Nonparametric change point detection in regression

Valeriy Avanesov

submitted: February 6, 2020

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: valeriy.avanesov@wias-berlin.de

Nonparametric change point detection in regression

Valeriy Avanesov

Abstract

This paper considers the prominent problem of change-point detection in regression. The study suggests a novel testing procedure featuring a fully data-driven calibration scheme. The method is essentially a black box, requiring no tuning from the practitioner. The approach is investigated from both theoretical and practical points of view. The theoretical study demonstrates proper control of first-type error rate under $\mathbb{H}_0$ and power approaching 1 under $\mathbb{H}_1$. The experiments conducted on synthetic data fully support the theoretical claims. In conclusion, the method is applied to financial data, where it detects sensible change-points. Techniques for change-point localization are also suggested and investigated.

1   Introduction

The current study works on a problem of change point detection, which applications range from neuroimaging [9] to finance [17, 10, 20, 32]. In many fields practitioners have to deal with the processes subject to an abrupt unpredictable change, hence arises the need to detect and localize such changes. In the writing we refer to the former problem as *break detection* and the latter as *change-point localization*, effectively adopting the terminology suggested in [4]. The importance of the topic promotes an immense variety of considered settings and obtained results on the topic [18, 2, 31, 28, 1, 48, 50, 15, 25, 16].

In the current paper we focus on break detection and change point localization in regression. Typically, in a regression setting a dataset of pairs of (possibly) multivariate covariates and univariate responses is considered, while the goal is to approximate the functional dependence between the two. Here we assume, the data points are separated in time. The problem at hand is whether the functional dependence stayed the same over time and if not, when did the break take place. This setting has been attracting a plethora of attention for decades now. Most researches consider linear [36, 26, 7, 6, 8, 35, 30, 27] or piece-wise constant regression [5, 29, 34]. A recent paper [43] allows for a generalized linear model, leaving the proper choice of a parametric model to the practitioner. In contrast, we develop a fully nonparametric method, eliminating the need to choose a parametric family. Some papers (e.g. [36, 6, 43]), however, rely on a fairly general framework of Likelihood-Ratio test, which we employ in our study as well. Further, some researchers (see [21] for example) propose a test statistic, yet leave the choice of the critical value to the practitioner, while we also suggest a fully data driven way to obtain them. Naturally, the problem has also received a nonparametric treatment [46, 22, 42, 24]. Having suggested a test statistic, [24, 22] propose to rely on an asymptotic distribution in order to acquire its quantiles, while [46, 42] also propose a bootstrap approach and justify its validity. On the other hand, the suggested methodologies do not yield change-point localization and are not applicable in case of multiple breaks.

Contribution of our work consists in a novel break detection approach in regression which is:

- fully nonparametric

- fully data-driven

- working in black-box mode: has virtually no tuning parameters

- capable of multiple break detection

- naturally suitable for change-point localization

- featuring formal results bounding first type error rate (from above) and power (from below)

- performing well on simulated and on real-world data.

Formally, we consider the pairs of deterministic multidimensional covariates $X_i \in \mathcal{X}$ and corresponding univariate responses $y_i \in \mathbb{R}$ for $i \in 1..N$, where $\mathcal{X}$ is a compact in $\mathbb{R}^p$. We wish to test a null hypothesis

$$\mathbb{H}_0 = \{\forall i : y_i = f^*(X_i) + \varepsilon_i\}$$

versus an alternative (only a single break is allowed for simplicity, Section 2.1 suggests a generalization)

$$\mathbb{H}_1 = \left\{ \begin{array}{ll} \exists \tau, f_1^* \neq f_2^* : & y_i = f_1^*(X_i) + \varepsilon_i \text{ if } i < \tau \\ & y_i = f_2^*(X_i) + \varepsilon_i \text{ otherwise} \end{array} \right\},$$

where ε_i denote centered independent identically distributed noise. The functions f^*, f_1^* and f_2^*, mapping from the compact $\mathcal{X}$ to $\mathbb{R}$, are assumed to be unknown along with the distribution of ε_i.

The approach relies on Likelihood Ratio test statistic. Assume for now, the break could happen only at the time t. Then it makes sense to consider n data points to the left and n data points to the right of t and consider the ratio of likelihoods $A_n(t)$ of $2n$ points under a single model and under a pair of models explaining the portions of data to the left and to the right of t separately. Yet the break can happen at any moment, so we consider the test statistics for all possible time moments simultaneously. Finally, in order to resolve the issue of the proper choice of the window size n we suggest to consider multiple window sizes $n \in \mathfrak{N} \subset \mathbb{N}$ at once (e. g. powers of 2).

The paper is organized as follows. Section 2 describes the approach. Further, the approach receives a formal treatment in Section 3. Finally, the behavior of the approach is empirically examined in Section 4.

2 The approach

Let us introduce some notation first. Denote the maximal and the minimal window sizes as $n_+ := \max \mathfrak{N}$ and $n_- := \min \mathfrak{N}$. Define a set of central points for each window size n as $\mathbb{T}_n := \{n, n+1, .., N-n\}$. Further, for each $n \in \mathfrak{N}$ and $t \in \mathbb{T}_n$ define vectors $y_n^l(t)$ composed of the responses $\{y_i\}_{i=t-n+1}^{t}$ belonging to the window to the left of t. Correspondingly, vectors $y_n^r(t)$ are composed of $\{y_i\}_{i=t+1}^{t+n}$. The concatenation of these two vectors is denoted as $y_n(t)$. Also, we use $X_n^l(t)$, $X_n^r(t)$ and $X_n(t)$ to denote the tuples of covariates corresponding to $y_n^l(t)$, $y_n^r(t)$ and $y_n(t)$ respectively. For each window size $n \in \mathfrak{N}$ and central point $t \in \mathbb{T}_n$ we define the test statistic

$$A_n(t) := L\left(y_n^l(t), X_n^l(t)\right) + L\left(y_n^r(t), X_n^r(t)\right) - L(y_n(t), X_n(t)),$$

where L is a likelihood function which is defined below. Intuitively, the statistic should take extremely large values when the two portions of data before and after t are much better explained by a pair of distinct models than by a common one. As we aim to construct a nonparametric approach, we define L relying on a well known technique named Gaussian Process Regression [37]. Formally, we model the noise with a normal distribution and impose the zero-mean Gaussian Process prior with covariance function $k(\cdot, \cdot)$ on the regression function

$$f \sim \mathcal{GP}\left(0, \rho k(\cdot, \cdot)\right),$$
$$y_j \sim \mathcal{N}\left(f(X_j), \sigma^2\right) \text{ for } j \in 1..M,$$

where M is the number of response-covariate pairs under consideration and ρ is a regularization parameter (see (3.1) and (3.2) for its choice). Integrating f out we can easily see, the joint distribution of responses $y \in \mathbb{R}^M$ given the covariates $X = \{X_j\}_{j=1}^M$ is modelled as a multivariate normal distribution with zero mean and covariance matrix $K\left(X\right) \in \mathbb{R}^{M \times M}$, such that $K\left(X\right)_{jj'} := \rho k(X_j, X_{j'}) + \sigma^2 \delta_{jj'}$, where $\delta_{jj'}$ is Kronecker delta. This observation followed by taking the logarithm and abolishing the non-random additive constants leads to the following definition of the likelihood L:

$$L(y, X) := -\frac{1}{2} y^T K(X)^{-1} y.$$

Remark 2.1. *The suggested approach shares its local nature with the ones presented in [4, 3, 43] as they use only a portion of the dataset (of size 2n) to construct a test statistic for time t. Alternatively, one could use the whole dataset as in [47], yet, this is not the best option in presence of multiple breaks. Consider a setting where a function f_1^* changes to f_2^* and back to f_1^* shortly afterwards. The long tails might "water down"the test statistic. To that end a method called Wild Binary Segmentation suggests to choose multiple random continuous sub-datasets of random lengths [21]. Unfortunately, this might lead to excessively long sub-datasets and significantly increase computational complexity. Our approach is free of either of these issues. Also see Remark 3.1 for another motivation for an approach of a local nature.*

The approach being suggested rejects the $\mathbb{H}_0$ if for some window size $n \in \mathfrak{N}$ and some central point $t \in \mathbb{T}_n$ the test statistic $A_n(t)$ exceeds its corresponding critical level $x_{n,\alpha}(t)$ given the significance level α. Formally, the rejection set is

$$\{\exists n \in \mathfrak{N}, t \in \mathbb{T}_n : A_n(t) > x_{n,\alpha}(t)\}. \tag{2.1}$$

As the joint distribution of $A_n(t)$ is unknown, we mimic it with a residual bootstrap scheme in order to allow for the proper choice of the critical levels. First, let us choose some subset of indices $\mathcal{I}^\flat \subseteq 1..N$ we use for bootstrap. We assume the response-covariate pairs $\{(y_i, X_i)\}_{i \in \mathcal{I}^\flat}$ follow the same distribution, hence we require $\mathcal{I}^\flat$ to be located either to the left, or to the right from τ (we presume the former without loss of generality). Given a collection of pairs $\{(y_i, X_i)\}_{i \in \mathcal{I}^\flat}$, we construct estimates $\hat{y}_i$ of $\mathbb{E}\left[y_i\right]$ and the corresponding residuals $\hat{\varepsilon}_i := y_i - \hat{y}_i$. Now define the bootstrap counterpart of the response y_i as

$$y_i^\flat = \hat{y}_i + \varepsilon_i^\flat, \text{ with } \varepsilon_i^\flat := s_i \hat{\varepsilon}_{j_i},$$

where for all $i \in 1..N$ we draw j_i independently and uniformly with replacement from $\mathcal{I}^\flat$ and s_i are independently and uniformly drawn from $\{-1, 1\}$. At this point we can trivially define the bootstrap statistics $A_n^\flat(t)$ in the same way their real-world counterparts $A_n(t)$ are defined by plugging in $y_i^\flat$

instead of y_i. Next, using $\mathbb{P}^\flat$ to denote the bootstrap probability measure, we define the quantile functions for each $\mathrm{x} \in [0, 1]$

$$z_{n,\mathrm{x}}^\flat(t) := \inf\left\{z : \mathbb{P}^\flat\left\{A_n^\flat(t) > z\right\} \leq \mathrm{x}\right\}.$$

Finally, we correct the significance level α for multiplicity

$$\alpha^* := \sup\left\{\mathrm{x} : \mathbb{P}^\flat\left\{\exists n \in \mathfrak{N}, t \in \mathbb{T}_n : A_n^\flat(t) > z_{n,\mathrm{x}}^\flat(t)\right\} \leq \alpha\right\} \qquad (2.2)$$

and define the critical levels $x_{n,\alpha}^\flat(t) := z_{n,\alpha^*}^\flat(t)$.

