UNIVERSITY OF CALIFORNIA SAN DIEGO

Template Matching and Texture Segmentation: Theory, Methods and Algorithms

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics with a Specialization in Statistics

by

Lin Zheng

Committee in charge:

Professor Ery Arias-Castro, Chair
Professor Truong Quang Nguyen
Professor Ronghui (Lily) Xu
Professor Danna Zhang
Professor Wenxin Zhou

2021

The dissertation of Lin Zheng is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep and sincere gratitude to my advisor Professor Ery Arias-Castro, for his continuous support of my Ph.D. study and research, for his motivation, patience and professional knowledge in mathematics and statistics. His sparking wisdom and deep insight into the problem helped me tame the seemingly impossible research projects. He provided me invaluable guidance throughout my research, and encouraged me to carry on research projects independently. It was a great privilege and honor to work and study under his supervision.

Besides, I would like to thank the rest of my doctoral committee: Professor Truong Quang Nguyen, Professor Ronghui (Lily) Xu, Professor Danna Zhang and Professor Wenxin Zhou for offering me excellent courses, insightful comments and encouragement on my research. Outside UCSD, I am immensely grateful for the numerous helpful discussions with Mr. Samer Fasheh, Ms. Daksha Agarwal and Dr. Cristina Marinovici during my internship at Microsoft.

I would also like to express my gratitude to all my friends in the Department of Mathematics in UCSD, including but not limited to Zheng, Rong, Yuchao, Tuo, Chao and Yingjia for sharing fruitful research ideas, discussions and relaxing time with me. Moreover, I would like to thank all faculty and staff members in the Department of Mathematics who provided me tremendous guidance and help during my Ph.D. life.

Last but not least, I would like to thank my whole family for their support, endless love and understanding of my PhD study and all my career. In particular, my parents brought me to the world, raised me up, and they are the role models teaching me to learn, be independent, be strong, and pursue the better. As the most important part of my life, they have always been supportive of every decision I made. My life during the five years would not be such an incredible journey without them.

Chapter 2, in full, is a version of the paper " Template Matching and Change Point Detection by M-estimation", Arias-Castro, Ery; Zheng, Lin. The manuscript has been submitted

for publication in a major statistical journal. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is a version of the paper "Template Matching with Ranks", Arias-Castro, Ery; Zheng, Lin. The manuscript has been submitted for publication in a major statistical journal. The dissertation author was the primary investigator and author of this material.

Chapter 4, in full, is a version of the paper "Some Theory for Texture Segmentation", Zheng, Lin. The manuscript has been submitted for publication in a major journal in electrical engineering. The dissertation author was the primary investigator and author of this material.

VITA

2013	B. S. in Mathematics, Nankai University, Tianjin China
2016	M. S. in Statistics, Nankai University, Tianjin China
2017-2021	Graduate Teaching Assistant, University of California San Diego
2021	Ph. D. in Mathematics with a Specialization in Statistics, University of California San Diego

PUBLICATIONS

Tao Wang and Lin Zheng, "A Robust Variable Screening Method for High-dimensional Data", *Journal of Applied Statistics*, 44.10, (2017): 1839-1855.

Ery Arias-Castro and Lin Zheng, "Template Matching and Change Point Detection by M-estimation", *Submitted*, 2020.

Ery Arias-Castro and Lin Zheng, "Template Matching with Ranks", *Submitted*, 2020.

Lin Zheng, "Some Theory for Texture Segmentation", *Submitted*, 2020.

ABSTRACT OF THE DISSERTATION

Template Matching and Texture Segmentation: Theory, Methods and Algorithms

by

Lin Zheng

Doctor of Philosophy in Mathematics with a Specialization in Statistics

University of California San Diego, 2021

Professor Ery Arias-Castro, Chair

Template matching, as an important topic in image processing, is the process of matching a clean and noiseless template to an observed, typically noisy signal. This topic is closely related to the problem of matching two or more noisy signals, sometimes referred to as 'aligning' or 'registering' the signals, and to methodology in spatial statistics falling under the umbrella name of 'scan statistic'. When the template has a point of discontinuity, matching a template to the signal can be interpreted as detecting the location of the discontinuity, a more specialized task more broadly referred to as 'change-point detection' in statistics. In Chapter 2, we study a standard mathematical model for matching a template to a noisy signal by M-estimation. While the most popular method may still be based on maximizing the (Pearson) correlation, the estimators

we study can be made much more robust to heavy-tailed noise or the presence of outlying observations. We draw on standard empirical process theory and decision theory, as expounded in [VdV98], to derive limit distributions and minimax convergence rates of M-estimators in a wide array of situations. In Chapter 3, we suggest a rank-based method for matching a template to a noisy signal, and study its asymptotic properties using some well-established techniques in empirical process theory combined with Hájek's projection method. The resulting estimator of the shift is shown to achieve a parametric rate of convergence and to be asymptotically normal.

Texture segmentation, as another crucial role in image processing, is the process of partitioning the image into differently textured regions. In Chapter 4, we present our related work on texture segmentation. We provide some theoretical guarantees for the prototypical approach which consists in extracting local features in the neighborhood of a pixel and then applying a clustering algorithm for grouping the pixel according to these features. The proposed algorithms work on both stationary and non-stationary random fields.

Chapter 1

Introduction

1.1 Template matching

A basic task in signal processing is the matching of a template, aka filter or pattern, to a noisy signal [Bru09, Tur60]. There has been an extensive amount of research in this very broad area, as applications are many, from locating blood vessels and tumors in medical imaging to more sophisticated tasks such as locating a human face or 'recognizing' objects like cars in images [HL01, BMP02, Low99, SWP05].

While the literature on template matching is very large on the side of methodology and applications [ZF03, HH01, SDP13], there is a sizable literature that develops theory for such problems. We note that the signals are typically smoothed before the alignment is carried out, and that the Fourier transform plays an important role, and the noise is often assumed to be Gaussian or to have sub-Gaussian tails. This is in contrast to our setting in which no smoothing is used and the Fourier transform plays no role. We consider the following model for template matching,

$$Y_i = f(X_i - \theta^*) + Z_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a known function referred to as the *template*, and $\theta^* \in \mathbb{R}$ is the unknown *shift* of interest. The noise $Z_1, \ldots, Z_n$ are assumed iid with density ϕ. We work with minimal assumptions on the noise distribution. We mention a more recent line of work on this problem which focuses on more complex and noisy settings arising in applications such as cryo-electron microscopy [PWB+19, WS13, PWBM18].

Our goal is, in signal processing terminology, to match the template f to the signal Y, which in statistical terms consists in the estimation of the shift θ^*. We consider an *M-estimator* defined implicitly as the solution to the following optimization problem

$$\hat{\theta} := \arg\min_{\theta} \sum_{i=1}^{n} \mathsf{L}(Y_i - f(X_i - \theta)), \tag{1.2}$$

where L is a loss function chosen by the analyst. Popular choices are

1. Squared error loss
$$\mathsf{L}(y) = y^2. \tag{1.3}$$

2. Absolute-value loss
$$\mathsf{L}(y) = |y|. \tag{1.4}$$

3. Huber loss
$$\mathsf{L}(y) = \begin{cases} \frac{1}{2}y^2 & \text{if } |y| \le c, \\ c|y| - \frac{1}{2}c^2 & \text{if } |y| > c. \end{cases} \tag{1.5}$$

4. Tukey loss
$$\mathsf{L}(y) = \begin{cases} 1 - (1 - (y/c)^2)^3 & \text{if } |y| \le c, \\ 1 & \text{if } |y| > c. \end{cases} \tag{1.6}$$

Besides M-estimators for template matching, we also consider estimators based on ranks. Rank-based methods are, of course, classical in statistics [ŠSH99, HM10, Leh06, GC11]. The

theory of rank tests is particularly well-developed, featuring some prominent methods such as the Wilcoxon/Mann–Whitney and Kruskal–Wallis two- and multi-sample tests, and essentially all (other) distribution-free tests for goodness-of-fit such as the Kolmogorov–Smirnov test, and more closely related to our topic here, the Spearman and Kendall rank correlation tests for independence. The theory of rank-based estimators is also well developed, although perhaps not as well-known [HM10]. The most famous example may be the Hodges–Lehmann estimator, which is derived from the Wilcoxon signed-rank test. In multiple linear regression, the asymptotic linearity and resulting normality of certain rank-based estimators is established in a number of publications [Jur71, HW88, KM93, GKS96, Dra88]. Closer to our setting, some papers consider the use of ranks for the detection and/or localization of one or multiple change-points [Dar76, GH98, Huš97, LYFLLC15, Ger18, WWZ20, ACCTW18].

Rank-based estimators go by the name of *R-estimators* in the statistics literature. In our setting, for matching the template f to the signal Y, the most direct route to such an estimator is to replace the response values with their ranks, yielding

$$\hat{\theta}_{\text{rank}} := \arg\max_{\theta} \sum_{i=1}^{n} R_i f(x_i - \theta), \tag{1.7}$$

where R_i denotes the rank of Y_i in $\{Y_1, \ldots, Y_n\}$ in increasing order.

1.2 Texture segmentation

We address the problem of texture segmentation, which fits within the larger area of image segmentation, with a particular focus on images that contain textures. The goal is to partition the image, i.e., group the pixels into differently textured regions. At least in recent decades, texture segmentation methods are almost invariably based on extracting local features around each pixel, such as SIFT [Low99], which are then fed into a clustering algorithm, such as k-means. An emblematic approach in this context is that of [SM00], who used the pixel value as feature,

arguably the simplest possible choice, and applied a form of spectral clustering to group the pixels.

On the one hand, the process of extracting features has undergone some important changes over the years, ranging from Gabor filters or wavelets, multi-resolution or multiscale aggregation approaches, to the use deep learning. In this thesis, we extract features by collecting local second moment information on patches. We do so in a stylized setting which is nonetheless a reasonable mathematical model for the problem of texture segmentation. Markov random fields (MRF) are common models for textures [CJ83, GG86], and arguably the most popular in theoretical texture analysis [RH05, ACBLV18, Ver10a, Ver10b, VV09]. This is the model that we use. Although supplanted by the more recent feature extraction methods mentioned above, which in recent years are invariably nonparametric, Gaussian MRF in particular remain the most commonly-used parametric model for textures, also used in the development of methodology not too long ago [CC85, ZWM98, MC91, PS06]. When textures are modeled by stationary Gaussian MRF, what characterizes them is the covariance structure, so that in congruence with adopting Gaussian MRF as models for textures, when assumed stationary the feature we extract is the (local) covariance. When textures are not assumed stationary, we also incorporate location as an additional feature, as the covariance structure may change within a textured region.

On the other hand, the basic approach calls for applying a clustering algorithm to the extracted features. Features are typically represented by (possibly high-dimensional) feature vectors, as is the case with the features that we work with, and thus a large number of clustering methods are applicable, some of them coming with theoretical guaranties such as k-means [AV07], Gaussian mixture models [Das99, VW04, HK13], hierarchical clustering [DL05, Das10], including single-linkage clustering [AC11], and spectral clustering [NJW02]. In this work, we use k-means in the context of stationary textures and singe-linkage clustering in the context of non-stationary textures.

1.3 Thesis structure

The dissertation is organized as follows. In Chapter 2, we consider the fundamental problem of matching a template to a signal. We do so by M-estimation, which encompasses procedures that are robust to gross errors (i.e., outliers). Using standard results from empirical process theory, we derive the convergence rate and the asymptotic distribution of the M-estimator under relatively mild assumptions. We also discuss the optimality of the estimator, both in finite samples in the minimax sense and in the large-sample limit in terms of local minimaxity and relative efficiency. Although most of the paper is dedicated to the study of the basic shift model in the context of a random design, we consider many extensions towards the end of the paper, including more flexible templates, fixed designs, the agnostic setting, and more.

In Chapter 3, we continue to focus on the problem of matching a template to a noisy signal. Motivated by some recent proposals in the signal processing literature, we suggest a rank-based method and derive asymptotic (normal) limit distribution of the R-estimator by empirical process theory and Hájek's projection method. The rate of convergence for our R-estimator is shown to be parametric and minimax optimal. Some numerical simulations corroborate these findings.

In Chapter 4, we address the problem of texture segmentation using approaches via extracting local features and clustering, which facilitates a sharp analysis by leveraging the rich theory on clustering. On the one hand, for stationary textures, which we model with Gaussian Markov random fields, we construct the feature for each pixel by calculating the sample covariance matrix of its neighborhood patch and cluster the pixels by an application of k-means to group the covariance matrices. We show that this generic method is consistent. On the other hand, for non-stationary fields, we include the location of the pixel as an additional feature and apply single-linkage clustering. We again show that this generic and emblematic method is consistent. We complement our theory with some numerical experiments performed on both generated and natural textures.

Chapter 2

Template Matching and Change Point Detection by M-estimation

2.1 Introduction

Scan statistic In statistics, template matching problem has been considered in different variants. Such is the literature on the scan statistic [GB12, GPW09, GNW01], where the focus has been on the detection of the presence of the filter somewhere in the noisy signal, rather than the estimation of its location (when present), for which theory has been developed, including first-order performance bounds [Wal10, DMM03, ACDH05, ACCD11, ACCTW18] with a minimax decision theory perspective, as well as more refined results studying and even establishing the limit distribution [NW04, PGKS05, GZ04, WG14, HP06, PWM18, KMW20, SAC16]. Some of this has some nontrivial intersections with the study of the maximum of various types of random walks and similar processes [Sha95, Jia02, BK06, SV95, Kab11]. In this literature, there are comparatively very few papers that tackle the problem of estimating the location of the template: [JCL10] establish a consistency result for the location of very short intervals, while [Kou17]

shows that the scan statistic is, as a location estimator, consistent near the signal-to-noise ratio required for mere detection.

Change point detection Discontinuities are features of great importance in practice, and it is no coincidence that the template that is most often considered in the scan statistic literature is the indicator of an interval — or multiple intervals as in [JCL10, HJ10, FMS14] — or a rectangle or other shape as in [ACCD11, Wal10]. A discontinuity is sometimes called a change point, and there is also a large body of work in that area. While some of the work on change point detection happens under the umbrella of sequential analysis where data are streaming in [Sie13], what we are discussing here is most closely related to the 'offline' or 'a posteriori' setting where all data are readily available [TOV20]. This spans, in itself, a very large literature [CG11, CH97, BD13, BN93]. First-order performance bounds are available, for example in [FMS14, HS10], but distributional limits are more scarce, at least when it comes to the location of the discontinuities. [Hin70] derives the asymptotic distribution of the maximum likelihood ratio when everything else about the model is known. [YA89] obtain the asymptotic distribution of the least squares estimator of a piecewise constant signal, and this is extended to other U-type statistics in [Dör11, Mau18, Fer01, Bai97]. [Fer94, DÖ1] consider estimating the location of a change of distribution in a sequence of random variables. We note that all this work is done in the context of a deterministic design corresponding to a regular grid (as in a signal processing setting).

Alignment The problem of matching a (clean/noiseless) template to a signal is closely related to the problem of matching two or more noisy signals, sometimes referred to as aligning or registering or synchronizing the signals. While this literature is also very large on the side of methodology and applications [ZF03, HH01, SDP13], there is a sizable literature that develops theory for such problems. Some of these theoretical developments were made in the context of functional data, which is where in statistics such problems are most prominent, and where the problem is sometimes called 'self-modeling' or 'shape invariant modeling' [LSM72]. In this

context, consistency results and/or rates of convergence are obtained in [KG88, KG92, WG99], while distributional limits are derived in [HM90, GLM07, BGV09, TIR11, KE95], and also in [Vim10], where questions of efficiency are considered in a semi-parametric model were only smoothness assumptions are made on the common 'shape'. [CD12, CD15] consider a hypothesis testing problem in that setting. We note that the signals are typically smoothed before alignment, and that the Fourier transform plays an important role, and the noise is usually assumed to be Gaussian or to have sub-Gaussian tails, while in our setting, no smoothing is used and the Fourier transform is not included, and we work with minimal assumptions on the noise distribution.

2.1.1 Model

We assume a regression model with additive error

$$Y_i = f(X_i - \theta^*) + Z_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a known function referred to as the *template*, and $\theta^* \in \mathbb{R}$ is the unknown *shift* of interest. The design points $X_1, \ldots, X_n$ are assumed to be iid with density λ. The noise or measurement error variables $Z_1, \ldots, Z_n$ are assumed iid with density ϕ, and independent of the design points. Note that in this model $(X_1, Y_1), \ldots, (X_n, Y_n)$ independent and identically distributed.

Assumption 1. We assume everywhere that f is compactly supported and, for convenience, càdlag with none or finitely many discontinuities. In particular, f is bounded. And to make sure the shift parameter is identifiable, we also assume that, for any $\theta \neq \theta^*$, $f(\cdot - \theta) \neq f(\cdot - \theta^*)$ on a set of positive measure under λ, or equivalently, $\int (f(x - \theta^*) - f(x - \theta))^2 \lambda(x) \mathrm{d}x > 0$ whenever $\theta \neq \theta^*$.

Assumption 2. We assume everywhere that λ is compactly supported.

Assumption 3. We assume everywhere that ϕ is even, so that the noise is symmetric about 0.

The assumption that the template and the design density both have compact support is

for convenience — although it already applies in most of the settings encountered in practice. It allows us to effectively restrict the parameter space to a compact interval of the real line. Indeed, take A large enough that f and λ are both supported on $[-A, A]$. Then the model (2.1) is effectively parameterized by $\theta \in [-2A, 2A]$ as the model is the same, namely $Y_i = Z_i$, whenever θ is outside that interval. The assumption that the noise is symmetric is not needed everywhere, but already covers interesting situations.

2.1.2 Goal and methods

In signal processing terminology, we focus on matching the template f to the signal Y, which in statistical terms consists in the estimation of the shift θ^*. We consider an *M-estimator* defined as follows

$$\hat{\theta} := \arg\min_{\theta} \sum_{i=1}^{n} \mathsf{L}(Y_i - f(X_i - \theta)), \tag{2.2}$$

where L is a loss function chosen by the analyst. A popular choice is the squared error loss, $\mathsf{L}(y) = y^2$, which defines the least squares estimator

$$\hat{\theta} := \arg\min_{\theta} \sum_{i=1}^{n} (Y_i - f(X_i - \theta))^2. \tag{2.3}$$

This is the maximum likelihood estimator when the noise distribution is Gaussian. Another popular choice is the absolute-value loss, $\mathsf{L}(y) = |y|$, which defines the least absolute-value estimator

$$\hat{\theta} := \arg\min_{\theta} \sum_{i=1}^{n} |Y_i - f(X_i - \theta)|. \tag{2.4}$$

This is the maximum likelihood estimator when the noise distribution is Laplace. Other popular losses include the Huber loss and the Tukey loss. The Huber loss is of the form

$$\mathsf{L}(y) = \begin{cases} \frac{1}{2}y^2 & \text{if } |y| \le c, \\ c|y| - \frac{1}{2}c^2 & \text{if } |y| > c. \end{cases} \tag{2.5}$$

Like the squared error and absolute-value losses, this loss is even and convex. The Tukey loss is of the form

$$\mathsf{L}(y) = \begin{cases} 1 - (1 - (y/c)^2)^3 & \text{if } |y| \le c, \\ 1 & \text{if } |y| > c. \end{cases} \tag{2.6}$$

The Tukey loss is also even, but not convex as it is in fact bounded. In both cases, $c \ge 0$ is a parameter traditionally chosen to maximize the efficiency of the estimator relative to the maximum likelihood estimator under a particular noise distribution (often Gaussian).

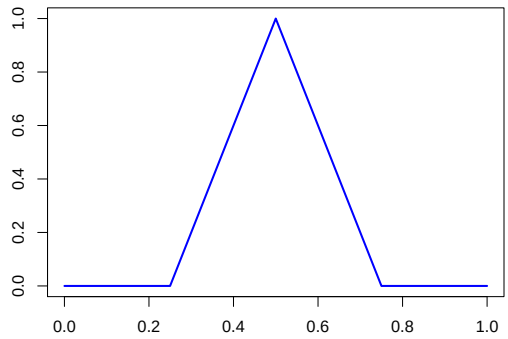
Assumption 4. We assume everywhere that L is non-negative, even, non-decreasing on away from the origin, and because these are the losses used in practice, we also assume that L is either Lipschitz or has a Lipschitz derivative. (We obviously assume that L is not constant.)

Remark 1. If in addition $c := \int \exp(-L(z))\mathrm{d}z < \infty$, then $\phi(z) := c^{-1}\exp(-L(z))$ satisfies Assumption 3, and $\hat{\theta}$ in (2.2) is the maximum likelihood estimator when this is the noise distribution. This additional condition is for example fulfilled when the loss is even and convex.

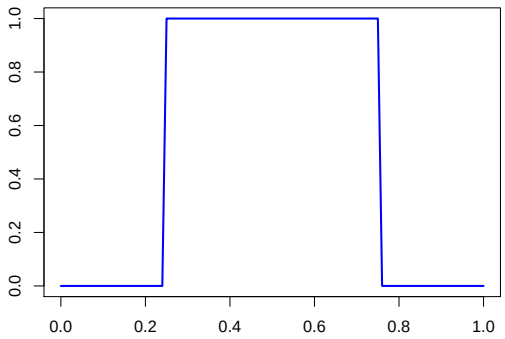
2.1.3 Content

Our focus will be on a generic M-estimator and its asymptotic properties (as $n \to \infty$). In Section 2.2, we establish the consistency of the M-estimator under mild assumptions. In Section 2.3, we consider the 'smooth setting', which corresponds here to the case where the template is Lipschitz. We obtain distributional limits and convergence rates in $\sqrt{n}$, and discuss some optimality properties: local asymptotic minimaxity, relative efficiency, and finite-sample

minimaxity. See Figure 2.1(a) for an illustration. In Section 2.4, we consider the 'non-smooth setting', which in particular includes the case where the template is discontinuous. Similarly, we obtain the limit distribution and the convergence rate, which is in n when the template is piecewise Lipschitz with at least one discontinuity. See Figure 2.1(b) for an illustration. In Section 2.5 we consider more flexible models for which we derive similar results. In Section 2.6, we discuss a number of variants and extensions, including the setting where the design is fixed. In Section 2.7, we present the result of some numerical experiments, which are only meant to probe our theory in finite samples.



(a) When the template is Lipschitz, and the other model components satisfy some mild assumptions, the model is 'smooth'. The M-estimator is $\sqrt{n}$-consistent and asymptotically normal.



(b) When the template is piecewise Lipschitz with at least one discontinuity, and the other model components satisfy some mild assumptions, the M-estimator is n-consistent and its asymptotic distribution, although not well-defined, is essentially the minimum of a marked Poisson process.

Figure 2.1: Two emblematic templates: a Lipschitz template on the left representing a 'smooth' setting, and a piecewise Lipschitz template on the right representing a 'non-smooth' setting.

2.2 Consistency

Having chosen to work with a particular loss function L, define

$$m_\theta(x,y) := \mathsf{L}(y - f(x-\theta)), \tag{2.7}$$

so that the estimator in (2.2) can be equivalently defined via

$$\hat{\theta}_n := \arg\min_{\theta} \widehat{M}_n(\theta), \qquad \widehat{M}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i, Y_i), \tag{2.8}$$

where we have added the subscript n to emphasize that the estimator is being computed on a sample of size n. So that we can carry out a large-sample analysis, we require that

$$\mathbb{E}[m_\theta(X,Y)] < \infty, \quad \text{for all } \theta, \tag{2.9}$$

where (X,Y) is a generic observation and $\mathbb{E}$ is the expectation with respect to the distribution of (X,Y) under θ^*, the true value of the shift. Under Assumption 1 and Assumption 4, the following can be seen to be sufficient

$$\mathbb{E}[\mathsf{L}(Z)] < \infty, \tag{2.10}$$

where Z is a generic noise random variable. For the squared error loss, the requirement is that the noise distribution have a finite second moment, while for the absolute-value and the Huber losses, the requirement is that the noise distribution have a finite first moment, and the requirement is automatically fulfilled when the loss is bounded. When (2.9) holds, $\widehat{M}_n(\theta)$ has a well-defined expectation under θ^* given by

$$M(\theta) := \mathbb{E}[m_\theta(X,Y)]. \tag{2.11}$$

In particular, for any θ, $\widehat{M}_n(\theta)$ converges to $M(\theta)$ in probability by the law of large numbers, and thus one anticipates that, under some additional conditions perhaps, $\hat{\theta}_n$ would converge to a minimizer of M.

Assumption 5. M introduced in (2.11) is well-defined. Moreover, θ^* is the unique minimum of M and $\inf\{M(\theta) : |\theta - \theta^*| \geq \delta\} > M(\theta^*)$ for every $\delta > 0$.

With Assumption 3 and Assumption 4 in place, Assumption 5 is satisfied for the squared error, absolute-value, and Huber losses, and also for the Tukey loss if in addition the noise

distribution is unimodal. See Lemma 10.

Theorem 1 (Consistency)**.** *Under the basic assumptions,* $\hat{\theta}_n$ *is consistent for* θ^*.

Proof. The requirements for applying [VdV98, Th 5.7] are *(i)* M takes its minimum at θ^* and is bounded away from its minimum at θ away from θ^* — which is implied in Assumption 5; and *(ii)* $\widehat{M}_n$ converges to M uniformly, meaning that

$$\sup_{\theta} |\widehat{M}_n(\theta) - M(\theta)| \to 0, \quad n \to \infty, \tag{2.12}$$

in probability — which is due to the fact that the function class $\{m_\theta : \theta \in \mathbb{R}\}$ is Glivenko–Cantelli. This latter property is essentially established in [VdV98, Ex 19.8], where the assumption of continuity can be relaxed to continuity almost everywhere, which holds here by the fact that f and L are at least piecewise continuous. [VdV98, Ex 19.8] works under the assumption that the parameter space is compact, and this is effectively the case here. □

2.3 Smooth setting

We start by analyzing the situation where f is Lipschitz. It turns out that, when this is the case, under mild assumptions on the design and noise distributions as well as the loss function, the situation is 'standard' for a parametric model in the sense that the M-estimator is $\sqrt{n}$ - consistent and asymptotically normal. In addition, the model is 'smooth' in the sense of being quadratic mean differentiable [VdV98, Sec 5.5], which then implies that the maximum likelihood estimator — which can coincide with the M-estimator as mentioned in Remark 1 — is locally asymptotic minimax and efficient.

Remark 2. We remind the reader that a Lipschitz function is differentiable almost everywhere with a bounded derivative, and that it is the integral of that derivative (i.e., it is absolutely continuous). For such a function g, we will denote by g' its derivative and by $|g'|_\infty$ its Lipschitz constant

(which bounds the derivative when it exists).

2.3.1 Asymptotic normality

We first consider a smooth loss, encompassing the squared error loss and the Huber loss, among others.

Theorem 2. *Suppose the basic assumptions are in place. Assume, in addition, that f is Lipschitz, that λ is Lipschitz on an open set containing the support of $f(\cdot - \theta^*)$, and that L has a Lipschitz first derivative. Unless L itself is Lipschitz, assume that the noise distribution has finite second moment. Then $\sqrt{n}(\hat{\theta}_n - \theta^*)$ is asymptotically normal with mean 0 and variance τ^2, with*

$$\tau^2 := \frac{C_{\phi,L}}{\mathbb{E}[f'(X-\theta^*)^2]}, \quad C_{\phi,L} := \frac{\mathbb{E}[L'(Z)^2]}{\mathbb{E}[L''(Z)]^2}. \tag{2.13}$$

Remark 3. The assumption that λ is Lipschitz is not needed, as can be seen by using the approach of Section 2.4. However, it makes for a particularly straightforward proof.

Proof. Assume $\theta^* = 0$ without loss of generality. In [VdV98, Th 5.23], the first condition is that $\theta \mapsto m_\theta(x,y)$ be differentiable at 0 for almost every (x,y), which is clearly the case here, as it is differentiable with derivative

$$\dot{m}_\theta(x,y) = f'(x-\theta)L'(y - f(x-\theta)). \tag{2.14}$$

If L itself is Lipschitz, we have

$$\begin{aligned} |m_{\theta_1}(x,y) - m_{\theta_2}(x,y)| &= |L(y - f(x-\theta_1)) - L(y - f(x-\theta_2))| && (2.15)\\ &\le |L'|_\infty |f(x-\theta_1) - f(x-\theta_2)| && (2.16)\\ &\le |L'|_\infty |f'|_\infty |\theta_1 - \theta_2| && (2.17)\\ &=: \overline{m}(x,y)\,|\theta_1 - \theta_2|. && (2.18) \end{aligned}$$

Otherwise, we have

$$|m_{\theta_1}(x,y) - m_{\theta_2}(x,y)| = |\mathsf{L}(y - f(x-\theta_1)) - \mathsf{L}(y - f(x-\theta_2))| \tag{2.19}$$

$$\leq |\mathsf{L}''|_\infty \left(|y - f(x)| + 2|f|_\infty\right) |f'|_\infty |\theta_1 - \theta_2| \tag{2.20}$$

$$=: \overline{m}(x,y)\,|\theta_1 - \theta_2|. \tag{2.21}$$

In the second inequality we used the fact that $|y - f(x-\theta)| \leq |y - f(x)| + |f|_\infty$ for all θ and that $|\mathsf{L}'(y)| \leq |\mathsf{L}''|_\infty |y|$ for all y since $\mathsf{L}'(0) = 0$. In either case, $\mathbb{E}[\overline{m}(X,Y)^2] < \infty$, so that the second condition of [VdV98, Th 5.23] is satisfied.

The third condition in that theorem is that $M(\theta)$ admits a Taylor expansion of order two at θ^* with nonzero second order term. Our assumptions imply that we can differentiate once inside the expectation defining M to obtain its first derivative. To see this, note that $\theta \mapsto m_\theta(x,y)$ has first derivative given by (2.14), which is dominated by $|f'|_\infty |\mathsf{L}'|_\infty$ if L is Lipschitz, and by $|f'|_\infty |\mathsf{L}''|_\infty (|y - f(x)| + 2|f|_\infty)$ otherwise, which is integrable in either case. Hence, M is differentiable, with

$$\begin{aligned}
M'(\theta) &= \mathbb{E}[\dot{m}_\theta(X,Y)] \\
&= \mathbb{E}[f'(X-\theta)\mathsf{L}'(Y - f(X-\theta))] \\
&= \int\int f'(x-\theta)\mathsf{L}'(f(x) - f(x-\theta) + z)\phi(z)\lambda(x)\mathrm{d}z\mathrm{d}x \\
&= \int\int f'(x)\mathsf{L}'(f(x+\theta) - f(x) + z)\phi(z)\lambda(x+\theta)\mathrm{d}z\mathrm{d}x.
\end{aligned}$$

Thus, by applying a simple change of variables we have transferred θ away from f' at the cost of burdening λ. But our assumptions are exactly what we need to differentiate inside the integral, since the integrand is differentiable with derivative

$$f'(x)\left\{f'(x+\theta)\mathsf{L}''(f(x+\theta) - f(x) + z)\lambda(x+\theta) + \mathsf{L}'(f(x+\theta) - f(x) + z)\lambda'(x+\theta)\right\}\phi(z), \tag{2.22}$$

which is dominated by

$$|f'|_\infty \left\{ |f'|_\infty |\mathsf{L}''|_\infty |\lambda|_\infty + |\mathsf{L}'|_\infty |\lambda'|_\infty \right\} \phi(z), \tag{2.23}$$

if L is Lipschitz, and otherwise by

$$|f'|_\infty \left\{ |f'|_\infty |\mathsf{L}''|_\infty |\lambda|_\infty + |\mathsf{L}''|_\infty (|z| + 2|f|_\infty) |\lambda'|_\infty \right\} \phi(z), \tag{2.24}$$

and this is integrable in either case. (Recall that λ is compactly supported.) Hence, M is indeed twice differentiable, with

$$\begin{aligned} M''(0) &= \int \int f'(x) \left\{ f'(x) \mathsf{L}''(z) \lambda(x) + \mathsf{L}'(z) \lambda'(x) \right\} \phi(z) \mathrm{d}z \mathrm{d}x \\ &= \int \int f'(x)^2 \mathsf{L}''(z) \lambda(x) \phi(z) \mathrm{d}z \mathrm{d}x \\ &= \mathbb{E}[f'(X)^2] \mathbb{E}[\mathsf{L}''(Z)], \end{aligned}$$

because $\int_{-\infty}^{\infty} \mathsf{L}'(z) \phi(z) \mathrm{d}z = 0$ due to the fact that L' is odd and ϕ is even.

The last condition in [VdV98, Th 5.23] is that $\hat{\theta}_n$ be consistent, which we have already established in Theorem 1.

Therefore, [VdV98, Th 5.23] applies, and implies that $\sqrt{n}(\hat{\theta}_n - \theta^*)$ is asymptotically normal with mean 0 and variance

$$\frac{\mathbb{E}[\dot{m}_0(X,Y)^2]}{M''(0)^2}, \tag{2.25}$$

the latter reducing to (2.13) after some simplifications. □

Theorem 2 applies to the squared error, the Huber, and the Tukey losses with no additional conditions other than the basic assumptions. For squared error loss, the factor $C_{\phi,\mathsf{L}}$ in the

asymptotic variance is given by

$$C_{\text{squared}} := \mathbb{E}[Z^2] = \text{noise variance}. \tag{2.26}$$

For the Huber loss (2.5),

$$C_{\text{huber}} := \frac{\mathbb{E}[Z^2 \wedge c^2]}{\mathbb{P}(|Z| \le c)^2}. \tag{2.27}$$

Theorem 2 does not apply to the absolute-value loss. But very similar arguments can be used to obtain a normal limit distribution for that loss.

Theorem 3. *In the context of Theorem 2, assume that* L *is the absolute-value loss and that* ϕ *is continuous and strictly positive at the origin. Then the same conclusion holds with*

$$C_{\phi,\mathsf{L}} = \frac{1}{4\phi(0)^2}. \tag{2.28}$$

Proof. Assume $\theta^* = 0$ without loss of generality.

In [VdV98, Th 5.23], the first condition is that $\theta \mapsto m_\theta(x,y)$ be differentiable at 0 for almost every (x,y), which is clearly the case here, as it is differentiable with derivative

$$\dot{m}_\theta(x,y) = f'(x-\theta)\,\operatorname{sign}(y - f(x-\theta)). \tag{2.29}$$

We also have

$$\begin{aligned} |m_{\theta_1}(x,y) - m_{\theta_2}(x,y)| &= \big||y - f(x-\theta_1)| - |y - f(x-\theta_2)|\big| && (2.30)\\ &\le |f(x-\theta_1) - f(x-\theta_2)| && (2.31)\\ &\le |f'|_\infty\, |\theta_1 - \theta_2|. && (2.32) \end{aligned}$$

Hence, the second condition of [VdV98, Th 5.23] is satisfied.

Our assumptions imply that we can differentiate once inside the expectation defining M

to obtain its first derivative

$$\begin{aligned} M'(\theta) &= \mathbb{E}[\dot{m}_\theta(X,Y)] \\ &= \mathbb{E}[f'(X-\theta)\,\mathrm{sign}(Y-f(X-\theta))] \\ &= \int\int f'(x-\theta)\,\mathrm{sign}(f(x)-f(x-\theta)+z)\phi(z)\lambda(x)\mathrm{d}z\mathrm{d}x \\ &= \int\int f'(x)\,\mathrm{sign}(f(x+\theta)-f(x)+z)\phi(z)\lambda(x+\theta)\mathrm{d}z\mathrm{d}x \\ &= \int f'(x)\big(1-2\Phi(f(x)-f(x+\theta))\big)\lambda(x+\theta)\mathrm{d}x, \end{aligned}$$

where $\Phi(z) := \int_{-\infty}^{z}\phi(u)\mathrm{d}u$. The integrand is differentiable with derivative

$$f'(x)\big\{2f'(x+\theta)\phi(f(x)-f(x+\theta))\lambda(x+\theta)+\big(1-2\Phi(f(x)-f(x+\theta))\big)\lambda'(x+\theta)\big\} \tag{2.33}$$

which is bounded and therefore dominated as θ varies. The remaining arguments are as in the proof of Theorem 2. □

Remark 4. The expression for the asymptotic variance for the absolute-value loss given could have been anticipated based on the corresponding expression for the Huber loss given in (2.27). Indeed, as $c \to 0$, the Huber loss (2.5) converges pointwise to the absolute-value loss, so that we could have speculated that the same would be true for the asymptotic variances. It turns out that this prediction would have been correct, as it is indeed the case that

$$\mathbb{E}[Z^2 \wedge c^2]/\mathbb{P}(|Z| \le c)^2 \xrightarrow{c\to 0} 1/4\phi(0)^2 \tag{2.34}$$

when ϕ may be taken continuous at the origin. It is also the case that the Huber loss converges pointwise to the squared error loss when $c \to \infty$ instead, and that

$$\mathbb{E}[Z^2 \wedge c^2]/\mathbb{P}(|Z| \le c)^2 \xrightarrow{c\to\infty} \mathbb{E}[Z^2]. \tag{2.35}$$

Remark 5. The constant $C_{\phi,\mathsf{L}}$ in Theorem 2 and Theorem 3 is the asymptotic variance in the classical location model where $Y_i = \theta^* + Z_i$. The only way in which the model we assume (2.1) is different asymptotically is in the denominator in (2.13).

2.3.2 Local asymptotic minimaxity

Under the same smoothness condition on the template as in Section 2.3.1, namely, assuming that f has a bounded derivative, and under some smoothness assumption on the noise density (and not the design density this time), the statistical model is smooth in the sense of being *quadratic mean differentiable (QMD)*. Indeed, under θ, the joint distribution of (X,Y) has density

$$p_\theta(x,y) := \lambda(x)\phi(y - f(x-\theta)). \tag{2.36}$$

The function $\theta \mapsto p_\theta(x,y)$ is differentiable when f and ϕ are. The *information* at θ is defined as

$$\begin{aligned} I_\theta &:= \int\int \frac{\dot{p}_\theta(x,y)^2}{p_\theta(x,y)} \mathrm{d}y\mathrm{d}x && (2.37)\\ &= \int\int \frac{\lambda(x) f'(x-\theta)^2 \phi'(y - f(x-\theta))^2}{\phi(y - f(x-\theta))} \mathrm{d}y\mathrm{d}x && (2.38)\\ &= \int \lambda(x) f'(x-\theta)^2 \mathrm{d}x \times \int \frac{\phi'(y)^2}{\phi(y)} \mathrm{d}y, && (2.39) \end{aligned}$$

and when it is finite and continuous, as it is the case here, the model is QMD [LR05, Th 12.2.1].

When a model is QMD then, under additional mild assumptions, the MLE behaves in a standard way: *If $\hat{\theta}_n$ denotes the MLE, then $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal with zero mean and variance $1/I_\theta$ under θ.* These additional assumptions are fulfilled, for example, when $(x,y) \mapsto \sup_\theta \dot{p}_\theta(x,y)^2/p_\theta(x,y)$ is square integrable [VdV98, Th 5.39]. If the M-estimator is the MLE, meaning when $\phi(y) \propto \exp(-\mathsf{L}(y))$, this is the case under the conditions of Theorem 2.

It turns out this behavior is best possible asymptotically as we describe next. In the present

context, we say that an estimator $\hat{\theta}_n$ is *locally asymptotically minimax (LAM)* at some θ_0 if

$$\lim_{\delta\to 0}\liminf_{n\to\infty}\sup_{|\theta-\theta_0|<\delta}\mathbb{E}_\theta\left[\sqrt{n}|\hat{\theta}_n-\theta|\right] \tag{2.40}$$

achieves the minimum possible value among all estimators. Here $\mathbb{E}_\theta$ denotes the expectation with respect to p_θ. We say that an estimator is LAM if it is LAM at every θ_0. And, indeed, the MLE is LAM under the same conditions, meaning those of [VdV98, Th 5.39], which again hold in the context of Theorem 2. Hence, we may conclude the following.

Theorem 4. *Under the conditions of Theorem 2, and assuming in addition that* $\phi(y)\propto\exp(-\mathsf{L}(y))$*, the M-estimator is LAM.*

2.3.3 Relative efficiency

Although the Huber estimator and the least absolute-value estimator can be motivated as maximum likelihood estimators in their own right, they are often considered as robust compromises to the otherwise preferred least squares estimator. The rationale behind this is in general flimsy, as it amounts to assuming that the noise distribution is Gaussian, as in that case the least squares estimator is the MLE and thus LAM. In signal processing applications, however, the normal assumption can be justified. In any case, this is the perspective we adopt.

We assume therefore that the noise distribution is Gaussian, making the least squares estimator the gold standard for asymptotic comparisons. In that context, if $\hat{\theta}_n$ is another estimator such that $\sqrt{n}(\hat{\theta}_n-\theta)$ is asymptotically normal with mean 0 and variance τ^2 under θ, then its relative efficiency with respect to the MLE is

$$\frac{\tau^2}{\tau^2_{\text{mle}}}, \tag{2.41}$$

and gives, asymptotically, how many more samples this estimator requires in order to achieve the

same precision as the MLE in terms of the width of asymptotically-valid confidence intervals at an arbitrary level of confidence. In our context, τ_{mle} is obtained from (2.13) with (2.26).

Using the expression (2.27), the relative efficiency of the Huber estimator with parameter c is

$$\frac{\mathbb{E}[Z^2 \wedge c^2]}{\mathbb{E}[Z^2]\,\mathbb{P}(|Z| \le c)^2}. \tag{2.42}$$

As anticipated, this is greater than 1 for all values of c and converges to 1 as $c \to \infty$ (as the Huber loss converges to the squared error loss). Using the expression (2.28), the relative efficiency of the least absolute-value estimator is

$$\frac{1}{4\phi(0)^2\,\mathbb{E}[Z^2]}. \tag{2.43}$$

In line with Remark 5, in both cases we recover the relative efficiency for the problem of estimating a normal mean.

2.3.4 Finite-sample minimaxity

When the model is smooth, the M-estimator is locally asymptotically minimax when it is the MLE, as we saw in Section 2.3.2, and more generally achieves the optimal $\sqrt{n}$ rate of convergence only losing in the leading constant, as detailed in Section 2.3.3. What about in finite samples?

