Log-concave Density Estimation and Bump Hunting for i.i.d. Observations

Inauguraldissertation
zur Erlangung der Doktorwürde
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

und

der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von

Kaspar Rufibach

von Guttannen BE

Leiter der Arbeit:

Prof. Dr. rer. nat. L. Dümbgen
Institut für mathematische Statistik und
Versicherungslehre der Universität Bern

und

Prof. Dr. rer. nat. A. Munk
Institut für mathematische Stochastik der
Georg-August-Universität zu Göttingen

Log-concave Density Estimation and Bump Hunting for i.i.d. Observations

Inauguraldissertation
zur Erlangung der Doktorwürde
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

und

der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von

Kaspar Rufibach

von Guttannen BE

Leiter der Arbeit:

Prof. Dr. rer. nat. L. Dümbgen
Institut für mathematische Statistik und
Versicherungslehre der Universität Bern

und

Prof. Dr. rer. nat. A. Munk
Institut für mathematische Stochastik der
Georg-August-Universität zu Göttingen

Von den beteiligten Fakultäten angenommen.

Bern, 18. Mai 2006 — Die Dekane:
Prof. Dr. Paul Messerli

Göttingen, 18. Mai 2006 — Prof. Dr. Ina Kersten

Contents

Abstract

The first part of this thesis is concerned with the estimation of a univariate density f nonparametrically via maximum likelihood from a given ordered sample $X_1, \ldots, X_n$ of independent and identically distributed random variables having distribution function F. It is well known that such an estimator $\widehat{f}_n$ does only exist if additional assumptions are made, i.e. the maximum likelihood function needs some regularization. We will impose the shape constraint of log-concavity, a natural generalization of many parametric densities such as Normal, Gamma, Laplace or Generalized Pareto. We show that such an estimator exists, is unique and that the estimated log-density $\widehat{\varphi}_n$ is supported by $[X_1, X_n]$ and piecewise linear with knots at some of the observation points. We provide two characterizations of the estimator, both of them involving the empirical distribution function of the sample. The first of these characterizations is essential for the proof of our main result: a uniform rate of convergence of $\widehat{f}_n$ on a fixed compact interval T as n goes to infinity. Under standard assumptions this rate is of probabilistic order $(\log(n)/n)^{2/5}$. But we also prove adaptivity with respect to the unknown smoothness of the underlying density f in terms of Hölder-continuity.

The result above, together with considerations about the modulus of continuity of a uniform empirical process, can be used to show that the integral of $\widehat{f}_n$, the distribution function estimator $\widehat{F}_n$, is asymptotically equivalent to the empirical distribution function $\mathbb{F}_n$ of the sample. Consequently, $\widehat{F}_n$ can be viewed as an efficient smoother of the empirical distribution function, if the underlying density is indeed log-concave. Log-concavity of the density function f immediately implies potentially desired properties for functions derived from it, such as the tail function $1 - F$ or the hazard rate function $f/(1 - F)$. The first is again log-concave and the latter is monotone non-decreasing. As an application of the above theorem we give an upper bound for the uniform rate of convergence for a monotone hazard rate estimator.

Then, methods are provided to find $\widehat{f}_n$ numerically via iterative algorithms. To this end, the piecewise linearity of $\widehat{\varphi}_n$ is exploited to embed the problem of minimizing the negative log-likelihood functional into a high- but finite-dimensional convex optimization framework. We compare four different algorithms, including two standard

approaches from convex optimization. It turns out that a suitable modification of the iterative convex minorant algorithm is very efficient in solving this optimization problem.

The second part is devoted to bump hunting, a term used for procedures to identify regions where a density exhibits either a convex or concave behavior. For certain reasons we reformulate the problem in that we seek to detect regions of log-convexity and log-concavity. First we analyze a specific two-parameter model regarding its power properties in a test for log-concavity vs. log-convexity. Then we use this model to approximate the density on all intervals spanned by a pair of observations. All these local tests are then combined in a global multiscale statistic, yielding two sets of intervals whereon one can claim with probability at least $1 - \alpha$ as n tends to infinity that the underlying density is either log-convex or log-concave. We further introduce an additive correction term into the global test statistic in order to prevent it to be dominated by the local statistics stemming from small intervals. The chosen multiscale approach ensures that all statements hold simultaneously. From the collections of the above intervals a lower bound for the number of bumps and dips of the underlying density can be derived. To our knowledge, this is the first multiscale test in density estimation exhibiting all these properties (asymptotically holding the significance level, simultaneous statements, additive correction term) at once. However, the proposed method relies on an unproven assumption about the quantiles of the limiting distribution and is therefore a first approach to the problem. A detailed theoretical analysis of its properties, especially those of the limiting distribution of the multiscale test statistic, is still lacking.
Assuming that a non-degenerate limiting distribution for the multiscale test statistic exists we provide its quantiles, gained from numerical simulations. We also describe a worst case distribution to input in the statistic when doing Monte-Carlo simulations. The procedure is illustrated with some examples.

Preface

First and foremost I would like to thank the main supervisor of this work, Lutz Dümbgen, for his guidance, patience, his always-open office door, sharing his ideas with me and finally proposing me to work on a very exciting, fruitful and challenging topic on the cutting edge of today's statistical research. Even more, if possible, I am grateful to him for making possible visits to San Francisco, Karlsruhe, Oslo, Stanford and Göttingen.
I thank Axel Munk for being the co-supervisor and providing me the possibility to spend several weeks in Göttingen.
Sincere thanks is given to Geurt Jongbloed for taking the Koreferat at short notice.
I am indebted to the people at the SAKK Coordinating Centre, especially Shu-Fang Hsu Schmitz, for giving me the freedom to handle my duties there very flexibly, especially in Summer 2005.
My fellow colleagues at IMSV deserve special mentioning, for the pleasant atmosphere at work and endless assistance and support of many different kinds: mathematics, informatics or things of real life. Not to forget playing cards! Above all, Sämi Müller gave me invaluable assistance in many respects, but especially while preparing my successful postdoc grant application.
Special thanks to Günther Walther for inviting me to Stanford.
Finally, I owe invaluable thanks to my parents for making my studies possible and to Nadja, for ♡ and enduring my bad temper when things sometimes did not progress as I wanted them to.
This work was supported by the Swiss National Science Foundation.

Kaspar Rufibach
Bern, Mai 2006

Chapter 1

Introduction

1.1 Density estimation in general

The first part of this thesis is concerned with a standard problem in statistics: The estimation of an unknown univariate probability density function (pdf) f. Typically, one considers a sample $X_1, \ldots, X_n$ of independent, identically distributed real-valued random variables with common density f and the aim is to get an estimate

$$\widehat{f} = \widehat{f}(.; X_1, \ldots, X_n)$$

for f from the data. Denote by $\mathbb{F}_n$ the empirical distribution function of the sample $X_1, \ldots, X_n$. In what follows, all asymptotic statements are to be understood when the sample size n tends to infinity.
The following sections review some general methods in density estimation.

1.2 Kernel density estimation

A standard tool in nonparametric density estimation are kernel estimators $\widehat{f}_{nh}$,

$$\widehat{f}_{nh}(x) \;:=\; \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k\Big(\frac{x - X_i}{h}\Big), \quad x \in \mathbb{R}$$

where $h > 0$ is the bandwidth and $k : \mathbb{R} \to \mathbb{R}$ the kernel function. The main advantage of kernel estimators is that they are easy computable, independent from the assumptions made on f. However, in general using kernels poses at least one major problem, namely the selection of a kernel and an appropriate bandwidth in order to

avoid oversmoothing (hiding relevant features of f, e.g. modes) or undersmoothing (producing artifacts). Asymptotic results under standard assumptions on the kernel typically depend on the smoothness of f. Suppose f is m-times differentiable and choose the bandwidth $h = h(n)$ in order to balance the variance and the bias term in the mean squared error. The rate of convergence of $\widehat{f}_{nh} - f$ at a fixed point is then $O_{\mathrm{p}}(n^{-m/(2m+1)})$, a rate that approaches the "parametric rate" $n^{-1/2}$ (see below) as $m \to \infty$.

1.3 Parametric density estimation

Here and subsequently we will concentrate on methods for density estimation relying on the maximum likelihood principle. Therefore introduce the negative maximum log-likelihood functional as:

$$\begin{aligned} L_n(f) &:= -n \int \log f(x) \, \mathrm{d}\mathbb{F}_n(x) \\ &= -\sum_{i=1}^{n} \log f(X_i). \end{aligned}$$

In classical parametric estimation, f is assumed to belong to a class $\mathcal{F}_1$ of densities, where

$$\mathcal{F}_1 = \{g_{\boldsymbol{\theta}} \ : \ \boldsymbol{\theta} \in \Theta\}$$

with a given subset Θ of $\mathbb{R}^d$ and $\theta \mapsto g_{\boldsymbol{\theta}}$ a continuous function from Θ into $\mathrm{L}_1(\mathbb{R})$. The dimension d is usually fixed and small compared to n. Our problem of estimating f then reduces to estimate $\boldsymbol{\theta} \in \Theta$ from the data $X_1, \ldots, X_n$, via minimizing $L_n(g_{\boldsymbol{\theta}})$ over all $\boldsymbol{\theta} \in \Theta$:

$$\widehat{\boldsymbol{\theta}}_n := \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} L_n(g_{\boldsymbol{\theta}}).$$

If possible this can be done analytically, otherwise numerically. Under standard assumptions the rate of convergence of $\widehat{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}$ is of order $O_{\mathrm{p}}(n^{-1/2})$.

1.4 Roughness penalized density estimation

When talking about nonparametric maximum likelihood estimation, it is not evident how to actually get an estimator. One can make $L_n(g)$ arbitrary small over all $g \in \mathrm{L}_1(\mathbb{R})$ that are continuous, i.e. the continuity assumption is too weak, the class of densities over which $L_n(g)$ is minimized needs to be made smaller. A general approach to achieve this is via penalizing. Add a penalty term $R = R(g)$ to the negative log-likelihood functional to get a penalized version $L_n^P(g, R)$ of the maximum log-likelihood functional:

$$L_n^P(g, R, \lambda) \;:=\; L_n(g) + \lambda R(g)$$

where $\lambda > 0$ is a Lagrange-multiplier sequence decreasing to 0. Roughness penalized density estimators are then defined as

$$\widehat{f}_{n,2}(R, C, \lambda) \;:=\; \operatorname*{arg\,min}_{g \in \mathcal{F}_2(R,C)} L_n^P(g, R, \lambda)$$

where $\mathcal{F}_2(R, C)$ is the following family of functions:

$$\mathcal{F}_2(R, C) \;:=\; \{f \;:\; f \text{ is a continuous pdf and } R(f) \leq C\}$$

for $C \in (0, \infty)$ a fixed constant. In principle, λ may be chosen such that

$$R(\widehat{f}_{n,2}(R, C, \lambda)) = C.$$

Since C is usually unknown, λ is often determined by other means.
One of the most famous choices for R is the first roughness penalty functional by Good (1971):

$$R_G(g) := \int_{-\infty}^{\infty} \left| \frac{\mathrm{d}}{\mathrm{d}x} \sqrt{g(x)} \right|^2 \mathrm{d}x,$$

where $R_G(g) = \infty$ if the derivative of $\sqrt{g}$ is not square integrable on $\mathbb{R}$. According to Eggermont and LaRiccia (2001), R_G has remarkably good practical and theoretical properties. For instance, under the assumptions

$$R(f) < \infty \qquad \int_{\mathbb{R}} f''(x)\,\mathrm{d}x < \infty \qquad \int_{\mathbb{R}} |x|^m f(x)\,\mathrm{d}x < \infty \text{ for some } m > \kappa > 1$$

on the true density f, Eggermont and LaRiccia (1999) prove that convergence in the space $\mathrm{L}_1(\mathbb{R})$ happens at a rate of $O_{\mathrm{p}}(n^{-2/5})$, so one that is similar to that for kernel estimators under comparable assumptions.

1.5 Density estimation under qualitative assumptions

A different approach to density estimation is to assume certain shape restrictions for f, such as monotonicity, unimodality or convexity. These restrictions are often plausible, sometimes even theoretically justified and they share the following common property. Defining the estimators as

$$\widehat{f}_{n,3} := \arg\min_{g \in \mathcal{F}_3} L_n(g)$$

where $\mathcal{F}_3$ is the family of densities satisfying the given constraint(s), e.g.

$$\mathcal{F}_3 = \begin{cases} \{f \ : \ f \text{ is a monotone decreasing pdf on } (0,\infty)\} \\ \\ \{f \ : \ f \text{ is a convex decreasing pdf on } (0,\infty)\}, \end{cases}$$

it can be shown that $\widehat{f}_{n,3}$ must be piecewise linear with the number of knots being at most n. These properties can be used to construct a penalty term and to consider estimation under qualitative assumptions as a penalized estimation problem where the class $\mathcal{F}_2(R,C)$ is generalized to $\mathcal{F}_2(R)$, defined as

$$\mathcal{F}_2(R) \ := \ \bigcup_{C>0} \mathcal{F}_2(R,C).$$

To summarize, both methods, roughness penalization and shape constraints, impose some sort of regularization on the maximum log-likelihood functional in order to get a meaningful estimator.
Nonparametric maximum likelihood estimation of density functions restricted by qualitative assumptions has received much attention in the last decades and in the following sections we briefly summarize these developments.

1.6 Monotone density estimation

For applications of monotone density estimation consult e.g. Barlow et al. (1972) or Robertson, Wright, and Dykstra (1988).
Maximum likelihood estimation of a monotone density was first studied by Grenander (1956), who found that a function $\widehat{f}_G$ is the nonparametric maximum likelihood estimator (NPMLE) if and only if it is the left derivative of the concave majorant

of the empirical cumulative distribution function. Grenander's was continued by Prakasa Rao (1969) who established asymptotic distribution theory for $\widehat{f}_n - f$ at a fixed point $x_o > 0$:

$$n^{1/3}\Big(\widehat{f}_n(x_o) - f(x_o)\Big) \to_{\mathcal{D}} 16\Big|f(x_o)f'(x_o)\Big|^{1/3} Z,$$

where Z is distributed as the location of maxima of the process $(\mathbb{W}(t) - t^2)_{t\in(0,\infty)}$ with $\mathbb{W}$ being Brownian Motion starting at 0. Groeneboom (1985) resumed the asymptotic distribution theory and examined the limiting distribution in great detail (Groeneboom, 1988) whereas Groeneboom, Hooghiemstra and Lopuhaä (1999) and Kulikov and Lopuhaä (2005a) concentrated on limit theory in the space $\mathrm{L}_1(\mathbb{R})$. The pointwise rate of convergence, $O_{\mathrm{p}}(n^{-1/3})$, is slow compared e.g. to that of a regular parametric problem where one obtains $O_{\mathrm{p}}(n^{-1/2})$. The rate of convergence with respect to uniform norm is further decelerated by a factor $\log(n)$. This result is not directly proven but a special case of a theorem derived by Jonker and van der Vaart (2001). They assumed that f possesses a derivative that is bounded, strictly negative and bounded away from zero. The supremum distance between the empirical distribution function $\mathbb{F}_n$ and its concave majorant $\widehat{F}_G$ was investigated by Kiefer and Wolfowitz (1976) who proved that this difference disappears (in probability) at a rate $o_{\mathrm{p}}((\log n)^{5/6}n^{-2/3})$. This result has recently been extended by Kulikov and Lopuhäa (2005b) in the sense that they investigated the whole process

$$n^{2/3}\Big(\widehat{F}_n(t) - \mathbb{F}_n(t)\Big)_{t\in[0,1]}.$$

1.7 Unimodal density estimation

Remember that a density f on the real line is unimodal if there exists a number $M = M(f)$ such that f is non-decreasing on $(-\infty, M]$ and non-increasing on $[M, \infty)$. In case the true mode is known a priori, unimodal density estimation boils down to monotone estimation, by estimating the true underlying distribution function F by the distribution function $\widehat{F}_n$ that is the least concave majorant of $\mathbb{F}_n$ on the interval $[M, \infty)$ and the greatest convex minorant on $(-\infty, M]$. The density f is then estimated by the left derivative $\widehat{f}_n$ of $\widehat{F}_n$. In case none of the observations equals M, this estimator maximizes the likelihood (but must not be continuous at M).

The situation is completely different if M is not known. In that case, the likelihood can be maximized to ∞ by placing an arbitrary large mode at some fixed observation, meaning that consistent estimation of f at the mode is not possible. This phenomena is called "spiking". Several methods were proposed to remedy this problem. Wegman (1970) introduced a modal interval of fixed length ε on which the density is assumed to be flat (this estimator is inconsistent except the true density f also has a modal interval of at least length ε), ensuring that the density can not exceed $1/\varepsilon$. Woodroofe and Sun (1993) penalized the ordinary maximum likelihood estimator (MLE), resulting in a consistent density estimator. Bickel and Fan (1996) showed that estimating the mode first and then plug it into their smooth maximum likelihood procedure does not change the asymptotic behavior of this estimator. The meaning of "smooth" here is that they optimize the maximum likelihood functional (given the true or estimated mode) not over the class of all unimodal densities, but over the class of all continuous piecewise linear densities with mode at one of the X_i to get a linear spline MLE. To circumvent the spiking problem, they further propose to group the data before computing their MLE. As for the spiking problem, Meyer and Woodroofe (2004) generalize Wegman's idea by introducing an estimator that is concave over an interval containing the mode. This interval may be chosen a priori or through an algorithm.
The combination of shape constraints and smoothing was continued by Eggermont and LaRiccia (2000). In order to improve the slow rate of convergence of $n^{-1/3}$ in the space $\mathrm{L}_1(\mathbb{R})$ for arbitrary unimodal densities, they derived a Grenander type estimator by taking the derivative of the least concave majorant of the distribution function corresponding to a kernel estimator rather than the empirical distribution function, yielding a rate of convergence of $O_{\mathrm{p}}(n^{-2/5})$. They introduced log-concavity in density estimation (see below), but instead of a shape constraint for the density as a property of the kernel A_h (h is the bandwidth), exploiting a key property of log-concave density functions ($\mathrm{d}F_o$ is the true density):

> *The log-concavity is sensible since then the convolution $A_h * \mathrm{d}F_o$ is unimodal whenever f_o is unimodal, by the celebrated result of Ibragimov (1956).*

Additionally, $A_h * \mathrm{d}\mathbb{F}_n$ is then continuous. Examples for log-concave kernels are Epanechnikov, Gaussian or two-sided Exponential. In their book of 2001, Eggermont and LaRiccia treated a similar case, replacing unimodality by log-concavity (of the

density f) and they presumed, whether smoothing with the log-concave kernel A_h is really necessary to get a "good" rate of convergence in the space $\mathrm{L}_1(\mathbb{R})$ and how to actually compute a log-concave density estimator. The second of these questions is answered in Chapter 4 of this thesis.
Renouncing on a continuity assumption on f, Van der Vaart and Van der Laan (2003) complemented the work by investigating the interplay of isotonization and kernel estimation, showing that the limit distribution at a fixed point is more concentrated for the isotonized kernel than using either isotonization or smoothing exclusively (but the rate of convergence is not improved).
For a discussion of other approaches than maximum likelihood consult e.g. Hall and Huang (2002) and the references therein.

1.8 Convex density estimation

Convex density estimation was pioneered by Anevski (1994) (later published as Anevski, 2003). The problem arose in a study of migrating birds discussed by Hampel (1987). Jongbloed (1995) established lower bounds for minimax rates of convergence and rates of convergence for a "sieved MLE". Groeneboom, Jongbloed, and Wellner (2001b) almost completely cleaned up the situation providing a characterization of the estimator as well as consistency and limiting behavior at a fixed point of positive curvature of the function to be estimated. They do this not only for maximum likelihood but also for least squares density estimation and the corresponding regression problems as well. They found that in all cases the estimators have to be piecewise linear with knots between the observation points. They show for the (rescaled) distance between the maximum likelihood estimator $\widehat{f}_n$ and the true density at a fixed point $x_o > 0$ that

$$n^{2/5}\Big(\widehat{f}_n(x_o) - f(x_o)\Big) \to_{\mathcal{D}} (1/24)^5\Big(f^2(x_o)f''(x_o)\Big)^{1/5}\mathcal{H}''(0)$$

where $\mathcal{H}$ is a stochastic process connected to Brownian Motion and further detailed in Groeneboom, Jongbloed, and Wellner (2001a). Apparently, they assumed existence and positivity of the true density's second derivative f'', what together with the convexity assumption enables one to estimate f at a fairly better rate of $O_{\mathrm{p}}(n^{-2/5})$ than that in the non-smoothed monotone and unimodal case. Precisely, they assumed that the true density f is twice continuously differentiable, convex,

and decreasing on $[0, \infty)$. Note that here again the estimator is inconsistent at 0 (which corresponds to the mode in the given situation).
It would be of great surprise if the rate of convergence with respect to uniform norm was not $(\log(n)/n)^{2/5}$, but to our knowledge no proof for this result has ever been published.
Balabdaoui and Wellner (2004a-d) treated a unifying and extending approach. Let k be a non-negative integer and G be a distribution function on $(0, \infty)$. Then

$$f(x) \quad = \quad \int_0^\infty \frac{k}{y^k}(y-x)_+^{k-1}\,\mathrm{d}G(y), \quad x \geq 0$$

is monotone (decreasing) if $k = 1$ and convex and decreasing if $k = 2$. They figured out the details for all finite k, with the final aim to solve the case $k = \infty$ (completely monotone densities).
Although a characterization of $\widehat{f}_n$ in the convex case exists (but is not as simple as the least concave majorant in the monotone case), actual calculation of $\widehat{f}_n$ is not straight-forward and has to be done numerically. Several attacks to the problem were made. Jongbloed (1998) proposed an algorithm to minimize a smooth convex (likelihood-) function over a convex cone in $\mathbb{R}^n$, well applicable to convex density estimation. Another successful approach was chosen by Terlaky and Vial (1998), using interior point methods. Dümbgen, Freitag, and Jongbloed (2006) presented a new method specially tailored to find piecewise linear functions with only a few knot points. They examined unimodal distribution function estimation with censored data, but the methods should be applicable in the convex density case as well.

1.9 Log-concavity

In this thesis we will impose a quite natural shape constraint on the density f to be estimated: log-concavity, meaning that the density f to be estimated can be represented as

$$f(x) = \exp\varphi(x), \quad x \in \mathbb{R}$$

for some concave function $\varphi : \mathbb{R} \to [-\infty, \infty)$. This class is rather flexible in the sense that it generalizes many densities of common parametric distributions, such as Normal, Uniform, Logistic, χ^2 or Laplace. Many other distributions have log-concave densities for broad ranges of the parameter values: Gamma, Beta, Weibull

or the Generalized Pareto distribution. Tables detailing these issues can be found in Section 2.4 and in Bagnoli and Bergstrom (1989, later published as Bagnoli and Bergstrom, 2005). The latter paper also offers a concise summary of the main properties of log-concave density functions, their corresponding distribution functions, and their applications in reliability and many fields of economic theory. Further applications of log-concavity in reliability can be found in the standard book by Barlow and Proschan (1975). The book by Devroye (1986) offers a whole chapter about random number generation for random variables having a log-concave density. Voting theory and the theory of imperfect competition is the field of application in a pair of papers by Caplin and Nalebuff (1991a, 1991b). A nice discussion of (multivariate) log-concavity, log-convexity and the differences between both is provided by An (1995, 1998). He further details the connection between log-concavity/-convexity and the properties inherited by functions derived from such densities under more general assumptions than Bagnoli and Bergstrom (1989, 2005). We will exploit the connection between a log-concave density and the corresponding hazard function λ in Section 3.6 to derive a new consistent estimator of λ.

In his first paper, An also describes an indirect goodness-of-fit test for log-concavity, based on the hazard rate.

A key reference in connection with log-concavity of functions is the book by Karlin (1968) about total positivity, a concept generalizing log-concavity (log-concave functions correspond to totally positive functions of order 2).

Note that every log-concave density is automatically unimodal. Although certainly the class of log-concave densities is much smaller than that of unimodal, if ever one can estimate a log-concave density one gets a method to circumvent the problems described in Section 1.7 of either trying out many modes or spiking at a known mode.

Although being very flexible and an apparent generalization of several parametric models, not much on log-concave density estimation has been published. So far only Walther (2000) attacked the problem and used the iterative convex minorant algorithm (as introduced by Jongbloed, 1998 for the estimation of a convex decreasing density on $(0, \infty)$) for estimation of a logarithmically concave density.

Walther further conjectures:

> *The theoretical properties of a log-concave MLE are similar to those of the MLE of a concave density, and the arguments in Groeneboom, Jongbloed, and Wellner (2001b) suggest that the uniform rate of convergence is* $O_{\mathrm{p}}((\log(n)/n)^{2/5}$.

One of our results is indeed the verification of this conjectured rate of convergence, see Section 3.3. Walther describes the MLE $\widehat{f_c}$ under the assumption that the true density is of the form

$$f_c(x) \;=\; \exp\Big(\phi(x) + c|x|^2\Big), \quad x \in [0,1]$$

for some concave function ϕ and $c \geq 0$. He suggests a bootstrap test to assess log-concavity based on $(\widehat{f_c})_{c\in\mathcal{C}}$, where $\mathcal{C}$ is some finite set of nonnegative numbers. Absence of log-concavity indicated by the test is interpreted as a mixture of several log-concave distributions. In Walther (2001), testing for log-concavity is transformed in testing for monotonicity, enabling the application of the monotone estimation device described in Section 1.6. The price to pay for this indirect procedure is that deviations of log-concavity can only hardly be localized and visualized. In Part 2 of this thesis we present a new method to make possible this visualization.
As pointed out by Bagnoli and Bergstrom (1989, 2005), a distribution function received from a log-concave density function is again log-concave, the converse being not true. Sengupta and Paul (2004) considered testing for log-concavity of a distribution function versus the alternative that it is not, where they need to restrict their attention to such distribution functions having a point mass at 0. According to the above mentioned authors, direct maximum likelihood estimation of a log-concave distribution function is not possible without further restrictions, most likely because this class is simply too big.
Note that by imposing log-concavity on the density, two of the major problems arising in monotone and convex density estimation, namely spiking (both) resulting in non-consistency points and discontinuity of the estimator (only monotone), do not come up. Together with the fact that many parametric models are automatically log-concave, an in-depth analysis of log-concave density estimation is overdue and one step in this direction is the aim of this thesis.

1.10 Bump hunting

The second part of this thesis leaves the field of density estimation and is concerned with what has been named "bump hunting".

In the analysis of univariate data, researchers often want to infer qualitative characteristics of the density function of their data. Examples for such characteristics are local extrema, inflection points or regions where the density function is monotone (mode hunting) or convex (bump hunting). Kernel density estimates, pioneered by Silverman (1981), prevail in problems of this type. Silverman's method is constructed such that the number of modes of the underlying density f is a decreasing function of the bandwidth of a normal kernel (the only admissible in this specific case). Critical values to test the null hypothesis whether f has, say, k modes versus the alternative of having more than k modes are then found through a simple bootstrap procedure. This principle can be generalized in various ways, one of them being SiZer (Chaudhuri and Marron, 1999; 2000). This method goes further in the sense that it combines kernels using a broad range of bandwidths. However, in this approach it is not clear how to combine conclusions from kernel estimates at different scales. Furthermore, the correction term for small scales derived by Dümbgen and Spokoiny (2001) is not applied, meaning that the global view is possibly dominated by the tests stemming from short intervals. Instead, Chaudhuri and Marron restrict their attention to kernel bandwidths h such that $h \geq \varepsilon > 0$ for a fixed positive ε.

Other approaches are excess masses, see e.g. Cheng and Hall (1998) and the references therein, maximum likelihood as in Walther (2001) or the "dip test", proposed by Hartigan and Hartigan (1985).

For mode hunting, Dümbgen and Walther (2006) proposed a procedure that simultaneously provides confidence statements with guaranteed significance level for arbitrary sample size (i.e. also for finite n, not only asymptotically). They applied a multiscale approach in the spirit of Dümbgen and Spokoiny (2001) and Dümbgen (2002) by introducing a test statistic derived from a simple parametric model. This statistic is evaluated on local spacings (i.e. on every interval spanned by two observations) and all these test statistics are then combined to get a multiscale test. To reach significance, even for finite n, Dümbgen and Walther (2006) provided a quite remarkable deterministic inequality (Proposition 1 in their paper). They also derived the limiting distribution for their global test statistic as the sample size increases, by extending results from Dümbgen and Spokoiny (2001) to a more general

class of stochastic processes. However, critical values are generated via Monte Carlo simulations.

In Part 2 we propose a bump hunting method in the same spirit. We equally introduce a relatively simple local parametric model and combine all test statistics calculated on local spacings to get a global multiscale test. Commonly, to "hunt bumps" means to identify intervals where the density f is either convex or concave, at best with a certain confidence. However, our focus here is on log-concavity and log-convexity. Beneath better mathematical tractability observe that by taking the logarithm non-concave densities with only one bump, e.g. the gaussian density, become purely concave, meaning that the region of the sole bump could possibly be detected easier because it is not "contaminated" by non-concave regions. To the best of our knowledge, no one has up to now chosen such an approach to the problem.

However, compared to the mode hunting case, at least one major difference has to be ascertained. Dümbgen and Walther (2006) received their local test statistics using the general parametric model

$$f_\lambda(x) \quad = \quad 1 + \lambda(x - 1/2), \quad x \in [0, 1],$$

for $\lambda \in \mathbb{R}$. Their test statistic is then the Neyman-Pearson locally most powerful test in this model for the null hypothesis $\lambda \leq 0$ versus the alternative $\lambda > 0$. Evidence for a non-decrease, say, is then simply received from testing this null hypothesis $\lambda \leq 0$. To detect log-concavity we propose the following parametric model:

$$f_{\theta,\eta}(x) := C(\theta, \eta) \exp\Big(\theta x + \eta x^2/2\Big), \quad x \in [0, 1] \tag{1.1}$$

for $\theta \in \mathbb{R}, \eta \in \mathbb{R}$, where $C(\theta, \eta)$ is a normalizing constant. Log-concavity is then postulated if a statistical test decides on $\eta < 0$. Unfortunately, in this model one has somehow to deal with the nuisance parameter θ: Either by considering a test statistic using "the worst" of all possible $\theta \in \mathbb{R}$, resulting probably in a considerable loss of power, or to estimate θ. This approach presumably yields more power, however only with the major drawback that all results are only asymptotically valid.

We motivate a test statistic to perform a test for η in (1.1) and give some further consistency justifications for the specific test statistic. That the method works is illustrated with some examples.

Part I

Log-concave density estimation

Chapter 2

Log-concave densities

In this short chapter, we introduce some fundamental properties of log-concave densities. Parametric examples for log-concave densities are given.

2.1 Log-concave densities and unimodality

Throughout the first part of this thesis X will denote a random variable having distribution function F. If we talk about densities they are always meant to be defined with respect to Lebesgue measure. We assume that F possesses a density f such that

$$f(x) = \exp \varphi(x)$$

for some concave function $\varphi : \mathbb{R} \to [-\infty, \infty)$. Such densities f are given the name log-concave and we will use this term also for the random variable X itself. The following lemmas summarize three key properties of log-concave densities.

Lemma 2.1.1. *Suppose the random variable X has a log-concave density function f on $\mathbb{R}$. Then f is also unimodal, i.e. there exists a number $m \in \mathbb{R}$ such that f is non-decreasing on $(-\infty, m]$ and non-increasing on $[m, \infty)$.*

To be able to state the following results properly, define the convolution $a * b$ of two density functions $a, b \in \mathrm{L}_1(\mathbb{R})$ at $x \in \mathbb{R}$ as

$$(a * b)(x) := \int_{\mathbb{R}} a(t)b(x - t)\,\mathrm{d}t.$$

Lemma 2.1.2. *The convolution $l_1 * l_2$ of two log-concave densities l_1 and l_2 is again log-concave.*

Even more surprising is the fact that convolutions of unimodal and log-concave densities remain unimodal and that this property can even be used to characterize log-concavity.

Theorem 2.1.3. *A density function l is log-concave if and only if its convolution $l * u$ with any unimodal density function u is again unimodal.*

The latter results are both due to Ibragimov (1956), where Theorem 2.1.3 is generally referred to as "Ibragimov's Theorem". Historically, Ibragimov introduced the term "strongly unimodal" for densities exhibiting the property stated in the theorem and showed that the class of strongly unimodal and log-concave densities coincide.
A survey of the connections between log-concavity and unimodality can be found in the book by Barndorff-Nielsen (1978).

2.2 Tail behavior

One of the key properties of a log-concave random variable X is the existence of all of its moments. The precise, and even stronger, statement is detailed in the next lemma.

Lemma 2.2.1. *There exist constants $a_o \in \mathbb{R}$ and $b_o > 0$ such that for all $x \in \mathbb{R}$ one has:*

$$\varphi(x) \;\leq\; a_o - b_o|x|.$$

In particular,

$$\int \exp(t_o|\varphi|)\,\mathrm{d}F \;<\; \infty \quad \text{whenever } t_o < 1.$$

Moreover, for any polynomial p and any number $t_o \in (0,1)$ there exists a constant $c_o > 0$ such that

$$\begin{aligned}
\int_r^\infty p(|\varphi|)\,\mathrm{d}F &\leq c_o \exp\Big(t_o\varphi(r)\Big) \quad \text{and} \\
\int_{-\infty}^{-r} p(|\varphi|)\,\mathrm{d}F &\leq c_o \exp\Big(t_o\varphi(-r)\Big) \quad \text{for all } r \geq 0.
\end{aligned}$$

2.3 Derived functions

Log-concavity of the density function f immediately implies the same or similar properties for functions derived from f such as the distribution function F, tail function $1 - F$ or hazard function λ. Such connections under somewhat restrictive smoothness conditions on the density were e.g. elaborated in Bagnoli and Bergstrom (1989, 2005). An (1995) expanded their work to densities that need not necessarily be differentiable. For illustrative purposes, we will pick one of these functions derived from the density, namely the hazard function λ.

Lemma 2.3.1. *Define the hazard rate function λ as*

$$\lambda(x) \quad := \quad \frac{f(x)}{1 - F(x)}$$

for x in the interval $I := \{y \,:\, F(y) < 1\}$. If f is log-concave, then λ is monotone non-decreasing on I.

The proof of this lemma can be found in Bagnoli and Bergstrom (1989, 2005, Proposition 1) for smooth densities and in the more general form stated in the lemma the proof was given by An (1995, Corollary 2).