If the method rejects $\mathbb{H}_0$, one can localize the change-point as follows. First, define the earliest central point, where $\mathbb{H}_0$ is rejected

$$\tilde{\tau}^n := \min\{t \in \mathbb{T}_n : A_n(t) > x_{n,\alpha}^\flat(t)\}.$$

Now, if $A_n(t) > x_{n,\alpha}^\flat(t)$, the change point is located in the interval $[t-n, t+n)$ (up to the significance level α). Therefore, we suggest to define the earliest detecting window

$$n_* := \arg\min_{n \in \mathfrak{N}}(\tilde{\tau}^n + n)$$

and use the following change-point location estimator

$$\hat{\tau} := \arg\max_{t \in [\tilde{\tau}^{n_*} - n_*, \tilde{\tau}^{n_*} + n_*)} A_{n_+}(t). \qquad (2.3)$$

Remark 2.2. *To sum it up, the practitioner needs to provide a set of window sizes $\mathfrak{N}$ and the parameters driving the Gaussian Process prior – covariance functoin $k(\cdot, \cdot)$ and the standard deviation of noise σ. In the absence of additional knowledge about the problem it is viable to use an exponential grid, e.g. $\mathfrak{N} = \{5, 10, 20, \dots\}$ (this strategy is employed in the experimental study, see Section 4). The choice of covariance function $k(\cdot, \cdot)$ and variance σ^2 is of great importance as well. Thankfully, Gaussian Process Regression is known to be a "fairly reliable black-box function approximator" [19]. Typically, a parametric family of covariance functions $\{k_\theta(\cdot, \cdot)\}_{\theta \in \Theta}$ is considered and the optimal combination of hyper-parameters θ and σ^2 is chosen via evidence maximization (see Section 4.5.1 in [37] for details). Eventually, the user only has to provide the confidence level α. Therefore, we conclude our approach is a "black-box".*

Remark 2.3. *The estimates $\hat{y}_i$ may be obtained with any regression method as long as they are consistent under $\mathbb{H}_0$. As we strive for a nonparametric methodology, Gaussian Process Regression trained on $\{(y_i, X_i)\}_{i \in \mathcal{I}^\flat}$ is suggested. The theoretical results can be trivially adapted to any kind of a consistent regressor used instead.*

Remark 2.4. *In practice it may be computationally difficult to obtain enough samples of the bootstrap statistics $A_n^\flat(t)$ for the large number of quantiles to be simultaneously estimated. Alternatively, we suggest to choose the critical levels $x_{n,\alpha}^\flat(t) = x_{n,\alpha}^\flat$ independently of the central point t, effectively replacing the rejection region (2.1) with*

$$\left\{\exists n \in \mathfrak{N} : \max_{t \in \mathbb{T}_n} A_n(t) > x_{n,\alpha}^\flat\right\}$$

as the smaller number of quantiles can be reliably estimated based on much fewer number of the samples drawn. Clearly, this may lead to some drop of sensitivity.

Remark 2.5. *The method can be easily extended for break detection in multivariate regression. In that case one can consider $A_n^l(t)$ for l-th component of outcome, alter the calibration scheme accordingly and make multiplicity correction (2.2) also account for the dimensionality of responses (not only for the windows and break locations).*

2.1 Multiple break detection

In spite of the fact that we allow for at most one break, the local nature of the test statistic $A_n(t)$ allows for a straightforward application of the test in presence of multiple breaks as well. Again, consider a dataset $\{(X_i, y_i)\}_{i=1}^N$ but assume $\mathbb{H}_1$ allows for multiple change-points $\{\tau_k\}_{k=1}^K$ (K is unknown). Formally, extending the notation $\tau_0 \coloneqq 1$ and $\tau_{K+1} = N$,

$$
\mathbb{H}_1 \coloneqq \left\{
\begin{array}{l}
\exists \{f_k^*\}_{k=1}^{K+1} : f_k^* \neq f_{k+1}^* \\
y_i = f_k^*(X_i) + \varepsilon_i \text{ if } \tau_{k-1} \leq i < \tau_k \\
\text{for all } k
\end{array}
\right\}.
$$

Then we estimate the location of the first change-point as

$$
\hat{\tau}_1 \coloneqq \arg \max_{t \in [\tilde{\tau}^{n_*} - n_*, \tilde{\tau}^{n_*} + n_*)} A_{n_+}(t).
$$

Next, the procedure is recursively called on the rest of the dataset $\{(X_i, y_i)\}_{i=\tilde{\tau}^{n_*} + n_*}^N$.

3 Theoretical results

This section is devoted to the theoretical results. Namely, Section 3.2 presents the bootstrap validity result, claiming that the critical levels $x_{n,\alpha}^{\flat}(t)$ yielded by the calibration procedure are indeed chosen in accordance with the critical level α. The sensitivity result is reported in Section 3.3. It defines the minimal window width sufficient for the detection of a break and is also followed by a corollary providing change-point localization guaranties.

3.1 Assumptions and definitions

In order to state the theoretical results we need to formulate some assumptions and definitions. Particularly, we rely on definition of sub-Gaussian variables and vectors.

Definition 3.1 (Sub-Gaussianity). *We say a centered random variable x is sub-Gaussian with $\mathfrak{g}^2$ if*

$$
\mathbb{E}\left[\exp(sx)\right] \leq \exp\left(\mathfrak{g}^2 s^2 / 2\right), \quad \forall s \in \mathbb{R}.
$$

We say a centered random vector X is sub-Gaussian with $\mathfrak{g}^2$ if for all unit vectors u the product $\langle u, X \rangle$ is sub-Gaussian with $\mathfrak{g}^2$.

Further, we consider two broad classes of smooth functions: Sobolev and Hölder.

Definition 3.2 (Sobolev and Hölder classes). *Consider an orthonormal basis $\{\psi_j\}$ in $L_2(\mathbb{R}^p)$ and a function $f = \sum_j f_j \psi_j \in L_2(\mathbb{R}^p)$. We call it $\aleph$-smooth Sobolev if*

$$
\exists B : \sum_{j=1}^{\infty} j^{2\aleph} f_j^2 \leq B^2
$$

and we call it $\aleph$-smooth Hölder if

$$
\exists B : \sum_{j=1}^{\infty} j^{\aleph} |f_j| \leq B^2.
$$

These properties drive the choice of the regularization parameter ρ. Namely, for sample size M large enough we choose

$$\rho = \frac{B^2}{\log M} \tag{3.1}$$

if the function is Sobolev and

$$\rho = \left(\frac{B^2}{\log M}\right)^{2\aleph/(2\aleph+1)} \left(\frac{1}{M}\right)^{1/(2\aleph+1)} \tag{3.2}$$

if the function is $\aleph$-Hölder.

Throughout the paper we use a variety of norms. We use $\|\cdot\|$ to denote Euclidean norm of a vector or a spectral norm of a matrix. Further, $\|\cdot\|_\infty$ refers to sup-norm for both vectors and matrices (the maximal absolute value of an element), as well as functions (the maximal absolute value of an element of its image), while $\|\cdot\|_F$ stands for Frobenius norm of a matrix.

The result Lemma F.1 (by [49]) we rely upon imposes the following two assumptions.

Assumption 3.1. *Let there exist C_ψ and L_ψ s.t. for eigenfunctions $\{\psi_j(\cdot)\}_{j=1}^\infty$ of covariance function $k(\cdot,\cdot)$*

$$\max_j \|\psi_j\|_\infty \le C_\psi$$

and for all $t, s \in \mathbb{R}^p$

$$|\psi_j(t) - \psi_j(s)| \le jL_\psi \|t - s\|.$$

Assumption 3.2. *Let for the eigenvalues $\{\mu_j\}_{j=1}^\infty$ of covariance function $k(\cdot,\cdot)$ exist positive c and C s.t. $cj^{-2\aleph} \le \mu_j \le Cj^{-2\aleph}$ for $\aleph > 1/2$.*

Matérn kernel with smoothness index $\aleph - 1/2$ satisfy these assumptions. In [49] the authors claim, their results also hold for kernels with non-polynomially decaying eigenvalues, like RBF and polynomial kernels. And as long as we do not use these assumptions in our proofs directly, so do ours.

Finally, we introduce the assumptions required by our machinery.

Assumption 3.3. *Let $\tilde{K}_n(t)^{-1}$ have the same elements as $K(X_n(t))^{-1}$ with exception for the diagonal and $\mathrm{diag}\,\tilde{K}_n(t)^{-1} = 0$. Assume, exists a positive γ s.t. for all $t \in \mathbb{T}_n$ for $n \to \infty$*

$$\left\|\tilde{K}_n(t)^{-1}\mathbb{E}\left[y_n(t)\right]\right\|_\infty = O(n^\gamma). \tag{3.3}$$

It would be natural to expect $K(X_n(t))^{-1}$ in (3.3) instead of $\tilde{K}_n(t)^{-1}$, e.g.

$$\left\|K(X_n(t))^{-1}\mathbb{E}\left[y_n(t)\right]\right\|_\infty = O(n^\gamma). \tag{3.4}$$

On the one hand, if the design $\{X_i\}_{i=1}^N$ is regular, (e.g. a uniform grid), (3.4) implies (3.3), yet in general, particularly (3.3) is the desired assumption. We prove the bootstrap validity result (Theorem 3.1) using our Gaussian approximation Lemma B.3. There we have to treat the diagonal and off-diagonal elements of the quadratic forms separately. This is reminiscent of the results in [23] where they study an asymptotic distribution of a single quadratic form (we, in contrast, work with a joint distribution of numerous quadratic forms).

Assumption 3.4. *Let there exist a positive constant C independent of n s.t. $\forall t$*

$$\|K(X_n(t))\| < C.$$

Informally, Assumption 3.3 does not let the GP prior be too unrealistic, while Assumption 3.4 prohibits concentrations of measurements in a local area. Neither would we like Assumption 3.4 violated looking from a practical perspective, as it ensures $K(X_n(t))$ being well-conditioned.

3.2 Bootstrap validity

In this section we demonstrate closeness of measures $\mathbb{P}$ and $\mathbb{P}^\flat$ in some sense which is a theoretical justification of our choice of the calibration scheme.

Theorem 3.1. *Let $\mathbb{H}_0$, Assumption 3.1, Assumption 3.2 hold, ε_i be sub-Gaussian with $\mathfrak{g}^2$. Let f^* be $\aleph$-smooth Sobolev and $\kappa := (\aleph - 1/2) / (2\aleph)$ or $\aleph$-smooth Hölder and $\kappa := \aleph/(2\aleph + 1)$. Let n_-, n_+, $|\mathfrak{N}|$ and N grow. Also assume for some positive γ and δ*

$$\frac{\log^{15} (|\mathfrak{N}|\, N)}{n_-^{1-6\gamma}} = o(1), \tag{3.5}$$

$$\left(\frac{\log |\mathcal{I}^\flat|}{|\mathcal{I}^\flat|}\right)^\kappa n_+^{1/2+\delta+\gamma} = o(1) \tag{3.6}$$

and finally, let Assumption 3.3 hold for γ. Then on a set of arbitrarily high probability

$$\sup_{c_n(t)} \big|\mathbb{P}\left\{\forall n \in \mathfrak{N}, t \in \mathbb{T}_n : A_n(t) < c_n(t)\right\} -$$
$$\mathbb{P}^\flat\left\{\forall n \in \mathfrak{N}, t \in \mathbb{T}_n : A_n^\flat(t) < c_n(t)\right\}\big| = o(1).$$

Proof of the theorem is given in Section A. The strategy of the proof is typical for bootstrap validity results. First, we approximate the joint distribution of the test statistics $\{A_n(t)\}_{n\in\mathfrak{N},t\in\mathbb{T}_n}$ with a distribution of some function of a high-dimensional Gaussian vector. This step is handled with our Gaussian approximation result Lemma B.4. Next, the same is done for their bootstrap counterparts $\{A_n^\flat(t)\}_{n\in\mathfrak{N},t\in\mathbb{T}_n}$ using a different Gaussian vector. Finally, we build the bridge between the two approximating distributions using the fact that the mean and variance of these Gaussian vectors are close to each other (see Lemma C.1). The assumptions (3.5) and (3.6) enforce negligibility of the remainder terms involved in Lemma B.4 and Lemma C.1 respectively. In turn, the Gaussian approximation result (Lemma B.4) is obtained using a novel, significantly tailored version of Lindeberg principle [33, 38, 11, 12]. The proof of Gaussian comparison result (Lemma C.1) is inspired by the technique used in [45]. We use Slepian smart interpolant too, yet applying it in a non-trivial way. We believe, Lemma B.4 can also be proven via Slepian smart interpolant instead of Lindeberg principle, which might yield slightly better convergence rate. We leave this for the future research.