In the present context, we define the risk of an estimator $\hat{\theta}_n$ at θ as

$$\mathsf{risk}(\hat{\theta}_n, \theta) := \mathbb{E}_\theta\left[|\hat{\theta}_n - \theta|\right]. \tag{2.44}$$

The *minimax risk* is then defined as

$$R_n^* := \inf_{\hat{\theta}_n} \sup_{\theta} \mathsf{risk}(\hat{\theta}_n, \theta), \tag{2.45}$$

where the infimum is over all estimators taking in a sample of size n.

Lemma 1. *Suppose the basic assumptions are in place. Assume, in addition, that f is Lipschitz, that λ is Lipschitz on an open set containing the support of $f(\cdot - \theta^*)$, and that $t \mapsto B(t) := \int \phi(y) \log \phi(y-t) \mathrm{d}y$ has a derivative which is Lipschitz near the origin. Then the minimax risk satisfies $\sqrt{n} R_n^* \gtrsim 1$.*

Proof. Assume without loss of generality that $\theta^* = 0$. Recall the notation used for the joint density of (X, Y) set in (2.36). By [Tsy09, Th 2.1 and Th 2.2], it suffices to prove that $\mathsf{KL}(p_\theta, p_0) \lesssim 1/n$ when $\sqrt{n}|\theta| \lesssim 1$. Here KL denotes the Kullback–Leibler divergence. We have

$$\mathsf{KL}(p_\theta, p_0) = A(\theta) - A(0), \tag{2.46}$$

where

$$A(\theta) := \mathbb{E}[-\log p_\theta(X, Y)] \tag{2.47}$$

$$= -\int B(f(x-\theta) - f(x))\lambda(x)\mathrm{d}x. \tag{2.48}$$

Note that $|f(x-\theta) - f(x)| \le |f'|_\infty |\theta|$ and that θ is taken small. Hence, since B is continuously differentiably in a neighborhood of the origin and f is almost surely differentiable, $\theta \mapsto B(f(x-\theta) - f(x))\lambda(x)$ is almost surely differentiable near the origin. Its derivative is

$$f'(x-\theta) B'(f(x-\theta) - f(x))\lambda(x), \tag{2.49}$$

which is bounded since f' and λ are. Also, B' is continuous and therefore locally bounded. Hence,

by dominated convergence, A is differentiable with

$$A'(\theta) = \int f'(x-\theta)B'(f(x-\theta)-f(x))\lambda(x)\mathrm{d}x \tag{2.50}$$

$$= \int f'(x)B'(f(x)-f(x+\theta))\lambda(x+\theta)\mathrm{d}x, \tag{2.51}$$

after a change of variable. We also have that $\theta \mapsto f'(x)B'(f(x)-f(x+\theta))\lambda(x+\theta)$ is almost surely differentiable near the origin, with derivative

$$f'(x)\big\{-f'(x+\theta)B''(f(x)-f(x+\theta))\lambda(x+\theta)+B'(f(x)-f(x+\theta))\lambda'(x+\theta)\big\}, \tag{2.52}$$

which is bounded for similar reasons. Hence, by dominated convergence, A is twice differentiable with bounded second derivative. Note that $A'(0) = B'(0)\int f'(x)\lambda(x)\mathrm{d}x$ with $B'(0) = 0$ since B is differentiable and attains its maximum at 0, the latter because $B(0)-B(t) = \int \phi(y)\log(\phi(y)/\phi(y-t))\mathrm{d}y$ is the Kullback–Leibler divergence from $\phi(\cdot - t)$ to ϕ. We thus obtain

$$A(\theta)-A(0) \le \tfrac{1}{2}|A''|_\infty \theta^2. \tag{2.53}$$

This in turn implies that $\mathsf{KL}(p_\theta, p_0) \le C\theta^2$ for some $C > 0$. And $C\theta^2 \le 1/n$ when $\sqrt{n}|\theta| \le 1/\sqrt{C}$. □

Hence, when the model is smooth, the minimax rate of convergence is also $\sqrt{n}$. And this rate is achieved by the M-estimator under essentially the same conditions.

Corollary 1 (Minimaxity)**.** *Assume that the conditions of Theorem 2 or Theorem 3 hold, as well as those of Lemma 1. Then the M-estimator achieves the minimax convergence rate.*

Proof. It turns out that under the conditions of Theorem 2 or Theorem 3, the M-estimator satisfies

[VdV98, Cor 5.53][1]

$$\mathbb{E}_\theta[\sqrt{n}|\hat{\theta}_n - \theta|] = O(1). \tag{2.54}$$

(Note that this is *not* a consequence of the normal limit established in Theorem 2 or Theorem 3.) In our case, we can bound the expectation in (2.54) uniformly over θ. This follows trivially from the arguments provided in the proof of [VdV98, Cor 5.53] and further details are omitted. Hence, $\sup_\theta \mathsf{risk}(\hat{\theta}_n, \theta) = O(1/\sqrt{n})$, and in particular, $\hat{\theta}_n$ achieves the minimax convergence rate established in Lemma 1. □

2.4 Non-smooth setting

When f is not Lipschitz, the situation can be nonstandard. We saw in Theorem 1 that the M-estimator remains consistent under the basic assumptions, but it turns out that it can have a consistency rate other than $\sqrt{n}$. Because nonstandard, the analysis is a little bit more involved, yet as it turns out all the tools we need are again available in [VdV98]. The model is no longer smooth and we do not go into questions of local asymptotic minimaxity or relative efficiency. However, it turns out that the M-estimator remains rate-minimax.

Although other situations may be of interest, for the sake of concreteness we will consider two cases (and some sub-cases). In the first case, we take f to be α-Holder for some $0 < \alpha \leq 1$, meaning

$$H := \sup_{x_1, x_2} \frac{|f(x_1) - f(x_2)|}{|x_1 - x_2|^\alpha} < \infty. \tag{2.55}$$

When this holds for $\alpha = 1$ the function is simply Lipschitz. In the second case, we take f to be piecewise α-Holder. We will let $\mathcal{D}$ denote the discontinuity set of f which, remember, we are assuming is finite.

[1] The statement of [VdV98, Cor 5.53] says that $\sqrt{n}(\hat{\theta}_n - \theta)$ is bounded in probability under p_θ, but this is done via proving that the conditions of Th 5.52 are fulfilled, and although Th 5.52 provides a bound in probability, a simple modification of the arguments underlying this result yield a bound in expectation under the same exact conditions. That bound in expectation gives (2.54) in the present context.

Everywhere, the basic assumptions are in place.

2.4.1 Preliminaries

We start with some preliminary results from which everything else follows. The following function will play an important role

$$\Delta(\theta_1, \theta_2) = \int (f(x-\theta_1) - f(x-\theta_2))^2 \lambda(x) \mathrm{d}x. \tag{2.56}$$

We will also use the following notation

$$\tilde{m}_\theta(x,z) = \mathsf{L}(z + f(x-\theta^*) - f(x-\theta)), \tag{2.57}$$

which is useful because $m_\theta(x,y) = \tilde{m}_\theta(x, y - f(x-\theta^*))$.

The first result is an expansion of the function M of (2.11).

Lemma 2. *If* L *has a bounded and almost everywhere continuous second derivative, then*

$$M(\theta) = M(\theta^*) + A(\theta)\Delta(\theta, \theta^*), \tag{2.58}$$

where A *is bounded and satisfies* $A(\theta) \to C_0 := \frac{1}{2}\int L''\phi$ *as* $\theta \to \theta^*$. *If* L *is the absolute-value loss, and* ϕ *is locally bounded and continuous at 0, then this continues to holds, except that* $C_0 := \phi(0)$.

Proof. Assume $\theta^* = 0$ without loss of generality and let $\Delta(\theta)$ be short for $\Delta(\theta, \theta^*)$. We start with the first situation. Motivated by the fact that $f(x) - f(x-\theta)$ tends to be small when θ is small, we derive

$$\tilde{m}_\theta(x,z) = \mathsf{L}(z + f(x) - f(x-\theta)) \tag{2.59}$$

$$= \mathsf{L}(z) + \mathsf{L}'(z)(f(x) - f(x-\theta)) + \tfrac{1}{2}\mathsf{L}''(z + \zeta(z,x,\theta))(f(x) - f(x-\theta))^2, \tag{2.60}$$

for some $|\zeta(z,x,\theta)| \le |f(x) - f(x-\theta)|$. Integrating over x and z, we obtain

$$\begin{aligned} M(\theta) &= M(0) + \int \int_{\mathbb{R}} \tfrac{1}{2} \mathsf{L}''(z + \zeta(z,x,\theta))(f(x) - f(x-\theta))^2 \phi(z) \lambda(x) \mathrm{d}z \mathrm{d}x \\ &= M(0) + C_0 \Delta(\theta) \pm \int \int_{\mathbb{R}} \gamma(z, |f(x) - f(x-\theta)|)(f(x) - f(x-\theta))^2 \phi(z) \lambda(x) \mathrm{d}z \mathrm{d}x, \end{aligned}$$

where we have used the fact that $\int_{\mathbb{R}} \mathsf{L}'(z)\phi(z)\mathrm{d}z = 0$, and where we used the notation

$$\gamma(z,a) := \sup_{|b| \le a} |\mathsf{L}''(z+b) - \mathsf{L}''(z)|. \tag{2.61}$$

By our assumptions on L'', $\gamma(z,a) \to 0$ when $a \to 0$ for almost every z. In addition, γ is uniformly bounded. Hence, by dominated convergence, the last integral is $o(\Delta(\theta))$ as $\theta \to 0$.

We now turn to the situation where L is the absolute-value loss. We first note that

$$\int_{\mathbb{R}} |z + a| \phi(z) \mathrm{d}z = 2\Phi_2(z) + a, \tag{2.62}$$

where

$$\Phi_2(z) := \int_{-\infty}^{z} \Phi(u) \mathrm{d}u, \quad \Phi(z) := \int_{-\infty}^{z} \phi(u) \mathrm{d}u. \tag{2.63}$$

Note that $\Phi_2' = \Phi$ and $\Phi_2'' = \phi$. Using this, we derive

$$\begin{aligned} M(\theta) &= M(0) + \int \left\{ 2\left[\Phi_2(f(x) - f(x-\theta)) - \Phi_2(0)\right] + f(x) - f(x-\theta) \right\} \mathrm{d}x \\ &= M(0) + C_0 \Delta(\theta) + \int \phi(\zeta(x,\theta))(f(x) - f(x-\theta))^2 \lambda(x) \mathrm{d}x, \end{aligned}$$

for some $|\zeta(x,\theta)| \le |f(x) - f(x-\theta)|$. Note that $|\zeta(x,\theta)| \le 2|f|_\infty$, so that $\phi(\zeta(x,\theta))$ is bounded. Moreover, by the fact that ϕ is continuous at 0, and that f is continuous almost everywhere, it holds that $\phi(\zeta(x,\theta)) \to 0$ as $\theta \to 0$ for almost all x. We can thus apply dominated convergence to find that the last integral is $o(\Delta(\theta))$ as $\theta \to 0$. □

Another notion that will be important is that of an envelope. We say that a function V is an envelope for a function class $\mathcal{M}$ if any function in $\mathcal{M}$ is, in absolute value, pointwise bounded by V, or in formula, $|g(x)| \le V(x)$ for all x and all $g \in \mathcal{M}$. Below,

$$\mathcal{M}_\delta := \left\{ m_\theta - m_{\theta^*} : |\theta - \theta^*| \le \delta \right\}, \tag{2.64}$$

and

$$\mathcal{D}(\delta) := \{ x = d + t : d \in \mathcal{D}, |t| \le \delta \}, \tag{2.65}$$

where δ is thought of as being small.

Lemma 3. *Assume either that* L *has a Lipschitz derivative and that the noise distribution has finite second moment; or that* L *itself is Lipschitz. If* f *is* α*-Holder, then* $\mathcal{M}_\delta$ *admits an envelope of the form* $V(x,y) := \overline{m}(z)\delta^\alpha$ *where* $z := y - f(x - \theta^*)$ *and* $\int \overline{m}^2 \phi < \infty$. *If* f *is only piecewise* α*-Holder, then the same is true but with* $V(x,y) := \overline{m}(z)(\delta^\alpha + \mathbb{I}\{x \in \mathcal{D}(\delta)\})$.

Proof. Assume $\theta^* = 0$ without loss of generality, so that θ below satisfies $|\theta| \le \delta$. For x, y, we let $z = y - f(x)$.

First, assume that f *is* α*-Holder* and let H be defined as in (2.55). If the loss has a Lipschitz derivative, we have

$$\begin{aligned} |m_\theta(x,y) - m_0(x,y)| &= |\mathsf{L}(z + f(x) - f(x-\theta)) - \mathsf{L}(z)| \\ &\le |\mathsf{L}'(z+\zeta)| \times |f(x) - f(x-\theta)| \\ &\le |\mathsf{L}''|_\infty (|z| + 2|f|_\infty) \times H|\theta|^\alpha, \end{aligned}$$

for some $|\zeta| \le |f(x) - f(x-\theta)| \le 2|f|_\infty$. We then conclude with $|\theta| \le \delta$. If the loss is Lipshitz, we

have

$$|m_\theta(x,y) - m_0(x,y)| \le |L'|_\infty |f(x) - f(x-\theta)|$$
$$\le |L'|_\infty H|\theta|^\alpha,$$

and we conclude in the same way.

Next, assume that f is piecewise α-Holder. In this situation, let H be defined as in (2.55), except away from discontinuities. Based on what we just did, it suffices to show that

$$|f(x) - f(x-\theta)| \le C_1(\delta^\alpha + \mathbb{I}\{x \in \mathcal{D}(\delta)\}), \tag{2.66}$$

when $|\theta| \le \delta$ for some constant C_1. Indeed, either there are no discontinuity point between x and $x-\theta$, in which case $|f(x) - f(x-\theta)| \le H|\theta|^\alpha$; or there is a discontinuity point, say d, between x and $x-\theta$, so that $|x-d| \le |\theta| \le \delta$, implying that $|f(x) - f(x-\theta)| \le 2|f|_\infty \mathbb{I}\{x \in \mathcal{D}(\delta)\}$. □

The next result is a complexity bound for $\mathcal{M}_\delta$. The complexity is in terms of bracketing numbers [VdV98, Sec 19.2]. Two functions g_1, g_2 such that $g_1 \le g_2$ pointwise define a bracket made of all functions g such that $g_1 \le g \le g_2$. It is said to be an ε-bracket with respect to $L^2(\mu)$, for a positive measure μ, if $\int (g_2 - g_1)^2 d\mu \le \varepsilon^2$. Given a class of functions $\mathcal{M}$, its ε-bracketing number with respect to $L^2(\mu)$ is the minimum number of ε-brackets needed to cover $\mathcal{M}$ (meaning to include any function in that class). In our context, the measure is the underlying sample distribution, meaning p_{θ^*} in the notation introduced in (2.36). We let $N_\delta(\varepsilon)$ denote the ε-bracketing number of $\mathcal{M}_\delta$ with respect to $L^2(p_{\theta^*})$.

Lemma 4. *There is a constant $C > 0$ such that the following holds. Assume the loss is as in Lemma 3. If f is α-Holder, then $N_\delta(\varepsilon) \le C\delta\varepsilon^{-1/\alpha}$. If f is piecewise α-Holder, and λ is bounded, then $N_\delta(\varepsilon) \le C\delta\varepsilon^{-1/(\alpha \wedge 1/2)}$.*

Proof. Assume $\theta^* = 0$ without loss of generality, so that any value of the parameter below is in

$[-\delta, \delta]$. Let $\theta_j = j\delta/k$ for $j = -k, \ldots, k$. For x, y, we let $z = y - f(x)$ below. We rely on the proof of Lemma 3.

First, assume that f is α-Holder. We have

$$|m_\theta(x,y) - m_{\theta_0}(x,y)| \le \overline{m}(z)|\theta - \theta_0|^\alpha,$$

where $\overline{m}(z)$ is square integrable. For a given θ, let j be such that $|\theta - \theta_j| \le \delta/k$. Then

$$|m_\theta(x,y) - m_{\theta_j}(x,y)| \le \overline{m}(z)|\theta - \theta_j|^\alpha \le \overline{m}(z)(\delta/k)^\alpha,$$

implying that

$$A_j(x,y) \le m_\theta(x,y) \le B_j(x,y), \tag{2.67}$$

where

$$A_j(x,y) := m_{\theta_j}(x,y) - \overline{m}(z)(\delta/k)^\alpha$$
$$B_j(x,y) := m_{\theta_j}(x,y) + \overline{m}(z)(\delta/k)^\alpha.$$

Note that $B_j - A_j$ has L^2 norm of order $(\delta/k)^\alpha$ since $\overline{m}$ is square integrable. Choosing $k \ge C_2\delta/\varepsilon^{1/\alpha}$ for a large enough C_2 makes it an ε-bracket. And these k brackets together cover $\mathcal{M}_\delta$.

Next, assume that f is piecewise α-Holder. Assume for expediency that f has a single discontinuity, and let d denote the location of that discontinuity. Following the corresponding arguments in Lemma 3, we can prove that

$$|m_\theta(x,y) - m_{\theta_0}(x,y)| \le \overline{m}(z)(|\theta - \theta_0|^\alpha + \mathbb{I}\{\theta_0 \le x - d \le \theta\}),$$

where $\overline{m}(z)$ is square integrable. For a given θ, let j be such that $\theta_j \le \theta \le \theta_{j+1}$. Then

$$A_j(x,y) \le m_\theta(x,y) \le B_j(x,y), \tag{2.68}$$

where

$$\begin{aligned} A_j(x,y) &:= m_{\theta_j}(x,y) - \overline{m}(z)((\delta/k)^\alpha + \mathbb{I}\{\theta_j \le x - d \le \theta_{j+1}\}) \\ B_j(x,y) &:= m_{\theta_j}(x,y) + \overline{m}(z)((\delta/k)^\alpha + \mathbb{I}\{\theta_j \le x - d \le \theta_{j+1}\}). \end{aligned}$$

The difference satisfies

$$\begin{aligned} &\int (B_j(x,y) - A_j(x,y))^2 \phi(y - f(x))\lambda(x)\mathrm{d}y\mathrm{d}x \\ &\le 4\int \overline{m}(z)^2\phi(z)\mathrm{d}z \times 2\int ((\delta/k)^{2\alpha} + \mathbb{I}\{\theta_j \le x - d \le \theta_{j+1}\})\lambda(x)\mathrm{d}x \\ &\lesssim (\delta/k)^{2\alpha} + |\lambda|_\infty(\delta/k), \end{aligned}$$

and so has L^2 norm of order $(\delta/k)^\alpha + (\delta/k)^{1/2} \asymp (\delta/k)^{\alpha\wedge 1/2}$, uniformly in j. Therefore, choosing $k \ge C_3\delta/\varepsilon^{1/(\alpha\wedge 1/2)}$ for a large enough C_3 ensures that $B_j - A_j$ has L^2 norm bounded by ε. And the k brackets that these pairs of functions define as j ranges through $\{-k,\dots,k\}$, together, cover the space $\mathcal{M}_\delta$. □

From here proceed in reverse order compared to Section 2.3: We first study the rate of convergence and then discuss the limit distribution.

2.4.2 Rate of convergence

The bracketing integral of a function class $\mathcal{M}$ with respect to $L^2(\mu)$ is defined as the integral of the square root of the logarithm of the corresponding ε-bracketing number as a function

of ε. In particular, we introduce the bracketing integral of $\mathcal{M}_\delta$, denoted

$$J_\delta(t) := \int_0^t \sqrt{\log N_\delta(\varepsilon)}\,d\varepsilon. \tag{2.69}$$

A simple adaptation of the arguments underlying [VdV98, Th 5.52], in combination with [VdV98, Cor 19.35], gives the following result.

Lemma 5. *Suppose there are constants $a > b \geq 0$ and $C > 0$ such that, for $\delta > 0$ small enough,*

$$\inf_{|\theta-\theta^*|\geq\delta} M(\theta) - M(\theta^*) \geq \delta^a/C, \tag{2.70}$$

and

$$J_\delta(\infty) \leq C\delta^b. \tag{2.71}$$

Then the M-estimator satisfies

$$\mathbb{E}\left[n^{\frac{1}{2(a-b)}}|\hat{\theta}_n - \theta^*|\right] = O(1). \tag{2.72}$$

With the preceding lemmas, we are able to establish an upper bound on the rate of convergence of the M-estimator.

Theorem 5. *Suppose that either L has a bounded and almost everywhere continuous second derivative; or is the absolute-value loss and ϕ is locally bounded and continuous at 0.*

- *Suppose f is α-Holder and*

$$\sup_{|\theta-\theta^*|\leq\delta} \Delta(\theta,\theta^*) \asymp \delta^{2\alpha}. \tag{2.73}$$

  Then the M-estimator is r_n-consistent with $r_n = n^{1/2\alpha}$.

- *Suppose f is discontinuous and piecewise α-Holder with $\alpha \geq 1/2$, and λ is bounded and continuous at the points of discontinuity of $f(\cdot - \theta^*)$, and strictly positive at one or more of*

these locations. Then the M-estimator is r_n-consistent with $r_n = n$.

Proof. Consider the first situation. We apply Lemma 5. By Lemma 2 and (2.73), the condition (2.70) is satisfied with $a = 2\alpha$. Hence, it suffices to show that (2.71) holds with $b = \alpha$. Indeed, by Lemma 3, we have $J_\delta(\infty) = J_\delta(C\delta^\alpha)$ for C large enough, and by Lemma 4, we have

$$J_\delta(C\delta^\alpha) \le \int_0^{C\delta^\alpha} \sqrt{\log(C\delta/\varepsilon^{1/\alpha})}\mathrm{d}\varepsilon \asymp \delta^\alpha. \tag{2.74}$$

We now turn to the second situation. By of Lemma 3, we have $J_\delta(\infty) = J_\delta(C\delta^{1/2})$ for C large enough, and by Lemma 4, we have

$$J_\delta(C\delta^{1/2}) \le \int_0^{C\delta^{1/2}} \sqrt{\log(C\delta/\varepsilon^2)}\mathrm{d}\varepsilon \asymp \delta^{1/2}, \tag{2.75}$$

so that (2.71) holds with $b = 1/2$. It thus suffices to show that (2.70) holds with $a = 1$. Assume for simplicity that f has only one discontinuity, and therefore of the form $f(x) = f_1(x)\mathbb{I}\{x < d\} + f_2(x)\mathbb{I}\{x \ge d\}$ with f_1 and f_2 being α-Holder. For example, consider a situation where $\theta > 0 = \theta^*$. We have

$$\Delta(\theta, \theta^*) = \int_{-\infty}^{d} + \int_{d}^{d+\theta} + \int_{d+\theta}^{\infty} \quad (f(x) - f(x-\theta))^2\lambda(x)\mathrm{d}x. \tag{2.76}$$

(In what follows, remember that λ is compactly supported.) In the 1st and 3rd integral, $|f(x) - f(x-\theta)| \le H|\theta|^\alpha$ since f does not have a discontinuity over the corresponding ranges, $x \in [0, d)$ and $x \in (d+\theta, 1]$, respectively. Hence, these two integrals are of order $|\theta|^{2\alpha}$, which is at most of order $O(\theta)$ since $\alpha \ge 1/2$. For the 2nd or middle integral, we use the fact that $(f(x) - f(x-\theta))\lambda(x) = (f_2(x) - f_1(x-\theta))^2\lambda(x) \to (f_2(d) - f_1(d))^2\lambda(d)$ when $x \in [d, d+\theta]$ and $\theta \to 0$, so that the integral is $\sim |\theta|(f_2(d) - f_1(d))^2\lambda(d)$ by dominated convergence. More generally, if $\mathcal{D}$ denotes the points of discontinuity of f, extending these arguments gives

$$\Delta(\theta, \theta^*) \sim |\theta|D, \quad \text{as } \theta \to 0 = \theta^*, \quad \text{where } D := \sum_{d \in \mathcal{D}} (f(d^+) - f(d^-))^2\lambda(d)\ . \tag{2.77}$$

From this we conclude. □

Of course, we have only obtained an upper bound on the consistency rate. But as we will see in the next subsection, that rate is sharp.

Remark 6. We note that (2.73) is not automatically satisfied even when the function f is exactly α-Holder (and not $(\alpha+\eta)$-Holder for any $\eta > 0$). Indeed, consider the case where f coincides near the origin with $x \mapsto |x|^\alpha$ for some fixed $0 < \alpha < 1$, and is otherwise Lipschitz. Then such an f is exactly α-Holder, and yet, it can be shown that

$$\sup_{|\theta-\theta^*|\leq\delta} \Delta(\theta,\theta^*) \asymp \begin{cases} \delta^2 & \text{if } \alpha > 1/2, \\ \delta^2 \log(1/\delta) & \text{if } \alpha = 1/2, \\ \delta^{2\alpha+1} & \text{if } \alpha < 1/2. \end{cases} \tag{2.78}$$

2.4.3 Minimaxity

We can easily adapt the arguments underlying the information bound stated in Lemma 1 to the settings that interest us in the present section where f is not Lipschitz. We do so and obtain the following.

Lemma 6. *Suppose the basic assumptions are in place. Assume, in addition, that* ϕ *has finite first moment and that B is as in Lemma 1. If* r_n *is such that* $\sup\{\Delta(\theta,\theta^*) : r_n|\theta-\theta^*| \leq 1\} \lesssim 1/n$, *the minimax rate satisfies* $r_n R_n^* \gtrsim 1$.

Proof. Assume without loss of generality that $\theta^* = 0$ and let $\Delta(\theta)$ be short for $\Delta(\theta,\theta^*)$. In the notation introduced in the proof of Lemma 1, it suffices to prove that $A(\theta) - A(0) \lesssim 1/n$ when $r_n|\theta| \leq 1$. We saw in the proof of that lemma that $B'(0) = 0$. This together with the fact that B' is Lipschitz near the origin implies that there is a positive constant C_1 such that $|B(t) - B(0)| \leq C_1 t^2$

when $|t|$ is small enough. We use that to derive

$$A(\theta) - A(0) = \int \left\{B(0) - B(f(x-\theta) - f(x))\right\}\lambda(x)\mathrm{d}x \tag{2.79}$$

$$\leq C_1 \int (f(x-\theta) - f(x))^2 \lambda(x)\mathrm{d}x \tag{2.80}$$

$$= C_1 \Delta(\theta). \tag{2.81}$$

From this we conclude. □

Corollary 2. *In the context of Lemma 6, if f is α-Holder, then $n^{1/2\alpha} R_n^* \gtrsim 1$; if instead f is discontinuous and piecewise α-Holder with $\alpha \geq 1/2$, then $nR_n^* \gtrsim 1$.*

Proof. As we saw in the proof of Theorem 5, if f is α-Holder, $\Delta(\theta, \theta^*) \lesssim |\theta - \theta^*|^{2\alpha}$; while if instead f is discontinuous and piecewise α-Holder with $\alpha \geq 1/2$, $\Delta(\theta, \theta^*) \lesssim |\theta - \theta^*|$. We then apply Lemma 6 to conclude. □

2.4.4 Limit distribution

To establish a limit distribution, we follow the standard strategy which starts by showing that a properly normalized version of the empirical process $\widehat{M}_n$ converges to a Gaussian process, and is followed by an application of the argmax continuous mapping theorem.

The following is a direct consequence of [VdV98, Th 19.28].

Lemma 7. *Consider a set of functions on some Euclidean space and let μ denote a Borel probability measure. Consider a class of functions $\mathcal{W}_{n,T} := \{w_{n,t} : n \geq 1, t \in [-T, T]\}$ satisfying the following:*

(i) The class has a square integrable envelope.

(ii) The limit $g(t) := \lim_n \sqrt{n} \int w_{n,t} \mathrm{d}\mu$ is well-defined for all t.

(iii) There is a function ω_1 *with* $\lim_{t\to 0+}\omega_1(t) = 0$ *and a sequence* $\eta_n \to 0$ *such that*

$$v_n(s,t) := \int (w_{n,s} - w_{n,t})^2 d\mu \le \omega_1(|s-t|) + \eta_n; \tag{2.82}$$

(iv) The limit $v(s,t) := \lim_n v_n(s,t)$ *is well-defined for all* s,t.

(v) The class has bracketing integral $J_{n,T}(\delta) \le \omega_2(\delta)$ *for some function* ω_2 *with* $\lim_{t\to 0+}\omega_2(t) = 0$.

Then $W_{n,t} := \sum_{i=1}^n w_{n,t}$ *converges weakly to* $g(t) + G_t$ *on* $[-T,T]$, *where* G_t *denotes the centered Gaussian process with variance function* v. *Note that* $v(s,t) \le \omega_1(|s-t|)$.

The following is a special case of [VdV98, Cor 5.58].

Lemma 8. *Suppose that a sequence of processes* W_n *defined on the real line converges weakly in the uniform topology on every bounded interval to a process* W *with continuous sample paths each having a unique minimum point* h *(almost surely). If* h_n *minimizes* W_n, *and* (h_n) *is uniformly tight, then* h_n *converges weakly to* h.

We apply these two lemmas to obtain a limit distribution for the M-estimator when the template is α-Holder. For expediency, we only consider smoother losses.

Theorem 6. *Suppose that* L *has a bounded and almost everywhere continuous second derivative. Assume that* f *is* α*-Holder and such that*

$$r^{2\alpha}\Delta(\theta^* + s/r, \theta^* + t/r) \to \Delta_0(s,t), \quad r\to\infty, \tag{2.83}$$

where Δ_0 *is a continuous function such that* $\Delta_0(s,t) \ne 0$ *when* $s \ne t$. *Then* $n^{1/2\alpha}(\hat\theta_n - \theta^*)$ *converges weakly to* $\arg\min_t\{g(t) + G_t\}$ *where* $g(t) := C_0\Delta_0(t,0)$ *with* $C_0 := \frac{1}{2}\int L''\phi$ *and* G_t *is the centered Gaussian process on the real line with variance function* $C_1\Delta_0(s,t)$ *with* $C_1 := \int (\mathsf{L}')^2\phi$.

Proof. As usual, assume that $\theta^* = 0$ without loss of generality, and let $z = y - f(x)$.

We first prove that the process $\sqrt{n}(\widehat{M}_n(t/r_n) - \widehat{M}_n(0))$ — where $r_n = n^{1/2\alpha}$ is the rate of convergence established in Theorem 5 — converges weakly to $g(t) + G_t$. For this it is enough to prove that it converges on every interval of the form $[-T, T]$, and we do so by applying Lemma 7 with $w_{n,t} := \sqrt{n}(m_{t/r_n} - m_0)$. Note that $\mu = p_0$, the distribution of (X, Y) under $\theta = 0$. Lemma 3 and a rescaling argument gives *(i)*. But to be sure, using Lemma 2, we have

$$|w_{n,t}(x,y)| \le \sqrt{n}\left\{|\mathsf{L}'(z)||f(x) - f(x - t/r_n)| + \tfrac{1}{2}|\mathsf{L}''|_\infty (f(x) - f(x - t/r_n))^2\right\} \tag{2.84}$$

$$\le \sqrt{n}\left\{|\mathsf{L}'(z)|H(T/r_n)^\alpha + \tfrac{1}{2}|\mathsf{L}''|_\infty (H(T/r_n)^\alpha)^2\right\} \tag{2.85}$$

$$\le C(\mathsf{L}'(|z|) + 1) =: \overline{w}(x,y). \tag{2.86}$$

We then conclude with the fact that $\overline{w}$ is square integrable by the usual assumptions.

For *(ii)*, using Lemma 2, we have

$$\sqrt{n} \int w_{n,t} \mathrm{d}\mu = \sqrt{n} \times \sqrt{n}(M(t/r_n) - M(0)) \tag{2.87}$$

$$\sim n C_0 \Delta(t/r_n, 0), \quad n \to \infty, \tag{2.88}$$

$$= n C_0 r_n^{-2\alpha} r_n^{2\alpha} \Delta(t/r_n, 0) \tag{2.89}$$

$$\to C_0 \Delta_0(t, 0) = g(t), \quad n \to \infty. \tag{2.90}$$

For *(iii)*, using (2.59), we have for $s,t \in [-T,T]$,

$$\int (w_{n,s} - w_{n,t})^2 \mathrm{d}\mu = n \int (m_{s/r_n} - m_{t/r_n})^2 \mathrm{d}p_0 \tag{2.91}$$

$$\leq n\Big[2C_1\Delta(s/r_n, t/r_n) + |\mathsf{L}''|_\infty^2\{\Delta(s/r_n,0)^2 + \Delta(t/r_n,0)^2\}\Big] \tag{2.92}$$

$$\leq n\Big[2C_1(H|s/r_n - t/r_n|^\alpha)^2 + |\mathsf{L}''|_\infty^2\{(H|s/r_n|^\alpha)^4 + (H|t/r_n|^\alpha)^4\}\Big] \tag{2.93}$$

$$\leq nC\Big(|s-t|^{2\alpha}/r_n^{2\alpha} + (T/r_n)^{4\alpha}\Big) \tag{2.94}$$

$$= C\Big(|s-t|^{2\alpha} + T^{4\alpha}/n\Big), \tag{2.95}$$

using the fact that $r_n^{2\alpha} = n$. We conclude that (2.82) holds with $\omega_1(s,t) := C|s-t|^{2\alpha}$ and $\eta_n := CT^{4\alpha}/n$.

For *(iv)*, we refine these arguments using dominated convergence, to get for any $s,t \in \mathbb{R}$,

$$v_n(s,t) = n \int (m_{s/r_n} - m_{t/r_n})^2 \mathrm{d}p_0 \tag{2.96}$$

$$\sim nC_1\Delta(s/r_n, t/r_n), \quad n \to \infty, \tag{2.97}$$

$$= nC_1 r_n^{-2\alpha} r_n^{2\alpha} \Delta(s/r_n, t/r_n) \tag{2.98}$$

$$\to C_1\Delta_0(s,t) = v(s,t), \quad n \to \infty. \tag{2.99}$$

For *(v)*, we use Lemma 4 and a rescaling argument to bound the bracketing number of this function class by $C(T/r_n)(\varepsilon/\sqrt{n})^{-1/\alpha} = CT\varepsilon^{-1/\alpha}$, so that its bracketing integral at δ is bounded by $\omega_2(\delta) := \int_0^\delta \sqrt{\log(CT/\varepsilon^{-1/\alpha})}\mathrm{d}\varepsilon$.

The fact that $\sqrt{n}(\widehat{M}_n(t/r_n) - \widehat{M}_n(0))$ converges weakly to $g(t) + G_t$ is useful to us because $\hat{\theta}_n = \hat{t}_n/r_n$ where $\hat{t}_n = \arg\max_t \sqrt{n}(\widehat{M}_n(t/r_n) - \widehat{M}_n(0))$. To conclude, we only need to show that Lemma 8 applies. On the one hand, the process $g(t) + G_t$ has continuous sample paths. This is because g is continuous — since $g(t) \propto \Delta_0(t,0)$ and Δ_0 is assumed continuous — and G_t is Gaussian with variance function v satisfying $v(s,t) \leq C|s-t|^{2\alpha}$ and this is enough for G_t to have continuous sample paths according to [Dud67, Th 7.1]. On the other hand, the process has a

unique minimum point since $v(s,t) \neq 0$ when $s \neq t$ — since $v \propto \Delta_0$ and Δ_0 satisfies that property — which is sufficient by [KP90, Lem 2.6]. And we have shown in Theorem 5 that $\hat{t}_n$ is uniformly tight. □

Theorem 6 also applies when $\alpha = 1$, meaning when the template is Lipschitz, and in fact implies Theorem 2 in that case. Thus, we have established that, when the template f is Holder, the M-estimator converges weakly to the minimizer of a Gaussian process which does not depend on the noise distribution except through the constants C_0 and C_1. When the template f is discontinuous, the situation is qualitatively different: the limit process does not have a unique minimum point and the M-estimator does not have a limit distribution, and the limit process is far from 'universal' but rather depends heavily on the noise distribution.

Instead of the more commonly-used argmax theorem (Lemma 8), which takes place in the uniform topology, we will use the following mild variant of [Fer04, Th 3], which takes place in the Skorohod topology.

Lemma 9. *Suppose that a sequence of processes W_n defined on the real line converges weakly in the Skorohod topology on every bounded interval to a process W with sample paths each achieving their minimum somewhere in $[\underline{h}, \overline{h}]$, and only there, with $\underline{h}$ and $\overline{h}$ being continuous random variables. Assume also that W_n and W have right and left limits at every point, and that $\mathbb{P}(W \text{ is continuous at } x) = 1$ for all x. If h_n minimizes W_n, and (h_n) is uniformly tight, then, for every x,*

$$\mathbb{P}(\underline{h} \leq x) \leq \liminf_n \mathbb{P}(h_n \leq x) \leq \limsup_n \mathbb{P}(h_n \leq x) \leq \mathbb{P}(\overline{h} \leq x).$$

It turns out that the limit process is here a marked Poisson process, and with the help of this variant of the argmax continuous mapping theorem, we obtain the following.

Theorem 7. *Suppose that either L has a bounded and almost everywhere continuous second derivative. Assume that f is discontinuous and piecewise α-Holder with $\alpha > 1/2$, and λ is bounded and continuous at the points of discontinuity of $f(\cdot - \theta^*)$, and strictly positive at one*

or more of these locations. Then $n(\hat{\theta}_n - \theta^*)$ *dominates* $\underline{t}$ *and is dominated by* $\bar{t}$ *asymptotically, where* $[\underline{t}, \bar{t}]$ *is the closure of the minimum point set of* $W = \sum_{d \in \mathcal{D}} W_d$*, where* $\mathcal{D}$ *denotes the discontinuity set of* $f(\cdot - \theta^*)$ *and the* W_d*'s are independent with* W_d *being the double-sided marked Poisson process on* $\mathbb{R}$ *with intensity* $\lambda(d)$ *and mark distribution that of* $\mathsf{L}(Z + \delta_d) - \mathsf{L}(Z)$*, where* $\delta_d := f(d^+ - \theta^*) - f(d^- - \theta^*)$.

Remark 7. The values that W takes at its discontinuity points are irrelevant, and in particular it does not need to be taken càdlag as is the norm. In the proof we set things up so that it is, but this is only for convenience.

Proof. As usual, assume that $\theta^* = 0$ without loss of generality. When the meaning of (x, y) is clear from context, we use z as a shorthand for $y - f(x)$. We assume for convenience that f is càglad, the reverse of càdlag, meaning that at every point it is continuous from the left and has a limit from the right. We do is in order for the Poisson processes that follow to be càdlag.

We first prove that the process $W_n(t) := n(\widehat{M}_n(t/n) - \widehat{M}_n(0))$ converges weakly to $W(t)$. Note that n is the rate of convergence established in Theorem 5. Indeed, take x, and also $t > 0$ smaller than the separation between any two discontinuity points of f. Two cases are possible. Either there is $d \in \mathcal{D}$ such that $x - t \le d < x$, in which case we use the fact that

$$\begin{aligned} f(x) - f(x-t) &= f(x) - f(d^+) + \delta_d + f(d^-) - f(x-t) \\ &= \pm H(x-d)^\alpha + \delta_d \pm H(d - (x-t))^\alpha \\ &= \delta_d \pm 2Ht^\alpha, \end{aligned}$$

implying that

$$\mathsf{L}(z + f(x) - f(x-t)) - \mathsf{L}(z) = \mathsf{L}(z + \delta_d) - \mathsf{L}(z) \pm \mathsf{L}'(z + \delta_d) 2Ht^\alpha \pm \tfrac{1}{2} |\mathsf{L}''|_\infty (2Ht^\alpha)^2 \tag{2.100}$$

$$= B_d(z) \pm \overline{m}(z) t^\alpha, \quad B_d(z) := \mathsf{L}(z + \delta_d) - \mathsf{L}(z), \tag{2.101}$$

where $\overline{m} \geq 0$ and $\int \overline{m}^2 \phi < \infty$. Otherwise,

$$f(x) - f(x-t) = \pm H(x-(x-t))^{\alpha} = \pm H t^{\alpha},$$

implying that

$$\mathsf{L}(z + f(x) - f(x-t)) - \mathsf{L}(z) = \mathsf{L}'(z)(f(x) - f(x-t)) \pm \tfrac{1}{2}|\mathsf{L}''|_{\infty}(Ht^{\alpha})^2 \tag{2.102}$$

$$= \mathsf{L}'(z)(f(x) - f(x-t)) \pm \overline{m}(z)t^{2\alpha}, \tag{2.103}$$

for a possibly different non-negative, square integrable function $\overline{m}$. We simply take the pointwise maximum of the two. Hence,

$$\begin{aligned} m_t(x,y) - m_0(x,y) &= \mathsf{L}(z + f(x) - f(x-t)) - \mathsf{L}(z) \\ &= \sum_{d \in \mathcal{D}} \Big\{ (B_d(z) \pm \overline{m}(z)t^{\alpha}) \mathbb{I}\{x \in (d, d+t]\} \\ &\qquad + (\mathsf{L}'(z)(f(x) - f(x-t)) \pm \overline{m}(z)t^{2\alpha}) \mathbb{I}\{x \notin (d, d+t]\} \Big\}, \end{aligned}$$

and therefore,

$$n(\widehat{M}_n(t/n) - \widehat{M}_n(0)) = \sum_{d \in \mathcal{D}} W_{n,d}(t) \pm (t/n)^{\alpha} \sum_{i=1}^{n} \overline{m}(Z_i) \mathbb{I}\{d < X_i \leq d + t/n\} \tag{2.104}$$

$$+ \sum_{d \in \mathcal{D}} A_{n,d}(t) \pm (t/n)^{2\alpha} \sum_{i=1}^{n} \overline{m}(Z_i), \tag{2.105}$$

where

$$W_{n,d}(t) := \sum_{i=1}^{n} \mathbb{I}\{d < X_i \leq d + t/n\} B_d(Z_i) \tag{2.106}$$

and

$$A_{n,d}(t) := \sum_{i=1}^{n} \mathsf{L}'(Z_i)(f(X_i) - f(X_i - t/n)) \mathbb{I}\{X_i \notin (d, d + t/n]\}. \tag{2.107}$$

A Poisson approximation gives that $W_{n,d}$ converges as a process to a marked Poisson process with intensity $\lambda(d)$ on $\mathbb{R}_+$ and mark distribution that of $B_d(Z)$, and these processes are independent of each other in the large-n limit. We elaborate in Lemma 11. So it suffices to show that the other three terms above are $o_P(1)$. The second term has absolute first moment equal to

$$\begin{aligned}
&(t/n)^\alpha n \mathbb{E}[\overline{m}(Z)] \mathbb{P}(d \le X < d + t/n) \\
&\sim (t/n)^\alpha n \mathbb{E}[\overline{m}(Z)] \lambda(d)(t/n) \\
&\asymp t^{\alpha+1}/n^\alpha \to 0.
\end{aligned}$$

For the third term, each $A_{n,d}$ has mean 0 by the fact that $\mathbb{E}[\mathsf{L}'(Z)] = 0$ and $X \perp\!\!\!\perp Z$, and it has second moment equal to

$$\begin{aligned}
&n \mathbb{E}[\mathsf{L}'(Z)^2] \mathbb{E}\left[(f(X) - f(X - t/n))^2 \mathbb{I}\{X \notin (d, d + t/n]\} \right] \\
&\le n \mathbb{E}[\mathsf{L}'(Z)^2] (H(t/n)^\alpha)^2 \\
&\asymp n/n^{2\alpha} \to 0,
\end{aligned}$$

since $\alpha > 1/2$. The fourth term has absolute first moment equal to

$$(t/n)^{2\alpha} n \mathbb{E}[\overline{m}(Z)] \asymp n/n^{2\alpha} \to 0,$$

since $\alpha > 1/2$. In all cases, we may thus apply Markov's inequality to get that these terms all converge to 0 in probability.

