{E}[Y^2|X = x]$ and f^*, respectively. The main drawback of this approach is that it is not always nonnegative for example, if different smoothing parameters are used in estimating those terms and adaptation to the unknown regression function f^* is still not available. Härdle and Tsybakov (1997) proposed a local polynomial regression estimates of those terms using the same bandwidth and the same kernel function. They established the asymptotic normality of local polynomial estimators of the regression function and the variance function.

2. **The residual-based method**: this approach consists of two steps. First, one estimates the regression function and computes the squared residuals $\hat{r} = (Y - \hat{f}(X))^2$ where $\hat{f}$ is the estimator

of f^*. Second, we estimate the variance function by solving the regression problem when the input is the feature X and the output variable is the residuals $\hat{r}$. For more details, see Fan and Yao (1998), Neumann (1994), Ruppert et al. (1997). It exists many ways to study this method. For instance, Fan and Yao (1998) applied a local linear regression in both steps and showed that their estimator is fully regression-adaptive to the unknown regression function. Using the local polynomial regression can be negative when the bandwidths are not selected appropriately. As a solution to this, Yu and Jones (2004) proposed estimators based on a localised normal likelihood, using a standard local linear form for estimating the mean and a local log-linear form for estimating the variance. Ziegelmann (2002) introduced an exponential estimator of the conditional variance in the second step to ensure the nonnegativity of the estimator of σ^2. Xu and Phillips (2011) used a reweighted local constant estimator (kernel estimator) based on maximising the empirical likelihood subject to a bias-reducing moment restriction. Moreover, such estimators have the form $\hat{\sigma}^2(X) = \sum_i \omega_i(X)(Y_i - \hat{f}(X_i))^2$ where $\omega_i(X)$ are weight functions (Denis et al. (2020), Kulik and Wichelhaus (2011)). Recently, Denis et al. (2020) used the previous estimator and focused on estimating the regression function and the variance function respectively, by kNN. Under mild assumptions, they provided the rate of convergence of the kNN estimator of the conditional variance function in supremum norm. The residual-based method can be regarded as a generalised difference-based estimator. For more details, see Fan and Yao (1998). Another line of work (see Evans and Jones (2008), Ferrario and Walk (2012), Liitiäinen et al. (2010, 2009)) have focused on nearest-neighbour-based estimators of $\mathbb{E}[\sigma^2(X)]$ and have analysed the properties of such an estimator both theoretically and numerically for various machine learning problems.

In this paper, we focus on the residual-based method to estimate the variance function since they appear more tractable. In particular, we develop an aggregation procedures for this task.

Aggregation methods: Aggregation is a popular approach in statistics and machine learning. This technique is well known to estimate the regression function in the homoscedastic or heteroscedastic model. We refer the reader to the baseline articles Audibert (2009), Bunea et al. (2007), Juditsky and Nemirovski (2000), Tsybakov (2003, 2014), Yuhong (2004). Given a set of estimators of regression function f^*, the aggregation constructs a new estimator, called the aggregate, which mimics, in a certain sense the behaviour of the best estimator in a class of estimates. There are several popular types of aggregation and we focus on two: the model selection aggregation ($\texttt{MS}$) which allows to select the best estimator from the set; the convex aggregation ($\texttt{C}$) where the goal is to select the best convex combination of functions in the set. In general, the aggregation procedures are based on sample splitting, that is, the original data set $\mathcal{D}_N$ is split into two independent data sequences $\mathcal{D}_k$ and $\mathcal{D}_l$ with $N = l + k \geq 1$. The first subsample $\mathcal{D}_k$ is exploited to build $M > 1$ competing estimators of the regression function f^* and $\mathcal{D}_l$ is used to aggregate them. Most of the work has focused on fixing the first sample, resulting in fixed estimators (the estimators are then seen as fixed functions). Under mild assumptions, the authors in Tsybakov (2003) showed that the optimal rates for $\texttt{MS}$ and $\texttt{C}$ aggregation *w.r.t.* L^2-error in gaussian regression model are of order $\frac{\log(M)}{N}$, and $\frac{M}{N}$ if $M \leq \sqrt{N}$, respectively, $\sqrt{\frac{1}{N}\log\left(\frac{M}{\sqrt{N}}+1\right)}$ if $M \geq \sqrt{N}$ in both cases.

In this paper, we consider aggregation methods for the conditional variance estimation. Up to our knowledge, such approaches have not been considered yet for this problem.

Reject option: Reject option is important in nonparametric statistic since it helps avoid uncertain prediction. It has been initially introduced in the classification setting (Chow, 1957, 1970, Denis and Hebiri, 2020, Herbei and Wegkamp, 2006, Lei, 2014, Nadeem et al., 2009, Vovk et al., 2005) where it has shown important improvements in the quality of prediction. It has been recently developed in the case of the regression model in the case where the rejection rate is controlled by the practitioner (see Denis et al. (2020)). There the authors provided a characterisation of the optimal rule (knowing the true distribution of the data) and demonstrated that it relies on thresholding the conditional variance function. More formally, it is defined as follows: given a rejection rate $\varepsilon \in (0, 1)$

$$\Gamma^*_\varepsilon(x) := \begin{cases} \{f^*(x)\} & \text{if } \sigma^2(x) \leq F^{-1}_{\sigma^2}(1-\varepsilon) \\ \emptyset & \text{otherwise} \end{cases},$$

where $F_{\sigma^2}^{-1}$ is the generalised inverse of the cumulative distribution function (CDF) F_{σ^2} of $\sigma^2(X)$. As can be observed, this optimal solution depends on several parameters: the rejection rate ε that is known in advance, the CDF F_{σ^2} that is efficiently estimated the empirical CDF, the regression function $f^*(x)$ for which efficient estimators exist in the literature, and the conditional variance σ^2. This last quantity is less considered in the literature and our goal is to build accurate estimators of the conditional variance that rely on aggregation. The ultimate purpose is then to make a sharper estimation of the optimal rule in the case of the rejection option in the regression setting.

1.2 Main contribution

We develop the notions of model selection aggregation and convex aggregation to estimate the conditional variance function. To our knowledge, this work is the first to deal with this setting. We consider two independent datasets: the first will be used to build the initial estimators of the variance function and the second to aggregate them. We call these estimators the `MS`-estimator and `C`-estimator. We consider the residual-based method to build the initial estimators which is based on estimating the regression function in the first step. We focus on estimating the regression function by model selection aggregation and convex aggregation. In this paper, the major part is then devoted to show the upper-bounds of L^2-error of the `MS`-estimator and `C`-estimator when the initial estimators can be arbitrary or verify very weak conditions such that boundeness. We establish that the rate of convergence for `MS` and `C` procedures is of order $O((\log(M_1)/N)^{1/8})$ when Y is unbounded and is of order $O((\log(M_1)/N)^{1/4})$ when Y is bounded. Finally, we obtain numerical results which show the performance of our procedures.

1.3 Outline

The paper is organised as follows. In the next section, the aggregation problems, the model selection and convex aggregations, are described in detail. Section 3 is focused on investigating the upper-bounds for the L^2-error of our procedures. Finally, Section 4 presents a numerical comparison of the proposed method w.r.t. the heteroscedastic model as well as a direct application to both the regression framework with reject option and the quantile regression.

Notations. We introduce some notation that is used throughout this paper. Let $p \geq 2$ be an integer, the set of integers $\{1, \ldots, p\}$ is denoted $[p]$. Let N be integer. For any function $f : \mathbb{R}^d \to \mathbb{R}$, we define the empirical norm $\|f\|_N^2 = \frac{1}{N}\sum_{i=1}^N |f(X_i)|^2$ and the supremum norm $\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)|$. Moreover, we denote by $\Lambda^p := \{\lambda \in \mathbb{R}^p : \lambda_j \geq 0, \sum_{j=1}^p \lambda_j = 1\}$ the simplex. Let $\|\cdot\|_{1,p}$ denote the ℓ_1 norm on $\mathbb{R}^p$, that is, $\|\lambda\|_{1,p} := \sum_{j=1}^p |\lambda_j|$. For the sake of simplicity, let $Z = (Y - f^*(X))^2$.

2 Aggregation estimators

In this section, we describe an estimation algorithm of the variance function σ^2 by aggregation. In particular, we focus on two types of aggregations: the model selection aggregation (`MS`), and the convex aggregation (`C`). These aggregation problems, (`MS`) and (`C`), have been considered to estimate the unknow regression function in the regression model. The objective is to estimate f^* by a combination of elements of a known set called dictionary made up of deterministic functions or preliminary estimators. The collection of estimators or algorithms is given and can be parametric, nonparametric or semi-parametric nature. Given a set of estimators, the `MS`-aggregation consists in constructing a new estimator which is approximately as good as the best estimator in the set, while the objective of `C`-aggregation is to construct a new estimator which is approximately as good as the best convex combination of the elements in the set, for more details see Audibert (2009), Bunea et al. (2007), Juditsky and Nemirovski (2000), Tsybakov (2003, 2014), Yuhong (2004). Besides, to construct an aggregate of σ^2, we first introduce two independent learning samples: $\mathcal{D}_n = \{(X_i^{'}, Y_i^{'}), i = 1, \ldots, n\}$ and $\mathcal{D}_N = \{(X_i, Y_i), i = 1, \ldots, N\}$ which consist of respectively, n and N i.i.d. copies of (X, Y).

2.1 Model selection aggregation

In this first paragraph, we detail how we perform MS-aggregation in order to estimate of the conditional variance function σ^2 by MS. It consists of two steps: one step for the estimation of the regression function f^* and a second one devoted to the estimation of σ^2. More precisely, in the first one we build M_1 estimators of the regression function $\hat{f}_1, \ldots, \hat{f}_{M_1}$ based on $\mathcal{D}_n$ with $2 \leq M_1 < \infty$. Then, we use the second dataset $\mathcal{D}_N$ to estimate f^* by MS: we select the optimal index, denoted $\hat{s}$ as follows

$$\hat{s} \in \underset{s \in [M_1]}{\operatorname{argmin}} \hat{\mathcal{R}}_N(\hat{f}_s), \text{ where } \hat{\mathcal{R}}_N(\hat{f}_s) = \frac{1}{N} \sum_{i=1}^{N} |Y_i - \hat{f}_s(X_i)|^2 \ , \tag{1}$$

and the aggregate of the regression function, denoted by $\hat{f}_{\texttt{MS}}$, is then given by

$$\hat{f}_{\texttt{MS}} := \hat{f}_{\hat{s}}. \tag{2}$$

In the second step, given the estimator $\hat{f}_{\texttt{MS}}$ built on $\mathcal{D}_N$, we construct using back the sample $\mathcal{D}_n$ M_2 estimators of the variance function σ^2, denoted $\hat{\sigma}^2_{\hat{s},1}, \ldots, \hat{\sigma}^2_{\hat{s},M_2}$, by residual-based method with $2 \leq M_2 < \infty$. Finally, based on $\mathcal{D}_N$ again, we select the optimal single, denoted $\hat{m}$, as follows

$$\hat{m} \in \underset{m \in [M_2]}{\operatorname{argmin}} \hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) \text{ where } \hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) = \frac{1}{N} \sum_{i=1}^{N} |\hat{Z}_i - \hat{\sigma}^2_{\hat{s},m}(X_i)|^2$$

with $\hat{Z}_i = \left(Y_i - \hat{f}_{\texttt{MS}}(X_i)\right)^2$. Therefore, the aggregate of the variance function, called MS-estimator and denoted $\hat{\sigma}^2_{\texttt{MS}}$, is defined as

$$\hat{\sigma}^2_{\texttt{MS}} := \hat{\sigma}^2_{\hat{s},\hat{m}}. \tag{3}$$

2.2 Convex aggregation

Convex aggregation procedures for nonparametric regression are discussed in Audibert (2004), Bunea et al. (2007), Tsybakov (2003). We describe here an algorithm for aggregating estimates of the conditional variance function σ^2 by C-aggregation. As for MS-aggregation, the construction of the aggregate of σ^2 needs two independent datasets $\mathcal{D}_n$ and $\mathcal{D}_N$. The aggregation still proceeds in two steps: one for estimating f^* and the second for the estimation of σ^2. Each step is again decomposed in two parts. Firstly, we consider M_1 estimators of the regression function f^*, $\{\hat{f}_1, \ldots, \hat{f}_{M_1}\}$, based on $\mathcal{D}_n$, and for any $\lambda \in \Lambda^{M_1}$ we define the linear combinations $\hat{f}_\lambda$

$$\hat{f}_\lambda = \sum_{j=1}^{M_1} \lambda_j \hat{f}_j.$$

Then, aggregates of the regression function based on the sample $\mathcal{D}_N$ have the form

$$\hat{f}_{\texttt{C}} := \hat{f}_{\hat{\lambda}} = \sum_{j=1}^{M_1} \hat{\lambda}_j \hat{f}_j \ , \tag{4}$$

where the estimator $\hat{\lambda}$ is defined by

$$\hat{\lambda} \in \underset{\lambda \in \Lambda^{M_1}}{\operatorname{argmin}} \hat{\mathcal{R}}_N(\hat{f}_\lambda).$$

Once $\hat{f}_{\texttt{C}}$ is obtained, we focus on the estimation of σ^2. Based on the sample $\mathcal{D}_n$, we build M_2 estimators for the conditional variance function by the residual-based method, denoted $\hat{\sigma}^2_{\hat{\lambda},1}, \ldots, \hat{\sigma}^2_{\hat{\lambda},M_2}$, and for any $\beta \in \Lambda^{M_2}$ we define $\hat{\sigma}^2_{\hat{\lambda},\beta}$ as follows

$$\hat{\sigma}^2_{\hat{\lambda},\beta} = \sum_{j=1}^{M_2} \beta_j \hat{\sigma}^2_{\hat{\lambda},j} \ .$$

Finally, based on $\mathcal{D}_N$, the aggregate estimate for σ^2 is given by

$$\hat{\sigma}^2_{\mathtt{C}} := \hat{\sigma}^2_{\hat{\lambda},\hat{\beta}} \ , \tag{5}$$

where the estimator $\hat{\beta}$ is defined by

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \Lambda^{M_2}} \hat{R}_N(\hat{\sigma}^2_{\hat{\lambda},\beta}), \ \text{ where } \ \hat{R}_N(\hat{\sigma}^2_{\hat{\lambda},\beta}) = \frac{1}{N}\sum_{i=1}^{N} |\hat{Z}_i - \hat{\sigma}^2_{\hat{\lambda},\beta}(X_i)|^2$$

with $\hat{Z}_i = (Y_i - \hat{f}_{\mathtt{C}}(X_i))^2$. We called $\hat{\sigma}^2_{\mathtt{C}}$ the C-estimator.

3 Main results

This section is devoted to studying the L^2-error of MS-estimator and C-estimator. Firstly, we introduce general conditions required on the model in Section 3.1. Secondly, we show the consistency of our methods in Sections 3.2 and 3.3.

3.1 Assumptions

The following assumptions are the bedrock of our theoretical analysis:

Assumption 1. *The functions f^* and σ^2 are bounded.*

Assumption 2. *Y is bounded or Y satisfies the gaussian model*

$$Y = f^*(X) + \sigma(X)\xi, \tag{6}$$

where ξ is independent of X and distributed according to a standard normal distribution.

These assumptions are relatively weak and play a key role in our approach. They allow to use the Hoeffding's inequality in the case of boundness of Y or ξ. In particular, it is important to emphasise that Assumptions 1 and 2 guarantee that the variable $Y - f^*(X)$ is sub-Gaussian (see Lemma 9 in the case where Y is bounded).

3.2 Upper bound for $\hat{\sigma}^2_{\mathtt{MS}}$

We study the L^2-error of the MS-estimator $\hat{\sigma}^2_{\mathtt{MS}}$. Let $\mathcal{R}(\hat{f}_s) = \mathbb{E}\left[|Y - \hat{f}_s(X)|^2\right]$ for all $s \in [M_1]$. We define s^* as follows

$$s^* \in \operatorname*{argmin}_{s\in[M_1]} \mathcal{R}(\hat{f}_s) \ . \tag{7}$$

Besides, we need the following assumptions in the case of MS-aggregation:

Assumption 3. *For all $s \in [M_1]$ and all $m \in [M_2]$, $\hat{f}_s$ and $\hat{\sigma}^2_{s,m}$ are bounded a.s $\mathcal{D}_n$. More precisely, there exist two positive constants K_1 and K_2 such that for all $n \in \mathbb{N}^*$*

$$\max_{s\in[M_1]} \|\hat{f}_s\|_\infty \le K_1, \ \ \textit{and} \ \ \max_{(s,m)\in[M_1]\times[M_2]} \|\hat{\sigma}^2_{s,m}\|_\infty \le K_2.$$

Assumption 4 (Separability hypothesis)**.** *There exists $\delta_0 > 0$ such that*

$$\delta^*\left(\mathcal{D}_n\right) = \min_{s\neq s^*}\{|\mathcal{R}(\hat{f}_s) - \mathcal{R}(\hat{f}_{s^*})|\} > \delta_0 \ .$$

Both assumptions are used to control the L^2-error of the MS-estimator $\hat{\sigma}^2_{\mathtt{MS}}$. Assumption 3 describes the boundedness of the estimators. It is in particular satisfied if the functions in the dictionaries are bounded. In our construction, the constants K_1 and K_2 do not need to be known. In practice, the response variable Y in the sample is finite and then it ensures that the candidates in the dictionaries are bounded. Assumption 4 is used for MS and helps us to ensure that the minimum of the risk $\mathcal{R}$ is well defined. It cannot be verified in practice since it depends on the distribution $\mathbb{P}$. Let $\mathbb{E}$ be the expectation which is taken with respect to both X and the samples $\mathcal{D}_n$ and $\mathcal{D}_N$. We establish the following result

Theorem 1. *Let $\hat{f}_{\mathtt{MS}}$ and $\hat{\sigma}^2_{\mathtt{MS}}$ be two* MS*-estimators of f^* and σ^2 defined in Eq.* (2) *and* (3) *respectively. Then, under Assumptions 1- 4, there exist two absolute constants $C_1 > 0$ and $C_2 > 0$ such that*

$$\mathbb{E}\left[|\hat{\sigma}^2_{\mathtt{MS}}(X) - \sigma^2(X)|^2\right] \leq \mathbb{E}\left[\min_{m\in[M_2]} \mathbb{E}_X\left[|\hat{\sigma}^2_{s^*,m}(X) - \sigma^2(X)|^2\right]\right] + C_1\left\{\min_{s\in[M_1]} \mathbb{E}\left[\|\hat{f}_s - f^*\|^2_N\right]\right\}^{1/2p} + C_2\left(\frac{\log(M_1)}{N}\right)^{1/4p},$$

with $p = 1$ if Y is bounded and $p = 2$ otherwise.