3.3 Sensitivity result

Consider a setting under $\mathbb{H}_1$. For simplicity, assume there is a single change point at τ. In order for the break to be detectable we have to impose some discrepancy condition on f_1^* and f_2^*. Moreover, in order to guarantee detection we have to require the choice of covariates X_i to make this discrepancy observed. Keeping that in mind we define the *observed break extent*

$$\mathfrak{B}_n^2 := \frac{1}{n} \sum_{i=\tau}^{\tau+n-1} (f_1^*(X_i) - f_2^*(X_i))^2. \tag{3.7}$$

Theorem 3.2. *Let the setting described above hold, ε_i be sub-Gaussian with $\mathfrak{g}^2$. Let f^*, f_1^*, f_2^* be $\aleph$-smooth Sobolev and $\kappa := (\aleph - 1/2) / (2\aleph)$ or $\aleph$-smooth Hölder and $\kappa := \aleph/(2\aleph + 1)$. Also let*

$n_* \in \mathfrak{N}$, $n_*, N \to +\infty$ and $\mathfrak{B}_{n_*} \to 0$. *Also impose Assumption 3.1, Assumption 3.2, Assumption 3.3 (for $t < \tau$), Assumption 3.4, (3.5), (3.6) and*

$$\mathfrak{B}_{n_*}^{-1} \left(\frac{\log n_*}{n_*} \right)^{\kappa} = o(1). \tag{3.8}$$

Then

$$\mathbb{P}\left\{ \mathbb{H}_0 \text{ is rejected} \right\} \to 1.$$

We defer proof to Appendix D. It is fairly straightforward. First, we bound the test statistics $A_n(\tau)$ with high probability, next we use Theorem 3.1 to also bound the critical levels $x_{n,\alpha}(\tau)$ and finally, we bound the test statistic $A_n(\tau)$ from below and make sure it exceeds the critical level. The assumption (3.8) essentially requires the observed break extent to exceed the precision of Gaussian Process Regression predictor.

Remark 3.1. *The sensitivity result gives rise to another motivation behind simultaneous consideration of wider and narrower windows (and also it is another argument for local statistics in the first place, also see Remark 2.1). Consider a hostile setting, where the values of functions f_1^* and f_2^* coincide for most of the arguments. For instance, let $\mathfrak{B}' := |f_1^*(X_\tau) - f_2^*(X_\tau)|$ and let $f_1^*(X_i) = f_2^*(X_i)$ for all $i > \tau$. Then by definition $\mathfrak{B}_n = \mathfrak{B}'/n$ and hence the assumption (3.8) implies*

$$\mathfrak{B}'^{-1} n_*^{1-\kappa} \log^{\kappa} n_* = o(1).$$

Clearly, a narrower window detects a smaller break of such a kind.

Remark 3.2. *In the setting allowing for multiple change-points (see Section 2.1) assumption (3.8) dictates the requirement for the minimal distance $\Delta_\tau := \min_{k,k':k \neq k'} |\tau_k - \tau_{k'}|$ between two consecutive change-points as $\Delta_\tau \geq 2n_* + |\mathcal{I}^\flat|$ which is sufficient for detection of all the change-points with probability approaching 1.*

Finally, we formulate a trivial corollary providing change-point localization guaranties.

Corollary 3.1. *Under the assumptions of Theorem 3.2*

$$\mathbb{P}\left\{ |\tilde{\tau} - \tau| \leq n_* \right\} \gtrsim 1 - \alpha.$$

4 Empirical Study

In this section we report the results of our experiments[1]. Section 4.1 presents the findings of the simulation study supporting the bootstrap validity and sensitivity results, as well as empirically justifying the simultaneous use of multiple windows and the change-point location estimator 2.3. In Section 4.2 we successfully apply the method to detect change-points in daily quotes of NASDAQ Composite index.

---

[1]The code is available at `github.com/akopich/gpcd`

Table 1: This table demonstrates average width of the narrowest detecting window n_* (from Theorem 3.2) may be reduced employing multiple window sizes at once without noticeable loss of power.

$\mathfrak{N}$	Power	n_*
$\{40\}$	1.0	40.0
$\{40, 20\}$	1.0	20.5
$\{40, 20, 10\}$	1.0	15.7
$\{40, 20, 10, 5\}$	1.0	15.9

4.1 Experiment on synthetic data

We consider functions $f^*(x) = f_1^*(x) = \sin(x)$ and $f_2^*(x) = \sin(x + \phi)$ for various choices of ϕ. Univariate covariates $\{x_i\}_{i=1}^{800}$ are shuffled 800 equidistant points between 0 and π. Under $\mathbb{H}_0$ the responses are sampled independently as $y_i \sim \mathcal{N}\left(f^*(x_i), 0.1^2\right)$. Under $\mathbb{H}_1$ we choose the change-point location $\tau = 700$ and sample $y_i \sim \mathcal{N}\left(f_1^*(x_i), 0.1^2\right)$ for $i < \tau$ and $y_i \sim \mathcal{N}\left(f_2^*(x_i), 0.1^2\right)$ for $i \geq \tau$. For our experiments we consider $\phi \in \{\pi/2, \pi/5, \pi/10, \pi/20, \pi/40\}$ and report the corresponding observed break extent $\mathfrak{B}_n$ (defined by (3.7)). In all the experiments $\mathcal{I}^\flat = \{1, 2, .., 500\}$, the confidence level α was chosen to be 0.01. We choose RBF kernel family

$$k_\theta(x_1, x_2) = \theta_1^2 \exp\left(-\frac{|x_1 - x_2|^2}{\theta_2^2}\right) \tag{4.1}$$

and choose optimal parameters θ and σ^2 via evidence maximization using $\{x_i\}_{i \in \mathcal{I}^\flat}$.

The suggested approach has demonstrated proper control of the first type error rate in all the configurations we consider, keeping it below 0.015.

The power the test exhibits is shown on Figure 1. As expected, larger window size n and larger observed break extent $\mathfrak{B}_n$ correspond to higher power. At the same time, the Figure 2 summarizes root mean squared errors of the estimator $\hat{\tau}$ (defined by (2.3)). The estimator proves itself to be reliable when the power of the test is high. Generally, wider windows and larger observable break extent lead to higher accuracy of $\hat{\tau}$.

Further, in order to investigate the behavior of the method in a multiscale regime ($|\mathfrak{N}| > 1$) we use several choices of $\mathfrak{N}$ for a single $\phi = \pi/10$. Results, reported in the Table 1, exhibit a significant decrease in the average width of the narrowest detecting window n_* and hence an improvement in change-point localization thanks to simultaneous use of wider and narrower windows. This should be highly beneficial in presence of multiple change points, as it allows for smaller distance Δ_τ between them (see Section 2.1 and Remark 3.2).

4.2 Real-world dataset experiment

The prices of stock indexes are known to be subject to abrupt breaks [40, 41]. We consider a series X_t of closing daily prices of NASDAQ Composite index. The dataset spans from February 1990 until February 2019. We suggest to model the process using the following Stochastic Differential equation

$$\frac{dX_t}{X_t} = f(X_t)dt + \sigma dW_t, \sigma > 0,$$

where W_t denotes a Wiener process. Now we wish to test the dataset for the presence of breaks. In order to do so we employ the Euler–Maruyama method, effectively boiling the problem down to
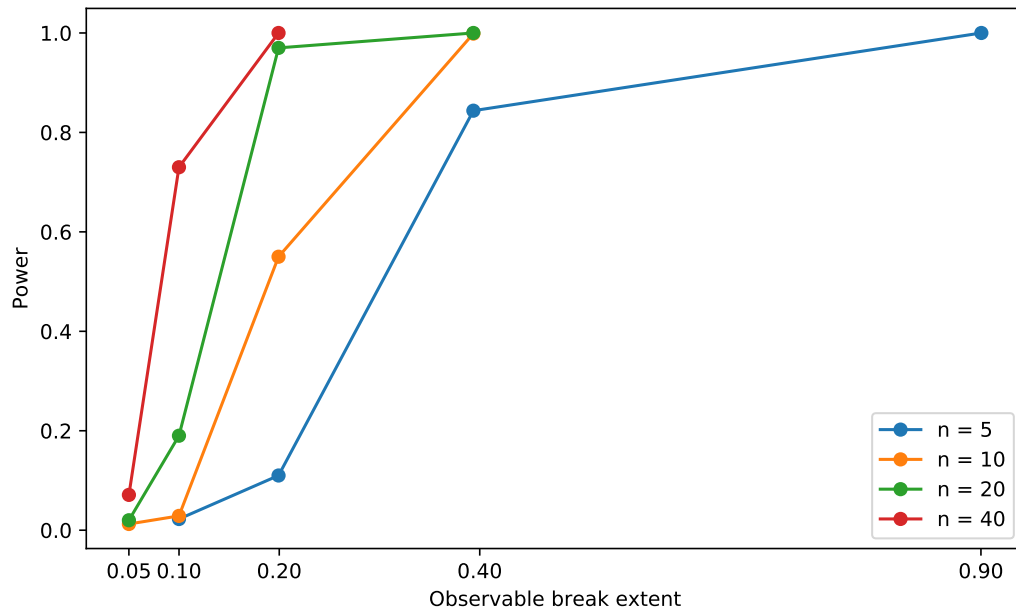
Figure 1: The plot demonstrates the dependence of the power of the test on the observable break extent $\mathfrak{B}_n$ (see (3.7)) for multiple window sizes n.
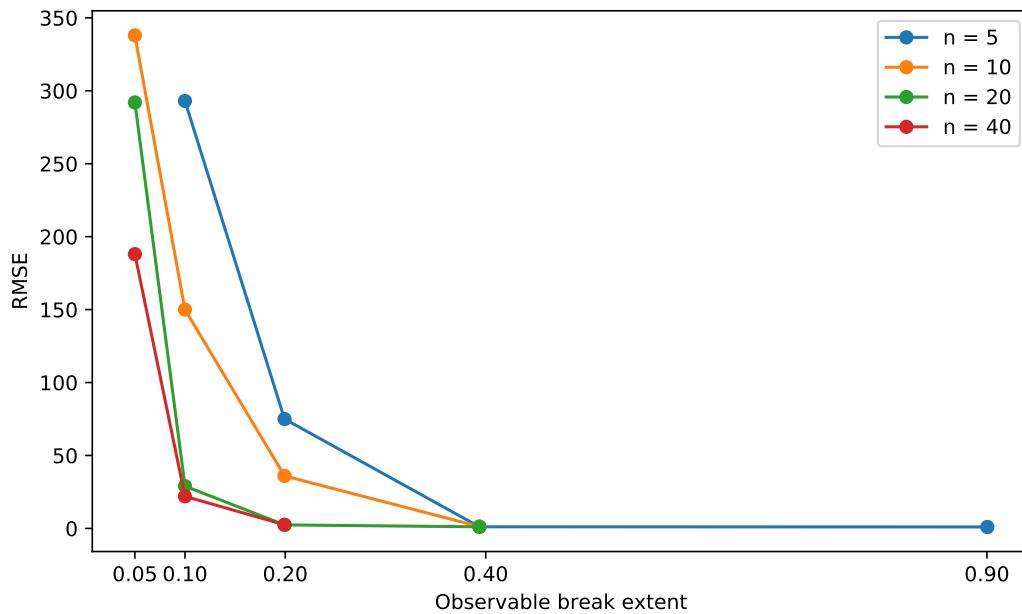