We focused on $t > 0$, but the same arguments apply to $t < 0$. Note that, when considering $t < 0$, $B_d^-(z) := \mathsf{L}(z - \delta_d) - \mathsf{L}(z)$ plays the role of $B_d(z)$, but $B_d^-(Z) \sim B_d(Z)$, using the fact that L is even and that Z is symmetric about 0, so that the mark distribution remains that of $B_d(Z)$.

We now apply Lemma 9. By construction, W_n and W both are piecewise continuous and thus have right and left limits at every point. And we proved that $n\hat{\theta}_n$ is uniformly tight

in Theorem 5. So we just have to show that W satisfies the other properties required in the lemma. We note that W is itsef a marked Poisson process with intensity $\sum_{d\in\mathcal{D}}\lambda(d)$ and mark distribution the mixture of the $B_d(Z)$ distributions with mixture weights proportional to the $\lambda(d)$'s. In particular, as is well-known, for every (fixed) x, W is continuous at x with probability 1. Further, the mark distribution has strictly positive mean, this being the case because $\mathbb{E}[B_d(Z)] > 0$ under Assumption 5. It follows from this that $W(t) \to +\infty$ as $t \to \pm\infty$ (almost surely), and given that W is piecewise constant with different values on each interval — since the distribution of $B_d(Z)$ is continuous under our assumptions — its minimum point set is a (bounded) interval defined by two discontinuity points, $\underline{t}$ and $\bar{t}$, which have continuous distributions. □

Remark 8. In change-point settings where the design is fixed, the minimizer of the limit process is often unique, in which case this is the limit of any empirical minimizer. This is the case, for example, in [YA89, Hin70, D$\ddot{9}$1]. In [Kos08, Sec 14.5.1], a function of the form $a\{x \le t\} + b\{x > t\}$ is fitted to the data (or in the language that we have been using, is 'matched' to the noisy signal). The design is random as it is here, and therefore the empirical minimizer is not unique in terms of t. To circumvent this issue, the smallest minimizer (in t) is used as the location estimator, which is then shown to converge to the smallest minimizer of the limit process, which in terms of t is also a compound Poisson process. See also [SS11, LBM09]. This approach does not seem applicable in our context. Indeed, the situation here is different in that f is not necessary piecewise constant, and in particular it is very possible that the empirical minimizer is unique, rendering the selection of the smallest minimizer superfluous and leaving the issue of multiple minima in the limit untouched. We thus opted for the approach proposed by [Fer04].

2.5 More flexible templates

Some situations may call for finding the best match in a family of templates. Assuming a parametric model, the model becomes

$$Y_i = f_{\theta^*}(X_i) + Z_i, \tag{2.108}$$

where the family $\{f_\theta : \theta \in \Theta\}$ is known. The smooth setting can be considered in a very general framework and we do so in Section 2.5.1. The non-smooth setting is, as usual, more delicate, so we content ourselves with considering a location-scale extension of our basic model (2.1) in Section 2.5.2 which seems popular in practice.

2.5.1 Smooth setting

We consider the model (2.108) in a rather general setting where $f_\theta : \mathbb{R}^p \to \mathbb{R}$ and Θ is a bounded subset of a Euclidean space. The same toolset can then be used to derive the asymptotic behavior of the M-estimator, defined as in (2.8) with

$$m_\theta(x, y) := \mathsf{L}(y - f_\theta(x)). \tag{2.109}$$

Throughout, the same basic assumptions are in place with appropriate modifications: In particular, in Assumption 1 we now assume that

$$\text{For any } \theta \neq \theta^*,\ f_\theta(\cdot) \neq f_{\theta^*}(\cdot) \text{ on a set of positive measure under } \lambda. \tag{2.110}$$

Here we consider the smooth case, which again corresponds to a template f which is Lipschitz.

Consistency

In view of Assumption 5, consistency ensues as soon as the function class $\{m_\theta : \theta \in \Theta\}$ is Glivenko–Cantelli, which is the case here since

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \le H\|\theta_1 - \theta_2\| \tag{2.111}$$

for all $\theta_1, \theta_2 \in \Theta$. See [VdV98, Ex 19.7].

Rate of convergence

To obtain a rate of convergence, we rely on Lemma 5 as before. We note that Lemma 2 remains valid, now with $\Delta(\theta, \theta^*) = \int (f_\theta(x) - f_{\theta^*}(x))^2 \lambda(x) \mathrm{d}x$, and by dominated convergence,

$$\Delta(\theta, \theta^*) \sim (\theta - \theta^*)^\top Q (\theta - \theta^*), \quad \theta \to \theta^*, \quad \text{where } Q := \mathbb{E}\left[\dot{f}_{\theta^*}(X) \dot{f}_{\theta^*}(X)^\top\right], \tag{2.112}$$

so that $\Delta(\theta, \theta^*) \asymp \|\theta - \theta^*\|^2$ as long as Q is full rank, which we assume henceforth. It is also the case that $\mathcal{M}_\delta$ has an envelope with square integral bounded by $C\delta^2$ and ε-bracketing number $N_\delta(\varepsilon) \le C\delta\varepsilon^{-d}$. The former is an immediate consequence of (2.111), while the latter is based on [VdV98, Ex 19.7] and a scaling argument. We are thus in a position to apply Lemma 5, with $a = 2$ and $b = 1$, to get that the M-estimator is $\sqrt{n}$-consistent in expectation, meaning that

$$\mathbb{E}\left[\sqrt{n}\|\hat{\theta}_n - \theta^*\|\right] = O(1). \tag{2.113}$$

And Lemma 6 applies (essentially verbatim) to establish this as the minimax rate of convergence under the same condition on ϕ.

Limit distribution

We can also obtain a limit distribution following the arguments underlying Theorem 6. With all these arguments in plain view, it is a simple endeavor to check that everything proceeds in the same way, based on the fact that

$$n\Delta(\theta^* + s/\sqrt{n}, \theta^* + t/\sqrt{n}) \xrightarrow{n\to\infty} \Delta_0(s,t) := (s-t)^\top Q(s-t). \tag{2.114}$$

Following that path, we arrive at the conclusion that

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xRightarrow{n\to\infty} \arg\min_t \{g(t) + G_t\}, \tag{2.115}$$

where $g(t) := C_0\Delta_0(t,0) = C_0 t^\top Qt$ and G_t is the centered Gaussian process with variance function $v(s,t) := C_1\Delta_0(s,t) = C_1(s-t)^\top Q(s-t)$. But this is the classical setting. Indeed, G_t can be represented as $U^\top Q^{1/2}t$ where U is a standard normal random vector. This is true because G_t and $Q^{1/2}U^\top t$, as processes, are both Gaussian with same mean and same variance function. Hence,

$$\arg\min_t \{g(t) + G_t\} \equiv \arg\min_t \{C_0 t^\top Qt + C_1^{1/2} Q^{1/2} U^\top t\} = -\frac{C_1^{1/2}}{2C_0} Q^{-1/2} U, \tag{2.116}$$

which is centered normal with covariance

$$(\mathbb{E}[\mathsf{L}'(Z)^2]/\mathbb{E}[\mathsf{L}''(Z)]^2)\,\mathbb{E}[\dot{f}_{\theta^*}(X)\dot{f}_{\theta^*}(X)^\top]^{-1}. \tag{2.117}$$

We have thus established this as the limit distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$.

The model is also quadratic mean differentiable [LR05, Th 12.2.2], with the information at θ being given by

$$I_\theta := \int \dot{f}_\theta(x)\dot{f}_\theta(x)^\top \lambda(x)\mathrm{d}x \times \int \frac{\phi'(y)^2}{\phi(y)}\mathrm{d}y\,. \tag{2.118}$$

The analog of Theorem 4 applies in that case, meaning that the M-estimator is locally asymptotically minimax when it coincides with the MLE.

Example 1. A special case which might be of interest in the context of matching a template to a signal is when $f_\theta(x) = S_\theta(f(T_\theta(x)))$, where for every θ, $S_\theta : \mathbb{R} \to \mathbb{R}$ and $T_\theta : \mathbb{R}^p \to \mathbb{R}^p$ are both diffeomorphisms, and such that $\theta \mapsto S_\theta(x)$ and $\theta \mapsto T_\theta(x)$ are differentiable for almost every x. As before, f is a known function. In this special case, and assuming without loss of generality that S_{θ^*} and T_{θ^*} are the identity functions for their respective spaces, we have

$$\dot{f}_{\theta^*}(x) = \dot{S}_{\theta^*}(f(x)) + \dot{T}_{\theta^*}(x)^\top \nabla f(x). \tag{2.119}$$

For instance, in the affine family of templates given by $af(B^{-1}(\cdot - b))$, the parameter is $\theta = (a, b, B)$ where $a > 0$, $b \in \mathbb{R}^p$ and $B \in \mathbb{R}^{p \times p}$ invertible, and the transformations are given by $S_\theta(y) = ay$, $T_\theta(x) = B^{-1}(x - b)$. Assuming without loss of generality that $\theta^* = (1, 0, I)$, S_{θ^*} and T_{θ^*} are the identity functions. Then $\dot{f}_{\theta^*}(x) = (f(x), -\nabla f(x), -\nabla f(x)x^\top)$. When the design is on the real line, that is when b and B are real numbers, this becomes $\dot{f}_{\theta^*}(x) = (f(x), -f'(x), -f'(x)x)$, in which case the asymptotic covariance matrix of the M-estimator is, according to (2.117), proportional to the inverse of the integral with respect to λ of the following matrix

$$\dot{f}_{\theta^*}(x)\dot{f}_{\theta^*}(x)^\top = \begin{pmatrix} f(x)^2 & -f(x)f'(x) & -xf(x)f'(x) \\ -f(x)f'(x) & f'(x)^2 & xf'(x)^2 \\ -xf(x)f'(x) & xf'(x)^2 & x^2 f'(x)^2 \end{pmatrix}. \tag{2.120}$$

2.5.2 Non-smooth setting

Unlike the smooth setting, non-smooth situations typically need a more custom treatment. Instead of attempting to do that, which if possible at all is beyond the scope of the present paper, we consider a relatively simple extension of our basic model (2.1) which already exhibits some

interesting features and which, simultaneously, happens to be popular in practice. Specifically we consider (2.108) where, for $\theta = (\beta, \xi, \nu) \in \Theta := \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+^*$, $f_\theta(x) = \beta f((x-\xi)/\nu)$. As before, f is fixed. We are thus considering the following amplitude-location-scale model

$$Y_i = \beta^* f((X_i - \xi^*)/\nu^*) + Z_i, \tag{2.121}$$

where f is known to the analyst and the goal is to estimate the three parameters β^*, ξ^*, ν^*. (Sometimes only one of these parameters is of interest, so that the others can be considered nuisance parameters. But this perspective does not fundamentally change the analysis.) The M-estimator is then defined as in (2.8) with

$$m_\theta(x,y) := \mathsf{L}(y - \beta f((X_i - \xi)/\nu)), \tag{2.122}$$

with the minimization now being over Θ. We of course assume (2.110), and in addition, that f_{θ^*} is not constant on the support of λ. So that the setting is non-smooth, we assume below that f is piecewise Lipschitz. For concreteness, we focus on a loss function which has a bounded second derivative. Otherwise, all the other basic assumptions are in place.

It turns out that the Euclidean norm, which was used in Section 2.3 and Section 2.4 as the absolute value, and in Section 2.5.1, is here *not* a good reference semimetric. Instead, we use the following

$$\mathsf{d}(\theta_1, \theta_2) := (\beta_1 - \beta_2)^2 + |\xi_1 - \xi_2| + |\nu_1 - \nu_2|. \tag{2.123}$$

The square-root of that would also do, but the main feature of this semimetric is that the amplitude is treated differently than the location and the scale, and this will be crucial when deriving rates of convergence (as we do below) since the rate of convergence for the estimation of the amplitude is in $\sqrt{n}$ while those for the location and scale are in n.

Consistency

Θ as defined above is not compact, and some odd things can happen. To avoid complications, we assume that the support of λ contains in its interior the support of f_{θ^*}. Besides being a somewhat natural restriction, with this assumption we are able to reduce the analysis to a compact subset $\Theta_0 \subset \Theta$. We do so in Lemma 12, where we make mild assumptions on the noise distribution and loss function, which are essentially the same underlying Assumption 5, as detailed in Lemma 10. Then consistency is, again, a consequence of the fact that $\{m_\theta : \theta \in \Theta_0\}$ is Glivenko–Cantelli — in fact, it is Donsker, as we argue below.

Rate of convergence

To obtain a rate of convergence, we rely on Lemma 5 as before. We note that Lemma 2 remains valid, now with $\Delta(\theta, \theta^*) = \int (f_\theta(x) - f_{\theta^*}(x))^2 \lambda(x) \mathrm{d}x$. Without loss of generality, assume that $\theta^* = (1, 0, 1)$ henceforth. Developing the square inside the integral, we get

$$\begin{aligned}
\Delta(\theta, \theta^*) = (1-\beta)^2 &\int f(x)^2 \lambda(x) \mathrm{d}x \\
&+ \beta^2 \int (f(x) - f(x-\xi))^2 \lambda(x) \mathrm{d}x \\
&+ \beta^2 \int (f(x-\xi) - f((x-\xi)/\nu))^2 \lambda(x) \mathrm{d}x \\
&+ 2(1-\beta)\beta \int f(x)(f(x) - f(x-\xi)) \lambda(x) \mathrm{d}x \\
&+ 2(1-\beta)\beta \int f(x)(f(x-\xi) - f((x-\xi)/\nu) \lambda(x) \mathrm{d}x \\
&+ 2\beta^2 \int (f(x) - f(x-\xi))(f(x-\xi) - f((x-\xi)/\nu) \lambda(x) \mathrm{d}x.
\end{aligned}$$

The first three terms on the RHS are the main terms, as we will see, with the other terms being negligible. The 1st term is $= C(1-\beta)^2$. For the following, consider (β, ξ, ν) approaching $(\beta^*, \xi^*, \nu^*) = (1, 0, 1)$. As we saw in (2.77), the 2nd term is $\sim |\xi| D$. Similar calculations establish that the 3rd term is $\sim |\nu - 1| D$, and also that the 4th term is $= O((1-\beta)\xi)$, that the 5th term is

$= O((1-\beta)(1-\nu))$, and that the 6th term is $= O(\xi(1-\nu))$. We thus conclude that

$$\Delta(\theta,\theta^*) \asymp \mathsf{d}(\theta,\theta^*), \quad \theta \to \theta^*. \tag{2.124}$$

In particular, (2.70) holds with $a = 1$.

We now look at the class $\mathcal{M}_\delta := \{m_\theta - m_{\theta^*} : \mathsf{d}(\theta,\theta^*) \le \delta\}$. Using similar arguments as those underlying Lemma 3, we derive

$$|m_\theta(x,y) - m_{\theta^*}(x,y)| \le \overline{m}(z)|f(x) - \beta f((x-\xi)/\nu)|,$$

where $z := y - f(x)$ and $\overline{m}$ is a square integrable function with respect to ϕ. We then have

$$\begin{aligned}
|f(x) - \beta f((x-\xi)/\nu)| &\le |f(x) - \beta f(x)| + |\beta f(x) - \beta f(x-\xi)| + |\beta f(x-\xi) - \beta f((x-\xi)/\nu)| \\
&\le |1-\beta||f(x)| + |\beta||f(x) - f(x-\xi)| + |\beta||f(x-\xi) - f((x-\xi)/\nu)| \\
&\le |1-\beta||f|_\infty + |\beta|C(|\xi| + \mathbb{I}\{x \in \mathcal{D}(\delta)\}) + |\beta|C(|\nu - 1| + \mathbb{I}\{x \in \mathcal{D}(\delta)\}),
\end{aligned}$$

where for the last two terms on the RHS we reasoned as in the proof of Lemma 3. When $\mathsf{d}(\theta,\theta^*) \le \delta$, we have $|\beta - 1| \le \delta^{1/2}$, while $|\xi| \le \delta$ and $|\nu - 1| \le \delta$, so that

$$|m_\theta(x,y) - m_{\theta^*}(x,y)| \le \overline{m}(z) \cdot C(\delta^{1/2} + \mathbb{I}\{x \in \mathcal{D}(\delta)\}).$$

The function on the RHS is therefore an envelope for the class $\mathcal{M}_\delta$. In terms of bracketing numbers, reasoning as above, we find that for any θ, θ_0 within δ of θ^*,

$$|m_\theta(x,y) - m_{\theta_0}(x,y)| \le \overline{m}(z) \cdot \big(|\beta - \beta_0| + \mathbb{I}\{\mathrm{dist}(x, \xi_0 + \nu_0\mathcal{D}) \le C(|\xi - \xi_0| \vee |\nu - \nu_0|)\}\big),$$

where again $\overline{m}$ is square integrable. For $\theta = (\beta,\xi,\nu)$, let $\beta_0 \in \{1 + (j/k)\delta^{1/2} : j = -k, \ldots, k\}$ such

that $|\beta - \beta_0| \le \delta^{1/2}/k$; let $\xi_0 \in \{(j/l)\delta : j = -l, \dots, l\}$ such that $|\xi - \xi_0| \le \delta/l$; and let $\nu_0 \in \{1 + (j/l)\delta : j = -l, \dots, l\}$ such that $|\nu - \nu_0| \le \delta/l$. Then

$$m_\theta(x,y) = m_{\theta_0}(x,y) \pm \overline{m}(z) \cdot \big(\delta^{1/2}/k + \mathbb{I}\{\mathrm{dist}(x, \xi_0 + \nu_0 \mathcal{D}) \le C\delta/l\}\big),$$

so that m_θ is in the bracket defined by the two functions (differing by $\pm$) on the RHS. The square integral of the difference between these two functions is $\asymp \delta/k^2 + \delta/l$. So this is bounded by ε^2, we choose $k \asymp \delta^{1/2}/\varepsilon$ and $l \asymp \delta/\varepsilon^2$. We conclude that we can cover $\mathcal{M}_\delta$ with $k \times l \times l = O(\delta^{1/2}/\varepsilon)^5$ ε-brackets, so that $\mathcal{M}_\delta$ has ε-bracketing number $N_\delta(\varepsilon) \le C(\delta^{1/2}/\varepsilon)^5$. And from this we deduce, as in the proof of Theorem 5, that (2.71) holds with $b = 1/2$.

We are thus in a position to apply Lemma 5 and get that the M-estimator $\hat{\theta}_n$ is n-consistent, which here implies that

$$\mathbb{E}\big[\sqrt{n}|\hat{\beta}_n - \beta^*|\big] = O(1), \qquad \mathbb{E}\big[n|\hat{\xi}_n - \xi^*|\big] = O(1), \qquad \mathbb{E}\big[n|\hat{\nu}_n - \nu^*|\big] = O(1).$$

In particular, $\hat{\beta}_n$ is $\sqrt{n}$-consistent while $\hat{\xi}_n$ and $\hat{\nu}_n$ are n-consistent. And these rates are minimax, as Lemma 6 applies essentially verbatim, all resting on the behavior of $\Delta(\theta, \theta^*)$ as $\theta \to \theta^*$.

Limit distribution

We can also obtain a limit distribution following the arguments underlying Theorem 7. We assume for convenience that f is càglad. In what follows we assume that ξ is sufficiently close to 0 and ν to 1 that intervals of the form $[d \wedge (\xi + \nu d), d \vee (\xi + \nu d)]$ where $d \in \mathcal{D}$ do not intersect. For $d \in \mathcal{D}$, define $\delta_d := f(d^+) - f(d^-)$, which is the 'jump' that f makes at d. In view of the rates of convergence that we obtained just above, we consider $\beta = 1 + r/\sqrt{n}$, $\xi = s/n$, and $\nu = 1 + t/n$, where r, s, t will remain fixed while $n \to \infty$. Using the fact that f is piecewise Lipschitz, we have

- If $d < x \le \xi + \nu d$, for some $d \in \mathcal{D}$, then

$$\begin{aligned} m_\theta(x,y) - m_{\theta^*}(x,y) &= \mathsf{L}(z + \delta_d \pm C/\sqrt{n}) - \mathsf{L}(z) \\ &= B_d^+(z) \pm \overline{m}(z)/\sqrt{n}, \quad B_d^+(z) := \mathsf{L}(z+\delta_d) - \mathsf{L}(z), \end{aligned}$$

  where $\overline{m}$ is generic for a square integrable function. To arrive there we used a Taylor development of L around $z + \delta_d$.

- Similarly, if $\xi + \nu d < x \le d$, for some $d \in \mathcal{D}$, then

$$m_\theta(x,y) - m_{\theta^*}(x,y) = B_d^-(z) \pm \overline{m}(z)/\sqrt{n}, \quad B_d^-(z) := \mathsf{L}(z-\delta_d) - \mathsf{L}(z).$$

- If there is no discontinuity point between x and $(x-\xi)/\nu$, then using the fact that

$$\mathsf{L}(z+t) - \mathsf{L}(z) = \mathsf{L}'(z)t + \frac{1}{2}\mathsf{L}''(z)t^2 \pm \frac{1}{2}\gamma(z,|t|)t^2,$$

  with γ defined in (2.61), and the fact that

$$f(x) - \beta f((x-\xi)/\nu) = (1-\beta)f(x) + R(x;\beta,\xi,\nu), \tag{2.125}$$

  with

$$|R(x;\beta,\xi,\nu)| \le C(|\xi| + |\nu - 1|) \le C/n, \tag{2.126}$$

  we have

$$\begin{aligned} m_\theta(x,y) - m_{\theta^*}(x,y) = \mathsf{L}'(z)(1-\beta)f(x) + \mathsf{L}'(z)R(x;\beta,\xi,\nu) + \tfrac{1}{2}\mathsf{L}''(z)(1-\beta)^2 f(x)^2 \\ \pm \tfrac{1}{2}\gamma(z, C/n)C/n \pm C/n^2. \end{aligned}$$

Hence, denoting $I_d(\xi,\nu') := (d \wedge (d+\xi+d(\nu'-1)), d \vee (d+\xi+d(\nu'-1))]$, we have

$$
\begin{aligned}
&n\big(\widehat{M}_n(\theta) - \widehat{M}_n(\theta^*)\big) \\
&= \sum_i \sum_{d\in\mathcal{D}} \Big(B_d^+(Z_i)\mathbb{I}\{d < X_i \le d + s/n + dt/n\} + B_d^-(Z_i)\mathbb{I}\{d + s/n + dt/n < X_i \le d\}\Big) \\
&\quad + \sum_i \sum_{d\in\mathcal{D}} (\overline{m}(Z_i)/\sqrt{n})\mathbb{I}\{X_i \in I_d(s/n,t/n)\} \\
&\quad + \sum_i \mathsf{L}'(Z_i)(1-\beta)f(X_i)\mathbb{I}\{X_i \notin \cup_d I_d(s/n,t/n)\} \\
&\quad + \sum_i \tfrac{1}{2}\mathsf{L}''(Z_i)(1-\beta)^2 f(X_i)^2\mathbb{I}\{X_i \notin \cup_d I_d(s/n,t/n)\} \\
&\quad + \sum_i \mathsf{L}'(Z_i)R(X_i;\beta,s/n,t/n)\mathbb{I}\{X_i \notin \cup_d I_d(s/n,t/n)\} \\
&\quad \pm \sum_i \tfrac{1}{2}\gamma(Z_i,C/n)(C/n)\mathbb{I}\{X_i \notin \cup_d I_d(s/n,t/n)\} \\
&\quad \pm \sum_i (C/n^2)\mathbb{I}\{X_i \notin \cup_d I_d(s/n,t/n)\}.
\end{aligned}
$$

As $n \to \infty$, the 1st sum on the RHS converges to $\sum_{d\in\mathcal{D}} N_d(s+td)$, where $\{N_d : d \in \mathcal{D}\}$ are independent two-sided marked Poisson processes on the real line, with N_d having intensity $\lambda(d)$ and mark distribution that of $B_d^+(Z)$. (Note that $B_d^+(Z) \sim B_d^-(Z)$.) This process can be equivalently expressed as $\sum_{d\in\mathcal{D}} N_d(s) + \sum_{d\in\mathcal{D}} \tilde{N}_d(t)$, where N_d is as before while $\tilde{N}_d$ has intensity $\lambda(d)d$ and same mark distribution, and all these processes are independent of each other. The 2nd sum has expectation

$$
n\sum_{d\in\mathcal{D}} (C/\sqrt{n})\,\mathbb{P}(X \in I_d(s/n,t/n)) \asymp 1/\sqrt{n}, \quad \text{since } \mathbb{P}(X \in I_d(s/n,t/n)) = O(1/n),
$$

and so it converges to 0 in probability by Markov's inequality. Because $\mathbb{P}(X \notin \cup_d I_d(s/n,t/n)) \to 1$, by the Lindeberg central limit theorem, the 3rd sum converges in distribution to the normal distribution with zero mean and variance $r^2\,\mathbb{E}[\mathsf{L}'(Z)^2]\,\mathbb{E}[f(X)^2]$. The 4th sum converges to $\frac{1}{2}r^2\,\mathbb{E}[\mathsf{L}''(Z)]\,\mathbb{E}[f(X)^2]$ in probability, essentially by the law of large numbers. (More rigorously,

via Chebyshev's inequality, for example.) The 5th sum has zero mean and variance bounded by $n\mathbb{E}[\mathsf{L}'(Z)^2]\mathbb{E}[R(X_i;\beta,s/n,t/n)^2] = O(1/n)$, and so converges to 0 in probability by Markov's inequality. The 6th sum is non negative and has expectation bounded by $C\mathbb{E}[\gamma(Z,C/n)] = o(1)$, by dominated convergence, so that it converges to 0 in probability by Markov's inequality. The 7th sum is simply $= O(1/n)$.

The minimization over β, or equivalently over r, corresponds in the limit to a classical setting, since the Gaussian process, as a (random) function of r, can be expressed as $\tilde{C}_1^{1/2} rU$, where $\tilde{C}_1 := \mathbb{E}[\mathsf{L}'(Z)^2]\mathbb{E}[f(X)^2]$ and U is standard normal, and the drift term is equal to $\tilde{C}_0 r^2$, $\tilde{C}_0 := \frac{1}{2}\mathbb{E}[\mathsf{L}''(Z)]\mathbb{E}[f(X)^2]$. In particular, we have established that

$\sqrt{n}(\hat{\beta}_n - \beta^*)$ *converges weakly to the normal distribution with zero mean and variance*

$$(\mathbb{E}[\mathsf{L}'(Z)^2]/\mathbb{E}[\mathsf{L}''(Z)]^2)\mathbb{E}[f(X)^2]^{-1}.$$

The minimization over ξ, or equivalently over s, corresponds in the limit to the minimum a compound Poisson process. In particular, as in Theorem 7, we can establish that

$n(\hat{\xi}_n - \xi^*)$ is asymptotically, and in distribution, between the two most extreme minimizers of $\sum_d N_d$.

And, similarly,

$n(\hat{\nu}_n - \nu^*)$ is asymptotically, and in distribution, between the two most extreme minimizers of $\sum_d \tilde{N}_d$.

Since the minimization over ξ and ν rely, in the limit, on independent processes, $\hat{\xi}_n$ and $\hat{\nu}_n$ are independent in the asymptote. Also, these processes are asymptotically independent of the Gaussian process driving the minimization over β, as these Poisson processes only rely on $O_P(1)$ data points. Hence,

$\hat{\beta}_n$, $\hat{\xi}_n$ and $\hat{\nu}_n$ are asymptotically mutually independent.

2.6 Variants and extensions

2.6.1 Fixed design

A fixed design is most common in signal processing, and also in change point analysis. Most, if not all, of our results can be proved with very similar tools in that context. So that there is a close correspondence with the basic setting of (2.1), in the context of a fixed design we assume that

$$Y_i = f(x_i - \theta^*) + Z_i, \quad i = 1, \dots, n, \tag{2.127}$$

where $x_i = \Lambda^{-1}(i/(n+1))$, Λ being the distribution with density λ and λ being as before. That the same results apply is due to the fact that the empirical process theory behind our results extends quite naturally (and easily) to the independent-but-not-necessarily-iid setting, which is exactly the extension needed since $(x_1, Y_1), \dots, (x_n, Y_n)$ are no longer iid but are still assumed independent. The additional complexity is the fact that now the underlying distribution depends on n: In terms of $(x_1, Z_1), \dots, (x_n, Z_n)$, it is the product of $\delta_{x_i} \otimes \phi$ over $i = 1, \dots, n$. But this presents no particular difficulty in our case.

These results from empirical process theory enable us to understand the large-n behavior of $\widehat{M}_n(\theta) - M_n(\theta)$, where

$$M_n(\theta) := \mathbb{E}[\widehat{M}_n(\theta)] = \frac{1}{n}\sum_{i=1}^{n} m_\theta(x_i), \qquad m_\theta(x) := \mathbb{E}[\mathsf{L}(Z + f(x - \theta^*) - f(x - \theta))].$$

But what is the behavior of M_n itself? As is clear, M_n is a Riemann sum, and when f is piecewise Lipschitz, for example, and L is as before, one can easily prove that

$$\sup_\theta |M_n(\theta) - M(\theta)| \le C/n, \tag{2.128}$$

where C depends f and L. The limit function M is defined exactly as before in (2.11), and we

have used the fact that the supremum is over a compact set when f and λ are compactly supported, which we assume them to be.

2.6.2 Periodic template

In the signal processing literature it is not uncommon to consider the periodic setting where the design points are effectively in a torus rather than the real line. This is equivalent to considering a template that is periodic. The torus can be taken to be the unit interval with algebra there done modulo 1. In that case, the design density λ is simply a density with support the unit interval, and the assumption that the template f is compact is waved. So that the shift is identifiable, we require that f is exactly 1-periodic meaning that $f(\cdot - \theta) \neq f(\cdot - \theta^*)$ unless $\theta = \theta^*$ (remember, modulo 1).

As can be easily verified, all our results apply in that setting with only minor modifications, if any at all.

Remark 9. In signal processing it is common to match a template to a signal by maximizing its convolution or Pearson correlation with the signal. Considering a regular design (also very common) where $x_i = i/n$, this amounts to defining the following estimator

$$\hat{t} := \operatorname*{arg\,max}_{t \in [n]} \sum_{i=1}^{n} f\left(\tfrac{i-t}{n}\right) Y_i \ . \tag{2.129}$$

The parameter θ corresponds to t/n and is here constrained to be on the grid, as is typically the case in signal processing. This estimator, as defined, is in fact the least squares estimator since

$$\sum_{i=1}^{n} \left(Y_i - f\left(\tfrac{i-t}{n}\right)\right)^2 = \sum_{i=1}^{n} Y_i^2 + \sum_{i=1}^{n} f\left(\tfrac{i-t}{n}\right)^2 - 2\sum_{i=1}^{n} f\left(\tfrac{i-t}{n}\right) Y_i \ , \tag{2.130}$$

and the first two sums on right-hand side do not depend on t. (For the second one, this is because f is 1-periodic.)

2.6.3 Agnostic setting

In practice, the model (2.1) could be completely wrong. Suppose, however, that the following holds

$$Y_i = g(X_i) + Z_i, \tag{2.131}$$

with otherwise the same assumptions on the design and noise. (This type of situation is sometimes referred to as a 'mispecified model' or 'improper learning'.) In that case, the function $\tilde{m}_\theta(x,z)$ of (2.57), which plays a crucial role in our derivations, takes the form

$$\tilde{m}_\theta(x,z) = \mathsf{L}(z + g(x) - f(x-\theta)). \tag{2.132}$$

It then becomes quite clear that most, if not all of our results extend to this case, if it is true that

$$\theta^* := \arg\min_\theta \mathbb{E}[\tilde{m}_\theta(X,Z)] \tag{2.133}$$

is uniquely defined. Almost no conditions on g are required other than, say, boundedness.

A case in point is where g, a function on the unit interval, is known to have a single discontinuity, say at d. Note that d is unknown. Then one may want to use a stump, $f(x) = a\mathbb{I}\{x > 0\}$, to locate the discontinuity. If one uses the squared error loss for example, it happens that θ^* is exactly the location of the discontinuity if it is the case that $g(x) < a/2$ when $x < d$ and $g(x) > a/2$ when $x > d$. And using an additional scale parameter as in (2.121) frees one from having to guess a good value for a. The point here is that a simple, parameteric model can be used to locate a feature of interest (a discontinuity in this example) in an otherwise 'nonparametric' setting.

2.6.4 Semi-parametric models

The models we discussed in this paper are all parametric. This was intentional, to keep the exposition focused. However, empirical process theory has developed an arsenal of tools for semi-parametric models [VdV98, BKRW98, Ch 25]. An emblematic example in the context of matching a template would be a class of piecewise Lipschitz functions: the parametric component of the model would be the location of the knots defining the intervals where the template is Lipschitz, while the nonparametric component would be the Lipschitz functions defining the template on these intervals. [Kor88] considered this problem in the continuous white noise model and showed that it is possible to locate the location of a discontinuity to what corresponds to $O(1/n)$ accuracy in our context by simply looking for unusually large slopes. This is similar to the approach we allude to at the end of Section 2.6.3. In any case, the same rate of convergence can be obtained for the M-estimator using tools similar to the ones we used in this paper; see [VdV98, Sec 5.8.1].

2.6.5 Alignment/registration

We mentioned in the Introduction the close relationship with the problem of signal alignment (aka registration). A typical setting, that most resembles our basic shift model (2.1), is where

$$Y_{ij} = f(X_{ij} - \theta_j) + Z_{ij}, \quad i \in [n], \quad j \in [s], \tag{2.134}$$

so that instead of a single sample of size n from (2.1), we are provided with s such samples, all shifted differently. Invariably, the function f is unknown, and not much needs to be assumed about f to estimate the shifts (i.e., align the signals) with $\sqrt{n}$ precision. Indeed, with some kernel smoothing, it is possible to design an estimator that is $\sqrt{n}$-consistent, as shown in [HM90, GLM07, Vim10, TIR11]. Although this convergence rate can be achieved with hardly any assumption of f, our work here, which we placed in the context of the change point analysis

literature, would indicate that this rate is suboptimal when f has a discontinuity.

2.6.6 Concentration bounds

The rate of convergence and the limit distribution, in each case, was established under very mild assumptions on the noise distribution. In particular, in terms of tail decay, we only assumed that it was sufficient for the expected loss to be finite at the true value of the parameter, i.e., $\mathbb{E}[\mathsf{L}(Z)] < \infty$. As we discussed earlier, for the losses considered here (and almost everywhere else in the literature), this implies that $\mathbb{E}[\mathsf{L}'(Z+c)^2] < \infty$ for every $c > 0$, which is all that was needed.

When some tail decay is assumed, then concentration bounds can be derived. The simplest and most straightforward example of that is when the loss function is bounded, as is the case for the Tukey loss. Then, under no additional assumption on the noise distribution, the M-estimator enjoys sub-Gaussian concentration. This can be seen from combining Talagrand's inequality for empirical processes [Tal96] and [vdVW96, Th 3.2.5] — the latter being a more general version of [VdV98, Th 5.52], which is the result at the foundation of Lemma 5.

2.7 Numerical experiments

We performed some basic experiments to probe our theory. We present the result of these experiments below, subdivided into 'smooth' and 'non-smooth' settings. The design distribution is the uniform distribution on the unit interval. We consider three noise distributions: Gaussian, Student t-distribution with 3 degrees of freedom, and Cauchy. And we consider four losses: squared error, absolute-value, Huber, and Tukey. We assume throughout that $\theta^* = 0$.

The implementation of Python codes used in numerical experiments is available on Github: https://github.com/zhenglin0266/Template-Matching

2.7.1 Smooth setting

We consider the following two filters:

$$\text{Template } A\text{:}\quad f(x) = \begin{cases} 4x-1 & 0.25 \le x < 0.5, \\ 3-4x & 0.5 \le x < 0.75, \\ 0 & \text{otherwise} \end{cases} \tag{2.135}$$

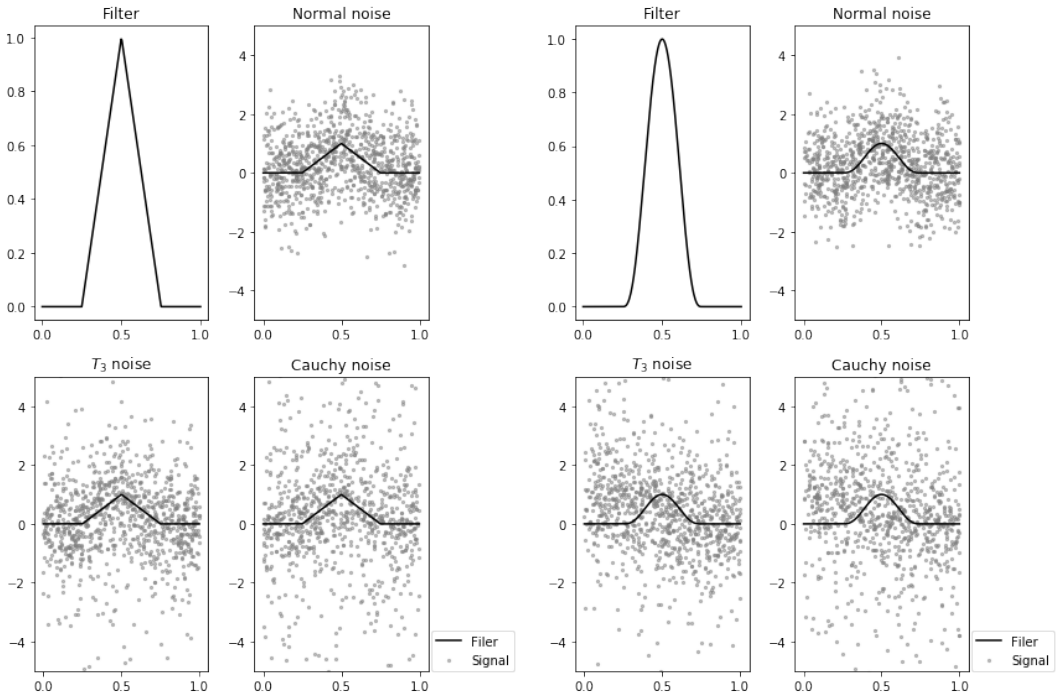
and

$$\text{Template } B\text{:}\quad f(x) = \max\{0, (1-(4x-2)^2)^3\}. \tag{2.136}$$

Template A is Lipschitz, while Template B is even smoother. Our motivation for considering Template B is to verify that more smoothness does not change things much (as predicted by our theory). See Figure 2.2 for an illustration.

We used a sample of size $n = 10000$ and repeated each scenario (combination of template, noise distribution, and loss function) 200 times. We show the mean of $|\sqrt{n}(\hat{\theta}_n - \theta^*)|$ in Table 2.1. Box plots of estimation error $|\hat{\theta}_n - \theta^*|$ are shown in Figure 2.3 and Figure 2.4. The distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$ is plotted in Figure 2.5 and Figure 2.6 as an histogram overlaid with the Gaussian distribution predicted by our asymptotic calculations. Out of curiosity, we also looked at the setting where the noise is Cauchy and yet we use squared error as loss in Figure 2.7. The result of these experiments are by and large congruent with our theory. In particular, there is no noticeable difference between the two templates.

We also ran experiments with varying sample size n. We focused on Template A with absolute-value loss and T_3 noise, to investigate the accuracy of the asymptotic distribution as n increases. As sample size we used $n \in \{100, 500, 1000, 5000, 10000\}$. To have a finer sense of the accuracy, we used 1000 repeats. We show the mean of $|\sqrt{n}(\hat{\theta}_n - \theta^*)|$ in Table 2.2, the box plot of



(a) Template A (b) Template B

Figure 2.2: Templates and noisy signals. Although the sample size is $n = 10000$, for the sake of clarity, we only include 1000 points and limit the range of the y-axis to $[-5,5]$.

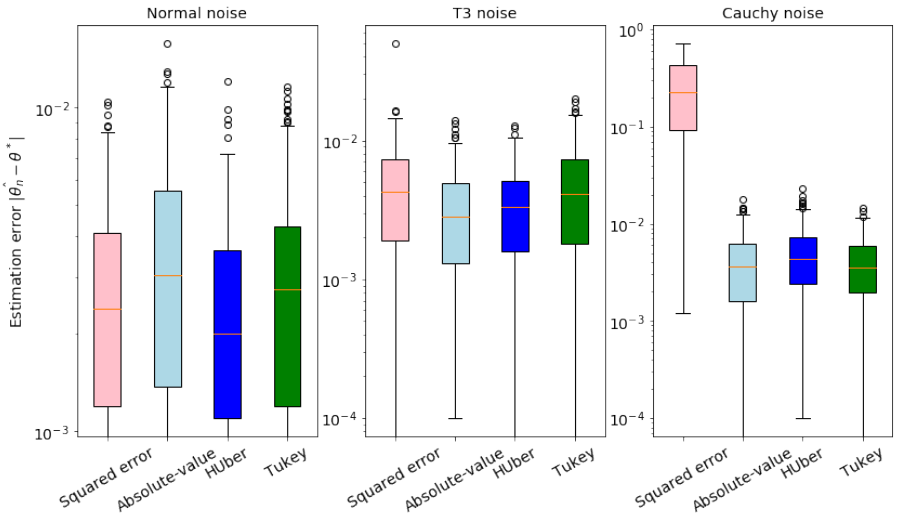


Figure 2.3: Box plot of estimation error $|\hat{\theta}_n - \theta^*|$ for Template A

Table 2.1: Mean of $|\sqrt{n}(\hat{\theta}_n - \theta^*)|$ with n = 10000 over 200 repeats.