The proof of this result is postponed to the Appendix. Let's give a sketch of the proof. The L^2-error is the exces risk of $\hat{\sigma}^2_{\mathtt{MS}}$ where $\mathbb{E}\left[|\hat{\sigma}^2_{\mathtt{MS}}(X) - \sigma^2(X)|^2\right] := \mathbb{E}\left[R(\hat{\sigma}^2_{\mathtt{MS}}) - R(\sigma^2)\right]$ with $R(\sigma^2) = \mathbb{E}\left[|Z - \sigma^2(X)|^2\right]$. We introduce the minimiser $\bar{\sigma}^2_{\mathtt{MS}} := \hat{\sigma}^2_{\hat{s},\bar{m}}$ where $\bar{m} \in \operatorname{argmin}_{m\in[M_2]} R(\hat{\sigma}^2_{\hat{s},m})$. We consider the decomposition $\mathbb{E}\left[|\hat{\sigma}^2_{\mathtt{MS}}(X) - \sigma^2(X)|^2\right] = \mathbb{E}\left[R(\hat{\sigma}^2_{\mathtt{MS}}) - R(\bar{\sigma}^2_{\mathtt{MS}})\right] + \mathbb{E}\left[R(\bar{\sigma}^2_{\mathtt{MS}}) - R(\sigma^2)\right]$. We control the two terms in the right side separately. The first one is the estimation error (variance term). To control it, we need to introduce $\tilde{\sigma}^2_{\mathtt{MS}} := \hat{\sigma}^2_{\hat{s},\tilde{m}}$ where $\tilde{m} \in \operatorname{argmin}_{m\in[M_2]} R_N(\hat{\sigma}^2_{\hat{s},m})$, with $R_N(\hat{\sigma}^2_{\hat{s},m}) = \frac{1}{N}\sum_{i=1}^N |Z_i - \hat{\sigma}^2_{\hat{s},m}(X_i)|^2$. The upper bound of the variance depends on the L^2-error of the aggregate $\hat{f}_{\mathtt{MS}}$ with respect to the empirical norm. The second one is the approximation error. Its upper-bound is linked to $\mathbb{P}(\hat{s} \neq s^*)$.

Theorem 1 gives the upper-bound for L^2-error of $\hat{\sigma}^2_{\mathtt{MS}}$. This bound consists of two parts: the first part is the bias of MS-estimator $\hat{\sigma}^2_{\mathtt{MS}}$ and depends on the deterministic selector s^*; the second part is composed of the two remaining terms and corresponds to the estimation error (variance). The first term is the bias term of $\hat{f}_{\mathtt{MS}}$ expressed in terms of the empirical norm $\|\cdot\|^2_N$, and the second one characterises the price to pay for MS-aggregation and is of order $(\log(M_1)/N)^{1/4p}$ where $p = 1$ if Y is bounded and $p = 2$ otherwise. Note that this rate is slower than in the case of the estimation of the regression function f^*. This slow rate is due to the double aggregation that we need to perform for the estimation of the conditional variance function.

3.3 Upper bound for $\hat{\sigma}^2_{\mathtt{C}}$

In this part, we focus in studying the L^2-error of $\hat{\sigma}^2_{\mathtt{C}}$. The construction of $\hat{\sigma}^2_{\mathtt{C}}$ needs the following estimators $\{\hat{f}_i\}_{i=1}^{M_1}$ and $\{\hat{\sigma}_{\lambda,i}\}_{i=1}^{M_2}$. We require the following conditions

Assumption 5. *For all $i \in [M_1]$, all $\lambda \in \Lambda^{M_1}$ and all $j \in [M_2]$, $\hat{f}_i$ and $\hat{\sigma}^2_{\lambda,j}$ are bounded a.s. $\mathcal{D}_n$.*

Assumption 6. *Suppose that there exists a constant $K \geq 0$ such that for every $j \in [M_2]$*

$$\mathbb{E}\left[|\hat{\sigma}^2_{\lambda_1,j}(X) - \hat{\sigma}^2_{\lambda_2,j}(X)|\right] \leq K\|\lambda_1 - \lambda_2\|_{1,M_1}, \quad \forall \lambda_1, \lambda_2 \in \Lambda^{M_1} \;\; a.s\; \mathcal{D}_n.$$

Assumption 5 describes the boundedness of the candidates as Assumption 3. Assumption 6 is a strong condition. However, it holds, for instance, for estimators of the form $\hat{\sigma}^2_{\lambda,j}(X) = \sum_i \omega_i(X)(Y_i - \hat{f}_\lambda(X_i))^2$ where $\omega_i(X)$ are weight functions, that are nonnegative and sum to one. The next theorem is the main result of this section, it displays the upper-bound of L^2-error for $\hat{\sigma}^2_{\mathtt{C}}$.

Theorem 2. *Let $\hat{f}_{\mathtt{C}}$ and $\hat{\sigma}^2_{\mathtt{C}}$ be two* C*-estimators of f^* and σ^2 defined in Eq.* (4) *and* (5) *respectively. Then, under Assumptions 1, 2, 5, and 6, there exist two absolute constants $C_1 > 0$ and $C_2 > 0$ such that*

$$\mathbb{E}\left[|\hat{\sigma}^2_{\mathtt{C}}(X) - \sigma^2(X)|^2\right] \leq \mathbb{E}\left[\inf_{\beta\in\Lambda^{M_2}} \mathbb{E}_X\left[|\hat{\sigma}^2_{\hat{\lambda},\beta}(X) - \sigma^2(X)|^2\right]\right] + C_1\left\{\inf_{\lambda\in\Lambda^{M_1}} \mathbb{E}\left[\|\hat{f}_\lambda - f^*\|^2_N\right]\right\}^{1/2p} + C_2\left(\frac{\log(M_1)}{N}\right)^{1/4p},$$

with $p = 1$ if Y is bounded and $p = 2$ otherwise.

As for Theorem 1, the upper-bound for the L^2-error of C-estimator $\hat{\sigma}^2_{\mathtt{C}}$ is composed of three terms. The first one is the bias term of $\hat{\sigma}^2_{\mathtt{C}}$ which depends on the random selector $\hat{\lambda}$, the second and third ones is a bound of the variance term that rely on the bias term of $\hat{f}_{\mathtt{C}}$ with respect to the empirical norm $\|\cdot\|^2_N$

and on the price to pay for convex aggregation which is of order $(\log(M_1)/N)^{1/4p}$ where $p = 1$ if Y is bounded and $p = 2$ otherwise.

We notice that both procedures $\texttt{MS}$ and $\texttt{C}$ have the same rate. Indeed, the variance term of $\hat{\sigma}^2_{\texttt{MS}}$ and $\hat{\sigma}^2_{\texttt{C}}$ is based on the upper bound for $\hat{f}_{\texttt{MS}}$ and $\hat{f}_{\texttt{C}}$. Moreover, the aggregates $\hat{f}_{\texttt{MS}}$ and $\hat{f}_{\texttt{C}}$ have the same rate which is of order $(\log(M_1)/N)^{1/2}$ with respect to the empirical norm $\|\cdot\|^2_N$, see Proposition 1 and Proposition 2 in the Appendix. Let us now compare with the rates of $\hat{f}_{\texttt{MS}}$ and $\hat{f}_{\texttt{C}}$ with respect to L^2-error and L^2-risk. For the Gaussian and bounded regression model, the rate of the variance term of $\hat{f}_{\texttt{MS}}$ and $\hat{f}_{\texttt{C}}$ is of order $\frac{\log(M_1)}{N}$ and $\frac{M_1}{N}$ if $M_1 \leq \sqrt{N}$, respectively, $\sqrt{\frac{1}{N}\log\left(\frac{M_1}{\sqrt{N}}+1\right)}$ if $M_1 \geq \sqrt{N}$ in both cases (see Bunea et al., 2007, Lecué, 2013, Lecué and Mendelson, 2009, Tsybakov, 2003). We can deduce that our rates are very slow because our procedures need to estimate at the same time the unknown regression function f^* and the variance function σ^2 by aggregation procedures.

4 Numerical results

This section is devoted to the numerical analysis of our procedures. In Section 4.1, we describe four heteroscedastic models in the gaussian case and two models when Y is bounded. Second, we evaluate the performances of $\texttt{MS}$-estimator and $\texttt{C}$-estimator for different examples in Section 4.2. Once we have calibrated our estimate of the variance function σ^2, we exploit it to consider the problem of regression with reject option and the quantile regression in Section 4.4.

4.1 Data

Our numerical study relies on synthetic data:
Heteroscedastic models: we propose four examples of heteroscedastic models (6):

- Model 1: let $a \in \{1/4, 1\}$ and $X = (X_1, X_2, X_3)$ have a uniform distribution on $[0,1]^3$. Let
  1. $f^*(X) = 0.1\cos(X_1) + \exp(-X_3^2)$;
  2. $\sigma^2(X) = a\left(0.1 + \exp(-7(X_1-0.2)^2) + \exp(-10(X_2-0.5)^2 + \exp(-50(X_3-0.9)^2)\right)$.
- Model 2: let $X = (X_1, \ldots, X_{10})$ have a uniform distribution on $[0,1]^{10}$. We define
  1. $f^*(X) = 0.1 + \exp(-X_1^2) + 0.2\sin(X_2 + X_3 + X_4 + 0.1X_5^2)$;
  2. $\sigma^2(X) = \frac{1}{2}(0.5 + \sqrt{X_1(1-X_2)} + 0.8X_3X_4 + X_5X_6X_7^2 + 0.9\exp(-500(X_8+X_9+X_{10}-0.5)^2))^2$.
- Model 3: sparse model
  1. $f^*(X) = X\beta, \quad \beta \in \mathbb{R}^p$;
  2. $\sigma^2(X) = \frac{1}{2}\left(0.3 + \sqrt{X_1(1-X_1)}\sin\left(\frac{2.1\pi}{X_2+0.05}\right) + 0.5X_3 + X_4\right)^2$.

Bounded Y: we consider the following regression model when Y is bounded

$$Y = f^*(X) + \sigma(X)\varrho$$

where ϱ have a uniform distribution on $[-\sqrt{3}, \sqrt{3}]$. We give the following examples of models :

- Model 4: let X have a uniform distribution on $[0,1]^2$ and
  1. $f^*(X) = X_1 + \exp(-X_2^2)$;
  2. $\sigma^2(X) = 0.01 + X_1\exp\left(-(X_2-0.9)^2\right)$.
- Model 5: let $X = (X_1, X_2, X_3)$ have a uniform distribution on $[0,1]^3$ and
  1. $f^*(X) = X_1 + X_2 + 0.5\cos(X_3)$;
  2. $\sigma^2(X) = \left(0.3 + \sqrt{X_1(1-X_1)}\sin\left(\frac{(2.1)\pi}{X_2+0.05}\right) + X_3\right)^2$.

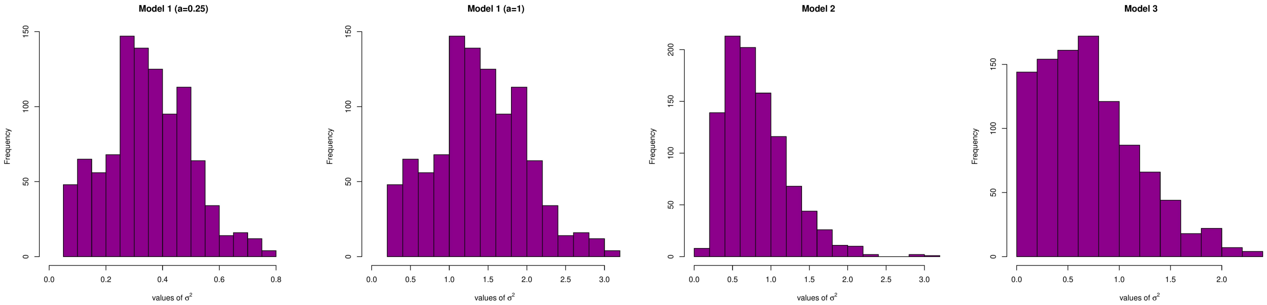


Figure 1: Histogram of values of σ^2 in Gaussian models.

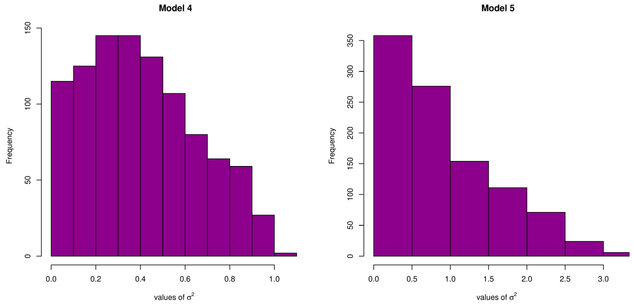


Figure 2: Histogram of values of σ^2 in regression models for bounded responses.

We describe the previous models. We display in Figures 1 and 2 the histograms of the variance function for every model. Model 1 is a multivariate model in which the regression and variance functions are regular functions. In the case $a = 1/4$, the problem of estimation of σ^2 is hard since it takes a large proportion of values smaller than 1, while the case $a = 1$ is simpler because about 76.3% of the values of σ^2 are larger than 1 and 0.04% larger than 3. Moreover, Model 2 is also a multivariate model where we introduce higher order terms in the variance function. In this sense, the estimation of the variance function is hard since in addition, there are only 28% of values of σ^2 greater than 1. In Model 3 we consider a sparse model for the regression function where X is an $N \times p$ matrix (p is the number of predictors) with independent uniform entries, $\beta \in \mathbb{R}^p$ is a vector of weights, and $\xi \in \mathbb{R}^N$ is a standard Gaussian noise vector and is independent of the feature X. We fix $p = 50$. The vector β is chosen to be s-sparse where $s < p$, that means β has only first s coordinates different from 0; $\beta_i = \mathbb{1}_{\{i \leq s\}}$. Here, we choose $s = 14$. In addition, the variance function in this model is less difficult. Indeed, σ^2 takes only 24.8% values greater than 1. Finally, the last two examples are two models when Y is bounded. Model 4 is bivariate regression model where the estimation of σ^2 is difficult (about 99.8% of the values are less than 1). Lastly, considering Model 5, the values of σ^2 are between 0 and 3.12. There are 36.6% of values that are larger than 1. From this perspective, the estimation of the variance function is less complicated. However, the presence of higher order terms makes the problem harder.

4.2 Benefit of aggregation

In this section, we improve the classical methods based on residual-based approach by considering aggregation. In the same time we compare `MS` and `C` aggregation.

4.2.1 Machines and simulation scheme

The construction of the aggregates $\hat{\sigma}^2_{\texttt{MS}}$ and $\hat{\sigma}^2_{\texttt{C}}$ is described in Sections 2.1 and 2.2. We recall that we focus on the residual-based method to compute the candidates of the variance function σ^2. One of the advantages of using the aggregation approach is that the collection of candidates is chosen by the practitioner and can be arbitrary. We build three dictionaries $\mathcal{F} = \{\hat{f}_s\}_{s=1}^{12}$, $\mathcal{G}_1 = \big\{\hat{\sigma}^2_{\hat{s},12}\big\}_{m=1}^{12}$ and $\mathcal{G}_2 = \{\hat{\sigma}^2_{\hat{\lambda},j}\}_{j=1}^{12}$ that contain 12 machines each: the random forest with different number of trees (ntree=50, 150, 500), the kNN with different values of k ($k = 7, 13, 22$), the Lasso with different values of tuning parameter ($\lambda = 0.5, 2$), the Ridge with different values of tuning parameter ($\lambda = 0.9, 3$), regression tree and the Elastic Net regression with a penalty term $\lambda = 1$ and a parameter $\alpha = 0.6$ that compromises between the ℓ_1 and the ℓ_2 terms in the penalty. The first dictionary is exploited to compute the aggregates $\hat{f}_{\mathtt{MS}}$ and $\hat{f}_{\mathtt{C}}$ while the last two are used to calculate respectively, $\hat{\sigma}^2_{\mathtt{MS}}$ and $\hat{\sigma}^2_{\mathtt{C}}$ with those 12 machines. For the 12 algorithms, we use the following R packages:

- Regression tree (R package `tree`, Ripley (2019));
- k-nearest neighbours regression (R package `FNN`, Li (2019));
- RandomForest regression (R package `randomForest`, Liaw and Wiener (2002));
- Lasso regression (R package `glmnet`, Friedman et al. (2010));
- Ridge regression (R package `glmnet`);
- Elastic Net regression (R package `glmnet`).

Other parameters are set by default. In addition to that, we use `Optim` function in `R` which is based on method `BFGS` to compute $\hat{\lambda}$ and $\hat{\beta}$. Now, we evaluate the performances of $\hat{\sigma}^2_{\mathtt{MS}}$ and $\hat{\sigma}^2_{\mathtt{C}}$ on previous models. Besides, we provide estimation of the L^2-error for $\hat{\sigma}^2_{\mathtt{MS}}$ and $\hat{\sigma}^2_{\mathtt{C}}$ and repeat independently $L = 100$ times the following steps:

1. simulate three datasets $\mathcal{D}_n$, $\mathcal{D}_N$ and $\mathcal{D}_T$ with $n \in \{100, 1000\}$, $N \in \{100, 1000\}$ and $T = 1000$;
2. based on $\mathcal{D}_n$, we compute the dictionary $\mathcal{F}$, and then based on $\mathcal{D}_N$, we compute the aggregates $\hat{f}_{\mathtt{MS}}$ (that is $\hat{s}$) and $\hat{f}_{\mathtt{C}}$ (that is $\hat{\lambda}$) of the regression function f^* provided in Eqs (2) and (4);
3. based on $\mathcal{D}_n$ and $\hat{f}_{\mathtt{MS}}$ (resp. $\hat{f}_{\mathtt{C}}$), we compute the collection $\mathcal{G}_1$ (resp. $\mathcal{G}_2$) and we calculate $\hat{\sigma}^2_{\mathtt{MS}}$ and $\hat{\sigma}^2_{\mathtt{C}}$ on $\mathcal{D}_N$;
4. based on $D_n \cup D_N$: firstly, we compute the collection $\mathcal{F}$[1]; secondly, for each estimate $\hat{f}_s$ in $\mathcal{F}$ we calculate the estimators $\{\hat{\sigma}^2_{s,m}\}_{1 \leq m \leq 12}$ of σ^2 corresponding to the 12 procedures in $\mathcal{F}$;
5. finally, over $\mathcal{D}_T$, we compute the empirical L^2-error of the aggregates $\hat{\sigma}^2_{\mathtt{MS}}$ and $\hat{\sigma}^2_{\mathtt{C}}$. On the other hand, we compute the collection of estimators of the variance function $\{\sigma^2_{s,m}\}_{1 \leq s,m \leq 12}$ and we choose the best among them in terms of empirical L^2-error (always on the datatest $\mathcal{D}_T$). That means, we take the smallest of empirical L^2-error as follow: for all $(s, m) \in [12] \times [12]$
$$\min \frac{1}{T} \sum_{i=1}^{T} \left(\hat{\sigma}^2_{s,m}(X_i) - \sigma^2(X_i)\right)^2 \quad , \tag{8}$$
and denote the best method. Finally, we compare it with our aggregation methods.