Figure 2: The plot demonstrates the dependence of the root mean squared error of change point localization on the observable break extent $\mathfrak{B}_n$ (see (3.7)) for multiple window sizes n.

a regression problem with univariate covariates $x_t := X_t$ and the corresponding responses $y_t := \frac{X_{t+1} - X_t}{X_t}$. Further we apply the scheme suggested in Section 2.1 with $\alpha = 0.01$, $\mathfrak{N} = \{20\}$, $\mathcal{I}^\flat = \{1..300\}$ and the kernel family (4.1). The method detects three breaks and all of them may be related to the known events. Namely, computer virus CIH has activated itself and attacked Windows 9x in August 1998, burst of the dot-com bubble and 2008 financial crisis.

A Proof of the bootstrap validity result

Proof of Theorem 3.1. Apply Lemma B.4 to $A_n(t)$ and $A_n^\flat(t)$, next apply Lemma C.1 and via triangle inequality obtain on a set of probability at least $1 - 2\exp(-\mathrm{u}^2)$

$$\sup_{c_t} \left| \mathbb{P}\left\{A_n(t) < c_t\right\} - \mathbb{P}^\flat\left\{A_n^\flat(t) < c_t\right\} \right| \le 2R_A + R_C,$$

where R_A and R_C come from Lemma B.4 and Lemma C.1 respectively. Now observe

$$|J| = \sum_{n \in \mathfrak{N}} |\mathbb{T}_n| \le |\mathfrak{N}| N$$

and using (3.5) conclude $R_A = o(1)$. Clearly, the ratio entering the definition of R_C is bounded $\sqrt{n}/\mathfrak{s} = O(1)$ (in the same way as in the proof of Lemma B.4). Next we use Lemma F.1 and obtain on a set of probability at least $1 - \left|\mathcal{I}^\flat\right|^{-10}$

$$\left\| \mathbb{E}\left[y\right] - \hat{y} \right\|_\infty \le O(\Delta_f),$$

where

$$\Delta_f := \left(\frac{\log\left|\mathcal{I}^\flat\right|}{\left|\mathcal{I}^\flat\right|} \right)^\kappa.$$

Now observe that the following holds for Δ_μ and Δ_Σ involved in Lemma C.1 by construction of Z and $\tilde{Z}$ (coming from the gaussian approximation and defined by (B.3))

$$\Delta_\mu \le \left(\Delta_f + \|f^*\|_\infty \right) \Delta_f,$$

$$\Delta_\Sigma \le 4\left| \mathrm{Var}\left[\varepsilon_1\right] - \mathrm{Var}\left[\varepsilon_1^\flat\right] \right| \|f^*\|_\infty^2 + \left| \mathrm{Var}\left[\varepsilon_1\right] - \mathrm{Var}\left[\varepsilon_1^\flat\right] \right|^2.$$

Further, Lemma C.3 yields the bound $\left| \mathrm{Var}\left[\varepsilon_1\right] - \mathrm{Var}\left[\varepsilon_1^\flat\right] \right| = O(\Delta_f^2)$. Assumption (3.5) implies $\gamma < 1/6$. Then (3.6) it turn implies

$$\left(\frac{\log\left|\mathcal{I}^\flat\right|}{\left|\mathcal{I}^\flat\right|} \right)^\kappa n_+^{\delta/2+\gamma} = o(1).$$

Finally, choose $\Delta = n^{-\delta/2}$ (involved in the definition of R_C, see Lemma C.1), recall assumption (3.6) and conclude $R_C = o(1)$. $\qquad\square$

B Gaussian Approximation

Consider a random vector $x \in \mathbb{R}^N$ of independent components centered at $\mu = \mathbb{E}\left[x\right]$. Introduce $x_n(j)$ for even $n = 2m \in \mathbb{N}$ and $j \in J := \{m, m+1, .., N-m\}$ denoting a vector composed

of $\{x_i\}_{i=j-m+1}^{j+m}$. Also, assume $|J|$ symmetric matrices $B(j) \in \mathbb{R}^{n \times n}$ are given and define a map $S : \mathbb{R}^N \to \mathbb{R}^{|J|}$ s. t. $S(x)_j := \frac{1}{\sqrt{n}} \langle x_n(j), B(j) x_n(j) \rangle$.

Two ingredients of paramount importance are soft-max function $F_\beta : \mathbb{R}^{|J|} \to \mathbb{R}$

$$F_\beta(w) := \log \left(\sum_{j \in J} \exp\left(\beta w_j \right) \right) \text{ for } \beta \in \mathbb{R}$$

and a smooth indicator function g_Δ with three bounded derivatives s.t. $|x| \geq \Delta \Rightarrow g_\Delta(x) = 1[x > 0]$. Also let $g := g_1$ and $g(x/\Delta) = g_\Delta(x)$. An example of such function along with bounds for its derivatives is provided in [43].

Consider the following decomposition of matrices $B(j)$ into diagonal matrices and matrices with zeroes down their diagonals

$$B(j) = E(j) + diag(D(j)), \text{ where } E(j)_{kk} = 0 \ \forall k. \tag{B.1}$$

Further, consider a vector $X \in \mathbb{R}^N$ s.t. $x_i^2 = X_i$ for all $i = 1..N$. And introduce notation $X_n(j)$ similar to $x_n(j)$. Now consider a vector Z denoting vectors x and X stacked. Clearly, there is a map $Q : \mathbb{R}^{2N} \to \mathbb{R}^{|J|}$ s.t.:

$$S(x)_j = Q(Z)_j := Q(x, X)_j := \frac{1}{\sqrt{n}} \langle x_n(j), E(j) x_n(j) \rangle + \frac{1}{\sqrt{n}} \langle D(j), X_n(j) \rangle \tag{B.2}$$

for all $j = 1..|J|$. Also define an independent vector

$$\tilde{Z} \sim \mathcal{N}\left(\mathbb{E}\left[Z \right], \text{Var}\left[Z \right] \right) \tag{B.3}$$

and denote the first half of the vector as $\tilde{x}$ and the second as $\tilde{X}$.

Our proof employs a novel version of the Lindeberg principle [33, 38, 11, 12] tuned for the problem at hand. Typically, Lindeberg principle suggests to "replace" random variables with their Gaussian counterparts one by one. Here we have to "replace" each n-th component of x along with the component of X being its square starting with the 1-st one, repeat starting with the 2-nd one and so on repeating the procedure n times. Namely, in the first step we "replace" components with indexes $1, n+1, 2n+1$ and so on. On the second step we "replace" components with indexes $2, n+2, 2n+2$ and so on. And further in the same manner. Or more formally, consider a sequence of vectors $x^i \in \mathbb{R}^N$ for $i = 0..n$ s. t. $x^0 = x$ and $\forall i > 0 : x_{kn+i}^i = \tilde{x}_{kn+i}$ for all $k \in \{0, 1, 2, .., \lceil N/n \rceil - 1\}$ and $x_j^i = x_j^{i-1}$ for j s.t. $\nexists k \in \{0, 1, 2, .., \lceil N/n \rceil - 1\} : kn + i = j$. Denote the indexes of components which were replaced at step i as I_i. Also define a vector $\mathring{x}^i$ s.t. $\mathring{x}_j^i = 0$ for $j \in I_i$ and $\mathring{x}_j^i = x_j^i$ for the rest of j. Define sequence of X^i and $\mathring{X}^i$ in a similar way. Finally, let Z^i denote the vectors x^i and X^i stacked together and $\mathring{Z}_i$ denote stacked vectors $\mathring{x}^i$ and $\mathring{X}^i$. Note, $Z^n = \tilde{Z}$.

Lemma B.1. *Choose $i = 1..n$. Consider a function $\phi : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ defined as*

$$\phi(a, b) := F_\beta \left(Q \left(\mathring{x}^i + a, \mathring{X}^i + b \right) \right),$$

where $j \notin I_i \Rightarrow a_j = 0, \ b_j = 0$ and $Q(\cdot)$ is defined by (B.2). *Further, using decomposition* (B.1) *assume for some positive L:*

$$\max_j \left\| E(j)(\mathring{x}_n^i(j) + a_n(j)) \right\|_\infty < L \tag{B.4}$$

and denote

$$\mathfrak{L} := \max\left\{2L, \max_{j \in J} \|D(j)\|_\infty\right\}. \tag{B.5}$$

Then

$$\left|\sum_{j \in I_i} \left(\partial_{a_j} + \partial_{b_j}\right) \phi(a, b)\right| \le \varphi' := \frac{2\mathfrak{L}}{\sqrt{n}},$$

$$\left|\sum_{j,j' \in I_i} \left(\partial^2_{a_j a_{j'}} + \partial^2_{b_j b_{j'}}\right) \phi(a, b)\right| \le \varphi'' := \frac{4\beta \mathfrak{L}^2}{n}$$

$$\left|\sum_{j,j',j'' \in I_i} \left(\partial^3_{a_j a_{j'} a_{j''}} + \partial^3_{b_j b_{j'} b_{j''}}\right) \phi(a, b)\right| \le \varphi''' := \frac{12\beta^2 \mathfrak{L}^3}{n^{3/2}}.$$

Proof. Proof of this result consists in direct differentiation followed by application of Lemma A.2 from [13] providing bounds for the first three derivatives of soft-max function. □

Lemma B.2. *Let assumptions of Lemma B.1 hold. Then for an independent Gaussian vector $\tilde{Z}$ (defined by* (B.3)*)*

$$\left|\mathbb{E}\left[g_\Delta \circ F_\beta \circ S(x)\right] - \mathbb{E}\left[g_\Delta \circ F_\beta \circ Q(\tilde{Z})\right]\right| \le \frac{n}{6}\mathcal{Z}m_3,$$

where m_3 is the sum of the maximal third centered absolute moments of x and $\tilde{Z}$, while $\mathcal{Z}$ is defined in Lemma B.5 and $Q(\cdot)$ is defined by (B.2)*.*