Template	Noise	Loss Function			
		Squared error	Absolute-value	Huber	Tukey
Template A	Normal	0.2791	0.3705	0.2620	0.3301
	T_3	0.5168	0.3496	0.3634	0.5053
	Cauchy	28.2535	0.4355	0.5203	0.4113
Template B	Normal	0.2511	0.3326	0.2453	0.2918
	T_3	0.4524	0.3236	0.3293	0.3286
	Cauchy	48.2211	0.3958	0.3982	0.3462

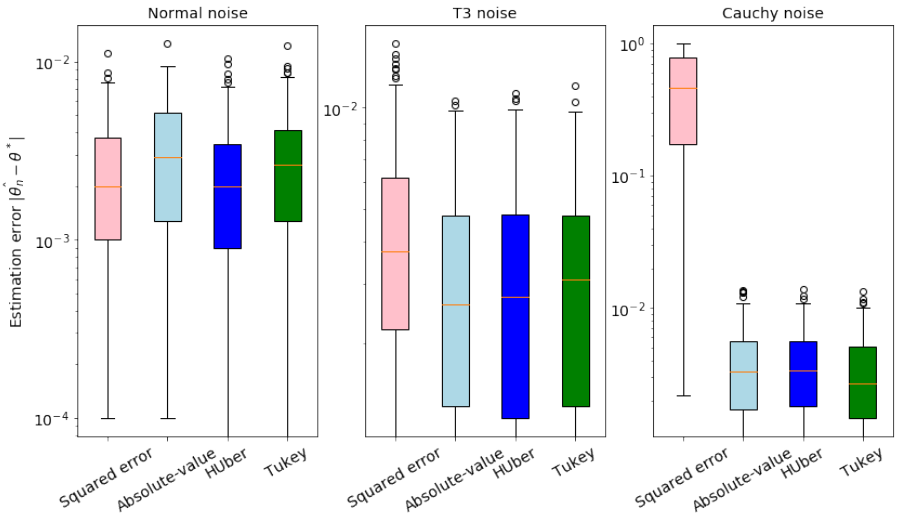


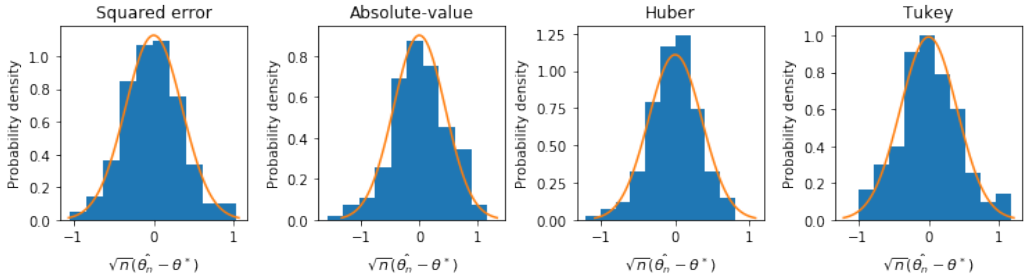
Figure 2.4: Box plot of estimation error $|\hat{\theta}_n - \theta^*|$ for Template B

$|\hat{\theta}_n - \theta^*|$ in Figure 2.8, and the histogram of $\sqrt{n}(\hat{\theta}_n - \theta^*)$ in Figure 2.9.

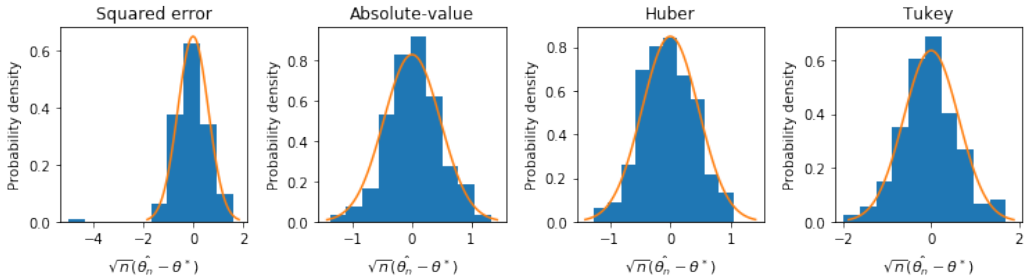
Table 2.2: Mean of $|\sqrt{n}(\hat{\theta}_n - \theta^*)|$ for Template A with absolute-value loss and T_3 noise based on 1000 repeats.

n	100	500	1000	5000	10000
Mean of $\lvert\sqrt{n}(\hat{\theta}_n - \theta^*)\rvert$	0.5704	0.4286	0.4278	0.3766	0.3889

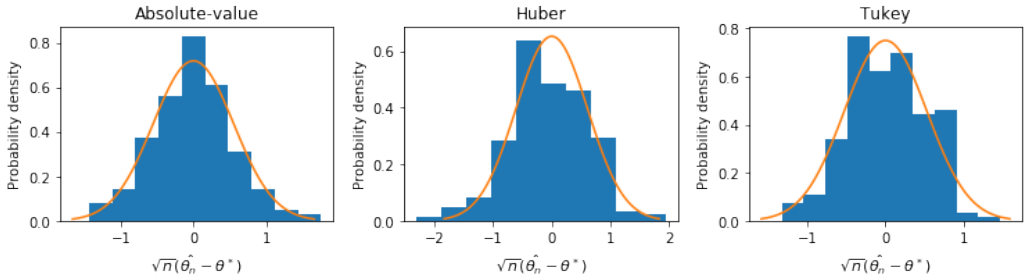
In Figure 2.9, as $n \to \infty$, the distributions of $\sqrt{n}(\hat{\theta}_n - \theta^*)$ approaches to the normal distribution shown in the theory. Figure 2.8 indicates estimation error $|\hat{\theta}_n - \theta^*|$ decreases as n increases.



(a) Normal noise

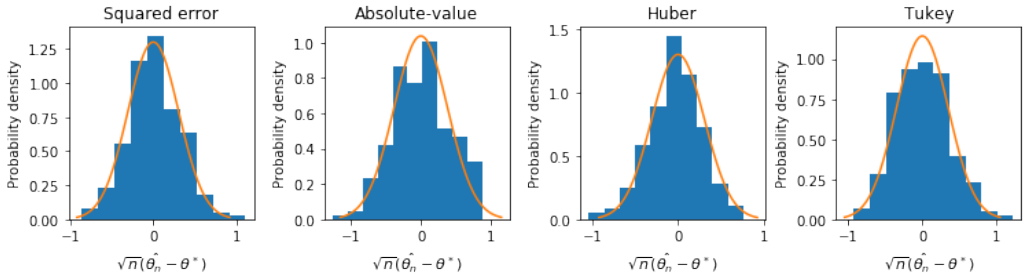


(b) T_3 noise

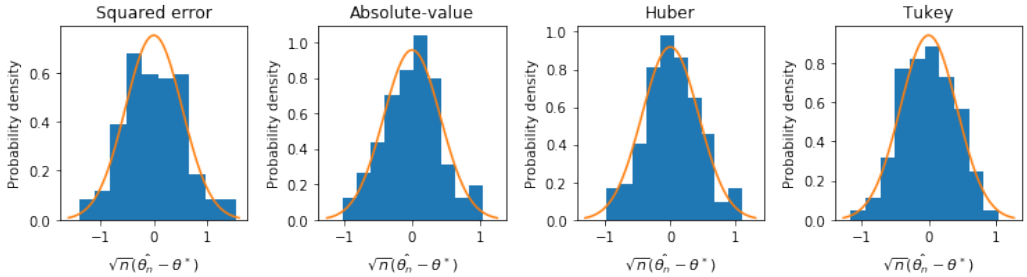


(c) Cauchy noise

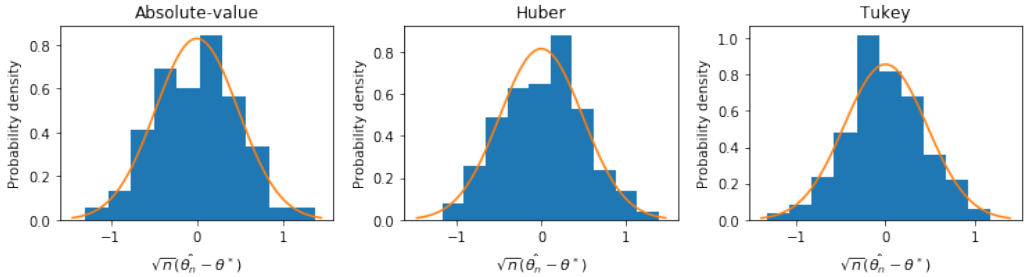
Figure 2.5: Distribution under Template A. The histogram presents the distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$. The orange bell-shaped curve is the density of normal distribution predicted by the theory.



(a) Normal noise



(b) T_3 noise



(c) Cauchy noise

Figure 2.6: Distribution under Template B. The histogram presents the distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$. The orange bell-shaped curve is the density of normal distribution predicted by the theory.

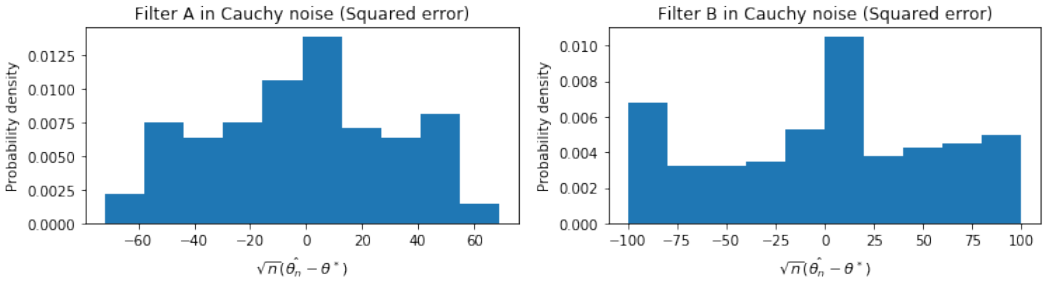


Figure 2.7: Distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$, for Templates A and B, with squared error loss under the Cauchy distribution as noise distribution. Note that our theory is silent on this setting.

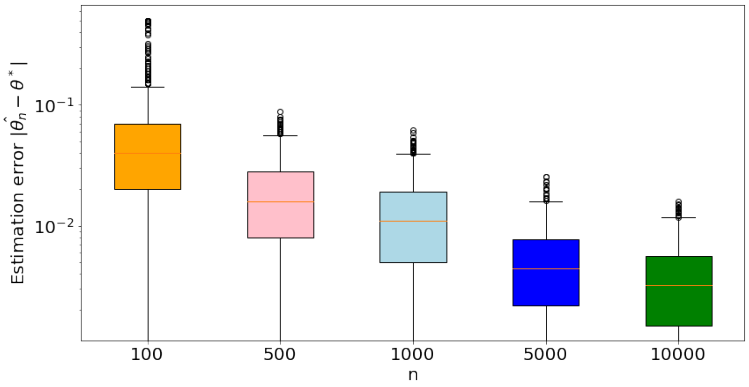


Figure 2.8: Box plot of the estimation error $|\hat{\theta}_n - \theta^*|$ for Template A with absolute-value loss and T_3 noise based on 1000 repeats.

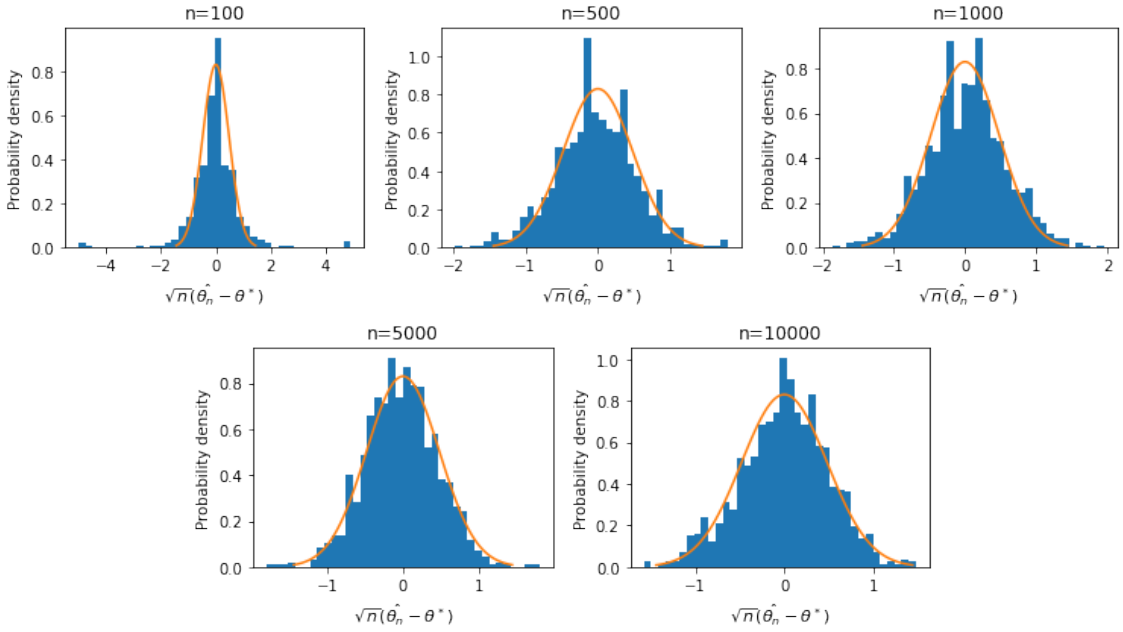


Figure 2.9: Distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$ for Template A with absolute-value loss and T_3 noise based on 1000 repeats.

2.7.2 Non-smooth setting

We consider the following three filters:

$$\text{Template } C\text{:}\quad f(x) = \begin{cases} 1 & 0.25 \le x < 0.75, \\ 0 & \text{otherwise,} \end{cases} \tag{2.137}$$

$$\text{Template } D\text{:}\quad f(x) = \begin{cases} 1 & 0.2 \le x < 0.4 \text{ or } 0.6 \le x < 0.8, \\ 0 & \text{otherwise} \end{cases} \tag{2.138}$$

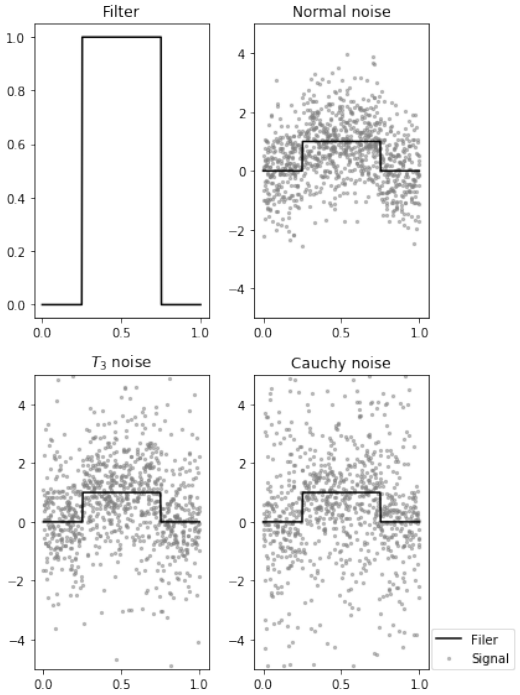
and

$$\text{Template } E\text{:}\quad f(x) = \begin{cases} 4x - 1 & 0.25 \le x < 0.5, \\ 0 & \text{otherwise.} \end{cases} \tag{2.139}$$

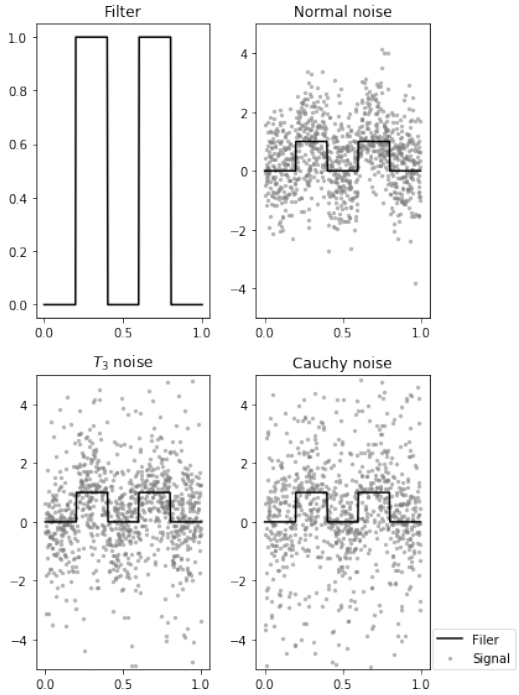
Template C is a piecewise constant function with two discontinuities. Template D is another piecewise constant function with more discontinuities. Template E is a half-triangle with one discontinuity. See Figure 2.10 for an illustration. Our theory predict that what matters is the number and size of the discontinuities.

We show the mean of $|n(\hat{\theta}_n - \theta^*)|$ in Table 2.3, the estimation error $|\hat{\theta}_n - \theta^*|$ is shown as a box plot in Figures 2.11, 2.12, and 2.13, and the distribution of $n(\hat{\theta}_n - \theta^*)$ is plotted in Figures 2.14, 2.15, and 2.16.

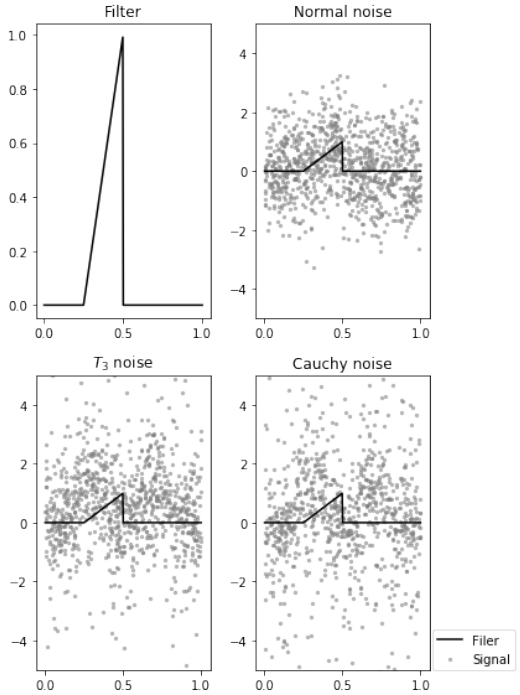
We also ran some experiments with varying n as before. We focused on Template E with the absolute-value loss under the T_3 noise distribution. The mean of $|n(\hat{\theta}_n - \theta^*)|$ is reported in Table 2.4, a box plot of the estimation error $|\hat{\theta}_n - \theta^*|$ is given in Figure 2.17, and the distribution of $n(\hat{\theta}_n - \theta^*)$ is plotted in Figure 2.18.



(a) Template C



(b) Template D



(c) Template E

Figure 2.10: Templates and noisy signals. Although the sample size is $n = 10000$, for the sake of clarity, we only include 1000 points and limit the range of the y-axis to $[-5, 5]$.

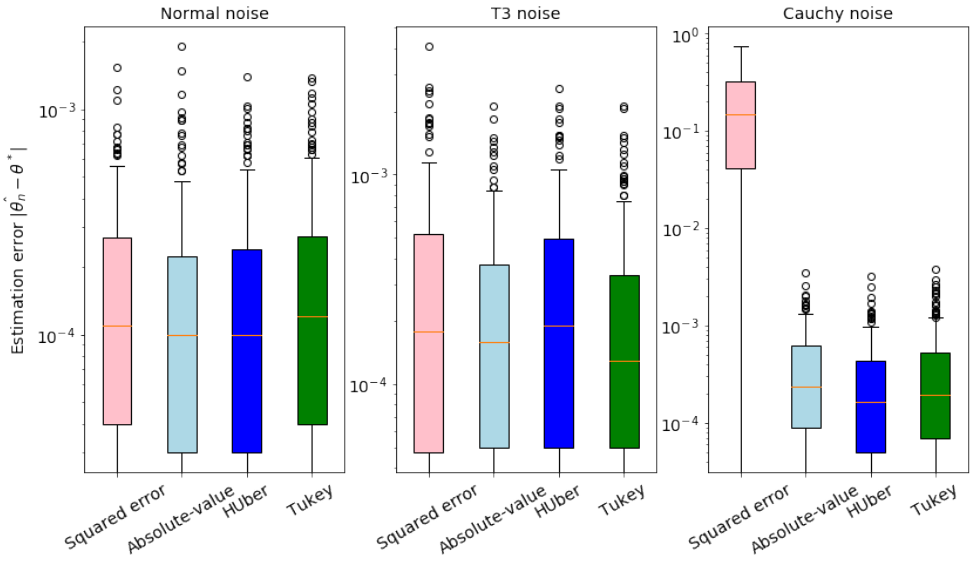


Figure 2.11: Box plot of estimation error $|\hat{\theta}_n - \theta^*|$ for Template *C* based on 200 repeats.

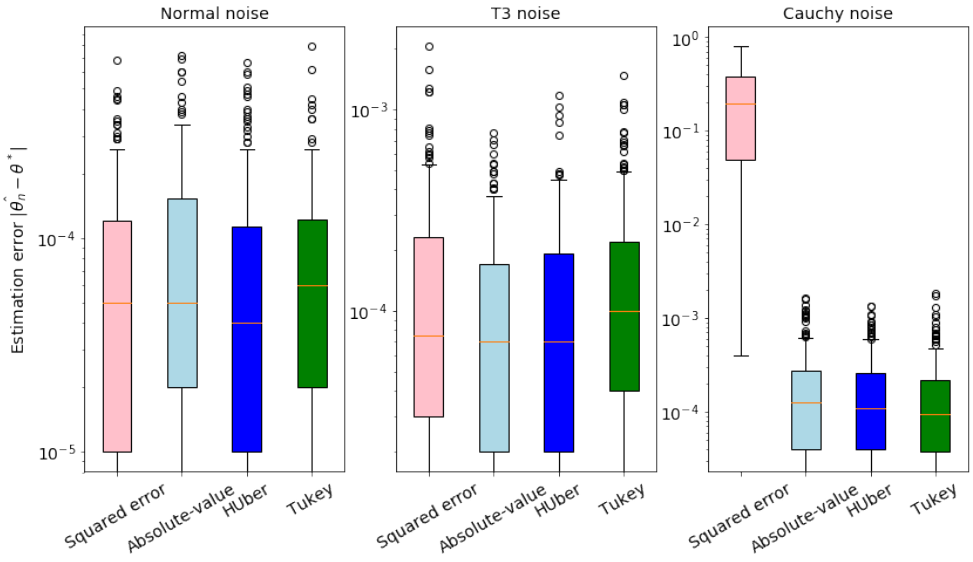


Figure 2.12: Box plot of estimation error $|\hat{\theta}_n - \theta^*|$ for Template *D* based on 200 repeats.

Table 2.3: Mean of $|n(\hat{\theta}_n - \theta^*)|$ based on 200 repeats.

Template	Noise	Loss Function			
		Squared error	Absolute-value	Huber	Tukey
Template C	Normal	1.876	1.849	1.868	2.120
	T_3	3.889	2.739	3.550	2.761
	Cauchy	2326.050	4.362	3.310	4.204
Template D	Normal	0.868	1.080	0.897	0.877
	T_3	1.802	1.220	1.344	1.813
	Cauchy	2766.000	2.278	2.078	1.862
Template E	Normal	3.498	3.798	3.414	4.008
	T_3	7.307	5.366	5.720	5.734
	Cauchy	4798.370	9.102	8.839	7.759

Table 2.4: Mean of $|n(\hat{\theta}_n - \theta^*)|$ for Template E with absolute-value loss under T_3 noise based on 1000 repeats.

n	100	500	1000	5000	10000
Mean of $\lvert n(\hat{\theta}_n - \theta^*)\rvert$	10.1872	5.0808	5.7420	5.7146	5.0682

2.8 Acknowledgments

Chapter 2, in full, is a version of the paper " Template Matching and Change Point Detection by M-estimation", Arias-Castro, Ery; Zheng, Lin. The manuscript has been submitted for publication in a major statistical journal. The dissertation author was the primary investigator and author of this material.

2.9 Appendix

Lemma 10. *Suppose Assumptions 1, 3, and 4 are in place. Then, Assumption 5 is fulfilled if* (i) *the loss is strictly convex, or* (ii) *the noise distribution is unimodal and the loss is either Lipschitz or has a Lipschitz derivative and* ϕ *has finite first moment. Assumption 5 is also fulfilled for the Huber loss and the absolute-value loss if the noise distribution has a unique median.*

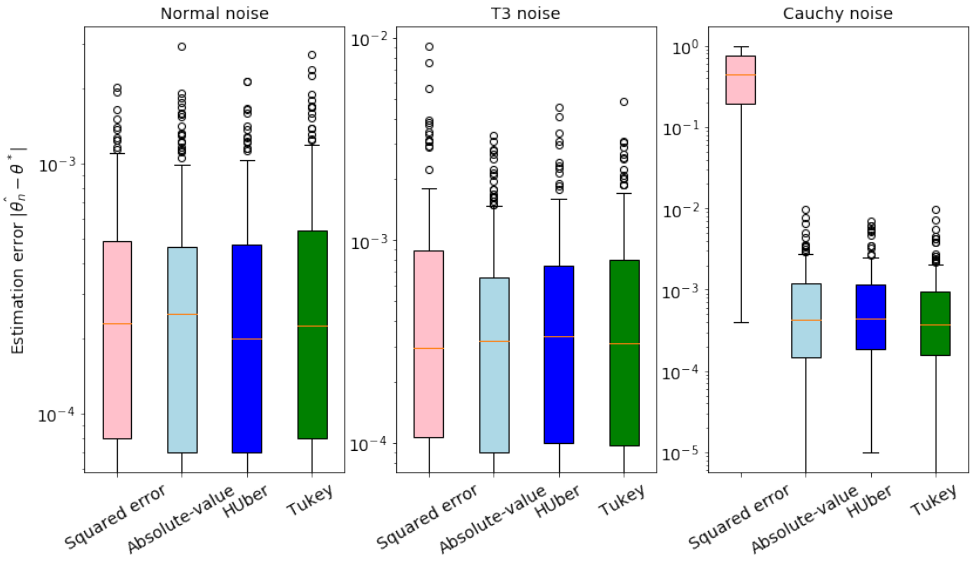


Figure 2.13: Box plot of estimation error $|\hat{\theta}_n - \theta^*|$ for Template E based on 200 repeats.

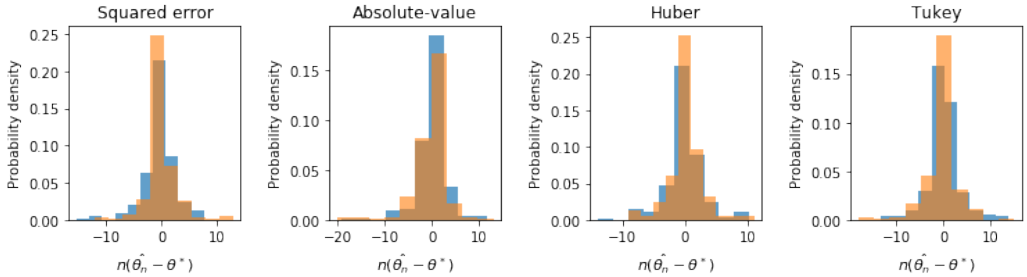
Proof. Assume $\theta^* = 0$ without loss of generality. For x and y given, we let $z = y - f(x)$.

The function $\theta \mapsto m_\theta(x,y)$ is piecewise continuous under our assumptions. Moreover, it is dominated since, by the properties of L,

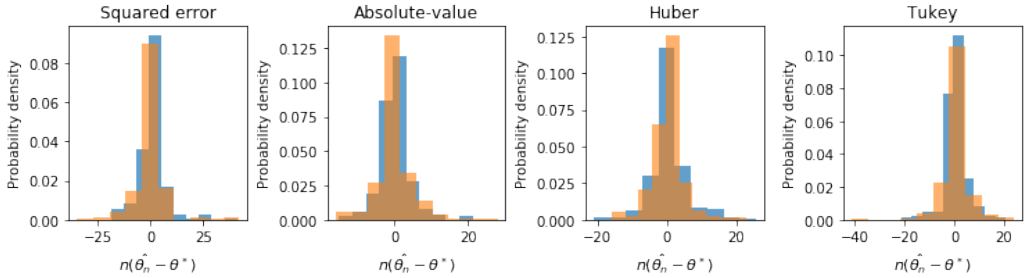
$$\begin{aligned} m_\theta(x,y) &= \mathsf{L}(y - f(x-\theta)) \\ &= \mathsf{L}(y - f(x) + f(x) - f(x-\theta)) \\ &\le \mathsf{L}(|y - f(x)| + |f(x) - f(x-\theta)|) \\ &\le \overline{m}(x,y) := \mathsf{L}(|y - f(x)| + 2|f|_\infty), \end{aligned}$$

and $\mathbb{E}[\overline{m}(X,Y)] < \infty$ by (2.10) and the fact that the loss is slowly increasing. Hence, M is continuous by dominated convergence. In addition, by the fact that f is compactly supported, $m_\theta(x,y) \to \mathsf{L}(y)$ when $\theta \to \pm\infty$, and so we also have $M(\theta) \to M(\infty) := \mathbb{E}[\mathsf{L}(Y)]$ as $\theta \to \pm\infty$. We have

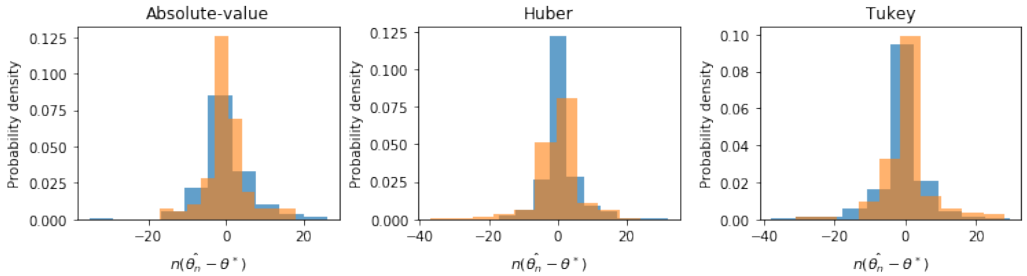
$$M(\theta) = \mathbb{E}[\mathsf{L}(Z + f(X) - f(X-\theta))], \qquad M(\infty) = \mathbb{E}[\mathsf{L}(Z + f(X))].$$



(a) Normal noise

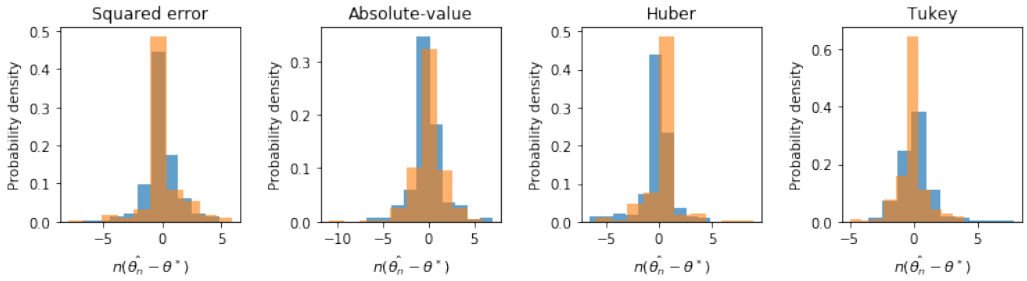


(b) T_3 noise

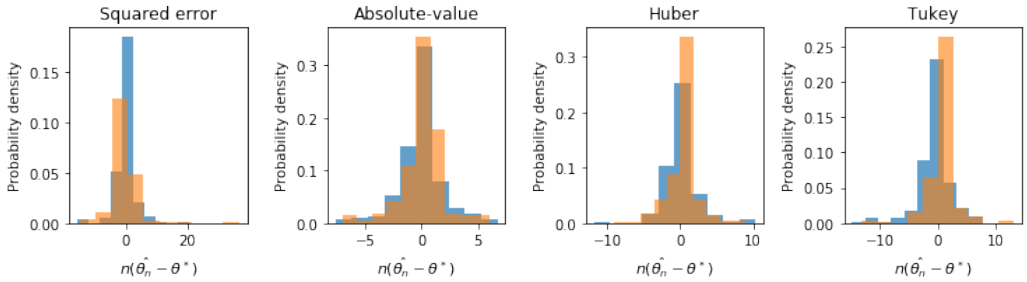


(c) Cauchy noise

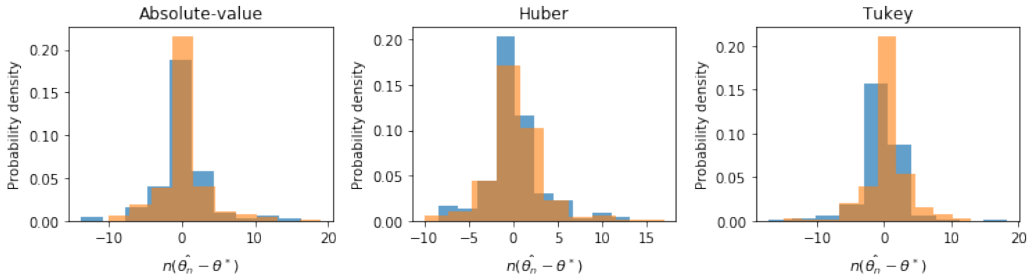
Figure 2.14: The blue histogram represents the distribution of $n(\hat{\theta}_n - \theta^*)$ for Template C. The orange histograms shows the simulated density of the midpoint of the minimizer interval of the marked Poisson process predicted by the theory. Based on 200 repeats.



(a) Normal noise

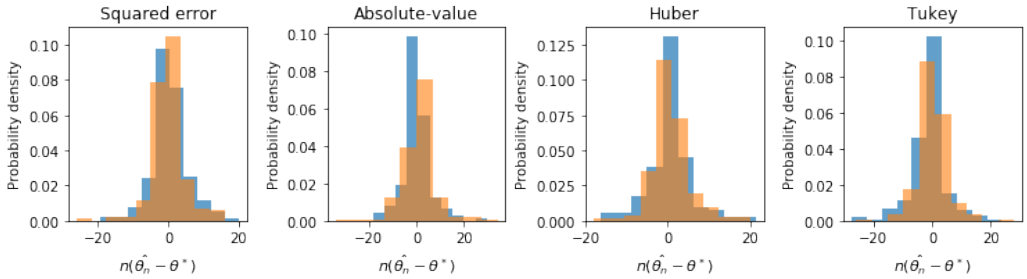


(b) T_3 noise

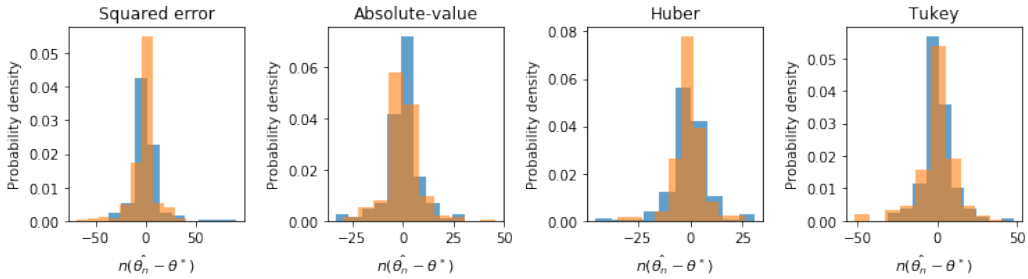


(c) Cauchy noise

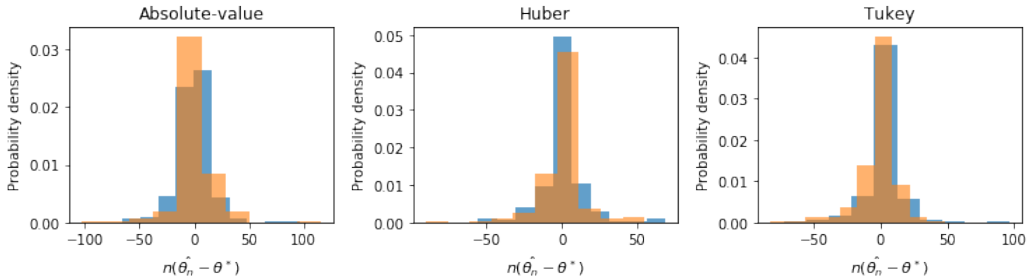
Figure 2.15: The blue histogram represents the distribution of $n(\hat{\theta}_n - \theta^*)$ for Template D. The orange histograms shows the simulated density of the midpoint of the minimizer interval of the marked Poisson process predicted by the theory. Based on 200 repeats.



(a) Normal noise



(b) T_3 noise



(c) Cauchy noise

Figure 2.16: The blue histogram represents the distribution of $n(\hat{\theta}_n - \theta^*)$ for Template E. The orange histograms shows the simulated density of the midpoint of the minimizer interval of the marked Poisson process predicted by the theory. Based on 200 repeats.

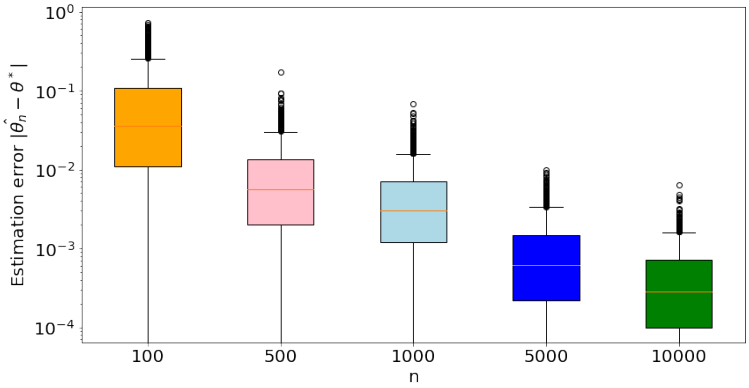


Figure 2.17: Box plot of estimation error $|\hat{\theta}_n - \theta^*|$ for Template E with absolute-value loss under T_3 noise based on 1000 repeats.

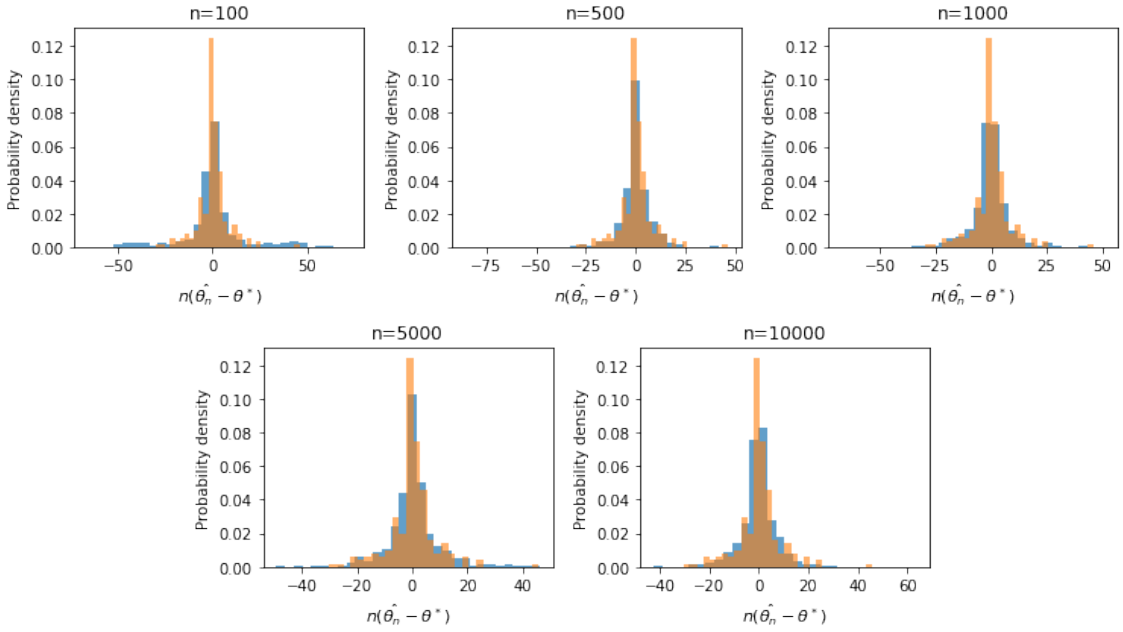


Figure 2.18: The blue histograms present the distribution of $n(\hat{\theta}_n - \theta^*)$ for Template E with absolute-value loss and T_3 noise. The orange histograms show the simulated density of marked Poisson process predicted by the theory. Based on 1000 repeats.

The identifiability condition in Assumption 1 implies that $\mathbb{P}(f(X) - f(X-\theta) \neq 0) > 0$ when $\theta \neq 0$ and $\mathbb{P}(f(X) \neq 0) > 0$, so by conditioning on X, taking into account that X and Z are independent, it suffices to prove that $g(a) := \mathbb{E}[\mathsf{L}(Z+a)]$ achieves its minimum uniquely at $a = 0$. Since g is even, it suffices to prove that $g(a) > g(0)$ for any $a > 0$.

Assume L *is convex.* Then

$$\begin{aligned} g(a) = \mathbb{E}[\mathsf{L}(Z+a)] &= \mathbb{E}[\mathsf{L}(Z-a)] \\ &= \mathbb{E}[\tfrac{1}{2}\mathsf{L}(Z+a) + \tfrac{1}{2}\mathsf{L}(Z-a)] \\ &\geq \mathbb{E}[\mathsf{L}(Z)] = g(0), \end{aligned}$$

where in the second equality we used the fact that the distribution of Z is symmetric about 0 and that L is even, and in the inequality we used the fact that L is convex. That inequality is strict when the loss is strictly convex. This covers the squared error loss, for example.

For the Huber loss (2.5), the inequality is strict unless $Z - a \geq c$ or $Z + a \leq -c$ with probability 1, meaning that $\mathbb{P}(Z \in [-a-c, a+c]) = 0$; for the absolute-value loss, the inequality is strict unless $\operatorname{sign}(Z-a) = \operatorname{sign}(Z+a)$ with probability 1, meaning that $\mathbb{P}(Z \in [-a, a]) = 0$. In either case, this is only possible if the noise distribution does not put any mass in $[-a, a]$, which would then imply that any point in this entire interval is a median of the noise distribution.