2.4 Examples of parametric log-concave densities

The class of log-concave densities comprises many well-known parametric densities, see Table 2.1. In Bagnoli and Bergstrom (1989, 2005) calculations necessary to verify log-concavity of a specific density function, eventually only for certain parameter values, are carried out, i.e. they check for many smooth enough parametric densities that $(\log f)'' \leq 0$.
The Generalized Pareto distribution (GPD) appears in extreme value theory as an adequate parametric model for exceedances, see e.g. Reiss and Thomas (2001).

Table 2.1: Some parametric log-concave densities

Type	Density function $f(x)$	Support	Parameters[a]
Uniform	$(b-a)^{-1}$	$[a,b]$	$a,b \in \mathbb{R}; a<b$
Normal	$(\sqrt{2\pi}\sigma)^{-1}\exp\left(-(x-\mu)^2/(2\sigma)\right)$	$(-\infty,\infty)$	$\mu \in \mathbb{R}, \sigma > 0$
Gamma	$b^a\Gamma(a)^{-1}x^{a-1}\exp\left(-bx\right)$	$[0,\infty)$	$a \geq 1, b > 0$
Beta	$\Gamma(a+b)(\Gamma(a)\Gamma(b))^{-1}x^{a-1}(1-x)^{b-1}$	$[0,1]$	$a \geq 1, b \geq 1$
Fréchet	$ax^{-(1+a)}\exp\left(-x^{-a}\right)$	$[0,\infty)$	$a \geq 0$
Gumbel	$\exp(-x)\exp\left(-e^{-x}\right)$	$(-\infty,\infty)$	
GPD	$(1+\gamma x)^{-(1+1/\gamma)}$	$[0,1/\lvert\gamma\rvert)$	$-1 \leq \gamma < 0$
Logistic	$\exp(-x)(1+\exp(-x))^{-2}$	$(-\infty,\infty)$	
Laplace	$(1/2)\exp\left(-\lvert x\rvert\right)$	$(-\infty,\infty)$	

[a] Parameter values such that f is log-concave

2.5 Proofs

Proof of Lemma 2.1.1: The function φ is concave. Together with the fact that f is a probability density, i.e. $\int_{\mathbb{R}} \exp\varphi = 1$, it can not happen that $\varphi(x) \not\to -\infty$ for $|x| \to \infty$, implying unimodality of φ, i.e. there exists a $j \in \mathbb{R}$ such that $\varphi(x)$ is non-decreasing in $x \leq j$ and non-increasing in $x \geq j$. The result follows via monotonicity of the exponential function. □

Proof of Lemma 2.2.1. The crucial point here is that φ can be bounded from above by a piecewise linear function with one knot. Without loss of generality let φ be upper semi-continuous. After an affine transformation, if necessary, we assume w.l.o.g. (see Section 3.2) that

$$\max_{t\in\mathbb{R}} \varphi(t) = \varphi(0) \leq 0.$$

Then by Lemma 2.1.1 there exists a number $r_o > 0$ such that $\varphi(\pm r_o) \leq \varphi(0) - 1$. By concavity of φ, for any $x \geq r_o$,

$$\varphi(x) \;\leq\; \varphi(r_o) + \frac{\varphi(r_o)-\varphi(0)}{r_o - 0}\,(x - r_o) \;\leq\; \varphi(0) - 1 - \frac{(x-r_o)}{r_o} \;< -|x|/r_o + \varphi(0).$$

Analogously, $\varphi(x) < -|x|/r_o + \varphi(0)$ for $x \leq -r_o$. Since $\varphi(x) \leq \varphi(0) \leq -|x|/r_o + \varphi(0) + 1$ whenever $|x| \leq r_o$, the first assertion is true with $a_o = \varphi(0) + 1$ and

$b_o = 1/r_o$. Then the second assertion follows from

$$\begin{aligned}\int \exp(t_o|\varphi|)\,\mathrm{d}F &= \int \exp\Big((1-t_o)\varphi(x)\Big)\,\mathrm{d}x \\ &\le \int \exp\Big(a_o(1-t_o) - b_o(1-t_o)|x|\Big)\,\mathrm{d}x \;<\; \infty.\end{aligned}$$

As for the last part, note first that

$$p(|\varphi|)f \;=\; p(|\varphi|)\exp(-|\varphi|) \;\le\; \exp(c_o - t_o|\varphi|) \;=\; \exp(c_o)\exp(t_o\varphi)$$

for a suitable constant c_o. Since $\int p(|\varphi|)\,\mathrm{d}F$ is finite, it suffices to consider numbers r that are greater than or equal to, say, r_o above. Since the slope of φ is not larger than $-1/r_o$ on $[r_o, \infty)$,

$$\begin{aligned}\int_r^\infty p(|\varphi|)\,\mathrm{d}F &\le \exp(c_o)\int_0^\infty \exp\Big(t_o\varphi(r+z)\Big)\,\mathrm{d}z \\ &\le \exp(c_o)\int_0^\infty \exp\Big[t_o\Big(\varphi(r) - z/r_o\Big)\Big]\,\mathrm{d}z \\ &= \exp(c_o)\int_0^\infty \exp\Big(-(t_o/r_o)z\Big)\,\mathrm{d}z\,\exp\Big(t_o\varphi(r)\Big) \\ &= \exp(c_o)(r_o/t_o)\exp\Big(t_o\varphi(r)\Big).\end{aligned}$$

Analogously one can show that $\int_{-\infty}^{-r} p(|\varphi|)\,\mathrm{d}F \le \exp(c_o)(r_o/t_o)\exp(t_o\varphi(-r))$. □

Chapter 3

Maximum likelihood estimation

In this chapter we introduce the maximum likelihood estimator of a log-concave density. At first we prove its existence and uniqueness. Then we provide two characterizations for this estimator and give some results about uniform rate of convergence. These asymptotic results are then extended to functions derived from the density estimator, namely the distribution and hazard function.

3.1 General framework

Our goal is to estimate a univariate log-concave density function f based on a random sample of size $n > 1$. Let $X_1 < \ldots < X_n$ be the corresponding order statistics. For any such density f on $\mathbb{R}$, the negative log-likelihood functional at f, our parameter of interest, is defined as

$$\begin{aligned} L_n(f) &:= -n \int \log f(x) \, \mathrm{d}\mathbb{F}_n(x) \\ &= -\sum_{i=1}^{n} \log f(X_i) \end{aligned} \tag{3.1}$$

where $\mathbb{F}_n$ stands for the empirical distribution function:

$$\mathbb{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}, \quad x \in \mathbb{R}.$$

The indicator function 1_A for a condition A is defined as

$$1_A = \begin{cases} 1 & \text{if } A \text{ holds,} \\ 0 & \text{else.} \end{cases}$$

The NPMLE is then defined as the minimizer of the functional in (3.1) over all log-concave probability densities. In order to relax the constraint of f being a probability density and to get a criterion function to minimize over all concave functions in general, we focus on $\varphi = \log f$ and employ the standard trick of adding a Lagrange-term to the log-likelihood functional defined in (3.1). This leads to

$$\Psi_n(\varphi) \quad = \quad -n \int \varphi(x)\, \mathrm{d}\mathbb{F}_n(x) + n \int \exp \varphi(x)\, \mathrm{d}x. \tag{3.2}$$

Define $\widehat{\varphi}_n$ as the minimizer of this functional over the set of all concave functions:

$$\widehat{\varphi}_n \quad := \quad \underset{\varphi \text{ concave}}{\arg\min}\, \Psi_n(\varphi)$$

and let

$$\widehat{f}_n = \exp(\widehat{\varphi}_n)$$

be the corresponding maximum likelihood estimator of f. The distribution function $\widehat{F}_n$ of $\widehat{f}_n$ is given by

$$\widehat{F}_n(x) := \int_{-\infty}^{x} \widehat{f}_n(u)\, \mathrm{d}u.$$

Since

$$\begin{aligned} 0 \quad &= \quad \frac{d}{dt}\bigg|_{t=0} \Psi_n(\widehat{\varphi}_n + t) \\ &= \quad -n + n \int \widehat{f}_n(x)\, \mathrm{d}x, \end{aligned}$$

the Lagrange term guarantees in fact a probability density.

3.2 Basic Properties of $\widehat{\varphi}_n$ and $\widehat{f}_n$

Existence and uniqueness

First of all we need to show that $\widehat{\varphi}_n$ is a meaningful estimator: Theorem 3.2.1 guarantees existence and uniqueness of $\widehat{\varphi}_n$ and states an interesting key property of it.

Theorem 3.2.1. *The NPMLE $\widehat{\varphi}_n$ exists and is unique. It is piecewise linear and continuous on $[X_1, X_n]$ with changes of slope only at observation points. Moreover, $\widehat{\varphi}_n = -\infty$ for $x \notin [X_1, X_n]$.*

The piecewise linearity of $\widehat{\varphi}_n$ is analogous to the case of estimating a convex decreasing density, treated extensively by Groeneboom, Jongbloed, and Wellner (2001b). But in the latter case the knots of the estimated density are situated strictly between the observations. Theorem 3.2.1 further entails that $\widehat{f}_n$ is completely determined by the vector

$$\widehat{\boldsymbol{\varphi}} = \Big(\widehat{\varphi}_n(X_i)\Big)_{i=1}^{n}.$$

Hence, the infinite-dimensional problem of finding the minimizer of Ψ_n over all concave functions boils down to a finite (but high) dimensional task which is elaborated in Chapter 4.

Characterizations

We give two characterizations of the estimator $\widehat{f}_n$. The first via special perturbation functions and the second by connecting the empirical distribution function of the sample with the distribution function derived from the estimator.

Theorem 3.2.2. *Let $\tilde{\varphi}_n$ be a concave piecewise linear function on $[X_1, X_n]$ with knots only at $\{X_1, \ldots, X_n\}$. Moreover, let $\tilde{\varphi}_n = -\infty$ on $\mathbb{R}\backslash[X_1, X_n]$. Then $\tilde{\varphi}_n = \widehat{\varphi}_n$ if, and only if,*

$$\int \Delta(x)\, \mathrm{d}\mathbb{F}_n(x) \;\leq\; \int \Delta(x) \exp \tilde{\varphi}_n(x)\, \mathrm{d}x. \tag{3.3}$$

for any $\Delta : \mathbb{R} \to \mathbb{R}$ such that $\tilde{\varphi}_n + t\Delta$ is concave for some $t > 0$.

For functions Δ that are continuous, piecewise linear and have the same knots as $\tilde{\varphi}_n$, one gets even equality in (3.3).

The characterization in terms of distribution functions is given in the following theorem. Let $h_n : [X_1, X_n] \to \mathbb{R}$ be a piecewise linear continuous function, such that the knots coincide with some of the observation points $X_1 < \ldots < X_n$. The set of knots $\mathcal{S}(h_n)$ of h_n is then defined as follows:

$$\mathcal{S}(h_n) \;:=\; \{t \in (X_1, X_n) : h_n'(t-) > h_n'(t+)\} \cup \{X_1, X_n\}.$$

Recall that $\widehat{\varphi}_n$ is an example for such a function h_n.

Theorem 3.2.3. *Let $\tilde{\varphi}_n$ be as in Theorem 3.2.2 and define*

$$\tilde{F}_n(x) \;:=\; \int_{-\infty}^{x} \exp \tilde{\varphi}_n(t)\, \mathrm{d}t.$$

In addition, it is assumed that $\tilde{F}_n(X_n) = 1$. Then, $\tilde{\varphi}_n = \widehat{\varphi}_n$ and thus $\tilde{F}_n = \widehat{F}_n$, if and only if for arbitrary $a < t < b$ with $a, b \in \mathcal{S}(\tilde{\varphi}_n)$,

$$\int_a^t \tilde{F}_n(r)\, dr \;\leq\; \int_a^t \mathbb{F}_n(r)\; \mathrm{d}r\,, \tag{3.4}$$

$$\int_t^b \tilde{F}_n(r)\, dr \;\geq\; \int_t^b \mathbb{F}_n(r)\; \mathrm{d}r\,, \tag{3.5}$$

$$\int_a^b \tilde{F}_n(r)\, dr \;=\; \int_a^b \mathbb{F}_n(r)\; \mathrm{d}r\,. \tag{3.6}$$

Note that (3.4) follows directly from (3.5) and (3.6). In Figure 3.1 we illustrate the behavior of the process

$$D(t) \;:=\; \int_{X_1}^{t} (\widehat{F}_n - \mathbb{F}_n)(r)\; \mathrm{d}r, \quad t \in [X_1, X_n].$$

The characterization of $\widehat{f}_n$ in Theorem 3.2.3 as the second derivative of the integral of the empirical distribution function coincides with that of the least squares estimator of a convex decreasing density, specified in Lemma 2.2 of Groeneboom, Jongbloed, and Wellner (2001b). The convex case analogue of (3.3) can be found in the cited paper, Lemma 2.4.

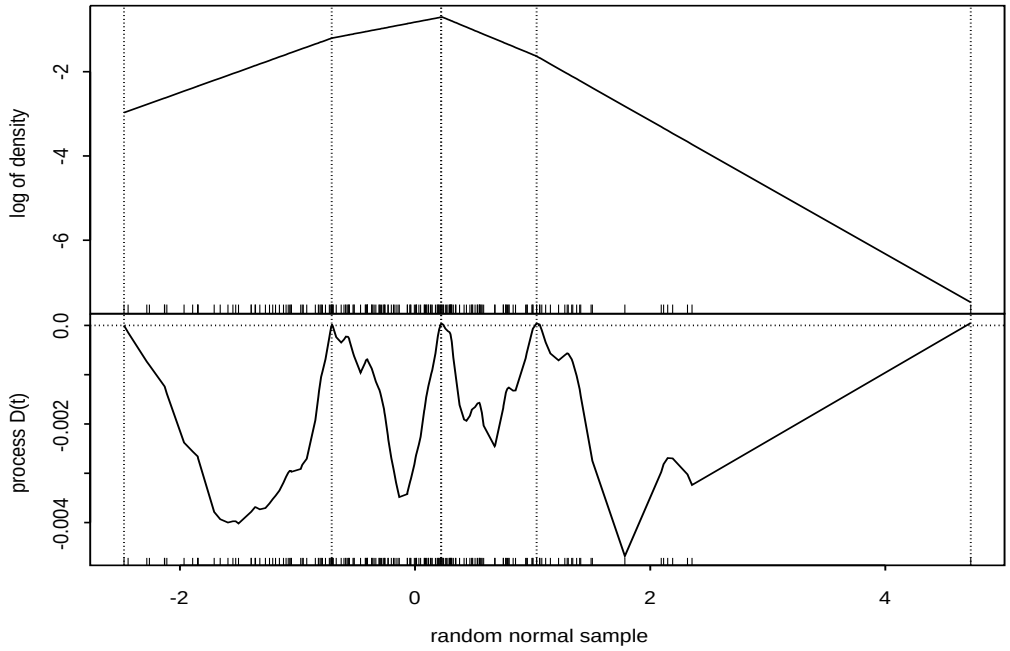

Figure 3.1: The process $D(t)$ for a normal random sample of size 200.

Further properties of $\widehat{f}_n$

For an arbitrary distribution function G on the real line let

$$\begin{aligned} \mu(G) &:= \int u \, \mathrm{d}G(u) \\ \mathrm{Var}(G) &:= \int \Big(u - \mu(G)\Big)^2 \, \mathrm{d}G(u) \end{aligned}$$

denote the mean and the variance, provided that $\int |u| \, \mathrm{d}G(u) < \infty$. Then the following corollary can be derived from Theorem 3.2.2.

Corollary 3.2.4. *Setting $\Delta(x) = x$ and $\Delta(x) = -x^2$ in (3.3) one obtains:*

$$\mu(\widehat{F}_n) \;=\; \mu(\mathbb{F}_n) \quad \textit{and} \quad \mathrm{Var}(\widehat{F}_n) \;\leq\; \mathrm{Var}(\mathbb{F}_n).$$

The distribution function estimator $\widehat{F}_n$ has the highly appealing feature of being very close the the empirical distribution function $\mathbb{F}_n$ at all knot points of $\widehat{\varphi}_n$.

Corollary 3.2.5. *Choosing* $\Delta(x) := 1_{\{x<q\}}$ *or* $\Delta(x) := -1_{\{x\leq q\}}$ *for* $q \in \mathcal{S}(\widehat{\varphi}_n)$ *yields:*

$$\widehat{F}_n \;\in\; \big[\mathbb{F}_n - n^{-1}, \mathbb{F}_n\big] \quad \textit{on} \quad \mathcal{S}(\widehat{\varphi}_n).$$

This fact, together with Characterization 2 in Theorem 3.2.3 finally entails:

$$\widehat{F}_n(X_1) \;=\; 0 \quad \textit{and} \quad \widehat{F}_n(X_n) \;=\; 1.$$

Equivariance

Finally, let us mention that our estimators are affine equivariant in the following sense. To explicitly express the dependence of the log-likelihood function on $X_1, \ldots, X_n$ write

$$\Psi_n(\varphi) \;=\; \Psi_n(\varphi; X_1, \ldots, X_n).$$

Replacing the observations $X_1, \ldots, X_n$ by $\ddot{X}_i := a + bX_i$ for all $i = 1, \ldots .n$ and $a \in \mathbb{R}$ and $b > 0$ and defining

$$\ddot{\varphi}(x) = \varphi\Big(\frac{x-a}{b}\Big) - \log b \qquad \ddot{\mathbb{F}}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{\ddot{X}_i \leq x\}}, \quad x \in \mathbb{R}$$

we have:

$$\begin{aligned}
\Psi_n(\varphi; X_1, \ldots, X_n) &= -n\int \varphi(x)\,\mathrm{d}\mathbb{F}_n(x) + n\int \exp\varphi(x)\,\mathrm{d}x \\
&= -n\int \Big(\ddot{\varphi}(x) + \log b\Big)\,\mathrm{d}\ddot{\mathbb{F}}_n(x) + n\int \exp\Big(\varphi[(y-a)/b]\Big) b^{-1}\,\mathrm{d}y \\
&= \Psi_n(\ddot{\varphi}; \ddot{X}_1, \ldots, \ddot{X}_n) + n\log b. \qquad (3.7)
\end{aligned}$$

Consequently, minimizing the function $\Psi_n(\varphi; X_1, \ldots, X_n)$ over all piecewise linear functions φ with knots at some of the observations yields the same solution as minimizing (3.7) w.r.t to functions $\ddot{\varphi}$ (where this latter functions are also piecewise linear with knots at some of the observation points). Because of this equivariance we may and do assume from now on that

$$\max_{x\in\mathbb{R}} \varphi(x) \;=\; \varphi(0) \;=\; -1. \qquad (3.8)$$

This will be convenient later on when we use $|\varphi| \geq 1$ as a weight function.

3.3 Uniform consistency of $\widehat{f}_n$

Let us introduce some notation. Define

$$\rho_n := (\log n)/n$$

and the uniform norm of a function g on an interval I by

$$\|g\|_\infty^I := \sup_{x \in I} |g(x)|.$$

With $T := [A, B]$ we always denote a fixed compact interval on $\mathbb{R}$, where $A < B$. The set of knots of $\widehat{\varphi}_n$ on an interval $T \subset \mathbb{R}$ is written as:

$$\mathcal{S}(\widehat{\varphi}_n) \cap T =: \{s_1, \ldots, s_{M(n)}\}.$$

A function $g : T \to \mathbb{R}$ belongs to the Hölder smoothness class $\mathcal{H}^{\beta,L}(T)$ with exponent $\beta \in [1, 2]$ and some constant $L > 0$ if for all $x, y \in T$ we have

$$\begin{aligned} |g(x) - g(y)| &\leq L|x - y| \qquad \text{if } \beta = 1, \\ |g'(x) - g'(y)| &\leq L|x - y|^{\beta-1} \quad \text{if } \beta > 1. \end{aligned}$$

Finally, convergence in probability and in law are written as $\to_{\mathrm{p}}$ and $\to_{\mathcal{D}}$ (equality likewise).

Groeneboom, Jongbloed, and Wellner (2001b) proved uniform consistency of the estimator of a convex density on $(0, \infty)$ as well as its rate of convergence of $n^{-2/5}$ at a fixed point $x_o > 0$ under the following smoothness conditions on the true density f: $f'(x_o) < 0$, $f''(x_o) > 0$, and f'' is continuous in a neighborhood of x_o. The key in the proof was the explicit characterization of the estimator $\widehat{f}_n$ and a lemma about pointwise consistency.

On the other hand, under similar assumptions, Dümbgen, Freitag, and Jongbloed (2004) established a rate of uniform convergence of $(\log(n)/n)^{2/5}$ for concave least squares regression using perturbation functions that are piecewise linear and continuous.

What we do here is transforming the latter result to maximum likelihood estimation of log-concave densities under some Hölder smoothness conditions on the true density function f. We give theorems for uniform convergence on a compact interval, for the density estimator $\widehat{f}_n$, the distribution function estimator $\widehat{F}_n$ derived from it (Section 3.5), and the hazard rate estimator $\widehat{\lambda}_n$ (Section 3.6).

To conclude, we point out the difference to the general approach of van de Geer (2000) to derive consistency and rates. While she uses entropy numbers for the family of all potential density functions we consider a much smaller class of "caricatures" for the difference between estimated and true density. Namely, our caricatures in the proof of Theorem 3.3.1 are piecewise linear functions with at most three knots.

Theorem 3.3.1. *Assume for the log-density $\varphi = \log f$ that $\varphi \in \mathcal{H}^{\beta,L}(T)$ for some exponent $\beta \in [1,2]$ and T a compact subinterval of $\{f > 0\}$. Then,*

$$\begin{aligned} \max_{t \in T} (\widehat{f}_n - f)(t) &= O_p\left(\rho_n^{\beta/(2\beta+1)}\right), \\ \max_{t \in \left[A+\rho_n^{1/(2\beta+1)}, B-\rho_n^{1/(2\beta+1)}\right]} (f - \widehat{f}_n)(t) &= O_p\left(\rho_n^{\beta/(2\beta+1)}\right). \end{aligned} \tag{3.9}$$

Note that a concave function φ is automatically Lipschitz-continuous (i.e. Hölder-continuous with exponent $\beta = 1$) on any interval $T = [A, B]$ with $A > \inf\{\varphi > -\infty\}$ and $B < \sup\{\varphi > -\infty\}$. This entails:

Corollary 3.3.2. *For any continuous log-concave density f,*

$$\|\widehat{f}_n - f\|_\infty^{\mathbb{R}} \to_{\mathrm{p}} 0 \quad \textit{and} \quad \|\widehat{F}_n - F\|_\infty^{\mathbb{R}} \to_{\mathrm{p}} 0.$$

In the convex density case treated by Groeneboom, Jongbloed, and Wellner (2001b), the rate of convergence of $\widehat{f}_n$ to f at a fixed point (under the assumption $\beta = 2$) is $O_{\mathrm{p}}(n^{-2/5})$. It would therefore be no surprise if the uniform rate in that situation would be equal to the log-concave case, as generally the rate of convergence is slowed down by a log-factor when considering uniform instead of pointwise convergence. Furthermore, our proof for a uniform rate of convergence should be adaptable to convex density estimation (where this result is still lacking).

3.4 Distance between consecutive knots of $\widehat{\varphi}_n$: the gap problem

The next lemma about the maximal distance of two consecutive knots of $\widehat{\varphi}_n$ plays a crucial role in the proof of Theorem 3.5.1. However, it also deserves its own merits, as it specifies how fast two consecutive knot points of $\widehat{\varphi}_n$ are approaching each other.

Theorem 3.4.1. *Let $s_{i-1}, s_i \in \mathcal{S}(\widehat{\varphi}_n)$ be two arbitrary consecutive knots of $\widehat{\varphi}_n$ on $T := [A, B]$ where $\varphi \in \mathcal{H}^{\beta,L}(T)$ for some $\beta \in (1, 2]$. Assume $\varphi'(x) - \varphi'(y) \geq C(y - x)$ for $C > 0$ and $A \leq x < y \leq B$. Then:*

$$\sup_{i=2,\ldots,M(n)} (s_i - s_{i-1}) = O_{\mathrm{p}}\Big(\rho_n^{\beta/(4\beta+2)}\Big).$$

This result completely corresponds to convex density estimation, as the rate of convergence of two consecutive knots is of order root of the pointwise rate of the density estimator (anticipating the log-concave pointwise rate from the uniform rate in Theorem 3.3.1). However, there the knots are between observation points what makes it much more difficult to receive a result that compares to Theorem 3.4.1. In fact, in proving the result about the pointwise limiting distribution in Groeneboom, Jongbloed, and Wellner (2001b), the distance about the distance of two consecutive knots is the key result in the whole proof.

The situation is different for density estimation under a monotonicity constraint. The Grenander density estimator $\widehat{f}_G$ is the left-sided derivative of the least concave majorant $\widehat{F}_G$ of the empirical distribution function, implying that the jumps of the estimator are at observation points. In Jonker and van der Vaart (2001) appears a uniform rate of convergence for $\widehat{f}_G$ together with the distance between two consecutive changes of slope of $\widehat{F}_G$ as a corollary of a more general statement about monotone estimation with censored data. These two rates of convergence are equal, up to a log-factor for the uniform rate, namely $O_{\mathrm{p}}(n^{-1/3})$.

In estimation of k-monotone densities, Balabdaoui and Wellner (2004d) derived the rate of convergence of the difference between two consecutive knots in a neighborhood of a fixed point $x_o > 0$ only assuming that a certain unproven conjecture about the upper bound on the error in a particular Hermite interpolation problem holds true. Clearly, as k-monotone densities are a generalization of convex decreasing densities, the whole limiting distribution theory again relies on the solution of the gap problem and therefore on the abovementioned conjecture. Note that Balabdaoui and Wellner introduced the term "gap problem".

Theorem 3.4.1 solves a gap problem in log-concave density estimation, via some relatively fundamental geometrical considerations (see the proof of the theorem on p. 57). However, the crucial point in our case is that the knot points of the estimator $\widehat{\varphi}_n$ are at some of the observations X_i, and not strictly inbetween as in all k-monotone cases for $k \geq 2$.

3.5 Uniform consistency of $\widehat{F}_n$

Note that log-concavity is preserved under integration, see Bagnoli and Bergstrom (1989 and 2005, Theorem 1). Using Theorem 3.3.1 together with Theorem 3.4.1 and a theorem elaborated in Stute (1982) about the modulus of continuity of a uniform empirical process, one can deduce an at least rate of convergence for the difference between the integrated density estimator $\widehat{F}_n$ and the empirical distribution function $\mathbb{F}_n$. Two things are important to note. First, the proof of the theorem reveals why the case $\beta = 1$ has to be excluded. Second, additionally to the conditions in Theorem 3.3.1, the derivative of the log-density, which is well-defined (because $\beta > 1$), has to be bounded from below.

Theorem 3.5.1. *Assume $\varphi'(x) - \varphi'(y) \geq C(y-x)$ for $C > 0$ and $A \leq x < y \leq B$. Suppose that $\varphi \in \mathcal{H}^{\beta,L}(T)$ for some $\beta \in (1, 2]$. Then,*

$$\begin{aligned} \max_{t \in T} (\widehat{F}_n - \mathbb{F}_n)(t) &= o_{\mathrm{p}}\left(n^{-1/2}\right), \\ \max_{t \in [A+\rho_n^{\beta/(4\beta+2)}, B-\rho_n^{\beta/(4\beta+2)}]} (\mathbb{F}_n - \widehat{F}_n)(t) &= o_{\mathrm{p}}\left(n^{-1/2}\right). \end{aligned} \tag{3.10}$$

The interval in (3.10) is slightly shorter (for finite n) than that in (3.9). This ensures that we have at least one knot between A and the place where the maximum occurs (same for B).
Using Theorem 3.5.1 together with the well known Dvoretzky-Kiefer-Wolfowitz inequality (Theorem A.3.1) we easily get the following corollary.

Corollary 3.5.2. *Under the same assumptions as in Theorem 3.5.1 we have:*

$$\max_{t \in [A+\rho_n^{\beta/(4\beta+2)}, B-\rho_n^{\beta/(4\beta+2)}]} |(\widehat{F}_n - F)(t)| = O_{\mathrm{p}}\left(n^{-1/2}\right).$$

In most simulations we looked at, the estimator $\widehat{F}_n$ satisfied the inequality

$$\|\widehat{F}_n - \mathbb{F}_n\|_\infty^{\mathbb{R}} \leq \|F - \mathbb{F}_n\|_\infty^{\mathbb{R}}. \tag{3.11}$$

However, one can construct counterexamples showing that (3.11) may be violated, even if the right hand side is multiplied with any fixed constant $C > 1$. The latter findings are in contrast to "Marshall's Lemma" about the Grenander estimator $\widehat{F}_G$.

Lemma 3.5.3 (Marshall (1970)). *Suppose that F is concave on $[0,\infty)$ such that $F(0)=0$. The least concave majorant $\widehat{F}_G$ of $\mathbb{F}_n$ then satisfies:*

$$\|\widehat{F}_G - F\|_\infty^{[0,\infty)} \;\leq\; \|\mathbb{F}_n - F\|_\infty^{[0,\infty)}.$$

Note that the distribution function estimator $\widehat{f}_G$ corresponding to $\widehat{F}_G$ is a piecewise constant monotone decreasing function. Kiefer and Wolfowitz (1976) showed that

$$\|\widehat{F}_G - \mathbb{F}_n\|_\infty^{[0,\infty)} \;=\; o_{\mathrm{p}}(n^{-2/3}(\log n)^{5/6}).$$

Kulikov and Lopuhaä (2005b) derived the limiting process of

$$G_n(t) \;:=\; n^{2/3}\Big(\widehat{F}_G(t) - \mathbb{F}_n(t)\Big)_{t\in[0,1]}.$$

Note that $\widehat{F}_G$ is quite well accessible through its characterization as concave majorant of $\mathbb{F}_n$. However, to derive similar results in the log-concave (and convex) case one has presumably to rely on the characterization of the estimator given in Theorem 3.2.3.

Theorem 3.5.1 assures that essentially the empirical distribution function and the estimator $\widehat{F}_n$ are equivalent up to a fast rate, at least on a fixed compact interval T. Together with Theorem 3.5.4 this reveals a remarkable advantage of the log-concave density estimator over kernel estimators. If the latter are constructed with a non-negative even kernel and a bandwidth of optimal order $O(n^{-1/5})$, then the uniform distance between integrated density estimator $\widehat{F}_{n,h}$ and the true distribution function F is only of order $O_{\mathrm{p}}(n^{-2/5})$, i.e. even worse than the simple empirical distribution function while in the log-concave case the parametric rate $O_{\mathrm{p}}(n^{-1/2})$ is attained.

Theorem 3.5.4. *Let k be a nonnegative and symmetric kernel and K its normalized integral:*

$$K(r) \;:=\; \int_{-\infty}^{r} k(x)\,\mathrm{d}x \quad \textit{such that } K(\infty) = 1.$$

For a bandwidth $h = h(n)$ such that $h \downarrow 0$ and $nh \to \infty$, the integrated kernel density estimator is defined as

$$\widehat{F}_{n,h}(x) := \int_{\mathbb{R}} K(x-y)\,\mathrm{d}\mathbb{F}_n(y)$$

for any $x \in \mathbb{R}$. Then, if the true density f has bounded derivative f' at any fixed $x_o \in \mathbb{R}$,

$$\widehat{F}_{n,h}(x_o) \;=\; F(x_o) + O_{\mathrm{p}}(n^{-1/2}) + O_{\mathrm{p}}(h^2 f'(x_o)). \tag{3.12}$$

If f' is strictly positive at x_o, choosing $h = O_{\mathrm{p}}(n^{-1/5})$ in (3.12) yields:

$$\widehat{F}_{n,h}(x_o) \;=\; F(x_o) + O_{\mathrm{p}}(n^{-2/5}).$$

3.6 A monotone hazard rate estimator

The estimation of a monotone hazard rate is already described in the book by Robertson, Wright, and Dykstra (1988). They directly solve an isotonic estimation problem similar to that for the Grenander density estimator.
Recently, there has again grown some interest in the estimation of a monotone hazard rate, see Hall et al. (2001) and Hall and van Keilegom (2005). Methods used there relied upon suitable modifications of kernel estimators and Silverman's "increasing bandwidth" approach, proposed in 1981. However, with the aid of Lemma 2.3.1 and defining

$$\widehat{\lambda}_n(x) \;=\; \frac{\widehat{f}_n(x)}{1 - \widehat{F}_n(x)} \quad \text{for } x < X_n$$

yields a simple plug-in monotone hazard rate estimator and gives raise to the following theorem.

Theorem 3.6.1. *Under the same assumptions as in Theorem 3.3.1 we have that* $\widehat{\lambda}_n$ *is a non-decreasing function on* $(-\infty, X_n)$. *Furthermore,*

$$\begin{aligned} \max_{t\in T} (\widehat{\lambda}_n - \lambda)(t) &= O_p\left(\rho_n^{\beta/(2\beta+1)}\right), \\ \max_{t\in\left[A+\rho_n^{1/(2\beta+1)}, B-\rho_n^{1/(2\beta+1)}\right]} (\lambda - \widehat{\lambda}_n)(t) &= O_p\left(\rho_n^{\beta/(2\beta+1)}\right). \end{aligned}$$

Find graphical illustrations for all the estimators $\widehat{f}_n, \widehat{\varphi}_n, \widehat{F}_n$ and $\widehat{\lambda}_n$ in Chapter 4.

3.7 Proofs

Before coming to the proofs let us mention that vectors in $\mathbb{R}^n$ are written as $\boldsymbol{x} = (x_1, \ldots, x_n)$ and that the L_2-norm for a vector $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$\|\boldsymbol{x}\|_2 := \Big(\sum_{i=1}^{n} x_i^2\Big)^{1/2}.$$

Existence and uniqueness

Proof of Theorem 3.2.1. We start with proving piecewise linearity of $\widehat{\varphi}_n$. Fix an arbitrary concave function φ with $\Psi_n(\varphi) < \infty$, and define $\bar{\varphi}$ by requiring that $\bar{\varphi}(X_i) = \varphi(X_i)$ for all $i = 1, \ldots, n$, while $\bar{\varphi}$ is linear between successive observations. Further let $\bar{\varphi} \equiv -\infty$ outside $[X_1, X_n]$. The concavity of φ then entails that $\varphi \geq \bar{\varphi}$. Consequently,

$$\Psi_n(\bar{\varphi}) \;\leq\; \Psi_n(\varphi) \tag{3.13}$$

with strict inequality unless $\bar{\varphi} = \varphi$. Thus minimizers of Ψ_n must have the form of $\bar{\varphi}$.