Proof. Clearly, for $f := g_\Delta \circ F_\beta \circ Q$,

$$f(Z) - f(\tilde{Z}) = \sum_{i=1}^n f\left(Z^{i-1}\right) - f\left(Z^i\right)$$

and hence

$$\left|\mathbb{E}\left[f(Z) - \mathbb{E}\left[f(\tilde{Z})\right]\right]\right| \le \sum_{i=1}^n \left|\mathbb{E}\left[f\left(Z^{i-1}\right)\right] - \mathbb{E}\left[f\left(Z^i\right)\right]\right|. \tag{B.6}$$

The rest of the proof consists in bounding an arbitrary summand on the right hand side. In order to do so we use Taylor expansion of second degree for $f\left(Z^{i-1}\right)$ and $f\left(Z^i\right)$ around $\mathbb{E}\left[Z\right]$ with Lagrange remainder. Given equality of the first two moments of Z and $\tilde{Z}$, we conclude, the first two terms cancel out. Hence, using Lemma B.5 we immediately obtain

$$\left|\mathbb{E}\left[f\left(Z^{i-1}\right)\right] - \mathbb{E}\left[f\left(Z^i\right)\right]\right| \le \frac{1}{6}\mathcal{Z}m_3. \tag{B.7}$$

Combination of (B.6) and (B.7) establishes the claim. □

Lemma B.3. *Let assumptions of Lemma B.1 hold. Then*

$$\sup_c \left|\mathbb{P}\left\{S(x) \le c\right\} - \mathbb{P}\left\{Q(\tilde{Z}) \le c\right\}\right| \le R_A$$

$$:= 84 \log^{3/2} |J| \left(1 + \frac{\sqrt{n}}{\mathfrak{s}}\right) \frac{\mathfrak{L}^{3/4}}{n^{1/8}},$$

where $\mathfrak{s}$ comes from Lemma E.1 and $Q(\cdot)$ is defined by (B.2)*.*

Proof. Choose $\Delta = \log|J|/\beta$. Then for an arbitrary constant vector $c \in \mathbb{R}^{|J|}$

$$
\begin{aligned}
\mathbb{P}\{S(x) \le c\} &\le \mathbb{E}\left[g_\Delta(F_\beta(S(x) - c + \Delta))\right] \\
&\le \mathbb{E}\left[g_\Delta(F_\beta(Q(\tilde{Z}) - c + \Delta))\right] + \frac{n}{6}\mathcal{Z}m_3 \\
&\le \mathbb{P}\left\{Q(\tilde{Z}) \le c - 2\Delta\right\} + \frac{n}{6}\mathcal{Z}m_3 \\
&\le \mathbb{P}\left\{Q(\tilde{Z}) \le c\right\} + 12\Delta\frac{\sqrt{n}}{\mathfrak{s}}\left(\sqrt{4\log|J|} + 2\right) + \frac{n}{6}\mathcal{Z}m_3 \\
&\le \mathbb{P}\left\{Q(\tilde{Z}) \le c\right\} + 24\Delta\frac{\sqrt{n}}{\mathfrak{s}}\sqrt{4\log|J|} + \frac{768\mathfrak{L}^3\log^2|J|}{\Delta^3\sqrt{n}}.
\end{aligned}
$$

Here he have consequently used Lemma B.6, Lemma B.2, Lemma B.6 again and Lemma E.1 (which also defines $\mathfrak{s}$). The last step uses that $\log 3 > 1$. Now we choose

$$
\Delta = 8\left(\frac{144\log^{3/2}|J|\,\mathfrak{L}}{\sqrt{n}}\right)^{1/4}
$$

and obtain

$$
\begin{aligned}
\mathbb{P}\{S(x) \le c\} &\le \mathbb{P}\left\{Q(\tilde{Z}) \le c\right\} + \left(84\log^{3/2}|J|\frac{\sqrt{n}}{\mathfrak{s}} + 13\log^{7/8}|J|\right)\frac{\mathfrak{L}^{3/4}}{n^{1/8}} \\
&\le \mathbb{P}\left\{Q(\tilde{Z}) \le c\right\} + 84\log^{3/2}|J|\left(1 + \frac{\sqrt{n}}{\mathfrak{s}}\right)\frac{\mathfrak{L}^{3/4}}{n^{1/8}}.
\end{aligned}
$$

Similar reasoning yields a chain of "larger-or-equal"inequalities which, combined with the one above, finalizes the proof. □

Lemma B.4. *Let $x - \mathbb{E}[x]$ be sub-Gaussian and matrices $B(j)$ have bounded spectrum. Also assume for some positive γ*

$$
\max_{j \in J}\|B(j)\mu_n(j)\|_\infty = O(n^\gamma).
$$

Then for any positive u on a set of probability at least $1 - \exp(-\mathrm{u}^2)$ for N and n going to infinity

$$
\sup_c\left|\mathbb{P}\{S(x) \le c\} - \mathbb{P}\left\{Q(\tilde{Z}) \le c\right\}\right| = O\left(\log^{3/2}|J|\frac{\mathrm{u}^{3/4} + \log^{3/8}|J| + n^{3\gamma/4}}{n^{1/8}}\right),
$$

where $Q(\cdot)$ is defined by (B.2).

Proof. Application of Lemma B.7 to matrices $E(j)$ yields the bound on $\mathfrak{L}$ defined by (B.5)

$$
\mathfrak{L} = O(\mathrm{t} + n^\gamma)
$$

on a set of probability at least $1 - |J|\,e^{-\mathrm{t}^2}$.

Investigation of $\mathfrak{s}$ defined in Lemma E.1 yields $\sqrt{n}/\mathfrak{s} = O(1)$. Really, $\|B(j)\|_F$ is a sum of squared eigenvalues (which are bounded) and $\|\mathbb{E}[x]\|^2 \le n\|\mathbb{E}[x]\|_\infty^2$. Now we apply Lemma B.3

$$
\sup_c\left|\mathbb{P}\{S(x) \le c\} - \mathbb{P}\left\{Q(\tilde{Z}) \le c\right\}\right| \le R_A = O\left(\log^{3/2}|J|\frac{\mathrm{t}^{3/4} + n^{3\gamma/4}}{n^{1/8}}\right).
$$

Now change the variable $\mathrm{u}^2 := \mathrm{t}^2 - \log|J|$

$$R_A = O\left(\log^{3/2}|J|\,\frac{\mathrm{u}^{3/4} + \log^{3/8}|J| + n^{3\gamma/4}}{n^{1/8}}\right).$$

$\square$

Lemma B.5. *In terms of Lemma B.1 for function*

$$\zeta := g_\Delta \circ \phi$$

it holds that

$$\sum_{j\in I_i} \left(\partial_{a_j} + \partial_{b_j}\right)\zeta(a,b) \le 2\Delta^{-1}\varphi',$$

$$\left|\sum_{j,j'\in I_i} \left(\partial^2_{a_j a_{j'}} + \partial^2_{b_j b_{j'}}\right)\zeta(a,b)\right| \le 8\Delta^{-2}\varphi'^2 + 2\Delta^{-1}\varphi'',$$

and

$$\left|\sum_{j,j',j''\in I_i} \left(\partial^3_{a_j a_{j'} a_{j''}} + \partial^3_{b_j b_{j'} b_{j''}}\right)\zeta(a,b)\right| \le \mathcal{Z}$$

$$:= 32\Delta^{-3}\varphi'^3 + 24\Delta^{-2}\varphi''\varphi' + 2\Delta^{-1}\varphi'''.$$

Proof. The proof consists in direct differentiation and bounding using Lemma B.1 and equation (53) from [43]. Intermediate differentiation steps can be found in the proof of Lemma A.14 [43]. $\square$

The following lemma justifies the smoothing relying on smooth indicator g_Δ and soft-max F_β. Its proof can be found in [13].

Lemma B.6. *Let $\Delta = \log|J|\,/\beta$, then for arbitrary vector x :*

$$g_\Delta(F_\beta(x - \Delta)) \le 1\left\{\|x\|_\infty \ge 0\right\} \le g_\Delta(F_\beta(x + \Delta)).$$

The next lemma establishes prerequisites for inequality (B.8).

Lemma B.7. *Consider a symmetric matrix A with the largest eigenvalue Λ. Let ε be a vector of independent sub-Gaussian with $\mathfrak{g}^2$ elements. Then on a set of probability at least $1 - \exp(-\mathrm{t}^2)$*

$$\|A\varepsilon\|_\infty \le \Lambda\mathrm{t}. \tag{B.8}$$

Proof. For a given unit vector a, as far as the components of ε are independent and sub-Gaussian, $a^T\varepsilon$ is sub-Gaussian with $\mathfrak{g}^2$ as well. Hence,

$$\mathbb{P}\left\{\left|a^T\varepsilon\right| > t\right\} \le p := 2\exp(-\mathrm{t}^2/\mathfrak{g}^2)$$

and therefore,

$$\|A\varepsilon\|_\infty \le \|A\varepsilon\| \le \Lambda\mathrm{t}.$$

$\square$

C   Gaussian comparison

Notation of this section follows the notation of Section B. Proof of the following result was inspired by the proof of Theorem 1 in [45].

Lemma C.1. *Consider two $2N$-dimensional normal vectors $Z \sim \mathcal{N}\left(\mu, \Sigma\right)$ and $\tilde{Z} \sim \mathcal{N}\left(\tilde{\mu}, \tilde{\Sigma}\right)$. Denote $\Delta_\mu := \|\mu - \tilde{\mu}\|_\infty$ and $\Delta_\Sigma := \left\|\Sigma - \tilde{\Sigma}\right\|_\infty$. Use notation of Lemma B.1. Then for any constant vector c and positive Δ holds*

$$\left|\mathbb{P}\left\{Q(Z) < c\right\} - \mathbb{P}\left\{Q(\tilde{Z}) < c\right\}\right| \leq \Delta_\mu \left(\frac{\mathfrak{L}\sqrt{n}}{\Delta}\right) + 16\Delta_\Sigma \left(\frac{\mathfrak{L} + \log|J|\,\mathfrak{L}^2}{\Delta^2}\right)$$
$$+ \frac{8\sqrt{n}}{\mathfrak{s}}\Delta\sqrt{4\log|J|},$$

where $\mathfrak{s}$ comes from Lemma E.1.