Assume that the loss is Lipschitz. In that case, by dominated convergence, g is differentiable with $g'(a) = \mathbb{E}[\mathsf{L}'(Z+a)]$, and all we have to prove is that $g'(a) \geq 0$ when the noise distribution is unimodal. (Recall that we work with $a > 0$.) The uniqueness would then come from the fact that g is not constant. We have

$$\begin{aligned} g'(a) &= \int_{-\infty}^{\infty} \mathsf{L}'(z+a)\phi(z)\,\mathrm{d}z \\ &= \int_{0}^{\infty} \mathsf{L}'(z)\big\{\phi(z-a) - \phi(z+a)\big\}\,\mathrm{d}z. \end{aligned}$$

We claim that the integrand is non-negative. Indeed, $\mathsf{L}'(z) \geq 0$ for $z \geq 0$ since L is non-decreasing away from the origin per Assumption 4. Also, using the fact that ϕ is non-increasing on $[0, \infty)$, we reason as follows: if $z \geq a$, then $z + a \geq z - a \geq 0$, implying $\phi(z+a) \leq \phi(z-a)$; and if $z \leq a$, then $a + z \geq a - z \geq 0$, implying $\phi(a+z) \leq \phi(a-z)$, in turn implying $\phi(z+a) \leq \phi(z-a)$ since ϕ is even. Our claim is thus established. This covers the Tukey loss, for example.

Assume that the loss has a Lipschitz derivative. In that case we can reduce this to the previous case by observing that g above remains differentiable as long as ϕ has a finite first moment. Indeed,

$$\mathbb{E}[|\mathsf{L}'(Z+a)|] \leq \mathbb{E}[|\mathsf{L}''|_\infty |Z+a|] \leq |\mathsf{L}''|_\infty (\mathbb{E}[|Z|] + a) < \infty, \tag{2.140}$$

for all $a \geq 0$. □

Lemma 11. *Suppose that, on the real line, $X_1, \ldots, X_n$ are iid from some density λ, and independently, that $A_1, \ldots, A_n$ are iid with distribution Ψ. Fix a finite subset $d_1 < \cdots < d_p$ where λ can be taken to be continuous. Define $W_{n,j}(t) = \sum_i \{d_j < X_i \leq d_j + t/n\} A_i$. Then $W_{n,j}$ converges weakly to the marked Poisson process with intensity $\lambda(d_j)$ on the positive real line and mark distribution Ψ. Moreover, $W_{n,1}, \ldots, W_{n,p}$ are asymptotically independent.*

Proof. It is suffices to establish the statement on intervals. Without loss of generality, we consider the unit interval, so that we only consider $t \in [0,1]$. Everywhere, n is large enough that $d_{j+1} - d_j > 1/n$ for all j. This guarantees that, for any $t \in [0,1]$, the intervals $[d_j, d_j + t/n]$ are disjoint. Let $\{A_{j,i} : i \geq 1, j = 1, \ldots, p\}$ be iid from Ψ. Then it's easy to see that $W_{n,1}, \ldots, W_{n,p}$, jointly, have the same distribution as $\tilde{W}_{n,1}, \ldots, \tilde{W}_{n,p}$, where

$$\tilde{W}_{n,j}(t) := \sum_{i=1}^{N_{n,j}(t)} A_{j,i}, \quad N_{n,j}(t) := \#\{i : d_j < X_i \leq d_j + t/n\}. \tag{2.141}$$

It is thus sufficient to show that $N_{n,j}$ converges weakly to the Poisson process with intensity $\lambda(d_j)$

on the positive real line and that $N_{n,1}, \ldots, N_{n,p}$ are asymptotically independent. By [Bil99, Th 12.6], it suffices to look at the finite-dimensional distributions. Therefore, fix $0 \le t_1 < \cdots < t_k \le 1$ and $\{n_{j,s} : j = 1, \ldots, p; s = 1, \ldots, k\}$ non-negative integers. Based on the 'balls-in-bins' dependency structure of these counts, we have

$$\mathbb{P}\left(N_{n,j}(t_s) \le n_{j,s}, \forall\, j \in [p], s \in [k]\right) \tag{2.142}$$

$$= \mathbb{P}\left(N_{n,1}(t_s) \le n_{1,s}, \forall s \in [k]\right) \tag{2.143}$$

$$\times\, \mathbb{P}\left(N_{n,2}(t_s) \le n_{2,s}, \forall s \in [k] \mid N_{n,1}(t_k) \le n_{1,k}\right) \tag{2.144}$$

$$\times \cdots \tag{2.145}$$

$$\times\, \mathbb{P}\left(N_{n,p}(t_s) \le n_{p,s}, \forall s \in [k] \mid N_{n,1}(t_k) \le n_{1,k}, \ldots, N_{n,p-1}(t_k) \le n_{p-1,k}\right), \tag{2.146}$$

with

$$\mathbb{P}\left(N_{n,j}(t_s) \le n_{j,s}, \forall s \in [k] \mid N_{n,1}(t_k) \le n_{1,k}, \ldots, N_{n,j-1}(t_k) \le n_{j-1,k}\right)$$

$$\begin{cases} \ge \mathbb{P}\left(N_{n,j}(t_s) \le n_{j,s}, \forall s \in [k]\right) \\ \le \mathbb{P}\left(N_{n-r_j,j}(t_s) \le n_{j,s}, \forall s \in [k]\right), \quad r_j := n_{1,k} + \cdots + n_{j-1,k}. \end{cases}$$

Note that $r_1, \ldots, r_p$ are all fixed, and it's easy to convince oneself that it suffices at this point to show that, for each j,

$$\mathbb{P}\left(N_{n,j}(t_s) \le n_{j,s}, \forall s \in [k]\right) \xrightarrow{n\to\infty} \mathbb{P}\left(N_j(t_s) \le n_{j,s}, \forall s \in [k]\right), \tag{2.147}$$

were N_j is a Poisson process with intensity $\lambda(d_j)$ on the positive real line, in other words, that $N_{n,j}$ converges weakly to N_j. We do so by looking at the probability mass function instead of the

cumulative distribution function. With $t_1 < \cdots < t_k$ and $\{n_{j,s}\}$ generic, we have

$$\begin{aligned}
&\mathbb{P}\left(N_{n,j}(t_1) = n_{j,1}, \ldots, N_{n,j}(t_k) = n_{j,k}\right) \\
&= \mathbb{P}\left(N_{n,j}(t_1) = n_{j,1}, N_{n,j}(t_2) - N_{n,j}(t_1) = n_{j,2} - n_{j,1}, \ldots, N_{n,j}(t_k) - N_{n,j}(t_{k-1}) = n_{j,k} - n_{j,k-1}\right) \\
&\sim \mathbb{P}\left(N_{n,j}(t_1) = n_{j,1}\right) \times \mathbb{P}\left(N_{n,j}(t_2) - N_{n,j}(t_1) = n_{j,2} - n_{j,1}\right) \\
&\qquad \times \cdots \times \mathbb{P}\left(N_{n,j}(t_k) - N_{n,j}(t_{k-1}) = n_{j,k} - n_{j,k-1}\right), \quad \text{as } n \to \infty,
\end{aligned}$$

where the approximation follows the same arguments used above to prove the asymptotic independence of the various count processes. We then conclude with the fact that, for any $m \geq 0$ integer,

$$\begin{aligned}
\mathbb{P}\left(N_{n,j}(t_s) - N_{n,j}(t_{s-1}) = m\right) &= \#\{i : d_j + t_{s-1}/n < X_i \leq d_j + t_s/n\} \\
&\xrightarrow{n\to\infty} \lambda(d_j)(t_s - t_{s-1}) = \mathbb{P}\left(N_j(t_s) - N_j(t_{s-1}) = m\right),
\end{aligned}$$

using the fact that λ is continuous at d_j. □

Lemma 12. *Consider the setting of Section 2.5.2. Assume in addition that* $\mathbb{E}[\mathsf{L}(Z)] < \mathsf{L}(\infty)$ *and that* $\mathbb{E}[\mathsf{L}(Z)] < \mathbb{E}[\mathsf{L}(Z+a)]$ *for all* $a \neq 0$*. Then there is a compact subset* $\Theta_0 \subset \Theta$ *of the form* $[-B,B] \times [-B,B] \times [1/B,B]$ *for some* $B \geq 1$*, such that, with probability tending to 1, any minimizer* $\hat{\theta}_n$ *of* $\widehat{M}_n$ *is in* Θ_0*.*

Proof. Let $a = \inf\{x : f(x) > 0\}$ and $b = \sup\{x : f(x) > 0\}$, and assume that λ is supported on $[-c,c]$. Since we know that f_{θ^*} has support contained within the support of λ, we can restrict attention to those θ such that f_θ has support contained within the support of λ, which is equivalent to ξ and ν satisfying $\xi + a\nu \geq -c$ and $\xi + b\nu \leq c$. Note that this implies that $-c|b+a|/(b-a) \leq \xi \leq c|b+a|/(b-a)$ and $\nu \leq 2c/(b-a)$.

We now show that there is $\nu_0 > 0$ such that, if $\nu \leq \nu_0$, then $\widehat{M}_n(\beta,\xi,\nu)$ is, asymptotically, bounded away (and above) from $\widehat{M}_n(\theta^*)$, regardless of β and ξ. Henceforth, we assume without

loss of generality that $\theta^* = (1, 0, 1)$.

On the one hand, by the law of large numbers, in probability as $n \to \infty$,

$$\widehat{M}_n(\theta^*) \to M(\theta^*) = \mathbb{E}[\mathsf{L}(Z)]. \tag{2.148}$$

On the other hand,

$$\inf_{\beta,\xi,\nu\le\nu_0} \widehat{M}_n(\beta,\xi,\nu) = \inf_{\beta,\xi,\nu\le\nu_0} \frac{1}{n}\sum_i \mathsf{L}\big(Z_i + f(X_i) - \beta f((X_i-\xi)/\nu)\big) \tag{2.149}$$

$$\ge \inf_{\beta,\xi,\nu\le\nu_0} \frac{1}{n}\sum_i \mathsf{L}\big(Z_i + f(X_i)\big)\mathbb{I}\{X_i \notin \xi \pm \nu K\} \tag{2.150}$$

$$= \inf_{|I|\le 2\nu_0 K} \frac{1}{n}\sum_i \mathsf{L}\big(Z_i + f(X_i)\big)\mathbb{I}\{X_i \notin I\}, \tag{2.151}$$

where I above is a closed interval of length $|I|$. The inequality relies on the fact that $\beta f((x-\xi)/\nu) = 0$ when $x \notin \xi \pm \nu K$ and the fact that the loss function is non-negative. Now,

$$\frac{1}{n}\sum_i \mathsf{L}\big(Z_i + f(X_i)\big)\mathbb{I}\{X_i \notin I\} = \frac{1}{n}\sum_i g_I(X_i, Z_i), \quad g_I(x,z) := \mathsf{L}\big(z + f(x)\big)\mathbb{I}\{x \notin I\}, \tag{2.152}$$

and the class of functions $\{g_I : |I| \le 2\nu_0 K\}$ is clearly Glivenko–Cantelli. Hence,

$$\inf_{|I|\le 2\nu_0 K} \frac{1}{n}\sum_i \mathsf{L}\big(Z_i + f(X_i)\big)\mathbb{I}\{X_i \notin I\} \xrightarrow{n\to\infty} \inf_{|I|\le 2\nu_0 K} \mathbb{E}[\mathsf{L}\big(Z + f(X)\big)\mathbb{I}\{X \notin I\}]. \tag{2.153}$$

In turn, by dominated convergence, as $\nu_0 \to 0$, the last expression converges to $\mathbb{E}[\mathsf{L}(Z + f(X))]$, which is strictly larger than $\mathbb{E}[\mathsf{L}(Z)]$ by (the equivalent of) Assumption 5. In particular, if ν_0 is so small that the last infimum is $> \mathbb{E}[\mathsf{L}(Z)]$, we guarantee that, with probability tending to 1, $\hat{\nu}_n \ge \nu_0$.

So far, we showed that we can restrict (ξ, ν) in order for f_θ to have support inside that of λ, and that we may add a lower bound on ν (away from 0). It remains to show that, under these conditions, we may also add a bound on $|\beta|$. Fix $\varepsilon > 0$ and define $S_\varepsilon = \{x : |f(x)| > \varepsilon\}$. Then, with

the infimum over ξ and ν restricted as described above, for $\beta_0 > 0$ and $z_0 > 0$, we have

$$
\begin{aligned}
&\inf_{|\beta|\geq\beta_0,\xi,\nu} \widehat{M}_n(\beta,\xi,\nu) \\
&\geq \inf_{|\beta|\geq\beta_0,\xi,\nu} \quad \frac{1}{n}\sum_i \inf_{|a_i|\geq\varepsilon} \mathsf{L}\big(Z_i + f(X_i) - \beta a_i\big)\mathbb{I}\{X_i \in \xi + \nu S_\varepsilon\} + \frac{1}{n}\sum_i \mathsf{L}\big(Z_i + f(X_i)\big)\mathbb{I}\{X_i \notin \xi + \nu S_0\} \\
&\geq \inf_{\xi,\nu} \quad \frac{1}{n}\sum_i \mathsf{L}((\beta_0\varepsilon - Z_i - f(X_i))_+)\mathbb{I}\{X_i \in \xi + \nu S_\varepsilon\} + \frac{1}{n}\sum_i \mathsf{L}\big(Z_i + f(X_i)\big)\mathbb{I}\{X_i \notin \xi + \nu S_0\} \\
&\xrightarrow{n\to\infty} \inf_{\xi,\nu} \quad \mathbb{E}[\mathsf{L}((\beta_0\varepsilon - Z - f(X))_+)\mathbb{I}\{X \in \xi + \nu S_\varepsilon\}] + \mathbb{E}[\mathsf{L}(Z + f(X))\mathbb{I}\{X \notin \xi + \nu S_0\}] \\
&\xrightarrow{\beta_0\to\infty} \inf_{\xi,\nu} \quad \mathsf{L}(\infty)\,\mathbb{P}(X \in \xi + \nu S_\varepsilon) + \mathbb{E}[\mathsf{L}(Z + f(X))\mathbb{I}\{X \notin \xi + \nu S_0\}] \\
&\xrightarrow{\varepsilon\to 0} \inf_{\xi,\nu} \quad \mathsf{L}(\infty)\,\mathbb{P}(X \in \xi + \nu S_0) + \mathbb{E}[\mathsf{L}(Z + f(X))\mathbb{I}\{X \notin \xi + \nu S_0\}].
\end{aligned}
$$

The first limit uses a uniform law of large numbers as we did above, only here it is based on the fact that the collection of sets of the form $\xi + \nu S_\varepsilon$, as ξ and ν vary as they do here, has finite VC dimension, and similarly for the collection of sets of the form $\xi + \nu S_0$. We compare the last expression with $M(\theta^*)$ to get

$$
\begin{aligned}
&\mathsf{L}(\infty)\,\mathbb{P}(X \in \xi + \nu S_0) + \mathbb{E}[\mathsf{L}(Z + f(X))\mathbb{I}\{X \notin \xi + \nu S_0\}] - M(\theta^*) && (2.154)\\
&= (\mathsf{L}(\infty) - \mathbb{E}[\mathsf{L}(Z)])\,\mathbb{P}(X \in \xi + \nu S_0) + \mathbb{E}[\varphi(f(X))\mathbb{I}\{X \notin \xi + \nu S_0\}] =: Q(\xi,\nu), && (2.155)
\end{aligned}
$$

where $\varphi(a) := \mathbb{E}[\mathsf{L}(Z + a)] - \mathbb{E}[\mathsf{L}(Z)]$. By assumption, $\mathsf{L}(\infty) - \mathbb{E}[\mathsf{L}(Z)] > 0$ and $\varphi(a) > 0$ when $a \neq 0$, so that $Q(\xi,\nu) = 0$ if and only if $\mathbb{P}(X \in \xi + \nu S_0) = 0$ and $\mathbb{P}(X \notin \xi + \nu S_0, X \in S_0) = 0$, which together imply that $\mathbb{P}(X \in S_0) = 0$, which would contradict the fact that $f \not\equiv 0$ on the support of λ. Hence, $Q(\xi,\nu) > 0$, and since Q is continuous by dominated convergence, and (ξ,ν) is in a compact set, we have that $\inf_{\xi,\nu} Q(\xi,\nu) > 0$. All in all, we are able to conclude that there is $\beta_0 > 0$ such that, with probability tending to 1, $|\hat{\beta}_n| \leq \beta_0$. And this is the only thing that was left to prove to establish the lemma. □

Chapter 3

Template Matching with Ranks

3.1 Introduction

While in previous chapter we proposed and analyzed M-estimators for template matching, in this chapter we consider R-estimators instead. The former includes what is perhaps the most widely used method which consists in maximizing the Pearson correlation of the signal with a shift of the template; see (3.2) below. We focus here on the rank variant of this approach, which is an example of the latter category and consists, instead, in replacing the signal with the corresponding ranks before maximizing the correlation over shifts of the template; see (3.4).

In the signal and image processing literature per se, where a lot of the work on template matching resides, rank-based methods have been attracting some attention in recent years. [KD09] propose a rank variant of the well-known feature extractor SIFT, while [XXJ+20] propose a rank-based local self-similarity feature descriptor for use in synthetic-aperture radar (SAR) imaging. [AY02] present a face recognition approach using the rank correlation of Gabor filtered images, while [GF16] apply the Spearman rank correlation for template matching in face photo-sketch recognition. [KSC08] construct a matched-filter object detection algorithm based on the Spearman

rank correlation to detect Ca^{2+} sparks in biochemical applications. A number of papers use ranks to align images, a task also known as 'stereo matching' [BB01, BBKC99, CCL12, GG12] — a problem we will not address here but which also has a sizable literature in statistics; we provide some pointers to the literature in our recent paper [ACZ20].

3.1.1 Model and methods

We consider a standard model for template matching, where we observe a shift of the template with additive noise,

$$Y_i = f(x_i - \theta^*) + Z_i, \quad i = 1, \dots, n, \tag{3.1}$$

where $x_i := i/n$ denote the *design points*, $f : [0,1] \to \mathbb{R}$ is a known 1-periodic function referred to as the *template*, and $\theta^* \in \mathbb{R}$ is the unknown *shift* of interest. The *noise* or measurement error variables $Z_1, \dots, Z_n$ are assumed iid with density ϕ and distribution function denoted Φ.

Assumption 6. We assume everywhere that f is 1-periodic, and in fact exactly so in the sense that $f(\cdot - \theta) \neq f(\cdot - \theta^*)$ on a set of positive measure whenever $\theta \neq \theta^* \mod 1$. We also assume that f is Lipschitz continuous.

Assumption 7. We assume everywhere that ϕ is even, so that the noise is symmetric about 0.

Our goal is, in signal processing terminology, to match the template f to the signal Y, which in statistical terms consists in the estimation of the shift θ^*. One of the most popular ways to do so is via maximization of the Pearson correlation, leading to the estimator

$$\hat{\theta} := \arg\max_{\theta} \sum_{i=1}^{n} Y_i f(x_i - \theta). \tag{3.2}$$

In the present context, this is equivalent to the least squares estimator in that the same estimator is

also the solution to the following least squares problem

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (Y_i - f(x_i - \theta))^2. \tag{3.3}$$

Note that this corresponds to the maximum likelihood estimator when the noise distribution is Gaussian. Of course, other loss functions can be used, some of them leading to methods that are robust to noise distributions with heavy tails or to the presence of gross errors (outliers) in the signal. Our recent paper [ACZ20] studies these so-called *M-estimators* in great detail.

In the present paper we consider, instead, estimators based on ranks. These go by the name of *R-estimators* in the statistics literature. The most direct route to such an estimator is to replace the response values with their ranks, yielding

$$\hat{\theta}_{\text{rank}} := \arg\max_{\theta} \sum_{i=1}^{n} R_i f(x_i - \theta), \tag{3.4}$$

where R_i denotes the rank of Y_i in $\{Y_1, \ldots, Y_n\}$ in increasing order. Doing so is sometimes called the 'rank transformation' in the signal processing literature. Note that this is similar in spirit to replacing the Pearson correlation in (3.2) with the Spearman rank correlation, except that we do not rank the values of the template itself. We find that there is no real reason to want to replace the template values with the corresponding ranks as the template is assumed to be free of noise.

3.1.2 Content

The rest of the paper is devoted to studying the asymptotic ($n \to \infty$) properties of the estimator we propose in (3.4), which we will refer to as the R-estimator. In Section 3.2, we derive some basic results describing the asymptotic behavior of this estimator. In more detail, in Section 3.2.1 we discuss the consistency of this estimator, which we are not fully able to establish but is clearly supported by computer simulations; in Section 3.2.2 we derive a rate of

convergence for our R-estimator, which happens to be parametric and also minimax optimal; and in Section 3.2.3 we derive a normal limit distribution of the same estimator, as well as its asymptotic relative efficiency with respect to maximum likelihood estimator under Gaussian noise. Section 3.3 summarizes the result of some numerical experiments we performed to probe our asymptotic theory in finite samples. Section 3.4 provides a discussion of possible extensions. The mathematical proofs are gathered in Section 3.5.

3.2 Theoretical properties

In this section we study the asymptotic properties of the estimator defined in (3.4). We first study its consistency in Section 3.2.1; bound its rate of convergence in Section 3.2.2; and derive its asymptotic (normal) limit distribution in Section 3.2.3.

3.2.1 Consistency

The estimator in (3.4) can be equivalently defined via

$$\hat{\theta}_n := \arg\max_{\theta} \widehat{M}_n(\theta), \qquad \widehat{M}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{n} f(x_i - \theta), \tag{3.5}$$

where we have added the subscript n to emphasize that the estimator is being computed on a sample of size n. To understand the large-sample behavior of this estimator, we need to understand that of $\widehat{M}_n$ as a function of θ, and this leads directly to empirical process theory.

Lemma 13. *In probability,*

$$\sup_{\theta \in \mathbb{R}} \left| \widehat{M}_n(\theta) - M(\theta) \right| \xrightarrow{n \to \infty} 0,$$

where

$$M(\theta) := \int_0^1 \int_0^1 \Phi_2(f(x_0) - f(x)) f(x_0 + \theta^* - \theta) \mathrm{d}x \mathrm{d}x_0, \tag{3.6}$$

with $\Phi_2(t) := \int_{-\infty}^{\infty} \Phi(t+z)\phi(z)\mathrm{d}z$.

With the uniform convergence of $\widehat{M}_n$ to M established in Lemma 13, and with the fact that M is continuous and 1-periodic, it suffices that θ^* be the unique maximizer of M for the estimator $\hat{\theta}_n$ defined in (3.5) to converge to θ^* in probability, that is, to be consistent [VdV98, Th 5.7]. We are able, under an additional mild assumption on the noise distribution, to prove that θ^* is a local maximizer of M.

Proposition 1. *M is twice differentiable, with $M'(\theta^*) = 0$, and if ϕ is positive everywhere, $M''(\theta^*) < 0$, in which case θ^* is a local maximizer of M.*

Unfortunately, we are not able to verify that θ^* is indeed a global maximizer. However, since it is borne out by our numerical experiments (Section 3.3), we conjecture this is true, possibly under some additional (reasonable) conditions on the model. In the meantime, we formulate this as an assumption, which henceforth forms part of our basic assumptions.

Assumption 8. θ^* is the unique maximum point of M defined in (3.6).

Theorem 8. *Under the basic assumptions, $\hat{\theta}_n$ converges in probability to θ^* as $n \to \infty$.*

3.2.2 Rate of convergence and minimaxity

Besides consistency, we derive the estimator's rate of convergence in this section. The rate turns out to be parametric, i.e., the convergence of the estimator to the true value of the parameter is in $O(\sqrt{n})$.

Theorem 9. *Under the basic assumptions, $\hat{\theta}_n$ is $\sqrt{n}$-consistent.*

The parametric rate of $\sqrt{n}$ happens to be minimax optimal in the present setting. This is established in [ACZ20, Cor 3.9] in the context of a random design corresponding to the situation where the design points, instead of being the grid points spanning the unit interval, are generated as an iid sample from the uniform distribution on the unit interval. But similar arguments carry over. We omit details and refer the reader to the discussion in [ACZ20, Sec 6.1].

Corollary 3. *The R-estimator achieves the minimax rate of convergence.*

3.2.3 Limit distribution and asymptotic relative efficiency

In addition to obtaining a rate of convergence, we are also able to derive the asymptotic distribution, which happens to be normal.

Theorem 10. *Under the basic assumptions, strengthened with the assumption that f is continuously differentiable, $\sqrt{n}(\hat{\theta}_n - \theta^*)$ converges weakly to the centered normal distribution with variance $\gamma^2/M''(\theta^*)^2$, where*

$$\gamma^2 := \int_0^1 \int_{-\infty}^{\infty} \left[\int_0^1 (f'(x) - f'(x_0))\, \Xi(f(x) - f(x_0), z) \mathrm{d}x \right]^2 \phi(z) \mathrm{d}z \mathrm{d}x_0,$$

with $\Xi(w,z) := \Phi(z+w) - \Phi_2(w)$, and

$$M''(\theta^*) = -\int_0^1 \int_0^1 f'(x)^2 \phi_2(f(x) - f(x_0)) \mathrm{d}x dx_0,$$

with $\phi_2(t) := \Phi_2'(t) = \int_{-\infty}^{\infty} \phi(z+t)\phi(z)\mathrm{d}z$.

Now that the R-estimator is known to be asymptotically normal with an explicit expression for the asymptotic variance (after standardization), we can consider its (Pitman) efficiency relative to the more popular estimator based on maximizing the Pearson correlation (3.2) (which coincides with the MLE when the noise is Gaussian). Indeed, this estimator (denoted $\tilde{\theta}_n$ now) was studied in [ACZ20, Sec 3] in the setting of a random design. Adapting the arguments there, which are very

similar to (and in fact simpler than) those used here, we find that $\sqrt{n}(\tilde{\theta}_n - \theta^*)$ is asymptotically normal with mean zero and variance $\sigma^2 / \int_0^1 f'(x)^2 \mathrm{d}x$, where σ^2 is the noise variance. With Theorem 10, we are thus able to conclude the following.

Corollary 4. *Suppose that the noise has finite variance* σ^2*. Then the asymptotic efficiency of the R-estimator* (3.4) *relative to the standard estimator* (3.2) *is given by*

$$\frac{\sigma^2 / \int_0^1 f'(x)^2 \mathrm{d}x}{\gamma^2 / M''(\theta^*)^2}.$$

The result, in fact, continues to hold even when the noise has infinite variance, and in that case the asymptotic relative efficiency of the R-estimator relative to the more common estimator is infinite.

3.3 Numerical experiments

We performed some simple numerical experiments to probe the asymptotic theory developed in the previous sections. The implementation of Python codes used in numerical experiments is available on Github: https://github.com/zhenglin0266/Template-Matching-with-Ranks

We note that the implementation of the R-estimator (3.4) is completely straightforward, as it only requires replacing the observations with their ranks before the usual template matching by maximization of the correlation over shifts as in (3.2), which is typically implemented by a fast Fourier transform. This ease of computation is in contrast with rank methods for, say, linear regression which are computationally demanding and require dedicated algorithms and implementations — difficulties that may explain the very limited adoption of such methods in practice.

In our experiments, we considered three noise distributions: Gaussian distribution, Student t-distribution with 3 degrees of freedom, and Cauchy distribution. We took $\theta^* = 0$ throughout,

which is really without loss of generality since the two methods we compared — the standard method based on maximizing the Pearson correlation and the rank-based method that we study — are translation equivariant. We chose to work with the following three filters:

$$\text{Template } A\text{:}\quad f(x) = \begin{cases} 4x-1 & 0.25 \le x < 0.5, \\ 3-4x & 0.5 \le x < 0.75, \\ 0 & \text{otherwise,} \end{cases} \tag{3.7}$$

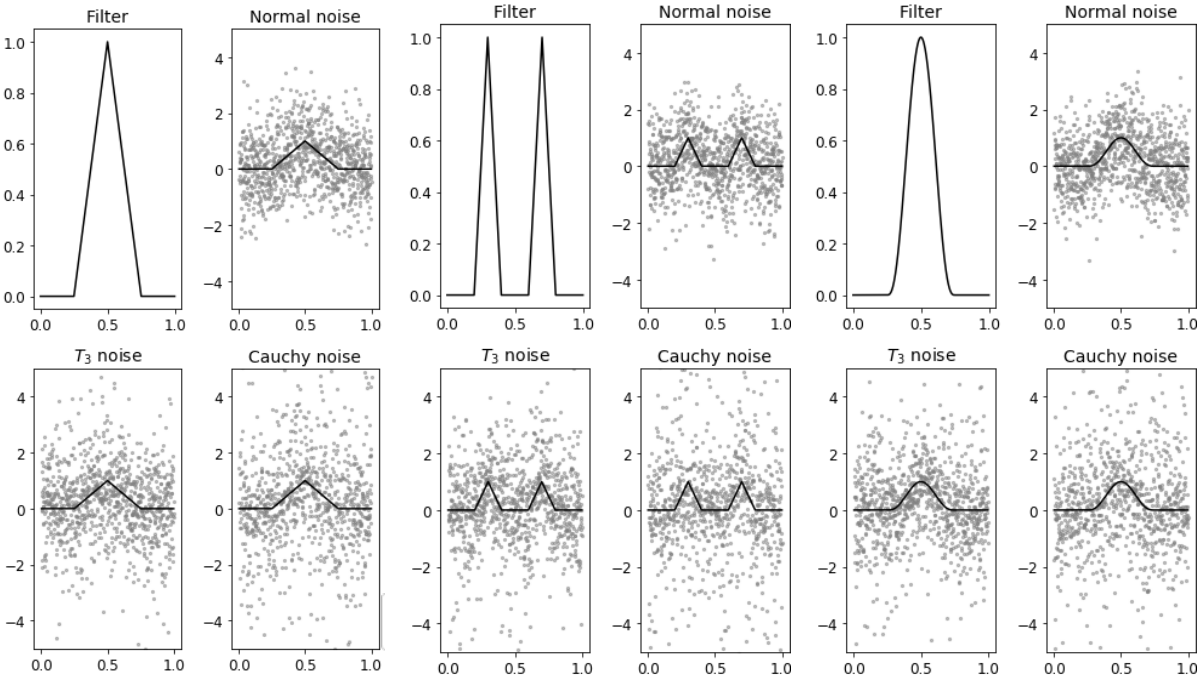
$$\text{Template } B\text{:}\quad f(x) = \begin{cases} 10x-2 & 0.2 \le x < 0.3, \\ 4-10x & 0.3 \le x < 0.4, \\ 10x-6 & 0.6 \le x < 0.7, \\ 8-10x & 0.7 \le x < 0.8, \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

and

$$\text{Template } C\text{:}\quad f(x) = \max\{0, (1-(4x-2)^2)^3\}. \tag{3.9}$$

All are Lipschitz, with Template C being even smoother. See Figure 3.1 for an illustration.

We set the sample size at $n = 10000$. Each setting, defined by a choice of filter and of noise distribution, was repeated 500 times. Box plots of estimation error $|\hat{\theta}_n - \theta^*|$ are presented in Figures 3.2, 3.3 and 3.4, while the distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$ is depicted in Figures 3.5, 3.6 and 3.7 via histograms. The results are congruent with what the theory predicts: The rank-based method is slightly inferior to the standard method when the noise is Gaussian (exactly when the standard method coincides with the MLE), while it is superior when the noise distribution has heavier tails; and the histograms overlay nicely with the predicted asymptotic distribution. The asymptotic relative efficiency of R-estimator relative to the method based on maximizing the



(a) Template A (b) Template B (c) Template C

Figure 3.1: Templates and noisy signals. Although the sample size is $n = 10000$, for the sake of clarity, we only include 1000 points and limit the range of the y-axis to $[-5,5]$.

Pearson correlation is displayed in Table 3.1.

Table 3.1: Asymptotic relative efficiency of the R-estimator (3.4) to the more common estimator (3.2). Note that the latter is asymptotically best in the setting of Gaussian noise as it then coincides with the maximum likelihood estimator for a 'smooth' model.

Template	Noise		
	Normal	Student t_3	Cauchy
Template A	0.949	2.008	∞
Template B	0.940	1.992	∞
Template C	0.948	2.008	∞

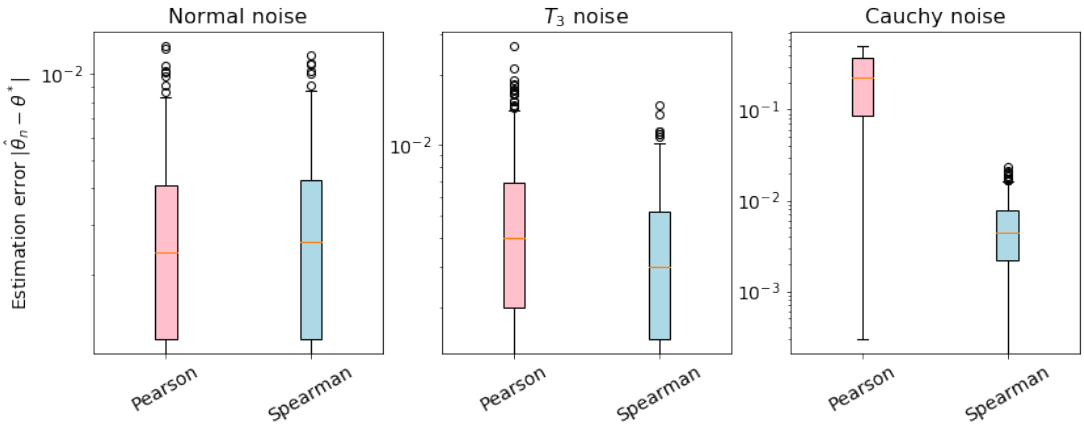


Figure 3.2: Box plot of estimation error $|\hat{\theta}_n - \theta^*|$ for Template A

3.4 Discussion

Our main goal in this paper was to show that a standard rank-based approach to template matching is viable and amenable to study using well-established techniques in mathematical statistics — some basic results in empirical process theory and the projection method of Hájek. This provides some theoretical foundation for related approaches proposed in recent years in the signal processing literature. We chose to keep the exposition contained and, in particular, have focused on the 'smooth setting' of a Lipschitz template. We leave the equally important case of a discontinuous template for future work (likely by others). Based on our previous work [ACZ20] —

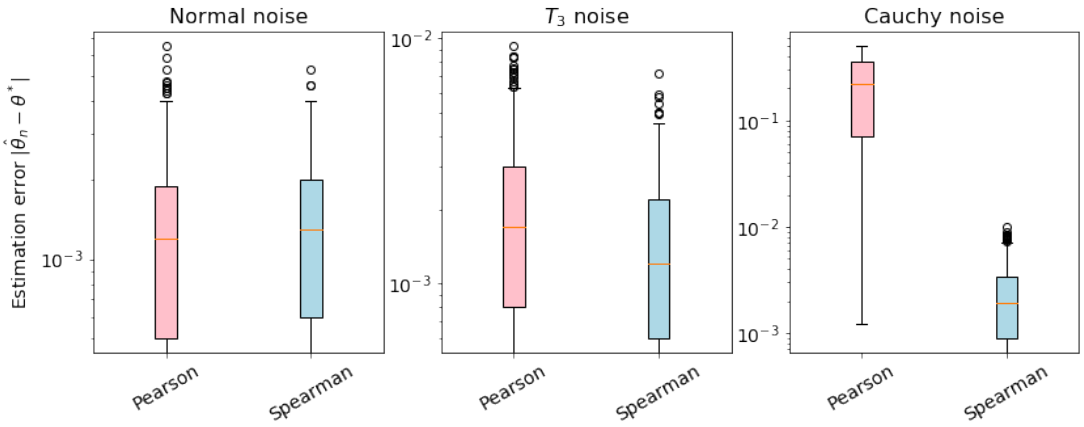


Figure 3.3: Box plot of estimation error $|\hat{\theta}_n - \theta^*|$ for Template B

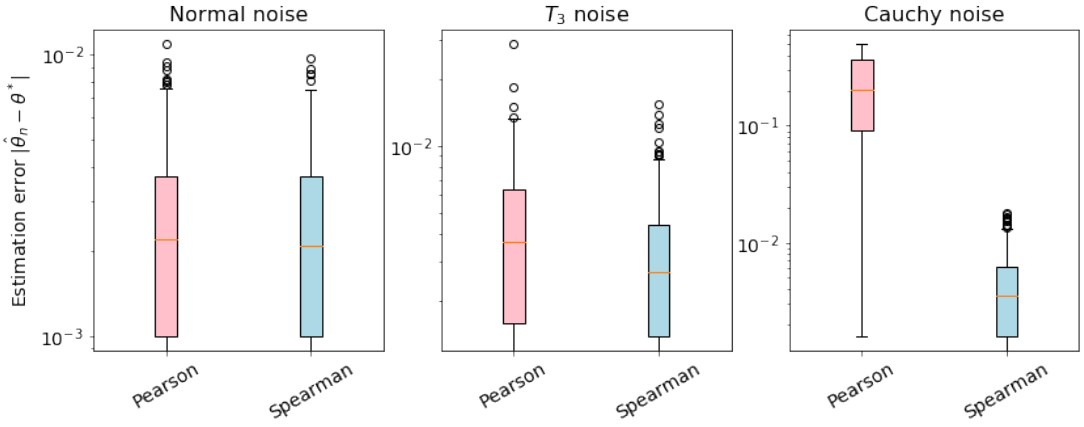
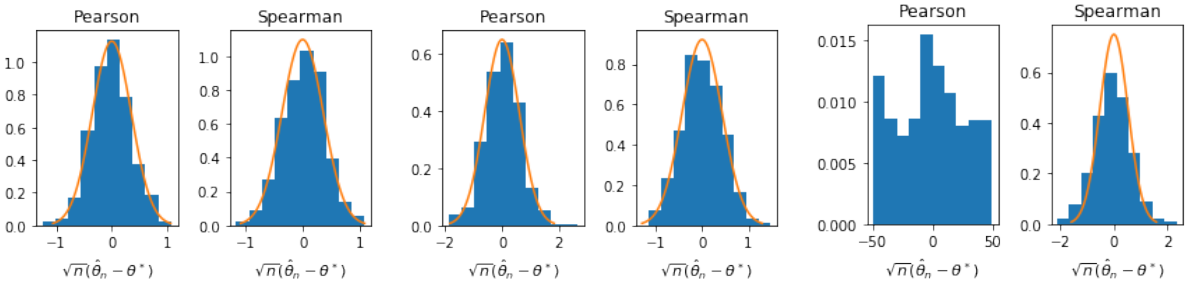
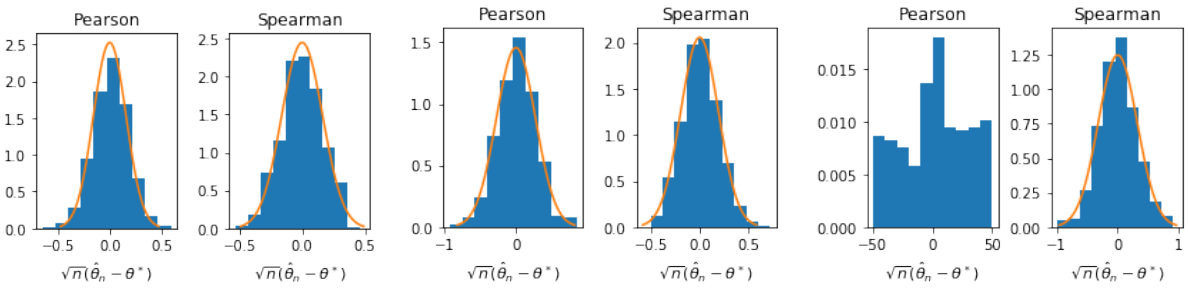


Figure 3.4: Box plot of estimation error $|\hat{\theta}_n - \theta^*|$ for Template C



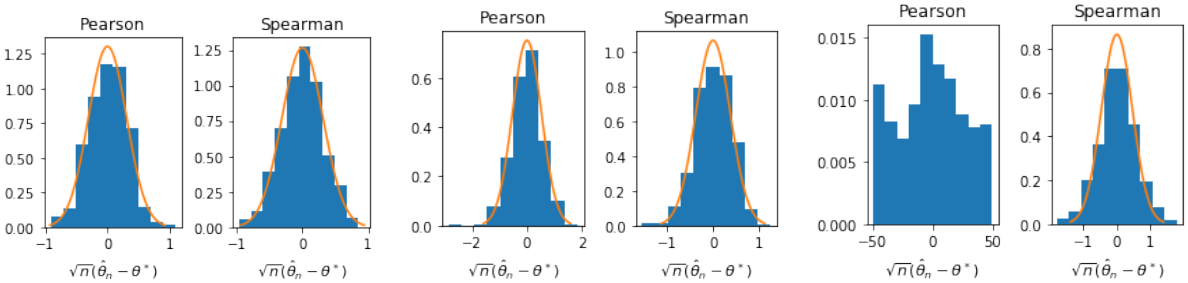
(a) Normal noise (b) T_3 noise (c) Cauchy noise

Figure 3.5: Distribution under Template A. The histogram presents the distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$. The orange bell-shaped curve is the density of normal distribution predicted by the theory.



(a) Normal noise (b) T_3 noise (c) Cauchy noise

Figure 3.6: Distribution under Template B. The histogram presents the distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$. The orange bell-shaped curve is the density of normal distribution predicted by the theory.



(a) Normal noise (b) T_3 noise (c) Cauchy noise

Figure 3.7: Distribution under Template C. The histogram presents the distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$. The orange bell-shaped curve is the density of normal distribution predicted by the theory.

where we did study this case in detail — and on our work here — in particular Hájek's projection technique used in the proof of Theorem 10 — we do believe that the study of this case is within the reach of similarly standard tools.

Some of the other extensions discussed in our previous work are also relevant here. In particular, while we focused on shifts in the context of 1D signals, other settings are possible, including shifts in 2D or 3D signals (i.e., images), as well as other transformations. We omit details and simply affirm that such extensions are also amenable to a similar mathematical analysis.