In order to prove existence of $\widehat{\varphi}_n$, we only consider concave functions φ satisfying the constraints just derived. Moreover it suffices to consider the case that $\int \exp \varphi(x)\,\mathrm{d}x = 1$. For if $\varphi = \varphi_o + t$ with $\exp(\varphi_o)$ being a probability density and some number $t \neq 0$, it follows from (3.2) that

$$\Psi_n(\varphi) \;=\; \Psi_n(\varphi_o) + n\Big(\exp(t) - t - 1\Big) \;>\; \Psi_n(\varphi_o).$$

For the remainder of this proof, any such function φ is identified with the vector

$$\boldsymbol{\varphi} := \Big(\varphi(X_i)\Big)_{i=1}^{n} \;\in\; \mathbb{R}^n.$$

Note that the functional $\varphi \mapsto \Psi_n(\varphi)$ is continuous. Thus for the existence of a minimizer it suffices to show that

$$\Psi_n(\varphi) \;\to\; \infty$$

whenever $\|\boldsymbol{\varphi}\|_2 \to \infty$. For that purpose, let $(\boldsymbol{\varphi}^{(k)})_{k=1}^{\infty}$ be a sequence of such vectors satisfying

$$\|\boldsymbol{\varphi}^{(k)}\|_2 \;\to\; \infty$$

and

$$\varphi_i^{(k)} \;\to\; \gamma_i \in [-\infty, \infty] \quad \text{for } i = 1, \ldots, n.$$

Suppose first that $\gamma_i < \infty$ for all i. Then $\gamma_i = -\infty$ for at least one index i, so that $\Psi_n(\boldsymbol{\varphi}^{(k)}) = -\sum_{i=1}^n \varphi_i^{(k)} + n$ tends to infinity.
Secondly, suppose there exists an index j with $\gamma_j = \infty$. Let $j > 1$. The piecewise linearity of the function $\varphi^{(k)}$ entails that

$$\begin{aligned} 1 &\geq \int_{X_{j-1}}^{X_j} \exp\Big(\varphi^{(k)}(x)\Big)\,\mathrm{d}x \\ &= (X_j - X_{j-1}) \exp\left(\varphi_j^{(k)}\right) \frac{1 - \exp\left(-\delta_k\right)}{\delta_k} \\ &\geq (X_j - X_{j-1}) \exp\left(\varphi_j^{(k)}\right) (1 + \delta_k)^{-1}, \end{aligned}$$

where $\delta_k := \varphi_j^{(k)} - \varphi_{j-1}^{(k)}$. The latter inequality is a consequence of

$$\frac{1 - e^{-x}}{x} \;\geq\; \frac{1}{1 + x} \quad \text{for } x \geq 0.$$

Thus δ_k is bounded from below by $(X_j - X_{j-1}) \exp(\varphi_j^{(k)}) - 1$. Consequently, $\gamma_j = \infty$ entails that

$$\begin{aligned} -\varphi_j^{(k)} - \varphi_{j-1}^{(k)} &= -2\varphi_j^{(k)} + \delta_k \\ &\geq -2\varphi_j^{(k)} + (X_j - X_{j-1}) \exp(\varphi_j^{(k)}) - 1 \\ &\to \infty. \end{aligned}$$

Analogously, if $j < n$, then $-\varphi_j^{(k)} - \varphi_{j+1}^{(k)}$ tends to infinity. These considerations show that $\Psi_n(\varphi^{(k)}) \to \infty$.
For uniqueness observe that Ψ_n is a strictly convex functional in φ in the sense that

$$\Psi_n\Big((1 - \lambda)\varphi^1 + \lambda\varphi^2\Big) < (1 - \lambda)\Psi_n(\varphi^1) + \lambda\Psi_n(\varphi^2)$$

for $\lambda \in (0, 1)$ and concave functions $\varphi^1, \varphi^2 : \mathbb{R} \mapsto [-\infty, \infty)$ such that $\int \exp \varphi^i < \infty$ and $\mathrm{Leb}\{\varphi^1 \neq \varphi^2\} > 0$. This is a consequence of the strict convexity of the exponential function. □

Characterizations

To simplify notation in the following proofs, let us introduce three function classes. For a concave function $g_n : [X_1, X_n] \to \mathbb{R}$, let $\mathcal{D}^1(g_n)$ be the class of all functions Δ such that $g_n + t\Delta$ is concave for some $t > 0$. Define $\mathcal{D}^2(g_n)$ as the family of piecewise linear (not necessarily continuous) functions Δ such that any knot q of Δ has one of the two following properties:

$$q \in \mathcal{S}(g_n) \quad \text{and} \quad \Delta(q) = \liminf_{r \to q} \Delta(r), \tag{3.14}$$

$$\Delta(q) = \lim_{r \to q} \Delta(r) \quad \text{and} \quad \Delta'(q-) \geq \Delta'(q+). \tag{3.15}$$

Finally, $\mathcal{D}^3(g_n)$ shall be the subset of $\mathcal{D}^2(g_n)$ consisting of all continuous and piecewise linear functions with knots only in $\mathcal{S}(g_n)$. See Figure 3.2 for two examples of admissible perturbation functions in $\mathcal{D}^2(g_n)$.

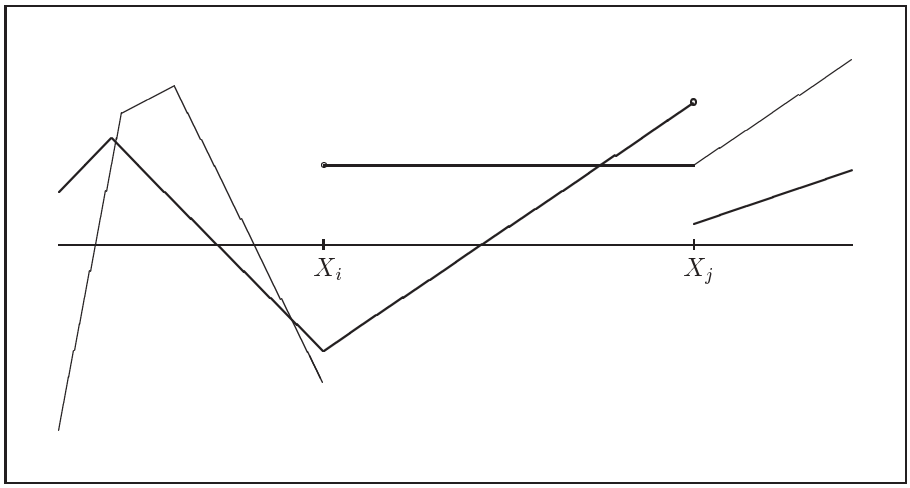

Figure 3.2: Two examples for admissible perturbation functions $\Delta \in \mathcal{D}^2(g_n)$.

In Theorem 3.2.2 perturbation functions $\Delta \in \mathcal{D}^1(\widehat{\varphi}_n)$ are used to characterize the estimator $\widehat{\varphi}_n$. We can generalize and specify inequality (3.3) to the even more general classes $\mathcal{D}^2(\widehat{\varphi}_n)$ and $\mathcal{D}^3(\widehat{\varphi}_n)$, see the following lemma.

Lemma 3.7.1. *Inequality (3.3) is also valid for functions $\Delta \in \mathcal{D}^2(\tilde{\varphi}_n)$. For functions $\Delta \in \mathcal{D}^3(\tilde{\varphi}_n)$, we even get an equality.*

Proof of Lemma 3.7.1. Suppose that $\Delta \in \mathcal{D}^2(\widehat{\varphi}_n)$. In this case there are continuous, piecewise linear functions Δ_k for $k \in \mathbb{N}$ converging pointwise isotonically to Δ and having the following property: Any knot point q of Δ_k either belongs to $\mathcal{S}(\widehat{\varphi}_n)$, or $\Delta_k'(q-) > \Delta_k'(q+)$; see Figure 3.3.

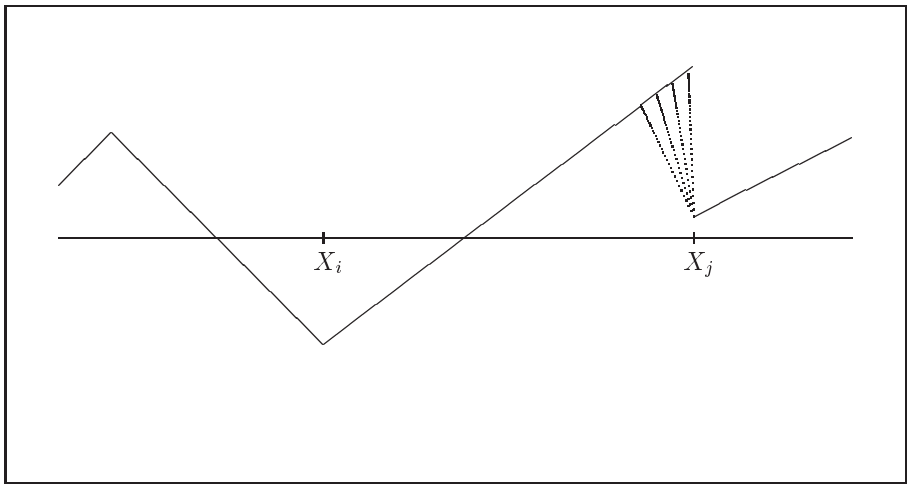

Figure 3.3: An example for an admissible perturbation function Δ and some approximations Δ_k.

Thus $\widehat{\varphi}_n + t\Delta_k$ is concave for sufficiently small $t > 0$. Consequently, since $\Delta_1 \leq \Delta_k \leq \Delta$ for all k, it follows from dominated convergence (Theorem A.1.1) and (3.3) that

$$\int \Delta \, \mathrm{d}\mathbb{F}_n \;=\; \lim_{k\to\infty} \int \Delta_k \, \mathrm{d}\mathbb{F}_n \;\leq\; \lim_{k\to\infty} \int \Delta_k(x) \widehat{f}_n(x) \, \mathrm{d}x \;=\; \int \Delta(x) \widehat{f}_n(x) \, \mathrm{d}x.$$

Finally, if $\Delta \in \mathcal{D}^3(\widehat{\varphi}_n)$, one may apply (3.3) to $\pm\Delta$ and obtains equality in (3.3). □

Proof of Theorem 3.2.2. First suppose $\tilde{\varphi}_n$ is a minimizer of Ψ_n. This entails for any function $\Delta \in \mathcal{D}^1(\tilde{\varphi}_n)$ that the corresponding directional derivative of Ψ_n must be non-negative:

$$\begin{aligned} 0 \;&\leq\; \lim_{t\downarrow 0} \frac{\Psi_n(\tilde{\varphi}_n + t\Delta) - \Psi_n(\tilde{\varphi}_n)}{t} \\ &=\; n\Big(- \int \Delta \, \mathrm{d}\mathbb{F}_n + \int \Delta(x) \exp \tilde{\varphi}_n(x) \, \mathrm{d}x \Big). \end{aligned}$$

As for the other direction let g be a concave function such that $\Psi_n(g) < \infty$ and define $g(r) - \tilde{\varphi}_n(r) = 0$ for $r \in \{-\infty, \infty\}$. Then:

$$\begin{aligned}
n^{-1}\Big(\Psi_n(g) - \Psi_n(\tilde{\varphi}_n)\Big) &= \\
&= \int \exp g(x)\,\mathrm{d}x - \int \Big(g(x) - \tilde{\varphi}_n(x)\Big)\,\mathrm{d}\mathbb{F}_n(x) - \int \tilde{f}_n(x)\,\mathrm{d}x \\
&= \int \exp\Big(g(x) - \tilde{\varphi}_n(x)\Big)\tilde{f}_n(x)\,\mathrm{d}x - \int \Big(g(x) - \tilde{\varphi}_n(x)\Big)\,\mathrm{d}\mathbb{F}_n(x) - \int \tilde{f}_n(x)\,\mathrm{d}x \\
&\geq \int \Big(1 + g(x) - \tilde{\varphi}_n(x)\Big)\tilde{f}_n(x)\,\mathrm{d}x - \int \Big(g(x) - \tilde{\varphi}_n(x)\Big)\,\mathrm{d}\mathbb{F}_n(x) - \int \tilde{f}_n(x)\,\mathrm{d}x \\
&= \int \Big(g(x) - \tilde{\varphi}_n(x)\Big)\tilde{f}_n(x)\,\mathrm{d}x - \int \Big(g(x) - \tilde{\varphi}_n(x)\Big)\,\mathrm{d}\mathbb{F}_n(x) \qquad (3.16)
\end{aligned}$$

by the inequality $\exp(x) \geq 1 + x$ for $x \in \mathbb{R}$. But the class of functions

$$\{g - \tilde{\varphi}_n,\ \ g \text{ concave}\}$$

is equivalent to the class $\mathcal{D}^1(\tilde{\varphi}_n)$, so (3.16) is only positive if (3.3) holds for all functions in $\mathcal{D}^1(\tilde{\varphi}_n)$, entailing that $\tilde{\varphi}_n$ is effectively the minimizer of Ψ_n. □

Proof of Theorem 3.2.3. First, we provide a formula about integration of a special class of functions. Assume G to be an arbitrary distribution function. Suppose $\Delta : \mathbb{R} \to \mathbb{R}$ can be written as follows.

$$\Delta(x) \;=\; \Delta_o + \int_{-\infty}^{x} \Delta'(t)\,\mathrm{d}t$$

where Δ' is a bounded and measurable function with bounded support. Then, using Fubini's Theorem:

$$\begin{aligned}
\int_{\mathbb{R}} \Delta\,\mathrm{d}G(x) &= \Delta_o + \int_{\mathbb{R}}\int_{\mathbb{R}} \Delta'(t) 1_{\{t<x\}}\,\mathrm{d}t\,\mathrm{d}G(x) \\
&= \Delta_o + \int_{\mathbb{R}} \Delta'(t)\Big(\int_{\mathbb{R}} 1_{\{t<x\}}\,\mathrm{d}G(x)\Big)\,\mathrm{d}t \\
&= \Delta_o + \int_{\mathbb{R}} \Delta'(t)[1 - G(t)]\,\mathrm{d}t. \qquad (3.17)
\end{aligned}$$

Equality (3.17) is specifically valid for piecewise linear and continuous functions Δ with bounded support.

Suppose $\tilde{\varphi}_n$ is a minimizer of Ψ_n. Then inequalities (3.4)-(3.6) follow from Theorem 3.2.2 applied to

$$\begin{aligned}\Delta_1(x) &= \min\{(b-x)_+, b-t\} \\ &= (b-t) + \int_{-\infty}^{x} -1_{\{t\le r\le b\}}\,\mathrm{d}r\end{aligned}$$

and

$$\begin{aligned}\Delta_2(x) &= \min\{(x-a)_+, t-a\} \\ &= \int_{-\infty}^{x} 1_{\{a\le r\le t\}}\,\mathrm{d}r\end{aligned}$$

and remembering that $a, b \in \mathcal{S}(\tilde{\varphi}_n)$.
As for the other direction let us just continue calculations in (3.16) as follows. Recall from the proof of Theorem 3.2.1 the function $\bar{g}$ which is concave and piecewise linear with knots only at the observations $X_1, \ldots, X_n$. Using inequalities (3.13), (3.16), the assumption $\tilde{F}_n(X_n) = 1$, and (3.17) then yields:

$$\begin{aligned}n^{-1}\Big(\Psi_n(g) - \Psi_n(\tilde{\varphi}_n)\Big) &\ge n^{-1}\Big(\Psi_n(\bar{g}) - \Psi_n(\tilde{\varphi}_n)\Big) \\ &\ge \int \Big(g(x) - \tilde{\varphi}_n(x)\Big)\tilde{f}_n(x)\,\mathrm{d}x - \int \Big(g(x) - \tilde{\varphi}_n(x)\Big)\,\mathrm{d}\mathbb{F}_n(x) \\ &= \int_{\mathbb{R}} \Big(\mathbb{F}_n(x) - \tilde{F}_n(x)\Big)\Big(\bar{g}'(x) - \tilde{\varphi}_n'(x)\Big)\,\mathrm{d}x \\ &= -\int_{X_1}^{X_n}\int_{X_1}^{t} \Big(\mathbb{F}_n(x) - \tilde{F}_n(x)\Big)\,\mathrm{d}x\,\mathrm{d}\Big(\bar{g}'(t) - \tilde{\varphi}_n'(t)\Big)\end{aligned}$$

using integration by parts where $\bar{g}'$ and $\tilde{\varphi}_n'$ can be interpreted as left-sided derivatives. Note that the outer integration over $\mathrm{d}(\bar{g}'(t) - \tilde{\varphi}_n'(t))$ is just a sum over the knot points. The assumption on $\tilde{\varphi}_n$ entails that

$$\begin{aligned}&\int_{X_1}^{X_n}\int_{X_1}^{t} \Big(\mathbb{F}_n(x) - \tilde{F}_n(x)\Big)\,\mathrm{d}x\,\mathrm{d}\tilde{\varphi}_n'(t) = \\ &= \sum_{s\in\mathcal{S}(\tilde{\varphi}_n)} (\tilde{\varphi}'(s+) - \varphi'(s)) \int_{X_1}^{s} \Big(\mathbb{F}_n(x) - \tilde{F}_n(x)\Big)\,\mathrm{d}x \\ &= 0\end{aligned}$$

by (3.6). Define for $i = 2, \ldots, n$ the right-most knot of $\tilde{\varphi}_n$ left of X_i as

$$s_i = \max_j\{s_j \in \mathcal{S}(\tilde{\varphi}_n) : s_j < X_i\}.$$

Introduce $\bar{g}_i'' = \bar{g}'(X_i+) - \bar{g}'(X_i) < 0$ and use the calculations from above to get:

$$\begin{aligned} & n^{-1}\Big(\Psi_n(g) - \Psi_n(\tilde{\varphi}_n)\Big) \\ & \geq \; -\int_{X_1}^{X_n} \int_{X_1}^{t} \Big(\mathbb{F}_n(x) - \tilde{F}_n(x)\Big)\, \mathrm{d}x\, \mathrm{d}\bar{g}'(t) \\ & = \; \sum_{i=2}^{n} (-\bar{g}_i'') \Big[\int_{X_1}^{s_i} \Big(\mathbb{F}_n(x) - \tilde{F}_n(x)\Big)\, \mathrm{d}x + \int_{s_i}^{X_i} \Big(\mathbb{F}_n(x) - \tilde{F}_n(x)\Big)\, \mathrm{d}x\Big] \\ & \geq \; 0 \end{aligned}$$

by (3.4) and (3.6). □

Uniform consistency of $\widehat{f}_n$

Proof of Theorem 3.3.1: The proof consists of several lemmas. To lift the fog spread by the technical details, we summarize the ingredients. First, define $\mathcal{D}_m$ as the family of all piecewise linear functions on $\mathbb{R}$ with at most m knots. Second, verify that the class $\mathcal{D}_m$ indeed contains useful perturbation functions (for a fixed m, Lemma 3.7.3) in the sense of providing sufficiently accurate "caricatures" for the difference $\widehat{\varphi}_n - \varphi$. Finally, bound the moment generating function of a random variable specified there (Lemma 3.7.5) to show that the supremum norm of a suitably weighted empirical process $(w_n(\Delta) \int \Delta \, \mathrm{d}(\mathbb{F}_n - F))_{\Delta \in \mathcal{D}_m}$ is bounded in probability (Lemma 3.7.4). This last step is done by approximating elements of $\mathcal{D}_m$ by linear functions from a finite family (Lemma 3.7.6) to be followed by some bracketing argument. Finally, to prove the theorem, use Lemma 3.7.2. This claim about the difference of two concave functions (one of which is sufficiently smooth) was introduced in slightly different form in Dümbgen (1998, Lemma 5.2) and readopted in Dümbgen, Freitag, and Jongbloed (2004, Lemma 2). For completeness, we also give a proof of this lemma.

It is important to note that thanks to inequality (3.3) we can concentrate our attention in Lemma 3.7.4 on the rescaled supremum of the standard empirical process $(F(t) - \mathbb{F}_n(t))_{t \in \mathbb{R}}$ rather than having to deal with $(F(t) - \widehat{F}_n(t))_{t \in \mathbb{R}}$, what in fact would be a much more difficult task.

Recall that, according to (3.8), φ is assumed to satisfy $\varphi \leq -1$. In order to be able to state the following results rigorously we define two auxiliary quantities for any

function h on the real line:

$$W(h) := \|h/\varphi\|_\infty^{\mathbb{R}} \qquad\qquad \sigma(h) := \left(\int_{\mathbb{R}} h(x)^2 \, \mathrm{d}F(x)\right)^{1/2}.$$

The first key ingredient in the proof of Theorem 3.3.1 is a statement about the difference of two concave functions, one of which is sufficiently smooth.

Lemma 3.7.2. *For any $\beta \in [1,2]$ and $L > 0$ there exists a positive constant $K = K(\beta, L)$ with the following property: Suppose that g and $\widehat{g}$ are concave and real-valued functions on a compact interval $T = [A, B]$, where $g \in \mathcal{H}^{\beta,L}(T)$. For any $\varepsilon > 0$ let $\delta := K \min\{B - A, \varepsilon^{1/\beta}\}$. Then*

$$\sup_{t\in T}(\widehat{g} - g) \;\geq\; \varepsilon \quad \text{or} \quad \sup_{t\in[A+\delta,B-\delta]} (g - \widehat{g}) \;\geq\; \varepsilon$$

implies that

$$\inf_{t\in[c,c+\delta]} (\widehat{g} - g)(t) \;\geq\; \varepsilon/4 \quad \text{or} \quad \inf_{t\in[c,c+\delta]} (g - \widehat{g})(t) \;\geq\; \varepsilon/4$$

for some $c \in [A, B - \delta]$.

This is followed by the specification of "useful" perturbation functions Δ.

Lemma 3.7.3. *Let $\varphi - \widehat{\varphi}_n \geq \varepsilon$ or $\widehat{\varphi}_n - \varphi \geq \varepsilon$ on some interval $[c, c + \delta] \subset T$ with length $\delta > 0$. Then there exists a function $\Delta \in \mathcal{D}_3$ each knot of which satisfies condition (3.14) or (3.15) and a positive constant $K = K(f, T)$ such that*

$$\begin{aligned} &\widehat{\varphi}_n - \varphi \leq -\varepsilon\Delta \;\text{ if }\; \varphi - \widehat{\varphi}_n \geq \varepsilon \;\text{ on }\; [c, c+\delta] \\ &\widehat{\varphi}_n - \varphi \geq -\varepsilon\Delta \;\text{ if }\; \widehat{\varphi}_n - \varphi \geq \varepsilon \;\text{ on }\; [c, c+\delta], \end{aligned} \tag{3.18}$$

$$\operatorname{sign}(\Delta) \;=\; \operatorname{sign}(\varphi - \widehat{\varphi}_n) \;\text{ on }\; \{x \,:\, \Delta(x) \neq 0\}\,, \tag{3.19}$$

$$\Delta \;\leq\; 1 \;\text{ on }\; \mathbb{R} \tag{3.20}$$

$$\int_c^{c+\delta} \Delta^2(x)\,\mathrm{d}x \;\geq \delta/3, \tag{3.21}$$

$$W(\Delta) \;\leq\; K(f,T)\max\{1, \delta^{-1/2}\}\sigma(\Delta). \tag{3.22}$$

Lemma 3.7.4 shows that $\mathbb{F}_n$ is close to F uniformly over the function class $\mathcal{D}_m$.

Lemma 3.7.4. *For any $\kappa \in [2/3, 1)$ there exists a constant $B = B(\kappa, f)$ such that*

$$S_n(m) \;:=\; \sup_{\Delta \in \mathcal{D}_m} \frac{\left|\int \Delta \,\mathrm{d}(\mathbb{F}_n - F)\right|}{\sigma(\Delta) m^{1/2} \rho_n^{1/2} + W(\Delta) m \rho_n^{\kappa}} \le B$$

with probability tending to one as $n \to \infty$.

The additional term $W(\Delta) m \rho_n^{\kappa}$ in the denominator is necessary to prevent $S_n(m)$ from becoming "too big" in case $\sigma(\Delta)$ is very small. This latter problem can occur when the perturbation function Δ has small support.

Proof of Theorem 3.3.1

Now, to prove the theorem let $G = G(\kappa, f, T) > 0$ be a generic constant whose value may be different in different expressions. Since the exponential function is Lipschitz–continuous on any halfline $(-\infty, m]$, we may and do replace $(f, \widehat{f}_n)$ with $(\varphi, \widehat{\varphi}_n)$. Suppose that

$$\sup_{t \in T} (\widehat{\varphi}_n - \varphi)(t) \;\ge\; C\varepsilon_n$$

or

$$\sup_{t \in [A+\delta_n, B-\delta_n]} (\varphi - \widehat{\varphi}_n)(t) \;\ge\; C\varepsilon_n$$

for some $C > 0$, where $\varepsilon_n := \rho_n^{\beta/(2\beta+1)}$ and $\delta_n := \rho_n^{1/(2\beta+1)} = \varepsilon_n^{1/\beta}$. It follows from Lemma 3.7.2 with $\varepsilon := C\varepsilon_n$ that for sufficiently large n and $C \ge K(f,T)^{-\beta}$, there is a (random) interval $[c_n, c_n + \delta_n] \subset T$ on which either $\widehat{\varphi}_n - \varphi \ge (C/4)\varepsilon_n$ or $\varphi - \widehat{\varphi}_n \ge (C/4)\varepsilon_n$. But then by Lemma 3.7.3 there is a (random) function $\Delta_n \in \mathcal{D}_3 \subset \mathcal{D}^2(\widehat{\varphi}_n)$ fulfilling (3.18)-(3.22). For this Δ_n we have by (3.3)

$$\begin{aligned} \int_{\mathbb{R}} \Delta_n(x) \,\mathrm{d}(F - \mathbb{F}_n)(x) &\ge \int_{\mathbb{R}} \Delta_n(x)(f - \widehat{f}_n)(x) \,\mathrm{d}x \\ &= \int_{\mathbb{R}} \Delta_n(x) f(x) \Big(1 - \exp[\widehat{\varphi}_n(x) - \varphi(x)]\Big) \,\mathrm{d}x \end{aligned} \tag{3.23}$$

From (3.18) and the assumption above we get on the interval $[c_n, c_n + \delta_n]$

$$\widehat{\varphi}_n - \varphi \;\ge\; -(C/4)\varepsilon_n \Delta_n$$

if $\widehat{\varphi}_n - \varphi \geq (C/4)\varepsilon_n$ and

$$\widehat{\varphi}_n - \varphi \;\; \leq \;\; -(C/4)\varepsilon_n \Delta_n$$

if $\varphi - \widehat{\varphi}_n \geq (C/4)\varepsilon_n$. This together with (3.19) and the fact that the function $1 - \exp(x)$ is decreasing for $x \in \mathbb{R}$ implies that (3.23) is not smaller than

$$\int_{\mathbb{R}} \Delta_n(x) f(x) \Big(1 - \exp[-(C/4)\varepsilon_n \Delta_n(x)]\Big) \, \mathrm{d}x = 4(C\varepsilon_n)^{-1} \int_{\mathbb{R}} \tilde{\Delta}_n(x) f(x) \Big(1 - \exp[-\tilde{\Delta}_n(x)]\Big) \, \mathrm{d}x$$

with $\tilde{\Delta}_n := (C/4)\varepsilon_n \Delta_n$. Using Taylor expansion one can verify the inequalities

$$x[1 - \exp(-x)] \geq \begin{cases} x^2 & \text{if } x \leq 0 \\ x^2/(1+x) & \text{if } x > 0. \end{cases}$$

Combining this with the above derivations yields

$$\begin{aligned} &\int_{\mathbb{R}} \Delta_n(x) \, \mathrm{d}(F - \mathbb{F}_n)(x) \geq \\ &\qquad 4(C\varepsilon_n)^{-1} \int_{\{\tilde{\Delta}_n \leq 0\}} \tilde{\Delta}_n^2(x) f(x) \, \mathrm{d}x + 4(C\varepsilon_n)^{-1} \int_{\{\tilde{\Delta}_n > 0\}} \frac{\tilde{\Delta}_n^2(x) f(x)}{1 + \tilde{\Delta}_n(x)} \, \mathrm{d}x \\ &\geq \; (C/4)\varepsilon_n \int_{\{\Delta_n \leq 0\}} \Delta_n^2(x) f(x) \, \mathrm{d}x + \frac{(C/4)\varepsilon_n}{1 + (C/4)\varepsilon_n} \int_{\{\Delta_n > 0\}} \Delta_n^2(x) f(x) \, \mathrm{d}x \\ &\geq \; \frac{(C/4)\varepsilon_n}{1 + (C/4)\varepsilon_n} \sigma^2(\Delta_n) \end{aligned}$$

by (3.20). This entails, together with (3.21) and (3.22),

$$\begin{aligned} S_n(3) \;&\geq\; \frac{\int_{\mathbb{R}} \Delta_n(x) \, \mathrm{d}(F - \mathbb{F}_n)(x)}{3^{1/2}\sigma(\Delta_n)\rho_n^{1/2} + 3W(\Delta_n)\rho_n^{\kappa}} \\ &\geq\; \frac{(C/4)\varepsilon_n \sigma^2(\Delta_n)}{(3^{1/2}\sigma(\Delta_n)\rho_n^{1/2} + G\delta^{-1/2}\sigma(\Delta_n)\rho_n^{\kappa})(1 + (C/4)\varepsilon_n)} \\ &=\; \frac{GC\varepsilon_n \sigma(\Delta_n)}{(\rho_n^{1/2} + \delta_n^{-1/2}\rho_n^{\kappa})(1 + (C/4)\varepsilon_n)} \\ &\geq\; \frac{CG\varepsilon_n \delta_n^{1/2}}{(\rho_n^{1/2} + \delta_n^{-1/2}\rho_n^{\kappa})(1 + (C/4)\varepsilon_n)}. \end{aligned}$$

Consequently, the fact that $S_n(3) \leq B(\kappa, f)$ implies

$$C \leq G(\rho_n^{1/2} + \delta_n^{-1/2}\rho_n^{\kappa})\varepsilon_n^{-1}\delta_n^{-1/2}(1 + (C/4)\varepsilon_n)$$

wherefrom we deduce

$$\begin{aligned} C &\leq G(1 + \rho_n^{\kappa-(\beta+1)/(2\beta+1)})(1 - G\rho_n^{\beta/(2\beta+1)} - G\rho_n^{\kappa-1/(2\beta+1)})^{-1} \\ &= O(1). \end{aligned}$$

Now the assertion follows from Lemma 3.7.4. □

Proof of Lemma 3.7.3. Again, the proof of this Lemma is very much inspired by that of Lemma 3 in Dümbgen, Freitag, and Jongbloed (2004). It is worth noting that here we are also incorporating non-continuous functions, what brings down the number of knots which are necessary for the Δ's from 6 to 3. The crucial point in all the cases we have to distinguish is to construct a $\Delta \in \mathcal{D}_3$ satisfying (3.18).

Case 1: Let $\widehat{\varphi}_n - \varphi \geq \varepsilon$ on $[c, c+\delta]$. Then a function $\Delta \in \mathcal{D}_3$ will do. From Theorem 3.2.1 we know that $\widehat{\varphi}_n$ is piecewise linear.

Case 1a: Suppose $[c, c+\delta] \cap \mathcal{S}(\widehat{\varphi}_n)$ contains (at least) one point X_o. Then we force $\Delta \in \mathcal{D}_3$ to have knots at $c, X_o, c+\delta$, where $\Delta = 0$ on the set $(-\infty, c] \cup [c+\delta, \infty)$, and $\Delta(X_o) = -1$. Requirements (3.18), (3.19), and (3.21) are readily verified. To establish (3.22) note that $W(\Delta) \leq \|\Delta\|_\infty^{\mathbb{R}} \leq 1$.

Case 1b: Suppose $[c, c+\delta] \cap \mathcal{S}(\widehat{\varphi}_n) = \emptyset$. Let $(c_o, d_o) \supset (c, c+\delta)$ be the maximal open interval on which $\varphi - \widehat{\varphi}_n$ is concave. Then there exists a linear function $\tilde{\Delta} < 0$ such that $\tilde{\Delta} \geq \varphi - \widehat{\varphi}_n$ on (c_o, d_o) and $\tilde{\Delta} \leq -\varepsilon$ on $[c, c+\delta]$. Next let $(c_1, d_1) := \{\tilde{\Delta} < 0\} \cap (c_o, d_o)$. Now we define $\Delta \in \mathcal{D}_2$ via

$$\Delta(x) := \begin{cases} 0 & \text{if } x \in (-\infty, c_1) \cup (d_1, \infty), \\ \tilde{\Delta}/\varepsilon & \text{if } x \in [c_1, d_1]. \end{cases}$$

This function Δ satisfies $\varphi - \widehat{\varphi}_n \leq \varepsilon\Delta \leq 0$ on $[X_1, X_n]$, what establishes (3.18) and (3.19). As for (3.21) note that $|\Delta| \geq 1$ on $[c, c+\delta]$. In order to verify (3.22) introduce $\mathcal{P}$, the class of piecewise linear functions such that for every element of $\mathcal{P}$ the interval $[c, c+\delta]$ is fully contained in its support. Let us assume for the moment that

$$\sup_{\delta>0, \Delta \in \mathcal{P}} \min\{1, \delta\}^{1/2} \frac{W(\Delta)}{\sigma(\Delta)} \tag{3.24}$$

is unbounded. But then there exist sequences δ_n and Δ_n such that

$$\min\{1,\delta_n\}^{1/2}\frac{W(\Delta_n)}{\sigma(\Delta_n)} \quad \to \quad \infty$$

as $n \to \infty$. Furthermore, assume w.l.o.g. that Δ_n can be written as

$$\Delta_n(x) \quad = \quad \frac{\beta_n}{d_n - c_n}(x - c_n)1_{\{c_n \le x \le d_n\}}$$

for sequences β_n, c_n and d_n. Since W and σ are both semi-norms, β_n can be set to 1 for all n. As for the other sequences we have $\delta_n \to \delta \in [0,1]$, $c_n \to c_1 \in [-\infty, B]$, and $d_n \to d_1 \in [A, \infty]$. Elementary calculations yield:

$$\begin{aligned}\int_{c_n}^{d_n} \Delta_n^2(x) f(x)\,\mathrm{d}x &\ge 3^{-1}\min_{x\in T} f(x)(d_n - c_n) \\ &= G(d_n - c_n).\end{aligned}$$

Since by Lemma 2.2.1 and equivariance (see Section 3.2) for $x \in \mathbb{R}$

$$|\varphi(x)| \quad \ge \quad \max\{1, -a_o + b_o|x|\} \tag{3.25}$$

we can write:

$$\begin{aligned}\min\{1,\delta_n\}^{1/2}\frac{W(\Delta_n)}{\sigma(\Delta_n)} &\le \frac{G\min\{1,\delta_n\}^{1/2}}{(d_n - c_n)^{1/2}} \sup_{x\in[c_n,d_n]} \frac{x - c_n}{\max\{1, -a_o + b_o|x|\}(d_n - c_n)} \\ &=: R_1(f, T, \delta_n, c_n, d_n).\end{aligned}$$

Note that this latter function is continuous in its last three arguments. Now, assuming that $c_n \to c_1, d_n \to c_1$ for $c_1 \in T$ immediately entails that $\delta_n \to 0$. But then, as $n \to \infty$,

$$\begin{aligned}R_1(f, T, \delta_n, c_n, d_n) &\le \frac{G}{\max\{1, -a_o + b_o|c_1|\}} \\ &= G.\end{aligned}$$

If one considers either the case $c_n \to -\infty, d_n \to d_1 \in [A, \infty), \delta_n \to \delta \in [0,1]$ or $c_n \to c_1 \in (-\infty, B], d_n \to \infty, \delta_n \to \delta \in [0,1]$ one even gets that

$$\begin{aligned}R_1(f, T, \delta_n, c_n, d_n) &= R_1(f, T) \\ &\to 0.\end{aligned}$$

But these considerations imply that

$$\min\{1, \delta_n\}^{1/2} \frac{W(\Delta_n)}{\sigma(\Delta_n)}$$

is at least bounded, what contradicts assumption (3.24). This establishes (3.22). For an illustration consult Figure 3.4.