From these experiments, we compute the means and standard deviations of both empirical L^2-errors $\widehat{\mathrm{Err}}$ for $\hat{\sigma}^2_{\mathtt{MS}}$, $\hat{\sigma}^2_{\mathtt{C}}$ and the best method and we display the boxplot of the empirical L^2-error.

4.2.2 Results

We present our results in Figures 3-8 and Tables 1 and 2. We make several observations. First, the convex aggregation method is better than the model selection aggregation method in all models. The best can only serve as a benchmark to see the performance of a real estimator. For this reason, it is perfectly natural that "best" has better performances than our aggregation procedures on the test sample $\mathcal{D}_T$ because the selection of the best couple (s, m) (see Eq. (8)) depends precisely on $\mathcal{D}_T$. Second, we notice that when n and N are enough, the `MS`-estimator $\hat{\sigma}^2_{\mathtt{MS}}$ and the `C`-estimator $\hat{\sigma}^2_{\mathtt{C}}$ have a similar

[1] Note that this set of estimators differ from the dictionary computed in step 2. since it is computed in the whole data $D_n \cup D_N$. We abuse in the notation to avoid extra notation that are irrelevant for the understanding.

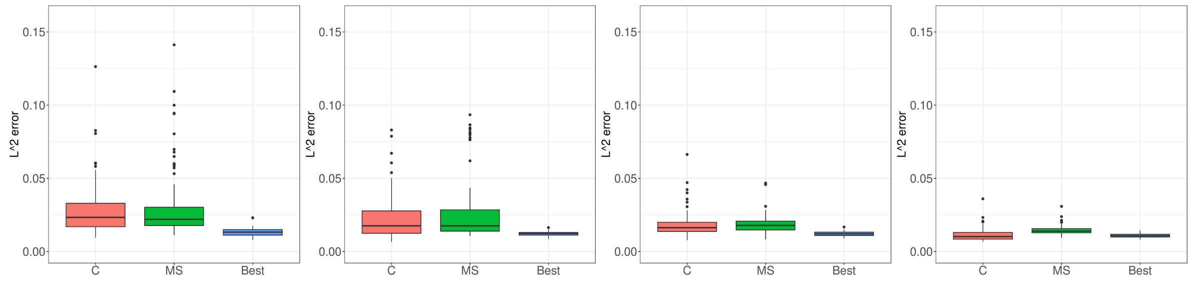
performance, that is close to the performance of the best method. These results reflect our theory: the consistency of MS-estimator and the C-estimator. Third, we observe that the empirical L^2-error of $\hat{\sigma}^2_{\texttt{MS}}$ and $\hat{\sigma}^2_{\texttt{C}}$ decreases faster in the simpler models (with respect to the estimation of the variance function) when n and N increase (see the evolution of the boxplots in Figures 4 and 8 as compared to Figures 3 and 6. In addition, our numerical results highlight an interesting fact: when we split data, it is advantageous to put more data in the second dataset $\mathcal{D}_N$ used in the aggregation step. Indeed, it seems as illustrated in Table 2 that the methods have better performance for large samples $\mathcal{D}_N$ is all cases. As an example, the mean error in Model 1 with $a = 1$ for C-aggregation is 0.33 when $n = 1000$ and $N = 100$ and 0.26 when $n = 100$ and $N = 1000$.

	$n = N = 100$			$n = N = 1000$		
	C	MS	Best	C	MS	Best
Model	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$
Model 1 ($a = 0.25$)	0.028 (0.018)	0.031 (0.023)	0.013 (0.003)	0.011 (0.004)	0.014 (0.003)	0.011 (0.001)
Model 1 ($a = 1$)	0.407 (0.214)	0.428 (0.279)	0.200 (0.44)	0.155 (0.044)	0.200 (0.040)	0.164 (0.013)
Model 2	0.247 (0.133)	0.272 (0.180)	0.110 (0.025)	0.106 (0.046)	0.100 (0.093)	0.070 (0.010)
Model 3	0.287 (0.092)	0.302(0.125)	0.218 (0.019)	0.194 (0.021)	0.198 (0.044)	0.164 (0.011)
Model 4	0.032 (0.027)	0.034 (0.036)	0.010 (0.005)	0.010 (0.005)	0.011 (0.003)	0.009 (0.001)
Model 5	0.382 (0.116)	0.405 (0.168)	0.264 (0.032)	0.209 (0.040)	0.223 (0.024)	0.178 (0.016)

Table 1: Average and standard deviation of the empirical L^2-error of the three estimators with $n = N$.

	$n = 1000$ $N = 100$			$n = 100,\ N = 1000$		
	C	MS	Best	C	MS	Best
Model	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$
Model 1 ($a = 0.25$)	0.023 (0.015)	0.028 (0.023)	0.012 (0.002)	0.018 (0.008)	0.019 (0.006)	0.012 (0.002)
Model 1 ($a = 1$)	0.335 (0.265)	0.381 (0.343)	0.170 (0.014)	0.262 (0.090)	0.278 (0.081)	0.169 (0.018)
Model 2	0.193 (0.132)	0.227 (0.189)	0.074 (0.013)	0.159 (0.055)	0.148 (0.054)	0.073 (0.010)
Model 3	0.252 (0.082)	0.278 (0.149)	0.180 (0.015)	0.259 (0.029)	0.270 (0.035)	0.179 (0.015)
Model 4	0.021 (0.014)	0.026 (0.027)	0.009 (0.002)	0.019 (0.012)	0.017 (0.015)	0.009 (0.002)
Model 5	0.295 (0.144)	0.336 (0.209)	0.195 (0.019)	0.313 (0.079)	0.317 (0.095)	0.194 (0.015)

Table 2: Average and standard deviation of the empirical L^2-error of the three estimators with $n \neq N$.

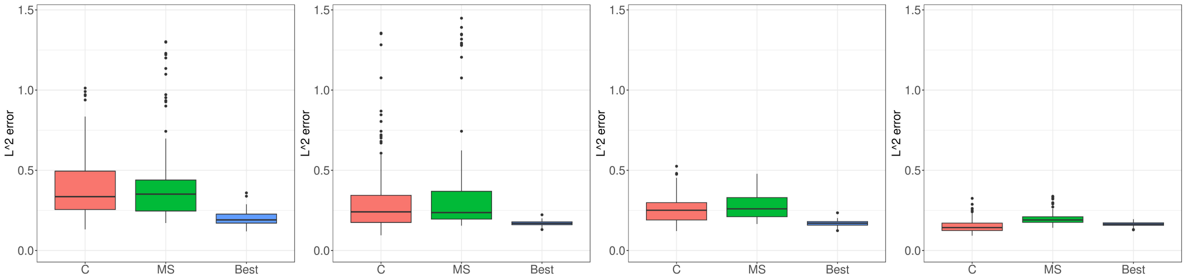


(a) $n = N = 100$ (b) $n = 1000, N = 100$ (c) $n = 100, N = 1000$ (d) $n = N = 1000$

Figure 3: Boxplot of the empirical L^2-error of the estimators in Model 1 ($a = 0.25$)

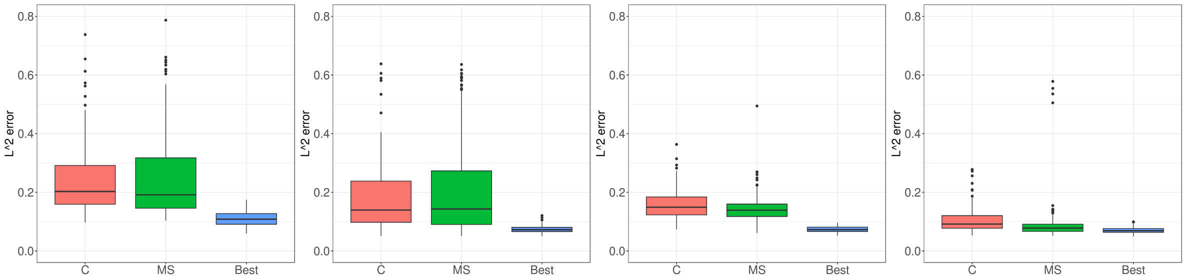
4.3 Real datasets

In this part, we consider two real datasets which are available on the UCI database. The first dataset is *Concrete Compressive Strength*. The concrete compressive strength is a highly nonlinear function of age and ingredients (Cement, Water, Blast Furnace Slag, ...). It contains 1030 observations of 8 numerical features. The output takes its values in $[2.330, 82.598]$. The second dataset is *Airfoil Self-Noise*. It



(a) $n = N = 100$ (b) $n = 1000, N = 100$ (c) $n = 100, N = 1000$ (d) $n = N = 1000$

Figure 4: Boxplot of the empirical L^2-error of the estimators in Model 1 ($a = 1$)



(a) $n = N = 100$ (b) $n = 1000, N = 100$ (c) $n = 100, N = 1000$ (d) $n = N = 1000$

Figure 5: Boxplot of the empirical L^2-error of the estimators in Model 2

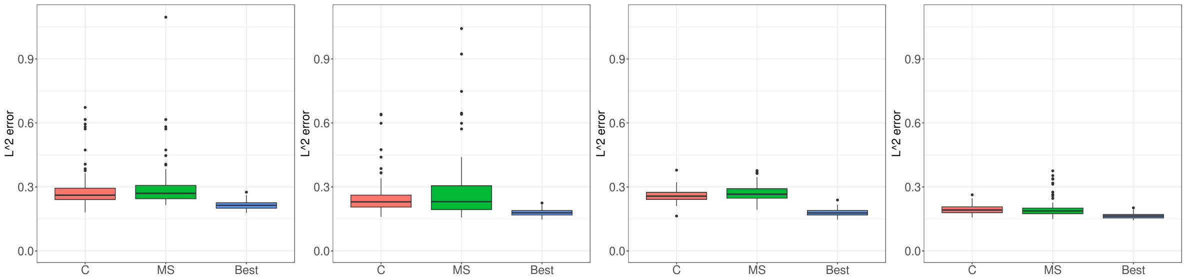
is obtained from a series of aerodynamic and acoustic tests of two and three-dimensional airfoil blade sections conducted in an anechoic wind tunnel. It contains 1503 observations of 5 numerical features. The output takes its values in $[103, 140]$. We display the histogram of an estimate of the variance function σ^2 produced by the random forest algorithm for both datasets in Figure 9. The estimated values of $\hat{\sigma}^2$ are large in two real datasets: 30% and 41% of the values are larger than 10 in Concrete Compressive Strength and Airfoil Self-Noise, respectively. Now, we use the same steps in Section 4.2.1 to illustrate the performance of our methods with the following modifications: we reduce the dictionaries $\mathcal{F}$, $\mathcal{G}_1$ and $\mathcal{G}_2$ into 4 candidates: Lasso, kNN, random forest and support vector machines[2] methods where the parameters of the first two algorithms are chosen by cross-validation and the last two are chosen by default from `glmnet`, `FNN`, `randomForest` and `e1071` packages. We set $k \in \{5, 10, 13, 15, 17, 22, 35, 50, 75, 85, 100, 125\}$ for kNN. In Step 4, based on $D_n \cup D_N$, we firstly compute the collection $\mathcal{F}$; secondly, for each estimate $\hat{f}_s$ in $\mathcal{F}$ we compute all possible true estimators $\{\hat{\sigma}^2_{s,m}\}_{1 \leq m \leq 4}$ of σ^2 corresponding to the 4 procedures in $\mathcal{F}$. In the last step, we compute the empirical L^2-risk of

- `MS`-method: $\frac{1}{T} \sum_{i=1}^{T} \left((Y_i - \hat{f}_{\texttt{MS}}(X_i))^2 - \hat{\sigma}^2_{\texttt{MS}}(X_i) \right)^2$;
- `C`-method: $\frac{1}{T} \sum_{i=1}^{T} \left((Y_i - \hat{f}_{\texttt{C}}(X_i))^2 - \hat{\sigma}^2_{\texttt{C}}(X_i) \right)^2$;
- `best`-method (best empirical L^2-risk): based on Step 4, we take the minimum of

$$\frac{1}{T} \sum_{i=1}^{T} \left((Y_i - \hat{f}_s(X_i))^2 - \hat{\sigma}^2_{s,m}(X_i) \right)^2 \quad \text{for all} \quad (s, m) \in [4] \times [4].$$

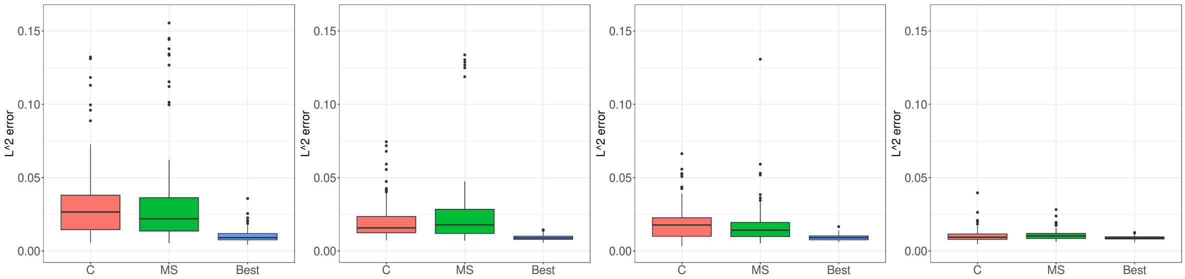
From this estimates, we compute the mean and the standard deviation of the empirical L^2-risk. The associated boxplots are given in Figure 10. Here, we fix $T = 200$. We take $n \in \{150, 415, 680\}$,

[2] We simplify it by `svm` and use the R package with default parameters: `e1071`.



(a) $n = N = 100$ (b) $n = 1000, N = 100$ (c) $n = 100, N = 1000$ (d) $n = N = 1000$

Figure 6: Boxplot of the L^2-error of the estimators in sparse model when $p = 50$, and $s = 14$.



(a) $n = N = 100$ (b) $n = 1000, N = 100$ (c) $n = 100, N = 1000$ (d) $n = N = 1000$

Figure 7: Boxplot of the L^2-error of the estimators in Model 4.

$N \in \{680, 415, 150\}$ for the first real dataset and $n \in \{150, 652, 1153\}$, $N \in \{1153, 651, 150\}$ for the second dataset. We observe that both of our aggregation methods achieve the same performance. Note that the aggregation methods are outperformed by the best method when n is large. This is mainly due to the fact that the method called " best" is computed without splitting the data (as explained before). Finally, we notice that it is advantageous to put a lot of data in the first dataset.

4.4 Applications of variance function

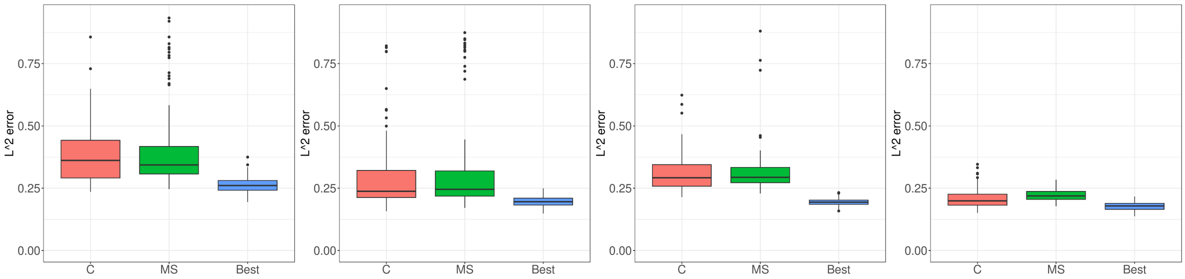
This section presents two applications of our aggregation methods: regression with reject option and quantile regression.

4.4.1 Regression with reject option

We illustrate our aggregation methods in the regression with reject option with two real datasets: we begin with a a brief description of regression with reject option.

In regression, many proposed estimation procedures aim to reduce prediction errors. However, even the most efficient methods make mistakes that can in some cases have serious consequences. Regression with reject option is a way to address the problem of estimating the uncertainty of a predictor. That means, we refuse to predict when the doubt in the predicted value is too great. Denis et al. (2020) formulate a general framework in regression with reject option and derive the optimal rule which relies on thresholding the conditional variance function where the rejection rate is fixed. Given a rejection rate $\varepsilon \in (0, 1)$, the ε-predictor is given by

$$\Gamma^*_\varepsilon(x) := \begin{cases} \{f^*(x)\} & \text{if } F_{\sigma^2}(\sigma^2(x)) \leq 1 - \varepsilon \\ \emptyset & \text{otherwise} \end{cases},$$



(a) $n = N = 100$ (b) $n = 1000, N = 100$ (c) $n = 100, N = 1000$ (d) $n = N = 1000$

Figure 8: Boxplot of the L^2-error of the estimators in Model 5.