Proof. The proof consists in a multiple use of Slepian smart interpolant. Denote the first and the second halves of vector Z as x and X and similarly introduce $\tilde{x}$ and $\tilde{X}$ being halves of $\tilde{Z}$. Further, consider n real values $\varphi_1, \varphi_2, .., \varphi_n$ and compose a vector of length N iterating over these values:

$$\psi(\varphi) := \psi\left(\{\varphi_i\}_{i=1}^n\right) := \left(\varphi_1, \varphi_2, .., \varphi_n, \varphi_1, \varphi_2 ...\right).$$

Denote $f := g_\Delta \circ F_\beta \circ Q$ and consider a function

$$\Psi(\psi) := \mathbb{E}\left[f(\underbrace{x \otimes \sqrt{\psi} + \tilde{x} \otimes \sqrt{1-\psi}}_{x(\psi)}, \underbrace{X \otimes \sqrt{\psi} + \tilde{X} \otimes \sqrt{1-\psi}}_{X(\psi)})\right],$$

where we use $\otimes$ to denote element-wise product and radicals are applied to vectors in an element-wise manner. Clearly,

$$\Psi(1) = \mathbb{E}\left[f(x, X)\right] \text{ and } \Psi(0) = \mathbb{E}\left[f(\tilde{x}, \tilde{X})\right]$$

and hence

$$\left|\mathbb{E}\left[f(x, X)\right] - \mathbb{E}\left[f(\tilde{x}, \tilde{X})\right]\right| = \left|\int_0^1 ... \int_0^1 \left(\prod_{i=1}^n \partial_{\varphi_i}\right) \Psi(\psi(\varphi)) \prod_{i=1}^n d\varphi_i\right|. \tag{C.1}$$

For the derivative we have

$$\left(\prod_{i=1}^n \partial_{\varphi_i}\right) \Psi(\psi(\varphi)) = \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^n \sum_{j=I_i} \partial_{x_j} f(x(\psi), X(\psi))\left(x_j \psi_i^{-1/2} - \tilde{x}_j(1-\psi_i)^{-1/2}\right)\right.$$
$$\left. + \partial_{X_j} f(x(\psi), X(\psi))\left(X_j \psi_i^{-1/2} - \tilde{X}_j(1-\psi_i)^{-1/2}\right)\right].$$

Next we apply Lemma C.2 (which applies only to centered vectors, thus the second term)

$$\left| \left(\prod_{i=1}^{n} \partial_{\varphi_i} \right) \Psi(\psi(\varphi)) \right| \le \frac{1}{2} \Delta_{\Sigma} \left| \mathbb{E} \left[\sum_{i=1}^{n} \sum_{j=I_i} \sum_{k=I_i} \left(\partial_j^2 + \partial_k^2 \right) f(x(\psi), X(\psi)) \right] \right|$$

$$+ \frac{1}{2} \Delta_{\mu} \left| \sum_{i=1}^{n} \sum_{j=I_i} \partial_{x_j} f(x(\psi), X(\psi)) \right|$$

$$=: \frac{1}{2} \left(T_1 + T_2 \right).$$

Now we make use of Lemma B.5 and Lemma B.1 and choose $\beta = \log |J| / \Delta$

$$T_1 \le 32 \Delta_{\Sigma} \left(\frac{\mathfrak{L} + \log |J| \, \mathfrak{L}^2}{\Delta^2} \right),$$

$$T_2 \le \Delta_{\mu} \left(\frac{2 \mathfrak{L} \sqrt{n}}{\Delta} \right).$$

Next, recalling (C.1) obtain

$$\left| \mathbb{E} \left[f(x, X) \right] - \mathbb{E} \left[f(\tilde{x}, \tilde{X}) \right] \right| \le \Delta_{\mu} \left(\frac{\mathfrak{L} \sqrt{n}}{\Delta} \right) + 16 \Delta_{\Sigma} \left(\frac{\mathfrak{L} + \log |J| \, \mathfrak{L}^2}{\Delta^2} \right).$$

Finally, in order to move from smooth functions to indicators we employ reasoning identical to the one in Lemma B.3.

$$\mathbb{P} \{ Q(Z) < c \} \le \mathbb{E} \left[g_{\Delta} \circ F_{\beta} \left(Q(x, X) - c - \Delta \right) \right]$$

$$\le \mathbb{E} \left[g_{\Delta} \circ F_{\beta} \left(Q(\tilde{x}, \tilde{X}) - c - \Delta \right) \right] + \frac{1}{2} \left(T_1 + T_2 \right)$$

$$\le \mathbb{P} \left\{ Q(\tilde{Z}) < c - 2\Delta \right\} + \frac{1}{2} \left(T_1 + T_2 \right)$$

$$\le \mathbb{P} \left\{ Q(\tilde{Z}) < c \right\} + \frac{1}{2} \left(T_1 + T_2 \right) + \frac{8 \sqrt{n}}{\mathfrak{s}} \Delta \sqrt{4 \log |J|}.$$

Combination with a similar chain of larger-or-equal finalizes the proof. □

We use the same version of Stein's identity as the authors of [45] have.

Lemma C.2 (Stein's identity). *Let $X \in \mathbb{R}^p$ be a normal centered vector and function $f : \mathbb{R}^p \to \mathbb{R}$ be a C^1 function with finite first derivatives. Then for all $j = 1..p$*

$$\mathbb{E} \left[X_j f(X) \right] = \sum_{k=1}^{p} \mathbb{E} \left[X_j X_k \right] \mathbb{E} \left[\partial_k f(X) \right].$$

Proof. See Section A.6 of [44]. □

Lemma C.3. *Consider $y, \hat{y}, \varepsilon, \varepsilon^{\flat}$ defined in Section 2. Let*

$$\Delta_f := \| \mathbb{E} \left[y \right] - \hat{y} \|_{\infty} = O \left((\log |\mathcal{I}^{\flat}| / |\mathcal{I}^{\flat}|)^{\kappa} \right)$$

for some positive $\kappa \le 1/2$. Let ε_i be sub-Gaussian with $\mathfrak{g}^2$. Then on a set of probability at least $1 - |\mathcal{I}^{\flat}|^{10}$

$$\left| \mathrm{Var} \left[\varepsilon_1 \right] - \mathrm{Var}^{\flat} \left[\varepsilon_1^{\flat} \right] \right| = O \left(\Delta_f^2 \right).$$

Proof. By construction

$$
\mathrm{Var}^\flat\left[\varepsilon_1^\flat\right] = \frac{1}{|\mathcal{I}^\flat|}\sum_{i\in\mathcal{I}^\flat}\left(\hat{\varepsilon}_i - \underbrace{\frac{1}{|\mathcal{I}^\flat|}\sum_{i\in\mathcal{I}^\flat}\hat{\varepsilon}_i}_{\bar{\varepsilon}}\right)^2
$$

$$
= \frac{1}{|\mathcal{I}^\flat|}\sum_{i\in\mathcal{I}^\flat}\left(\mathbb{E}\left[y_i\right] - \hat{y}_i + \varepsilon_i - \bar{\varepsilon}\right)^2
$$

$$
= \frac{1}{|\mathcal{I}^\flat|}\sum_{i\in\mathcal{I}^\flat}\left(\mathbb{E}\left[y_i\right] - \hat{y}_i\right)^2 + \varepsilon_i^2 + \bar{\varepsilon}^2
$$

$$
+ 2\varepsilon_i\left(\mathbb{E}\left[y_i\right] - \hat{y}_i\right) - 2\bar{\varepsilon}\left(\mathbb{E}\left[y_i\right] - \hat{y}_i\right) - 2\bar{\varepsilon}\varepsilon_i.
$$

Now due to sub-Gaussianity for a positive e

$$
\forall i : \mathbb{P}\left\{|\varepsilon_i| > \mathrm{e}\right\} \leq 2\exp(-\mathrm{e}^2/\mathfrak{g}^2)
$$

and hence

$$
\mathbb{P}\left\{\underbrace{\exists i : |\varepsilon_i| > \mathrm{e}}_{\mathcal{E}}\right\} \leq p' := 2\left|\mathcal{I}^\flat\right|\exp(-\mathrm{e}^2/\mathfrak{g}^2).
$$

On set $\mathcal{E}$ Hoeffding inequality applies to ε_i and their squares:

$$
\mathbb{P}\left\{\left|\frac{1}{|\mathcal{I}^\flat|}\sum_{i\in\mathcal{I}^\flat}\varepsilon_i\right| > E\right\} \leq p'' := 2\exp\left(-2\left|\mathcal{I}^\flat\right|E^2/\mathrm{e}\right),
$$

$$
\mathbb{P}\left\{\left|\frac{1}{|\mathcal{I}^\flat|}\sum_{i\in\mathcal{I}^\flat}\varepsilon_i^2 - \mathrm{Var}\left[\varepsilon_i\right]\right| > E'\right\} \leq p''' := 2\exp\left(-2\left|\mathcal{I}^\flat\right|E'^2/\mathrm{e}^2\right).
$$

Therefore, with probability at least $1 - p' - p'' - p'''$

$$
|\bar{\varepsilon}| \leq \Delta_f + E
$$

and hence

$$
\left|\mathrm{Var}\left[\varepsilon_i\right] - \mathrm{Var}^\flat\left[\varepsilon_i^\flat\right]\right| \leq \Delta_f^2 + E' + 2E\Delta_f + 2\left(\Delta_f + E\right)\left(\Delta_f + 2E + E^2\right).
$$

Clearly, the choice

$$
\mathrm{e} = \sqrt{\log 2\left|\mathcal{I}^\flat\right|},
$$

$$
E = \sqrt{\log(2\left|\mathcal{I}^\flat\right|)/\left|\mathcal{I}^\flat\right|},
$$

$$
E' = \log\left(2\left|\mathcal{I}^\flat\right|\right)\bigg/\sqrt{\left|\mathcal{I}^\flat\right|}.
$$

makes p', p'', p''' polynomially decreasing. Substitution yields the claim.   □

D   Proof of sensitivity result

Proof of Theorem 3.2. Denoting the probability density functions in the world of the Gaussian Process Regression model as $p(\cdot)$ by construction we have

$$A_n(t) = \log \frac{p(y_n^l(t))p(y_n^r(t))}{p(y_n(t))}$$
$$= \log \frac{p(y_n^r(t))}{p(y_n^r(t)|y_n^l(t))}.$$

Further, denote $f_n^r(t) := \mathbb{E}\left[y_n^r(t)\right]$ and $\varepsilon_n^r(t) := y_n^r(t) - f_n^r(t)$. Define shorthand notation $K_n^r(t) := K(X_n^r(t))$ and $K_n^l(t) := K(X_n^l(t))$. Also let $\hat{f}_n^r(t)$ and $\hat{V}_n(t)$ denote predictive mean and variance of the Gaussian Process Regression for $X_n^r(t)$ given $X_n^l(t)$ and $y_n^l(t)$. Now recall the posterior is Gaussian:

$$p(y_n^r(t)|y_n^l(t)) = \mathcal{N}\left(y_n^r(t)|\hat{f}_n^r(t), \hat{V}_n(t)\right).$$

Define a norm $\|x\|_A := \left\|A^{1/2}x\right\|$ for an arbitrary positive-definite symmetric matrix A. Clearly, $\|x\|_A^2 = \langle x, Ax \rangle$. Now trivial algebra yields

$$A_n(t) \cong \underbrace{\left\|f_n^r(t) - \hat{f}_n^r(t)\right\|_{\hat{V}_n(t)^{-1}}^2}_{T_1} + \underbrace{\|\varepsilon_n^r(t)\|_{\hat{V}_n(t)^{-1} - K_n^r(t)^{-1}}^2}_{T_2}$$
$$+ \underbrace{2\langle \left(f_n^r(t) - \hat{f}_n^r(t)\right)\hat{V}_n(t)^{-1} - f_n^r(t)K_n^r(t)^{-1}, \varepsilon_n^r(t)\rangle}_{T_3},$$

where we use $\cong$ to denote "equality up to an additive deterministic constant". Consider a matrix $K(X_n(t))$ being a block 2×2 matrix with blocks of equal size:

$$K(X_n(t)) = \begin{pmatrix} K_n^l(t) & \mathcal{K} \\ \mathcal{K}^T & K_n^r(t) \end{pmatrix}.$$