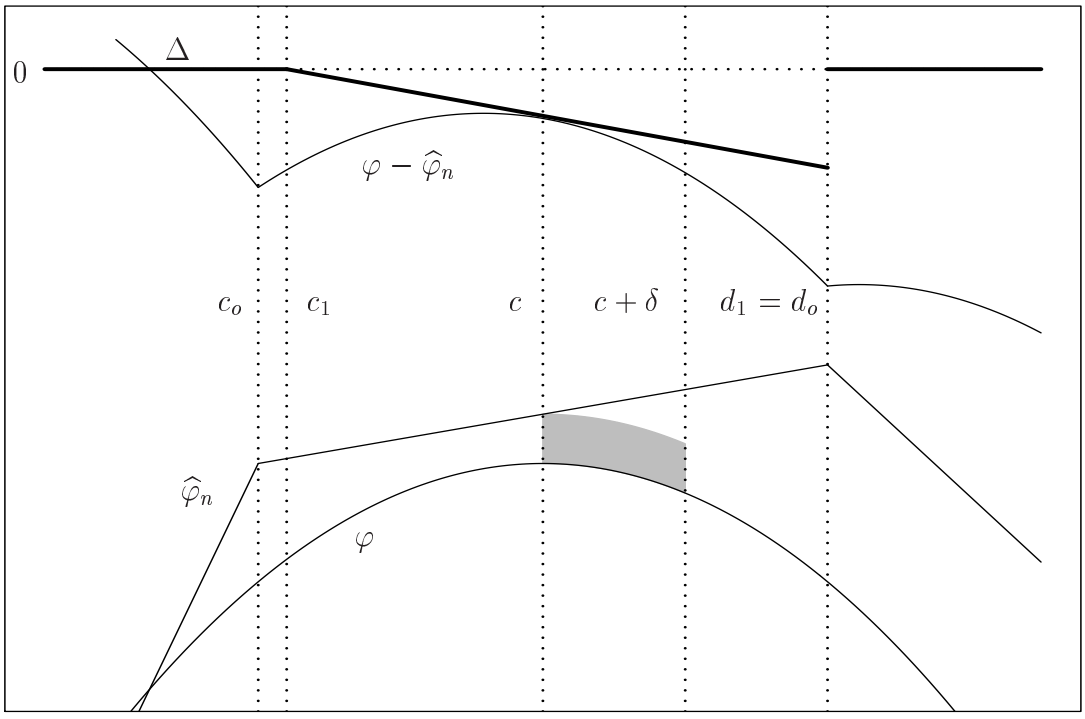

Figure 3.4: The perturbation function Δ in Case 1b.

Case 2: Let $\varphi - \widehat{\varphi}_n \geq \varepsilon$ on $[c, c+\delta]$. Let $[c_o, c]$ and $[c+\delta, d_o]$ be maximal intervals on which $\widehat{\varphi}_n$ is linear. Then define

$$\Delta(x) := \begin{cases} 0 & \text{if } x \in (-\infty, c_o) \cup (d_o, \infty), \\ 1 + \beta_1(x - x_o) & \text{if } x \in [c_o, x_o] \\ 1 + \beta_2(x - x_o) & \text{if } x \in [x_o, d_o], \end{cases}$$

where $x_o := c + \delta/2$ and $\beta_1 \geq 0$ is chosen such that either

$$\Delta(c_o) = 0 \qquad \text{and} \qquad (\varphi - \widehat{\varphi}_n)(c_o) \geq 0 \qquad \text{or}$$
$$(\varphi - \widehat{\varphi}_n)(c_o) < 0 \qquad \text{and} \qquad \operatorname{sign}(\Delta) = \operatorname{sign}(\varphi - \widehat{\varphi}_n) \text{ on } [c_o, x_o].$$

Analogously, $\beta_2 \leq 0$ is chosen such that

$$\begin{aligned} \Delta(d_o) = 0 \qquad &\text{and} \qquad (\varphi - \widehat{\varphi}_n)(d_o) \geq 0 \qquad \text{or} \\ (\varphi - \widehat{\varphi}_n)(d_o) < 0 \qquad &\text{and} \qquad \operatorname{sign}(\Delta) = \operatorname{sign}(\varphi - \widehat{\varphi}_n) \text{ on } [x_o, d_o]. \end{aligned}$$

By construction (3.18) and (3.21) are ensured. Moreover, $\int_c^{c+\delta} \Delta(x)^2 \, \mathrm{d}x \geq \delta/3$. Figure 3.5 gives an example. In order to verify (3.22) one can now apply the same reasoning as in Case 1b. Suppose that

$$\sup_{\delta>0, \Delta \in \mathcal{P}} \min\{1, \delta\}^{1/2} \frac{W(\Delta)}{\sigma(\Delta)} \tag{3.26}$$

is unbounded. Then there exist sequences δ_n and Δ_n such that

$$\min\{1, \delta_n\}^{1/2} \frac{W(\Delta_n)}{\sigma(\Delta_n)} \quad \to \quad \infty$$

as $n \to \infty$. For sequences $c_n, x_n, d_n, \beta_{1,n}, \beta_{2,n}$ write

$$\begin{aligned} \Delta_n(x) &= [1 + \beta_{1,n}(x - x_n)] 1_{\{c_n \leq x \leq x_n\}} + [1 + \beta_{2,n}(x - x_n)] 1_{\{x_n \leq x \leq d_n\}} \\ &=: \Delta_{1,n}(x) + \Delta_{2,n}(x) \end{aligned}$$

where $\delta_n \to \delta \in [0, 1], c_n \to c_o \in [-\infty, B], x_n \to c_o + \delta, d_n \to d_o \in [A, \infty], \beta_{1,n} \to \beta_1$, and $\beta_{2,n} \to \beta_2$. Define the function R_2 as follows, again using (3.25),

$$\begin{aligned} &\min\{1, \delta_n\}^{1/2} \frac{W(\Delta_{1,n})}{\sigma(\Delta_{1,n})} \\ &\leq \frac{\min\{1, \delta_n\}^{1/2} \|\Delta_{1,n}/\varphi\|_\infty^{\mathbb{R}}}{\left(\int_{\mathbb{R}} \Delta_{1,n}^2(x) f(x) \, \mathrm{d}x \right)^{1/2}} \\ &\leq \frac{G \min\{1, \delta_n\}^{1/2} \sup_{x \in [c_n, x_n]} [(1 + \beta_{1,n}(x - x_n)) / \max\{1, -a_o + b_o|x|\}]}{\left((x_n - c_n) + \beta_{1,n}(x_n - c_n)^2 + \beta_{1,n}^2 (x_n - c_n)^3/3) \right)^{1/2}} \\ &=: R_2(f, T, \delta_n, \beta_{1,n}, c_n, x_n). \end{aligned}$$

Again, $R_2(f, \beta_1, c_o, x_o)$ is continuous in its latter four arguments. The first case to look at is the following: $c_n \to c_o, x_n \to c_o$ (immediately implying $\delta_n \to 0$) and

$\beta_{1,n} \to \infty$. But then $R_2(f, T, \delta_n, \beta_{1,n}, c_n, x_n)$ is not bigger than

$$\begin{aligned}
&\frac{G\delta_n^{1/2} \sup_{x\in[c_n,x_n]}[(1+\beta_{1,n}\delta_n)/\max\{1, -a_o + b_o|x|\}]}{\Big((x_n - c_n) + \beta_{1,n}(x_n - c_n)^2 + \beta_{1,n}^2(x_n - c_n)^3/3)\Big)^{1/2}} \\
&\leq \frac{G\delta_n^{1/2} + G\delta_n^{3/2}\beta_{1,n}}{(\delta_n + \beta_{1,n}\delta_n^2 + \beta_{1,n}^2\delta_n^3/3)^{1/2}} \\
&= \frac{G}{(1 + \beta_{1,n}\delta_n + \beta_{1,n}^2\delta_n^2/3)^{1/2}} + \frac{G}{(\beta_{1,n}^{-2}\delta_n^{-2} + \beta_{1,n}^{-1}\delta_n^{-1} + 1/3)^{1/2}} \\
&\leq G
\end{aligned}$$

as $n \to \infty$. If on the other hand $\beta_{1,n} \to 0$, then

$$\begin{aligned}
R_2(f, T, \delta_n, \beta_{1,n}, c_n, x_n) &= \frac{G\delta_n^{1/2}(1 + o(1))}{\delta_n^{1/2}(1 + \beta_{1,n}\delta_n + \beta_{1,n}^2\delta_n^2/3)^{1/2}} \\
&= G
\end{aligned}$$

as $n \to \infty$. Finally, if $\beta_{1,n} \to \beta_1 \in (0, \infty)$, then $R_2(f, T, \delta_n, \beta_{1,n}, c_n, x_n) = G$. Similarly one can deal with the settings $c_n \to -\infty, x_n \to x_o$ and $c_n \to c_o, x_n \to \infty$, both these cases analyzed as above regarding the behavior of the sequence $\beta_{1,n}$. All this cases together yield that the function $R_2(f, T, \delta_n, \beta_{1,n}, c_n, x_n)$ is either bounded by a constant only depending on f and T or going to 0 as $n \to \infty$, contradicting (3.26). As in Case 1b this implies that

$$\min\{1, \delta_n\}^{1/2} \frac{W(\Delta_{1,n})}{\sigma(\Delta_{1,n})}$$

is bounded by a constant only depending on f and T. As a consequence we get

$$W(\Delta_{1,n}) \leq G \max\{1, \delta_n\}^{-1/2} \sigma(\Delta_{1,n}).$$

Similar considerations apply to $\Delta_{2,n}$. Noting that

$$\begin{aligned}
W(\Delta_n) &= \max\{W(\Delta_{1,n}), W(\Delta_{2,n})\} \\
&\leq G \max\{1, \delta_n\}^{-1/2} \max\{\sigma(\Delta_{1,n}), \sigma(\Delta_{2,n})\} \\
&\leq G \max\{1, \delta_n\}^{-1/2} \sigma(\Delta_n)
\end{aligned}$$

verifies (3.22). □

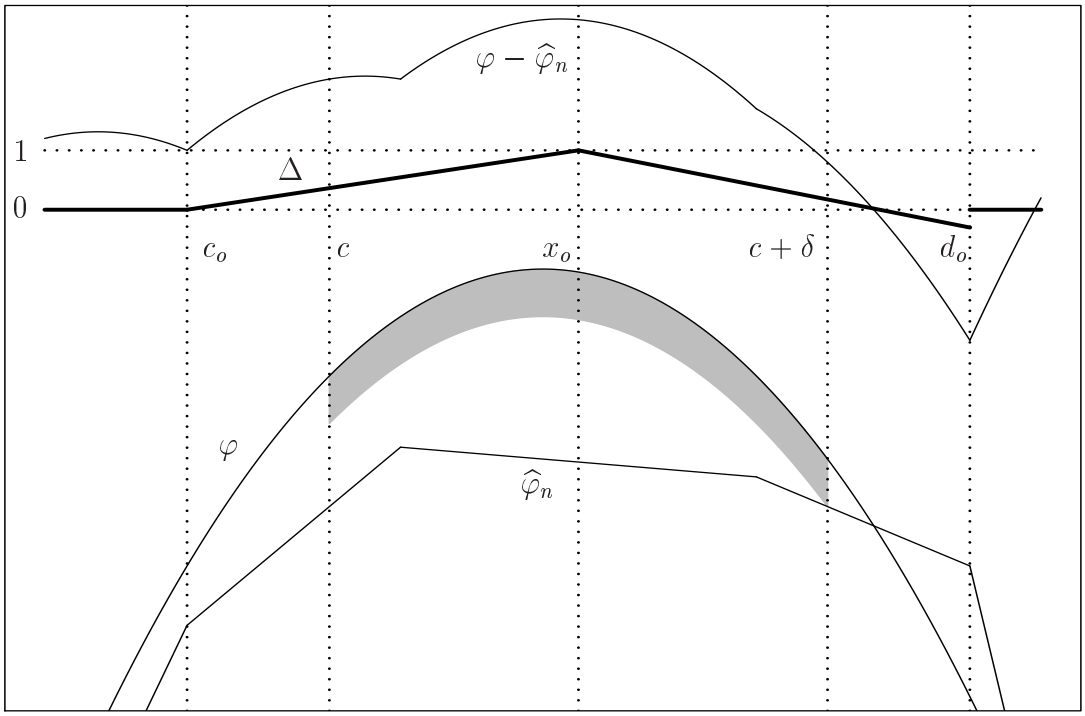

Figure 3.5: The perturbation function Δ in Case 2.

In order to prove Lemma 3.7.4 we derive first an auxiliary inequality for the moment generating function of a random variable with bounded exponential moment.

Lemma 3.7.5. *Let Y be a random variable such that $\mathbb{E}(Y) = 0$, $\mathbb{E}(Y^2) = \sigma^2$ and $\mathbb{E}\exp(|Y|) \leq 1 + C$. Then for arbitrary $\lambda \in (0,1)$ and $t \in \mathbb{R}$,*

$$\mathbb{E}\exp(tY) \;\leq\; 1 + \frac{\sigma^2 t^2}{2} + \frac{\sigma^{2\lambda} C^{1-\lambda} e^{1-\lambda} |t|^3}{(1-\lambda)^2 (1-\lambda-|t|)_+}.$$

This entails the following result for finite families of functions:

Lemma 3.7.6. *Let $\mathcal{H}_n$ be a finite family of functions h with $0 < W(h) < \infty$ such that $\#\mathcal{H}_n = O(n^p)$ for some $p > 0$. Then for any fixed $\lambda \in [0,1)$, $\kappa := (2-\lambda)/(3-2\lambda) \in [2/3, 1)$ and sufficiently large D,*

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{h \in \mathcal{H}_n} \frac{\left| \int h \, \mathrm{d}(\mathbb{F}_n - F) \right|}{\sigma(h)\rho_n^{1/2} + W(h)\rho_n^{\kappa}} \geq D \right) \;=\; 0.$$

Proof of Lemma 3.7.4. At first we consider the family $\mathcal{H}$ of all functions h of the form

$$h(x) \;=\; 1_{\{x\in J\}}(a+bx)$$

with any interval $J \subset \mathbb{R}$ and real constants a, b such that h is nonnegative. Given this family $\mathcal{H}$ we show now that for each $\kappa \in [2/3, 1)$ there exists a constant $C = C(\delta, f)$ such that

$$\sup_{h\in\mathcal{H}} \frac{\left|\int h \,\mathrm{d}(\mathbb{F}_n - F)\right|}{\sigma(h)\rho_n^{1/2} + W(h)\rho_n^{\kappa}} \;\le\; C \tag{3.27}$$

with probability tending to one as $n \to \infty$. Again, since $W(\cdot)$ and $\sigma(\cdot)$ are seminorms, we may replace $\mathcal{H}$ with the subfamily $\mathcal{H}_o$ of all functions $h \in \mathcal{H}$ such that $W(h) = 1$.

Now we use a bracketing argument. Let

$$-\infty = t_{n,0} < t_{n,1} < \cdots < t_{n,m(n)} = \infty,$$

and define $I_{n,j} := (t_{n,j-1}, t_{n,j}]$ for $1 \le j \le m(n)$. Here the points $t_{n,j}$ are chosen such that

$$\int_{t_{n,j-1}}^{t_{n,j}} \varphi(x)^2 f(x)\,\mathrm{d}x \;\le\; n^{-1}$$

with equality for $j = 1$ and $j = m(n)$. According to Lemma 2.2.1, the integral of $\varphi^2 f$ is finite. Thus we may and do assume that $m(n) = O(n)$. Moreover the last two inequalities in Lemma 2.2.1 imply that

$$\max_{1\le j<m(n)} |\varphi(t_{n,j})| \;=\; O(\log n).$$

For any $h \in \mathcal{H}_o$ we define functions $h_{n,\ell}, h_{n,u}$ as follows: Let $\{j, \ldots, k\}$ be the set of all indices $i \in \{1, \ldots, m(n)\}$ such that $\{h > 0\} \cap I_{n,i} \ne \emptyset$. Then we define

$$h_{n,\ell}(x) \;:=\; 1_{\{t_{n,j}<x\le t_{n,k-1}\}}(c+dx)$$

and

$$\begin{aligned} h_{n,u}(x) \;:=\;& 1_{\{x\in I_{n,j}\cup I_{n,k}\}}(1+|x|) \\ &+ 1_{\{t_{n,j}<x\le t_{n,k-1}\}} \min(c+dx+n^{-1/2}, 1+|\varphi(x)|), \end{aligned}$$

where $c, d \in \{zn^{-1/2} : z = 0, 1, 2, \ldots\}$ are chosen as large as possible such that $h_{n,\ell} \le h$. Figure 3.6 illustrates the situation.

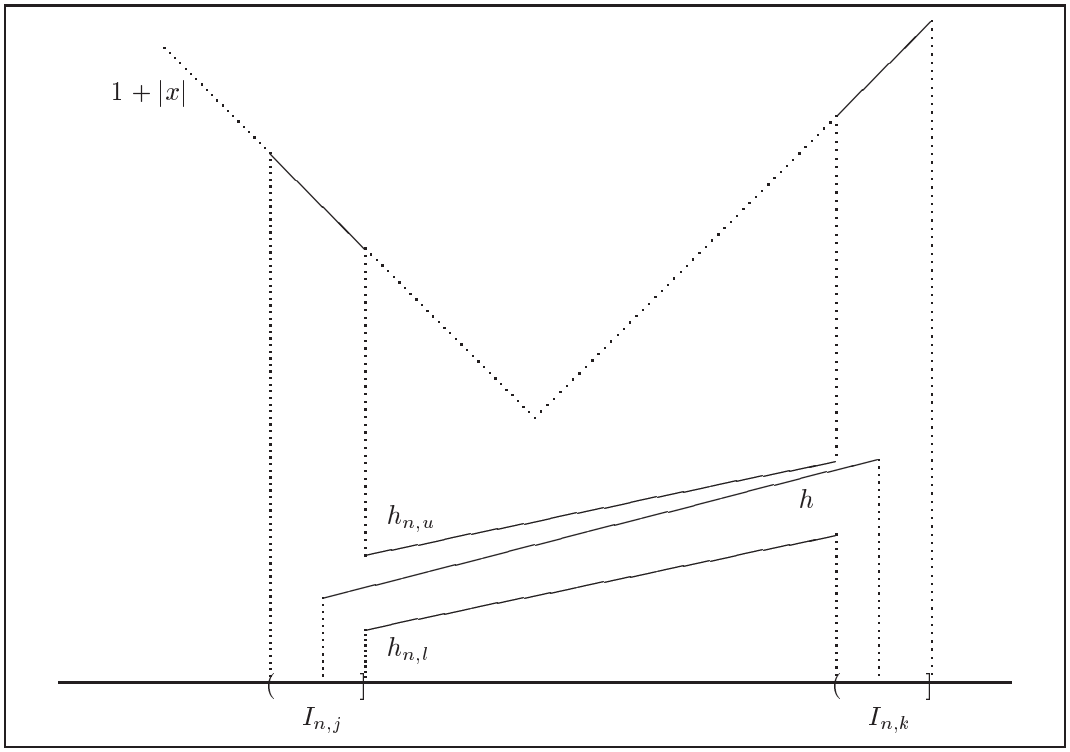

Figure 3.6: Construction of the brackets for h.

One easily verifies that $0 \le h_{n,\ell} \le h \le h_{n,u}$, $W(h_{n,u}) = 1$ and

$$\sigma(h_{n,u} - h_{n,\ell})^2 \;\le\; 3n^{-1}.$$

Moreover, the set $\mathcal{H}_n := \{h_{n,\ell}, h_{n,u} : h \in \mathcal{H}_o\}$ consists of $O(m(n)^2 n \log(n)^2) = o(n^4)$ different functions. For there are less than $m(n)^2$ possibilities for the index pair (j,k) and at most $(n^{1/2} \max_j |\varphi(t_{n,j})| + 1)^2$ possibilities for the pair (c,d).

It follows from Lemma 3.7.6 that for some suitable constant $D = D(\kappa, f)$,

$$\sup_{h \in \mathcal{H}_n} \frac{\left|\int h \,\mathrm{d}(\mathbb{F}_n - F)\right|}{\sigma(h)\rho_n^{1/2} + \rho_n^{\kappa}} \;\le\; D \tag{3.28}$$

with probability tending to one as $n \to \infty$. But for any $h \in \mathcal{H}_o$ the inequality (3.28)

entails that

$$\begin{aligned}
\int h \,\mathrm{d}(\mathbb{F}_n - F) &\leq \int h_{n,u} \,\mathrm{d}\mathbb{F}_n - \int h_{n,\ell} \,\mathrm{d}F \\
&= \int h_{n,u} \,\mathrm{d}(\mathbb{F}_n - F) + \int (h_{n,u} - h_{n,\ell}) \,\mathrm{d}F \\
&\leq D\Big(\sigma(h_{n,u})\rho_n^{1/2} + \rho_n^{\kappa}\Big) + 3^{1/2} n^{-1/2} \\
&\leq D\Big((\sigma(h) + 3^{1/2} n^{-1/2})\rho_n^{1/2} + \rho_n^{\kappa}\Big) + 3^{1/2} n^{-1/2} \\
&\leq (D+1)(\sigma(h)\rho_n^{1/2} + \rho_n^{\kappa})
\end{aligned}$$

for sufficiently large n. Thus we may take $C = D + 1$ in (3.27). In order to complete the proof of Lemma 3.7.4, consider any $\Delta \in \mathcal{D}_m$. There are $m' \leq 2m + 2$ disjoint intervals on which Δ is linear and either nonnegative or nonpositive. Thus we may write

$$\Delta = \sum_{j=1}^{m'} \lambda_j h_j$$

with functions $h_j \in \mathcal{H}$ having disjoint support and numbers $\lambda_j \in \{-1, 1\}$. Consequently,

$$\begin{aligned}
\sigma(\Delta)^2 &= \sum_{j=1}^{m'} \sigma(h_j)^2, \\
W(\Delta) &= \max_{j=1,\ldots,m'} W(h_j).
\end{aligned}$$

Thus (3.28), together with the Cauchy-Schwarz inequality, entails that

$$\begin{aligned}
\Big|\int \Delta \,\mathrm{d}(\mathbb{F}_n - F)\Big| &\leq \sum_{j=1}^{m'} \Big|\int h_j \,\mathrm{d}(\mathbb{F}_n - F)\Big| \\
&\leq \sum_{j=1}^{m'} C\Big(\sigma(h_j)\rho_n^{1/2} + W(h_j)\rho_n^{\kappa}\Big) \\
&\leq C\Big(\sum_{j=1}^{m'} \sigma(h_j)\rho_n^{1/2} + W(\Delta)m'\rho_n^{\kappa}\Big) \\
&\leq C\Big(\sigma(\Delta){m'}^{1/2}\rho_n^{1/2} + W(\Delta)m'\rho_n^{\kappa}\Big) \\
&\leq 4C\Big(\sigma(\Delta)m^{1/2}\rho_n^{1/2} + W(\Delta)m\rho_n^{\kappa}\Big)
\end{aligned}$$

what concludes the proof. □

Proof of Lemma 3.7.5. Note first that

$$\mathbb{E}\exp(tY) \;=\; \sum_{k=0}^{\infty} \frac{t^k}{k!}\,\mathbb{E}(Y^k) \;\le\; 1 + \frac{\sigma^2 t^2}{2} + \sum_{k=3}^{\infty} \frac{|t|^k}{k!}\,\mathbb{E}(|Y|^k).$$

It follows from Hölder's inequality that

$$\begin{aligned}
\mathbb{E}(|Y|^k) &= \mathbb{E}(|Y|^{\alpha}|Y|^{k-\alpha}) \qquad (\text{for } 0<\alpha<k)\\
&\le \mathbb{E}(|Y|^{\alpha/\lambda})^{\lambda}\,\mathbb{E}(|Y|^{(k-\alpha)/(1-\lambda)})^{1-\lambda}\\
&= \sigma^{2\lambda}\,\mathbb{E}(|Y|^{(k-2\lambda)/(1-\lambda)})^{1-\lambda} \qquad (\text{if } \alpha = 2\lambda).
\end{aligned}$$

Moreover, for $\ell \ge 1$,

$$\mathbb{E}(|Y|^{\ell}) \;=\; \mathbb{E}\Big((\exp(|Y|)-1)\,\frac{|Y|^{\ell}}{\exp(|Y|)-1}\Big) \;\le\; C\,\max_{y>0}\,\frac{y^{\ell}}{e^y-1} \;\le\; C\,\ell^{\ell}e^{1-\ell}.$$

For

$$\frac{d}{dy}\,\frac{y^{\ell}}{e^y-1} \;=\; \frac{\ell y^{\ell-1}(e^y-1-ye^y/\ell)}{(e^y-1)^2}$$

is strictly positive on $(0,z)$ and strictly negative on (z,∞), where z satisfies the equality $e^z-1 = ze^z/\ell$. Hence the maximum of $y^{\ell}/(e^y-1)$ over all $y>0$ is not greater than the maximum of $\ell z^{\ell-1}e^{-z}$ over all $z>0$, and the latter maximum equals $\ell(\ell-1)^{\ell-1}e^{1-\ell} \le \ell^{\ell}e^{1-\ell}$. Consequently,

$$\begin{aligned}
\mathbb{E}\exp(tY) &\le 1 + \frac{\sigma^2t^2}{2} + \sigma^{2\lambda}C^{1-\lambda}e^{1-\lambda}\sum_{k=3}^{\infty}\frac{|t|^k}{k!}\Big(\frac{k-2\lambda}{1-\lambda}\Big)^{k-2\lambda}e^{k-2\lambda}\\
&= 1 + \frac{\sigma^2t^2}{2} + \sigma^{2\lambda}C^{1-\lambda}e^{1+\lambda}\sum_{k=3}^{\infty}\frac{|t|^k}{k!}\Big(\frac{k-2\lambda}{1-\lambda}\Big)^{k-2\lambda}e^{-k}\\
&\le 1 + \frac{\sigma^2t^2}{2} + \sigma^{2\lambda}C^{1-\lambda}e^{1+\lambda}\sum_{k=3}^{\infty}\frac{|t|^k}{k!}\,3^{-2\lambda}\Big(\frac{k}{1-\lambda}\Big)^{k}e^{-k}\\
&< 1 + \frac{\sigma^2t^2}{2} + \sigma^{2\lambda}C^{1-\lambda}e^{1-\lambda}\sum_{k=3}^{\infty}\Big(\frac{|t|}{1-\lambda}\Big)^{k}\,\frac{k^ke^{-k}}{k!}\\
&\le 1 + \frac{\sigma^2t^2}{2} + \sigma^{2\lambda}C^{1-\lambda}e^{1-\lambda}\sum_{k=3}^{\infty}\Big(\frac{|t|}{1-\lambda}\Big)^{k}\\
&= 1 + \frac{\sigma^2t^2}{2} + \frac{\sigma^{2\lambda}C^{1-\lambda}e^{1-\lambda}|t|^3}{(1-\lambda)^2(1-\lambda-|t|)}.
\end{aligned}$$

□

Proof of Lemma 3.7.6. Since $W(ch) = cW(h)$ and $\sigma(ch) = c\sigma(h)$ for any $h \in \mathcal{H}_n$ and arbitrary constants $c > 0$, we may assume without loss of generality that $W(h) = 1$ for all $h \in \mathcal{H}_n$. Note that now $|h(x)| \le |\varphi(x)|$. Hence it follows from Lemma 2.2.1 that

$$\mathbb{E}\exp\Bigl(t_o|h(X) - \mathbb{E}\,h(X)|\Bigr) \;\le\; C_o \;:=\; \exp(t_o\,\mathbb{E}\,|\varphi(X)|)\,\mathbb{E}\exp(t_o|\varphi(X)|),$$

which is finite for $0 < t_o < 1$. Thus Lemma 3.7.5, applied to $Y := t_o(h(X) - \mathbb{E}\,h(X))$, implies that

$$\mathbb{E}\exp\Bigl[t\Bigl(h(X) - \mathbb{E}\,h(X)\Bigr)\Bigr] \;=\; \mathbb{E}\Bigl((t/t_o)Y\Bigr) \;\le\; 1 + \frac{\sigma(h)^2t^2}{2} + \frac{C_1\sigma(h)^{2\lambda}|t|^3}{(1 - C_2|t|)_+}$$

for arbitrary $h \in \mathcal{H}_n$, $t \in \mathbb{R}$ and constants C_1, C_2 depending on λ, t_o, C_o. Consequently,

$$\begin{aligned}
\mathbb{E}\exp\Bigl(t\int h\,\mathrm{d}(\mathbb{F}_n - F)\Bigr) &= \mathbb{E}\exp\Bigl((t/n)\sum_{i=1}^n (h(X_i) - \mathbb{E}\,h(X))\Bigr) \\
&= \Bigl(\mathbb{E}\exp\Bigl((t/n)(h(X) - \mathbb{E}\,h(X))\Bigr)\Bigr)^n \\
&\le \left(1 + \frac{\sigma(h)^2t^2}{2n^2} + \frac{C_1\sigma(h)^{2\lambda}|t|^3}{n^3(1 - C_2|t|/n)_+}\right)^n \\
&\le \exp\left(\frac{\sigma(h)^2t^2}{2n} + \frac{C_1\sigma(h)^{2\lambda}|t|^3}{n^2(1 - C_2|t|/n)_+}\right).
\end{aligned}$$

Now it follows from Markov's inequality that

$$\mathbb{P}\Bigl(\Bigl|\int h\,\mathrm{d}(\mathbb{F}_n - F)\Bigr| \ge \eta\Bigr) \;\le\; 2\exp\left(\frac{\sigma(h)^2t^2}{2n} + \frac{C_1\sigma(h)^{2\lambda}t^3}{n^2(1 - C_2t/n)_+} - t\eta\right) \tag{3.29}$$

for arbitrary $t, \eta > 0$. Specifically let $\eta = D(\sigma(h)\rho_n^{1/2} + \rho_n^\kappa)$ and set

$$t \;:=\; \frac{n\rho_n^{1/2}}{\sigma(h) + \rho_n^{\kappa-1/2}} \;\le\; n\rho_n^{1-\kappa} \;=\; o(n).$$

Then the bound (3.29) is not greater than

$$\begin{aligned}
&2\exp\left(\frac{\sigma(h)^2\log n}{2(\sigma(h) + \rho_n^{\kappa-1/2})^2} + \frac{C_1\sigma(h)^{2\lambda}\rho_n^{1/2}\log n}{(\sigma(h) + \rho_n^{\kappa-1/2})^3(1 - C_2\rho_n^{1-\kappa})_+} - D\log n\right) \\
&\le\; 2\exp\Bigl[\Bigl(\frac{1}{2} + \frac{C_1}{(1 - C_2\rho_n^{1-\kappa})_+} - D\Bigr)\log n\Bigr] \;=\; 2\exp\Bigl((O(1) - D)\log n\Bigr).
\end{aligned}$$

Consequently,

$$\mathbb{P}\left(\max_{h\in\mathcal{H}_n}\frac{\left|\int h\,\mathrm{d}(\mathbb{F}_n-F)\right|}{\sigma(h)\rho_n^{1/2}+W(h)\rho_n^{\kappa}}\geq D\right)$$
$$\leq \#\mathcal{H}_n 2\exp\Big((O(1)-D)\log n\Big) = O(1)\exp\Big((O(1)+p-D)\log n\Big) \to 0$$

as $n\to\infty$, provided that D is sufficiently large. □

Proof of Lemma 3.7.2: Define the linear approximation to g at t_o for $t\in T$ as:

$$\widetilde{g}(t) := \begin{cases} g(t_o) & \text{if } \beta=1,\\ g(t_o)+g'(t_o)(t-t_o) & \text{if } \beta>1.\end{cases}$$

The assumption that $g\in\mathcal{H}^{\beta,L}(T)$ then implies for $\beta=1$

$$|(\widetilde{g}-g)(t)| = |g(t_o)-g(t)|\leq L|t-t_o| \tag{3.30}$$

and for $\beta>1$

$$\begin{aligned}|(\widetilde{g}-g)(t)| &= |g(t_o)-g(t)+g'(t_o)(t-t_o)|\\ &\leq \int_t^{t_o}|g'(u)-g'(t_o)|\,\mathrm{d}u\\ &\leq L\int_t^{t_o}|u-t_o|^{\beta-1}\,\mathrm{d}u\\ &\leq (L/\beta)|t-t_o|^{\beta}.\end{aligned} \tag{3.31}$$

Case 1: Suppose that one has $(\widehat{g}-g)(t_o)\geq\varepsilon$ for a $t_o\in T$ such that, without loss of generality, $t_o\leq(A+B)/2$. Let $0<\delta\leq(B-A)/8$.