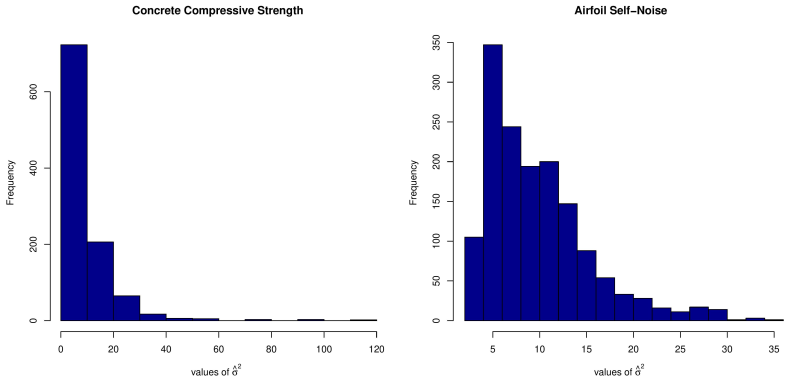


Figure 9: Histogram of the values of the estimates of the variance function.

where F_{σ^2} is the cumulative distribution function of $\sigma^2(X)$. Here, the quantity $|\Gamma^*_\varepsilon(X)|$ represents the cardinality of Γ^*_ε. If $|\Gamma^*_\varepsilon(X)| = 0$, that means no prediction is produced for X and $|\Gamma^*_\varepsilon(X)| = 1$ otherwise. The ε-predictor has rejection rate $r\left(\Gamma^*_\varepsilon\right)$ exactly ε

$$r\left(\Gamma^*_\varepsilon\right) := \mathbb{P}\left(|\Gamma^*_\varepsilon(X)| = 0\right) = \mathbb{P}\left(F_{\sigma^2}(\sigma^2(X)) \geq 1 - \varepsilon\right) = \varepsilon.$$

Moreover, the performance of Γ^*_ε is measured by the L_2-error when prediction is performed

$$\mathrm{Err}\left(\Gamma^*_\varepsilon\right) := \mathbb{E}\left[(Y - f^*(X))^2 \mid |\Gamma^*_\varepsilon(X)| = 1\right] \quad .$$

The L_2 error and the rejection rate of Γ^*_ε are working in two opposite directions *w.r.t.* ε, more precisely $\forall\, \varepsilon_1 \leq \varepsilon_2$

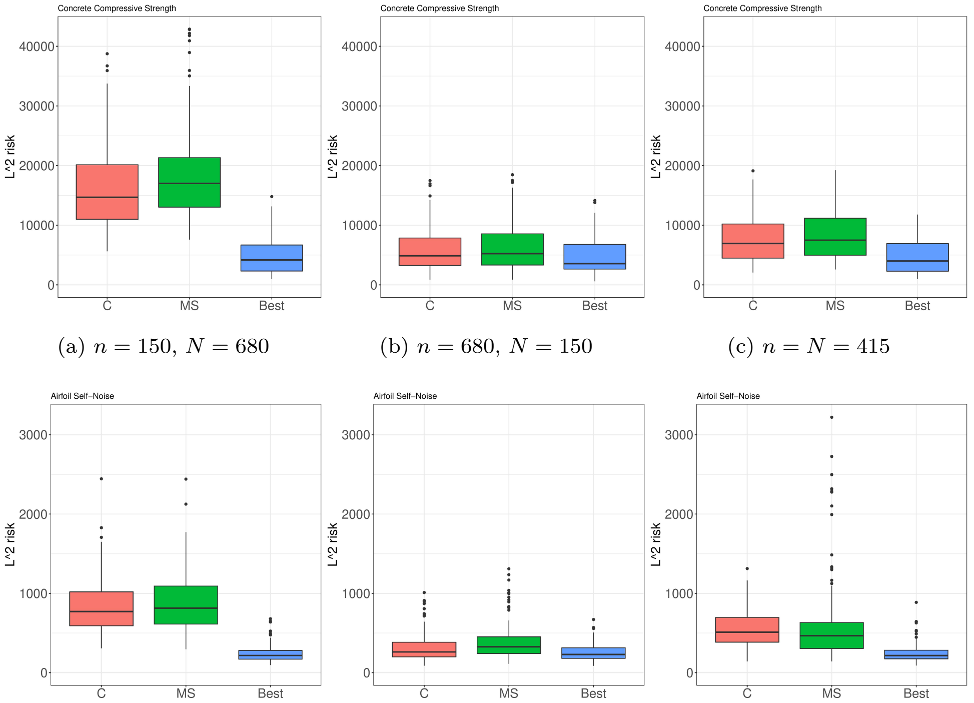
$$\mathrm{Err}\left(\Gamma^*_{\varepsilon_2}\right) \leq \mathrm{Err}\left(\Gamma^*_{\varepsilon_1}\right), \text{ and } r\left(\Gamma^*_{\varepsilon_1}\right) \leq r\left(\Gamma^*_{\varepsilon_2}\right) \quad .$$

The estimate of Γ^*_ε needs two independent samples $\mathcal{D}_{N_1}$ and $\mathcal{D}_M$ where $\mathcal{D}_M$ is composed of M independent copies of the feature X. The sample $\mathcal{D}_{N_1}$ will be used to construct estimators $\hat{f}$ and $\hat{\sigma}^2$ of f^* and σ^2. Besides, we consider the randomised prediction $\hat{\hat{\sigma}}^2(X) = \hat{\sigma}^2(X) + \zeta$ where $\zeta \sim \mathcal{U}([0, u])$ is independent of every other random variable with $u > 0$ is a small fixed real number. Thus, we use $\mathcal{D}_M$ to estimate F_{σ^2} which is given by the empirical distribution function of $\hat{\hat{\sigma}}^2$

$$\hat{F}_{\hat{\hat{\sigma}}^2}(\cdot) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}_{\{\hat{\sigma}^2(X_{N_1+i}) + \zeta_i \leq \cdot\}} \quad .$$

Finally, the *plug-in ε-predictor* is the predictor with reject option defined for each $x \in \mathbb{R}^d$ as

$$\hat{\Gamma}_\varepsilon(x) = \begin{cases} \left\{\hat{f}(x)\right\} & \text{if } \hat{F}_{\hat{\hat{\sigma}}^2}(\hat{\hat{\sigma}}^2(x)) \leq 1 - \varepsilon \\ \emptyset & \text{otherwise} \quad . \end{cases}$$



(a) $n = 150,\ N = 680$

(b) $n = 680,\ N = 150$

(c) $n = N = 415$

(d) $n = 150,\ N = 1153$

(e) $n = 1153,\ N = 150$

(f) $n = N = 652$

Figure 10: Boxplots of the L^2-risk of our aggregation methods and the best method.

This plug-in approach is shown to be consistent, see Denis et al. (2020). In particular, the plug-in ε-predictor asymptotically behaves as well as the the best predictor Γ_ε^* both in terms of risk and rejection rate. We may have some doubts in the associated prediction on two real datasets since the estimated variance is large. We evaluate the performance of the procedure on two real datasets considering the same algorithm for both estimation tasks (same approach to estimate the regression and variance functions) and build four plug-in ε-predictors based respectively, on support vector machines (`svm`), random forests (`rf`), and Lasso (`Lasso`) and kNN (`knn`) algorithms. We take $\zeta \sim \mathcal{U}([0, 10^{-7}])$. In particular, we run 100 times the procedure where we split the data each time in three: $\mathcal{D}_{N_1}$ with $N_1 = 780$ for Concrete Compressive Strength and $N_1 = 1253$ for Airfoil Self-Noise, $\mathcal{D}_M$ with $M = 150$ and $\mathcal{D}_T$ with $T = 100$. The dataset $\mathcal{D}_T$ is exploited to calculate the empirical rejection rate $\hat{r}$ and the empirical error $\widehat{\mathrm{Err}}$ for each $\varepsilon \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. From these estimations, we compute the average and standard deviation (between parentheses) of $\hat{r}$ and $\widehat{\mathrm{Err}}$. The results are reported in Table 3 with $\varepsilon \in \{0, 0.2, 0.5, 0.8\}$ and in Figure 11.

First of all, we recall that for $\varepsilon = 0$, the measure of risk of ε-predictor match with the error of the approach we use to estimate the regression function f^*. We remark that the best plug-in ε-predictor with $\varepsilon = 0$ is the `rf` method for concrete compressive strength and the `svm` method for airfoil self-noise. Their errors are 31.33 and 10.36, respectively. Notice that all the corresponding errors diminishes with ε and rejection rate is close to ε.

We recall that our aim is not to build a regression rule that reduces the error rate. The motivation for introducing the plug-in ε-predictor is only to improve the confidence on prediction. Since the construction of the optimal rule depends on the estimators of f^* and σ^2, poor estimators would lead to bad plug-in ε-predictor. Now, we hope that our aggregation methods improve the accuracy of the procedure.

For a comparative study, we evaluate the performance of the *plug-in ε-predictor* considering the same algorithm for both estimation tasks of estimating the regression and the variance functions: we build four plug-in ε-predictors based respectively, on support vector machines (`svm`), random forests (`rf`), and Lasso (`Lasso`), kNN (`knn`) algorithms. We compare these methods to our aggregation procedures `C` and `MS`. For each $\varepsilon \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, and each plug-in ε-predictor, we compute

Table 3: Performances of the four plug-in ε-predictors on the real datasets `Concrete compressive strength`, and `airfoil self-Noise`.

`Concrete compressive strength`

	`rf`		`svm`		`knn`		`Lasso`	
ε	$\widehat{\mathrm{Err}}$	$\hat{r}$	$\widehat{\mathrm{Err}}$	$\hat{r}$	$\widehat{\mathrm{Err}}$	$\hat{r}$	$\widehat{\mathrm{Err}}$	$\hat{r}$
0	31.33 (7.82)	0.00 (0.00)	46.83 (8.65)	0.00 (0.00)	87.32 (15.74)	0.00 (0.00)	111.72 (16.13)	0.00 (0.00)
0.2	20.99 (5.72)	0.21 (0.06)	33.98 (7.54)	0.20 (0.06)	65.61 (16.12)	0.20 (0.06)	89.91 (14.63)	0.19 (0.05)
0.5	13.47 (6.08)	0.48 (0.06)	21.73 (7.38)	0.50 (0.06)	46.71 (19.89)	0.50 (0.06)	66.86 (16.13)	0.50 (0.07)
0.8	7.26 (8.29)	0.81 (0.05)	18.25 (11.70)	0.81 (0.05)	28.58 (28.66)	0.79 (0.11)	52.32 (16.99)	0.81 (0.05)

`Airfoil Self-Noise`

	`rf`		`svm`		`knn`		`Lasso`	
ε	$\widehat{\mathrm{Err}}$	$\hat{r}$	$\widehat{\mathrm{Err}}$	$\hat{r}$	$\widehat{\mathrm{Err}}$	$\hat{r}$	$\widehat{\mathrm{Err}}$	$\hat{r}$
0	12.57 (1.99)	0.00 (0.00)	10.36 (2.68)	0.00 (0.00)	33.97 (4.45)	0.00 (0.00)	22.51 (3.62)	0.00 (0.00)
0.2	8.53 (1.48)	0.20 (0.05)	7.42 (2.09)	0.19 (0.05)	29.25 (4.03)	0.20 (0.06)	20.52 (3.17)	0.19 (0.06)
0.5	6.25 (1.41)	0.52 (0.07)	3.89 (1.45)	0.50 (0.06)	22.42 (4.10)	0.49 (0.06)	15.51 (3.37)	0.50 (0.06)
0.8	2.82 (1.00)	0.80 (0.05)	1.93 (0.93)	0.80 (0.05)	14.95 (7.12)	0.80 (0.07)	9.07 (3.19)	0.80 (0.05)

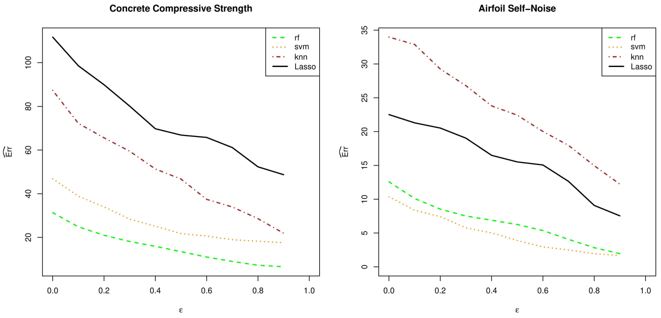


Figure 11: Visual description of the performance of four plug-in ε-predictors.

the empirical rejection rate $\hat{r}$ and the empirical error $\widehat{\mathrm{Err}}$. We take $\zeta \sim \mathcal{U}([0, 10^{-7}])$. So, we repeat independently 100 times the following steps:

1. simulate four datasets $\mathcal{D}_n$, $\mathcal{D}_N$, $\mathcal{D}_M$ and $\mathcal{D}_T$ with $M = 150$, $N = 150$ and $T = 100$;

2. based on $\mathcal{D}_n$, we compute the estimators in $\mathcal{F}$, and then based on $\mathcal{D}_N$, we compute the aggregates $\hat{f}_{\mathtt{MS}}$ and $\hat{f}_{\mathtt{CM}}$. Then, we compute the `knn`, `Lasso`, `rf` and `svm` estimators of the regression function on $\mathcal{D}_n \cup \mathcal{D}_N$;

3. based on $\mathcal{D}_n$ and $\hat{f}_{\mathtt{MS}}$ (resp. $\hat{f}_{\mathtt{CM}}$), we compute the estimators in $\mathcal{G}_1$ (resp. $\mathcal{G}_2$). Then, based on $\mathcal{D}_N$ we calculate $\hat{\sigma}^2_{\mathtt{MS}}$ and $\hat{\sigma}^2_{\mathtt{CM}}$. From $\mathcal{D}_n \cup \mathcal{D}_N$, we compute the `knn`, `Lasso`, `rf` and `svm` estimators of σ^2;

4. based on $\mathcal{D}_M$, we compute the empirical cumulative distribution function of the randomised estimators $\hat{\tilde{\sigma}}^2(X)$;

5. finally, over $\mathcal{D}_T$, we compute the empirical rejection rate $\hat{r}$ and the empirical error $\widehat{\mathrm{Err}}$ for the considered $\hat{\Gamma}_\varepsilon$.

From these estimations, we compute the average and standard deviation (between brackets) of $\hat{r}$ and $\widehat{\mathrm{Err}}$. The results are reported in Table 4 with $\varepsilon \in \{0, 0.2, 0.5, 0.8\}$ and in Figure 12. Our main observation is that the `C` plug-in ε-predictor has the same performance as `MS` plug-in ε-predictor. According to the rejection rate, we recall that our theory related to this point is distribution free and this is also observed in Table 4 since all rejection rate have the approximately prescribed level ε. Importantly, note that both aggregation-based methods require splitting the data (part for estimation and a part for aggregation) while the other plug-in ε-predictors (such as `rf`) do not. However, our plug-in ε-predictors based on aggregation have a similar performance as the best. This result validates the relevance of our strategy.

Table 4: Performances of the six plug-in ε-predictors on the real datasets `Concrete compressive strength`, and `Airfoil Self-Noise`.

`Concrete compressive strength`

	`rf`		`svm`		`knn`		`Lasso`		`C`		`MS`	
ε	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$
0	30.27 (7.16)	0.00 (0.00)	45.03 (9.03)	0.00 (0.00)	85.59 (16.11)	0.00 (0.00)	110.09 (16.01)	0.00 (0.00)	34.59 (7.74)	0.00 (0.00)	35.08 (7.44)	0.00 (0.00)
0.2	20.07 (5.65)	0.20 (0.05)	32.97 (7.03)	0.20 (0.05)	62.40 (13.93)	0.20 (0.06)	86.51 (11.72)	0.20 (0.05)	23.07 (6.70)	0.20 (0.05)	24.76 (6.91)	0.20 (0.05)
0.5	13.25 (5.38)	0.48 (0.07)	22.45 (7.16)	0.49 (0.07)	44.98 (15.11)	0.49 (0.07)	64.95 (11.08)	0.49 (0.07)	13.24 (3.87)	0.49 (0.06)	15.14 (4.18)	0.49 (0.07)
0.8	7.91 (8.26)	0.80 (0.05)	17.92 (13.69)	0.80 (0.04)	30.02 (23.85)	0.78 (0.09)	55.91 (20.09)	0.80 (0.06)	8.50 (5.33)	0.80 (0.05)	10.19 (9.77)	0.80 (0.05)

`Airfoil Self-Noise`

	`rf`		`svm`		`knn`		`Lasso`		`C`		`MS`	
ε	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$	$\widehat{\text{Err}}$	$\hat{r}$
0	12.80 (1.84)	0.00 (0.00)	10.49 (2.18)	0.00 (0.00)	33.78 (4.33)	0.00 (0.00)	23.23 (3.71)	0.00 (0.00)	10.30 (1.98)	0.00 (0.00)	10.77 (2.22)	0.00 (0.00)
0.2	8.67 (1.29)	0.20 (0.05)	7.02 (1.86)	0.21 (0.05)	29.15 (4.60)	0.20 (0.05)	21.16 (3.49)	0.20 (0.06)	6.19 (1.17)	0.21 (0.05)	5.82 (1.41)	0.21 (0.05)
0.5	6.08 (1.13)	0.49 (0.07)	3.84 (0.98)	0.48 (0.07)	21.95 (4.42)	0.49 (0.07)	14.83 (3.39)	0.49 (0.08)	4.09 (0.99)	0.49 (0.07)	3.73 (1.15)	0.49 (0.06)
0.8	2.92 (1.02)	0.80 (0.04)	1.97 (1.06)	0.80 (0.05)	14.25 (6.94)	0.81 (0.08)	8.76 (2.64)	0.80 (0.05)	1.68 (0.64)	0.80 (0.04)	1.67 (0.73)	0.80 (0.05)

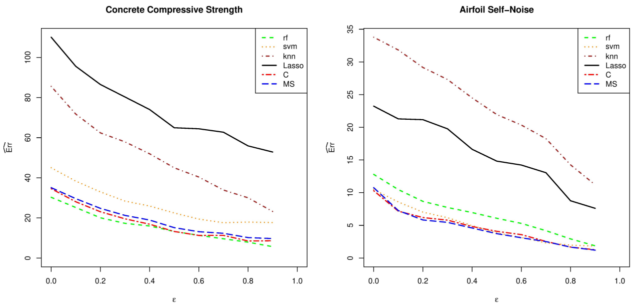


Figure 12: Visual description of the performance of six plug-in ε-predictors.