In signal processing, rank-based methods seem more prominently represented in the literature on signal registration. We are confident that the study of such methods is well within the range of established techniques in mathematical statistics. However, the situation becomes substantially more complex as 1) the setting is semi-parametric, and 2) some smoothing seems to be required to achieve good performance — as transpires from the statistics literature on the topic as mentioned in our previous work [ACZ20]. We leave a further exploration of rank-based methods for registration to future endeavors.

3.5 Proofs

3.5.1 Preliminaries

The following is an extension of the celebrated Glivenko–Cantelli theorem.

Lemma 14. *Suppose that* $\{Y_{i,n} : i \in [n], n \geq 1\}$ *are independent random variables with uniformly tight and equicontinuous distributions* $\{F_{i,n} : i \in [n], n \geq 1\}$. *Define*

$$\hat{F}_{[n]}(y) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_{i,n} \leq y\}, \qquad F_{[n]} := \mathbb{E}[\hat{F}_{[n]}] = \frac{1}{n}\sum_{i=1}^{n} F_{i,n}.$$

Then the following uniform convergence holds in probability

$$\sup_t |\hat{F}_{[n]}(t) - F_{[n]}(t)| \xrightarrow{n\to\infty} 0.$$

Proof. The proof arguments are very close to those supporting the more classical situation in which the random variables have the same distribution [VdV98, Th 19.1]. We provide a proof for completeness only.

Our assumptions mean 1) that for each $\varepsilon > 0$ there is K such that $\inf_i F_{i,n}(K) \geq 1-\varepsilon$ and $\sup_i F_{i,n}(-K) \leq \varepsilon$; and 2) that $\omega(\delta) := \sup_i \sup_t (F_{i,n}(t+\delta) - F_{i,n}(t))$ is continuous at 0; we then speak of ω as the modulus of continuity of the family $\{F_{i,n}\}$. Fix $\varepsilon > 0$ and let K be defined as above. Also, let δ be such that $\omega(\delta) \leq \varepsilon$ and set $-K = t_1 < t_2 < \cdots < t_{m-1} < t_m = K$ be such that $t_{j+1} - t_j \leq \delta$ for all j. We then have, regardless of n, that $F_{[n]}(K) \geq 1-\varepsilon$ and $F_{[n]}(-K) \leq \varepsilon$; and that ω is a modulus of continuity for $F_{[n]}$, meaning that $\sup_t (F_{[n]}(t+\delta) - F_{[n]}(t)) \leq \omega(\delta)$.

- For $t \leq t_1$, we have

$$|\hat{F}_{[n]}(t) - F_{[n]}(t)| \leq \hat{F}_{[n]}(t) + F_{[n]}(t) \leq \hat{F}_{[n]}(t_1) + \varepsilon.$$

- For $t \geq t_m$, we have

$$|\hat{F}_{[n]}(t) - F_{[n]}(t)| \leq (1 - \hat{F}_{[n]}(t)) + (1 - F_{[n]}(t)) \leq (1 - \hat{F}_{[n]}(t_m)) + \varepsilon.$$

- For $-K < t < K$, if j is such that $t_j < t \leq t_{j+1}$,

$$\hat{F}_{[n]}(t) - F_{[n]}(t) \leq \hat{F}_{[n]}(t_{j+1}) - F_{[n]}(t_j) \leq \hat{F}_{[n]}(t_{j+1}) - F_{[n]}(t_{j+1}) + \varepsilon,$$

and

$$\hat{F}_{[n]}(t) - F_{[n]}(t) \geq \hat{F}_{[n]}(t_j) - F_{[n]}(t_{j+1}) \leq \hat{F}_{[n]}(t_j) - F_{[n]}(t_j) - \varepsilon.$$

By Chebyshev's inequality, $\hat{F}_{[n]}(t) - F_{[n]}(t) \to 0$ in probability as $n \to \infty$ for every fixed $t \in \mathbb{R}$. In particular, $\max_j |\hat{F}_{[n]}(t_j) - F_{[n]}(t_j)| \to 0$ in probability, and under the event that this maximum is bounded by ε, we have $|\hat{F}_{[n]}(t) - F_{[n]}(t)| \leq 2\varepsilon$. Since $\varepsilon > 0$ is arbitrary, the proof is complete. □

The following is a simple result on functions defined as the linear combination of uniformly equicontinuous functions with random coefficients.

Lemma 15. *Suppose that* $\{B_{i,n} : i \in [n], n \geq 1\}$ *are independent such that* $|B_{i,n}| \leq K$ *and* $\mathbb{E}[B_{i,n}] = 0$ *for all* $i \in [n]$ *and all* $n \geq 1$*; and that* $\{f_{i,n} : i \in [n], n \geq 1\}$ *are uniformly bounded and uniformly equicontinuous functions either defined on a compact interval. Then, in probability,*

$$\left| \frac{1}{n} \sum_{i=1}^{n} B_{i,n} f_{i,n} \right|_\infty \xrightarrow{n \to \infty} 0.$$

Proof. The arguments are quite similar to those supporting Lemma 14. Define

$$S_n := \frac{1}{n} \sum_{i=1}^{n} B_{i,n} f_{i,n}.$$

Suppose without loss of generality that the functions are defined on the unit interval. Because they are uniformly equicontinuous, for any given $\varepsilon > 0$, there is $\delta > 0$ such that $\sup_n \sup_i |f_{i,n}(t) - f_{i,n}(s)| \leq \varepsilon$ for all $s, t \in [0,1]$ such that $|t - s| \leq \delta$. With $\varepsilon > 0$ fixed, and δ as such, let $0 = t_1 < t_2 < \cdots < t_{m-1} < t_m = 1$ be such that $t_{j+1} - t_j \leq \delta$ for all j. Then for any $t \in [0,1]$, if j is such that $t_j \leq t \leq t_{j+1}$,

$$|S_n(t) - S_n(t_j)| \leq \frac{1}{n} \sum_{i=1}^{n} |B_{i,n}||f_{i,n}(t) - f_{i,n}(t_j)| \leq K \sup_n \sup_i |f_{i,n}(t) - f_{i,n}(t_j)| \leq K\varepsilon.$$

In particular,

$$|S_n|_\infty \le \max_j |S_n(t_j)| + K\varepsilon.$$

Furthermore, by Chebyshev's inequality, in probability,

$$\max_j |S_n(t_j)| \xrightarrow{n\to\infty} 0.$$

Hence, in probability,

$$\limsup_n |S_n|_\infty \le K\varepsilon.$$

Since $\varepsilon > 0$ was chosen arbitrary, the proof is complete. □

The following is a well-known error bound for Riemann sums.

Lemma 16. *Assume that $f : [a,b] \to \mathbb{R}$ is continuous with modulus of continuity ω. Then*

$$\left| \int_a^b f(u)\mathrm{d}u - \frac{1}{m}\sum_{j=1}^{m} f(a + j(b-a)/m) \right| \le (b-a)\omega((b-a)/m)$$
$$\le \frac{(b-a)^2}{m}|f'|_\infty \quad \textit{if } f \textit{ is Lipschitz.}$$

Proof. This well-known result is a simple consequence of partitioning $[a,b]$ into sub-intervals of length $(b-a)/m$. □

The next two lemmas are refinements of Lemma 14 and Lemma 15. They clearly subsume them, but they are also much deeper, and we only provide proof sketches, relying on arguments borrowed from [VdV98].

Lemma 17. *In the context of Lemma 14, for some constant C,*

$$\mathbb{E}\Big[\sup_t |\hat{F}_{[n]}(t) - F_{[n]}(t)|\Big] \le C/\sqrt{n}.$$

Proof. The result is classical when the variables are not only independent, but also identically distributed, say $Y_{i,n} \sim F$ for all i and all n, and is a special case of so-called entropy bounds on the supremum of an empirical processes of the form

$$S_n := \sup_{g\in\mathcal{G}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (g(Y_{i,n}) - \mathbb{E}[g(Y_{i,n})]). \tag{3.10}$$

For example, assuming that the class $\mathcal{G}$ is uniformly bounded, Cor 19.35 in [VdV98] gives

$$\mathbb{E}[S_n] \le C_0 J(\mathcal{G}, F),$$

where C_0 is a constant and

$$J(\mathcal{G}, F) := \int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{G}, L^2(F))}\mathrm{d}\varepsilon,$$

$N(\varepsilon, \mathcal{G}, L^2(F))$ denoting the ε-bracketing number of the class $\mathcal{G}$ with respect to the $L^2(F)$ metric.[1]

The proof of that result takes several pages, but a close examination reveals that the 'identically distributed' property is not used in an essential way. Indeed, the assumption that the variables are iid is only used when applying Bernstein's concentration inequality (Lem 19.32 there), and it is well-known that the result applies in a generalized form to variables that are only independent, say $Y_{i,n} \sim F_{i,n}$. Everything follows from that, essentially verbatim, and yields

$$\mathbb{E}[S_n] \le C_0 J(\mathcal{G}, F_{[n]}).$$

[1] Two functions g_1, g_2 such that $g_1 \le g_2$ pointwise define a bracket made of all functions g such that $g_1 \le g \le g_2$. It is said to be an ε-bracket with respect to $L^2(F)$, for a positive measure F, if $\int (g_2 - g_1)^2 \mathrm{d}F \le \varepsilon^2$. Given a class of functions $\mathcal{G}$, its ε-bracketing number with respect to $L^2(F)$ is the minimum number of ε-brackets needed to cover $\mathcal{G}$.

It turns out that, for a given distribution function F, $J(\mathcal{G}, F)$ can be bounded based on the modulus of continuity of F (see Ex 19.6 in the same reference). And if ω is the modulus of continuity of $\{F_{i,n}\}$, then it is also a modulus of continuity for $F_{[n]}$, and with this we can bound $J(\mathcal{G}, F_{[n]})$ independently of n just based on ω.

When dealing with the empirical distribution function, which is our focus here, the class is taken to be $\mathcal{G} := \{\mathbb{I}\{y \le t\} : t \in \mathbb{R}\}$. For that class, $J(\mathcal{G}, F) < \infty$ for any distribution function, and this implies via the arguments above that $\sup_n J(\mathcal{G}, F_{[n]}) < \infty$, concluding the proof. □

Lemma 18. *Suppose that $\{B_i : i \ge 1\}$ are independent random variables that are centered and bounded in absolute value by K. And let $\{f_i : i \ge 1\}$ be L-Lipschitz functions on $[-t_0, t_0]$ with $f_i(0) = 0$ for all i. Then there is a some constant C such that, for any $n \ge 1$,*

$$\mathbb{E}\left[\sup_{|t| \le t_0} \left|\frac{1}{n}\sum_{i=1}^{n} B_i f_i(t)\right|\right] \le \frac{Ct_0}{\sqrt{n}}.$$

Proof. As in the proof of Lemma 17, we rely on entropy bounds. Here we use Dudley's entropy bound as presented in [GN16, Th 2.3.6]. For a given n, let $S(t) := \frac{1}{n}\sum_{i=1}^{n} B_i f_i(t)$, we have

$$S(t) - S(s) = \sum_{i=1}^{n} \frac{B_i}{n}(f_i(t) - f_i(s)), \tag{3.11}$$

with the variables $\frac{B_i}{n}(f_i(t) - f_i(s))$ being independent, centered, and bounded in absolute value by $(K/n)L|t-s|$. In [GN16], by Eq (3.8) and based on Def 2.3.5, the process $S(t)$ is sub-Gaussian on $[-t_0, t_0]$ with respect to the metric $\mathrm{d}(s,t) := (KL/\sqrt{n})|s-t|$. Because $S(0) = 0$, Th 2.3.6 there gives that

$$\mathbb{E}\Big[\sup_{|t| \le t_0} |S(t)|\Big] \le 4\sqrt{2}\int_0^{D/2} \sqrt{\log(2N(\varepsilon))}\mathrm{d}\varepsilon,$$

where D and $N(\varepsilon)$ are the diameter and ε-covering number of $[-t_0, t_0]$ with respect to d. Immediately, $D = (KL/\sqrt{n})(2t_0) = 2KLt_0/\sqrt{n}$, and $N(\varepsilon) \asymp (KL/\sqrt{n})(t_0/\varepsilon) \asymp D/\varepsilon$. With a simple change

of variable in the integral, this gives us

$$\mathbb{E}\Big[\sup_{|t|\le t_0}|S(t)|\Big] \le C_1 K L t_0/\sqrt{n},$$

for a universal constant C_1. □

3.5.2 Proof of Lemma 13

We assume without loss of generality that $\theta^* = 0$. We have

$$R_i = \sum_{j=1}^n \mathbb{I}\{Y_j \le Y_i\} = n\hat{\Psi}_n(Y_i), \quad \text{where } \hat{\Psi}_n(y) := \frac{1}{n}\sum_{j=1}^n \mathbb{I}\{Y_j \le y\}.$$

Note that $\hat{\Psi}_n$ is the empirical distribution function of $Y_1, \dots, Y_n$. Although these are not iid, they are independent, and the Glivenko–Cantelli theorem applies to give that, in probability,

$$\sup_{y\in\mathbb{R}} |\hat{\Psi}_n(y) - \mathbb{E}[\hat{\Psi}_n(y)]| \xrightarrow{n\to\infty} 0.$$

See Lemma 14 for details. Further, we have

$$\begin{aligned}
\mathbb{E}[\hat{\Psi}_n(y)] &= \frac{1}{n}\sum_{j=1}^n \mathbb{P}(Y_j \le y) \\
&= \frac{1}{n}\sum_{j=1}^n \mathbb{P}(Z_j \le y - f(x_j)) \\
&= \frac{1}{n}\sum_{j=1}^n \Phi(y - f(j/n)) \\
&\xrightarrow{n\to\infty} \Psi(y) := \int_0^1 \Phi(y - f(x))\mathrm{d}x,
\end{aligned}$$

where the convergence is by definition of the Riemann integral defining the limit. The convergence is in fact uniform in y. This comes from an application of Lemma 16 using with the fact that $x \mapsto \Phi(y - f(x))$ has derivative $f'(x)\phi(y - f(x))$, which has supremum norm bounded by

$|f'|_\infty |\phi|_\infty < \infty$ (independent of y). Hence, a simple application of the triangle inequality gives that, in probability,

$$A_{1,n} := \sup_{y \in \mathbb{R}} |\hat{\Psi}_n(y) - \Psi(y)| \xrightarrow{n \to \infty} 0. \tag{3.12}$$

This is useful to us because $R_i = n\Psi(Y_i) \pm nA_{1,n}$, which then triggers

$$\widehat{M}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \Psi(Y_i) f(x_i - \theta) \pm A_{2,n},$$

with $A_{2,n} := A_{1,n}|f|_\infty = o_P(1)$. We may thus focus on the first term on the right-hand side. We have

$$\frac{1}{n} \sum_{i=1}^{n} \Psi(Y_i) f(x_i - \theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\Psi(Y_i)] f(x_i - \theta) + Q_n(\theta),$$

with

$$Q_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} (\Psi(Y_i) - \mathbb{E}[\Psi(Y_i)]) f(x_i - \theta).$$

On the one hand, by a standard argument consisting in discretizing the values of θ and using the uniform continuity of f, we obtain

$$A_{3,n} := \sup_{\theta \in \mathbb{R}} |Q_n(\theta)| \xrightarrow{n \to \infty} 0,$$

in probability. See Lemma 15 for details. On the other hand,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Psi(Y_i)]f(x_i-\theta) &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\Psi(z+f(i/n))\phi(z)\mathrm{d}z\cdot f(i/n-\theta)\\ &\xrightarrow{n\to\infty}\int_0^1\int_{-\infty}^{\infty}\Psi(z+f(x))\phi(z)\mathrm{d}z\cdot f(x-\theta)\mathrm{d}x\\ &= \int_0^1\int_0^1\Phi_2(f(x)-f(t))f(x-\theta)\mathrm{d}t\mathrm{d}x = M(\theta),\end{aligned}$$

using again the definition of Riemann integral. In fact the convergence is uniform in θ, by an application of Lemma 16 to the function

$$g_\theta(x) := \int_{-\infty}^{\infty}\Psi(z+f(x))\phi(z)\mathrm{d}z\cdot f(x-\theta),$$

whose derivative can be bounded independently of θ as follows:

$$\begin{aligned}|g'_\theta(x)| &= \left|\int_{-\infty}^{\infty}f'(x)\Psi'(z+f(x))\phi(z)\mathrm{d}z\cdot f(x-\theta)+\int_{-\infty}^{\infty}\Psi(z+f(x))\phi(z)\mathrm{d}z\cdot f'(x-\theta)\right|\\ &\le |f'|_\infty|\Psi'|_\infty|f|_\infty+|\Psi|_\infty|f'|_\infty\\ &\le |f'|_\infty|\phi|_\infty|f|_\infty+|f'|_\infty,\end{aligned}$$

using the fact that ϕ is a density, that Ψ is a distribution function, and that $\Psi'(y)=\int_0^1\phi(y-f(x))\mathrm{d}x$ is non-negative and bounded by $|\phi|_\infty$. All combined, we can conclude that $\widehat{M}_n(\theta)$ indeed converges in probability as $n\to\infty$ to $M(\theta)$ uniformly in θ.

3.5.3 Proof of Proposition 1

Assume without loss of generality that $\theta^*=0$. Define

$$g(y) := \int_0^1\Phi_2(y-f(x))\mathrm{d}x,$$

so that

$$M(\theta) = \int_0^1 g(f(t))f(t-\theta)\mathrm{d}t.$$

Note that $0 \le g(y) \le 1$ for all y. When f is Lipschitz, it is absolutely continuous with bounded derivative, so that by dominated convergence, M is differentiable with derivative

$$M'(\theta) = -\int_0^1 g(f(t))f'(t-\theta)\mathrm{d}t = -\int_0^1 g(f(t+\theta))f'(t)\mathrm{d}t.$$

The reason we transferred θ to g is to be able to differentiate again. Indeed, g is also differentiable by dominated convergence, with derivative (recall that $\Phi' = \phi$)

$$g'(y) = \int_0^1 \int_{-\infty}^{\infty} \phi(z+y-f(x))\phi(z)\mathrm{d}z\mathrm{d}x,$$

which is bounded, so that M' in its second form is also differentiable (by dominated convergence again), with derivative

$$M''(\theta) = -\int_0^1 f'(t+\theta)g'(f(t+\theta))f'(t)\mathrm{d}t.$$

Therefore, M'' is twice differentiable (with bounded second derivative at that).

We now look at $\theta = 0$. Let G be the indeterminate integral of g. We have

$$\begin{aligned} M'(0) &= -\int_0^1 g(f(t))f'(t)\mathrm{d}t \\ &= -[G(f(1)) - G(f(0))] = 0, \end{aligned}$$

by the fact that f is 1-periodic. We also have

$$M''(0) = -\int_0^1 g'(f(t))f'(t)^2 \mathrm{d}t \leq 0,$$

by the fact that g' is non-negative (by simply looking at the integrand defining it, recalling that ϕ is a density). In fact the inequality is strict by our assumption on ϕ, as it forces g' to be strictly positive everywhere.

3.5.4 Proof of Theorem 9

When working with the raw responses $Y_1, \ldots, Y_n$, the result can be proved using [VdV98, Th 5.52]. Since we work with the ranks $R_1, \ldots, R_n$ instead, we elaborate, even though the core arguments are essentially the same. Assume without loss of generality that $\theta^* = 0$, so that we need to show that $\sqrt{n}\hat{\theta}_n$ is bounded in probability.

On the one hand, by definition, we have $\widehat{M}_n(\hat{\theta}_n) - \widehat{M}_n(0) \geq 0$. On the other hand, by consistency (since Theorem 8 applies), we have that $|\hat{\theta}_n|$ is small, and by the fact that M is close to quadratic in the neighborhood of 0 (Proposition 1), we have that $M(\hat{\theta}_n) - M(0) \leq -C_1\hat{\theta}_n^2$ for some constant $C_1 > 0$. Combined, these two observations yield

$$C_1\hat{\theta}_n^2 \leq \widehat{M}_n(\hat{\theta}_n) - M(\hat{\theta}_n) - (\widehat{M}_n(0) - M(0)),$$

with probability tending to 1.

Let $f_i(\theta) := f(x_i - \theta) - f(x_i)$. For any θ, we have

$$\widehat{M}_n(\theta) - M(\theta) - (\widehat{M}_n(0) - M(0)) \tag{3.13}$$

$$= \frac{1}{n}\sum_i \big(\hat{\Psi}_n(Y_i) - \mathbb{E}[\hat{\Psi}_n](Y_i)\big) f_i(\theta) \tag{3.14}$$

$$+ \frac{1}{n}\sum_i \big(\mathbb{E}[\hat{\Psi}_n](Y_i) - \Psi(Y_i)\big) f_i(\theta) \tag{3.15}$$

$$+ \frac{1}{n}\sum_i \big(\Psi(Y_i) - \mathbb{E}[\Psi(Y_i)]\big) f_i(\theta) \tag{3.16}$$

$$+ \frac{1}{n}\sum_i \mathbb{E}[\hat{\Psi}(Y_i)] f_i(\theta) - (M(\theta) - M(0)). \tag{3.17}$$

We saw in the proof of Lemma 13 that the terms in (3.15) and (3.17) are Riemannian sums and at most of order $O(1/n)$ uniformly in θ given that the $|f_i|_\infty \le 2|f|_\infty$. (This is crude, but enough for our purposes here.) For (3.14), we apply Lemma 17 together with Markov's inequality to get that, in probability as $n \to \infty$,

$$\sup_{y\in\mathbb{R}} |\hat{\Psi}_n(y) - \mathbb{E}[\hat{\Psi}_n](y)| \le C_2/\sqrt{n}.$$

With this, and the fact that $|f_i(\theta)| \le |f'|_\infty |\theta|$, we have that the quantity in (3.14) is bounded in absolute value by $(C_2/\sqrt{n})|f'|_\infty|\theta|$ for all θ, that is, this term is $O(|\theta|/\sqrt{n})$ uniformly in θ. Hence, if $S_n(\theta)$ denotes the term in (3.16), we have with probability tending to 1,

$$C_1\hat{\theta}_n^2 \le S_n(\hat{\theta}_n) + C_3(|\hat{\theta}_n|/\sqrt{n} + 1/n).$$

Let $C_4 > 0$ be such that $C_1\theta^2 - C_3(|\theta|/\sqrt{n} + 1/n) \ge \theta^2/C_4$ whenever $|\theta| \ge C_4/\sqrt{n}$, so that $S_n(\hat{\theta}_n) \ge \hat{\theta}_n^2/C_4$ when $\sqrt{n}|\hat{\theta}_n| \ge C_4$. Let J_0 be the smallest integer such that $2^{J_0} \ge C_4$. Then, for

$J \geq J_0$, we have

$$\begin{aligned}
\mathbb{P}(\sqrt{n}|\hat{\theta}_n| \geq 2^J) &= \sum_{j \geq J} \mathbb{P}(2^j < \sqrt{n}|\hat{\theta}_n| \leq 2^{j+1}) \\
&\leq \sum_{j \geq J} \mathbb{P}\Big(\max_{\sqrt{n}|\theta| \leq 2^{j+1}} S_n(\theta) \geq (2^j/\sqrt{n})^2/C_4 \Big) \\
&\leq \sum_{j \geq J} \frac{(C_5/\sqrt{n})(2^{j+1}/\sqrt{n})}{(2^j/\sqrt{n})^2/C_4} \\
&= C_6 2^{-J} \xrightarrow{J \to \infty} 0,
\end{aligned}$$

where C_5 is the constant of Lemma 18, and we used that lemma and Markov's inequality in the corresponding line. We can thus conclude that $\sqrt{n}\hat{\theta}_n$ is bounded in probability.

3.5.5 Proof of Theorem 10

We assume without loss of generality that $\theta^* = 0$. The derivation of the limiting distribution of the R-estimator follows via an application of the so-called argmax theorem. This standard route is described, for example, in [VdV98, Sec 5.9]. It goes like this. By a simple change of variables and by definition of $\hat{\theta}_n$, $h_n := \sqrt{n}\hat{\theta}_n$ maximizes

$$W_n(h) := r_n\big[\widehat{M}_n(h/\sqrt{n}) - \widehat{M}_n(0)\big].$$

This is true for any $r_n > 0$ and, with probability tending to one according to Theorem 9, it is true even if W_n is restricted to $[-a_n, a_n]$ for any given sequence $a_n \to \infty$. Suppose there is a choice of r_n that leads to the weak convergence of W_n to W in some appropriate sense, where W has a unique maximizer. Then it is reasonable to anticipate that h_n will converge to that maximizer in some way. This is indeed the case under some mild assumptions. The following is a special case of [VdV98, Cor 5.58].

Lemma 19. *Suppose that a sequence of processes W_n defined on $[-a_n, a_n]$ for some sequence*

$a_n \to \infty$ converges weakly in the uniform topology on every fixed compact interval to a process W with continuous sample paths each having a unique maximum point h^* (almost surely). If h_n maximizes W_n, and (h_n) is uniformly tight, then h_n converges weakly to h^*.

Back to our situation, we have established the tightness of $\{h_n\}$ in Theorem 9. It therefore remains to show that W_n converges weakly to an appropriate stochastic process for a proper choice of r_n. We will see that $r_n := n$ is the correct choice (up to an arbitrary multiplicative factor) and that the limit process is a simple Gaussian process. In what follows, we let $a_n \to \infty$ slowly (e.g., $a_n = \log n$).

So far, we have worked with the ranks using rather elementary means, but now we turn to more sophisticated tools. Specifically, we use the projection method of Hájek. The following is a special case of [Hó8, Th 4.2] with some minor modifications.

Lemma 20. *Suppose $Y_1, \dots, Y_n$ are independent with respective distribution functions $F_1, \dots, F_n$. Define $M = \sum_i b_i R_i / n$, where $R_1, \dots, R_n$ are the respective ranks of $Y_1, \dots, Y_n$, and $b_1, \dots, b_n$ are reals. Then, for a universal constant C,*

$$\mathbb{E}\Big[\Big(M - \mu - \sum_{i=1}^n V_i\Big)^2\Big] \le \frac{C}{n} \sum_{i=1}^n b_i^2,$$

where

$$\mu := \sum_{i=1}^n b_i \frac{1}{n} \sum_{j=1}^n \int F_j(t) \mathrm{d}F_i(t), \qquad V_i := \frac{1}{n} \sum_{i=1}^n (b_j - b_i) \int \big[\mathbb{I}\{Y_i \le t\} - F_i(t)\big] \mathrm{d}F_j(t).$$

This result thus provides an approximation of a linear combination of ranks (which are dependent) by a linear combination of independent random variables, and the latter is essentially ready for an application of a central limit theorem. We apply it to

$$W_n(h) = \sum_{i=1}^n f_i(h) \frac{R_i}{n}, \quad f_i(h) := f(x_i - h/\sqrt{n}) - f(x_i).$$

Note that f_i depends on n and recall that $x_i = i/n$. In Lemma 20, b_i corresponds here to $f_i(h)$ and F_i to $\Phi(\cdot - f(x_i))$. Hence, μ in the lemma is given by

$$\begin{aligned}
&\sum_{i=1}^{n} f_i(h) \frac{1}{n} \sum_{j=1}^{n} \int_{-\infty}^{\infty} \Phi(t - f(x_j)) \phi(t - f(x_i)) \mathrm{d}t \\
&= \sum_{i=1}^{n} f_i(h) \frac{1}{n} \sum_{j=1}^{n} \Phi_2(f(x_i) - f(x_j)) \\
&= \sum_{i=1}^{n} f_i(h) \left[\int_0^1 \Phi_2(f(x_i) - f(x)) \mathrm{d}x \pm \frac{|f'|_\infty}{2n} \right] \\
&= \sum_{i=1}^{n} (f(x_i - h) - f(x_i)) \int_0^1 \Phi_2(f(x_i) - f(x)) \mathrm{d}x \pm n \frac{|f'|_\infty a_n}{\sqrt{n}} \frac{|f'|_\infty}{n}.
\end{aligned}$$

In the 3rd equality, we used Lemma 16 and the fact that $x \mapsto \Phi_2(f(x_i) - f(x))$ has derivative $-f'(x)\phi_2(f(x_i) - f(x))$, whose supnorm is bounded by $|f'|_\infty$. Defining

$$g_h(t) := (f(t-h) - f(t)) \int_0^1 \Phi_2(f(t) - f(x)) \mathrm{d}x,$$

we have

$$\begin{aligned}
&\sum_{i=1}^{n} (f(x_i - h) - f(x_i)) \int_0^1 \Phi_2(f(x_i) - f(x)) \mathrm{d}x \\
&= \sum_{i=1}^{n} g_h(x_i) \\
&= n \left[\int_0^1 g_h(t) \mathrm{d}t \pm \frac{|g_h'|_\infty}{n} \right] \\
&= n \left[M(h/\sqrt{n}) - M(0) \right] \pm \omega_1(a_n/\sqrt{n}),
\end{aligned}$$

using Lemma 16 in the 3rd equality, and where $\omega_1(\varepsilon) := \sup_t \sup_{|s| \le \varepsilon} |f'(t+s) - f'(t)|$, which is the modulus of continuity of f'. Note that $\omega_1(\varepsilon) \to 0$ when $\varepsilon \to 0$ by the fact that f is assumed to

be continuously differentiable (and 1-periodic). Therefore, μ in the lemma is equal to

$$n\big[M(h/\sqrt{n}) - M(0)\big] \pm \omega_1(a_n/\sqrt{n}) \pm \frac{|f'|_\infty^2 a_n}{\sqrt{n}},$$

with the remainder terms tending to 0 and, for h fixed,

$$n\big[M(h/\sqrt{n}) - M(0)\big] \xrightarrow{n\to\infty} \tfrac{1}{2}M''(0)h^2.$$

With the fact that M'' is continuous under our assumption that f is continuously differentiable, we conclude that μ is equal to

$$\tfrac{1}{2}M''(0)h^2 \pm Q_{1,n}, \quad Q_{1,n} \xrightarrow{n\to\infty} 0 \text{ in probability.}$$

As for V_i in the lemma, it is equal to

$$\frac{1}{n}\sum_{j=1}^{n}(f_j(h) - f_i(h)) \int \big[\mathbb{I}\{Y_i \le t\} - \Phi(t - f(x_i))\big]\phi(t - f(x_j))\mathrm{d}t \tag{3.18}$$

$$= \frac{1}{n}\sum_{j=1}^{n}(f_j(h) - f_i(h)) \int \big[\mathbb{I}\{Z_i \le z + f(x_j) - f(x_i)\} - \Phi(z + f(x_j) - f(x_i))\big]\phi(z)\mathrm{d}z \tag{3.19}$$

$$\stackrel{d}{=} \frac{1}{n}\sum_{j=1}^{n}(f_j(h) - f_i(h))\big[\Phi(Z_i + f(x_j) - f(x_i)) - \Phi_2(f(x_j) - f(x_i))\big], \tag{3.20}$$

using the fact that ϕ is symmetric about 0. Note that

$$\begin{aligned} f_j(h) - f_i(h) &= f(x_j + h/\sqrt{n}) - f(x_j) - \big[f(x_i + h/\sqrt{n}) - f(x_i)\big] \\ &= (h/\sqrt{n})\big[f'(x_j) - f'(x_i) \pm \omega_1(h/\sqrt{n})\big] \\ &= (h/\sqrt{n})[f'(x_j) - f'(x_i)] \pm (a_n/\sqrt{n})\omega_1(a_n/\sqrt{n}). \end{aligned}$$

Hence, the quantity in (3.20) is equal to

$$\frac{1}{n}\sum_{j=1}^{n}(h/\sqrt{n})(f'(x_j)-f'(x_i))\big[\Phi(Z_i+f(x_j)-f(x_i))-\Phi_2(f(x_j)-f(x_i))\big]\pm(a_n/\sqrt{n})\omega_1(a_n/\sqrt{n})$$
$$=\frac{h}{\sqrt{n}}\big[\Lambda(x_i,Z_i)\pm\omega_1(1/n)\big]\pm(a_n/\sqrt{n})\omega_1(a_n/\sqrt{n}),$$

where

$$\begin{aligned}\Lambda(t,z)&:=\int_0^1(f'(x)-f'(t))\big[\Phi(z+f(x)-f(t))-\Phi_2(f(x)-f(t))\big]\mathrm{d}x\\&:=\int_0^1(f'(x)-f'(t))\Xi(f(x)-f(t),z)\mathrm{d}x,\end{aligned}$$

using Lemma 16. We thus conclude that $\sum_{i=1}^{n}V_i$ is equal, in distribution, to

$$\frac{h}{\sqrt{n}}\sum_{i=1}^{n}\Lambda(x_i,Z_i)\pm Q_{2,n},\quad Q_{2,n}\xrightarrow{n\to\infty}0\text{ in probability.}$$

By Markov's inequality and the fact that we only consider $W_n(h)$ for $|h|\le a_n$, we thus have that

$$W_n(h)\overset{d}{=}\tfrac{1}{2}M''(0)h^2\pm Q_{1,n}+\frac{h}{\sqrt{n}}\sum_{i=1}^{n}\Lambda(x_i,Z_i)\pm Q_{2,n}\pm Q_{3,n},$$

where

$$\mathbb{E}[Q_{3,n}^2]\le\frac{C}{n}\sum_{i=1}^{n}f_i(h)^2\le C(|f'|_\infty h/\sqrt{n})^2\le\frac{C|f'|_\infty^2a_n^2}{n},$$

so that $Q_{3,n}\to 0$ as $n\to 0$ in probability. More succinctly, therefore,

$$W_n(h)\overset{d}{=}\tfrac{1}{2}M''(0)h^2+\frac{h}{\sqrt{n}}\sum_{i=1}^{n}\Lambda(x_i,Z_i)+o_P(1).\tag{3.21}$$

Hence, by Slutsky's theorem in the form of [Kos08, Th 7.15], it suffices to look at

$$G_n(h) := \tfrac{1}{2}M''(0)h^2 + \frac{h}{\sqrt{n}}\sum_{i=1}^{n}\Lambda(x_i, Z_i), \tag{3.22}$$

which is an exceedingly simple process. (This is because we are effectively in a classical setting, even though the ranks obfuscate that.)

Indeed, note that $\Lambda(x_i, Z_i)$ is centered and bounded in absolute value by $2|f'|_\infty$, and

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[\Lambda(x_i, Z_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Lambda(x_i, Z_i)^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left[\int_0^1 (f'(x) - f'(x_i))\,\Xi(f(x) - f(x_i), z)\mathrm{d}x\right]^2 \phi(z)\mathrm{d}z \\
&\xrightarrow{n\to\infty} \gamma^2,
\end{aligned}$$

again applying Lemma 16. Therefore, by Lyapunov's central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Lambda(x_i, Z_i) \overset{n\to\infty}{\Longrightarrow} \mathcal{N}(0, \gamma^2).$$

From this, it follows that G_n converges weakly on every bounded interval to the Gaussian process given by

$$G(h) := \tfrac{1}{2}M''(0)h^2 + h\gamma U,$$

where U is a standard normal random variable. The limit process could not be simpler. In particular, G has continuous sample paths (in fact, its sample paths are parabolas), and recalling that $M''(0) < 0$ by Proposition 1, it is clear that G has a unique maximum point at $h^* := (\gamma/M''(0))U$. Note that h^* is normal with mean zero and variance $\gamma^2/M''(0)^2$. The proof of the theorem then

follows from an application of Lemma 19.

3.6 Acknowledgments

Chapter 3, in full, is a version of the paper “Template Matching with Ranks”, Arias-Castro, Ery; Zheng, Lin. The manuscript has been submitted for publication in a major statistical journal. The dissertation author was the primary investigator and author of this material.

Chapter 4

Some Theory for Texture Segmentation

4.1 Introduction

Texture segmentation, and image segmentation more generally, is an important task in computer vision and pattern recognition, being widely applied to areas such as scene understanding, remote sensing and autonomous driving [PP93, Zha06, RD93, LCF+19]. While the vast majority of the work in texture segmentation, as in image processing at large, is applied, we contribute some theory by establishing the consistency of the basic approach.

In this thesis, we address the problem of texture segmentation by extracting local features in the neighborhood of a pixel and then applying a clustering algorithm for grouping the pixel according to these features. The process of extracting features has undergone some important changes over the years, ranging from the use of sophisticated systems from applied harmonic analysis such as Gabor filters or wavelets [DH95, GPK02, JF91, Uns95, WH96, RH99] to multiresolution or multiscale aggregation approaches [GSBB03, MJ92], among others [MBLS01, HPB98], to the use deep learning, particularly in the form convolutional neural networks (CNN), whose success is attributed to the capability of CNN to learn a hierarchical representation of

raw input data [LSD15, RFB15, MNA16, BKC17]. In this work, we extract local features by collecting local second moment information on patches. For clustering, many algorithms and theoretical results have been reported in the literature, such as spectral clustering [NJW02], k-means++ [AV07] and single-linkage clustering [AC11]. In this thesis, we consider k-means and singe-linkage clustering.

The chapter is organized as follows. In Section 4.2, we consider stationary textures, which is done by the extraction of local second moment information on patches and the application of k-means. In Section 4.3, we consider non-stationary textures, where we also include location as a feature and we apply instead single-linkage clustering. In Section 4.4, we present the result of some numerical experiments, mostly there to illustrate the theory developed in the main part of the paper. Both synthetic and natural textures are considered.

4.2 Stationary Textures

In this section we consider textures to be stationary. The model we adopt and the method we implement are introduced in Section 4.2.1 and Section 4.2.2. We then establish in Section 4.2.3 the consistency of a simple incarnation of the basic approach.

4.2.1 Model

We have a pixel image X of size $n \times n$, that we assume is partitioned into two sub-regions $\mathcal{G}_0$ and $\mathcal{G}_1$ by curve $\bar{\mathcal{G}}$. $\mathcal{G}_0$ is a stationary Gaussian Markov random field with mean 0 and autocovariance matrix A_0. $\mathcal{G}_1$ is a stationary Gaussian Markov random field with mean 0 and autocovariance matrix A_1. In image X, we pick up n^2 pixels with equal intervals, and get observations

$$\{X_t\}, \quad t \in \mathcal{T} := \{1, 2, \cdots, n\}^2. \tag{4.1}$$

To estimate curve $\bar{\mathcal{G}}$, we need to cluster the n^2 pixels into two groups.

4.2.2 Methods

We define scanning patches as follows. To simplify the presentation assume n is the square of an integer (namely $n = m^2$ for some integer m). For $\forall t = (t_1, t_2) \in \mathcal{T}$, pick up patch S_t with size $(2m+1) \times (2m+1)$,

$$S_t = \begin{pmatrix} X_{t+(-m,-m)} & X_{t+(-m,-m+1)} & \cdots & X_{t+(-m,m)} \\ X_{t+(-m+1,-m)} & X_{t+(-m+1,-m+1)} & \cdots & X_{t+(-m+1,m)} \\ \vdots & \vdots & \cdots & \vdots \\ X_{t+(m,-m)} & X_{t+(m,-m+1)} & \cdots & X_{t+(m,m)} \end{pmatrix}, \tag{4.2}$$

Next, autocovariance is defined based on scanning patches. For $\forall t = (t_1, t_2) \in \mathcal{T}$ and $\forall i = (i_1, i_2) \in \mathcal{M} := \{-m, -m+1, \cdots, m-1, m\}^2$, define true autocovariance and sample autocovariance as follows

$$C_t(i) = \text{Mean of } \{\mathbb{E}(X_t \cdot X_{t+i}) \mid \text{both } X_t \text{ and } X_{t+i} \text{ are in } S_t\} \tag{4.3}$$

and

$$\hat{C}_t(i) = \text{Mean of } \{X_t \cdot X_{t+i} \mid \text{both } X_t \text{ and } X_{t+i} \text{ are in } S_t\}. \tag{4.4}$$

Denote the vectorizations of $\{C_t(i)\}_{i \in \mathcal{M}}$ and $\{\hat{C}_t(i)\}_{i \in \mathcal{M}}$ to be C_t and $\hat{C}_t$ respectively. Here C_t is the true feature of pixel X_t and $\hat{C}_t$ is the observed feature of pixel X_t.

Also based on scanning patches, we define following three sets

$$\mathcal{H}_0 = \{t \in \mathcal{T} \mid S_t \subset \mathcal{G}_0\}, \tag{4.5}$$

$$\mathcal{H}_1 = \{t \in \mathcal{T} \mid S_t \subset \mathcal{G}_1\} \tag{4.6}$$

and

$$\mathcal{H} = \{t \in \mathcal{T} \mid S_t \cap \mathcal{G}_0 \neq \varnothing \text{ and } S_t \cap \mathcal{G}_1 \neq \varnothing\}. \tag{4.7}$$

Here $\mathcal{G}_0$ and $\mathcal{G}_1$ are both stationary fields, so all elements in set $\{C_t\}_{\forall t \in \mathcal{H}_0}$ are the same and we denote it as C^0. Similarly, all elements in set $\{C_t\}_{\forall t \in \mathcal{H}_1}$ are the same and we denote it as C^1. Define template autocovariance $E = \begin{pmatrix} C^0 \\ C^1 \end{pmatrix}$.

Then we introduce membership matrix. Define $n^2 \times 2$ true membership matrix W such that for $\forall t = (t_1, t_2) \in \mathcal{T}$,

$$W_t = \text{the } n(t_1 - 1) + t_2 \, th \text{ row of matrix } W = \begin{cases} (1,0), & \text{if } t = (t_1, t_2) \in \mathcal{G}_0, \\ (0,1), & \text{if } t = (t_1, t_2) \in \mathcal{G}_1. \end{cases} \tag{4.8}$$

Also define the set of membership matrices $\mathcal{W}_{n,2}$ as follows

$$\mathcal{W}_{n,2} = \{n^2 \times 2 \text{ matrices with rows } (0,1) \text{ or } (1,0)\}. \tag{4.9}$$

Based on above calculations and definitions, we define k-means clustering estimation as

$$(\hat{W}, \hat{E}) = \underset{W \in \mathcal{W}_{n,2}, E \in \mathbb{R}_{2 \times (2m+1)^2}}{\arg\min} \sum_{t \in \mathcal{T}} \| (WE)_t - \hat{C}_t \|_\infty^2, \tag{4.10}$$

where $(WE)_t$ is the $n(t_1 - 1) + t_2 \, th$ row of matrix WE.