Case 1a: Assume that $(\widehat{g}-\widetilde{g})(t_o+\delta)\geq\varepsilon/2$. Since $\widehat{g}-\widetilde{g}$ is concave with $(\widehat{g}-\widetilde{g})(t_o)=(\widehat{g}-g)(t_o)\geq\varepsilon$, it follows that $(\widehat{g}-\widetilde{g})(t)\geq\varepsilon/2$ for all $t\in[t_o,t_o+\delta]$.

Case 1b: On the other hand, let $(\widehat{g}-\widetilde{g})(t_o+\delta)\leq\varepsilon/2$. The slope of $\widehat{g}-\widetilde{g}$ right of $t_o+\delta$ is then at most that of the line through $(\widehat{g}-\widetilde{g})(t_o)$ and $(\widehat{g}-\widetilde{g})(t_o+\delta)$, namely $-\varepsilon/(2\delta)$. This means that $(\widehat{g}-\widetilde{g})(t)\leq-\varepsilon/2$ if only $t\geq t_o+3\delta$.
Summarizing Cases 1a and 1b, we learn that there exists an interval $J\subset[t_o,t_o+4\delta]$ of length δ such that for all $t\in J$ we have $|\widetilde{g}-\widehat{g}|\geq\varepsilon/2$. By the triangle inequality

we get $\varepsilon/2 \leq |\widehat{g} - g| + |g - \widetilde{g}|$. Using Inequalities (3.30) and (3.31) this finally entails that

$$|\widehat{g} - g| \;\; \geq \;\; \varepsilon/2 - (L/\beta)(4\delta)^{\beta}.$$

The expression on the right is at least $\varepsilon/4$ if

$$\delta \leq (\beta/L)^{1/\beta} 4^{-1-1/\beta} \varepsilon^{1/\beta} =: K_1(\beta, L)\varepsilon^{1/\beta}.$$

Case 2: Now assume $(g - \widehat{g})(t_o) \geq \varepsilon$ for a $t_o \in [A+\delta, B-\delta]$ where $\delta \in (0, (B-A)/2]$. Thus, from (3.30) or (3.31) it follows the existence of ν_1 such that

$$g(t) - g(t_o) \geq \nu_1(t - t_o) - (L/\beta)|t - t_o|^{\beta}$$

and from the concavity of $\widehat{g}$ that of ν_2 with $\widehat{g}(t) - \widehat{g}(t_o) \leq \nu_2(t - t_o)$. Together this yields

$$(g - \widehat{g})(t) \;\; \geq \;\; \varepsilon + (\nu_1 - \nu_2)(t - t_o) - (L/\beta)|t - t_o|^{\beta} \geq \varepsilon - L\delta^{\beta}$$

for all t either in $[t_o, t_o+\delta]$ or $[t_o-\delta, t_o]$, depending on $\operatorname{sign}(\nu_1-\nu_2)$. Finally, $\varepsilon - L\delta^{\beta} \geq \varepsilon/4$ if $\delta \leq (3\varepsilon/(4L))^{1/\beta} =: K_2(\beta, L)\varepsilon^{1/\beta}$. Note that $K_1(\beta, L) \leq K_2(\beta, L)$ uniformly in β and L, so that we define $K(\beta, L) := \min\{K_1(\beta, L), K_2(\beta, L)\} = K_1(\beta, L)$. □
With the verification of this last lemma the proof of Theorem 3.3.1 is complete. □
Before coming to the proofs for $\widehat{F}_n$, we still owe that for Corollary 3.3.2.

Proof of Corollary 3.3.2: First, note that the statements are trivial outside $[X_1, X_n]$, by Theorem 3.2.1. The concave function $\varphi : (a, b) \to \mathbb{R}$ is automatically Lipschitz-continuous on any compact subinterval $[c, d]$ of (a, b), because

$$\frac{\varphi(d) - \varphi(c)}{d - c}$$

is, due to concavity of φ, uniformly bounded for any c, d. This fact, together with Theorem 3.3.1 and continuity of f entails uniform consistency of $\widehat{f}_n$. For the integrated density estimator $\widehat{F}_n$, write $|\widehat{F}_n - F| \leq \int |f - \widehat{f}_n|$ as

$$\int (\widehat{f}_n - f)_+ + \int (\widehat{f}_n - f)_- = 2 \int (f - \widehat{f}_n)_+ - \Big(\int f - \int \widehat{f}_n \Big).$$

On the right-hand side, the first term tends to zero by dominated convergence (Theorem A.1.1) applied to $f \geq (f - \widehat{f}_n)_+ \to_{\mathrm{p}} 0$. The second term equals zero. Actually, this is solely an application of what is known as Scheffé's Theorem. □

The gap problem

Proof of Theorem 3.4.1. To simplify things introduce a new coordinate system with origin $(s_{i-1}, \varphi(s_{i-1}))$. Suppose that for $\delta = \Delta s_i$ and $\varepsilon = K\rho_n^{\beta/(2\beta+1)}$ we have:

$$\varphi(\delta/2) - \varphi(\delta)/2 \leq 2\varepsilon. \tag{3.32}$$

Then the assumption about φ' together with (3.8) yields:

$$\begin{aligned} 2\varepsilon &\geq \varphi(\delta/2) - \varphi(\delta)/2 \\ &\geq 2^{-1}\Big(\int_0^{\delta/2} \varphi'(u)\,\mathrm{d}u - \int_{\delta/2}^{\delta} \varphi'(u)\,\mathrm{d}u\Big) \\ &= 2^{-1}\Big[\int_0^{\delta/2} \Big(\varphi'(u) - \varphi'(u+\delta/2)\Big)\,\mathrm{d}u\Big] \\ &\geq C(\delta^2/8). \end{aligned}$$

So we can conclude:

$$\delta \leq 2C^{-1/2}\varepsilon^{1/2}.$$

To prove assertion (3.32) recapitulate from Theorem 3.3.1 that

$$|(\varphi - \widehat{\varphi}_n)(x)| \leq \varepsilon. \tag{3.33}$$

Introduce for $x \in [0,\delta]$ the auxiliary functions $\iota(x) := (\varphi(\delta)/\delta)x$ and a parallelwise translated $\kappa(x)$: Define x_o as the left-most point in $[0,\delta]$ where $\widehat{\varphi}_n'(x) = \varphi(\delta)/\delta$ and using this $\kappa(x) := \iota(x) + (\varphi(x_o) - \iota(x_o))$. Then distinct three cases, depending on the number of intersections of φ and $\widehat{\varphi}_n$ in $(0,\delta)$.

Case 1: Let $\#\{x \in (0,\delta) : \widehat{\varphi}_n(x) = \varphi(x)\} = 2$. Then geometric considerations reveal that $(\varphi - \iota)(x)/2 \leq \varepsilon$ for $x \in [0,\delta]$ whenever (3.33) is true (and equality holds whenever $(\widehat{\varphi}_n - \iota)(x) = 2^{-1}(\kappa - \iota)(x)$ for all $x \in [0,\delta]$). Set $x = \delta/2$. For an illustration consult Figure 3.7.

Case 2: Let $\#\{x \in (0,\delta) : \widehat{\varphi}_n(x) = \varphi(x)\} = 1$. Again, $(\varphi - \iota)(x)/2 \leq \varepsilon$ for $x \in [0,\delta]$ but here $(\varphi - \iota)(x)/2 = \varepsilon$ e.g. in case $\widehat{\varphi}_n(0) = \kappa(0)$, $\widehat{\varphi}_n(\delta) = \iota(\delta)$, $\varphi'(\delta/2) = \kappa(\delta/2)$ and $x = \delta/2$. Figure 3.8 details the situation.

Case 3: Let $\#\{x \in (0,\delta) : \widehat{\varphi}_n(x) = \varphi(x)\} = 0$. In this last situation, we have w.l.o.g. that $(\kappa - \iota)(x) \leq \varepsilon$ for all $x \in [0,\delta/2]$ (otherwise mirror the situation) with equality whenever $\widehat{\varphi}_n(x) = \kappa(x)$ for all $x \in [0,\delta]$. This entails that $(\varphi - \iota)(x) \leq \varepsilon$ and with $x = \delta/2$ we get the assertion. □

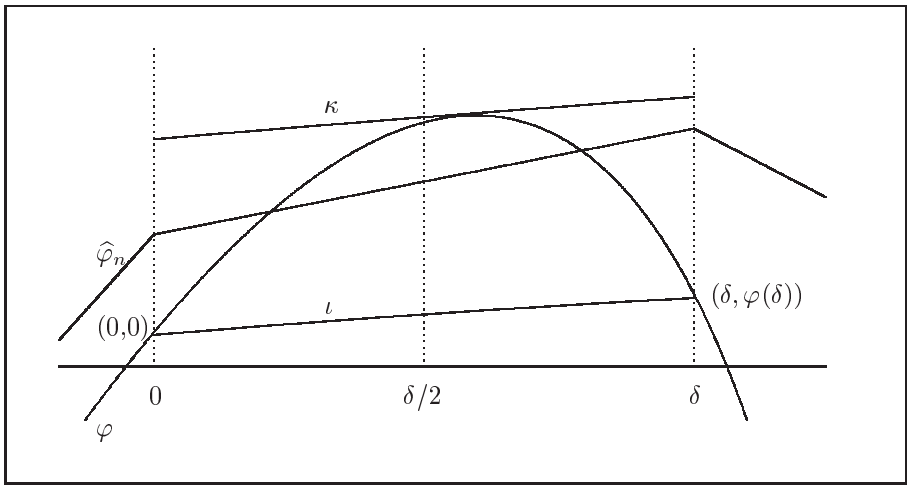

Figure 3.7: Illustration of Case 1 in the proof of Theorem 3.4.1.

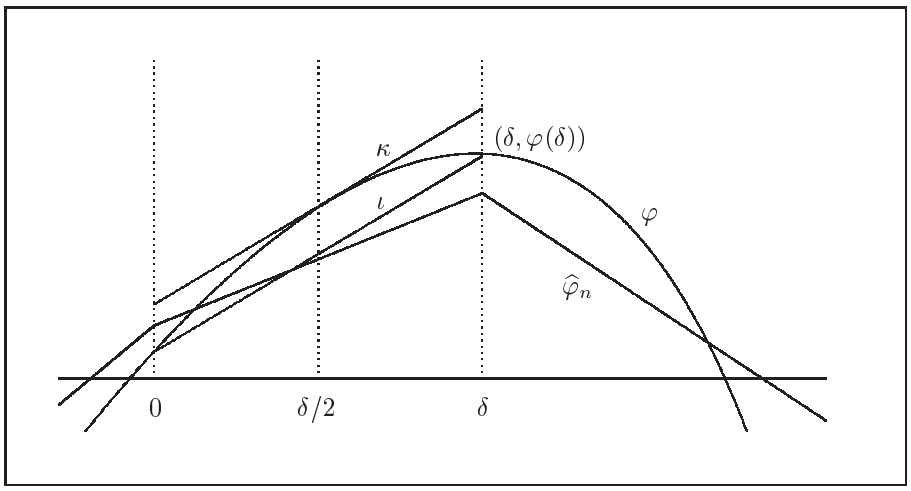

Figure 3.8: Illustration of Case 2.

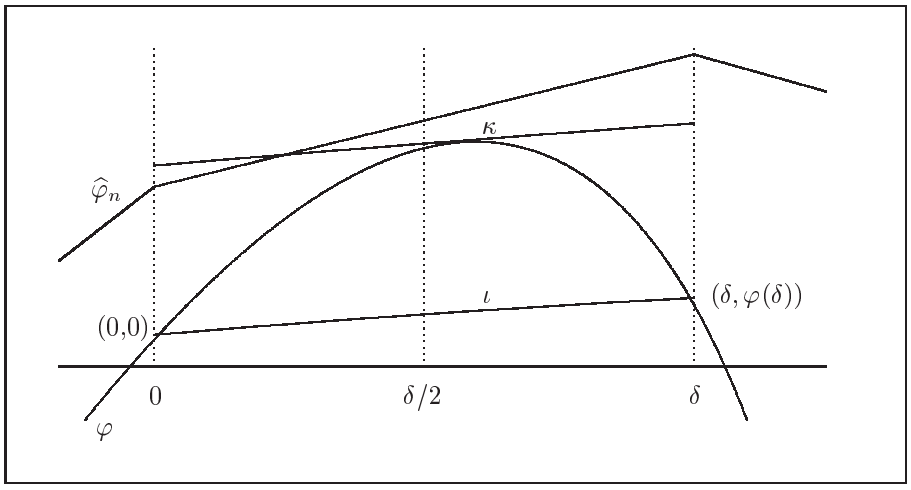

Figure 3.9: Illustration of Case 3.

Uniform consistency of $\widehat{F}_n$

Proof of Theorem 3.5.1: To simplify notation introduce:

$$\begin{aligned} r_n &:= \rho_n^{\beta/(4\beta+2)} \\ \omega(g,d) &:= \sup_{x\in[A+r_n,B-r_n-d]} \sup_{|h|\le d} |g(x+h)-g(x)| \end{aligned}$$

for $d>0$ and functions g bounded on $[A,B]$. The uniform empirical distribution function is denoted by $\mathbb{G}_n$ and $\mathbb{U}_n$ stands for a uniform empirical process. Then note that

$$\omega(F, r_n) \le r_n \|f\|_\infty^{\mathbb{R}}.$$

Consequently, together with

$$\omega(\mathbb{U}_n, r_n) = \sqrt{2r_n \log(r_n^{-1})}(1+o_{\mathrm{p}}(1))$$

guaranteed by Theorem A.2.1, we have (let $\mathrm{id}(x) := x$)

$$\begin{aligned} \omega(\mathbb{F}_n - F, r_n) &=_{\mathcal{D}} \omega\Big(\mathbb{G}_n(F) - \mathrm{id}, \omega(F, r_n)\Big) \\ &= n^{-1/2}\|f\|_\infty^{\mathbb{R}}\omega(\mathbb{U}_n, r_n) \\ &= n^{-1/2} O_{\mathrm{p}}\Big(\sqrt{(\log n)^{(5\beta+2)/(4\beta+2)} n^{-\beta/(4\beta+2)}}\Big) \\ &= o_{\mathrm{p}}(n^{-1/2}). \end{aligned}$$

The conditions on r_n imposed in the theorem are clearly fulfilled since

$$\begin{aligned} nr_n &= n^{(3\beta+2)/(4\beta+2)}(\log n)^{\beta/(4\beta+2)} \to \infty \\ \log(r_n^{-1})/\log\log n &= \Big((\log n)^{\beta/(4\beta+2)} - (\log\log n)^{\beta/(4\beta+2)}\Big)/\log\log n \to \infty \\ \log(r_n^{-1})/(nr_n) &= (1-o(1))n^{-(3\beta+2)/(4\beta+2)} \to 0. \end{aligned}$$

Together with Lemma 3.7.1 and Theorems 3.4.1 and 3.3.1 we have:

$$\begin{aligned}
&\sup_{x\in[A+r_n,B-r_n]} |(\widehat{F}_n - \mathbb{F}_n)(x)| \\
&\leq \sup_{i=2,\ldots,M} \sup_{x\in(s_{i-1},s_i]} \Big(|(\widehat{F}_n - F)(x) + (F-\mathbb{F}_n)(x) - (\widehat{F}_n - F)(s_{i-1}) - \\
&\qquad\qquad (F-\mathbb{F}_n)(s_{i-1})| + |(\widehat{F}_n - \mathbb{F}_n)(s_{i-1})| \Big) \\
&\leq \sup_{i=2,\ldots,M} \sup_{x\in(s_{i-1},s_i]} \Big(|\int_{s_{i-1}}^{x} (\widehat{f}_n - f)(t)\,\mathrm{d}t| + |(F-\mathbb{F}_n)(x) \\
&\qquad\qquad -(F-\mathbb{F}_n)(s_{i-1})| \Big) + n^{-1} \\
&\leq O_{\mathrm{p}}(r_n) \sup_{i=2,\ldots,M} \Big(\sup_{x\in(s_{i-1},s_i]} |(\widehat{f}_n - f)(x)| \Big) + \omega(F-\mathbb{F}_n, r_n) + n^{-1} \\
&= O_{\mathrm{p}}\Big(\rho_n^{3\beta/(4\beta+2)}\Big) + o_{\mathrm{p}}(n^{-1/2}) + n^{-1} \qquad (3.34) \\
&= o_{\mathrm{p}}(n^{-1/2}). \qquad \square
\end{aligned}$$

Note that for $\beta = 1$ the exponent of the first term in (3.34) equals $1/2$, so it is the logarithmic term in the nominator together with the assumption in Theorem 3.4.1 that prevents the expression to be of probabilistic order smaller than $n^{-1/2}$. However, Corollary 3.3.2 gives at least uniform consistency also for $\beta = 1$.

Integrated kernel estimator

Proof of Theorem 3.5.4: Write for $x_o \in \mathbb{R}$

$$\begin{aligned}
\widehat{F}_{n,h}(x_o) &= \Big(\widehat{F}_h(x_o) - \mathbb{E}\,\widehat{F}_h(X_1)\Big) + \Big(\mathbb{E}\,\widehat{F}_h(X_1) - F(x_o)\Big) + F(x_o) \\
&:= T_1(x_o) + T_2(x_o) + F(x_o).
\end{aligned}$$

Introduce a random variable Z independent of $X_1, \ldots, X_n$ and having density function k. Then:

$$\begin{aligned}
T_1(x_o) &= \frac{1}{n}\sum_{i=1}^{n}\Big[K\Big(\frac{x_o - X_i}{h}\Big) - \mathbb{E}\,K\Big(\frac{x_o - X_1}{h}\Big)\Big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big[P\Big(Z \le \frac{x_o - X_i}{h}|X_1, \ldots, X_n\Big) - P\Big(Z \le \frac{x_o - X_1}{h}\Big)\Big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(P(X_i \le x_o - hZ|X_1, \ldots, X_n) - P(X_1 \le x_o - hZ)\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int\Big(P(X_i \le x_o - hz|X_1, \ldots, X_n) - P(X_1 \le x_o - hz)\Big)k(z)\,\mathrm{d}z \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big[\int\Big(1_{\{X_i \le x_o - hz\}} - F(x_o - hz)\Big)k(z)\,\mathrm{d}z\Big] \\
&= \int\Big(\mathbb{F}_n(x_o - hz) - F(x_o - hz)\Big)k(z)\,\mathrm{d}z \\
&= O_{\mathrm{p}}(n^{-1/2})
\end{aligned}$$

by Theorem A.3.1. On the other hand for $T_2(x_o)$ one has:

$$\begin{aligned}
T_2(x_o) &= \int K\Big(\frac{x_o - y}{h}\Big)f(y)\,\mathrm{d}y - F(x_o) \\
&= \int_{\mathbb{R}}\Big(\int_{-\infty}^{(x_o - y)/h} k(t)\,\mathrm{d}t\Big)f(y)\,\mathrm{d}y - F(x_o) \\
&= \int_{\mathbb{R}}\Big(\int_{\mathbb{R}} 1_{\{y \le x_o - ht\}}k(t)\,\mathrm{d}t\Big)f(y)\,\mathrm{d}y - F(x_o) \\
&= \int_{\mathbb{R}} k(t)\Big(\int_{-\infty}^{x_o - ht} f(y)\,\mathrm{d}y\Big)\,\mathrm{d}t - F(x_o) \\
&= \int_{\mathbb{R}} k(t)\Big(F(x_o - ht) - F(x_o)\Big)\,\mathrm{d}t \\
&= h^2\int_{\mathbb{R}} k(t)\Big[(t^2/2)\Big(f'(x_o) + o(1)\Big)\Big]\,\mathrm{d}t \\
&= O(h^2 f'(x_o)).
\end{aligned}$$

It is important to note that this rate cannot be improved in the sense that the factor h^2 always appears. As a summary, choosing a bandwidth of optimal order $O(n^{-1/5})$:

$$\begin{aligned}\widehat{F}_{n,h}(x_o) &= F(x_o) + O_{\mathrm{p}}(n^{-1/2}) + O(h^2 f'(x_o)) \\ &= F(x_o) + O_{\mathrm{p}}(n^{-1/2}) + O(n^{-2/5}) \\ &= F(x_o) + O(n^{-2/5})\end{aligned}$$

as stated in the theorem. □

Proof of Theorem 3.6.1. The Theorem is in fact a corollary of Theorems 3.3.1 and 3.5.1 combined with Lemma 2.3.1. □

Chapter 4

Algorithms to find the density estimator

In this chapter, we describe several algorithms performing well in finding the log-concave density estimator $\widehat{f}_n$ of the true density f analyzed in Chapter 3. Some comparisons between the algorithms are reported.

4.1 Introduction

Suppose we want to estimate $\widehat{\varphi}_n$ introduced in Chapter 3 based on ordered observations $X_1 < X_2 < \ldots < X_n$. We show that this can be achieved through numerical minimization of the log-likelihood functional (3.2) over all concave functions φ, where we use that, according to Theorem 3.2.1, we only have to consider functions φ that are piecewise linear and have knots at some of the observation points.

The above task is typical for many estimation problems in statistics as it demands for the optimization of a (high-dimensional) objective function, the log-likelihood in our case. We show that, within a linearly constrained optimization framework, $\widehat{\varphi}_n$ and therewith the density estimator $\widehat{f}_n$ can be found.

In Walther (2002), maximum likelihood log-concave density estimation is described for the first time, in a multiscale context. He proposes the iterative convex minorant algorithm (ICMA) introduced by Groeneboom and Wellner (1992) to solve the maximization problem and considers it to perform better than interior point methods, in terms of speed and stability. We show that the interior point methods used for convex density estimation in Terlaky and Vial (1998) work in log-concave density estimation as well and give some simulation results comparing them to the ICMA and a new algorithm, recently proposed in Dümbgen, Freitag, and Jongbloed (2006).

4.2 Framework of numerical log-concave density estimation

We use the notation introduced in Chapter 3. We will estimate $\widehat{f}_n$ via its logarithm $\widehat{\varphi}_n$, i.e. we show how to find

$$\widehat{\varphi}_n \;:=\; \operatorname*{arg\,min}_{\varphi \text{ concave}} \Psi_n(\varphi).$$

According to Theorem 3.2.1 it is sufficient to know $\widehat{\varphi}_n$ only at the observation points $\boldsymbol{X} := (X_1, \ldots, X_n)$, even only at the points belonging to $\mathcal{S}(\widehat{\varphi}_n)$, the set of knots of $\widehat{\varphi}_n$. However, we have a priori no idea where the estimator $\widehat{\varphi}_n$ changes its slope. So denoting $\varphi(X_i)$ by φ_i and identifying the function φ with the vector

$$\boldsymbol{\varphi} \;:=\; (\varphi_i)_{i=1}^n,$$

we reparametrize $\boldsymbol{\varphi}$ by the successive slopes

$$\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\varphi}) := \Big(\varphi_1, \Big(\frac{\Delta\varphi_i}{\Delta X_i}\Big)_{i=2}^n\Big)$$

where $\Delta X_i := X_i - X_{i-1}$ for a vector $\boldsymbol{X} \in \mathbb{R}^n$ and $i = 2, \ldots, n$. Note that $\boldsymbol{\eta}$ is just an affine transformation of $\boldsymbol{\varphi}$, therefore not affecting the existence, uniqueness or location of the minimum of Ψ_n. In order to ensure concavity of φ, the corresponding vector $\boldsymbol{\eta} \in \mathbb{R}^n$ must belong to the cone

$$\mathcal{K}_\cap := \{\boldsymbol{\eta} \in \mathbb{R}^n : \eta_{i-1} \geq \eta_i, \quad i = 3, \ldots, n\}$$

where $\mathcal{K}_\cap$ is defined by $n-2$ inequalities. In other words, $(\eta_i)_{i=2}^n$ must be a decreasing sequence. The piecewise linearity now enables us to write the Lagrange term in (3.2) as

$$\begin{aligned} n\int \exp\varphi(x)\,\mathrm{d}x \;&=\; n\sum_{i=2}^n \int_{X_{i-1}}^{X_i} \exp\Big(\frac{\Delta\varphi_i}{\Delta X_i}(x - X_{i-1}) + \varphi_{i-1}\Big)\,\mathrm{d}x \\ &=\; ne^{\eta_1}\sum_{i=2}^n \exp\Big(\sum_{k=2}^{i-1}\Delta X_k\eta_k\Big)\frac{\exp(\Delta X_i\eta_i) - 1}{\eta_i} \end{aligned} \tag{4.1}$$

where $(\exp(0) - 1)/0$ is taken conventionally to be equal to one and $\sum_{k=i}^j q_k = 0$ if $j < i$. Note that (4.1) is now a sum rather than an integral, both depending on $\boldsymbol{\varphi}$.

The with $\boldsymbol{\eta}$ reparametrized log-likelihood function Ψ_n defined in (3.2) now details to:

$$\begin{aligned}\Psi_n(\boldsymbol{\eta}) &= -n\sum_{i=1}^{n}\varphi(X_i) + n\int \exp\varphi(x)\,\mathrm{d}x \\ &= -n\Big(n\eta_1 + \sum_{i=2}^{n}\sum_{k=2}^{i}\Delta X_k\eta_k\Big) + ne^{\eta_1}\sum_{i=2}^{n}\exp\Big(\sum_{k=2}^{i-1}\Delta X_k\eta_k\Big)\frac{\exp(\Delta X_i\eta_i)-1}{\eta_i}\end{aligned}$$

and the estimator we seek is then

$$\widehat{\boldsymbol{\eta}} := \operatorname*{arg\,min}_{\boldsymbol{\eta}\in\mathcal{K}_\cap}\Psi_n(\boldsymbol{\eta}).$$

In the case of convex density estimation as described in Terlaky and Vial (1998), the constraint of f being a probability density can be formulated as a simple linear equation, whereas in our case this results in the more complicated expression in (4.1).
Motivated by taking successive differences of the conditions in the definition of $\mathcal{K}_\cap$, we introduce the $m \times n$ matrix $\mathbf{B}$ with $m = n - 2$ as

$$\mathbf{B} = \begin{pmatrix} 0 & -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ & \vdots & & & \vdots & & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$$

and plugging in (4.1) into (3.2), the following optimization problem results:

$$\begin{aligned}&\text{minimize } \Psi_n(\boldsymbol{\eta}) \\ &\text{over } \boldsymbol{\eta}\in\mathbb{R}^n \text{ s.t. } \mathbf{B}\boldsymbol{\eta}\le\mathbf{0}\end{aligned} \tag{4.2}$$

where $\boldsymbol{\eta}\in\mathbb{R}^n$ is the variable in which the minimization is done and $\boldsymbol{x}\le\boldsymbol{y}$ for two vectors $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$ means that $x_i\le y_i$ for all $i=1,\ldots,n$. From Theorem A.4.1 recapitulate the necessary and sufficient Karush-Kuhn-Tucker (KKT) conditions for

$\widehat{\boldsymbol{\eta}}$ to be a solution of (4.2):

$$\begin{aligned} \nabla_{\boldsymbol{\eta}}\Psi_n + \mathbf{B}^\top \boldsymbol{v} &= \mathbf{0} && (4.3)\\ \mathbf{B}\widehat{\boldsymbol{\eta}} + \boldsymbol{s} &= \mathbf{0} && (4.4)\\ v_i s_i &= 0 \text{ for all } i = 1,\ldots,m && (4.5)\\ \boldsymbol{v} &\geq \mathbf{0} && (4.6)\\ \boldsymbol{s} &\geq \mathbf{0}. && (4.7) \end{aligned}$$

The vector $\boldsymbol{v} \in \mathbb{R}^m$ contains Lagrange-multipliers whereas $\boldsymbol{s} \in \mathbb{R}^m$ consists of slack variables. Furthermore,

$$\nabla_{\boldsymbol{\eta}}\Psi_n := \frac{\partial}{\partial \boldsymbol{\eta}}\Psi_n(\boldsymbol{\eta})$$

is the gradient of $\Psi_n = \Psi_n(\boldsymbol{\eta})$ w.r.t. $\boldsymbol{\eta}$. Let us introduce the feasible set $\mathcal{F}$ and the strictly feasible set $\mathcal{F}^\circ$:

$$\begin{aligned} \mathcal{F} &:= \{(\boldsymbol{\eta}, \boldsymbol{s}, \boldsymbol{v}) \in \mathbb{R}^{n+2m} : \nabla_{\boldsymbol{\eta}}\Psi_n + \mathbf{B}^\top \boldsymbol{v} = \mathbf{0},\ \mathbf{B}\boldsymbol{\eta} + \boldsymbol{s} = \mathbf{0},\ \boldsymbol{v} \geq \mathbf{0},\ \boldsymbol{s} \geq \mathbf{0}\}\\ \mathcal{F}^\circ &:= \{(\boldsymbol{\eta}, \boldsymbol{s}, \boldsymbol{v}) \in \mathbb{R}^{n+2m} : \nabla_{\boldsymbol{\eta}}\Psi_n + \mathbf{B}^\top \boldsymbol{v} = \mathbf{0},\ \mathbf{B}\boldsymbol{\eta} + \boldsymbol{s} = \mathbf{0},\ \boldsymbol{v} > \mathbf{0},\ \boldsymbol{s} > \mathbf{0}\}. \end{aligned}$$

Note that $\mathbf{B}\widehat{\boldsymbol{\eta}} + \boldsymbol{s} = \mathbf{0}$ for $\boldsymbol{s} \in [0,\infty)^m$ implies that $\mathbf{B}\widehat{\boldsymbol{\eta}} \leq \mathbf{0}$. Thus if $v_i > 0$ for a fixed $i \in \{1,\ldots,m\}$ then $s_i = 0$ and vice versa, by (4.5). This is known as the "complementary condition".

4.3 A primal log-barrier algorithm

The key idea of log-barrier algorithms is to introduce a barrier function h that penalizes the inequality constraints with ∞ whenever the constraints should not be satisfied. A function $h : \mathbb{R} \mapsto (-\infty, \infty]$ is a barrier function for the type of problems as in (4.2), if h is convex, continuous and nondecreasing and one has that $h(r) = \infty$ for all $r \geq 0$. The standard choice (inducing the name "log-barrier") for h is

$$h(r) := -\log(-r),$$

proposed by Fiacco and McCormick (1968). Introducing a tradeoff parameter $\mu > 0$, we thus obtain from (4.2) a barrier problem of the form:

$$\min_{\boldsymbol{\eta}\in\mathbb{R}^n} \Upsilon(\boldsymbol{\eta}, \mu) \qquad (4.8)$$

where

$$\begin{aligned}\Upsilon(\boldsymbol{\eta},\mu) &:= \Psi_n(\boldsymbol{\eta}) + \mu \sum_{i=1}^{m} h\Big((\mathbf{B}\boldsymbol{\eta})_i\Big) \\ &= \Psi_n(\boldsymbol{\eta}) - \mu \sum_{i=1}^{n-2} \log\Big(-(\mathbf{B}\boldsymbol{\eta})_i\Big).\end{aligned}$$

Similar to the inclusion of the equality constraint in (3.2), we add a Lagrange term to the criterion function to account for the inequality constraint $\mathbf{B}\boldsymbol{\eta} \leq \mathbf{0}$. Clearly the minimum of Ψ_n belongs to $\mathcal{F}^\circ$ and we can treat problem (4.2) actually as an unconstrained one. The proof of Theorem 3.2.1 together with the convexity of h entails that the function $\Upsilon(\boldsymbol{\eta},\mu)$ is strictly convex in $\boldsymbol{\eta}$ for all $\mu > 0$. Let $\widehat{\boldsymbol{\eta}}(\mu)$ denote the unique optimal point of problem (4.8) for a fixed $\mu > 0$. Collecting all these points yields a set $\mathcal{C}_p := \{\widehat{\boldsymbol{\eta}}(\mu) \ : \ \mu > 0\}$, called the "central path" of problem (4.8). The interior point log-barrier method roughly spoken follows this central path to reach an optimal solution. To accomplish this for a fixed μ, it takes repeatedly damped Newton steps in order to minimize the barrier function in (4.8), where a Newton step is as usual the minimizer of the local quadratic approximation of the objective function in (4.8). If for the specific μ the minimum is reached, μ is decreased in a controlled way. This procedure is repeated until a convergence criterion is met. Finally, the log-barrier algorithm almost boils down to an ordinary application of the Newton procedure to the function $\Upsilon = \Upsilon(\boldsymbol{\eta},\mu)$, the only speciality being the handling of μ. The Newton step, denoted by $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\eta},\mu)$, is given by

$$\boldsymbol{p} = -\mathbf{H}^{-1}\nabla_{\boldsymbol{\eta}}\Upsilon \tag{4.9}$$

where $\mathbf{H} = \mathbf{H}(\boldsymbol{\eta},\mu)$ is the Hessian matrix $\nabla^2_{\boldsymbol{\eta}\boldsymbol{\eta}}\Upsilon$ of the Lagrange-function in (4.8). To be able to measure the distance of the current iterate to the central path (and so to judge the appropriateness of a candidate), we follow the approach by Terlaky (1996), introducing the norm induced by $\mathbf{H}$:

$$\|\boldsymbol{p}\|_{\mathbf{H}} := \sqrt{\boldsymbol{p}^\top \mathbf{H}\boldsymbol{p}}.$$

The rationale behind introducing $\|.\|_{\mathbf{H}}$ is the following: ideally, we would like to measure the usual Euclidean difference between $\boldsymbol{\eta}(\mu)$ and the corresponding point on the central path $\widehat{\boldsymbol{\eta}}(\mu)$, but we do not know $\widehat{\boldsymbol{\eta}}(\mu)$. Straightforward calculation reveals, that

$$\|\boldsymbol{p}\|^2_{\mathbf{H}} = (\nabla_{\boldsymbol{\eta}}\Upsilon)^\top \mathbf{H}^{-1}(\nabla_{\boldsymbol{\eta}}\Upsilon).$$

This implies that if $\widehat{\boldsymbol{\eta}}$ is a minimizer of Υ (for a fixed μ) then $\|\boldsymbol{p}\|_{\mathbf{H}} = 0$ and $\|\boldsymbol{p}\|_{\mathbf{H}} > 0$ otherwise. So it makes sense to minimize Υ for a fixed μ as long as $\|\boldsymbol{p}\|_{\mathbf{H}}$ stays above a fixed constant (which signifies the current distance to the central path). After $\|\boldsymbol{p}\|_{\mathbf{H}}$ falling below this limit, μ is decreased and the procedure of minimizing $\boldsymbol{p}$ in $\mathbf{H}$-norm restarts. That this strategy is indeed successful guarantees Theorem A.4.2.