4.4.2 Quantile regression

In this section, we illustrate the performance of our aggregation methods for quantile regression. Quantile regression allows a comprehensive analysis of the relationships between a response Y and input variables X. It is interesting in the entire conditional distribution of the dependent variable, and not only on its mean. For more details see for instance (Koenker, 2005, Shan and Yang, 2009, Takeuchi et al., 2006). We recall that in our work we observe $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ such that

$$Y = f^*(X) + \sigma(X)\xi,$$

where ξ is the known noise with mean zero and unit variance. Let $\tau \in (0, 1)$. The conditional τ-quantile of Y given $X = x$, denoted by q_τ, is given by

$$q_\tau(Y|X) = \inf\{y : F_{Y|X}(y) \geq \tau\},$$

where $F_{Y|X}$ is the cumulative distribution function of $Y|X$. In particular, the quantity $q_\tau(Y|X)$ has the following form

$$q_\tau(Y|X) = f^*(X) + \sigma(X)F_\xi^{-1}(\tau),$$

where F_ξ is the cumulative distribution function of ξ and is known. The plug-in approach is a possible procedure to estimate q_τ. Given an estimator $\hat{f}$ of f^* and an estimator $\hat{\sigma}^2$ of σ^2, the plug-in estimator of q_τ is

$$\hat{q}_\tau(Y|X) = \hat{f}(X) + \hat{\sigma}(X)F_\xi^{-1}(\tau).$$

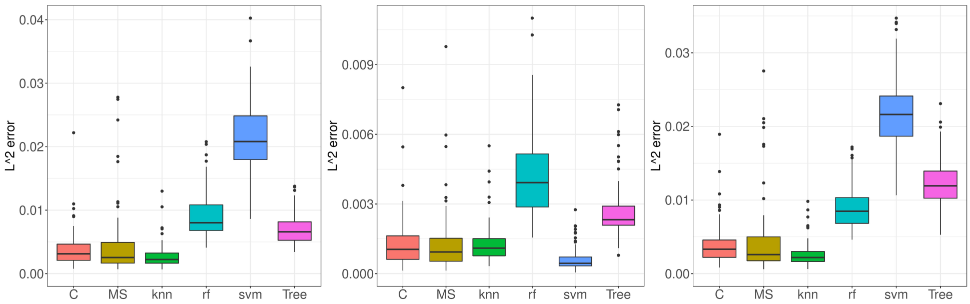
It is clear that a good estimate of the conditional τ-quantile is related to a good estimate of the regression function f^* and the variance function σ^2. We evaluate the performance of $\hat{q}_\tau(Y|X)$ according to four different estimations of the regression function and the variance function: random forests, kNN, tree regression, svm and our two aggregation approaches, in the following gaussian model which was introduced in Takeuchi et al. (2006)

$$Y = \text{sinc}(X) + 0.1\exp(1 - X)\xi,$$

where X is drawn uniformly from $[-1, 1]$, sinc is the normalised sinc function, and $\xi \sim \mathcal{N}(0, 1)$. We fix $maxnodes = 25$ for the `rf` algorithm, and $n = N = T = 1000$. The performance of $\hat{q}_\tau(Y|X)$ is measured by the empirical quadratic error. The performances obtained from 100 independent runs, computed using the same methods mentioned in the previous section, are provided in Table 5, Figure 13 and Figure 14 for three different quantiles $\tau \in \{0.1, 0.5, 0.9\}$. For $\tau \in \{0.1, 0.9\}$ this is an estimate of the first and last deciles and for $\tau = 0.5$ an estimate of the median.

Table 5: Performances of the six plug-in regression quantiles. We compute the means and standard deviations (between parentheses) of the L^2-error of $\hat{q}_\tau(Y|X)$.

	`C`	`MS`	`knn`	`rf`	`svm`	`Tree`
τ	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$	$\widehat{\text{Err}}$
0.1	0.0037 (0.0028)	0.0043 (0.0050)	0.0027 (0.0019)	0.0091 (0.0033)	0.0217 (0.0056)	0.0071 (0.0024)
0.5	0.0013 (0.0011)	0.0013 (0.0013)	0.0013 (0.0008)	0.0042 (0.0018)	0.0006 (0.0005)	0.0027 (0.0012)
0.9	0.0039 (0.0028)	0.0043 (0.0047)	0.0026 (0.0016)	0.0089 (0.0029)	0.0216 (0.0051)	0.0123 (0.0033)



(a) $\tau = 0.1$

(b) $\tau = 0.5$

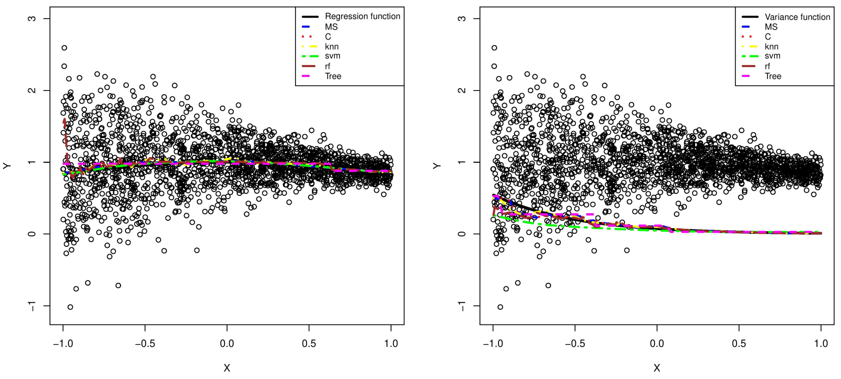
(c) $\tau = 0.9$

Figure 13: Boxplots of the L^2-error of six methods with $\tau \in \{0.1, 0.5, 0.9\}$.

Firstly, we can see here again that our both aggregation based approaches have the same performances. Secondly, the `svm` method which is built on the union of two samples $\mathcal{D}_n$ and $\mathcal{D}_N$ is the best method for the median problem estimation ($\tau = 0.5$) and `knn` method is the best procedure for the decile function. Finally, we can deduce that a good estimation of the regression and variance functions ensures a good estimation of the conditional τ-quantile.

5 Conclusion

In the regression setting, we estimated the variance function by the model selection and convex aggregation when the set of initial estimators are constructed by the residual based-method. We called the estimators of the two procedures the `MS`-estimator and `C`-estimator respectively. We established the consistency of our estimators under mild assumptions and provided rate of convergence for these



(a) Estimators of the regression function (b) Estimators of the variance function

Figure 14: Curves of six estimators of the regression function and the variance function.

methods in L_2-norm that are of order $O((\log(M_1)/N)^{1/8})$ when Y is satisfied the gaussian model; and $O((\log(M_1)/N)^{1/4})$ when Y is bounded.

Our theoretical bounds do not degrade with the dimension d of the inputs. The terms that depend on the dimension are the bias terms for which special treatment is required in the high-dimensional framework. Then, it would be reasonable to use methods that are able to exploit a (necessary) structure of sparsity in the model. An interesting source of inspiration may be the papers Dalalyan et al. (2013), Kolar and Sharpnack (2012) that nicely took advantage of the sparsity structure of the data in this case. Our convergence rates are slow compared to the classical framework of aggregation (see Tsybakov (2003)). A natural question might be to study the optimality of our aggregation procedures. In the same vein, it would be interesting to extend our aggregation method to the case of functional data (see Hu (2013), Ling and Vieu (2018)).

Acknowledgements. The author thanks the Editor in Chief, the Associate Editor, the reviewers, and his advisors Christophe Denis and Mohamed Hebiri for their valuable comments which have improved the former version of the paper.

References

Anderson, T. and Lund, J. (1997). Estimating continuous-time stochastic volatility models of the short-term interest rate. *Journal of Econometrics*, 77(2):343–377.

Audibert, J. (2004). Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. H. Poincaré Probab. Statist.*, 40(6):685–736.

Audibert, J. (2009). Robust linear least squares regression. *Annals of Statistics*, 37(4):1591–1646.

Biau, G. and Devroye, L. (2015). *Lectures on the Nearest Neighbor Method.* Springer Series in the Data Sciences. Springer New York.

Brown, L. and Levine, M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *Annals of statistics*, 35(5):2219–2232.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Aggregation for gaussian regression. *Annals of Statistics*, 35(4):1674–1697.

Chow, C. (1957). An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254.

Chow, C. (1970). On optimum error and reject trade-off. *IEEE Transactions on Information Theory*, 16:41–46.

Dalalyan, A., Hebiri, M., Meziani, K., and Salmon, J. (2013). Learning heteroscedastic models by convex programming under group sparsity. *Proceedings of the 30th International Conference on Machine Learning, PMLR*, 28(3):379–387.

Denis, C. and Hebiri, M. (2020). Consistency of plug-in confidence sets for classification in semi-supervised learning. *Journal of Nonparametric Statistics*, 32(1):42–72.

Denis, C., Hebiri, M., and Zaoui, A. (2020). Regression with reject option and application to knn. NeurIPS.

Devroye, L., Györfi, L., Lugosi, G., and Walk, H. (2018). A nearest neighbor estimate of the residual variance. *Electronic Journal of Statistics*, 12(1):1752–1778.

Evans, D. and Jones, A. J. (2008). Non-parametric estimation of residual moments and covariance. *Proceedings of the Royal Society, A*, 464:2831–2846.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87(420):998–1004.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3):645–660.

Ferrario, P. G. and Walk, H. (2012). Nonparametric partitioning estimation of residual and local variance based on first and second nearest neighbors. *Journal of Nonparametric Statistics*, 24:1019–1039.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22.

Gyorfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer-Verlag, New York.

Hall, P. and Carroll, R. (1989). Variance function estimation in regression: the effect of estimating the mean. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51(1):3–14.

Härdle, W. and Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics*, 81(1):223–242.

Herbei, R. and Wegkamp, M. (2006). Classification with reject option. *The Canadian Journal of Statistics*, 34(4):709–721.

Hu, Y. (2013). Nonparametric estimation of variance function for functional data under mixing conditions. *Communications in Statistics: Theory and Methods*, 42:1774–1786.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: a guide to practice. *Journal of Econometrics*, 142:615–635.

Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Annals of Statistics*, 28(3):681–712.

Koenker, R. (2005). *Quantile Regression*. Cambridge MA: Cambridge University Press.

Kolar, M. and Sharpnack, J. (2012). Variance function estimation in high-dimensions. In Langford, J. and Pineau, J., editors, *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, pages 1447–1454, New York, NY, USA. icml.cc / Omnipress.

Kulik, R. and Wichelhaus, C. (2011). Nonparametric conditional variance and error density estimation in regression models with dependent errors and predictors. *Electron. J. Statist.*, 5:856–898.

Lecué, G. (2013). Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli*, 19(5B):2153–2166.

Lecué, G. and Mendelson, S. (2009). Aggregation via empirical risk minimization. *Probability theory and related fields*, 145(3-4):591–613.

Lei, J. (2014). Classification with confidence. *Biometrika*, 101(4):755–769.

Li, S. (2019). Fnn: Fast nearest neighbor search algorithms and applications.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2:18–22.

Liitiäinen, E., Corona, F., and Lendasse, A. (2010). Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis*, 101:811–823.

Liitiäinen, E., Verleysen, M., Corona, F., and Lendasse, A. (2009). Residual variance estimation in machine learning. *Neurocomputing*, 72:3692–3703.

Ling, N. and Vieu, P. (2018). Nonparametric modelling for functional data: Selected survey and tracks for future. *Statistics*, 52(4):934–949.

Mammen, E., Nielsen, J., Scholz, M., and Sperlich, S. (2019). Conditional variance forecasts for long-term stock returns. *Machine learning in insurance*, 7(4).

Martins-Filho, C. and Yao, F. (2007). Nonparametric frontier estimation via local linear regression. *Journal of Econometrics*, 141:283–319.

Müller, H. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *The annals of statistics*, 15(2):610–625.

Nadeem, M., Zucker, J., and Hanczar, B. (2009). Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option. *Proceedings of the third International Workshop on Machine Learning in Systems Biology, PMLR*, 8:65–81.

Neumann, M. (1994). Fully data-driven nonparametric variance estimators. *Statistics*, 25:189–212.

Opsomer, J., Ruppert, D., Wand, M., Holst, U., and Hossjer, O. (1999). Kriging with nonparametric variance function estimation. *Biometrics*, 55(3):704–710.

Ripley, B. (2019). tree: Classification and regression trees.

Ruppert, D., Wand, M., Holst, U., and HöSJER, O. (1997). Local polynomial variance function estimation. *Technometrics*, 39(3):262–273.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *Annals of Statistics*, 43(4):1716–1741.

Shan, K. and Yang, Y. (2009). Combining regression quantile estimators. *Statistica Sinica*, 19(3):1171–1191.

Stone, C. (1977). Consistent nonparametric regression. *Annals of Statistics*, 5(4):595–620.

Takeuchi, I., Le, Q., Sears, T., and Smola, A. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(45):1231–1264.

Tsybakov, A. (2003). Optimal rates of aggregation. *Learning Theory and Kernel Machines*, 2777:303–313.

Tsybakov, A. (2008). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York.

Tsybakov, A. (2014). Aggregation and minimax optimality in high-dimensional estimation. *Proceedings of International Congress of Mathematicians*, 3:225–246.

Verzelen, N. and Gassiat, E. (2018). Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 24(4B):3683–3710.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer, New York.

Wang, L., D.Brown, L., Cai, T., and Levine, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *Annals of Statistics*, 36(2):646–664.

Xia, Y., Tong, H., and Li, W. K. (2002). Single-index volatility models and estimation. *Statistica Sinica*, 12(3):785–799.

Xu, K. and Phillips, P. B. (2008). Adaptive estimation of autoregressive models with time- varying variances. *Journal of Econometrics*, 142(1):265–280.

Xu, K. and Phillips, P. B. (2011). Tilted nonparametric estimation of volatility functions with empirical applications. *Journal of Business & Economic Statistics*, 29(4):518–528.

Yu, K. and Jones, M. (2004). Likelihood based-local linear estimation of the conditional variance function. *Journal of the American Statistical Association*, 99(465):139–144.

Yuhong, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47.

Ziegelmann, F. A. (2002). Nonparametric estimation of volatility functions: The local exponential estimator. *Econometric Theory*, 18:985–991.

Appendix

This section gathers the proof of our results.

A Proof of Theorem 1

Note that the quantity $\mathbb{E}\left[|\hat{\sigma}^2_{\texttt{MS}}(X) - \sigma^2(X)|^2\right]$ is the excess risk of the estimator $\hat{\sigma}^2_{\texttt{MS}}$ and defines as follows:

$$\mathbb{E}\left[|\hat{\sigma}^2_{\texttt{MS}}(X) - \sigma^2(X)|^2\right] := \mathbb{E}\left[R(\hat{\sigma}^2_{\texttt{MS}}) - R(\sigma^2)\right] \quad ,$$

where $R(\sigma^2) = \mathbb{E}\left[|Z - \sigma^2(X)|^2\right]$ is the true risk of the variance function. Besides, we introduce a minimiser of the risk R, denoted by $\bar{\sigma}^2_{\texttt{MS}}$ and given

$$\bar{\sigma}^2_{\texttt{MS}} := \hat{\sigma}^2_{\hat{s},\bar{m}} \ , \text{ where } \quad \bar{m} \in \underset{m\in[M_2]}{\operatorname{argmin}} R(\hat{\sigma}^2_{\hat{s},m}). \tag{9}$$

We consider the following decomposition

$$R(\hat{\sigma}^2_{\texttt{MS}}) - R(\sigma^2) = \underbrace{R(\hat{\sigma}^2_{\texttt{MS}}) - R(\bar{\sigma}^2_{\texttt{MS}})}_{\text{estimation error}} + \underbrace{R(\bar{\sigma}^2_{\texttt{MS}}) - R(\sigma^2)}_{\text{approximation error}} . \tag{10}$$

Each of these errors is obviously positive. The random term $R(\hat{\sigma}^2_{\texttt{MS}}) - R(\bar{\sigma}^2_{\texttt{MS}})$ is called the estimation error (or the variance). It measures how close $\hat{\sigma}^2_{\texttt{MS}}$ is to the best possible rule in $[M_2]$ in terms of the risk R. The deterministic term $R(\bar{\sigma}^2_{\texttt{MS}}) - R(\sigma^2)$ is called the approximation error (or the bias). We start with the following lemma

Lemma 1. *Let $\bar{\sigma}^2_{MS}$ be an aggregate defined in Equation* (9)*. Then,*

$$\mathbb{E}\left[R(\bar{\sigma}^2_{MS}) - R(\sigma^2)\right] = \mathbb{E}\left[\min_{m\in[M_2]} \mathbb{E}_X\left[|\hat{\sigma}^2_{\hat{s},m}(X) - \sigma^2(X)|^2\right]\right].$$

This result explicitly determines the approximation error.