Notice that $\hat{V}_n(t)$ is its Schur complement, thus $\lambda_{\max}(\hat{V}_n(t)) \le \lambda_{\max}(K(X_n(t))) \le C$ (the second inequality is due to Assumption 3.4). Using $\sigma^2 > 0$ we have $\lambda_{\min}(\hat{V}_n(t)) > 1/c$ and $\lambda_{\min}(K_n^r(t)) > 1/c$ for some c independent of n. To sum these observations up:

$$\exists c > 0 : \lambda_{\min}(\hat{V}_n(t)), \lambda_{\max}(\hat{V}_n(t)), \lambda_{\max}(K_n^r(t)^{-1}) \in (1/c, c).$$

Having established control over these eigenvalues, we are ready to bound the terms T_2 and T_3 from above under both $\mathbb{H}_0$ and $\mathbb{H}_1$, while T_1 should be bounded from above under $\mathbb{H}_0$ and from below under $\mathbb{H}_1$. Now we bound the test statistic $A_n(t)$ under $\mathbb{H}_0$. Denote $\Delta_f := \left\|f_n^r(t) - \hat{f}_n^r(t)\right\|_\infty$. Then

$$T_1 \le cn\Delta_f^2.$$

In order to bound the second term on a set of high probability we employ Lemma D.1 and obtain for a positive t

$$\mathbb{P}\left\{|T_2 - \mathbb{E}\left[T_2\right]| \ge 4\mathfrak{g}^2 c\sqrt{n\mathrm{t}}\right\} \le \exp(-\mathrm{t}). \tag{D.1}$$

The third term will be controlled using sub-Gaussianity of $\varepsilon_n^r(t)$. For any unit vector u and positive e

$$\mathbb{P}\left\{|\langle u, \varepsilon_n^r(t)\rangle| \geq \mathfrak{g}\sqrt{\mathrm{e}}\right\} \leq 2\exp(-\mathrm{e})$$

and clearly,

$$\left\|\left(f_n^r(t) - \hat{f}_n^r(t)\right)\hat{V}_n(t)^{-1} - f_n^r(t)K_n^r(t)^{-1}\right\| \leq 4\sqrt{n}c(F + \Delta_f),$$

where $F := \max\{\|f\|_\infty, \|f_1^*\|_\infty, \|f_2^*\|_\infty\}$. Hence, on a set of probability at least $1 - 2\exp(-\mathrm{e})$

$$|T_3| \leq 8\sqrt{n}\mathfrak{g}\mathrm{e}c(F + \Delta_f). \tag{D.2}$$

Finally, we choose $\mathrm{t} := 10\log n$ and $\mathrm{e} := 10\log n$. Now under $\mathbb{H}_0$ bound Δ_f by Lemma F.1, recall $\kappa < 1/2$ and obtain

$$|A_n(t) - \mathbb{E}[T_2]| = O\left(n^{1-2\kappa}\log^{2\kappa} n + \sqrt{n\log n}\right)$$
$$= O\left(n^{1-2\kappa}\log^{2\kappa} n\right)$$

on a set of probability at least $1 - 3/n^{10} \to 1$ as $n \to +\infty$. Now we use Theorem 3.1 along with the fact that for n large enough $\alpha > R_A + 3/n^{10}$ and obtain on a set of probability approaching 1

$$\left|x_{n,\alpha}^\flat(t) - \mathbb{E}[T_2]\right| = O\left(n^{1-2\kappa}\log^{2\kappa} n\right).$$

On the other hand, under $\mathbb{H}_1$ the bounds (D.1) and (D.2) still hold and

$$T_1 \geq \lambda_{\min}(\hat{V}_n(t))\left\|f_n^r(t) - \hat{f}_n^r(t)\right\|^2$$
$$\geq \frac{1}{c}n\left(\mathfrak{B}_n^2 - \Delta_f^2\right).$$

Finally, choose $n = n_*$, $t = \tau$, and recall assumption (3.8) to conclude $A_n(\tau) > x_{n,\alpha}^\flat(\tau)$ for large n with probability approaching 1. $\qquad\square$

The following result bounds a quadratic form of a sub-Gaussian vector with high probability. It is a direct corollary of Theorem 1.1 (Hanson-Wright inequality) stated in [39].

Lemma D.1. *Consider a vector $x \in \mathbb{R}^n$ sub-Gaussian with $\mathfrak{g}^2$ and a positive-definite matrix A of size $n \times n$. Let there be a constant Λ independent of n s.t. $\lambda_{\max}(A) \leq \Lambda$. Then for a positive t, large enough n and some absolute positive c*

$$\mathbb{P}\left\{|\langle x, xA\rangle - \mathbb{E}[\langle x, xA\rangle]| \geq c\mathfrak{g}^2\Lambda\sqrt{n\mathrm{t}}\right\} \leq \exp(-\mathrm{t}).$$

E   Anti-concentration inequality

This section uses notation introduced in Section B.

Lemma E.1. *Consider a $2p$-dimensional Gaussian vector $z = (x, X)$, where x and X are p-dimensional. Further, let $\mathrm{Var}[x] = \sigma^2 I_p$ and $Cov(x_j, X_j) = Cov(\eta, \eta^2)$ for arbitrary $1 \leq j \leq p$ and $\eta \sim \mathcal{N}(0, \sigma^2)$. Finally, let $\mathrm{Var}[X] = \mathrm{Var}[\eta^2]I_p$. Then for an arbitrary vector C and $\delta \in \mathbb{R}$*

$$\mathbb{P}\left\{Q(z) < C + \delta\right\} - \mathbb{P}\left\{Q(z) < C\right\} \leq 3\sqrt{\frac{n}{\mathfrak{s}^2}}\delta\left(\sqrt{4\log p} + 2\right)$$

where

$$\mathfrak{s}^2 := \left(\min_{j \in J} \|B(j)\|_F^2 + \|\mathbb{E}[x]\|^2 \right) \sigma^2$$

and the map $Q(\cdot)$ is defined by (B.2).

Proof. Introduce an isotropic Gaussian vector $\tilde{z} = \operatorname{Var}[z]^{-1/2} z \operatorname{Var}[z]^{-1/2}$ and notice, $\sigma^2/3 \leq \operatorname{Var}[\tilde{z}_j] \leq 3\sigma^2$ for all j. Applying Lemma E.2 to $\tilde{z}$ yields the claim. $\qquad \square$

The rest of the proofs of this section mostly follow the Nazarov's inequality proof presented in [14].

Define a map $u : \mathbb{R}^{2p} \to \mathcal{U}$, where $\mathcal{U} := \mathbb{R}^{(p+5)p/2}$:

$$u\left((x_1, x_2, .., x_p)^T, (X_1, X_2, .., X_p)^T\right)$$
$$= (x_1, x_2, .., x_p, X_1, X_2, .., X_p, x_1^2, x_2^2, .., x_p^2, \sqrt{2}x_1 x_2, \sqrt{2}x_1 x_3, .., \sqrt{2}x_{p-1}x_p)^T.$$

With a slight abuse of notation we will use u to denote both the map and an element of its image.

Lemma E.2. *Consider $x \sim \mathcal{N}(0, I_{2p})$ and $a_1, a_2, ..a_{p(p+5)/2} \in \{u \in \mathcal{U} : \|u\| = 1\}$ along with $b_1, b_2, .., b_{p(p+5)/2} \in \mathbb{R}$. Then for all positive δ:*

$$\mathbb{P}\left\{\forall j : a_j^T u(x) \leq b_j + \delta\right\} - \mathbb{P}\left\{\forall j : a_j^T u(x) \leq b_j\right\} \leq \delta\left(\sqrt{4\log p} + 2\right).$$

Proof. Define a set $K(t) := \left\{u \in \mathcal{U} : \forall j \ a_j^T u \leq b_j + t\right\}$, and a function $G(t) := \mathbb{P}\{u(x) \in K(t)\}$. G is absolutely continuous distribution function and hence

$$G(\delta) - G(0) = \int_0^\delta G'_+(t)dt,$$

where G'_+ denotes the right derivative of G. Essentially, the proof boils down to the following lemma.

Lemma E.3.
$$\lim_{\delta \downarrow 0} \frac{\mathbb{P}\{u(x) \in K(\delta) \backslash K(0)\}}{\delta} \leq \delta\left(\sqrt{4\log p} + 2\right).$$

Proof. Denote $K := K(0)$ and note it is a convex polyhedron. Denote a projector onto K as P_K: $\|x - P_K x\| = \min_{y \in K} \|x - y\|$. Now for a (proper) face F of K define

$$N_F = \left\{u \in \mathcal{U} \backslash K : P_K u \in \operatorname{relint}(F)\right\},$$

$$N_F(\delta) = N_F \bigcap K(\delta).$$

Clearly, $K(\delta) \backslash K = \bigcup_{F:\text{face of } K} N_F(\delta)$. Clearly,

$$\exists C > 0 : N_F \subset \{u \in \mathcal{U} : \operatorname{dist}(u, F) \leq C\delta\},$$

hence for any face F of dimensionality less than $\dim \mathcal{U} - 1$ for $\delta \downarrow 0 : \gamma_p(N_F) := \mathbb{P}\{u(x) \in N_F\} = o(\delta)$. Hence,

$$\gamma_p(K(\delta) \backslash K) = \gamma_p\left(\bigcup_{F:\text{facet of } K} N_F(\delta)\right) + o(\delta) \text{ as } \delta \downarrow 0.$$

Now it is left to prove that

$$\sum_{F:\text{facet of } K} \lim_{\delta \downarrow 0} \frac{\gamma_p(N_F(\delta))}{\delta} \leq \sqrt{4 \log p} + 2.$$

By Lemma E.4

$$\lim_{\delta \downarrow 0} \frac{\gamma_p(K(\delta) \backslash K)}{\delta} = \sum_{F:\text{facet of } K} \int_{u(x) \in F} f_{2p}(x) d\sigma(u(x))),$$

where $f_{2p}(\cdot)$ denotes the density of $\mathcal{N}(0, I_{2p})$. Consider facets F such that $\text{dist}(0, F) > 4 \log p$. Choose $\bar{h} = \text{dist}(0, F)v$ (or flip the sign if $\bar{h} \notin F$) and denote $\bar{x} = u^{-1}(\bar{h})$. Further, since $\|\bar{x}\| \geq \sqrt{4 \log p}$,

$$f_{2p}(x) = f_1(\|\bar{x}\|) f_{2p-1}(k) \leq \frac{f_{2p-1}(k)}{p^2}$$

and given the number of facets is less than p^2,

$$\sum_{\substack{F:\text{facet of } K \\ \text{dist}(0,F) > 4 \log p}} \int_{u(x) \in F} f_{2p}(x) d\sigma(u(x))) \leq 1. \tag{E.1}$$

Now turn to the facets F s.t. $\text{dist}(0, F) \leq 4 \log p$. By Lemma E.4,

$$\int_{u(x) \in F} f_{2p}(x) d\sigma(u(x))) \leq \left(\sqrt{4 \log p} + 1\right) \gamma_p(N_F).$$

The final observation is based on the fact that N_F are disjoint and $\gamma_p(\mathcal{U}) = 1$

$$\sum_{\substack{F:\text{facet of } K \\ \text{dist}(0,F) \leq 4 \log p}} \int_{u(x) \in F} f_{2p}(x) d\sigma(u(x))) \leq \sqrt{4 \log p} + 1$$

and its combination with (E.1) completes the proof. $\qquad \square$

$\square$

Lemma E.4.

$$\lim_{\delta \downarrow 0} \frac{\gamma_p(N_F(\delta))}{\delta} = \int_{u(x) \in F} f_{2p}(x) d\sigma(u(x))) \leq \left(\sqrt{\text{dist}(0, F)} + 1\right) \gamma_p(N_F),$$

where $d\sigma$ is the standard surface measure on F.