Putting all these ingredients together, a central path-following log-barrier algorithm can be described as follows:

input:

$\varepsilon \in \mathbb{R}_+$	:	accuracy parameter
$\tau \in (0,1)$	:	proximity parameter
$\theta \in (0,1)$	:	reduction parameter
$\mu_o \in \mathbb{R}_+$	:	initial barrier value
$\boldsymbol{\eta}_o$	:	feasible point such that $\|\boldsymbol{p}(\boldsymbol{\eta}_o, \mu_o)\|_{\mathbf{H}} \leq \tau$ and $\Psi_n(\boldsymbol{\eta}_o) < \infty$
T_1, T_2	:	maximal number of iterations for outer and inner loop

begin: $\mu := \mu_o; I_1 := 0; I_2 := 0; \boldsymbol{\eta} := \boldsymbol{\eta}_o$
while $\mu > \varepsilon/(4n)$ **and** $I_1 \leq T_1$ **do** (outer loop)
  $\mu := (1-\theta)\mu$
  $I_1 := I_1 + 1$
  $I_2 := 0$
  while $\|\boldsymbol{p}\|_{\mathbf{H}} \geq \tau$ **and** $I_2 \leq T_2$ **do** (inner loop)
    $\boldsymbol{p} :=$ solution of (4.9)
    $\tilde{\alpha} := \arg\min_{0<\alpha\leq\alpha_o}\{\Upsilon(\boldsymbol{\eta} + \alpha\boldsymbol{p}, \mu) : \boldsymbol{\eta} + \alpha\boldsymbol{p} \in \mathcal{F}^\circ\}$
    $\boldsymbol{\eta} := \boldsymbol{\eta} + \tilde{\alpha}\boldsymbol{p}$
    $I_2 := I_2 + 1$
  end (inner loop)
end (outer loop)
end.

The start vector $\boldsymbol{\eta}_o$ in the Newton procedure has to be in $\mathcal{F}$, i.e. the corresponding function φ_o must be concave. We used a quadratic interpolation to the logarithm of a kernel density estimate of the data as a first guess for our algorithm. Other approaches, such as a simple fit of a parametric log-concave density (e.g. Normal, Gamma) are also conceivable and work as well.

As an approximation to the Hessian of Υ in (4.9) we used its diagonal. It is well known (see Terlaky, 1996), that this reduced Hessian to be inverted in equation (4.9) becomes ill-conditioned as μ approaches 0. We did not encounter problems in that direction.

The upper bound α_o in the computation of $\tilde{\alpha}$ is calculated as

$$\alpha_o := 0.99 \min_{i \in \{2,\ldots,n\}} |\Delta\eta_i / \Delta p_i|,$$

so slightly below the limit beyond that a new candidate falls off $\mathcal{F}^\circ$. The step length $\tilde{\alpha}$ of the Newton step $\boldsymbol{p}$ is found via a search on a set of equidistant points.

4.4 A PRIMAL-DUAL ALGORITHM

Recapitulating the KKT conditions (4.3)-(4.7), one can derive another class of algorithms known as "primal-dual interior point methods". Introduce the mapping $F : \mathbb{R}^{n+2m} \mapsto \mathbb{R}^{n+2m}$ as:

$$F\begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{s} \\ \boldsymbol{v} \end{pmatrix} := \begin{pmatrix} \nabla_{\boldsymbol{\eta}} \Psi_n + \mathbf{B}^\top \boldsymbol{v} \\ \mathbf{B}\boldsymbol{\eta} + \boldsymbol{s} \\ \mathrm{diag}(\boldsymbol{v})\boldsymbol{s} \end{pmatrix}$$

where $\mathrm{diag}(\boldsymbol{x})$ is a diagonal matrix having the vector $\boldsymbol{x}$ on the diagonal. To see how a primal-dual algorithm works, introduce further the following system of (in-)equalities, for a fixed $\mu > 0$ and a vector $\boldsymbol{z}^\mu := (\boldsymbol{\eta}^\mu, \boldsymbol{s}^\mu, \boldsymbol{v}^\mu)$:

$$\begin{aligned} \nabla_{\boldsymbol{\eta}} \Psi_n + \mathbf{B}^\top \boldsymbol{v}^\mu &= \mathbf{0} \\ \mathbf{B}\boldsymbol{\eta}^\mu + \boldsymbol{s}^\mu &= \mathbf{0} \\ v_i^\mu s_i^\mu &= \mu \quad \text{for all} \quad i = 1, \ldots, m \\ \boldsymbol{v}^\mu &> \mathbf{0} \\ \boldsymbol{s}^\mu &> \mathbf{0}. \end{aligned} \tag{4.10}$$

These conditions differ from the original KKT conditions (4.3)-(4.7) in the term μ on the right hand side of (4.10) and the requirement that $\boldsymbol{z}^\mu$ be strictly feasible.

The central path in this case is defined as

$$\mathcal{C}_{pd} \;:=\; \{\boldsymbol{z}^\mu \,:\, \mu > 0\}.$$

An iterate in the primal-dual algorithm solves, for a fixed μ, the equation

$$F\begin{pmatrix} \boldsymbol{\eta}^\mu \\ \boldsymbol{s}^\mu \\ \boldsymbol{v}^\mu \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mu\boldsymbol{e} \end{pmatrix}, \tag{4.11}$$

where $\boldsymbol{e}$ is a vector of all 1's in appropriate dimension. One can conjecture that, as $\mu \to 0$, the corresponding vectors $\boldsymbol{z}^\mu$ approach $\boldsymbol{z}^*$ where $\boldsymbol{z}^*$ is the vector that meets the KKT conditions

$$F(\boldsymbol{z}^*) = \mathbf{0}.$$

That this strategy, implemented in the algorithm below, is indeed successful, guarantees Theorem 3.2 in Wright (1998). Note that (4.10) implies that $\boldsymbol{z}^\mu$ approaches the boundary of the feasible set $\mathcal{F}$, without actually ever leaving $\mathcal{F}^\circ$.
Looking at (4.11), we are now in the position to apply, for every fixed μ, an ordinary Newton procedure to F. For ease of simplicity, we will omit the dependence of $\boldsymbol{z}$ on μ. To get the Newton direction $\mathrm{d}\boldsymbol{z} = (\,\mathrm{d}\boldsymbol{\eta},\, \mathrm{d}\boldsymbol{s},\, \mathrm{d}\boldsymbol{v})$, the equation we actually solve is:

$$\frac{\partial F}{\partial \boldsymbol{z}} \begin{pmatrix} \mathrm{d}\boldsymbol{\eta} \\ \mathrm{d}\boldsymbol{s} \\ \mathrm{d}\boldsymbol{v} \end{pmatrix} + F(\boldsymbol{z}) = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mu\boldsymbol{e} \end{pmatrix}.$$

Computed explicitly, using the definition of F, this transforms to:

$$\begin{pmatrix} \nabla_{\eta\eta}\Psi_n & \mathbf{0} & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} & \mathbf{S} \end{pmatrix} \begin{pmatrix} \mathrm{d}\boldsymbol{\eta} \\ \mathrm{d}\boldsymbol{s} \\ \mathrm{d}\boldsymbol{v} \end{pmatrix} = - \begin{pmatrix} \nabla_{\boldsymbol{\eta}}\Psi_n + \mathbf{B}^\top \boldsymbol{v} \\ \mathbf{B}\boldsymbol{\eta} + \boldsymbol{s} \\ \mathbf{V}\boldsymbol{s} - \mu\boldsymbol{e} \end{pmatrix} \tag{4.12}$$

where we introduced the abbreviations $\mathbf{V} := \mathrm{diag}(\boldsymbol{v})$, $\mathbf{S} := \mathrm{diag}(\boldsymbol{s})$ and $\mathbf{I} := \mathrm{diag}(\boldsymbol{e})$. The Hesse matrix of Ψ_n w.r.t. to $\boldsymbol{\eta}$ is denoted by $\nabla_{\eta\eta}\Psi_n$. Formula (4.12) yields the

following equations:

$$\begin{aligned} \nabla_{\boldsymbol{\eta}\boldsymbol{\eta}}\Psi_n \,\mathrm{d}\boldsymbol{\eta} + \mathbf{B}^\top \,\mathrm{d}\boldsymbol{v} &= -\nabla\Psi_n - \mathbf{B}^\top \boldsymbol{v} \\ \mathbf{B}\,\mathrm{d}\boldsymbol{\eta} + \,\mathrm{d}\boldsymbol{s} &= -\mathbf{B}\boldsymbol{\eta} - \boldsymbol{s} \\ \mathbf{V}\,\mathrm{d}\boldsymbol{s} + \mathbf{S}\,\mathrm{d}\boldsymbol{v} &= \mathbf{V}\boldsymbol{s} - \mu\boldsymbol{e}. \end{aligned}$$

From these equations we finally get a closed system of formulas to calculate $\mathrm{d}\boldsymbol{z}$ iteratively:

$$\begin{aligned} \mathrm{d}\boldsymbol{\eta} &= -(\nabla_{\boldsymbol{\eta}\boldsymbol{\eta}}\Psi_n + \mathbf{B}^\top\mathbf{V}\mathbf{S}^{-1}\mathbf{B})^{-1}[\mathbf{B}^\top(\mathbf{V}\mathbf{S}^{-1}\mathbf{B}\boldsymbol{\eta} + \boldsymbol{v} - \mathbf{S}^{-1}\mu\boldsymbol{e}) - \nabla_{\boldsymbol{\eta}}\Psi_n] \\ \mathrm{d}\boldsymbol{s} &= -\mathbf{B}\boldsymbol{\eta} - \boldsymbol{s} - \mathbf{B}\,\mathrm{d}\boldsymbol{\eta} \\ \mathrm{d}\boldsymbol{v} &= -\mathbf{S}^{-1}(\mathbf{V}\boldsymbol{s} - \mu\boldsymbol{e}) - \mathbf{V}\mathbf{S}^{-1}\,\mathrm{d}\boldsymbol{s}. \end{aligned}$$

The only matrix for which inversion is not trivial is $(\nabla_{\boldsymbol{\eta}\boldsymbol{\eta}}\Psi_n + \mathbf{B}^\top\mathbf{V}\mathbf{S}^{-1}\mathbf{B})$, but this matrix is symmetric and positive definite, by convexity of Ψ_n and complementarity. This guarantees invertibility at every step. The detailed algorithmic procedure is as follows.

input:

$\varepsilon \in \mathbb{R}_+$	:	accuracy parameter
$\bar{\mu} \in \mathbb{R}_+$	:	lower bound for μ
$\gamma \in \mathbb{R}_+$	:	determines reduction of Newton step length via σ
$\boldsymbol{\eta}_o$	:	start vector, as in Section 4.3
T_1	:	maximal number of Newton steps

begin: $\bar{\mu} := 10^{-5}/(m+n); \gamma := (1-\sigma)^{-1}; I_1 := 0; I_2 := 0$

  while $\mu \geq \varepsilon$ **and** $I_1 * I_2 \leq T_1$ **do**

    $I_1 := I_1 + 1$

    $\mu := \max\Big(\Pi(m, \boldsymbol{s}, \boldsymbol{v}, \sigma), \bar{\mu}\Big)$

    Compute $\mathrm{d}\boldsymbol{z}$ as given in (4.13)

    $\tilde{\alpha} := \gamma \max\{\alpha > 0 \,:\, \boldsymbol{z} + \alpha\,\mathrm{d}\boldsymbol{z} \in \mathcal{F}\}$

    $\boldsymbol{z} := \boldsymbol{z} + \tilde{\alpha}\,\mathrm{d}\boldsymbol{z}$

  end

end.

The function Π calculates a new target value for μ in every iteration in the following way (according to Terlaky and Vial, 1998):

input:

$m, \boldsymbol{s}, \boldsymbol{v}, \sigma$

begin:

$E =: m \min_{i=1,\dots,m}\{v_i s_i\}/(\boldsymbol{v}^\top \boldsymbol{s})$

$\rho =: \left(\|\nabla_{\boldsymbol{\eta}}\Psi_n + \mathbf{B}^\top \boldsymbol{v}\|_2 + \|\mathbf{B}\boldsymbol{\eta} + \boldsymbol{s}\|_2\right)^{1/2}$

if: $E \geq \sigma$ **then:**

$S =: \rho/(\boldsymbol{v}^\top \boldsymbol{s} + \rho)$

else: $S =: 1$

$\mu =: S\,(\boldsymbol{v}^\top \boldsymbol{s}/m)$

end.

The lower bound $\bar{\mu}$ for μ is introduced to prevent μ from getting too small, i.e. to avoid that the current iterate is too close to the boundary of $\mathcal{F}$. If $S = 1$ then the new μ is simply the average of all pairwise products $v_i s_i$. Otherwise, almost all these products are approximately equal (resp. the minimum is a substantial proportion of the average), implying that none of the constraints are already "active", therefore μ can be decreased more rapidly.

Finally, note that if $(\boldsymbol{\eta}^*, \boldsymbol{s}^*, \boldsymbol{v}^*)$ is a solution of (4.3)-(4.7) for the current μ, then

$$\begin{aligned} \boldsymbol{\eta}^{*\top} \nabla_{\boldsymbol{\eta}}\Psi_n &= \boldsymbol{\eta}^{*\top}\left(\nabla_{\boldsymbol{\eta}}\Psi_n + \mathbf{B}^\top \boldsymbol{v}^*\right) + \boldsymbol{v}^{*\top}(-\mathbf{B}\boldsymbol{\eta}^*) \\ &= \boldsymbol{v}^{*\top} \boldsymbol{s}^*, \end{aligned}$$

so that with the definition of ρ we sort of measure how far we still are from the minimum. The number μ is generally known as "duality gap". Finally, the parameter $\gamma := (1-\sigma)^{-1}$ guarantees that $\tilde{\alpha}$ is such that $\boldsymbol{z} + \tilde{\alpha}\,\mathrm{d}\boldsymbol{z} \in \mathcal{F}^\circ$.

4.5 The modified iterative convex minorant algorithm

The ICMA was first presented in Groeneboom and Wellner (1992) and further detailed in Jongbloed (1998). It is especially tailored for minimizing a smooth convex function like Ψ_n over a convex cone such as our well-known $\mathcal{K}_\cap$. It simply minimizes the quadratic approximation to the functional under consideration (as an ordinary Newton procedure) with respect to a monotonicity constraint by using the pool adjacent violaters algorithm (PAVA, see e.g. Robertson, Wright, and Dykstra, 1988). To ensure convergence of the algorithm, one again needs to shorten the canonical Newton-direction, see Jongbloed (1998, Lemma 1). Additionally, we make use of the more general algorithmic framework provided by Dümbgen, Freitag, and Jongbloed (2006) that generalizes ICMA-like algorithms via supplementing the line search by a Hermite interpolation.

Recapitulate that Ψ_n is strictly convex and continuously differentiable on $\{\Psi_n < \infty\}$. Suppose $\mathbf{W}(\boldsymbol{x})$ is a positive definite diagonal matrix, depending continuously on $\boldsymbol{x}$ where $\boldsymbol{x} \in \mathcal{K}_\cap$. Introduce an algorithmic mapping $\mathbf{B} : \mathcal{K} \to \mathcal{K}$ where $\mathcal{K} := \{\Psi_n < \infty\} \cap \mathcal{K}_\cap$. Our goal is again to find

$$\widehat{\boldsymbol{\eta}} := \operatorname*{arg\,min}_{\boldsymbol{\eta} \in \mathcal{K}_\cap} \Psi_n(\boldsymbol{\eta}),$$

a unique point by the strict convexity of Ψ_n. Now approximate Ψ_n locally around $\boldsymbol{\delta}_o$ by the quadratic function $\tilde{\Psi}_n$:

$$\begin{aligned} \tilde{\Psi}_n(\boldsymbol{\delta}) &= \tilde{\Psi}_n(\boldsymbol{\delta}|\boldsymbol{\delta}_o) \\ &:= \Psi_n(\boldsymbol{\delta}_o) + \nabla_{\boldsymbol{\delta}} \Psi_n(\boldsymbol{\delta}_o)^\top (\boldsymbol{\delta} - \boldsymbol{\delta}_o) + 2^{-1} (\boldsymbol{\delta} - \boldsymbol{\delta}_o)^\top \mathbf{W}(\boldsymbol{\delta}_o)(\boldsymbol{\delta} - \boldsymbol{\delta}_o) \end{aligned} \quad (4.13)$$

where $\nabla_{\boldsymbol{\delta}} h(\boldsymbol{\delta}_o)$ denotes the gradient with respect to $\boldsymbol{\delta}$ at $\boldsymbol{\delta}_o$ for a function $h : \mathbb{R}^n \to \mathbb{R}$. This map provides a first guess $\boldsymbol{B}_1$ for $\boldsymbol{B}$:

$$\boldsymbol{B}_1 := \boldsymbol{B}_1(\boldsymbol{\delta}_o) := \operatorname*{arg\,min}_{\boldsymbol{\delta} \in \mathcal{K}_\cap} \tilde{\Psi}_n(\boldsymbol{\delta}). \quad (4.14)$$

If $\boldsymbol{B}_1 = \boldsymbol{\delta}_o$ we are done and set $\boldsymbol{B}(\boldsymbol{\delta}_o) = \boldsymbol{\delta}_o$. Note that this only happens if already $\boldsymbol{\delta}_o = \widehat{\boldsymbol{\eta}}$. Otherwise, apply the following robustificating line search procedure. Define the function H as

$$\begin{aligned} H(t) &:= H(t, \boldsymbol{\delta}_o, \boldsymbol{B}_1) \\ &:= \Psi_n\Big(\boldsymbol{\delta}_o + t(\boldsymbol{B}_1 - \boldsymbol{\delta}_o)\Big) - \Psi_n(\boldsymbol{\delta}_o). \end{aligned}$$

for $t \in [0, t_1]$ where $t_1 := t_1(\boldsymbol{\delta}_o, \boldsymbol{B}_1) = 2^{-m}$ with m the smallest positive integer such that $H(2^{-m}) \leq 0$. Finally, introduce a Hermite interpolation $\tilde{H}$ of H:

$$\begin{aligned} \tilde{H}(t) &= \tilde{H}(t|t_1, \boldsymbol{\delta}_o, \boldsymbol{B}_1) \\ &:= H'(0)t + \Big(t_1^{-2}H(t_1) - t_1^{-1}H'(0)\Big)t^2. \end{aligned}$$

This interpolation is constructed such that $\tilde{H}(0) = H(0) = 0$, $\tilde{H}'(0) = H'(0) > 0$, $\tilde{H}(t_1) = H(t_1) \geq 0$ and it attains its maximum over $[0, t_1]$ at

$$\begin{aligned} t_2 = t_2(t_1, \boldsymbol{\delta}_o, \boldsymbol{B}_1) &:= \arg\max_{[0,t_1]} \tilde{H}(t) \\ &= \min\Big\{\frac{t_1^2 H'(0)}{2(H'(0)t_1 - H(t_1))}, t_1\Big\} \\ &= \min\Big\{\Big(2 - 2\frac{H(t_1)}{H'(0)t_1}\Big)^{-1}, 1\Big\}t_1. \end{aligned}$$

By defining

$$\begin{aligned} \boldsymbol{B}(\boldsymbol{\delta}_o) &:= \boldsymbol{\delta}_o + t_2(\boldsymbol{B}_1 - \boldsymbol{\delta}_o) \\ &= (1 - t_2)\boldsymbol{\delta}_o + t_2\boldsymbol{B}_1 \end{aligned} \tag{4.15}$$

we get a new candidate. This procedure is justified by Theorem A.6.1. The assumptions in this theorem can easily be verified for Ψ_n and $\mathbf{B}$. Below we give pseudo-code for the ICMA.

input:

$\varepsilon \in \mathbb{R}_+$: accuracy parameter

$\boldsymbol{\delta}_o$: start vector such that $\boldsymbol{\delta}_o \in \mathcal{K}_\cap$ and $\Psi_n(\boldsymbol{\delta}_o) < \infty$

T_1, T_2 : maximal number of respective iterations

begin: $I_1 := 0; I_2 := 0; \boldsymbol{\delta} := \boldsymbol{\delta}_o; D = 2n\varepsilon$

while $|D| > n\varepsilon$ **and** $I_1 \leq T_1$ **do**

$I_1 := I_1 + 1$

$\boldsymbol{p} :=$ solution of (4.14)

$\boldsymbol{\delta}^* := \boldsymbol{\delta} + \boldsymbol{p}$

$D := \Psi_n(\boldsymbol{\delta})^\top \boldsymbol{p}$

$I_2 := 0$

while $\Psi_n(\boldsymbol{\delta}^*) > \Psi_n(\boldsymbol{\delta})$ **and** $I_2 \leq T_2$ **do** (Robustification)

$\boldsymbol{\delta}^* := (\boldsymbol{\delta} + \boldsymbol{\delta}^*)/2$

$D := D/2$

$I_2 := I_2 + 1$

end

$t^* := \Big[2 - 2\Big(\Psi_n(\boldsymbol{\delta}^*) - \Psi_n(\boldsymbol{\delta})\Big)/D\Big]^{-1}$

if $t^* < 1$ **then** (Hermite interpolation)

$\boldsymbol{\delta} := (1 - t^*)\boldsymbol{\delta}^* + t^*\boldsymbol{\delta}$

else $\boldsymbol{\delta} := \boldsymbol{\delta}^*$

end

The crucial point in the above algorithm is the minimization in (4.14), because of the constraint $\boldsymbol{\delta} \in \mathcal{K}_\cap$. We used the weighted PAVA (wPAVA) to accomplish this task. For details on the wPAVA consult Section A.5. To see how the wPAVA can be used to solve (4.14), recapitulate that the matrix $\mathbf{W}(\boldsymbol{\delta}_o)$ is diagonal, i.e.

$$\mathbf{W}(\boldsymbol{\delta}_o) := \operatorname{diag}(\boldsymbol{w})$$

for a vector $\boldsymbol{w} \in \mathbb{R}^n$. For ease of simple notation, introduce the abbreviation $\boldsymbol{g} := \nabla_{\boldsymbol{\delta}} \Psi_n(\boldsymbol{\delta}_o)$.

Inserting this in (4.13), the function $\tilde{\Psi}_n$ can then be written as:

$$\begin{aligned}\tilde{\Psi}_n(\boldsymbol{\delta}) &= \Psi_n(\boldsymbol{\delta}) + \sum_{i=1}^n g_i(\delta_i - \delta_{0,i}) + \frac{1}{2}\sum_{i=1}^n w_i(\delta_i - \delta_{0,i})^2 \\ &= \Psi_n(\boldsymbol{\delta}) + \frac{1}{2}\sum_{i=1}^n w_i\Big([(\delta_i - \delta_{0,i}) + g_i/w_i]^2 - (g_i/w_i)^2\Big) \\ &= \Psi_n(\boldsymbol{\delta}) - \frac{1}{2}\sum_{i=1}^n (g_i/w_i)^2 + \frac{1}{2}\sum_{i=1}^n w_i\Big(\delta_i - (\delta_{0,i} - g_i/w_i)\Big)^2.\end{aligned}$$

Thus, minimization of $\tilde{\Psi}_n$ over $\boldsymbol{\delta} \in \mathcal{K}_\cap$ is equivalent to the problem

$$\min_{\delta_2 \geq \ldots \geq \delta_n} \sum_{i=1}^n w_i\Big(\delta_i - (\delta_{0,i} - g_i/w_i)\Big)^2. \tag{4.16}$$

Setting $\delta_1 := \delta_{0,1} - g_1/w_1$, the weighted wPAVA is exactly what the doctor ordered to solve (4.16). For the matrix $\mathbf{W}$, we used an approximation to the complete Hessian, namely its diagonal. Robustification is necessary to guarantee conditions **B1** and **B2** of Theorem A.6.1. Dümbgen, Jongbloed and Freitag (2003) mention that numerical experiments suggested that inclusion of the Hermite interpolation improves the speed of convergence of the algorithm.

4.6 A problem-adapted algorithm

The algorithms presented so far are developed to solve general minimization problems under linear constraints, without taking into account very much the character of the problem.

A main property of all nonparametric density estimators under shape constraints (monotone, convex, log-concave) treated so far in literature is some sort of piecewise linearity with only a few knots, be at observation points or in between. See Section 3.4 and the comments there.

Dümbgen, Freitag, and Jongbloed (2006) proposed a Newton-type algorithm especially tailored for this situation. To avoid expensive inversion of huge matrices, an "oracle" guesses (at every iteration), where the knots of $\widehat{\varphi}_n$ most likely are situated and inversion only has to be performed on a subspace of $\mathbb{R}^n$ with the number of guessed knots as dimension. This new procedure was inspired by the support reduction algorithm, developed to minimize concave functions over convex cones, introduced by Groeneboom, Jongbloed, and Wellner (2003).

For our problem to fit in this new algorithmic framework, a reparametrization is necessary. Instead of a functional $\Psi_n : \mathbb{R}^n \to [-\infty, \infty)$, we need a new functional $\Psi_n : \Theta \to [-\infty, \infty)$ where $\Theta = [0, \infty)^n$. To accomplish this, introduce a vector $\boldsymbol{\theta}$, consisting mainly of the successive slope differences of the function under consideration:

$$\boldsymbol{\theta}(\boldsymbol{\varphi}) \;:=\; \Big(\varphi_1, \eta_2, -(\Delta\eta_i)_{i=3}^n\Big).$$

This $\boldsymbol{\theta}$ apparently comes up to the desired property of lying in Θ when looking at its entries $3, \ldots, n$. The first two components are just "free riders" which do not affect any calculations done for the algorithm. The aforementioned oracle for the current iterate $\boldsymbol{\theta}$ is then

$$\mathcal{I}(\boldsymbol{\theta}) := \{1, 2\} \cup \{j = 3, \ldots, n : \theta_j \geq \varepsilon(\boldsymbol{\theta})\},$$

where $\varepsilon(\boldsymbol{\theta}) > 0$ will be given later. To avoid cumbersome notation, define vectors

$$\begin{aligned} \boldsymbol{a} &:= \nabla_{\boldsymbol{\theta}} \Psi_n(\boldsymbol{\theta}), \\ \boldsymbol{b} &:= \operatorname{diag}(\mathbf{B}(\boldsymbol{\theta})) \end{aligned}$$

where $\mathbf{B}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \Psi_n(\boldsymbol{\theta})$ and $\boldsymbol{a} = \operatorname{diag}(\mathbf{A})$ is the vector consisting of the diagonal elements of a matrix $\mathbf{A}$. Given $\mathbf{B} = \mathbf{B}(\boldsymbol{\theta})$ and $\mathcal{I} = \mathcal{I}(\boldsymbol{\theta})$, we introduce sub-matrices $\mathbf{B}_{(1)}$ and $\mathbf{B}_{(2)}$:

$$\begin{aligned} \mathbf{B}_{(1)} &:= (B_{ij})_{i,j \in \mathcal{I}} \\ \mathbf{B}_{(2)} &:= \operatorname{diag}\Big((B_{ii})_{i \notin \mathcal{I}}\Big). \end{aligned}$$

Analogously define for any $\boldsymbol{y} \in \mathbb{R}^n$ sub-vectors $\boldsymbol{y}_{(1)} := (y_i)_{i \in \mathcal{I}}$ and $\boldsymbol{y}_{(2)} := (y_i)_{i \notin \mathcal{I}}$. The quadratic approximation to our functional Ψ_n we seek to minimize over

$$\{\boldsymbol{\theta}^* \;:\; \theta_j^* \geq 0 \text{ for } j \notin \mathcal{I}(\boldsymbol{\theta})\}$$

for a given $\boldsymbol{\theta}$ is then, similarly to (4.13),

$$Q(\boldsymbol{\theta}^* | \boldsymbol{\theta}) = \sum_{k=1,2} \Big(\boldsymbol{a}_{(k)}^T (\boldsymbol{\theta}_{(k)}^* - \boldsymbol{\theta}_{(k)}) + 2^{-1} (\boldsymbol{\theta}_{(k)}^* - \boldsymbol{\theta}_{(k)})^\top \mathbf{B}_{(k)} (\boldsymbol{\theta}_{(k)}^* - \boldsymbol{\theta}_{(k)}) \Big).$$

The arg min of this function can explicitly be computed as

$$\begin{aligned} \boldsymbol{p}(\boldsymbol{\theta}, \mathcal{I})_{(1)} &= \mathbf{B}_{(1)}^{-1} \boldsymbol{a}_{(1)} \\ \boldsymbol{p}(\boldsymbol{\theta}, \mathcal{I})_{(2)} &= \Big((\theta_i + a_i / b_i)^+ \Big)_{i \notin \mathcal{I}} - \boldsymbol{\theta}_{(2)}. \end{aligned}$$

To prevent the point $\boldsymbol{\theta}^* = \boldsymbol{\theta} + \boldsymbol{p}(\boldsymbol{\theta}, \mathcal{I})$ lying outside the cone Θ, replace it by

$$\boldsymbol{\theta} + t(\boldsymbol{\theta}, \mathcal{I})\boldsymbol{p}(\boldsymbol{\theta}, \mathcal{I}), \tag{4.17}$$

where $t := t(\boldsymbol{\theta}, \mathcal{I}) \in (0, 1]$ is chosen as large as possible to ensure that $\boldsymbol{\theta} + t\boldsymbol{p} \in \Theta$. Supplemented by the line search procedure already described in Section 4.5, this algorithm indeed converges to $\widehat{\boldsymbol{\theta}} := \boldsymbol{\theta}(\widehat{\boldsymbol{\varphi}})$.
We still owe the definition of the bound $\varepsilon(\boldsymbol{\theta})$, above which a θ_i is considered a potential candidate for being a knot of $\widehat{\varphi}_n$: similar to the latter paper, we used

$$2^{-1} \max_{i=3,\ldots,n} \left| \left((\theta_i + a_i/b_i)^+ \ - \theta_i \right)_{i=3}^{n} \right|.$$

A schematic algorithm looks exactly like that of the ICMA, except that the Newton step is calculated according to (4.17) instead of (4.14).
An apparent difference between the latter three and this new algorithm is the necessity of computation of not only the diagonal but the elements of the Hessian for all elements B_{ij} with $i, j \in \mathcal{I}$. However, the performance of the algorithm seems to depend on the ability to correctly choose the elements in B_{ij} with $i, j \in \mathcal{I}$.

4.7 Numerical examples

To test the algorithms, we implemented them in R, Version 2.1.1 and sampled random numbers z_k for $k = 1, \ldots, n$ for $n \in \{50, 100, 500, 1000\}$ drawn from the three distribution laws in Table 4.1.

Table 4.1: Distribution laws we sampled from.

Law	Density function	Range	Parameters
$\mathcal{N}(0,1)$	$(2\pi)^{-1/2} \exp(-z^2/2)$	$\mathbb{R}$	
$\Gamma(2,1)$	$z \exp z$	$[0, \infty)$	
Generalized Laplace(b)[a]	$K(b)(\exp(-\lvert z\rvert)1_{\{\lvert z\rvert \geq 1\}} + \exp(1/b)1_{\{\lvert z\rvert < 1\}})$	$\mathbb{R}$	$b > 0$

[a] Normalizing constant for the Generalized Laplace law is $K(b) = (2(b+1)\exp(-1/b))^{-1}$

The Normal law is chosen due to its universality and infinite support and the $\Gamma-$law because it has an infinite derivative of the log-density at 0. We introduce what we call generalized Laplace law to show that the algorithms also work for a genuine log-linear density and to assess the effect of non-differentiability points. To be able to compare the performance of the algorithms, we proceeded as follows.

1. Run the log-barrier algorithm with the settings specified below and measure its running time t_i^1 using the first argument of the R-function `system.time()` (user CPU time in seconds).

2. Run the other three algorithms until either the value of the log-likelihood or the time spent for the log-barrier algorithm was reached and measure the respective times t_i^2, t_i^3, t_i^4.

3. Repeat this for $i = 1, \ldots, 10$ times and report $t^j_{\min} := \min_{i=1,\ldots,10} t_i^j$, $\bar{t}^j := (\sum_{k=1}^{10} t_k^j)/10$ and $t^j_{\max} := \max_{i=1,\ldots,10} t_i^j$ for $j = 1, \ldots, 4$. As other measures of the quality of the estimators beneath the value of the log-likelihood we calculated for $j = 1, \ldots, 4$ the following mean errors (ME):

$$ME^j_\infty := (1/10) \sum_{i=1}^{10} \max_{k=1,\ldots,n} |\widehat{f}_i^j(z_k) - f(z_k)|$$

and

$$ME^j_1 := (1/10) \sum_{i=1}^{10} \sum_{k=1}^{n} (z_k - z_{k-1}) |\widehat{f}_i^j(z_i) - f(z_i)|.$$

Simulations were run on a Dell desktop with 1.8 GHz and 512 MB RAM. We imposed the settings detailed in Table 4.2.

Table 4.2: Settings for the ICMA and log-barrier algorithm.