Proof of Lemma 1. For all $m \in [M_2]$, the excess risk of $\hat{\sigma}^2_{\hat{s},m}$ is given as follows

$$\mathbb{E}\left[|Z - \hat{\sigma}^2_{\hat{s},m}(X)|^2\right] - \mathbb{E}\left[|Z - \sigma^2(X)|^2\right] = \mathbb{E}_X\left[|\hat{\sigma}^2_{\hat{s},m}(X) - \sigma^2(X)|^2\right]. \tag{11}$$

We apply *min* in Equation (11) and we get

$$\mathbb{E}\left[R(\bar{\sigma}^2_{\mathtt{MS}}) - R(\sigma^2)\right] \quad = \quad \mathbb{E}\left[\min_{m\in[M_2]} \mathbb{E}_X\left[|\hat{\sigma}^2_{\hat{s},m}(X) - \sigma^2(X)|^2\right]\right].$$

□

Proof of Theorem 1. We thank the decomposition in Eq. (10), we have

$$\mathbb{E}\left[|\hat{\sigma}^2_{\mathtt{MS}}(X) - \sigma^2(X)|^2\right] = \mathbb{E}\left[R(\hat{\sigma}^2_{\mathtt{MS}}) - R(\bar{\sigma}^2_{\mathtt{MS}})\right] + \mathbb{E}\left[R(\bar{\sigma}^2_{\mathtt{MS}}) - R(\sigma^2)\right]. \tag{12}$$

Step 1. Study of the term $\mathbb{E}\left[R(\bar{\sigma}^2_{\mathtt{MS}}) - R(\sigma^2)\right]$. We begin with the following Lemma

Lemma 2. *Let $\hat{s}$ and s^* be two estimators defined in* (1) *and* (7)*, respectively. Then, under Assumptions 1, 2, 3 and 4, there exists an absolute constant C such that*

$$\mathbb{P}\left(\hat{s} \neq s^*\right) \leq C\left(\frac{\log(M_1)}{N}\right)^{1/2}.$$

Proof. Under Assumption 4, we have firstly $\delta^*\left(\mathcal{D}_n\right) = \min_{s\neq s^*}\left\{|\mathcal{R}(\hat{f}_{s^*}) - \mathcal{R}(\hat{f}_s)|\right\} > \delta_0 > 0$. Recall that $\mathcal{R}(\hat{f}_{s^*}) \leq \mathcal{R}(\hat{f}_{\hat{s}})$. On the event $\{\hat{s} \neq s^*\}$, we have two cases

- $\hat{\mathcal{R}}_N(\hat{f}_{\hat{s}}) < \mathcal{R}(\hat{f}_{s^*})$, and then
$$\delta^*\left(\mathcal{D}_n\right) \leq |\mathcal{R}(\hat{f}_{\hat{s}}) - \mathcal{R}(\hat{f}_{s^*})| \leq |\hat{\mathcal{R}}_N(\hat{f}_{s^*}) - \mathcal{R}(\hat{f}_{s^*})| \leq \max_{s\in[M_1]} |\hat{\mathcal{R}}_N(\hat{f}_s) - \mathcal{R}(\hat{f}_s)|.$$
- $\hat{\mathcal{R}}_N(\hat{f}_{\hat{s}}) \geq \mathcal{R}(\hat{f}_{s^*})$, and then
$$\delta^*\left(\mathcal{D}_n\right) \leq |\hat{\mathcal{R}}_N(\hat{f}_{\hat{s}}) - \mathcal{R}(\hat{f}_{\hat{s}})| + |\hat{\mathcal{R}}_N(\hat{f}_{s^*}) - \mathcal{R}(\hat{f}_{s^*})| \leq 2\max_{s\in[M_1]} |\hat{\mathcal{R}}_N(\hat{f}_s) - \mathcal{R}(\hat{f}_s)|.$$

Therefore,

$$\mathbb{P}\left(\hat{s} \neq s^*\right) \leq \mathbb{P}\left(\max_{s\in[M_1]} |\hat{\mathcal{R}}_N(\hat{f}_s) - \mathcal{R}(\hat{f}_s)| \geq \delta_0/2\right)$$

We control this term using Bernstein's inequality. We check that the conditions for Bernstein's inequality are satisfied. For all $s \in [M_1]$, set $V_i(s) = |Y_i - \hat{f}_s(X_i)|^2 = |f^*(X_i) - \hat{f}_s(X_i) + \sigma(X_i)\xi_i|^2$ for all $i = 1, \ldots, N$. First, Assumptions 1 and 3 ensure that there exist a positive constants L_1 and L_2 such that $|f^*(X) - \hat{f}_s(X)| \leq L_1$ and $|\sigma^2(X)| \leq L_2$. Second, note that since the variables $V_i(s)$ are i.i.d. and by the elementary inequality $(x+y)^4 \leq 2^3(x^4+y^4)$ for all $x, y \in \mathbb{R}$, by Lemma 4, and by the elementary inequality $x^4 + y^4 \leq (x+y)^4$ for all $x, y \geq 0$ we have

$$\sum_{i=1}^{N} \mathbb{E}\left[V_i^2(s)\right] \leq 2^3 \sum_{i=1}^{N} \mathbb{E}\left[|f^*(X) - \hat{f}_s(X)|^4 + \sigma^4(X_i)\xi_i^4\right] \leq 2^7 N(L_1 + \sqrt{L_2})^4 := v_N \ ,$$

and for $k \geq 3$ we follow the elementary inequality $(x+y)^{2k} \leq 2^{2k-1}(x^{2k}+y^{2k})$ for all $x, y \in \mathbb{R}$, Lemma 4, and the following elementary inequality $x^{2k} + y^{2k} \leq (x+y)^{2k}$ for all $x, y \geq 0$

$$\begin{aligned}
\sum_{i=1}^{N} \mathbb{E}\left[(V_i^k(s) \vee 0)\right] &= \sum_{i=1}^{N} \mathbb{E}\left[|f^*(X_i) - \hat{f}_s(X_i) + \sigma(X_i)\xi_i|^{2k}\right] \\
&\leq 2^{2k-1} \sum_{i=1}^{N} \mathbb{E}\left[|f^*(X_i) - \hat{f}_s(X_i)|^{2k} + |\sigma^2(X_i)|^k |\xi_i|^{2k}\right] \\
&\leq \frac{1}{2} 2^{2k} N\left(L_1^{2k} + 2^{k+1}(\sqrt{L_2})^{2k}(k)!\right) \\
&\leq \frac{1}{2} 2^{3k+1} N\left(L_1 + \sqrt{L_2}\right)^{2k} k! \\
&\leq \frac{1}{2} v_N c^{k-2} k! \ .
\end{aligned}$$

where $c := 8\left(L_1 + \sqrt{L_2}\right)^2$. Using the Bernstein's inequality (Lemma 7), we get for all $s \in [M_1]$

$$\mathbb{P}\left(|\hat{\mathcal{R}}_N(\hat{f}_s) - \mathcal{R}(\hat{f}_s)| \geq \frac{\delta_0}{2}\right) \leq 2\exp\left(-\frac{N\delta_0^2}{2^{10}(L_1+\sqrt{L_2})^4 + 4c\delta_0}\right)$$

By union bound on $s \in [M_1]$, we obtain

$$\mathbb{P}\left(\hat{s} \neq s^*\right) \leq 2\exp\left(\log(M_1) - \frac{N\delta_0^2}{2^{10}(L_1+\sqrt{L_2})^4 + 4c\delta_0}\right) \leq C\left(\frac{\log(M_1)}{N}\right)^{1/2} \quad ,$$

where C is a positive constant which depends on L_1, L_2 and δ_0. □

By Lemmas 1 and 2, and under Assumptions 1 and 3 we get

$$\begin{aligned}\mathbb{E}\left[R(\bar{\sigma}^2_{\mathtt{MS}}) - R(\sigma^2)\right] &= \mathbb{E}\left[\min_{m\in[M_2]} \mathbb{E}_X\left[|\hat{\sigma}^2_{\hat{s},m}(X) - \sigma^2(X)|^2\left\{\mathbb{1}_{\{\hat{s}=s^*\}} + \mathbb{1}_{\{\hat{s}\neq s^*\}}\right\}\right]\right] \\ &\leq \mathbb{E}\left[\min_{m\in[M_2]} \mathbb{E}_X\left[|\hat{\sigma}^2_{s^*,m}(X) - \sigma^2(X)|^2\right]\right] + C\left(\frac{\log(M_1)}{N}\right)^{1/2}\end{aligned}$$

where C is a constant which depends on K_2 , σ^2 and the constant in Lemma 2.

Step 2. Study of the term $\mathbb{E}\left[R(\hat{\sigma}^2_{\mathtt{MS}}) - R(\bar{\sigma}^2_{\mathtt{MS}})\right]$. To treat the estimation error, we introduce an aggregate $\tilde{\sigma}^2_{\mathtt{MS}}$ which is based on minimisation of the empirical risk of R

$$\tilde{\sigma}^2_{\mathtt{MS}} := \hat{\sigma}^2_{\hat{s},\tilde{m}} \quad , \text{ where } \quad \tilde{m} \in \operatorname*{argmin}_{m\in[M_2]} R_N(\hat{\sigma}^2_{\hat{s},m}) \quad ,$$

with $R_N(\hat{\sigma}^2_{\hat{s},m}) = \frac{1}{N}\sum_{i=1}^N |Z_i - \hat{\sigma}^2_{\hat{s},m}(X_i)|^2$. Moreover, we consider the decomposition

$$\mathbb{E}\left[R(\hat{\sigma}^2_{\mathtt{MS}}) - R(\bar{\sigma}^2_{\mathtt{MS}})\right] = \mathbb{E}\left[R(\hat{\sigma}^2_{\mathtt{MS}}) - R(\tilde{\sigma}^2_{\mathtt{MS}})\right] + \mathbb{E}\left[R(\tilde{\sigma}^2_{\mathtt{MS}}) - R(\bar{\sigma}^2_{\mathtt{MS}})\right].$$

Step 2.1. Study of the term $\mathbb{E}\left[R(\tilde{\sigma}^2_{\mathtt{MS}}) - R(\bar{\sigma}^2_{\mathtt{MS}})\right]$. We decompose the term $\mathbb{E}\left[R(\tilde{\sigma}^2_{\mathtt{MS}}) - R(\bar{\sigma}^2_{\mathtt{MS}})\right]$ into two positive terms

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{\mathtt{MS}}) - R(\bar{\sigma}^2_{\mathtt{MS}})\right] = \mathbb{E}\left[R(\tilde{\sigma}^2_{\mathtt{MS}}) - R_N(\tilde{\sigma}^2_{\mathtt{MS}})\right] + \mathbb{E}\left[R_N(\tilde{\sigma}^2_{\mathtt{MS}}) - R(\bar{\sigma}^2_{\mathtt{MS}})\right]. \tag{13}$$

We use the fact that $R_N(\tilde{\sigma}^2_{\mathtt{MS}}) \leq R_N(\bar{\sigma}^2_{\mathtt{MS}})$ in Eq. (13), and we get the uniform bound

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{\mathtt{MS}}) - R(\bar{\sigma}^2_{\mathtt{MS}})\right] \leq 2\mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]} |R_N(\hat{\sigma}^2_{s,m}) - R(\hat{\sigma}^2_{s,m})|\right].$$

Then using Assumption 2, for some $(s,m) \in [M_1]\times[M_2]$, set $T_i(s,m) = |Z_i - \hat{\sigma}^2_{s,m}(X_i)|^2 = |\sigma^2(X_i)\xi_i^2 - \hat{\sigma}^2_{s,m}(X_i)|^2$ for all $i = 1,\dots,N$. First, note that since the variables $T_i(s,m)$ are i.i.d. , conditionally on $\mathcal{D}_n$ we have

$$\begin{aligned}|R_N(\hat{\sigma}^2_{s,m}) - R(\hat{\sigma}^2_{s,m})| &= \left|\frac{1}{N}\sum_{i=1}^N (T_i(s,m) - \mathbb{E}[T_i(s,m)])\right| \\ &\leq \left|\frac{1}{N}\sum_{i=1}^N (T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|\leq L\}}\right| \\ &\quad + \left|\frac{1}{N}\sum_{i=1}^N (T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|> L\}}\right|\end{aligned}$$

for any $L > 0$. Therefore, conditionally on $\mathcal{D}_n$

$$\begin{aligned}\mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]} |R_N(\hat{\sigma}^2_{s,m}) - R(\hat{\sigma}^2_{s,m})|\right] &\leq \mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]} \left|\frac{1}{N}\sum_{i=1}^N (T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|\leq L\}}\right|\right] \\ &\quad + \mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]} \left|\frac{1}{N}\sum_{i=1}^N (T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|> L\}}\right|\right].\end{aligned} \tag{14}$$

Step 2.1.1. We control the first term on the r.h.s. of Eq. (14). On the event $\{|\xi| \leq L\}$ and under Assumptions 1 and 3, we get $|T_i(s,m)| \leq c_1 L^4 + 2K_2^2$ for all $i = 1, \dots, N$ for some $c_1 > 0$ that depends on σ^2. Conditionally on $\mathcal{D}_n$, we apply Hoeffding's inequality, for all $(s,m) \in [M_1] \times [M_2]$, and all $t \geq 0$

$$\mathbb{P}\left(\left| \frac{1}{N} \sum_{i=1}^{N} (T_i(s,m) - \mathbb{E}[T_i(s,m)]) \mathbb{1}_{\{|\xi_i| \leq L\}} \right| \geq t \right) \leq 2 \exp\left(- \frac{Nt^2}{2(c_1 L^4 + 2K_2^2)^2} \right) ,$$

Conditionally on $\mathcal{D}_n$, by a union bound on $(s,m) \in [M_1] \times [M_2]$, we deduce that for all $t \geq 0$

$$\mathbb{P}\left(\max_{(s,m) \in [M_1] \times [M_2]} \left| \frac{1}{N} \sum_{i=1}^{N} (T_i(s,m) - \mathbb{E}[T_i(s,m)]) \mathbb{1}_{\{|\xi_i| \leq L\}} \right| \geq t \right) \leq 2 \exp\left(\log(M_1 M_2) - \frac{Nt^2}{2(c_1 L^4 + 2K_2^2)^2} \right).$$

We apply Lemma 6. Then, there exists a positive constant $\mathbf{c}$ such that

$$\mathbb{E}\left[\max_{(s,m) \in [M_1] \times [M_2]} \left| \frac{1}{N} \sum_{i=1}^{N} (T_i(s,m) - \mathbb{E}[T_i(s,m)]) \mathbb{1}_{\{|\xi_i| \leq L\}} \right| \right] \leq \mathbf{c} \left(c_2 L^4 + c_3 \right) \left(\frac{\log(M_1 M_2)}{N} \right)^{1/2} ,$$

where c_2 is a positive constant that depends on c_1 and c_3 depends on K_2.

Step 2.1.2. We control the second term on the r.h.s. of Eq. (14). By union bound on $(s,m) \in [M_1] \times [M_2]$, by Cauchy–Schwarz inequality, under Assumptions 1, 2 and 3, and Lemma 3 we obtain

$$\begin{aligned} & \mathbb{E}\left[\max_{(s,m) \in [M_1] \times [M_2]} \left| \frac{1}{N} \sum_{i=1}^{N} (T_i(s,m) - \mathbb{E}[T_i(s,m)]) \mathbb{1}_{\{|\xi_i| > L\}} \right| \right] \\ \leq & \sum_{s=1}^{M_1} \sum_{m=1}^{M_2} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[|T_i(s,m) - \mathbb{E}[T_i(s,m)]| \mathbb{1}_{\{\xi_i > L\}}] \\ \leq & \sum_{s=1}^{M_1} \sum_{m=1}^{M_2} \frac{1}{N} \sum_{i=1}^{N} \sqrt{\mathbb{E}[|T_i(s,m) - \mathbb{E}[T_i(s,m)]|^2] \mathbb{P}(|\xi_i| > L)]} \\ \leq & c M_1 M_2 \sqrt{\mathbb{P}(|\xi_1| > L)} \\ \leq & c M_1 M_2 \frac{\exp(-L^2/4)}{L^{1/2}} \quad , \end{aligned}$$

where c is a positive constant which depends on ξ, σ^2 and K_2.
Combining the results of the **Step 2.1.1** and **Step 2.1.2** in Eq. (14)

$$\mathbb{E}\left[\max_{(s,m) \in [M_1] \times [M_2]} |R_N(\hat{\sigma}^2_{s,m}) - R(\hat{\sigma}^2_{s,m})| \right] \leq \mathbf{c}(c_2 L^4 + c_3) \left(\frac{\log(M_1 M_2)}{N} \right)^{1/2} + c M_1 M_2 \frac{\exp(-L^2/4)}{L^{1/2}}.$$

Choosing $L = 2\sqrt{\log(N)}$ and we get

$$\mathbb{E}\left[\max_{(s,m) \in [M_1] \times [M_2]} |R_N(\hat{\sigma}^2_{s,m}) - R(\hat{\sigma}^2_{s,m})| \right] \leq C \left(\frac{\log(N)^4 \log(M_1 M_2)}{N} \right)^{1/2} ,$$

where C is a positive constant that depends on c_2 and $\mathbf{c}$, and **Step 2.1.2** is finished.

We combine the results of the **Step 2.1.1** and **Step 2.1.2** and we get the following bound

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{\mathtt{MS}}) - R(\bar{\sigma}^2_{\mathtt{MS}}) \right] \leq 2C \left(\frac{\log(N)^4 \log(M_1 M_2)}{N} \right)^{1/2} .$$

Remark 1. *It is clear that when Y is bounded, there exists an absolute constant $C > 0$ such that*

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{\mathit{MS}}) - R(\bar{\sigma}^2_{\mathit{MS}}) \right] \leq C \left(\frac{\log(M_1 M_2)}{N} \right)^{1/2} .$$

Step 2.2. Study of the term $\mathbb{E}\left[R(\hat{\sigma}^2_{\mathtt{MS}}) - R(\tilde{\sigma}^2_{\mathtt{MS}})\right]$. We start with the following decomposition

$$\mathbb{E}\left[R(\hat{\sigma}^2_{\mathtt{MS}}) - R(\tilde{\sigma}^2_{\mathtt{MS}})\right] = \mathbb{E}\left[R(\hat{\sigma}^2_{\mathtt{MS}}) - R_N(\hat{\sigma}^2_{\mathtt{MS}})\right] + \mathbb{E}\left[R_N(\hat{\sigma}^2_{\mathtt{MS}}) - R_N(\tilde{\sigma}^2_{\mathtt{MS}})\right] + \mathbb{E}\left[R_N(\tilde{\sigma}^2_{\mathtt{MS}}) - R(\tilde{\sigma}^2_{\mathtt{MS}})\right] . \quad (15)$$

We use the same arguments in **Step 2.1.** to control the first term and the last term on the r.h.s. of Eq. (15), and we get the following bound

$$\begin{aligned}\mathbb{E}\left[R(\hat{\sigma}^2_{\mathtt{MS}}) - R_N(\hat{\sigma}^2_{\mathtt{MS}})\right] + \mathbb{E}\left[R_N(\tilde{\sigma}^2_{\mathtt{MS}}) - R(\tilde{\sigma}^2_{\mathtt{MS}})\right] &\leq 2\mathbb{E}\left[\max_{(s,m)\in[M_1]\times[M_2]} |R_N(\hat{\sigma}^2_{s,m}) - R(\hat{\sigma}^2_{s,m})|\right] \\ &\leq C\left(\frac{\log(N)^4\log(M_1M_2)}{N}\right)^{1/2} .\end{aligned}$$

Remark 2. *If Y is bounded, there exists an absolute constant $C > 0$ such that*

$$\mathbb{E}\left[R(\hat{\sigma}^2_{MS}) - R_N(\hat{\sigma}^2_{MS})\right] + \mathbb{E}\left[R_N(\tilde{\sigma}^2_{MS}) - R(\tilde{\sigma}^2_{MS})\right] \leq C\left(\frac{\log(M_1M_2)}{N}\right)^{1/2} .$$

We now study the second term on the r.h.s. of Eq. (15). For that, we need the following decomposition

$$\mathbb{E}\left[R_N(\hat{\sigma}^2_{\mathtt{MS}}) - R_N(\tilde{\sigma}^2_{\mathtt{MS}})\right] = \mathbb{E}\left[R_N(\hat{\sigma}^2_{\mathtt{MS}}) - \hat{R}_N(\hat{\sigma}^2_{\mathtt{MS}})\right] + \mathbb{E}\left[\hat{R}_N(\hat{\sigma}^2_{\mathtt{MS}}) - R_N(\tilde{\sigma}^2_{\mathtt{MS}})\right] . \quad (16)$$