Proof. Parametrize every $h \in F$ as

$$h = h(\bar{h}, z) = \bar{h} + z_1 q_1 + z_2 q_2 + \ldots + z_{p(p+5)/2-1} q_{p(p+5)/2-1},$$

where $\bar{h}$ is an arbitrary element of F, while q_j form an orthonormal basis on $F - \bar{h}$. Further, choose a unit outward normal vector v to ∂K at F. Then we can parametrize N_F

$$y = y(\bar{h}, z, t) = h(\bar{h}, z) + tv \text{ for } t > 0.$$

Now

$$\gamma_p(N_F) = \int_{y=y(\bar{h},z,t)\in N_F} f_{2p}\left(u^{-1}(y)\right) dzdt$$

$$= \int_{h=h(\bar{h},z)\in F} \left(\int_0^{+\infty} f_{2p}\left(u^{-1}(y)\right) dt\right) dz \tag{E.2}$$

and in the same way

$$\gamma_p(N_F(\delta)) = \int_0^\delta \left(\int_{h=h(\bar{h},z)\in F} f_{2p}\left(u^{-1}(y)\right) dz\right) dt.$$

Thus

$$\lim_{\delta\downarrow 0} \frac{\gamma_p(N_F(\delta))}{\delta} = \int_{h=h(\bar{h},z)\in F} f_{2p}\left(u^{-1}(y)\right) dz,$$

which proves the equality in the claim.

Now for any $h \in F$ and v exist vectors x and n such that $u^{-1}(h+tv) = x + t'n$ for $t' = \sqrt{t}$. Then

$$\int_0^{+\infty} f_{2p}\left(u^{-1}(h+tv)\right) dt = \int_0^{+\infty} f_{2p}\left(x+t'n\right) dt'$$

$$\geq f_{2p}(x) \int_0^{+\infty} e^{-t|x^T n|} e^{-t^2/2} dt \geq \frac{f_{2p}(x)}{|x^T n|+1}. \tag{E.3}$$

Combination of (E.2) and (E.3) yields

$$\gamma_p(N_F) \geq \int_{h\in F} \frac{1}{|x^T n|+1} f_{2p}(x)dz, \text{ where } x = u^{-1}(h).$$

Now choose $h = \text{dist}(0,F)$, note $|x^T n| = \sqrt{\text{dist}(0,F)}$ and establish the claim. □

F Consistency of Gaussian Process Regression by [49]

In this section we quote a consistency result for predictions of Gaussian Process Regression.

Lemma F.1 (Corollary 2.1 in [49]). *Assume, ε_i are sub-Gaussian. Let f^* be $\aleph$-smooth Sobolev and $\kappa := (\aleph - 1/2)/(2\aleph)$ or $\aleph$-smooth Hölder and $\kappa := \aleph/(2\aleph+1)$. Further let $k(\cdot,\cdot)$ satisfy Assumption 3.1 and Assumption 3.2. Then, for the training sample size n going to infinity with probability at least $1 - n^{-10}$ we have*

$$\left\|f^* - \hat{f}\right\|_\infty = O\left(\left(\frac{\log n}{n}\right)^\kappa\right),$$

where $\hat{f}$ denotes the predictive function.

References

[1] Alexander Aue, Siegfried Hörmann, Lajos Horváth, and Matthew Reimherr. Break detection in the covariance structure of multivariate time series models. *Ann. Statist.*, 37(6B):4046–4087, 12 2009.

[2] Alexander Aue and Lajos Horváth. Structural breaks in time series. *Journal of Time Series Analysis*, 34(1):1–16, 2013.

[3] Valeriy Avanesov. Structural break analysis in high-dimensional covariance structure. *arXiv e-prints*, page arXiv:1803.00508, March 2018.

[4] Valeriy Avanesov and Nazar Buzun. Change-point detection in high-dimensional covariance structure. *Electron. J. Statist.*, 12(2):3254–3294, 2018.

[5] Jushan Bai. Estimating multiple breaks one at a time. *Econometric Theory*, 13(3):315–352, 1997.

[6] Jushan Bai. Likelihood ratio tests for multiple structural changes. *Journal of Econometrics*, 91(2):299 – 323, 1999.

[7] Jushan Bai and Pierre Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.

[8] Jushan Bai and Pierre Perron. Critical values for multiple structural change tests. *The Econometrics Journal*, 6(1):72–78, 2003.

[9] Danielle S. Bassett, Nicholas F. Wymbs, Mason a. Porter, Peter J. Mucha, Jean M. Carlson, and Scott T. Grafton. Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 108(18):7641, 2010.

[10] Luc Bauwens, Sébastien Laurent, and Jeroen V K Rombouts. Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, 21(1):79–109, jan 2006.

[11] Sourav Chatterjee. A simple invariance theorem. *arXiv Mathematics e-prints*, page math/0508213, August 2005.

[12] Sourav Chatterjee. A generalization of the lindeberg principle. *Ann. Probab.*, 34(6):2061–2076, 11 2006.

[13] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 12 2013.

[14] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Detailed proof of Nazarov's inequality. *arXiv e-prints*, page arXiv:1711.10696, November 2017.

[15] Haeran Cho. Change-point detection in panel data via double cusum statistic. *Electron. J. Statist.*, 10(2):2000–2038, 2016.

[16] Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society Series B*, 77(2):475–507, 2015.

[17] Mihaela Şerban, Anthony Brockwell, John Lehoczky, and Sanjay Srivastava. Modelling the dynamic dependence structure in multivariate financial time series. *Journal of Time Series Analysis*, 28(5):763–782, 2007.

[18] M. Csörgö and L. Horváth. *Limit theorems in change-point analysis*. Wiley series in probability and statistics. J. Wiley & Sons, Chichester, New York, 1997.

[19] Marc Peter Deisenroth and Jun Wei Ng. Distributed Gaussian processes. *32nd International Conference on Machine Learning, ICML 2015*, 2:1481–1490, 2015.

[20] Robert F. Engle, Victor K. Ng, and Michael Rothschild. Asset pricing with a factor-arch covariance structure. Empirical estimates for treasury bills. *Journal of Econometrics*, 45(1-2):213–237, 1990.

[21] Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, 42(6):2243–2281, 12 2014.

[22] Zhonghao Fu and Yongmiao Hong. A model-free consistent test for structural change in regression possibly with endogeneity. *Journal of Econometrics*, 211(1):206 – 242, 2019. Annals Issue in Honor of Jerry A. Hausman.

[23] Friedrich Goetze and A Tikhomirov. Asymptotic distribution of quadratic forms and applications. *Journal of Theoretical Probability*, 15:423–475, 04 2002.

[24] Javier Hidalgo. A nonparametric conditional moment test for structural stability. *Econometric Theory*, 11(4):671–698, 1995.

[25] Moritz Jirak. Uniform change point tests in high dimension. *Ann. Statist.*, 43(6):2451–2483, 12 2015.

[26] Hyune-Ju Kim. Tests for a Change-Point in Linear Regression. 23(1989), 1994.

[27] Jeankyung Kim and Hyune-Ju Kim. Consistent model selection in segmented line regression. *Journal of Statistical Planning and Inference*, 170:106–116, 2016.

[28] M. Lavielle and G. Teyssière. Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306, 2006.

[29] Housen Li, Axel Munk, and Hannes Sieling. Fdr-control in multiscale change-point segmentation. *Electron. J. Statist.*, 10(1):918–959, 2016.

[30] Miguel A. Martinez-Beneito, Gonzalo García-Donato, and Diego Salmerón. A Bayesian Join-point regression model with an unknown number of break-points. *Annals of Applied Statistics*, 5(3):2150–2168, 2011.

[31] David S. Matteson and Nicholas A. James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

[32] Thomas Mikosch, Søren Johansen, and Eric Zivot. Handbook of Financial Time Series. *Time*, 468(1996):671–693, 2009.

[33] V. Paulauskas and A. Rackauskas. *Approximation Theory in the Central Limit Theorem: Exact Results in Banach Spaces*. Mathematics and its Applications. Springer Netherlands, 2012.

[34] Florian Pein, Hannes Sieling, and Axel Munk. Heterogeneous Change Point Inference. *arXiv e-prints*, page arXiv:1505.04898, May 2015.

[35] Zhongjun Qu and Pierre Perron. Estimating and testing structural changes in multivariate regressions. *Econometrica*, 75(2):459–502, 2007.

[36] Richard E. Quandt. Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association*, 55(290):324–330, 1960.

[37] Rasmussen and Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[38] V. I. Rotar'. Limit theorems for polylinear forms. *Journal of Multivariate Analysis*, 9(4):511–530, December 1979.

[39] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18:9 pp., 2013.

[40] Florian Stimberg, Manfred Opper, Guido Sanguinetti, and Andreas Ruttor. Inference in continuous-time change-point models. In J Shawe-Taylor, R S Zemel, P L Bartlett, F Pereira, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2717–2725. Curran Associates, Inc., 2011.

[41] Florian Stimberg, Andreas Ruttor, and Manfred Opper. Bayesian inference for change points in dynamical systems with reusable states - a chinese restaurant process approach. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1117–1124, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[42] Liangjun Su and Zhijie Xiao. Testing structural change in time-series nonparametric regression models. *Statistics and its interface*, 1, 10 2006.

[43] A Suvorikova and V Spokoiny. Multiscale Change Point Detection. *Theory of Probability & Its Applications*, 61(4):665–691, 2017.

[44] M. Talagrand. *Spin Glasses: A Challenge for Mathematicians: Cavity and Mean Field Models*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge A Series of Modern Surveys in Mathematics. Springer, 2003.

[45] Victor V. Chernozhukov, Denis Chetverikov, and Kengo Kato. Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, 162, 2013.

[46] Michael Vogt. Testing for structural change in time-varying nonparametric regression models. *Econometric Theory*, 31(4):811–859, 2015.

[47] L. Ju. Vostrikova. Detecting "disorder" in multidimensional random processes. *Soviet Mathematics: Doklady*, 24:55–59, 1981.

[48] Yao Xie and David Siegmund. Sequential multi-sensor change-point detection. *Ann. Statist.*, 41(2):670–692, 04 2013.

[49] Yun Yang, Anirban Bhattacharya, and Debdeep Pati. Frequentist coverage and sup-norm convergence rate in Gaussian process regression. pages 1–43, 2017.

[50] Changliang Zou, Guosheng Yin, Long Feng, and Zhaojun Wang. Nonparametric maximum likelihood approach to multiple change-point problems. *Ann. Statist.*, 42(3):970–1002, 06 2014.