Algorithm	ε	τ	θ	μ	$T1$	$T2$
log-barrier	10^{-10}	0.9	0.1	0.1	8	25
primal-dual	10^{-10}				200	20
problem-adapted	10^{-10}				200	
ICMA	10^{-10}				200	20

Simulation results for the three distributional laws in Table 4.1 were very similar, find details in Tables 4.3 to 4.5.
The ICMA clearly performs best over all sample sizes and distributional laws. All methods are able to find the minimum of the negative maximum likelihood in principal, i.e. if given enough time. In all simulations, the ICMA was the sole algorithm

Table 4.3: Results for the $\mathcal{N}(0,1)$ law.

n	Algorithm	$t^j_{\min}$	$\bar{t}^j$	$t^j_{\max}$	$\bar{LL}^j$	ME^j_∞	ME^j_1
50	ICMA	0.98	1.26	1.63	114.18	0.12	$1.88 \cdot 10^{-1}$
	log-barrier	0.97	1.34	2.51	114.22	0.12	$1.88 \cdot 10^{-1}$
	interior-point	0.98	1.27	1.62	114.39	0.12	$1.91 \cdot 10^{-1}$
	prob-adap	1.00	1.29	1.64	114.78	0.11	$1.75 \cdot 10^{-1}$
100	ICMA	1.88	3.20	4.74	232.43	0.09	$1.39 \cdot 10^{-1}$
	log-barrier	3.58	4.01	4.67	232.48	0.09	$1.38 \cdot 10^{-1}$
	interior-point	3.67	4.06	4.70	232.78	0.09	$1.41 \cdot 10^{-1}$
	prob-adap	3.67	4.09	4.75	233.29	0.07	$1.28 \cdot 10^{-1}$
500	ICMA	19.55	43.70	62.10	1192.48	0.05	$6.78 \cdot 10^{-2}$
	log-barrier	194.69	197.90	203.03	1192.62	0.05	$6.78 \cdot 10^{-2}$
	interior-point	196.26	199.30	204.39	1193.22	0.05	$8.47 \cdot 10^{-2}$
	prob-adap	195.50	199.53	204.83	1193.37	0.04	$6.27 \cdot 10^{-2}$
1000	ICMA	48.59	130.10	226.17	2358.29	0.04	$5.17 \cdot 10^{-2}$
	log-barrier	1022.08	1047.21	1070.09	2358.49	0.04	$5.21 \cdot 10^{-2}$
	interior-point	1027.03	1066.98	1088.09	2359.24	0.04	$5.34 \cdot 10^{-2}$
	prob-adap	968.76	996.44	1015.42	2358.97	0.03	$5.08 \cdot 10^{-2}$

to reach the log-likelihood value of the log-barrier algorithm (by far), whereas the other two were interrupted when reaching the time limit set by the log-barrier algorithm (note that reaching the time limit does not imply consuming exactly the same amount of seconds, because time was only compared at the beginning of a whole iteration). Quality of the estimates measured by $\bar{LL}^j$, ME^j_∞ and ME^j_1 was similar for all algorithms. As reveals Figure 4.1, the performance of the problem adapted algorithm was inferior to the others. We attribute this mainly to the structure of the Hessian, which in our case (in the contrary to that in Dümbgen, Freitag, and Jongbloed, 2006) is not as sparse as necessary for this algorithm to perform well. We seem to have many non-negligible off-diagonal entries of the Hessian. Furthermore, this algorithm operates on a different parametrization, eventually causing higher computational resource consumption.

Figure 4.1 shows typical shapes of log-likelihood curves for a single run for $n = 1000$ resulting from the estimation of a $\Gamma-$density.

After all, Figures 4.2, 4.3, and 4.4 display the estimated densities $\widehat{f}_n$ and the log-densities $\widehat{\varphi}_n$ for all three distribution laws for a sample size of 500 where the parameter for the generalized Laplace law was chosen to be $b = 1$ (for all plots: estimators are drawn in solid and functions to be estimated in dashed lines). Note the piecewise

Table 4.4: Results for the $\Gamma(2,1)$ law.

n	Algorithm	$t^j_{\min}$	$\bar{t}^j$	$t^j_{\max}$	$\bar{LL}^j$	ME^j_∞	ME^j_1
50	ICMA	0.61	1.12	1.48	121.68	0.16	$1.490 \cdot 10^{-1}$
	log-barrier	1.02	1.19	1.47	121.76	0.17	$1.49 \cdot 10^{-1}$
	interior-point	1.01	1.21	1.45	121.82	0.16	$1.45 \cdot 10^{-1}$
	prob-adap	1.00	1.20	1.49	122.31	0.18	$1.39 \cdot 10^{-1}$
100	ICMA	0.68	1.92	3.75	251.08	0.14	$1.42 \cdot 10^{-1}$
	log-barrier	3.64	3.95	4.24	251.17	0.14	$1.43 \cdot 10^{-1}$
	interior-point	3.70	4.00	4.29	251.34	0.15	$1.53 \cdot 10^{-1}$
	prob-adap	3.72	40.00	4.27	252.20	0.21	$1.62 \cdot 10^{-1}$
500	ICMA	15.16	34.37	49.03	1277.59	0.20	$9.73 \cdot 10^{-2}$
	log-barrier	190.67	198.70	206.81	1277.77	0.20	$9.76 \cdot 10^{-2}$
	interior-point	192.78	200.16	205.32	1278.54	0.21	$1.07 \cdot 10^{-1}$
	prob-adap	192.37	200.09	204.95	1279.33	0.22	$1.20 \cdot 10^{-1}$
1000	ICMA	34.97	66.73	132.16	2538.06	0.23	$9.54 \cdot 10^{-2}$
	log-barrier	1025.13	1042.66	1059.86	2538.09	0.23	$9.63 \cdot 10^{-2}$
	interior-point	1022.61	1060.10	1110.74	2539.59	0.24	$9.98 \cdot 10^{-2}$
	prob-adap	982.10	997.42	1015.47	2539.12	0.24	$1.20 \cdot 10^{-1}$

linearity of $\widehat{\varphi}_n$.

In light of Theorem 3.5.1 hardly any difference is visible on a plot displaying $\mathbb{F}_n$ and $\widehat{F}_n$. We therefore concentrate on the differences $\mathbb{F}_n - F$ and $\widehat{F}_n - F$ in Figure 4.6, recapitulate also Figure 3.1.

For all the algorithms, we did not encounter major problems up to sample sizes of 500 points. But for larger datasets and especially in case of the generalized laplace law for small b, observation points may get very close ($< 10^{-3}$) to each other, causing numerical instabilities in the inversion of matrices. In this case, it is advisable to adopt the clustering scheme described in Terlaky and Vial (1998). Replace the log-likelihood function Ψ_n and the original data $\boldsymbol{X} := (X_1, \ldots, X_n)$ by

$$-n \int w(X')\varphi(X') \, \mathrm{d}\mathbb{F}_n(X')$$

and $\boldsymbol{X'} := (X'_1, \ldots, X'_n)$, where the latter vector is constructed starting at X_1. If the distance to X_2 is smaller than some (small) resolution number $\delta > 0$, then replace X_1 and X_2 by their mean X'_1 and define $w_1 = 2$. Continue this procedure up to n and so get $\boldsymbol{X}'$ and $\boldsymbol{w}$ of length $n' \leq n$. This clustering is only a minor change in the optimization problem, but a powerful remedy against poor condition numbers in the linear systems that have to be solved to find the Newton directions.

Table 4.5: Results for the Generalized Laplace(b) law.

n	Algorithm	$t^j_{\min}$	$\bar{t}^j$	$t^j_{\max}$	$\bar{LL}^j$	ME^j_∞	ME^j_1
50	ICMA	0.35	1.02	1.42	140.69	0.08	$2.17 \cdot 10^{-1}$
	log-barrier	1.09	1.26	1.47	140.70	0.08	$2.17 \cdot 10^{-1}$
	interior-point	1.10	1.28	1.48	140.83	0.07	$2.18 \cdot 10^{-1}$
	prob-adap	1.14	1.29	1.50	141.80	0.07	$2.43 \cdot 10^{-1}$
100	ICMA	1.53	2.69	4.50	282.74	0.07	$1.54 \cdot 10^{-1}$
	log-barrier	3.72	4.18	5.00	282.77	0.07	$1.54 \cdot 10^{-1}$
	interior-point	3.81	4.23	5.05	283.18	0.07	$1.59 \cdot 10^{-1}$
	prob-adap	3.86	4.26	5.1	284.72	0.060	$1.82 \cdot 10^{-1}$
500	ICMA	21.71	33.70	45.41	1423.59	0.05	$1.93 \cdot 10^{-1}$
	log-barrier	190.91	196.36	199.97	1423.16	0.05	$1.94 \cdot 10^{-1}$
	interior-point	193.43	198.38	202.36	1424.55	0.05	$2.02 \cdot 10^{-1}$
	prob-adap	192.87	198.75	202.40	1426.07	0.05	$2.15 \cdot 10^{-1}$
1000	ICMA	19.47	67.19	139.63	2838.38	0.06	$3.34 \cdot 10^{-1}$
	log-barrier	1054.81	1064.25	1085.00	2836.28	0.06	$3.35 \cdot 10^{-1}$
	interior-point	1061.19	1081.19	1105.21	2838.81	0.06	$3.43 \cdot 10^{-1}$
	prob-adap	1000.80	1010.40	1024.22	2839.74	0.06	$3.51 \cdot 10^{-1}$

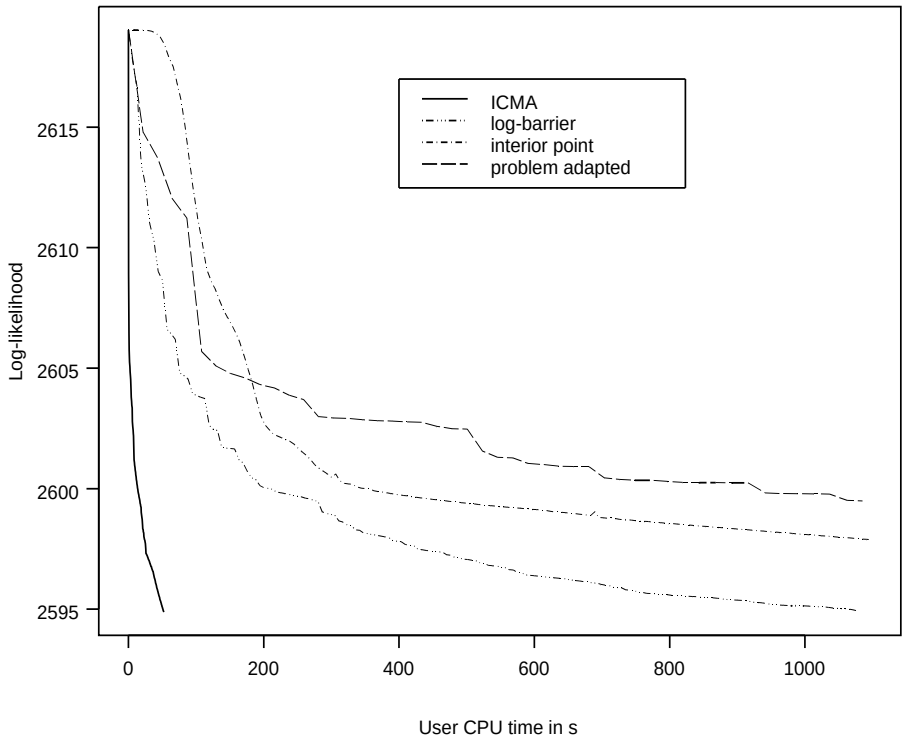

Figure 4.1: Log-likelihood functions for a run for $n = 1'000$.

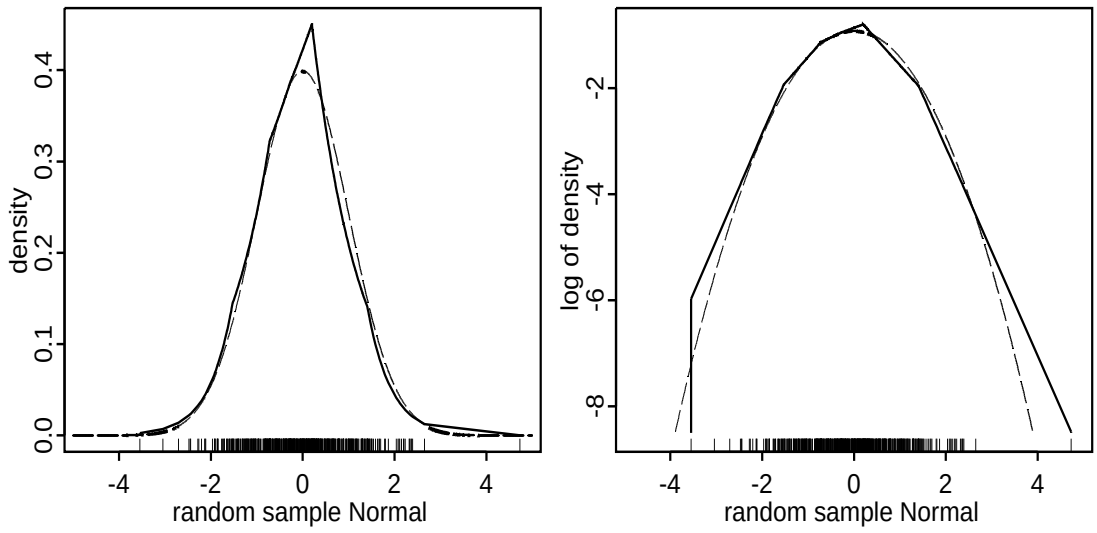

Figure 4.2: True and estimated Normal density and log-density.

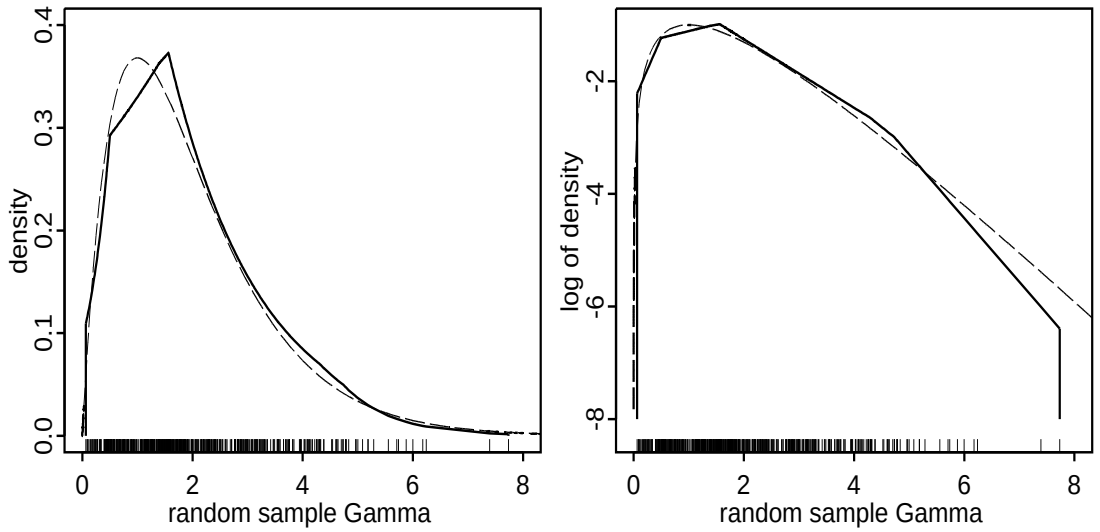

Figure 4.3: True and estimated $\Gamma-$density and log-density.

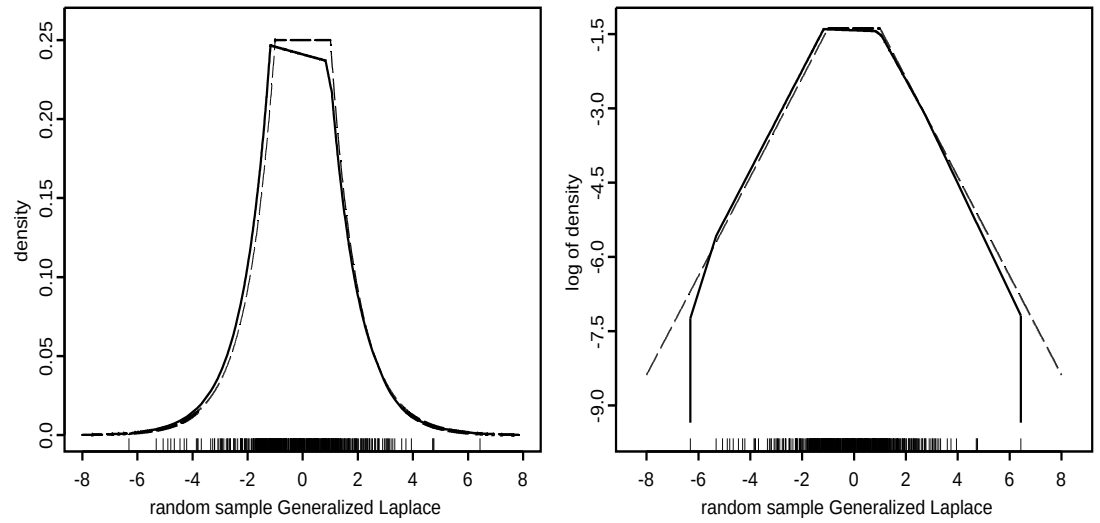

Figure 4.4: True and estimated Generalized Laplace density and log-density for $b = 1$.

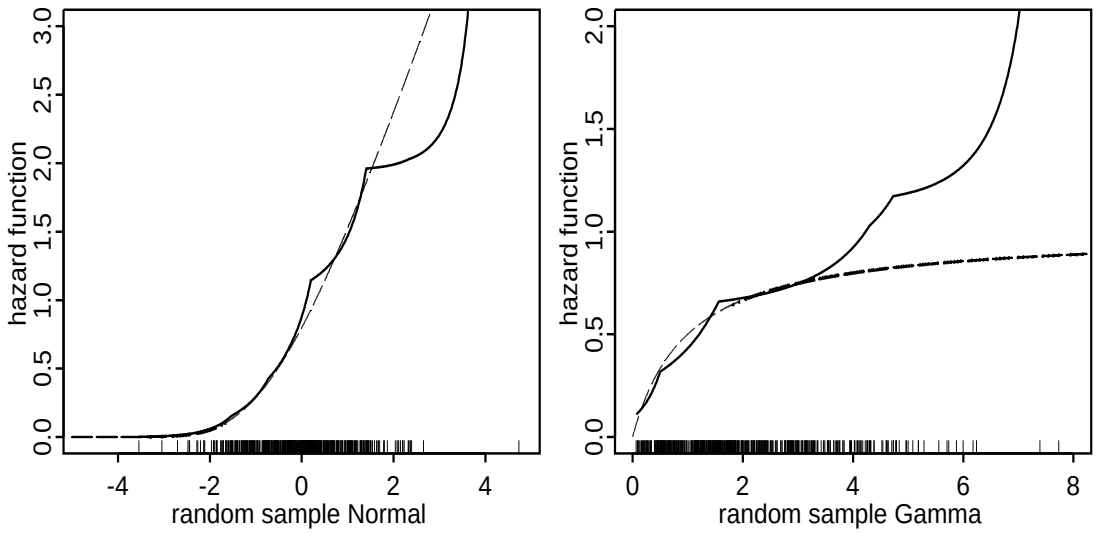

Figure 4.5: Hazard functions for Normal and Gamma sample.

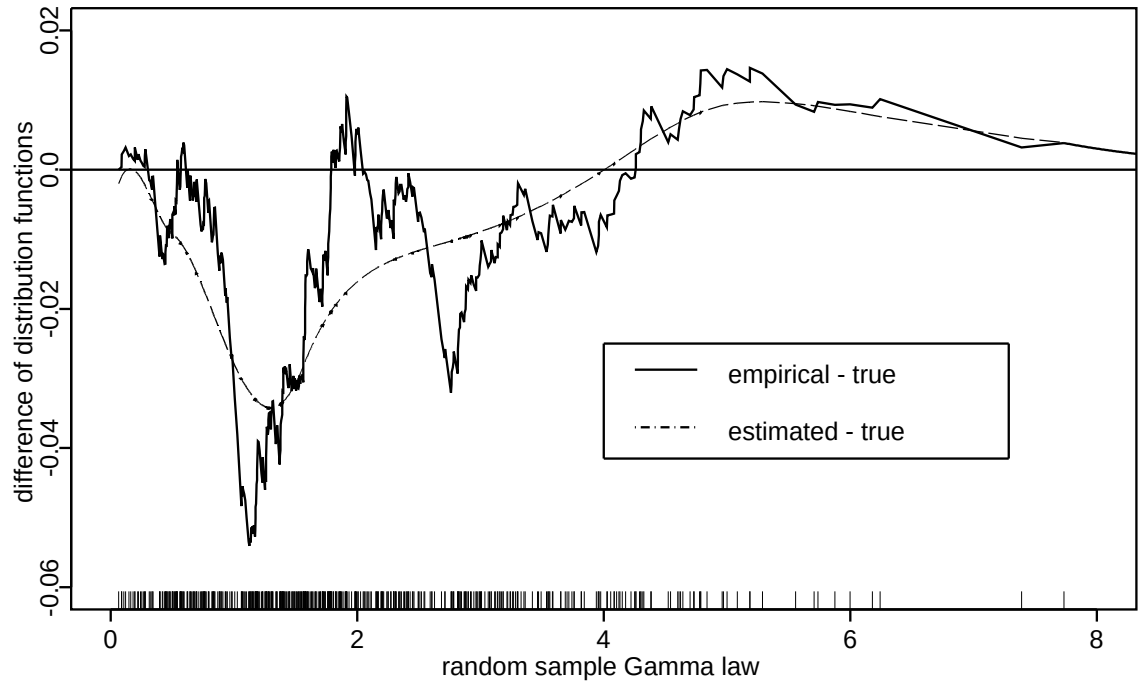

Figure 4.6: Differences of empirical and estimated to true distribution function for the Gamma law.

Part II

Bump hunting

Chapter 5

Bump hunting

The second part of this thesis proposes a method to detect regions, based on an i.i.d. sample drawn from a density f, where this density is either log-concave or log-convex. This implies lower bounds for the number of bumps and dips.

5.1 Exponential families

Let X be a random variable with distribution $P_{\boldsymbol{\theta}}$ on some measurable space $(\mathcal{X}, \mathcal{A})$ indexed by a parameter $\boldsymbol{\theta}$ ranging over an open subset Θ of $\mathbb{R}^p$. Let $p_{\boldsymbol{\theta}}$ be a density of $P_{\boldsymbol{\theta}}$ with respect to some dominating measure M. In what follows, we will choose Lebesgue measure for M. We additionally assume that $p_{\boldsymbol{\theta}}$ is a p-dimensional exponential family ($p \in \mathbb{N}$), i.e. it can be written as

$$p_{\boldsymbol{\theta}}(x) \;=\; c(\boldsymbol{\theta}) h(x) \exp\Big(\boldsymbol{\theta}^\top \boldsymbol{t}(x)\Big), \quad x \in \mathcal{X}$$

with a normalizing function $c : \Theta \to \mathbb{R}$

$$c^{-1}(\boldsymbol{\theta}) \;=\; \int_{\mathcal{X}} h(x) \exp\Big(\boldsymbol{\theta}^\top \boldsymbol{t}(x)\Big)\, \mathrm{d}x$$

and functions $h : \mathcal{X} \to \mathbb{R}$ and $\boldsymbol{t} : \mathcal{X} \to \mathbb{R}^p$. The "natural parameter space" for such a family is defined as

$$\mathcal{Y} \;=\; \{\boldsymbol{\theta} \in \mathbb{R}^p \;:\; c^{-1}(\boldsymbol{\theta}) < \infty\} \supset \Theta.$$

Define the expectation for a function $u : \mathbb{R} \to \mathbb{R}$ and the random variable X having density function $p_{\boldsymbol{\theta}}$ as

$$\mathbb{E}_{\boldsymbol{\theta}}\, u(X) = \int_{\mathcal{X}} u(t) p_{\boldsymbol{\theta}}(t)\, \mathrm{d}t. \tag{5.1}$$

Variances and covariances are written likewise. Expectations of vectors and matrices are to be understood componentwise.
Exponential families are very well studied, see e.g. Lehmann (1986, Sections 2.7 and 10.3) or van der Vaart (1998, Section 4.2). We summarize key properties of exponential families in the following lemma.

Lemma 5.1.1. *The function*

$$\boldsymbol{\theta} \to \int_{\mathcal{X}} h(x) \exp\Big(\boldsymbol{\theta}^\top \boldsymbol{t}(x)\Big)\, \mathrm{d}x$$

is infinitively often differentiable w.r.t. to $\boldsymbol{\theta}$ *and these derivatives can be found by interchanging integration and differentiation. Furthermore, for any* $\boldsymbol{u} \in \mathbb{R}^p$ *the Laplace transform is:*

$$\mathbb{E}_{\boldsymbol{\theta}} \exp\Big(\boldsymbol{u}^\top \boldsymbol{t}(X)\Big) \;=\; \frac{c(\boldsymbol{\theta})}{c(\boldsymbol{\theta}+\boldsymbol{u})}.$$

The last statement implies that $\mathbb{E} \exp[\boldsymbol{u}^\top \boldsymbol{t}(X)]$ exists if $\boldsymbol{\theta} + \boldsymbol{u} \in \mathcal{Y}$, meaning that $\boldsymbol{\theta}$ needs to be in the interior of $\mathcal{Y}$. If that is the case all moments of $\boldsymbol{t}(X)$ exist. Due to Lemma 5.1.1 the function $\log p_{\boldsymbol{\theta}}$ is infinitively often differentiable w.r.t. $\boldsymbol{\theta}$. For these two reasons the following definitions are justified for $x \in \mathcal{X}$:

$$\ell_{\boldsymbol{\theta}}(x) := \log p_{\boldsymbol{\theta}}(x) \quad \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(x) = (\partial/\partial\boldsymbol{\theta})\ell_{\boldsymbol{\theta}}(x) \quad \mathbf{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\Big(\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(X)\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(X)^T\Big),$$

where $\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}$ is denoted the "score function" and $\mathbf{I}$ the "Fisher information matrix" of the density function $p_{\boldsymbol{\theta}}$. Straightforward calculation using Lemma 5.1.1 reveals for the score function that $\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(x) = \boldsymbol{t}(x) - \mathbb{E}_{\boldsymbol{\theta}}\, \boldsymbol{t}(X)$. Therewith the following connection between the statistic $\boldsymbol{t}$ and $\mathbf{I}$ can be established:

$$\begin{aligned}
\mathbf{I}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}}\Big(\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(X)\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(X)^T\Big) \\
&= \mathbb{E}_{\boldsymbol{\theta}}\Big([\boldsymbol{t}(X) - \mathbb{E}_{\boldsymbol{\theta}}\, \boldsymbol{t}(X)][\boldsymbol{t}(X) - \mathbb{E}_{\boldsymbol{\theta}}\, \boldsymbol{t}(X)]^\top\Big) \\
&= \mathrm{Cov}_{\boldsymbol{\theta}}\, \boldsymbol{t}(X). \qquad (5.2)
\end{aligned}$$

We say that the exponential family is of "full rank" if this latter matrix $\mathrm{Cov}_{\boldsymbol{\theta}}\, \boldsymbol{t}(X)$ is non-singular. One can further derive the identity

$$\mathbb{E}_{\boldsymbol{\theta}}\, \ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(X) \;=\; -\mathbf{I}(\boldsymbol{\theta}) \qquad (5.3)$$

where $\ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(x) = (\partial/\partial\boldsymbol{\theta}^\top)\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(x)$. Now suppose we observe a sample $\boldsymbol{X} := (X_1, \ldots, X_n)$ of i.i.d. observations where all components $X_i, i = 1, \ldots, n$ have the same distribution as X. The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ based on a sample $\boldsymbol{X}$ is then defined as

$$\widehat{\boldsymbol{\theta}}_n \quad = \quad \arg\max_{\boldsymbol{\theta}\in\Theta} \widehat{L}_n(\boldsymbol{\theta}) \tag{5.4}$$

where

$$\widehat{L}_n(\boldsymbol{\theta}) \quad := \quad \sum_{i=1}^{n} \ell_{\boldsymbol{\theta}}(X_i)$$

is the log-likelihood function. Note that because the matrix

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell_{\boldsymbol{\theta}}(x) \quad = \quad -\operatorname{Cov}_{\boldsymbol{\theta}} \boldsymbol{t}(X)$$

is negative-definite, the function $\widehat{L}_n$ is strictly concave. This implies that if the exponential family $p_{\boldsymbol{\theta}}$ is of full rank and the true parameter $\boldsymbol{\theta}_o$ is in the interior of $\mathcal{Y}$, then with probability tending to one as $n \to \infty$ the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ defined by (5.4) exists, see e.g. Theorem 4.1 in van der Vaart (1998). Furthermore it exhibits the following asymptotic behavior:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) \quad \to_{\mathcal{D}} \quad \mathcal{N}_{\mathrm{p}}(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_o)^{-1}) \tag{5.5}$$

for $n \to \infty$ for every fixed $\boldsymbol{\theta}_o$ in the interior of $\mathcal{Y}$.
We intend to use a certain exponential family as a local parametric model in bump hunting. Therefore we need to generalize (5.5) to a triangular array of observations. Suppose we observe a sample $\boldsymbol{X}_n := (X_{1n}, \ldots, X_{nn})$ from $P_{\boldsymbol{\theta}_n}$. It is assumed that for a fixed n the elements of $\boldsymbol{X}_n$ are independent and identically distributed having the density $p_{\boldsymbol{\theta}_n}$ with parameter $\boldsymbol{\theta}_n \in \mathcal{Y}$ varying with n. The log-likelihood function is then generalized to

$$\widehat{L}_n(\boldsymbol{\theta}) \quad := \quad \sum_{i=1}^{n} \ell_{\boldsymbol{\theta}}(X_{in}).$$

Assume for the parameter $\boldsymbol{\theta}_n$ that it converges to $\boldsymbol{\theta}_o$ componentwise, at an arbitrary rate of convergence, i.e. for all $i = 1, \ldots, p$

$$\theta_{n,i} - \theta_{0,i} \quad = \quad o(1).$$

One can then extend statement (5.5) in the following sense.

Theorem 5.1.2. *Suppose that every element of* $\boldsymbol{X}_n := (X_{1n}, \ldots, X_{nn})$ *is i.i.d. having density function* $p_{\boldsymbol{\theta}_n}$. *Let* $p_{\boldsymbol{\theta}_n}$ *be an exponential family with full rank for every* n. *Then:*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) \quad \to_{\mathcal{D}} \quad \mathcal{N}_{\mathrm{p}}(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_o)^{-1}) \tag{5.6}$$

for $n \to \infty$.

5.2 Testing of composite hypotheses

To set up our multiscale test we will use a specific score test statistic in a specific two-parameter model. In this section we introduce score tests in exponential families in general and compare its power properties to a likelihood ratio test (LRT). We will furthermore assess the effect of nuisance parameters on the power of the above tests.

We adopt the setting of Section 5.1. To keep notation simple, let us split the Fisher matrix $\mathbf{I}$ as follows:

$$\mathbf{I}(\boldsymbol{\theta}) \quad := \quad \begin{pmatrix} \mathbf{I}^{11}(\boldsymbol{\theta}) & \mathbf{I}^{12}(\boldsymbol{\theta}) \\ \mathbf{I}^{21}(\boldsymbol{\theta}) & \mathbf{I}^{22}(\boldsymbol{\theta}) \end{pmatrix}$$

where

$$\begin{aligned} \mathbf{I}^{11}(\boldsymbol{\theta}) &= (\mathbf{I}_{ij}(\boldsymbol{\theta}))_{i,j=1,\ldots,p-1}, \\ \mathbf{I}^{12}(\boldsymbol{\theta}) &= (\mathbf{I}_{1,p}(\boldsymbol{\theta}), \ldots, \mathbf{I}_{p-1,p}(\boldsymbol{\theta}))^\top, \\ \mathbf{I}^{21}(\boldsymbol{\theta}) &= \mathbf{I}^{12}(\boldsymbol{\theta})^\top = (\mathbf{I}_{1,p}(\boldsymbol{\theta}), \ldots, \mathbf{I}_{p-1,p}(\boldsymbol{\theta})), \\ \mathbf{I}^{22}(\boldsymbol{\theta}) &= \mathbf{I}_{p,p}(\boldsymbol{\theta}). \end{aligned}$$

The following definition of a specific number will turn out be useful below:

$$\mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}) \quad = \quad \mathbf{I}^{22}(\boldsymbol{\theta}) - \mathbf{I}^{21}(\boldsymbol{\theta})\mathbf{I}^{11}(\boldsymbol{\theta})^{-1}\mathbf{I}^{12}(\boldsymbol{\theta}).$$

Given a vector $\boldsymbol{x} \in \mathbb{R}^p$ we write $\tilde{\boldsymbol{x}}$ for its first $p-1$ components: $\tilde{\boldsymbol{x}} = (x_1, \ldots, x_{p-1})$. Let $\boldsymbol{e}_p := (0, \ldots, 0, 1) \in \mathbb{R}^p$. For a fixed $\eta \in \mathbb{R}$ introduce the following set:

$$\Theta_\eta := \{\boldsymbol{\vartheta} \in \mathcal{Y} \ : \ \vartheta_p = \eta\}.$$

Then suppose we have an i.i.d. sample $\boldsymbol{X}_n = (X_{1n}, \ldots, X_{nn})$ where each component is distributed according to $P_{\boldsymbol{\theta}_n}$ introduced in Section 5.1. The row-wise "true"

parameter $\boldsymbol{\theta}_n \in \mathcal{Y}$ shall be converging to $\boldsymbol{\theta}_o \in \Theta_o$ componentwise, at a rate of convergence not yet further specified. Then consider the following test problem:

$$H_o : \boldsymbol{\theta} \in \Theta_o \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} \text{ is unrestricted}$$

which is equivalent to

$$H_o : \theta_p = 0 \quad \text{vs.} \quad H_1 : \theta_p \neq 0.$$

The test statistic we analyze first is the LRT statistic Λ_n

$$\Lambda_n \quad = \quad 2 \sup_{\theta \in \mathcal{Y}} \widehat{L}_n(\boldsymbol{\theta}) - 2 \sup_{\theta \in \Theta_o} \widehat{L}_n(\boldsymbol{\theta}).$$

Beneath the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ in the full model, introduce the estimator in the restricted model for an arbitrary fixed $\eta \in \mathbb{R}$:

$$\widehat{\boldsymbol{\theta}}_n^{\eta} \quad = \quad \arg\max_{\boldsymbol{\theta} \in \Theta_\eta} \widehat{L}_n(\boldsymbol{\theta}).$$

The likelihood ratio test statistic then becomes

$$\Lambda_n \quad = \quad 2\widehat{L}_n(\widehat{\boldsymbol{\theta}}_n) - 2\widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^0).$$

For a given significance level $\alpha \in (0,1)$, the null hypothesis H_o is rejected by the LRT if, and only if, $\Lambda_n \geq c_\alpha$ where $c_\alpha = c_\alpha(\widehat{\boldsymbol{\theta}}_n, \widehat{\boldsymbol{\theta}}_n^0) \in (1,\infty)$. If there exists a $c_\alpha \in (1,\infty)$ such that

$$\sup_{\boldsymbol{\theta} \in \Theta_o} P_{\boldsymbol{\theta}}(\Lambda_n \geq c_\alpha) \quad = \quad \alpha,$$

then we get a LRT of size α. However, it is often difficult to find a LRT with size α for a fixed finite n and one has to switch to tests of only asymptotic size α. This is what we do in the following theorem.