Using $\hat{R}_N(\hat{\sigma}^2_{\mathtt{MS}}) \leq \hat{R}_N(\tilde{\sigma}^2_{\mathtt{MS}})$ in Eq. (16), we obtain the following inequality

$$\mathbb{E}\left[R_N(\hat{\sigma}^2_{\mathtt{MS}}) - R_N(\tilde{\sigma}^2_{\mathtt{MS}})\right] \leq 2\mathbb{E}\left[\max_{m\in[M_2]} |\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})|\right] .$$

We control the term $\mathbb{E}\left[\max_{m\in[M_2]} |\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})|\right]$. By definition of $\hat{R}_N$ and R_N, and under Assumption 3, we get for all $m \in [M_2]$

$$|\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})| \quad \leq \quad \frac{1}{N}\sum_{i=1}^N |\hat{Z}_i - Z_i|^2 + \frac{2}{N}\sum_{i=1}^N |\hat{Z}_i - Z_i|(|Z_i| + K_2) \ ,$$

where K_2 is the bound of $\hat{\sigma}^2_{\hat{s},m}$. The upper-bound of $|\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})|$ does not depend on m, therefore

$$\mathbb{E}\left[\max_{m\in[M_2]} |\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})|\right] \leq \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N |\hat{Z}_i - Z_i|^2\right] + 2\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N |\hat{Z}_i - Z_i|(|Z_i| + K_2)\right] . \quad (17)$$

Note that, by Assumptions 1 and 3 we obtain for all $i = 1, \ldots, N$

$$|f^*(X_i) - \hat{f}_{\mathtt{MS}}(X_i)| \leq \|f^*\|_\infty + \max_{s\in[M_1]} \|\hat{f}_s\|_\infty \leq \|f^*\|_\infty + K_1 \leq L_1 < \infty.$$

Since $x^2 - y^2 = (x-y)(x+y)$, $(x+y)^2 \leq 2(x^2+y^2)$, we obtain the following inequality for all $i = 1, \ldots, N$

$$\begin{aligned}|\hat{Z}_i - Z_i|^2 &= |(Y_i - \hat{f}_{\mathtt{MS}}(X_i))^2 - (Y_i - f^*(X_i))^2|^2 \\ &= |(f^*(X_i) - \hat{f}_{\mathtt{MS}}(X_i))(2(Y_i - f^*(X_i)) + (f^*(X_i) - \hat{f}_{\mathtt{MS}}(X_i))|^2 \\ &\leq |f^*(X_i) - \hat{f}_{\mathtt{MS}}(X_i)|^2|(8|Y_i - f^*(X_i)|^2 + 2L_1^2| \ , \quad (18)\end{aligned}$$

Control of $\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N |\hat{Z}_i - Z_i|^2\right]$. First, since Assumptions 1-2 are satisfied, we have that for all $i = 1, \ldots, N$, $\mathbb{E}\left[|Y_i - f^*(X_i)|^4\right] \leq k_1 < \infty$. Second, by inequality (18), Cauchy-Schwarz inequality, Jensen's inequality, and under Assumptions 1, and 2, one gets

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}|\hat{Z}_i - Z_i|^2\right] &\leq 2L_1^2\mathbb{E}\left[\|\hat{f}_{\texttt{MS}} - f^*\|_N^2\right] + \frac{8}{N}\sum_{i=1}^{N}\mathbb{E}\left[|Y_i - f^*(X_i)|^2|f^*(X_i) - \hat{f}_{\texttt{MS}}(X_i)|^2\right] \\
&\leq 2L_1^2\mathbb{E}\left[\|\hat{f}_{\texttt{MS}} - f^*\|_N^2\right] + \frac{8}{N}\sum_{i=1}^{N}\sqrt{\mathbb{E}\left[|Y_i - f^*(X_i)|^4\right]}\sqrt{\mathbb{E}\left[|f^*(X_i) - \hat{f}_{\texttt{MS}}(X_i)|^4\right]} \\
&\leq 2L_1^2\mathbb{E}\left[\|\hat{f}_{\texttt{MS}} - f^*\|_N^2\right] + \frac{8\sqrt{k_1}L_1}{N}\sum_{i=1}^{N}\sqrt{\mathbb{E}\left[|f^*(X_i) - \hat{f}_{\texttt{MS}}(X_i)|^2\right]} \\
&\leq 2L_1^2\mathbb{E}\left[\|\hat{f}_{\texttt{MS}} - f^*\|_N^2\right] + 8\sqrt{k_1}L_1\sqrt{\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}|f^*(X_i) - \hat{f}_{\texttt{MS}}(X_i)|^2\right]} \\
&\leq C_1\sqrt{\mathbb{E}\left[\|\hat{f}_{\texttt{MS}} - f^*\|_N^2\right]}\ ,
\end{aligned}$$

where C_1 is a positive constant that depends on k_1 and L_1.

Control of $\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}|\hat{Z}_i - Z_i|(|Z_i| + K_2)\right]$. First, since Assumptions 1-2 are satisfied, we have that for all $i = 1, \ldots, N$, $\mathbb{E}\left[(|Y_i - f^*(X_i)|^2 + K_2)^2\right] \leq k_2 < \infty$. Second, by Cauchy-Schwarz inequality and Jensen's inequality, one gets

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[|\hat{Z}_i - Z_i|(|Z_i| + K_2)\right] &\leq \frac{1}{N}\sum_{i=1}^{N}\sqrt{\mathbb{E}\left[|\hat{Z}_i - Z_i|^2\right]}\sqrt{\mathbb{E}\left[(|Y_i - f^*(X_i)|^2 + K_1)^2\right]} \\
&\leq \frac{\sqrt{k_2}}{N}\sum_{i=1}^{N}\sqrt{\mathbb{E}\left[|\hat{Z}_i - Z_i|^2\right]} \\
&\leq \sqrt{k_2}\sqrt{\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}|\hat{Z}_i - Z_i|^2\right]} \\
&\leq C_2\mathbb{E}\left[\|\hat{f}_{\texttt{MS}} - f^*\|_N^2\right]^{1/4}\ ,
\end{aligned}$$

where C_2 is a positive constant that depends on C_1 and k_2. Thus, there exists an absolute constant C such that

$$\mathbb{E}\left[\max_{m\in[M_2]}|\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})|\right] \leq C\mathbb{E}\left[\|\hat{f}_{\texttt{MS}} - f^*\|_N^2\right]^{1/4}.$$

We need the following proposition:

Proposition 1. *Let $\hat{f}_{MS}$ the aggregate defined in Eq.* (2)*. Then, under Assumptions 2 and 3 there exists an absolute constant C such that*

$$\mathbb{E}\left[\|\hat{f}_{MS} - f^*\|_N^2\right] \leq \min_{s\in[M_1]}\mathbb{E}\left[\|\hat{f}_s - f^*\|_N^2\right] + C\left(\frac{\log(M_1)}{N}\right)^{1/2}.$$

This result studies the upper-bound of empirical norm risk of the aggregate $\hat{f}_{\texttt{MS}}$ and the proof of it exists in Tsybakov (2014). Besides, the Proposition 1 and the elementary inequality $(x + y)^{1/4} \leq x^{1/4} + y^{1/4}$ for all $x, y \geq 0$ give us the following inequality

$$\mathbb{E}\left[R(\hat{\sigma}^2_{\texttt{MS}}) - R(\tilde{\sigma}^2_{\texttt{MS}})\right] \leq C^{'}\left\{\min_{s\in[M_1]}\mathbb{E}\left[\|\hat{f}_s - f^*\|_N^2\right]\right\}^{1/4} + C^{"}\left(\frac{\log(M_1)}{N}\right)^{1/8}\ ,$$

where $C^{'}$ is a constant which depends on C_2 and $C^{"}$ is a constant which depends on C_2 and the constant in Proposition 1.

Merging the results of the **Step 1** and **Step 2** in Eq. (12) and we get the result.

Remark 3. *In the case where Y is bounded and from Eq.* (18)*, we observe that there exists a constant C_3 such that*

$$|\hat{Z}_i - Z_i|^2 \leq C_3 |f^*(X_i) - \hat{f}_{MS}(X_i)|^2 \ . \tag{19}$$

By Jensen's inequality twice an inequality (17) *and from Eq.*(19)*, one gets there exists an absolute constant C_4 such that*

$$\mathbb{E}\left[\max_{m\in[M_2]} |\hat{R}_N(\hat{\sigma}^2_{\hat{s},m}) - R_N(\hat{\sigma}^2_{\hat{s},m})|\right] \leq C_4 \mathbb{E}\left[\|\hat{f}_{MS} - f^*\|_N^2\right]^{1/2}. \tag{20}$$

Finally, we apply Proposition 1 in Eq.(20) *to get the result.*

□

B Proof of Proposition 1

From the definition of MS-estimator $\hat{f}_{\mathtt{MS}}$, we get by a simple algebra that, for any $s \in [M_1]$

$$\|\hat{f}_{\mathtt{MS}} - f^*\|_N^2 \leq \|\hat{f}_s - f^*\|_N^2 + 2 < \hat{f}_{\mathtt{MS}} - \hat{f}_s, Y - f^* >,$$

where $< \hat{f}_{\mathtt{MS}} - \hat{f}_s, Y - f^* > := \frac{1}{N}\sum_{i=1}^N \Big((\hat{f}_{\mathtt{MS}}(X_i) - \hat{f}_s(X_i))(Y_i - f^*(X_i)) \Big)$. Therefore, one gets for any $s \in [M_1]$

$$\mathbb{E}\left[\|\hat{f}_{\mathtt{MS}} - f^*\|_N^2\right] \leq \mathbb{E}\left[\|\hat{f}_s - f^*\|_N^2\right] + 2\mathbb{E}\left[< \hat{f}_{\mathtt{MS}} - \hat{f}_s, Y - f^* >\right]. \tag{21}$$

We control the second term in the r.h.s. of Eq (21). Firstly, we notice that

$$\mathbb{E}\left[< \hat{f}_{\mathtt{MS}} - \hat{f}_s, Y - f^* >\right] \leq \mathbb{E}\left[\max_{1\leq j\leq M_1} < \hat{f}_j - \hat{f}_s, Y - f^* >\right].$$

Secondly, since $Y - f^*$ is ρ-subgaussian where ρ is a positive constant that depends on $Y - f^*$, then the variables $< \hat{f}_j - \hat{f}_s, Y - f^* >$ is $\bar{\rho}$-subgaussian where $\bar{\rho}^2 = \frac{\rho^2\|\hat{f}_j - \hat{f}_s\|_N^2}{N}$. Moreover, under Assumption 3, it is clear that $\max_{1\leq j\leq M_1} \|\hat{f}_j - \hat{f}_s\|_N^2 \leq B$ where B is a constant which depends on K_1. Therefore, we use Lemma 5 and we get

$$\mathbb{E}\left[\max_{1\leq j\leq M_1} < \hat{f}_j - \hat{f}_s, Y - f^* >\right] \leq \rho\sqrt{B}\sqrt{\frac{2\log(M_1)}{N}}.$$

Thus,

$$\mathbb{E}\left[\|\hat{f}_{\mathtt{MS}} - f^*\|_N^2\right] \leq \min_{s\in[M_1]} \mathbb{E}\left[\|\hat{f}_s - f^*\|_N^2\right] + 2\rho\sqrt{B}\sqrt{\frac{2\log(M_1)}{N}}.$$

C Proof of Theorem 2

We introduce the following aggregates

$$\tilde{\sigma}^2_{\mathtt{C}} := \hat{\sigma}^2_{\hat{\lambda},\tilde{\beta}} \ , \text{ where } \ \tilde{\beta} \in \operatorname*{argmin}_{\beta\in\Lambda^{M_2}} R_N(\hat{\sigma}^2_{\hat{\lambda},\beta}) \ ,$$

and

$$\bar{\sigma}^2_{\mathtt{C}} := \hat{\sigma}^2_{\hat{\lambda},\bar{\beta}} \ , \text{ where } \ \bar{\beta} \in \operatorname*{argmin}_{\beta\in\Lambda^{M_2}} R(\hat{\sigma}^2_{\hat{\lambda},\beta}).$$

Consider the following decomposition

$$\mathbb{E}\left[|\hat{\sigma}^2_{\mathtt{C}}(X) - \sigma^2(X)|^2\right] = \mathbb{E}\left[R(\hat{\sigma}^2_{\mathtt{C}}) - R(\tilde{\sigma}^2_{\mathtt{C}})\right] + \mathbb{E}\left[R(\tilde{\sigma}^2_{\mathtt{C}}) - R(\bar{\sigma}^2_{\mathtt{C}})\right] + \mathbb{E}\left[R(\bar{\sigma}^2_{\mathtt{C}}) - R(\sigma^2)\right]. \tag{22}$$

Step 1. Study of the term $\mathbb{E}\left[R(\bar{\sigma}^2_{\mathtt{C}}) - R(\sigma^2)\right]$. We use the same proof of Lemma 1, and we get

$$\mathbb{E}\left[R(\bar{\sigma}^2_{\mathtt{C}}) - R(\sigma^2)\right] \leq \mathbb{E}\left[\inf_{\beta \in \Lambda^{M_2}} \mathbb{E}_X\left[|\hat{\sigma}^2_{\lambda,\beta}(X) - \sigma^2(X)|\right]\right].$$

Step 2. Study of the term $\mathbb{E}\left[R(\tilde{\sigma}^2_{\mathtt{C}}) - R(\bar{\sigma}^2_{\mathtt{C}})\right]$. We use the fact that $R_N(\tilde{\sigma}^2_{\mathtt{C}}) \leq R_N(\bar{\sigma}^2_{\mathtt{C}})$, and we get the uniform bound

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{\mathtt{C}}) - R(\bar{\sigma}^2_{\mathtt{C}})\right] \leq 2\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}\times\Lambda^{M_2}} |R_N(\hat{\sigma}^2_{\lambda,\beta}) - R(\hat{\sigma}^2_{\lambda,\beta})|\right].$$

Since Λ^{M_2} (resp. Λ^{M_1}) is compact, we have $\Lambda^{M_2} \subset \bar{B}(0,1)$ (the closed unit ball) (resp. $\Lambda^{M_1} \subset \bar{B}(0,1)$), and there exists an ϵ_2-net $\Lambda^{M_2}_{\epsilon_2}$ of Λ^{M_2} (resp. an ϵ_1-net $\Lambda^{M_1}_{\epsilon_1}$ of Λ^{M_1}) *w.r.t.* $\|\cdot\|_{1,M_2}$ (resp. $\|\cdot\|_{1,M_1}$) such that $|\Lambda^{M_2}_{\epsilon_2}| \leq (3/\epsilon_2)^{M_2}$ (resp. $|\Lambda^{M_1}_{\epsilon_1}| \leq (3/\epsilon_1)^{M_1}$). In particular, for all $\beta \in \Lambda^{M_2}$ (resp. $\lambda \in \Lambda^{M_1}$) there exists $\beta^{\epsilon_2} \in \Lambda^{M_2}_{\epsilon_2}$ (resp. $\lambda^{\epsilon_1} \in \Lambda^{M_1}_{\epsilon_1}$) such that $\|\beta - \beta^{\epsilon_2}\|_{1,M_2} \leq \epsilon_2$ (resp. $\|\lambda - \lambda^{\epsilon_1}\|_{1,M_1} \leq \epsilon_1$). From triangle inequality, one gets

$$\begin{aligned}|R_N(\hat{\sigma}^2_{\lambda,\beta}) - R(\hat{\sigma}^2_{\lambda,\beta})| \leq |R_N(\hat{\sigma}^2_{\lambda,\beta}) - R_N(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})| + |R_N(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}) - R_N(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}})| + |R_N(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}})| \\ + |R(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})| + |R(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda,\beta})|.\end{aligned}$$

1. **Control of** $|R(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda,\beta})|$**.** By Jensen's inequality, under assumptions 1- 2- 5 and $\mathbb{E}[\xi^2] = 1$ we obtain

$$\begin{aligned}|R(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda,\beta})| &\leq \mathbb{E}\left[\left||Z - \hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}(X)|^2 - |Z - \hat{\sigma}^2_{\lambda,\beta}(X)|^2\right|\right] \\ &= \mathbb{E}\left[|\left(\sum_{j=1}^{M_2}\left(\beta_j - \beta_j^{\epsilon_2}\right)\hat{\sigma}^2_{\lambda,j}(X)\right)\left(2Z - \hat{\sigma}^2_{\lambda,\beta}(X) - \hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}(X)\right)|\right] \\ &\leq C_1\epsilon_2\ ,\end{aligned}$$

   where C_1 is a constant which depends on the upper bounds of σ^2 and $\hat{\sigma}^2_{\lambda,j}$.

2. **Control of** $|R_N(\hat{\sigma}^2_{\lambda,\beta}) - R_N(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})|$**.** Since Assumptions 1, 2, and 3 are satisfied, we obtain

$$\begin{aligned}|R_N(\hat{\sigma}^2_{\lambda,\beta}) - R_N(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})| &\leq \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{M_2}|\beta_j - \beta_j^{\epsilon_2}||\hat{\sigma}^2_{\lambda,j}(X_i)|\right)|2\sigma^2(X_i)\xi_i^2 - \hat{\sigma}^2_{\lambda,\beta}(X_i) - \hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}(X_i)| \\ &\leq k\epsilon_2\left(\frac{C_2}{N}\sum_{i=1}^{N}\xi_i^2 + C_3\right),\end{aligned}$$

   where k is the bound of $\hat{\sigma}^2_{\lambda,j}$, C_2 is the constant which depends on σ^2 and C_3 is the constant which depends on the upper bounds $\hat{\sigma}^2_{\lambda,\beta}$ and $\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}$.