In practice k-means can not be solved exactly, however, there exists polynomial time algorithm which obtains approximation $(\hat{W}, \hat{E})$ satisfying following equation (4.11), such as $(1 + \varepsilon)$-approximate method in [KSS04].

$$\sum_{t \in \mathcal{T}} \| (\hat{W}\hat{E})_t - \hat{C}_t \|_\infty^2 \leq (1 + \varepsilon) \cdot \min_{W \in \mathcal{W}_{n,2}, E \in \mathbb{R}_{2 \times (2m+1)^2}} \sum_{t \in \mathcal{T}} \| (WE)_t - \hat{C}_t \|_\infty^2, \tag{4.11}$$

where $\hat{W} \in \mathcal{W}_{n,2}$ and $\hat{E} \in \mathbb{R}_{2\times(2m+1)^2}$. Thus we cluster the n^2 pixels into two groups by membership matrix estimation $\hat{W}$. As a summary, we provide the procedure of k-means algorithm in Algorithm 1.

Algorithm 1 Texture Segmentation with K-means Algorithm

Input: Observations $\{X_t\}_{t\in\mathcal{T}}$, approximation error ε.

Output: Membership matrix estimation $\hat{W}$.

1: For $\forall t = (t_1, t_2) \in \mathcal{T}$, pick up patch S_t.

2: Calculate sample autocovariance $\{\hat{C}_t(i)\}_{i\in\mathcal{M}}$ and obtain observed features $\{\hat{C}_t\}_{t\in\mathcal{T}}$.

3: Define template autocovariance E and the set of membership matrices $\mathcal{W}_{n,2}$.

4: Obtain k-means approximation solution $(\hat{W}, \hat{E})$ which satisfies (4.11).

5: **return** $\hat{W}$.

Define the set of 2×2 permutation matrices

$$\Phi_2 = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}, \tag{4.12}$$

then calculate

$$\hat{Q} = \arg\min_{Q\in\Phi_2} \sum_{t\in\mathcal{T}} \|(\hat{W}Q)_t - W_t\|_\infty^2. \tag{4.13}$$

Next, define the set of mistakenly clustered elements to be $\mathcal{R}$ as follows

$$\mathcal{R} = \{t \in \mathcal{T} : (\hat{W}\hat{Q})_t \neq W_t\}, \tag{4.14}$$

then clustering error rate is

$$|\mathcal{R}|/n^2 = \frac{1}{n^2} \sum_{t\in\mathcal{T}} \|(\hat{W}\hat{Q})_t - W_t\|_\infty^2. \tag{4.15}$$

4.2.3 Theory

Firstly we introduce following assumptions.

Assumption 9. Both $\mathcal{G}_0$ and $\mathcal{G}_1$ are wide-sence stationary Gaussian Markov random fields.

Assumption 10. Let $\Delta = \|C^0 - C^1\|_\infty$. For $\forall \beta > 1$,

$$\frac{(\log n)^\beta}{\Delta^2 n} \to 0 \quad \text{as} \quad n \to \infty. \tag{4.16}$$

Assumption 11. Define $C_{t0} = C^0$ for $\forall t \in \mathcal{G}_0$ and $C_{t0} = C^1$ for $\forall t \in \mathcal{G}_1$. With the same β in Assumption 10,

$$\sum_{t \in \mathcal{H}} \|C_t - C_{t0}\|_\infty^2 \le \frac{n(\log n)^\beta}{24}. \tag{4.17}$$

Next, before introducing the theory, we indicate the error bound of $\|\hat{C}_t - C_t\|_\infty$.

Lemma 21. *Under Assumption 9, for $\forall t \in \mathcal{T}$ and $\forall a > 0$, there exists a constant J such that*

$$\mathbb{P}(\|\hat{C}_t - C_t\|_\infty > a) \le 2(2\sqrt{n}+1)^2 \exp(-J \cdot a^2 n). \tag{4.18}$$

Proof. First let S_t^v be the vectorization of S_t, then S_t^v is a vector of length $(2m+1)^2$

$$S_t^v = \begin{pmatrix} X_{t+(-m,-m)} \\ \vdots \\ X_{t+(-m,m)} \\ X_{t+(-m+1,-m)} \\ \vdots \\ X_{t+(-m+1,m)} \\ \vdots \\ \vdots \\ \vdots \\ X_{t+(m,-m)} \\ \vdots \\ X_{t+(m,m)} \end{pmatrix}. \tag{4.19}$$

Then for $\forall i = (i_1, i_2) \in \mathcal{M}$, there exists a matrix A_i such that

$$\hat{C}_t(i) = \frac{1}{(2m+1-|i_1|)(2m+1-|i_2|)} (S_t^v)^T A_i S_t^v, \tag{4.20}$$

where A_i is a $(2m+1)^2 \times (2m+1)^2$ matrix with elements 0 and 1, and the number of 1 is less than $(2m+1-|i_1|)(2m+1-|i_2|)$.

Since the field is stationary, suppose $S_t^v \sim N(0, \Sigma)$, where Σ is non-negative. Let $\Sigma = U \Lambda U^T$ be the spectral decomposition of Σ. Define

$$Y_t = U^T S_t^v, \tag{4.21}$$

then

$$Y_t \sim N(0, U^T \Sigma U) = N(0, \Lambda). \tag{4.22}$$

So

$$\hat{C}_t(i) = \frac{1}{(2m+1-|i_1|)(2m+1-|i_2|)} (S_t^v)^T A_i S_t^v \tag{4.23}$$

$$= \frac{1}{(2m+1-|i_1|)(2m+1-|i_2|)} (UY_t)^T A_i (UY_t) \tag{4.24}$$

$$= \frac{1}{(2m+1-|i_1|)(2m+1-|i_2|)} Y_t^T (U^T A_i U) Y_t. \tag{4.25}$$

By Hanson-Wright inequality in [RV13], for $\forall a > 0$, there exist constants K_1 and J_1, such that

$$\mathbb{P}(|\hat{C}_t(i) - C_t(i)| > a) \tag{4.26}$$

$$= \mathbb{P}(|Y_t^T (U^T A_i U) Y_t - \mathbb{E}[Y_t^T (U^T A_i U) Y_t]| > a \cdot (2m+1-|i_1|)(2m+1-|i_2|)) \tag{4.27}$$

$$\leq 2\exp\Big(-J_1 \cdot \min\Big\{ \frac{a^2 (2m+1-|i_1|)^2 (2m+1-|i_2|)^2}{K_1^4 \|U^T A_i U\|_F^2}, \frac{a(2m+1-|i_1|)(2m+1-|i_2|)}{K_1^2 \|U^T A_i U\|_2} \Big\} \Big). \tag{4.28}$$

Next we focus on $\|U^T A_i U\|_F$ and $\|U^T A_i U\|_2$. Since U is an orthogonal matrix,

$$\|U^T A_i U\|_2 = \|A_i\|_2 \leq \|A_i\|_1 = \max_k \sum_{l=1}^{(2m+1)^2} |A_i(k,l)| = 1 \tag{4.29}$$

and

$$\|U^T A_i U\|_F = \sqrt{\text{Sum of eigenvalues of } (U^T A_i U)^T (U^T A_i U)} \tag{4.30}$$

$$\leq \sqrt{(2m+1)^2 \cdot \lambda_{max}((U^T A_i U)^T (U^T A_i U))} \tag{4.31}$$

$$= \sqrt{(2m+1)^2} \cdot \|U^T A_i U\|_2 \tag{4.32}$$

$$= 2m+1. \tag{4.33}$$

Then for $\forall i = (i_1, i_2) \in \mathcal{M}$, there exist constants K_1, J_1 and J such that

$$\mathbb{P}(|\hat{C}_t(i) - C_t(i)| > a) \tag{4.34}$$

$$\leq 2\exp\Big(-J_1 \cdot \min\Big\{\frac{a^2(2m+1-|i_1|)^2(2m+1-|i_2|)^2}{K_1^4(2m+1)^2}, \frac{a(2m+1-|i_1|)(2m+1-|i_2|)}{K_1^2}\Big\}\Big) \tag{4.35}$$

$$\leq 2\exp(-J \cdot \min\{a^2m^2, am^2\}). \tag{4.36}$$

So when a is small enough, we have

$$\mathbb{P}(|\hat{C}_t(i) - C_t(i)| > a) \leq 2\exp(-J \cdot a^2m^2). \tag{4.37}$$

Next since

$$\|\hat{C}_t - C_t\|_\infty = \max_{i \in \mathcal{M}} |\hat{C}_t(i) - C_t(i)|, \tag{4.38}$$

by Union bound, for $\forall a > 0$,

$$\mathbb{P}(\|\hat{C}_t - C_t\|_\infty > a) \leq 2(2m+1)^2 \exp(-J \cdot a^2m^2) \tag{4.39}$$

$$= 2(2\sqrt{n}+1)^2 \exp(-J \cdot a^2 n). \tag{4.40}$$

□

Lemma 22. *Let* $\Delta = \|C^0 - C^1\|_\infty$. *Define* $\mathcal{A}_k = \{t \in \mathcal{G}_k : \|(\hat{W}\hat{E})_t - C^k\|_\infty \geq \Delta/2\}, k = 0, 1,$ *and* $\mathcal{A}' = \mathcal{A}_0 \cup \mathcal{A}_1$, *we have* $\mathcal{T} \backslash \mathcal{A}' = (\mathcal{G}_0 \backslash \mathcal{A}_0) \cup (\mathcal{G}_1 \backslash \mathcal{A}_1)$. *Then all the elements in* $\mathcal{T} \backslash \mathcal{A}'$ *are clustered correctly.*

Proof. On the one hand, for $\forall t \in \mathcal{G}_0 \backslash \mathcal{A}_0$ and $\forall s \in \mathcal{G}_1 \backslash \mathcal{A}_1$, by contradiction, if $(\hat{W}\hat{E})_t = (\hat{W}\hat{E})_s$,

$$\Delta = \|C^0 - C^1\|_\infty \leq \|C^0 - (\hat{W}\hat{E})_t\|_\infty + \|(\hat{W}\hat{E})_t - (\hat{W}\hat{E})_s\|_\infty + \|(\hat{W}\hat{E})_s - C^1\|_\infty \tag{4.41}$$

$$< \Delta/2 + 0 + \Delta/2 \tag{4.42}$$

$$= \Delta, \tag{4.43}$$

which is conflicted by itself, so $(\hat{W}\hat{E})_t \neq (\hat{W}\hat{E})_s$. On the other hand, suppose $t, s \in \mathcal{G}_0 \backslash \mathcal{A}_0$ or $t, s \in \mathcal{G}_1 \backslash \mathcal{A}_1$, by contradiction, if $(\hat{W}\hat{E})_t \neq (\hat{W}\hat{E})_s$, $\hat{W}\hat{E}$ has at least three distinct rows, however, according to the structure of $\hat{W}\hat{E}$, it has exactly two distinct rows, which is a conflict. So $(\hat{W}\hat{E})_t = (\hat{W}\hat{E})_s$. Thus, all the elements in $\mathcal{T} \backslash \mathcal{A}'$ are clustered correctly. □

Next, we introduce the theory for k-means clustering algorithm.

Theorem 11. *Under Assumption 9, 10 and 11, consider k-means clustering in Algorithm 1. For* $\forall \beta > 1$*, there exists a constant J, as* $n \to \infty$*,*

$$\mathbb{P}\left(|\mathcal{R}|/n^2 > \frac{(\log n)^\beta}{\Delta^2 n}\right) \leq 2(2\sqrt{n}+1)^2 n^2 \exp(-J \cdot (\log n)^\beta) \to 0, \tag{4.44}$$

where $|\mathcal{R}|/n^2$ *is the clustering error rate. Here* $\frac{(\log n)^\beta}{\Delta^2 n} \to 0$ *as* $n \to \infty$*.*

Proof. By Algorithm 1 in Section 4.2.2, we have

$$\sum_{t \in \mathcal{T}} \|(\hat{W}\hat{E})_t - \hat{C}_t\|_\infty^2 \leq (1+\varepsilon) \cdot \min_{W \in \mathcal{W}_{n,2}, E \in \mathbb{R}_{2\times(2m+1)^2}} \sum_{t \in \mathcal{T}} \|(WE)_t - \hat{C}_t\|_\infty^2, \tag{4.45}$$

where $\hat{W} \in \mathcal{W}_{n,2}$, $\hat{E} \in \mathbb{R}_{2\times(2m+1)^2}$. Without loss of generality, set $\varepsilon < 1$, then

$$\sum_{t \in \mathcal{T}} \|(\hat{W}\hat{E})_t - \hat{C}_t\|_\infty^2 \leq (1+\varepsilon) \sum_{t \in \mathcal{T}} \|C_t - \hat{C}_t\|_\infty^2 \tag{4.46}$$

$$\leq 2 \sum_{t \in \mathcal{T}} \|C_t - \hat{C}_t\|_\infty^2. \tag{4.47}$$

On the one hand, by Triangle Inequality,

$$\sum_{t\in\mathcal{T}} \|(\hat{W}\hat{E})_t - C_{t0}\|_\infty^2 \le \sum_{t\in\mathcal{T}} (\|(\hat{W}\hat{E})_t - \hat{C}_t\|_\infty + \|\hat{C}_t - C_t\|_\infty + \|C_t - C_{t0}\|_\infty)^2 \tag{4.48}$$

$$\le 3\sum_{t\in\mathcal{T}} \|(\hat{W}\hat{E})_t - \hat{C}_t\|_\infty^2 + 3\sum_{t\in\mathcal{T}} \|\hat{C}_t - C_t\|_\infty^2 + 3\sum_{t\in\mathcal{T}} \|C_t - C_{t0}\|_\infty^2. \tag{4.49}$$

In addition, by Assumption 11 and (4.47),

$$\sum_{t\in\mathcal{T}} \|(\hat{W}\hat{E})_t - C_{t0}\|_\infty^2 \le 6\sum_{t\in\mathcal{T}} \|C_t - \hat{C}_t\|_\infty^2 + 3\sum_{t\in\mathcal{T}} \|\hat{C}_t - C_t\|_\infty^2 + 3\sum_{t\in\mathcal{H}} \|C_t - C_{t0}\|_\infty^2 \tag{4.50}$$

$$= 9\sum_{t\in\mathcal{T}} \|\hat{C}_t - C_t\|_\infty^2 + 3\sum_{t\in\mathcal{H}} \|C_t - C_{t0}\|_\infty^2 \tag{4.51}$$

$$\le 9\sum_{t\in\mathcal{T}} \|\hat{C}_t - C_t\|_\infty^2 + \frac{n(\log n)^\beta}{8}. \tag{4.52}$$

On the other hand,

$$\sum_{t\in\mathcal{T}} \|(\hat{W}\hat{E})_t - C_{t0}\|_\infty^2 \ge \sum_{t\in\mathcal{A}'} \frac{\Delta^2}{4} = \frac{\Delta^2(|\mathcal{A}_0| + |\mathcal{A}_1|)}{4} = \frac{|\mathcal{A}'|\Delta^2}{4}, \tag{4.53}$$

then we have

$$|\mathcal{A}'| \le \frac{36\sum_{t\in\mathcal{T}} \|\hat{C}_t - C_t\|_\infty^2 + \frac{n(\log n)^\beta}{2}}{\Delta^2}. \tag{4.54}$$

By Lemma 22,

$$\mathbb{P}\left(|\mathcal{R}|/n^2 > \frac{(\log n)^\beta}{\Delta^2 n}\right) = \mathbb{P}\left(|\mathcal{R}| > \frac{n(\log n)^\beta}{\Delta^2}\right) \le \mathbb{P}\left(|\mathcal{A}'| > \frac{n(\log n)^\beta}{\Delta^2}\right), \tag{4.55}$$

then by Lemma 21 and (4.54), there exists a constant J,

$$\mathbb{P}\Big(|\mathcal{R}|/n^2 > \frac{(\log n)^\beta}{\Delta^2 n}\Big) \tag{4.56}$$

$$\leq \mathbb{P}\Big(\frac{36\sum_{t\in\mathcal{T}}\|\hat{C}_t - C_t\|_\infty^2 + \frac{n(\log n)^\beta}{2}}{\Delta^2} > \frac{n(\log n)^\beta}{\Delta^2}\Big) \tag{4.57}$$

$$= \mathbb{P}\Big(\sum_{t\in\mathcal{T}}\|\hat{C}_t - C_t\|_\infty^2 > \frac{n(\log n)^\beta}{72}\Big) \tag{4.58}$$

$$\leq \sum_{t\in\mathcal{T}}\mathbb{P}\Big(\|\hat{C}_t - C_t\|_\infty^2 > \frac{(\log n)^\beta}{72n}\Big) \tag{4.59}$$

$$\leq 2(2\sqrt{n}+1)^2 n^2 \exp(-J\cdot(\log n)^\beta). \tag{4.60}$$

So for $\forall\beta > 1$, as $n \to \infty$,

$$\mathbb{P}\Big(|\mathcal{R}|/n^2 > \frac{(\log n)^\beta}{\Delta^2 n}\Big) \leq 2(2\sqrt{n}+1)^2 n^2 \exp(-J\cdot(\log n)^\beta) \to 0. \tag{4.61}$$

Thus, for stationary Gaussian random field, we obtain the error rate for k-means clustering algorithm. □

4.3 Non-stationary Textures

In this section we consider textures to be non-stationary. Here both $\mathcal{G}_0$ and $\mathcal{G}_1$ are non-stationary Gaussian Markov random fields with mean 0. We add location information into consideration and cluster the n^2 pixels into two groups by single-linkage algorithm. The algorithm is established in Section 4.3.1. Then we show the consistency of a simple incarnation of the basic approach in Section 4.3.2.

4.3.1 Methods

Pick up $\lfloor \frac{n}{2m+1} \rfloor \times \lfloor \frac{n}{2m+1} \rfloor$ pixels $\{X_u\}_{u \in \mathcal{U}}$ with equal intervals from $\{X_t\}_{t \in \mathcal{T}}$, where

$$\mathcal{U} = \left\{ u = (u_1, u_2) \,\middle|\, u_1 = (2m+1) \cdot t_1,\ u_2 = (2m+1) \cdot t_2,\ (t_1, t_2) = \left\{ 1, 2, \cdots, \left\lfloor \frac{n}{2m+1} \right\rfloor \right\}^2 \right\}. \quad (4.62)$$

Here $\mathcal{U}$ is a subset of $\mathcal{T}$. Similar to (4.2) in Section 4.2.2, for $\forall u \in \mathcal{U}$, pick up patch S_u as follows

$$S_u = \begin{pmatrix} X_{u+(-m,-m)} & X_{u+(-m,-m+1)} & \cdots & X_{u+(-m,m)} \\ X_{u+(-m+1,-m)} & X_{u+(-m+1,-m+1)} & \cdots & X_{u+(-m+1,m)} \\ \vdots & \vdots & \cdots & \vdots \\ X_{u+(m,-m)} & X_{u+(m,-m+1)} & \cdots & X_{u+(m,m)} \end{pmatrix}. \quad (4.63)$$

For $\forall u \neq v \in \mathcal{U}$, it is obvious that $|u_1 - v_1| \geq 2m+1$ or $|u_2 - v_2| \geq 2m+1$. So there is no overlap between S_u and S_v.

Next for $\forall u \in \mathcal{U}$, add location information into its true feature C_u. Denote C'_u as the new true feature of pixel X_u as follows

$$C'_u = \left(C_u, \frac{u_1}{n}, \frac{u_2}{n} \right), \quad (4.64)$$

where C'_u is a vector of length $(2m+1)^2 + 2$. Similarly, denote $\hat{C}'_u$ as the new observed feature of pixel X_u as follows

$$\hat{C}'_u = \left(\hat{C}_u, \frac{u_1}{n}, \frac{u_2}{n} \right), \quad (4.65)$$

where $\hat{C}'_u$ also is a vector of length $(2m+1)^2 + 2$.

Apply single-linkage algorithm in following steps. Firstly among $\mathcal{U}$, connect all pairs (u, v) with $\|\hat{C}'_u - \hat{C}'_v\|_\infty < \frac{(\log n)^{\beta/2}}{\sqrt{n}}$, where $1 < \beta < 2$. Next for $\forall u \in \mathcal{U}$, assign all the other pixels in S_u into the same cluster as X_u. Then for any pixel X_t which is still not clustered, find the pixel

X_u in $\{X_u\}_{u\in\mathcal{U}}$ with the smallest distance to X_t, and assign pixel X_t into the same cluster with X_u. Finally we obtain the clustering result.

As a summary, we provide the procedure of single-linkage algorithm in Algorithm 2.

Algorithm 2 Texture Segmentation with Single-linkage Algorithm

Input: Observations $\{X_t\}_{t\in\mathcal{T}}$.

Output: Single-linkage clustering result.

1: For $\forall u = (u_1, u_2) \in \mathcal{U}$, pick up patch S_u.

2: Calculate new observed features $\{\hat{C}'_u\}_{u\in\mathcal{U}}$.

3: Apply single-linkage algorithm based on $\{\hat{C}'_u\}_{u\in\mathcal{U}}$.

4: **return** Single-linkage clustering result.

Define the following set

$$\mathcal{V} = \{u \in \mathcal{U} \mid S_u \cap \mathcal{G}_0 = \varnothing \text{ or } S_u \cap \mathcal{G}_1 = \varnothing\} \tag{4.66}$$

and

$$\mathcal{W} = \{t \in \mathcal{T} \mid X_t \text{ is a pixel in patch } S_u \text{ where } u \in \mathcal{V}\}. \tag{4.67}$$

In next section, we can show that all the pixels in $\mathcal{V}$ can be clustered correctly with probability going to 1. Thus, all pixels in $\mathcal{W}$ can be clustered correctly with probability going to 1.

4.3.2 Theory

Firstly we introduce two assumptions on non-stationary level of the fields.

Assumption 12. For any X_t, X_s in the same sub-region,

$$\|C_t - C_s\|_\infty = O(\sqrt{\log n} \cdot \mathsf{D}(s,t)), \tag{4.68}$$

where $\mathsf{D}(s,t)$ is the distance between two pixels X_t and X_s

$$\mathsf{D}(s,t) = \sqrt{\left(\frac{s_1 - t_1}{n}\right)^2 + \left(\frac{s_2 - t_2}{n}\right)^2}. \tag{4.69}$$

Assumption 13. For any X_t, X_s in different sub-regions, if $\mathsf{D}(s,t) < \frac{\log n}{\sqrt{n}}$, there exists a constant K such that

$$\|C_t - C_s\|_\infty \geq \frac{K \cdot \log n}{\sqrt{n}}. \tag{4.70}$$

Next, we show that the single-linkage algorithm in above section is consistent.

Theorem 12. *Under Assumption 12 and Assumption 13, by single-linkage clustering in Algorithm 2, set threshold value* $b = \frac{(\log n)^{\beta/2}}{\sqrt{n}}$*, where* $1 < \beta < 2$*. Then as* $n \to \infty$*,*

$$\mathbb{P}(\textit{All pixels in } \mathcal{W} \textit{ are clustered correctly}) \geq 1 - n^3 \cdot \exp(-J \cdot (\log n)^\beta) \ \to 1. \tag{4.71}$$

Proof. Set threshold value $b = \frac{(\log n)^{\beta/2}}{\sqrt{n}}$, where $1 < \beta < 2$. For $\forall u \in \mathcal{U}$, denote u_+ as the pixel bordering and above u in $\mathcal{U}$, and denote u_- as the pixel bordering and below u in $\mathcal{U}$, then

$$\mathbb{P}(\text{All pixels in } \mathcal{W} \text{ are clustered correctly}) \tag{4.72}$$

$$\geq 1 - \mathbb{P}\Big(\max_{u \in \mathcal{V}} \max_{\substack{v \in \mathcal{V} \\ v,u \text{ in the same sub-region} \\ S_u \text{ borders on } S_v}} \|\hat{C}'_u - \hat{C}'_v\|_\infty > \frac{(\log n)^{\beta/2}}{\sqrt{n}}\Big) \tag{4.73}$$

$$- \mathbb{P}\Big(\min_{u \in \mathcal{V}} \min_{\substack{v \in \mathcal{V} \\ v,u \text{ in different sub-regions}}} \|\hat{C}'_u - \hat{C}'_v\|_\infty < \frac{(\log n)^{\beta/2}}{\sqrt{n}}\Big) \tag{4.74}$$

$$- \mathbb{P}\Big(\min_{u \in \mathcal{U} \setminus \mathcal{V}} \|\hat{C}'_{u_+} - \hat{C}'_{u_-}\|_\infty < \frac{2(\log n)^{\beta/2}}{\sqrt{n}}\Big). \tag{4.75}$$

By Union bound,

$$\mathbb{P}(\text{All pixels in } \mathcal{W} \text{ are clustered correctly}) \tag{4.76}$$

$$\geq 1 - \sum_{u \in \mathcal{V}} \sum_{\substack{v \in \mathcal{V} \\ v,u \text{ in the same sub-region} \\ S_u \text{ borders on } S_v}} \mathbb{P}\Big(\|\hat{C}'_u - \hat{C}'_v\|_\infty > \frac{(\log n)^{\beta/2}}{\sqrt{n}} \Big) \tag{4.77}$$

$$- \sum_{u \in \mathcal{V}} \sum_{\substack{v \in \mathcal{V} \\ v,u \text{ in different sub-regions}}} \mathbb{P}\Big(\|\hat{C}'_u - \hat{C}'_v\|_\infty < \frac{(\log n)^{\beta/2}}{\sqrt{n}} \Big) \tag{4.78}$$

$$- \sum_{u \in \mathcal{U} \setminus \mathcal{V}} \mathbb{P}\Big(\|\hat{C}'_{u_+} - \hat{C}'_{u_-}\|_\infty < \frac{2(\log n)^{\beta/2}}{\sqrt{n}} \Big). \tag{4.79}$$

We calculate above probability in three steps. Firstly, under Assumption 12, for $\forall u \in \mathcal{U}$,

$$\max_{\substack{v \in \mathcal{V} \\ v,u \text{ in the same sub-region} \\ S_u \text{ borders on } S_v}} \|C'_u - C'_v\|_\infty = O\Big(\sqrt{\frac{\log n}{n}} \Big), \tag{4.80}$$

then by Triangle Inequality,

$$\sum_{u \in \mathcal{V}} \sum_{\substack{v \in \mathcal{V} \\ v,u \text{ in the same sub-region} \\ S_u \text{ borders on } S_v}} \mathbb{P}\Big(\|\hat{C}'_u - \hat{C}'_v\|_\infty > \frac{(\log n)^{\beta/2}}{\sqrt{n}} \Big) \tag{4.81}$$

$$\leq \sum_{u \in \mathcal{V}} \sum_{\substack{v \in \mathcal{V} \\ v,u \text{ in the same sub-region} \\ S_u \text{ borders on } S_v}} \Big[\mathbb{P}\Big(\|\hat{C}'_u - C'_u\|_\infty > \frac{(\log n)^{\beta/2}}{3\sqrt{n}} \Big) + \mathbb{P}\Big(\|\hat{C}'_v - C'_v\|_\infty > \frac{(\log n)^{\beta/2}}{3\sqrt{n}} \Big) \tag{4.82}$$

$$+ \mathbb{P}\Big(\|C'_u - C'_v\|_\infty > \frac{(\log n)^{\beta/2}}{3\sqrt{n}} \Big) \Big] \tag{4.83}$$

$$\leq \sum_{u \in \mathcal{V}} \sum_{\substack{v \in \mathcal{V} \\ v,u \text{ in the same sub-region} \\ S_u \text{ borders on } S_v}} \Big[\mathbb{P}\Big(\|\hat{C}'_u - C'_u\|_\infty > \frac{(\log n)^{\beta/2}}{3\sqrt{n}} \Big) + \mathbb{P}\Big(\|\hat{C}'_v - C'_v\|_\infty > \frac{(\log n)^{\beta/2}}{3\sqrt{n}} \Big) \Big]. \tag{4.84}$$

By Lemma 21, for $1 < \beta < 2$, there exists a constant J, as $n \to \infty$,

$$\sum_{u \in \mathcal{V}} \sum_{\substack{v \in \mathcal{V} \\ v,u \text{ in the same sub-region} \\ S_u \text{ borders on } S_v}} \mathbb{P}\left(\|\hat{C}'_u - \hat{C}'_v\|_\infty > \frac{(\log n)^{\beta/2}}{\sqrt{n}} \right) \tag{4.85}$$

$$\leq |\mathcal{V}| \cdot 16(2\sqrt{n}+1)^2 \exp(-J \cdot (\log n)^\beta) \tag{4.86}$$

$$\leq 16n^2 \exp(-J \cdot (\log n)^\beta) \ \to 0. \tag{4.87}$$

Secondly, for $\forall u, v \in \mathcal{U}$ such that X_u, X_v are in different sub-regions, if $\mathsf{D}(u,v) \geq \frac{\log n}{\sqrt{n}}$, then

$$\|C'_u - C'_v\|_\infty \geq \frac{\log n}{2\sqrt{n}}. \tag{4.88}$$

If $\mathsf{D}(u,v) < \frac{\log n}{\sqrt{n}}$, under Assumption 13, there exists a constant K, such that

$$\|C'_u - C'_v\|_\infty \geq \frac{K \cdot \log n}{\sqrt{n}}. \tag{4.89}$$

So there exists a constant J, as $n \to \infty$,

$$\sum_{u \in \mathcal{V}} \sum_{\substack{v \in \mathcal{V} \\ v,u \text{ in different sub-regions}}} \mathbb{P}\left(\|\hat{C}'_u - \hat{C}'_v\|_\infty < \frac{(\log n)^{\beta/2}}{\sqrt{n}} \right) \tag{4.90}$$

$$\leq \sum_{u \in \mathcal{V}} \sum_{\substack{v \in \mathcal{V} \\ v,u \text{ in different sub-regions}}} \left[\mathbb{P}\left(\|C'_u - C'_v\|_\infty < \frac{3(\log n)^{\beta/2}}{\sqrt{n}} \right) + \mathbb{P}\left(\|\hat{C}'_u - C'_u\|_\infty > \frac{(\log n)^{\beta/2}}{\sqrt{n}} \right) \right. \tag{4.91}$$

$$\left. + \mathbb{P}\left(\|\hat{C}'_v - C'_v\|_\infty > \frac{(\log n)^{\beta/2}}{\sqrt{n}} \right) \right] \tag{4.92}$$

$$\leq \frac{2n^4}{(2\sqrt{n}+1)^4} \cdot \mathbb{P}\left(\|\hat{C}'_u - C'_u\|_\infty > \frac{(\log n)^{\beta/2}}{\sqrt{n}} \right) \tag{4.93}$$

$$\leq \frac{4n^4}{(2\sqrt{n}+1)^2} \cdot \exp(-J \cdot (\log n)^\beta) \ \to 0. \tag{4.94}$$

Thirdly, for $\forall u \in \mathcal{U}\backslash\mathcal{V}$, X_{u_+} and X_{u_-} are in different sub-regions, so by (4.88) and (4.89),

$$\sum_{u\in\mathcal{U}\backslash\mathcal{V}} \mathbb{P}\left(\|\hat{C}'_{u_+} - \hat{C}'_{u_-}\|_\infty < \frac{2(\log n)^{\beta/2}}{\sqrt{n}} \right) = 0. \tag{4.95}$$

Thus, there exists a constant J, as $n \to \infty$,

$$\mathbb{P}(\text{All pixels in } \mathcal{W} \text{ are clustered correctly}) \tag{4.96}$$

$$\geq 1 - 16n^2 \exp(-J\cdot(\log n)^\beta) - \frac{4n^4}{(2\sqrt{n}+1)^2}\cdot\exp(-J\cdot(\log n)^\beta) \tag{4.97}$$

$$\geq 1 - n^3 \cdot \exp(-J\cdot(\log n)^\beta) \ \to 1. \tag{4.98}$$

□

4.3.3 Example

Non-stationary Model

Follow the ideas in [HSK99], we define non-stationary Gaussian process as follows. For any pixels X_t and X_s, define non-stationary covariance between X_t and X_s to be

$$C(X_t, X_s) = \int_{\mathbb{R}^2} K_t(r) K_s(r) \mathrm{d}r, \tag{4.99}$$

where $K_t(\cdot)$ and $K_s(\cdot)$ are Gaussian kernel functions

$$K_t(r) = \frac{1}{2\pi|\Sigma_t|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(r-t)^T \Sigma_t^{-1} (r-t) \right] \tag{4.100}$$

and

$$K_s(r) = \frac{1}{2\pi|\Sigma_s|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(r-s)^T \Sigma_s^{-1} (r-s) \right]. \tag{4.101}$$

It is easy to check the covariance is non-negative definite. Then we create the non-stationary process by convoluting the white noise process $\phi(\cdot)$ with kernel function $K_t(\cdot)$,

$$X_t = \int_{\mathbb{R}^2} K_t(r)\, \mathrm{d}\phi(r) \quad \text{for} \quad t \in \mathcal{T}. \tag{4.102}$$

Next, we simplify the covariance matrix. Suppose $B \sim N(0, \Sigma_t)$ and $D \sim N(s, \Sigma_s)$, where B and D are independent. Let $g_B(\cdot)$, $g_D(\cdot)$ and $g_{D-B}(\cdot)$ denote the density functions of B, D and $D-B$. Similarly, $g_{B,D}(\cdot)$ and $g_{D-B,D}(\cdot)$ are join density functions. Then

$$C(X_t, X_s) = \int_{R^2} K_t(r) K_s(r) \mathrm{d}r = \int_{R^2} g_B(r-t) \cdot g_D(r)\, \mathrm{d}r \tag{4.103}$$

$$= \int_{R^2} g_{B,D}(r-t, r)\, \mathrm{d}r = \int_{R^2} g_{D-B,D}(t, r)\, \mathrm{d}r \tag{4.104}$$

$$= g_{D-B}(t) \cdot \int_R^2 g_D(r)\, \mathrm{d}r = g_{D-B}(t). \tag{4.105}$$

Since $D-B \sim N(s, \Sigma_t + \Sigma_s)$, we have

$$C(X_t, X_s) = g_{D-B}(t) = \frac{1}{2\pi |\Sigma_t + \Sigma_s|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (t-s)^T (\Sigma_t + \Sigma_s)^{-1} (t-s) \right]. \tag{4.106}$$

Assumptions and Theory on Example Model

Each pixel X_t has it own kernel function $K_t(\cdot)$, and the non-stationary process is controlled by kernel functions $\{K_t(\cdot)\}_{t \in \mathcal{T}}$. For each pixel X_t, it has Gaussian kernel function

$$K_t(r) = \frac{1}{2\pi |\Sigma_t|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (r-t)^T \Sigma_t^{-1} (r-t) \right]. \tag{4.107}$$

The only parameters of the non-stationary process are the covariance matrices $\{\Sigma_t\}_{t \in \mathcal{T}}$ of the kernel functions. Here we call $\{\Sigma_t\}_{t \in \mathcal{T}}$ as the "size" of the kernel functions. If all pixels X_t have the same "size" Σ_t, the field is stationary. For each pixel X_t, $t = (t_1, t_2)$ is a 2-dimension vector, so

its "size" Σ_t is a 2×2 matrix. Denote it as

$$\Sigma_t = \begin{pmatrix} a_t & b_t \\ c_t & d_t \end{pmatrix}. \tag{4.108}$$

For any pixels X_t and X_s, define

$$\mathsf{d}(s,t) = \max\{|a_t - a_s|, |b_t - b_s|, |c_t - c_s|, |d_t - d_s|\}. \tag{4.109}$$

Next, similar to Assumption 12 and Assumption 13, for the example kernel convolution model, we introduce assumptions directly on the "size" of kernel function.

Assumption 14. For $\forall X_t, X_s$ in the same sub-region,

$$\mathsf{d}(s,t) = O(\sqrt{\log n}\cdot \mathsf{D}(s,t)). \tag{4.110}$$

Assumption 15. For $\forall X_t, X_s$ in different sub-regions, if $\mathsf{D}(s,t) < \frac{\log n}{\sqrt{n}}$, there exists a constant K such that

$$\mathsf{d}(s,t) \geq \frac{K\cdot \log n}{\sqrt{n}}. \tag{4.111}$$

From Lemma 23, all conditions in Theorem 12 are satisfied, so Theorem 12 works here. Thus, under Assumption 14 and Assumption 15, single-linkage algorithm is consistent under the example kernel model.

4.4 Numerical Experiments

In our experiments, for the sake of stability, we use a form of size-constrained k-means [WCRS01, BBD00] , size-constrained single-linkage and size-constrained ward-linkage algorithms. The implementation of Python codes used in numerical experiments is available on

Github: https://github.com/zhenglin0266/Some-Theory-for-Texture-Segmentation

4.4.1 Synthetic Stationary Textures

In the section, we create several stationary random field models by moving average on Gaussian noise. Suppose white noise $Z_t \sim N(0,1)$. We generate four stationary random fields as follows

$$\text{Model 1:} \quad X_t = \sum_{i=-m}^{m} Z_{t+(i,i)}. \tag{4.112}$$

$$\text{Model 2:} \quad X_t = \sum_{i=-m}^{m} Z_{t+(-i,i)}. \tag{4.113}$$

$$\text{Model 3:} \quad X_t = \sum_{i=-m}^{m} Z_{t+(0,i)}. \tag{4.114}$$

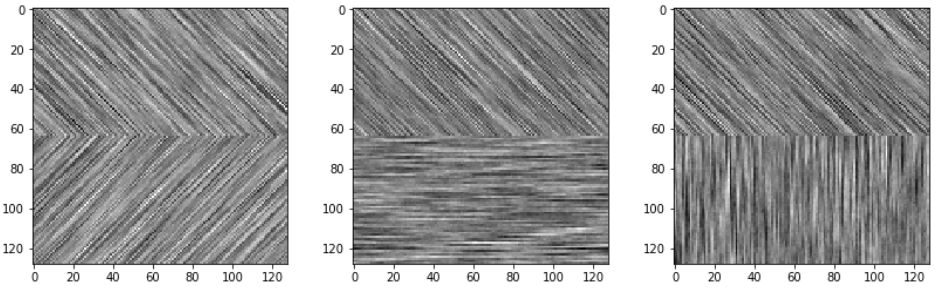
$$\text{Model 4:} \quad X_t = \sum_{i=-m}^{m} Z_{t+(i,0)}. \tag{4.115}$$

Based on above four models, after standardization and combination, we obtain six 128×128 mosaics, which are showed in Figure 4.1. Each mosaic contains two different textures, and it is divided by a straight line in the middle.

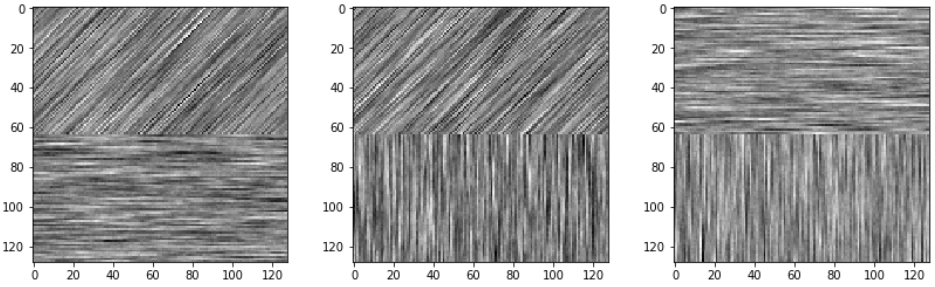
We apply $\{\hat{C}'_t\}_{t\in\mathcal{T}}$ as the observed features of pixels $\{X_t\}_{T\in\mathcal{T}}$. Consider three clustering algorithms: size-constrained single-linkage, size-constrained ward-linkage and size-constrained k-means algorithms. We present segmentation accuracy in Table 4.1. Single-linkage works but does not perform well, and ward-linkage improves it. K-means algorithm works best and segmentation accuracy is almost 1.

4.4.2 Natural Textures

In this section, We pick up textures from Brodatz album [Bro66].



(a) Model 1 vs Model 2 (b) Model 1 vs Model 3 (c) Model 1 vs Model 4



(d) Model 2 vs Model 3 (e) Model 2 vs Model 4 (f) Model 3 vs Model 4

Figure 4.1: GMRF texture mosaics

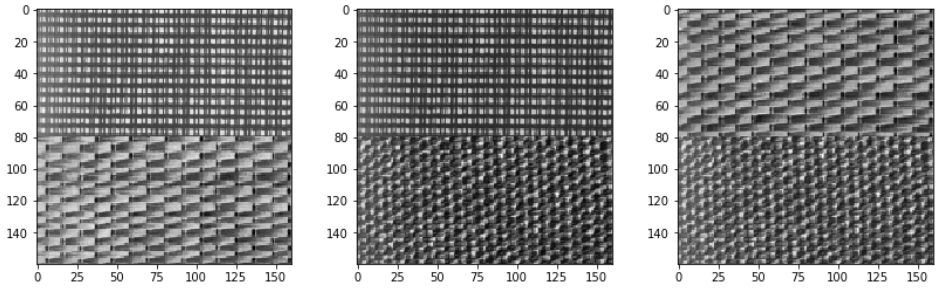
Table 4.1: Segmentation accuracy

Mosaic	Single-linkage	Ward-linkage	K-means
Model 1 vs Model 2	0.7362	0.9665	0.9868
Model 1 vs Model 3	0.9144	0.9632	0.9820
Model 1 vs Model 4	0.8329	0.9677	0.9867
Model 2 vs Model 3	0.9210	0.9657	0.9822
Model 2 vs Model 4	0.8190	0.9661	0.9868
Model 3 vs Model 4	0.7910	0.9626	0.9816
Mean Value	0.8358	0.9653	0.9844

Two Regions Divided by a Straight Line

We pick up three textures from Brodatz album [Bro66]: $D21$, $D55$ and $D77$. After standardization and combination, we obtain three 160×160 mosaics, which are showed in Figure 4.2. Each mosaic contains two different Brodatz textures, and they are divided by a straight

line in the middle.



(a) D21 vs D55 (b) D21 vs D77 (c) D55 vs D77

Figure 4.2: Texture mosaics

We apply $\{\hat{C}_t'\}_{t\in\mathcal{T}}$ as the observed features of pixels $\{X_t\}_{T\in\mathcal{T}}$. Consider three clustering algorithms: size-constrained single-linkage, size-constrained ward-linkage and size-constrained k-means algorithms. Segmentation results are shown in Figure 4.3. Also we present segmentation accuracy in Table 4.2. All three algorithms work perfectly and segmentation accuracy is almost 1.

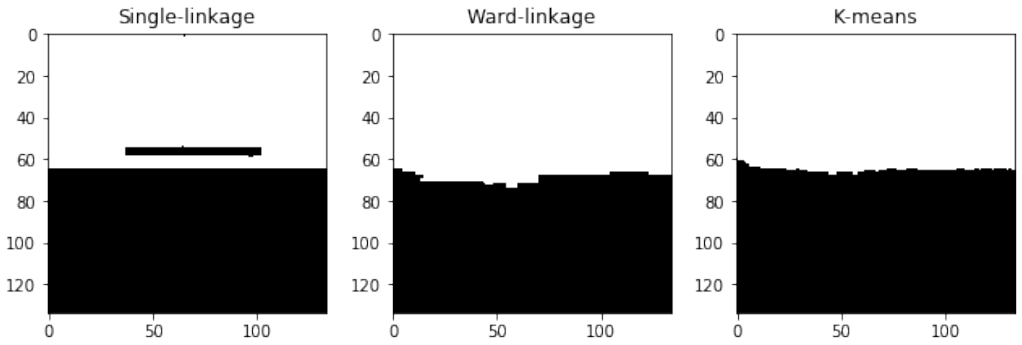
Table 4.2: Segmentation accuracy

Mosaics	Single-linkage	Ward-linkage	K-means
D21 vs D55	0.9703	0.9819	0.9891
D21 vs D77	0.9536	0.9592	0.9858
D55 vs D77	0.9914	0.9396	0.9928
Mean Value	0.9718	0.9602	0.9892

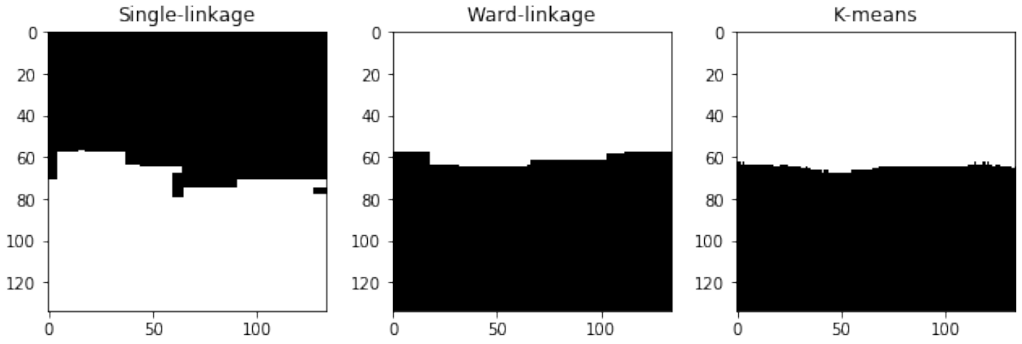
Two Regions Divided by a Curve

Same to Section 4.4.2, we still run simulations on textures $D21$, $D55$ and $D77$. Here in each mosaic, the textures are divided by a circle in the middle, as shown in Figure 4.4.

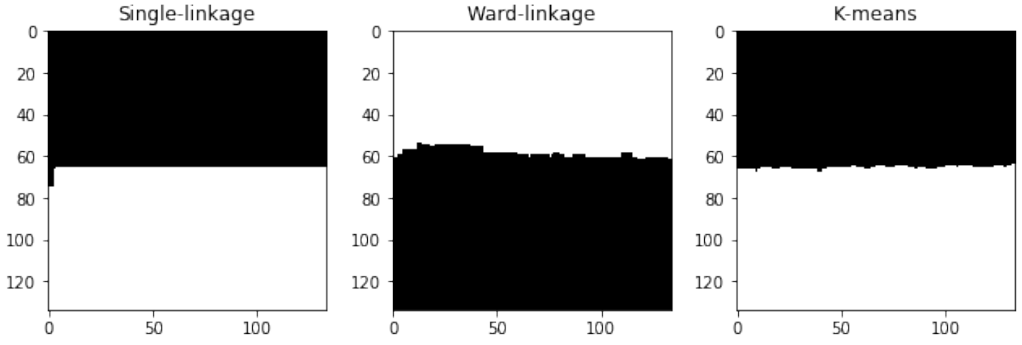
We apply $\{\hat{C}_t'\}_{t\in\mathcal{T}}$ as the observed features of pixels $\{X_t\}_{T\in\mathcal{T}}$. Consider three clustering algorithms: size-constrained single-linkage, size-constrained ward-linkage and size-constrained k-means algorithms. Segmentation results are shown in Figure 4.5. Also we present segmentation



(a) D21 vs D55

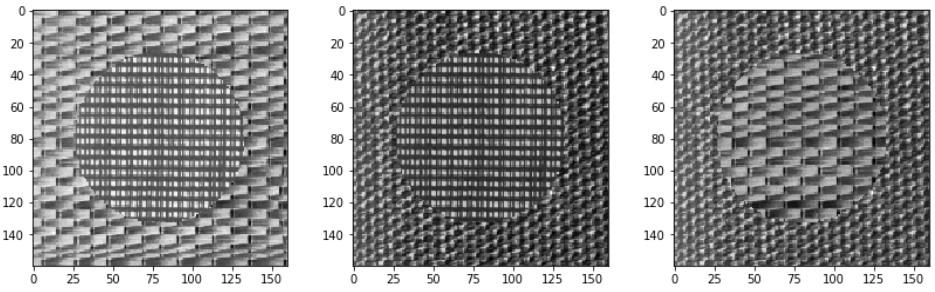


(b) D21 vs D77



(c) D55 vs D77

Figure 4.3: Segmentation results



(a) D21 vs D55 (b) D21 vs D77 (c) D55 vs D77

Figure 4.4: Texture mosaics

accuracy in Table 4.3. Single-linkage works but does not perform well, and ward-linkage improves it. K-means algorithm works perfectly and segmentation accuracy is almost 1.

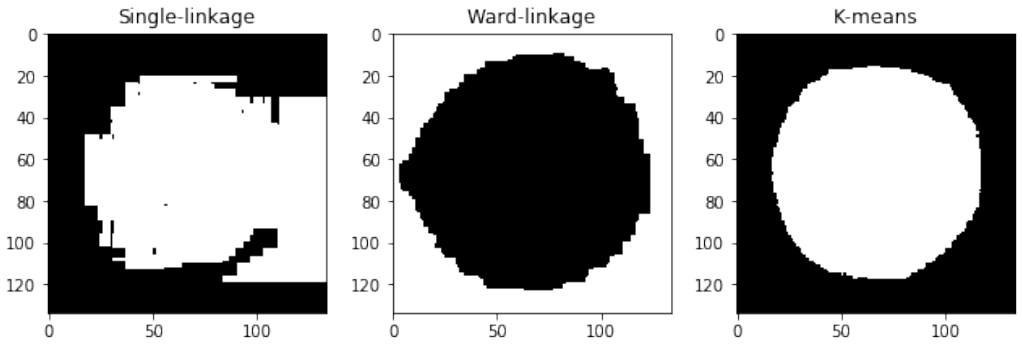
Table 4.3: Segmentation accuracy

Mosaic	Single-linkage	Ward-linkage	K-means
D21 vs D55	0.8093	0.9332	0.9562
D21 vs D77	0.8903	0.9563	0.9418
D55 vs D77	0.8295	0.8992	0.9739
Mean Value	0.8430	0.9296	0.9573

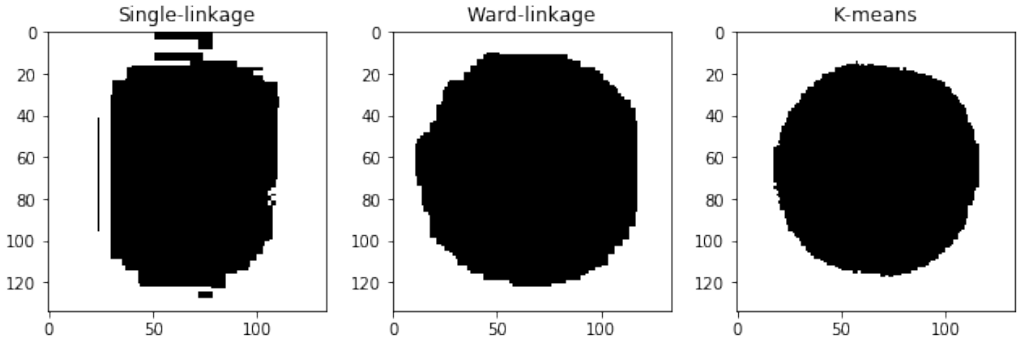
Multiple Regions

Here we construct three mosaics from eight textures in Brodatz album [Bro66]: $D4$, $D6$, $D20$, $D21$, $D34$, $D52$, $D55$ and $D77$. After standardization and combination, we obtain three 160×160 mosaics, which are showed in Figure 4.6. Each mosaic contains four different Brodatz textures, and they are divided by horizontal and vertical lines in the middle.

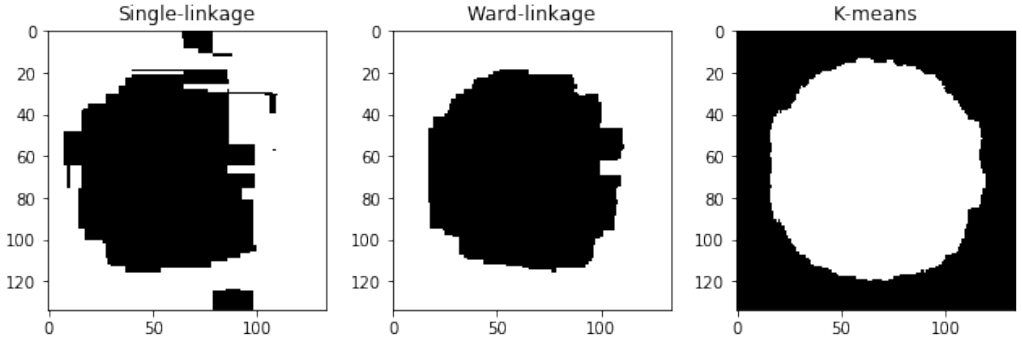
We apply $\{\hat{C}_t'\}_{t\in\mathcal{T}}$ as the observed features of pixels $\{X_t\}_{T\in\mathcal{T}}$. Consider three clustering algorithms: size-constrained single-linkage, size-constrained ward-linkage and size-constrained k-means algorithms. Segmentation results are shown in Figure 4.7. Also we present segmentation accuracy in Table 4.4. For multi-cluster mosaics, single-linkage works but does not perform well,



(a) D21 vs D55

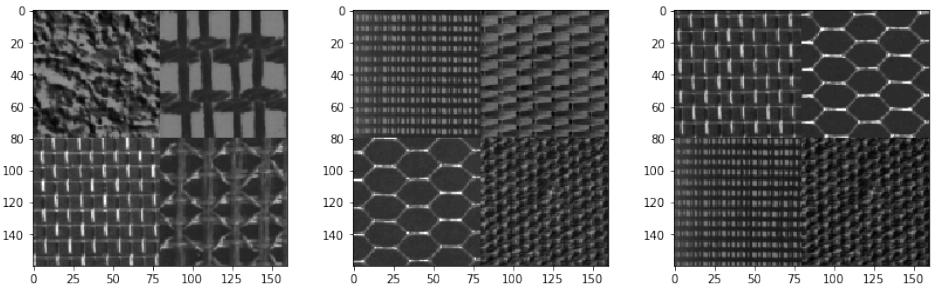


(b) D21 vs D77



(c) D55 vs D77

Figure 4.5: Segmentation results



(a) D4, D6, D20 & D52 (b) D21,D34, D55 & D77 (c) D6, D21, D34 & D77

Figure 4.6: Texture mosaics

and ward-linkage improves it. K-means algorithm works perfectly and segmentation accuracy is almost 1.

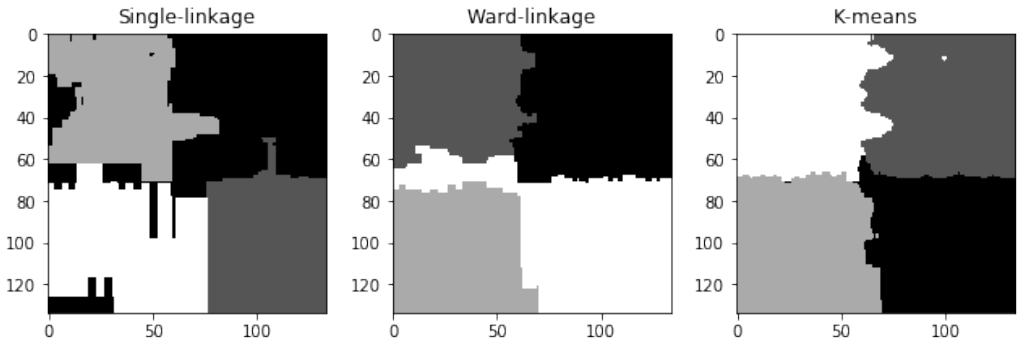
Table 4.4: Segmentation accuracy

Mosaic	Single-linkage	Ward-linkage	K-means
D04, D06, D20 & D52	0.8478	0.8985	0.9507
D21, D34, D55 & D77	0.7530	0.8969	0.9663
D06, D21, D34 & D77	0.9158	0.8767	0.9465
Mean Value	0.8389	0.8907	0.9545

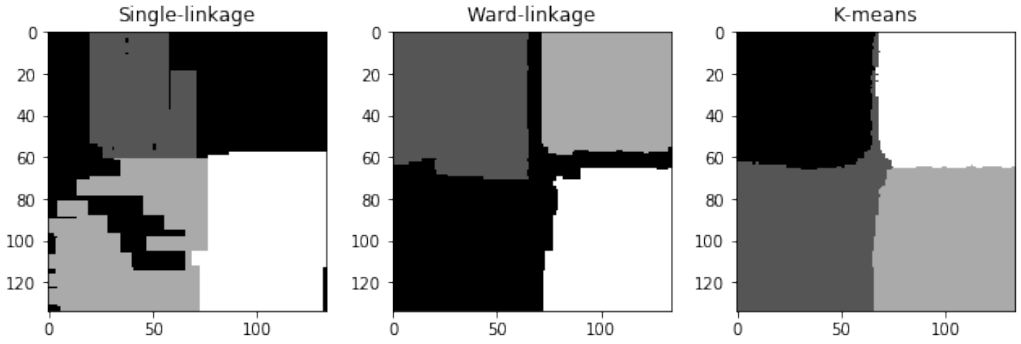
4.5 Discussion

For non-stationary textures, instead of single linkage clustering, we could also use DBSCAN algorithm. When MinPts = 2, DBSCAN is very similar to single-linkage algorithm, but DBSCAN includes a step for removing noisy observations. In this paper, these noisy observations could be the patches that overlap with the edge between the two sub-regions. In practice, in different settings and backgrounds, we can apply different MinPts values in DBSCAN algorithm.

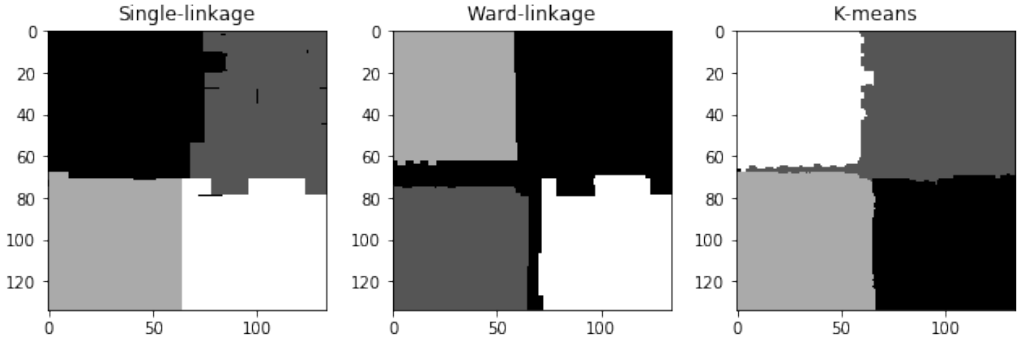
Also in Section 4.2 and Section 4.3, we only theoretically show the cases where there are only two sub-regions. Actually when there are more than two regions, the introduced algorithms also work. For example, we indicate the scene with four sub-regions in Section 4.4.2.



(a) D04, D06, D20 & D52



(b) D21, D34, D55 & D77



(c) D06, D21, D34 & D77

Figure 4.7: Segmentation results

In this paper, we do clustering based on local second moment information on patches. Actually beyond sample autocovariance, we could also use higher-order statistics or other features such as SIFT [Low99] for more complex textures that are not necessarily Gaussian, and speculate that similar results could also be obtained in these situations.

4.6 Acknowledgments

Chapter 4, in full, is a version of the paper "Some Theory for Texture Segmentation", Zheng, Lin. The manuscript has been submitted for publication in a major journal in electrical engineering. The dissertation author was the primary investigator and author of this material.

4.7 Appendix

Lemma 23. *For any pixels X_t, X_s in the image,*

$$\|C_t - C_s\|_\infty = O(\mathsf{d}(s,t)). \tag{4.116}$$

Proof. For any pixel X_t and $\forall i = (i_1, i_2) \in \mathcal{M}$,

$$C_t(i) = \text{Mean value of } \{C(X_t, X_{t+i}) \mid \text{both } X_t \text{ and } X_{t+i} \text{ are in } S_t\}. \tag{4.117}$$

Also we have

$$C(X_t, X_{t+i}) \tag{4.118}$$

$$= \frac{1}{(2\pi)^2 |\Sigma_t + \Sigma_{t+i}|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2} (\frac{i}{n}) (\Sigma_t + \Sigma_{t+i})^{-1} (\frac{i}{n})^T \right) \tag{4.119}$$

$$= \frac{1}{(2\pi)^2 \left| \begin{pmatrix} a_t + a_{t+i} & b_t + b_{t+i} \\ c_t + c_{t+i} & d_t + d_{t+i} \end{pmatrix} \right|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2} (\frac{i}{n}) \begin{pmatrix} a_t + a_{t+i} & b_t + b_{t+i} \\ c_t + c_{t+i} & d_t + d_{t+i} \end{pmatrix}^{-1} (\frac{i}{n})^T \right) \tag{4.120}$$

$$= \frac{1}{(2\pi)^2 \sqrt{(a_t + a_{t+i})(d_t + d_{t+i}) - (b_t + b_{t+i})(c_t + c_{t+i})}} \cdot \tag{4.121}$$

$$\exp\left(-\frac{1}{2(a_t + a_{t+i})(d_t + d_{t+i}) - (b_t + b_{t+i})(c_t + c_{t+i})} (\frac{i}{n}) \begin{pmatrix} d_t + d_{t+i} & -b_t - b_{t+i} \\ -c_t - c_{t+i} & a_t + a_{t+i} \end{pmatrix} (\frac{i}{n})^T \right). \tag{4.122}$$

$$\tag{4.123}$$

For $\forall t \in \mathcal{T}$ and $i \in \mathcal{M}$, let

$$I_{t,i} = (a_t + a_{t+i})(d_t + d_{t+i}) - (b_t + b_{t+i})(c_t + c_{t+i}), \tag{4.124}$$

then

$$C(X_t, X_{t+i}) \tag{4.125}$$

$$= \frac{1}{(2\pi)^2 \sqrt{I_{t,i}}} \cdot \exp\left(-\frac{1}{2I_{t,i}} (\frac{i}{n}) \begin{pmatrix} d_t + d_{t+i} & -b_t - b_{t+i} \\ -c_t - c_{t+i} & a_t + a_{t+i} \end{pmatrix} (\frac{i}{n})^T \right). \tag{4.126}$$

Also for $\forall t \in \mathcal{T}$ and $i \in \mathcal{M}$, let

$$R_{t,i} = \frac{1}{(2\pi)^2 \sqrt{I_{t,i}}} \tag{4.127}$$

and

$$Q_{t,i} = \exp\left(-\frac{1}{2I_{t,i}} (\frac{i}{n}) \begin{pmatrix} d_t + d_{t+i} & -b_t - b_{t+i} \\ -c_t - c_{t+i} & a_t + a_{t+i} \end{pmatrix} (\frac{i}{n})^T \right), \tag{4.128}$$

then

$$C(X_t, X_{t+i}) = R_{t,i} \cdot Q_{t,i}. \tag{4.129}$$

Similarly, for any pixel $X_s \neq X_t$,

$$C(X_s, X_{s+i}) = R_{s,i} \cdot Q_{s,i}. \tag{4.130}$$

Next, we work on the bound of $|C(X_t, X_{t+i}) - C(X_s, X_{s+i})|$. Since

$$|R_{t,i} - R_{s,i}| = \left| \frac{1}{(2\pi)^2 \sqrt{I_{t,i}}} - \frac{1}{(2\pi)^2 \sqrt{I_{s,i}}} \right| = O(\mathsf{d}(s,t)) \tag{4.131}$$

and

$$|Q_{t,i} - Q_{s,i}| = Q_{t,i} \cdot O\Big(\frac{\mathsf{d}(s,t)}{n}\Big) = O\Big(\frac{\mathsf{d}(s,t)}{n}\Big), \tag{4.132}$$

we have

$$\begin{aligned} |C(X_t, X_{t+i}) - C(X_s, X_{s+i})| &= |R_{t,i} \cdot Q_{t,i} - R_{s,i} \cdot Q_{s,i}| && (4.133) \\ &\leq R_{t,i} \cdot |Q_{t,i} - Q_{s,i}| + Q_{s,i} \cdot |R_{t,i} - R_{s,i}| && (4.134) \\ &\leq O\Big(\frac{\mathsf{d}(s,t)}{n}\Big) + O(\mathsf{d}(s,t)) && (4.135) \\ &= O(\mathsf{d}(s,t)). && (4.136) \end{aligned}$$

Then for any pixels X_t, X_s in the image, for $\forall i \in \mathcal{M}$,

$$|C_t(i) - C_s(i)| = O(\mathsf{d}(s,t)). \tag{4.137}$$

Thus,

$$\|C_t - C_s\|_\infty = O(\mathrm{d}(s,t)). \tag{4.138}$$

□

Bibliography

[AC11] Ery Arias-Castro. Clustering based on pairwise distances when the data is of mixed dimensions. *IEEE Transactions on Information Theory*, 57(3):1692–1706, 2011.

[ACBLV18] Ery Arias-Castro, Sébastien Bubeck, Gábor Lugosi, and Nicolas Verzelen. Detecting Markov random fields hidden in white noise. *Bernoulli*, 24(4B):3628–3656, 2018.

[ACCD11] Ery Arias-Castro, Emmanuel J. Candès, and Arnaud Durand. Detection of an anomalous cluster in a network. *The Annals of Statistics*, 39(1):278–304, 2011.

[ACCTW18] Ery Arias-Castro, Rui M Castro, Ervin Tánczos, and Meng Wang. Distribution-free detection of structured anomalies: Permutation and rank-based scans. *Journal of the American Statistical Association*, 113(522):789–801, 2018.

[ACDH05] Ery Arias-Castro, David L Donoho, and Xiaoming Huo. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Transactions on Information Theory*, 51(7):2402–2425, 2005.

[ACZ20] Ery Arias-Castro and Lin Zheng. Template matching and change point detection by M-estimation. *arXiv preprint arXiv:2009.04072*, 2020.

[AV07] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics*, pages 1027–1035, 2007.

[AY02] Olugbenga Ayinde and Yee-Hong Yang. Face recognition approach based on rank correlation of Gabor-filtered images. *Pattern Recognition*, 35(6):1275–1289, 2002.

[Bai97] Jushan Bai. Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, 79(4):551–563, 1997.

[BB01] Jasmine Banks and Mohammed Bennamoun. Reliability analysis of the rank transform for stereo matching. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 31(6):870–880, 2001.

[BBD00] Paul S Bradley, Kristin P Bennett, and Ayhan Demiriz. Constrained k-means clustering. *Technical Report*, Microsoft Research, Redmond, WA, 2000.

[BBKC99] Jasmine Banks, Mohammed Bennamoun, Kurt Kubik, and Peter Corke. A constraint to improve the reliability of stereo matching using the rank transform. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99*, volume 6, pages 3321–3324. IEEE, 1999.

[BD13] E Brodsky and Boris S Darkhovsky. *Nonparametric methods in change point problems*, volume 243. Springer Science+Business Media, 2013.

[BGV09] Jérémie Bigot, Fabrice Gamboa, and Myriam Vimond. Estimation of translation, rotation, and scaling between noisy images using the Fourier–Mellin transform. *SIAM Journal on Imaging Sciences*, 2(2):614–645, 2009.

[Bil99] Patrick Billingsley. *Convergence of probability measures*. John wiley & Sons, 1999.

[BK06] Michael V. Boutsikas and Markos V. Koutras. On the asymptotic distribution of the discrete scan statistic. *Journal of Applied Probability*, 43(4):1137–1154, 2006.

[BKC17] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[BKRW98] PJ Bickel, CAJ Klaassen, Y Ritov, and JA Wellner. *Efficient and adaptive estimation for semiparametric models*. Springer, 1998.

[BMP02] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[BN93] Michèle Basseville and Igor V Nikiforov. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[Bro66] Phil Brodatz. *Textures: A photographic album for artists and designers*. Dover Publications, 1966.

[Bru09] Roberto Brunelli. *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons, 2009.

[CC85] Rama Chellappa and Shankar Chatterjee. Classification of textures using Gaussian Markov random fields. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(4):959–963, 1985.

[CCL12] Jing Chen, Canhui Cai, and Cuihua Li. A multi-window stereo matching algorithm in rank tranform domain. In *2012 IEEE 11th International Conference on Signal Processing*, volume 2, pages 997–1000. IEEE, 2012.

[CD12] Olivier Collier and Arnak S Dalalyan. Minimax hypothesis testing for curve registration. *Electronic Journal of Statistics*, 6:1129–1154, 2012.

[CD15] Olivier Collier and Arnak S Dalalyan. Curve registration by nonparametric goodness-of-fit testing. *Journal of Statistical Planning and Inference*, 162:20–42, 2015.

[CG11] Jie Chen and Arjun K Gupta. *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science+Business Media, 2011.

[CH97] Miklós Csörgö and Lajos Horváth. *Limit theorems in change-point analysis*. John Wiley & Sons Inc, 1997.

[CJ83] George R Cross and Anil K Jain. Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:25–39, 1983.

[D̈91] L Dümbgen. The asymptotic behavior of some nonparametric change-point estimators. *The Annals of Statistics*, 19(3):1471–1495, 1991.

[Dar76] B S Darkhovskh. A nonparametric method for the a posteriori detection of the "disorder" time of a sequence of independent random variables. *Theory of Probability & Its Applications*, 21(1):178–183, 1976.

[Das99] Sanjoy Dasgupta. Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science*, pages 634–644. IEEE, 1999.

[Das10] Sanjoy Dasgupta. Hierarchical clustering with performance guarantees. In *Classification as a Tool for Research*, pages 3–14. Springer, 2010.

[DH95] Dennis Dunn and William E Higgins. Optimal Gabor filters for texture segmentation. *IEEE Transactions on Image Processing*, 4(7):947–964, 1995.

[DL05] Sanjoy Dasgupta and Philip M Long. Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences*, 70(4):555–569, 2005.

[DMM03] Agnès Desolneux, Lionel Moisan, and Jean-Michel Morel. Maximal meaningful events and applications to image analysis. *The Annals of Statistics*, 31(6):1822–1851, 2003.

[Dör11] Maik Döring. Convergence in distribution of multiple change point estimators. *Journal of Statistical Planning and Inference*, 141(7):2238–2248, 2011.

[Dra88] David Draper. Rank-based robust analysis of linear models. I. Exposition and review. *Statistical Science*, 3(2):239–271, 1988.

[Dud67] Richard M Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

[Fer94] Dietmar Ferger. Change-point estimators in case of small disorders. *Journal of Statistical Planning and Inference*, 40(1):33–49, 1994.

[Fer01] Dietmar Ferger. Exponential and polynomial tailbounds for change-point estimators. *Journal of Statistical Planning and Inference*, 92(1-2):73–109, 2001.

[Fer04] Dietmar Ferger. A continuous mapping theorem for the argmax-functional in the non-unique case. *Statistica Neerlandica*, 58(1):83–96, 2004.

[FMS14] Klaus Frick, Axel Munk, and Hannes Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society - Series B: Statistical Methodology*, 76(3):495–580, 2014.

[GB12] Joseph Glaz and Narayanaswamy Balakrishnan. *Scan statistics and applications*. Springer Science+Business Media, 2012.

[GC11] Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric statistical inference*. Springer, 2011.

[Ger18] Carina Gerstenberger. Robust Wilcoxon-type estimation of change-point location under short-range dependence. *Journal of Time Series Analysis*, 39(1):90–104, 2018.

[GF16] Christian Galea and Reuben A Farrugia. Face photo-sketch recognition using local and global texture descriptors. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 2240–2244. IEEE, 2016.

[GG86] Stuart Geman and Christine Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, volume 1, pages 1496–1517, 1986.

[GG12] Nan Geng and Qin Gou. Adaptive color stereo matching based on rank transform. In *2012 International Conference on Industrial Control and Electronics Engineering*, pages 1701–1704. IEEE, 2012.

[GH98] Edit Gombay and Marie Hušková. Rank based estimators of the change-point. *Journal of Statistical Planning and Inference*, 67(1):137–154, 1998.

[GKS96] Liudas Giraitis, Hira L Koul, and Donatas Surgailis. Asymptotic normality of regression estimators with long memory errors. *Statistics & Probability Letters*, 29(4):317–335, 1996.

[GLM07] Fabrice Gamboa, Jean-Michel Loubès, and Elie Maza. Semi-parametric estimation of shifts. *Electronic Journal of Statistics*, 1:616–640, 2007.

[GN16] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge: Cambridge University Press, 2016.

[GNW01] Joseph Glaz, Joseph Naus, and Sylvan Wallenstein. *Scan statistics*. Springer, 2001.

[GPK02] Simona E Grigorescu, Nicolai Petkov, and Peter Kruizinga. Comparison of texture features based on Gabor filters. *IEEE Transactions on Image Processing*, 11(10):1160–1167, 2002.

[GPW09] Joseph Glaz, Vladimir Pozdnyakov, and Sylvan Wallenstein. *Scan statistics: methods and applications*. Springer Science+Business Media, 2009.

[GSBB03] Meirav Galun, Eitan Sharon, Ronen Basri, and Achi Brandt. Texture segmentation by multiscale aggregation of filter responses and shape elements. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 716–725. IEEE, 2003.

[GZ04] Joseph Glaz and Zhenkui Zhang. Multiple window discrete scan statistics. *Journal of Applied Statistics*, 31(8):967–980, 2004.

[H68] Jaroslav Hájek. Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, 39(2):325–346, 1968.

[HH01] Joseph V Hajnal and Derek LG Hill. *Medical image registration*. CRC press, 2001.

[Hin70] D Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17, 1970.

[HJ10] Peter Hall and Jiashun Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686–1732, 2010.

[HK13] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, pages 11–20, 2013.

[HL01] Erik Hjelmås and Boon Kee Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236–274, 2001.

[HM90] W Härdle and JS Marron. Semiparametric comparison of regression curves. *The Annals of Statistics*, 18(1):63–89, 1990.

[HM10] Thomas P Hettmansperger and Joseph W McKean. *Robust nonparametric statistical methods*. Boca Raton, FL, USA: CRC Press, 2010.

[HP06] G Haiman and C Preda. Estimation for the distribution of two-dimensional discrete scan statistics. *Methodology and Computing in Applied Probability*, 8(3):373–382, 2006.

[HPB98] Thomas Hofmann, Jan Puzicha, and Joachim M Buhmann. Unsupervised texture segmentation in a deterministic annealing framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):803–818, 1998.

[HS10] Heping He and Thomas A Severini. Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, 16(3):759–779, 2010.

[HSK99] Dave Higdon, Jenise Swall, and J Kern. Non-stationary spatial modeling. *Bayesian Statistics*, 6(1):761–768, 1999.

[Hušk97] M Hušková. Limit theorems for rank statistics. *Statistics & Probability Letters*, 32(1):45–55, 1997.

[HW88] Siegfried Heiler and Reinhart Willers. Asymptotic normality of R-estimates in the linear model. *Statistics: A Journal of Theoretical and Applied Statistics*, 19(2):173–184, 1988.

[JCL10] X Jessie Jeng, T Tony Cai, and Hongzhe Li. Optimal sparse segment identification with application in copy number variation analysis. *Journal of the American Statistical Association*, 105(491):1156–1166, 2010.

[JF91] Anil K Jain and Farshid Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, 24(12):1167–1186, 1991.

[Jia02] Tiefeng Jiang. Maxima of partial sums indexed by geometrical structures. *The Annals of Probability*, 30(4):1854–1892, 2002.

[Jur71] Jana Jurečková. Nonparametric estimate of regression coefficients. *The Annals of Mathematical Statistics*, 42(4):1328–1338, 1971.

[Kab11] Zakhar Kabluchko. Extremes of the standardized Gaussian noise. *Stochastic Processes and their Applications*, 121(3):515–533, 2011.

[KD09] Georgios Kordelas and Petros Daras. Robust SIFT-based feature matching using Kendall's rank correlation measure. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 325–328. IEEE, 2009.

[KE95] Alois Kneip and Joachim Engel. Model estimation in nonlinear regression under shape invariance. *The Annals of Statistics*, 23(2):551–570, 1995.

[KG88] Alois Kneip and Theo Gasser. Convergence and consistency results for self-modeling nonlinear regression. *The Annals of Statistics*, 16(1):82–112, 1988.

[KG92] Alois Kneip and Theo Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, 20(3):1266–1305, 1992.

[KM93] Hira L Koul and Kanchan Mukherjee. Asymptotics of R-, MD- and LAD-estimators in linear regression models with long range dependent errors. *Probability Theory and Related Fields*, 95:535–553, 1993.

[KMW20] Claudia König, Axel Munk, and Frank Werner. Multidimensional multiscale scanning in exponential families: Limit theory and statistical consequences. *The Annals of Statistics*, 48(2):655–678, 2020.

[Kor88] AP Korostelev. On minimax estimation of a discontinuous signal. *Theory of Probability & Its Applications*, 32(4):727–730, 1988.

[Kos08] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer, New York, 2008.

[Kou17] Jiyao Kou. Identifying the support of rectangular signals in Gaussian noise. *arXiv preprint arXiv:1703.06226*, 2017.

[KP90] Jeankyung Kim and David Pollard. Cube root asymptotics. *The Annals of Statistics*, 18(1):191–219, 1990.

[KSC08] Cherrie HT Kong, Christian Soeller, and Mark B Cannell. Increasing sensitivity of Ca^{2+} spark detection in noisy images by application of a matched-filter object detection algorithm. *Biophysical Journal*, 95(12):6016–6024, 2008.

[KSS04] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 454–462. IEEE, 2004.

[LBM09] Yan Lan, Moulinath Banerjee, and George Michailidis. Change-point estimation under adaptive sampling. *The Annals of Statistics*, 37(4):1752–1791, 2009.

[LCF+19] Li Liu, Jie Chen, Paul Fieguth, Guoying Zhao, Rama Chellappa, and Matti Pietikäinen. From BoW to CNN: Two decades of texture representation for texture classification. *International Journal of Computer Vision*, 127(1):74–109, 2019.

[Leh06] Erich Leo Lehmann. *Nonparametrics: Statistical methods based on ranks*. Springer, New York, 2006.

[Low99] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157. IEEE, 1999.

[LR05] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science+Business Media, 2005.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[LSM72] WH Lawton, EA Sylvestre, and MS Maggio. Self modeling nonlinear regression. *Technometrics*, 14(3):513–532, 1972.

[LYFLLC15] Alexandre Lung-Yut-Fong, Céline Lévy-Leduc, and Olivier Cappé. Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la Société Française de Statistique*, 156(4):133–162, 2015.

[Mau18] René Mauer. *Least squares estimation in multiple change-point models*. PhD thesis, Technische Universität Dresden, 2018.

[MBLS01] Jitendra Malik, Serge Belongie, Thomas Leung, and Jianbo Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27, 2001.

[MC91] B S Manjunath and Rama Chellappa. Unsupervised texture segmentation using Markov random field models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):478–482, 1991.

[MJ92] Jianchang Mao and Anil K Jain. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition*, 25(2):173–188, 1992.

[MNA16] Fausto Milletari, Nassir Navab, and Seyed Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[NJW02] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2002.

[NW04] Joseph I Naus and Sylvan Wallenstein. Multiple window and cluster size scan procedures. *Methodology and Computing in Applied Probability*, 6(4):389–400, 2004.

[PGKS05] Vladimir Pozdnyakov, Joseph Glaz, Martin Kulldorff, and J Michael Steele. A martingale approach to scan statistics. *Annals of the Institute of Statistical Mathematics*, 57(1):21–37, 2005.

[PP93] Nikhil R Pal and Sankar K Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

[PS06] Christopher J Paciorek and Mark J Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics: The Official Journal of the International Environmetrics Society*, 17(5):483–506, 2006.

[PWB+19] Amelia Perry, Jonathan Weed, Afonso S Bandeira, Philippe Rigollet, and Amit Singer. The sample complexity of multireference alignment. *SIAM Journal on Mathematics of Data Science*, 1(3):497–517, 2019.

[PWBM18] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Message-passing algorithms for synchronization problems over compact groups. *Communications on Pure and Applied Mathematics*, 71(11):2275–2322, 2018.

[PWM18] Katharina Proksch, Frank Werner, and Axel Munk. Multiscale scanning in inverse problems. *The Annals of Statistics*, 46(6B):3569–3602, 2018.

[RD93] Todd R Reed and JM Hans Dubuf. A review of recent texture segmentation and feature extraction techniques. *CVGIP: Image Understanding*, 57(3):359–372, 1993.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[RH99] Trygve Randen and John Hakon Husoy. Texture segmentation using filters with optimized energy separation. *IEEE Transactions on Image Processing*, 8(4):571–582, 1999.

[RH05] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. London: Chapman and Hall–CRC Press, 2005.

[RV13] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013.

[SAC16] James Sharpnack and Ery Arias-Castro. Exact asymptotics for the scan statistic and fast alternatives. *Electronic Journal of Statistics*, 10(2):2641–2684, 2016.

[SDP13] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical imaging*, 32(7):1153–1190, 2013.

[Sha95] Qi Man Shao. On a conjecture of révész. *Proceedings of the American Mathematical Society*, 123(2):575–582, 1995.

[Sie13] David Siegmund. *Sequential analysis: tests and confidence intervals*. Springer Science+Business Media, 2013.

[SM00] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[SS11] Emilio Seijo and Bodhisattva Sen. Change-point in stochastic design regression and the bootstrap. *The Annals of Statistics*, 39(3):1580–1607, 2011.

[ŠSH99] Zbyněk Šidák, Pranab K Sen, and Jaroslav Hájek. *Theory of rank tests*. Academic Press, 2nd edition, 1999.

[SV95] D. Siegmund and E. S. Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, 23(1):255–271, 1995.

[SWP05] Thomas Serre, Lior Wolf, and Tomaso Poggio. Object recognition with features inspired by visual cortex. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 994–1000. IEEE, 2005.

[Tal96] Michel Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126(3):505–563, 1996.

[TIR11] Thomas Trigano, Uri Isserles, and Ya'acov Ritov. Semiparametric curve alignment and shift density estimation for biological data. *IEEE Transactions on Signal Processing*, 59(5):1970–1984, 2011.

[TOV20] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

[Tsy09] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science+Business Media, 2009.

[Tur60] George Turin. An introduction to matched filters. *IRE Transactions on Information Theory*, 6(3):311–329, 1960.

[Uns95] Michael Unser. Texture classification and segmentation using wavelet frames. *IEEE Transactions on Image Processing*, 4(11):1549–1560, 1995.

[VdV98] A W Van der Vaart. *Asymptotic statistics*. Cambridge: Cambridge University Press, 1998.

[vdVW96] Aad W van der Vaart and Jon A Wellner. *Weak convergence and empirical processes with applications to statistics*. Springer, 1996.

[Ver10a] Nicolas Verzelen. Adaptive estimation of stationary Gaussian fields. *The Annals of Statistics*, 38(3):1363–1402, 2010.

[Ver10b] Nicolas Verzelen. High-dimensional Gaussian model selection on a Gaussian design. In *Annales de I'IHP Probabilités et Statistiques*, volume 46, pages 480–524, 2010.

[Vim10] Myriam Vimond. Efficient estimation for a subclass of shape invariant models. *The Annals of Statistics*, 38(3):1885–1912, 2010.

[VV09] Nicolas Verzelen and Fanny Villers. Tests for Gaussian graphical models. *Computational Statistics & Data Analysis*, 53(5):1894–1905, 2009.

[VW04] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[Wal10] Guenther Walther. Optimal and fast detection of spatial clusters with scan statistics. *The Annals of Statistics*, 38(2):1010–1033, 2010.

[WCRS01] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, volume 1, pages 577–584, 2001.

[WG99] Kongming Wang and Theo Gasser. Synchronizing sample curves nonparametrically. *The Annals of Statistics*, 27(2):439–460, 1999.

[WG14] Xiao Wang and Joseph Glaz. Variable window scan statistics for normal data. *Communications in Statistics - Theory and Methods*, 43(10-12):2489–2504, 2014.

[WH96] Thomas P Weldon and William E Higgins. Design of multiple Gabor filters for texture segmentation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 4, pages 2243–2246. IEEE, 1996.

[WS13] Lanhui Wang and Amit Singer. Exact and stable recovery of rotations for robust synchronization. *Information and Inference: A Journal of the IMA*, 2(2):145–193, 2013.

[WWZ20] Yunlong Wang, Zhaojun Wang, and Xuemin Zi. Rank-based multiple change-point detection. *Communications in Statistics-Theory and Methods*, 49(14):3438–3454, 2020.

[XXJ+20] Xin Xiong, Qing Xu, Guowang Jin, Hongmin Zhang, and Xin Gao. Rank-based local self-similarity descriptor for optical-to-SAR image matching. *IEEE Geoscience and Remote Sensing Letters*, 17(10):1742–1746, 2020.

[YA89] Yi-Ching Yao and Siu-Tong Au. Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A*, 51(3):370–381, 1989.

[ZF03] Barbara Zitova and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003.

[Zha06] Yu Jin Zhang. *Advances in image and video segmentation*. IRM Press, 2006.

[ZWM98] Song Chun Zhu, Yingnian Wu, and David Mumford. Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998.