3. **Control of** $|R(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})|$**.** Under Assumptions 1, 2, 5, and 6, we get

$$\begin{aligned}|R(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}) - R(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})| &\leq \mathbb{E}\left[\sum_{j=1}^{M_2}\beta_j^{\epsilon_2}|\hat{\sigma}^2_{\lambda,j}(X) - \hat{\sigma}^2_{\lambda^{\epsilon_1},j}(X)||2\sigma^2(X)\xi^2 - \hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}(X) - \hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}(X)|\right] \\ &\leq \sum_{j=1}^{M_2}\beta_j^{\epsilon_2}\Bigg(2\mathbb{E}\left[\mathbb{E}\left[|\hat{\sigma}^2_{\lambda,j}(X) - \hat{\sigma}^2_{\lambda^{\epsilon_1},j}(X)|\sigma^2(X)\xi^2|\mathcal{D}_n, X\right]\right] \\ &+ \mathbb{E}\left[|\hat{\sigma}^2_{\lambda,j}(X) - \hat{\sigma}^2_{\lambda^{\epsilon_1},j}(X)||\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}(X) - \hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}(X)|\right]\Bigg) \\ &\leq \sum_{j=1}^{M_2}\beta_j^{\epsilon_2}\Bigg(2\mathbb{E}\left[|\hat{\sigma}^2_{\lambda,j}(X) - \hat{\sigma}^2_{\lambda^{\epsilon_1},j}(X)|\sigma^2(X)\mathbb{E}\left[\xi^2\right]\right] \\ &+ \mathbb{E}\left[|\hat{\sigma}^2_{\lambda,j}(X) - \hat{\sigma}^2_{\lambda^{\epsilon_1},j}(X)||\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}(X) - \hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}(X)|\right]\Bigg) \\ &\leq C_4\epsilon_1,\end{aligned}$$

where C_4 is constant which depends on K and the upper bounds of σ^2, $\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}$ and $\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}$.

4. **Control of $|R_N(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}) - R_N(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})|$.** We use the same way as 3. and we obtain

$$|R_N(\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}) - R_N(\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}})| \leq \epsilon_1 \left(\frac{C_2}{N} \sum_{i=1}^{N} \xi_i^2 + C_5 \right),$$

where C_5 is constant which depends on the upper bounds of $\hat{\sigma}^2_{\lambda^{\epsilon_1},\beta^{\epsilon_2}}$ and $\hat{\sigma}^2_{\lambda,\beta^{\epsilon_2}}$.

Therefore, we deduce that

$$\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}\times\Lambda^{M_2}} |R_N(\hat{\sigma}^2_{\lambda,\beta}) - R(\hat{\sigma}^2_{\lambda,\beta})| \right] \leq C_{k,C_2,C_3,C_4,C_5}(\epsilon_1+\epsilon_2) + \mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} |R_N(\hat{\sigma}^2_{\lambda,\beta}) - R(\hat{\sigma}^2_{\lambda,\beta})| \right].$$

For some $(\lambda,\beta) \in \Lambda^{M_1}_{\epsilon_1} \times \Lambda^{M_2}_{\epsilon_2}$, set $T_i(\lambda,\beta) = |Z_i - \hat{\sigma}^2_{\lambda,\beta}(X_i)|^2 = |\sigma^2(X_i)\xi_i^2 - \hat{\sigma}^2_{\lambda,\beta}(X_i)|^2$ for all $i = 1,\ldots,N$. Let $L > 0$. Since the variables $T_i(\lambda,\beta)$ are i.i.d. , we have

$$\begin{aligned}\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} |R_N(\hat{\sigma}^2_{\lambda,\beta}) - R(\hat{\sigma}^2_{\lambda,\beta})| \right] &\leq \mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} \left| \frac{1}{N}\sum_{i=1}^{N} (T_i(\lambda,\beta) - \mathbb{E}[T_i(\lambda,\beta)])\mathbb{1}_{\{|\xi_i|\leq L\}} \right| \right] \\ &+ \mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} \left| \frac{1}{N}\sum_{i=1}^{N} (T_i(\lambda,\beta) - \mathbb{E}[T_i(\lambda,\beta)])\mathbb{1}_{\{|\xi_i|> L\}} \right| \right].\end{aligned} \tag{23}$$

Step 2.1. We control the first term on the r.h.s. of Eq. (23). On the event $\{|\xi| \leq L\}$ and under assumptions 1, 2 and 5, we get $|T_i(\lambda,\beta)| \leq c_1 L^4 + \bar{c}_1$ for all $i = 1,\ldots,N$ where c_1 is a positive constant which depends on the upper bound of σ^2 and $\bar{c}_1$ depends on the upper bound of $\hat{\sigma}^2_{\lambda,\beta}$. Conditionally on $\mathcal{D}_n$, we apply Hoeffding's inequality, for all $(\lambda,\beta) \in \Lambda^{M_1}_{\epsilon_1} \times \Lambda^{M_2}_{\epsilon_2}$, and all $t \geq 0$

$$\mathbb{P}\left(\left| \frac{1}{N}\sum_{i=1}^{N} (T_i(\lambda,\beta) - \mathbb{E}[T_i(\lambda,\beta)])\mathbb{1}_{\{|\xi_i|\leq L\}} \right| \geq t \right) \leq 2\exp\left(-\frac{-Nt^2}{2(c_1L^4+\bar{c}_1)^2} \right) \quad ,$$

By a union bound on $(\lambda,\beta) \in \Lambda^{M_1}_{\epsilon_1} \times \Lambda^{M_2}_{\epsilon_2}$ and choosing $\epsilon_1 = \epsilon_2 = \frac{3}{N}$, we deduce that for all $t \geq 0$

$$\mathbb{P}\left(\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} \left| \frac{1}{N}\sum_{i=1}^{N} (T_i(\lambda,\beta) - \mathbb{E}[T_i(\lambda,\beta)])\mathbb{1}_{\{|\xi_i|\leq L\}} \right| \geq t \right) \leq 2\exp\left((M_1+M_2)\log(N) - \frac{-Nt^2}{2(c_1L^4+\bar{c}_1)^2} \right).$$

We apply Lemma 6. Then, there exists a positive constant $\mathbf{c}$ such that

$$\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} \left| \frac{1}{N}\sum_{i=1}^{N} (T_i(\lambda,\beta) - \mathbb{E}[T_i(\lambda,\beta)])\mathbb{1}_{\{|\xi_i|\leq L\}} \right| \right] \leq \mathbf{c}(c_2L^4+\bar{c}_2)\left(\frac{(M_1+M_2)\log(N)}{N} \right)^{1/2},$$

where c_2 is constant which depends on c_1 and $\bar{c}_2$ on $\bar{c}_1$.

Step 2.2. We control the second term on the r.h.s. of Eq. (23). Thanks to the boundness of σ^2 and $\hat{\sigma}^2_{\lambda,\beta}$ and $\mathbb{E}[\xi^4] = 3$, we get $\mathbb{E}[T_i(\lambda,\beta)] \leq c_3$ and $T_i(\lambda,\beta) \leq c_4\xi_i^4 + c_5$ for all $i = 1,\ldots,N$ where c_4 and c_5 are constants which depend on the upper bounds of σ^2 and $\hat{\sigma}^2_{\lambda,\beta}$, respectively. By Cauchy–Schwarz inequality and Lemma 3, we obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda^{M_1}_{\epsilon_1}\times\Lambda^{M_2}_{\epsilon_2}} \left| \frac{1}{N}\sum_{i=1}^{N} (T_i(s,m) - \mathbb{E}[T_i(s,m)])\mathbb{1}_{\{|\xi_i|> L\}} \right| \right] &\leq \frac{c_4}{N}\sum_{i=1}^{N} \mathbb{E}[\xi_i^4\mathbb{1}_{\{|\xi_i|>L\}}] + (c_3+c_5)\mathbb{P}(|\xi_1|>L) \\ &\leq \bar{c}_4\sqrt{\mathbb{P}(|\xi_1|>L)} + (c_3+c_5)\mathbb{P}(|\xi_1|>L) \\ &\leq \frac{\bar{c}_4\exp(-L^2/4)}{\sqrt{L}} + \frac{(c_3+c_5)\exp(-L^2/2)}{L} \quad ,\end{aligned}$$

where $\bar{c}_4$ is a positive constant that depends on c_4 and ξ.
Merging the results of the **Step 2.1** and **Step 2.2** in Eq.(23), and we obtain

$$\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda_{\epsilon_1}^{M_1}\times\Lambda_{\epsilon_2}^{M_2}} |R_N(\hat{\sigma}^2_{\lambda,\beta}) - R(\hat{\sigma}^2_{\lambda,\beta})|\right] \leq \mathbf{c}(c_2L^4+\bar{c}_2)\left(\frac{(M_1+M_2)\log(N)}{N}\right)^{1/2} + \frac{\bar{c}_4\exp(-L^2/4)}{\sqrt{L}} + \frac{(c_3+c_5)\exp(-L^2/2)}{L}.$$

Puting $L=\sqrt{2\log(N)}$, and we get

$$\mathbb{E}\left[\sup_{(\lambda,\beta)\in\Lambda_{\epsilon_1}^{M_1}\times\Lambda_{\epsilon_2}^{M_2}} |R_N(\hat{\sigma}^2_{\lambda,\beta}) - R(\hat{\sigma}^2_{\lambda,\beta})|\right] \leq c_6\left(\frac{(M_1+M_2)\log^5(N)}{N}\right)^{1/2} ,$$

where c_6 is constant which depends on c_2. Thus,

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{\mathsf{C}}) - R(\bar{\sigma}^2_{\mathsf{C}})\right] \leq C\left(\frac{(M_1+M_2)\log^5(N)}{N}\right)^{1/2} ,$$

where C is constant which depends on c_6 and $\mathbf{c}$.

Remark 4. *When Y is bounded, it is clear that there exists an absolute constant $C>0$*

$$\mathbb{E}\left[R(\tilde{\sigma}^2_{\mathcal{C}}) - R(\bar{\sigma}^2_{\mathcal{C}})\right] \leq C\left(\frac{(M_1+M_2)\log(N)}{N}\right)^{1/2} .$$

Step 3. Study of the term $\mathbb{E}\left[R(\hat{\sigma}^2_{\mathsf{C}}) - R(\tilde{\sigma}^2_{\mathsf{C}})\right]$. We use the same arguments of proof of Theorem 1 (**Step 2.2**), and we get that there exists two positive constants C_1 and C_2 such that

$$\mathbb{E}\left[R(\hat{\sigma}^2_{\mathsf{C}}) - R(\tilde{\sigma}^2_{\mathsf{C}})\right] \leq C_1\left\{\mathbb{E}\left[\|\hat{f}_{\mathsf{C}} - f^*\|_N^2\right]\right\}^{1/p} + C_2\alpha_N \ , \tag{24}$$

where $p=2$ if Y is bounded, $p=4$ otherwise, and

$$\alpha_N = \begin{cases} \left(\frac{(M_1+M_2)\log(N)}{N}\right)^{1/2} & \text{if } Y \text{ is bounded;} \\ \left(\frac{(M_1+M_2)\log^5(N)}{N}\right)^{1/2} & \text{otherwise.}\end{cases}$$

In the sequel, we give the following proposition

Proposition 2. *Let $\hat{f}_C$ be the aggregate defined in Eq. (4). Then, under Assumptions 2 and 5 there exists an absolute constant $C>0$*

$$\mathbb{E}\left[\|\hat{f}_C - f^*\|_N^2\right] \leq \min_{\lambda\in\Lambda^{M_1}} \mathbb{E}\left[\|\hat{f}_\lambda - f^*\|_N^2\right] + C\sqrt{\frac{\log(M_1)}{N}} .$$

The proof of this proposition is similar of the proof of Proposition 1. Thus, we apply Proposition 2 in inequality (24) and we get

$$\mathbb{E}\left[R(\hat{\sigma}^2_{\mathsf{C}}) - R(\tilde{\sigma}^2_{\mathsf{C}})\right] \leq C_1\left\{\min_{\lambda\in\Lambda^{M_2}} \mathbb{E}\left[\|\hat{f}_\lambda - f^*\|_N^2\right]\right\}^{1/p} + \bar{C}_1\phi_N^{\mathsf{C}}(M_1) \ ,$$

where $\bar{C}_1$ is a constant that depends on C_1 and the constant in Proposition 2, where $p=2$ if Y is bounded, $p=4$ otherwise, and

$$\phi_N^{\mathsf{C}}(M_1) = \begin{cases} \left(\frac{\log(M_1)}{N}\right)^{1/4} & \text{if } Y \text{ is bounded;} \\ \left(\frac{\log(M_1)}{N}\right)^{1/8} & \text{otherwise.}\end{cases}$$

Combining **Step** 1, **Step** 2 and **Step** 3 in Eq (22) yields the result.

D Technical lemmas

In this section, we gather several technical results which are used to derive the proof of results of this paper.

Lemma 3. *Let X be the standard gaussian distribution, then for any $x > 0$, it holds*

$$\mathbb{P}(X > x) \leq \frac{\exp(-x^2/2)}{\sqrt{2\pi}x} \quad , \quad \textit{and} \quad \mathbb{P}(|X| > x) \leq \sqrt{\frac{2}{\pi}}\frac{\exp(-x^2/2)}{x} \quad .$$

Proof. Since $X \sim \mathcal{N}(0,1)$, one gets

$$\mathbb{P}\left(X > x\right) = \frac{1}{\sqrt{2\pi}}\int_x^{+\infty} \exp(-u^2/2)du \leq \frac{1}{\sqrt{2\pi}}\int_x^{+\infty} \frac{u}{x}\exp(-u^2/2)du = \frac{\exp(-x^2/2)}{\sqrt{2\pi}x}.$$

The second inequality follows from symmetry and the last one using the union bound

$$\mathbb{P}(|X| > x) \leq 2\mathbb{P}\left(X > x\right).$$

□

Lemma 4. *Let $X \sim \mathcal{N}(0,1)$ and $k \geq 1$, then*

$$\mathbb{E}\left[|X|^{2k}\right] \leq 2^{k+1}k!.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[|X|^{2k}\right] = \int_0^{+\infty} \mathbb{P}\left(|X|^{2k} > t\right)dt = \int_0^{+\infty} \mathbb{P}\left(|X| > t^{\frac{1}{2k}}\right)dt \quad &\leq \quad 2\int_0^{+\infty} \exp\left(-t^{\frac{1}{k}}/2\right)dt \\
&\overset{u=t^{\frac{1}{k}}/2}{=} \quad 2^{k+1}k\int_0^{+\infty} u^{k-1}\exp(-u)du = 2^{k+1}k!.
\end{aligned}$$

□

Lemma 5. *Let $X_1, \dots, X_M$ be zero mean ν-subgaussian random variables,* i.e., $\mathbb{E}\left[\exp(rX_i)\right] \leq exp\left(\frac{r^2\nu^2}{2}\right)$ *for all $r > 0$. Then*

$$\mathbb{E}\left[\max_{1\leq i\leq M} X_i\right] \leq \nu\sqrt{2\log(M)}.$$

Proof. By Jensen's inequality, for any $r > 0$

$$\begin{aligned}
\mathbb{E}\left[\max_{1\leq i\leq N} X_i\right] = \frac{1}{r}\mathbb{E}\left[\log\left(\exp\left(r\max_{1\leq i\leq M} X_i\right)\right)\right] \quad &\leq \quad \frac{1}{r}\log\left(\mathbb{E}\left[\exp\left(r\max_{1\leq i\leq M} X_i\right)\right]\right) \\
&= \quad \frac{1}{r}\log\left(\mathbb{E}\left[\max_{1\leq i\leq M}\exp\left(rX_i\right)\right]\right) \\
&\leq \quad \frac{1}{r}\log\left(\sum_{i=1}^M \mathbb{E}\left[\exp\left(rX_i\right)\right]\right) \\
&\leq \quad \frac{1}{r}\log\left(\sum_{i=1}^M \mathbb{E}\left[\exp\left(\frac{r^2\nu^2}{2}\right)\right]\right) = \frac{\log(M)}{r} + \frac{\nu^2 r}{2} \quad ,
\end{aligned}$$

taking $r = \sqrt{\frac{2\log(M)}{\nu^2}}$ and we get the result. □

Lemma 6. *Let $N \in \mathbb{N}^*$, $a \geq 1$, b and c be two non negative real numbers. Consider Z a positive random variable such that*

$$\mathbb{P}\left(Z \geq t\right) \leq \min(1, \exp(a - bNt^2) \quad . \tag{25}$$

Then, there exists a constant $C > 0$ not depending of N such that

$$\mathbb{E}[Z] \leq C\left(\frac{a}{bN}\right)^{1/2}.$$

Proof. By condition (25), we have

$$\mathbb{E}[Z] \leq \int_0^{+\infty} \min(1, \exp(a - bNt^2)dt \leq \left(\frac{a}{bN}\right)^{1/2} + \int_{\left(\frac{a}{bN}\right)^{1/2}}^{+\infty} \exp(a - bNt^2)dt. \tag{26}$$

The following elementary inequality $(x-y)^2 \leq x^2 - y^2$ for all $x, y \geq 0$ yields to

$$\int_{\left(\frac{a}{bN}\right)^{1/2}}^{+\infty} \exp(a - bNt^2)dt \leq \int_{\left(\frac{a}{bN}\right)^{1/2}}^{+\infty} \exp\left(-bN\left(t - \left(\frac{a}{bN}\right)^{1/2}\right)^2\right)dt = \int_0^{+\infty} \exp\left(-bNu^2\right)du \leq C\left(\frac{1}{bN}\right)^{1/2}. \tag{27}$$

Combining Equation (27) in Equation (26) to yield the result. □

Lemma 7 (Bernstein's inequality)**.** *Let $T_1, \ldots, T_n$ be independent real valued random variables. Assume that there exists some positive numbers v and c such that*

$$\sum_{i=1}^{n} \mathbb{E}[T_i^2] \leq v \ ,$$

and for all integers $k \geq 3$

$$\sum_{i=1}^{n} \mathbb{E}[(T_i \vee 0)^k] \leq \frac{k!}{2} v c^{k-2} \ .$$

Let $S = \sum_{i=1}^{n}(T_i - \mathbb{E}[T_i])$, then for every any positive x, we have

$$\mathbb{P}\left(|S| \geq x\right) \leq 2\exp\left(-\frac{x^2}{2(v+cx)}\right) \ .$$

Lemma 8 (Hoeffding's inequality)**.** *Let $N \in \mathbb{N}^*$ and $a > 0$ be a real number. Let $X_1, \ldots, X_N$ be independent random variables having values in $[-a, a]$, then for all $t > 0$*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i])\right| > t\right) \leq 2\exp\left(-\frac{Nt^2}{2a^2}\right).$$

Lemma 9 (Hoeffding's Lemma)**.** *Let $X \in [a, b]$ be a bounded random variable with $\mathbb{E}[X] = 0$. Then, for all $\lambda \in \mathbb{R}$*

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$