Theorem 5.2.1. *Suppose the elements of $\boldsymbol{X}_n$ are independent and have density function $p_{\boldsymbol{\theta}_n}$ where $\boldsymbol{\theta}_n - \boldsymbol{\theta}_o = o(\mathbf{1})$. The statistic Λ_n has then the following asymptotic behavior:*

$$\Lambda_n \to_{\mathcal{D}} \begin{cases} \infty & \text{if } \sqrt{n}|\theta_{n,p}| \to \infty \\ \chi_1^2(\mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)h^2) & \text{if } \sqrt{n}|\theta_{n,p}| \to h \\ \chi_1^2(0) & \text{if } \sqrt{n}|\theta_{n,p}| \to 0 \end{cases}$$

where $h > 0$ and $\chi_1^2(p)$ is the non-central χ^2-distribution with one degree of freedom and non-centrality parameter p.

For a given significance level $\alpha \in (0,1)$ we reject the null hypothesis if Λ_n exceeds the critical value $\chi^2_{1;1-\alpha}$ where $\chi^2_{1;1-\alpha}$ is the $(1-\alpha)$-quantile of a χ^2-distribution with one degree of freedom. Such a test has then by construction asymptotic size α.
The (local, i.e. if not $\sqrt{n}|\theta_{n,p}| \to \infty$) power function π^L_n of the above test then satisfies, as $n \to \infty$,

$$\pi^L_n(\widehat{\boldsymbol{\theta}}^0_n, \widehat{\boldsymbol{\theta}}_n) - \pi^L\left(\mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)^{1/2}\sqrt{n}|\theta_{n,p}|\right) \;=\; o_{\mathrm{p}}(1).$$

Explicitly, the asymptotic power function is

$$\pi^L(p) \;=\; 1 - \chi^2_1(p^2, \chi^2_{1;1-\alpha})$$

where $\chi^2_1(p,.)$ is the χ^2-distribution function for one degree of freedom and non-centrality parameter $p \geq 0$.
Note that non-central χ^2-distributions are stochastically increasing in the non-centrality parameter, i.e. for two non-centrality parameters $p_1 < p_2$

$$\chi^2_1(p_1,.) \;\geq \chi^2_1(p_2,.),$$

implying that the LRT has good (local) power properties at large values of the non-centrality parameter.
The LRT introduced above is two-sided, i.e. in case of rejection of the null hypothesis, nothing about the sign of $\theta_{n,p}$ can be said. In our intended application to bump hunting however, it will be convenient to be able to make a statement about $\mathrm{sign}(\theta_{n,p})$ in case H_o is rejected, at least with a certain (asymptotic) confidence. The score test below is exactly what the doctor ordered. Its test statistic is defined as a normalized derivative of the profile log-likelihood function at $\eta = 0$:

$$S_n \;:=\; n^{-1/2}\frac{\partial}{\partial\eta}\widehat{L}_n(\widehat{\boldsymbol{\theta}}^\eta_n)\Big|_{\eta=0}.$$

The hypotheses we test are

$$H_o : \theta_{n,p} < 0 \quad \text{vs.} \quad H_1 : \theta_{n,p} \geq 0 \tag{5.7}$$

or vice versa. Again, as for the LRT, we can specify the limiting distribution for this statistic, depending on the behavior of $\theta_{n,p}$.

Theorem 5.2.2. *Under the assumptions of Theorem 5.2.1 the score test statistic S_n has the following asymptotic distribution:*

$$\mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)^{-1/2} S_n \to_{\mathcal{D}} \begin{cases} \pm\infty & \text{if } \sqrt{n}\theta_{n,p} \to \pm\infty \\ \mathcal{N}(\mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)^{1/2} h, 1) & \text{if } \sqrt{n}\theta_{n,p} \to h \\ \mathcal{N}(0,1) & \text{if } \sqrt{n}\theta_{n,p} \to 0 \end{cases}$$

for $h \in \mathbb{R}$.

In light of Theorem 5.2.2, for a given significance level $\alpha \in (0,1)$ the null hypothesis H_o in (5.7) is rejected if $\mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)^{-1/2} S_n \geq z_{1-\alpha}$ where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of a standard normal distribution. However, we do not know $\boldsymbol{\theta}_o$, but it seems clear, that a suitable consistent estimate of $\boldsymbol{\theta}_o$ can save us. For the specific two-parameter model elaborated in Section 5.3 this is detailed in Theorem 5.4.1.
As for the actual calculation of the score statistic S_n, observe the following. The log-likelihood function $\widehat{L}_n$ is a map from $\mathbb{R}^p \to \mathbb{R}$. Therefore:

$$\begin{aligned} \frac{\partial}{\partial \eta} \widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^{\eta}) &= \nabla \widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^{\eta})^{\top} \Big(\frac{\partial}{\partial \eta} \widehat{\boldsymbol{\theta}}_n^{\eta} \Big) \\ &= \boldsymbol{e}_p^{\top} \nabla \widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^{\eta}) \\ &= \Big(\nabla \widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^{\eta}) \Big)_p . \end{aligned}$$

This implies that

$$\begin{aligned} S_n &= n^{-1/2} \frac{\partial}{\partial \eta} \widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^{\eta}) \Big|_{\eta=0} \\ &= \Big(n^{-1/2} \sum_{i=1}^{n} \dot{\boldsymbol{\ell}}_{\widehat{\boldsymbol{\theta}}_n^0}(X_{1n}) \Big)_p . \end{aligned} \tag{5.8}$$

In other words, to calculate the score statistic S_n for a test on the p-th coordinate of $\boldsymbol{\theta}$, we can simply take the p-th coordinate of the score vector where we readily input the estimate under the constraint $\theta_{n,p} = 0$, namely $\widehat{\boldsymbol{\theta}}_n^0$.
Consider the general situation of tests involving a fixed number of parameters where some other nuisance parameter has to be estimated. Suppose further this nuisance parameter is estimated under the null using a $\sqrt{n}$-consistent estimator (e.g. maximum likelihood). It is well known that in this case likelihood ratio, score (and Wald)

tests are asymptotically equivalent under the null hypothesis, see e.g. Shao (2003, Section 4.5.2). In Theorems 5.2.1 and 5.2.2 we consider the more general situation of a "true" parameter $\boldsymbol{\theta}_n$ varying with n and one-parameter alternatives that lie in a $O(n^{-1/2})$-ball around the parameter $\theta_{n,p}$ we perform the test on.
The score statistic is designed to test the hypotheses (5.7) or vice versa, effectively entailing a statement about $\text{sign}(\theta_{n,p})$ in case of rejection of H_o, with asymptotic confidence $1-\alpha$. Using this, define a modified score test by combining two one-sided score tests using the test statistic S_n where each of the two tests is performed at half of the overall significance level α. For the local power function π_n^S in this case we have, according to Theorem 5.2.2 as $n \to \infty$,

$$\pi_n^S(\widehat{\boldsymbol{\theta}}_n^0, \widehat{\boldsymbol{\theta}}_n) - \pi^S\left(\mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)^{1/2}\sqrt{n}\theta_{n,p}\right) \;=\; o_{\mathrm{p}}(1).$$

To derive π^S, consider the case of testing the one-sided hypotheses in (5.7). According to 5.2.2, the asymptotic power function for testing at significance level $\alpha/2$ for any fixed $\alpha \in (0,1)$, $m \in \mathbb{R}$ and a random variable Z having a $\mathcal{N}(m,1)$ distribution, is

$$\begin{aligned} P(Z > z_{1-\alpha/2}) &= 1 - P(Z - m \le -z_{1-\alpha/2} - m) \\ &= 1 - \Phi(-z_{1-\alpha/2} - m) \end{aligned}$$

where $\Phi(.)$ is the standard normal distribution function. As we simply put together two one sided tests, testing either the hypotheses (5.7) or their reversed versions, we can write for the asymptotic power function for all $m \in \mathbb{R}$

$$\begin{aligned} \pi^S(m) &= [1 - \Phi(-z_{1-\alpha/2} - m)]1_{\{m \ge 0\}} + [1 - \Phi(-z_{1-\alpha/2} + m)]1_{\{m \le 0\}} \\ &= 1 - \Phi(-z_{1-\alpha/2} - |m|). \end{aligned}$$

Normal distributions with variance 1 (or in general with equal variance) are stochastically increasing in the mean, i.e. for two means $p_1 < p_2$

$$\Phi_1(p_1,) \;\ge\; \Phi_1(p_2, .)$$

entailing that, similar to the LRT, the score test has good local power properties for large values of $\mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)^{1/2}\sqrt{n}\theta_{n,p}$.
Recapitulate the asymptotic power functions for the above described tests, for a fixed significance level $\alpha \in (0,1)$ and any $p \in \mathbb{R}$,

$$\begin{aligned} \pi^L(p) &= 1 - \chi_1^2(p^2, \chi^2_{1;1-\alpha}) \\ \pi^S(p) &= 1 - \Phi(-z_{1-\alpha/2} - |p|). \end{aligned}$$

These two functions are almost identical, their difference decreases very fast with growing first argument. The only difference happens around 0, due to the fact that the score test is performed at half the significance level α compared to the LRT. Note that the power (against the considered local alternatives) for both tests introduced above is increased when $\mathbf{I}^{22\cdot1}(\boldsymbol{\theta}_o)$ increases. Recall the definition of $\mathbf{I}^{22\cdot1}(\boldsymbol{\theta})$

$$\mathbf{I}^{22\cdot1}(\boldsymbol{\theta}) \;=\; \mathbf{I}^{22}(\boldsymbol{\theta}) - \mathbf{I}^{21}(\boldsymbol{\theta})\mathbf{I}^{11}(\boldsymbol{\theta})^{-1}\mathbf{I}^{12}(\boldsymbol{\theta}).$$

Mathematical expressions simplify if one considers a model that has a diagonal Fisher matrix. Since in that case $\mathbf{I}^{12}(\boldsymbol{\theta}) = \mathbf{0}$ and consequently

$$\mathbf{I}^{22\cdot1}(\boldsymbol{\theta}) \;=\; \mathbf{I}^{22}(\boldsymbol{\theta}).$$

5.3 A SPECIFIC TWO-PARAMETER MODEL

This section is devoted to a specific two-parameter exponential family which serves as a building block for the multiscale test in Section 5.6. Let the random variable X_n have the univariate two-parameter density f_{θ_n,η_n} where

$$f_{\theta,\eta}(x) := C(\theta,\eta)\exp\Big(\theta x + \eta x^2/2\Big), \quad x \in [0,1] \tag{5.9}$$

for $\theta, \eta \in \mathbb{R}$ and a normalizing constant

$$C^{-1}(\theta,\eta) \;:=\; \int_0^1 \exp\Big(\theta x + \eta x^2/2\Big)\,\mathrm{d}x.$$

For the sequences of parameters we assume that $\theta_n \to \theta_o$ as well as $\eta_n \to 0$. Furthermore, for all n these sequences belong to the natural parameter space of $f_{\theta,\eta}$, i.e. $C^{-1}(\theta_n,\eta_n) < \infty$. Denote by X_∞ the random variable having density function $f_{\theta_o,0}$.

For n ordered i.i.d. observations $X_{1n} < \ldots < X_{nn}$ all having the same distribution as X_n, define a data vector $\boldsymbol{X}_n := (X_{1n},\ldots,X_{nn})$.

To embed this specific model in the framework of Sections 5.1 and 5.2 note that $f_{\theta,\eta}$ can be written as

$$f_{\boldsymbol{\theta}}(x) = c(\boldsymbol{\theta})h(x)\exp\Big(\boldsymbol{\theta}^\top \boldsymbol{t}(x)\Big)$$

with $\boldsymbol{\theta} := (\theta,\eta)$, $c(\boldsymbol{\theta}) := C(\theta,\eta)$, $h(x) := 1$ and $\boldsymbol{t}(x) := (x, x^2/2)$.

In bump hunting we will set up a multiscale test to assess log-concavity and log-convexity of a density, on specific intervals. The current two-parameter model will serve as basic element for this multiscale test. Based on a sample $\boldsymbol{X}_n$ a test

$$\begin{aligned} H_o &: f_{\theta_n,\eta_n} \text{ is log-linear vs.} \\ H_1 &: f_{\theta_n,\eta_n} \text{ is log-concave} \end{aligned}$$

translates into the following one-sided test for η_n:

$$\begin{aligned} H_o &: \eta_n = 0 \\ H_1 &: \eta_n < 0, \end{aligned}$$

where θ_n is unknown and takes the role of a nuisance parameter, i.e. needs to be estimated from the same sample $\boldsymbol{X}_n$. Testing for log-convexity is similar. Relying on the results of Section 5.2 we propose a score test, in order to be able to infer $\text{sign}(\eta_n)$ in case of rejection of H_o. The score test statistic in this specific problem is then, according to (5.8),

$$\begin{aligned} S_n &= \Big(n^{-1/2}\sum_{i=1}^{n} \dot{\boldsymbol{\ell}}_{\widehat{\theta}_n^0,0}(X_{in})\Big)_2 \\ &= (1/2)n^{1/2}\Big(\overline{X_{in}^2} - \mathbb{E}_{\widehat{\theta}_n^0,0}\, X_{1n}^2\Big) \end{aligned} \tag{5.10}$$

where we introduced the score vector

$$\dot{\boldsymbol{\ell}}_{\theta,\eta} := \frac{\partial}{\partial(\theta,\eta)} \log f_{\theta,\eta},$$

the maximum likelihood estimator $\widehat{\theta}_n^0$ of θ_n based on a sample $\boldsymbol{X}_n$ under the null hypothesis and an abbreviation for the mean

$$\overline{\boldsymbol{x}_i} = (1/n)\sum_{i=1}^{n} \boldsymbol{x}_i$$

for n vectors $\boldsymbol{x}_i \in \mathbb{R}^k$ (or n real numbers if $k = 1$). The estimator $\widehat{\theta}_n^0$ can be found using e.g. a Newton-Raphson procedure.

On p. 97 we discussed that a score test based on the statistic S_n is mathematically more convenient if the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta})$ is diagonal at the true parameter $\boldsymbol{\theta}_o$. When adopting the model (5.9) directly, the corresponding Fisher matrix

$$\mathbf{I}_n(\theta,\eta) = \begin{pmatrix} \text{Var}_{\theta,\eta}\, X_{1n} & \text{Cov}_{\theta,\eta}(X_{1n}, X_{1n}^2)/2 \\ \text{Cov}_{\theta,\eta}(X_{1n}, X_{1n}^2)/2 & \text{Var}_{\theta,\eta}(X_{1n}^2)/4 \end{pmatrix}$$

does clearly not have vanishing diagonal elements at $(\theta_o, 0)$, i.e. when $n \to \infty$. This is due to the fact that the covariance between X_∞ and X_∞^2 at $(\theta_o, 0)$ does not disappear.
In order to have mathematically convenient expressions, we therefore propose the following remedy. Instead of adopting the density function f_{θ_n,η_n} directly, replace it by $f^*_{\theta_n,\eta_n}$ where

$$f^*_{\theta,\eta}(x) := C^*(\theta,\eta)\exp\Big[\theta x + \eta\Big(x^2/2 - a(\theta)x - b(\theta)\Big)\Big] \tag{5.11}$$

for $x \in [0,1]$. The score vector corresponding to this density $f^*_{\theta,\eta}$ is

$$\dot{\boldsymbol{\ell}}^*_{\theta,\eta}(x) \;=\; \begin{pmatrix} x - a'(\theta)\eta x - \mathbb{E}_{\theta,\eta}(X_{1n} - a'(\theta)\eta X_{1n}) \\ T_\theta(x) - \mathbb{E}_{\theta,\eta}\, T_\theta(X_{1n}) \end{pmatrix}$$

where $T_\theta(x) := x^2/2 - a(\theta)x - b(\theta)$ for any $\theta \in \mathbb{R}$ and $x \in [0,1]$. The functions $a : \mathbb{R} \to \mathbb{R}$ and $b : \mathbb{R} \to \mathbb{R}$ are chosen such that

$$\begin{array}{lcll} \mathbb{E}_{\theta,0}\, T_\theta(X) & = & 0 & \text{and} \\ \mathbb{E}_{\theta,0}[T_\theta(X)X] & = & 0 & \end{array} \tag{5.12}$$

for all $\theta \in \mathbb{R}$ such that $C^{-1}(\theta,0) < \infty$ where X is distributed such that it exhibits a density function $f_{\theta,0}$. Properties of these latter functions are collected in 5.3.1. Deduce a modified score statistic according to (5.8) as follows:

$$\begin{aligned} S^*_n &= \Big(n^{-1/2}\sum_{i=1}^n \dot{\boldsymbol{\ell}}^*_{\widehat{\theta}^0_n,0}(X_{in})\Big)_2 \\ &= n^{1/2}\Big(\overline{T_{\widehat{\theta}^0_n}(X_{in})} - \mathbb{E}_{\widehat{\theta}^0_n,0}\, T_{\widehat{\theta}^0_n}(X_{in})\Big) \\ &= n^{1/2}\overline{T_{\widehat{\theta}^0_n}(X_{in})}. \end{aligned} \tag{5.13}$$

This construction immediately entails

$$\mathrm{Cov}_{\widehat{\theta}^0_n,\eta_n}\Big(X_{1n}, T_{\widehat{\theta}^0_n}(X_{1n})\Big) \;\to_{\mathrm{p}}\; 0, \tag{5.14}$$

implying that the Fisher matrix corresponding to (5.11) becomes diagonal as $n \to \infty$. By Theorem 5.2.2 we get

$$\frac{S^*_n}{(\mathrm{Var}_{\theta_o,0}\, X_{1n})^{1/2}} \;\to_{\mathcal{D}}\; \mathcal{N}\Big((\mathrm{Var}_{\theta_o,0}\, X_{1n})^{1/2}h, 1\Big)$$

when $\sqrt{n}\eta_n \to h$. However, θ_o is not known and has to be estimated. How this affects the test statistic is detailed in Section 5.4.
In Section 5.5 model (5.11) will be considered to derive a score test statistic enabling to test whether η_n is significantly different from 0. The difference between a score test statistic derived from $f_{\theta,\eta}$ to one received via $f^*_{\theta,\eta}$ is the different centering term, compare (5.10) to (5.13). Note that

$$a(\widehat{\theta}_n)\overline{X_{in}} + b(\widehat{\theta}_n)$$

consistently estimates $\eta_o = 0$ for an arbitrary consistent estimator $\widehat{\theta}_n$ of θ_o. For every n, the coefficient of the linear term θ_n takes the role of a nuisance parameter and must be estimated. The fact detailed in (5.14) ensures that estimation of θ_n and η_n are, at least asymptotically, "as independent as possible", i.e. do affect each other as little as possible.
To conclude this section, we owe the exact representations for the functions a and b. To omit these formulas being even more lengthy than they already are, introduce for $k = 0, 1, 2, \ldots$ and any $\theta \in \mathbb{R}$

$$H_k(\theta) \;=\; \int_0^1 x^k \exp(\theta x)\,\mathrm{d}x. \tag{5.15}$$

Using this abbreviation one can derive the following formulas for a and b from (5.12):

$$\begin{aligned} a(\theta) &= \frac{1}{2}\frac{H_1(\theta)H_2(\theta) - H_o(\theta)H_3(\theta)}{H_1(\theta)^2 - H_o(\theta)H_2(\theta)} \\ b(\theta) &= \frac{1}{2}\frac{H_1(\theta)H_3(\theta) - H_2(\theta)^2}{H_1(\theta)^2 - H_o(\theta)H_2(\theta)}. \end{aligned}$$

Some properties of these functions are collected in Lemma 5.3.1 and Figure 5.1 provides a plot.

Lemma 5.3.1. *For the function a we have the following limits:*

$$\lim_{\theta\to-\infty} a(\theta) = 0 \qquad \lim_{\theta\to\infty} a(\theta) = 1$$

and for b:

$$\lim_{\theta\to-\infty} b(\theta) = 0 \qquad \lim_{\theta\to\infty} b(\theta) = -1/2.$$

Furthermore, a is symmetric around 0: for any $\theta \in \mathbb{R}$ one has

$$a(-\theta) \;=\; 1 - a(\theta).$$

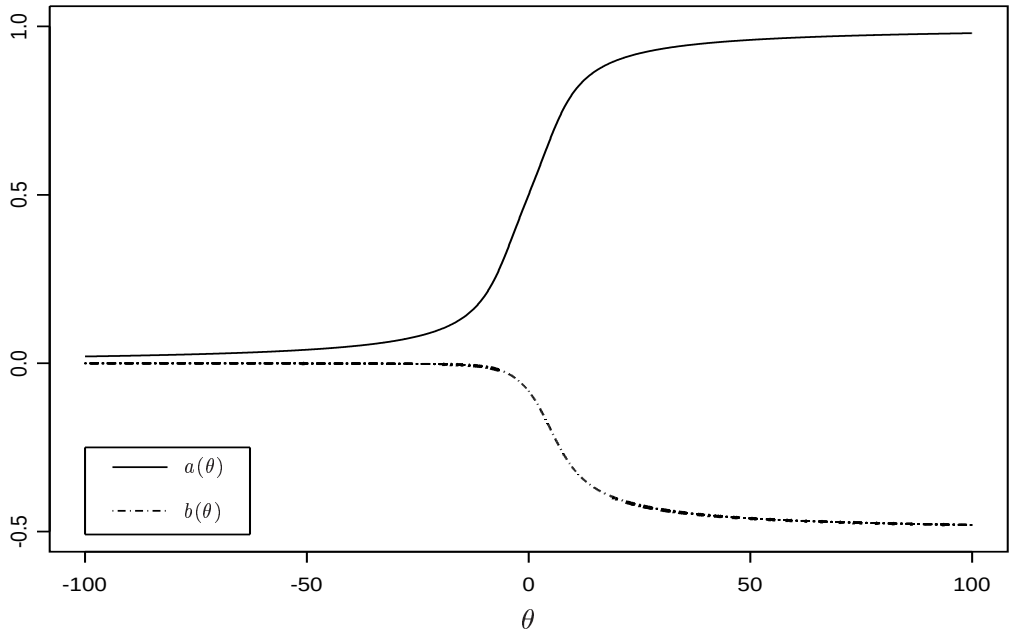

Figure 5.1: Centering functions $a(\theta)$ and $b(\theta)$.

Note that in Section 5.5 we rescale our original observations $X_1, \ldots, X_n$ such that they lie in $[0, 1]$. This is anticipated in the definition of the densities $f_{\theta,\eta}$ and $f^*_{\theta,\eta}$, as they will serve as a basis to introduce a multiscale test based on the rescaled observations. Clearly, the setting has implications on the precise form of a and b when defining them via (5.12). This latter definition provides one with the "simplest" form of these functions as well as the above densities, however it may not be optimal with regard to symmetry. If symmetry was the aim, one could rather concentrate on

$$\breve{T}_\theta(x) \quad = \quad (x - 1/2)^2/2 - \breve{a}(\theta)(x - 1/2) - \breve{b}(\theta)$$

where

$$\breve{a} = a - 1/2 \quad \text{and} \quad \breve{b} = a/2 + b + 1/8.$$

Remembering that T_θ is the score statistic derived from (5.11), the density having $\breve{T}_\theta - \mathbb{E}_{\theta,\eta} \breve{T}_\theta(X_{1n})$ as a score function is

$$\breve{f}_{\theta,\eta}(x) \quad = \quad \breve{C}(\theta, \eta) \exp\Big[\theta(x - 1/2) + \eta\Big((x - 1/2)^2/2 - \breve{a}(\theta)(x - 1/2) - \breve{b}(\theta)\Big)\Big],$$

what finally entails that the functions corresponding to (5.15) would be

$$\breve{H}_k(\theta) \;=\; \int_0^1 (x-1/2)^k \exp[\theta(x-1/2)]\,\mathrm{d}x.$$

Here, the integrand is a function that is centered around the midpoint 1/2 of the interval under consideration, and one can expect that the corresponding functions a and b exhibit "nicer" symmetry properties.

5.4 Analysis of local test statistic

We will now analyze the specific score test statistic introduced in the previous section.

To assess whether $\log f^*_{\theta_n,\eta_n}$ introduced in (5.11) is concave or convex on $[0,1]$, i.e. to test whether η_n, the coefficient of the quadratic part, is equal to or significantly different from 0, we will use, based on the arguments in the previous section, the following standardized score test statistic:

$$T_n(\boldsymbol{X}_n,\theta) \;:=\; n^{-1/2}\sum_{i=1}^{n}\frac{T_\theta(X_{in})}{[\mathrm{Var}_\theta\, T_\theta(X_{1n})]^{1/2}}$$

where we abbreviated

$$\mathrm{Var}_\theta\, T_\theta(X_{1n}) \;=\; \mathrm{Var}_{\theta,0}\, T_\theta(X_{1n})$$

as η will always be set to $\eta_o = 0$. Recall that the parameters of $f^*_{\theta_n,\eta_n}$ form sequences converging to θ_o and 0, i.e.

$$\theta_n - \theta_o = o(1) \text{ and } \eta_n = o(1). \tag{5.16}$$

Theorem 5.4.1. *Suppose that the elements of $\boldsymbol{X}_n := (X_{1n},\ldots,X_{nn})$ are i.i.d. distributed having density function $f^*_{\theta_n,\eta_n}$. Then, as $n\to\infty$:*

$$T_n(\boldsymbol{X}_n,\widehat{\theta}^0_n) \to_{\mathcal{D}} \begin{cases} \infty & \text{if } \sqrt{n}|\eta_n| \to \infty \\ \mathcal{N}\Big((\mathrm{Var}_{\theta_o} X_\infty)^{1/2}h, 1\Big) & \text{if } \sqrt{n}|\eta_n| \to h \\ \mathcal{N}(0,1) & \text{if } \sqrt{n}|\eta_n| \to 0. \end{cases}$$

Log-concavity or log-convexity of $f^*_{\theta_n,\eta_n}$ at a given siginificance level α will then be claimed if

$$\begin{aligned} T_n(\boldsymbol{X}_n, \widehat{\theta}^0_n) &\leq -z_{1-\alpha/2} \quad \text{and} \\ T_n(\boldsymbol{X}_n, \widehat{\theta}^0_n) &\geq z_{1-\alpha/2}, \end{aligned}$$

respectively. Theorem 5.4.1 delivers the justification for the use of $\widehat{\theta}^0_n$ as a plug-in estimator for the test statistic $T_n(\boldsymbol{X}_n, \theta)$.

5.5 (Log-)Density function approximated by local parabolas

Throughout the remainder of this chapter, we apply a setting similar to that in Dümbgen and Walther (2006). Suppose $Y_1 < \ldots < Y_m$ are ordered i.i.d. random variables with unknown distribution function F and density f on the real line. Assume that f is twice continuously differentiable on $\{f > 0\}$ and that this latter set is open. Sometimes it is a priori known that F is concentrated on an interval $[a, \infty), (-\infty, b]$ or $[a, b]$ where $-\infty < a < b < \infty$. If this is the case we add the point(s) $Y_0 := a$ or $Y_m =: b$ or both to our ordered sample, yielding an ordered data vector $X_0, \ldots, X_n$ where $n \in \{m-2, m-1, m\}$. For $0 \leq j < k \leq n+1$ with $k - j > 1$, the conditional joint distribution of $X_{j+1}, \ldots, X_{k-1}$, given the interval endpoints X_j and X_k, coincides with the joint distribution of the order statistics of $k - j - 1$ independent random variables with density

$$f_{jk}(x) \quad = \quad \frac{f(x)}{F(X_k) - F(X_j)} 1_{\{x \in \mathcal{I}_{jk}\}}$$

for intervals $\mathcal{I}_{jk} := (X_j, X_k)$. Rescaling the observations finally yields local order statistics:

$$X_{i:j,k} \quad := \quad \frac{X_i - X_j}{X_k - X_j}, \quad j \leq i \leq k.$$

Commonly, to "hunt bumps" means to identify such intervals $\mathcal{I}_{jk}$ where the density f is either convex or concave. However, our focus here is on log-concavity and -convexity. Beneath better mathematical tractability observe that by taking the logarithm non-concave densities with only one bump, e.g. the gaussian density, become purely concave, i.e. the whole line is a "bump region". Up to type 1 errors no spurious dips are then detected.

In this section we will describe how the log-density can locally be approximated by the parametric model in Section 5.3, implying local tests. The collection of all these tests on all intervals $\mathcal{I}_{jk}$ will then be used for multiscale testing in Section 5.6. Introduce two sequences of indices $j = j(n)$, $k = k(n)$ such that

$$j/n \to \gamma \text{ and } k/n \to \gamma \text{ while } k - j \to \infty \tag{5.17}$$

where $\gamma \in (0,1)$ determines the corresponding quantile x_γ, since $X_j \to_{\mathrm{p}} x_\gamma$ and $X_k \to_{\mathrm{p}} x_\gamma$ when $n \to \infty$.
By Taylor approximation we can write the log-density φ for any X_j, $j = 1, \ldots, n$ and $h \in \mathbb{R}$ as follows:

$$\varphi(X_j + h) = \varphi(X_j) + \varphi'(X_j)h + \varphi''(X_j)h^2/2 + r_j(h)h^2.$$

As φ is continuous (even twice differentiable) we have for the remainder

$$\|r_j\|_\infty^{[-\delta,\delta]} \to_{\mathrm{p}} 0$$

when $\delta \to 0$ (and $n \to \infty$, since $X_j \to_{\mathrm{p}} x_\gamma$). Using this, write f_{jk} as follows

$$\begin{aligned} f_{jk}(u) &= \frac{f(X_j + u\delta_{jk})}{\int_0^1 f(X_j + v\delta_{jk})\,\mathrm{d}v} 1_{\{u\in[0,1]\}} \\ &= \frac{\exp\varphi(X_j + u\delta_{jk})}{\int_0^1 \exp\varphi(X_j + v\delta_{jk})\,\mathrm{d}v} 1_{\{u\in[0,1]\}} \\ &= \frac{\exp\Big(h_{jk}(u) + r_j(u\delta_{jk})\delta_{jk}^2\Big)}{\int_0^1 \exp\Big(h_{jk}(v) + r_j(v\delta_{jk})\delta_{jk}^2\Big)\,\mathrm{d}v} 1_{\{u\in[0,1]\}} \end{aligned}$$

where we introduced

$$h_{jk}(x) = \varphi'(X_j)\delta_{jk}x + \varphi''(X_j)\delta_{jk}^2x^2/2 \tag{5.18}$$

for $x \in [0,1]$ and $\delta_{jk} = X_k - X_j$. Clearly,

$$\sup_{u\in[0,1]} |r_j(u\delta_{jk})| \to_{\mathrm{p}} 0 \tag{5.19}$$

as $n \to \infty$. Note that we normalize in order to get a density function on $[0,1]$. Additionally let

$$g_{jk}(u) = \frac{\exp h_{jk}(u)}{\int_0^1 \exp h_{jk}(v)\,\mathrm{d}v} 1_{\{u\in[0,1]\}}.$$

From (5.18) one can conclude that on an interval $\mathcal{I}_{jk}$, the parameters θ_n and η_n introduced in Section 5.3 are in detail, as $n \to \infty$:

$$\theta_n = \frac{\varphi'(x_\gamma)}{f(x_\gamma)} \frac{k-j}{n+1}(1 + o_{\mathrm{p}}(1)) \tag{5.20}$$

$$\eta_n = \frac{\varphi''(x_\gamma)}{2f(x_\gamma)^2} \Big(\frac{k-j}{n+1}\Big)^2 (1 + o_{\mathrm{p}}(1)), \tag{5.21}$$

since, according to the proof of Lemma 5.5.1,

$$\delta_{jk} = \frac{k-j}{n+1} f(x_\gamma)^{-1}(1 + o_{\mathrm{p}}(1)).$$

To give a legitimation for an approximation of a smooth enough log-density by a parabola, consider the total variation distance TV between two probability densities $f : \mathbb{R}^p \to \mathbb{R}$ and $g : \mathbb{R}^p \to \mathbb{R}$. For $\boldsymbol{x} \in \mathbb{R}^p$ define

$$\mathrm{TV}(f, g) := \int_{\mathbb{R}^p} |f(\boldsymbol{x}) - g(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x}.$$

Introduce the following joint densities:

$$\boldsymbol{f}_n(\boldsymbol{X}) := \prod_{i=j+1}^{k-1} f_{jk}(X_i)$$

$$\boldsymbol{g}_n(\boldsymbol{X}) := \prod_{i=j+1}^{k-1} g_{jk}(X_i).$$

The following lemma then specifies the asymptotic total variation distance between $\boldsymbol{f}_n(\boldsymbol{X})$ and $\boldsymbol{g}_n(\boldsymbol{X})$.

Lemma 5.5.1. *For $\boldsymbol{f}_n(\boldsymbol{X})$ and $\boldsymbol{g}_n(\boldsymbol{X})$ introduced above:*

$$\mathrm{TV}\Big(\boldsymbol{f}_n(\boldsymbol{X}), \boldsymbol{g}_n(\boldsymbol{X})\Big) = o_{\mathrm{p}}(1)$$

as $n \to \infty$.

Suppose we would like to test the hypothesis $H_o : \eta = 0$ vs. $H_1 : \eta = \eta_n > 0$. The above lemma implies, that the asymptotic power based on an i.i.d. sample of size $k - j - 1$ taken from f_{jk} is equal to the power for the same testing problem if we adopted a sample from g_{jk} instead. To be fully prepared for the statement of the theorem, introduce a so-called "perfect sequence of tests". A sequence of tests in the above hypothesis is called perfect, if for any sequence of alternatives η_n the power function $\pi_n(\eta_n)$ is tending to 1 and the size $\pi_n(\eta_o) = \pi_n(0)$ is tending to 0, as $n \to \infty$.

Theorem 5.5.2. *Suppose $\varphi''(x_\gamma) > 0$ and the sequences $j = j(n)$ and $k = k(n)$ are such that*

$$n^{1/5}\Big(\frac{k-j-1}{n}\Big) \;\to\; \infty \tag{5.22}$$

as $n \to \infty$. Then there exists a perfect sequence of tests for the hypothesis $H_o : \eta = 0$ vs. $H_1 : \eta = \eta_n > 0$ based on an i.i.d. sample of size $k-j-1$ where every random variable in the sample has density function f_{jk}.

To conclude, some words about the Condition (5.22). It seems not to be too stringent, since Definition (5.21) of η_n suggests that in order to be able to test for this latter parameter we anyway need enough observations in $\mathcal{I}_{jk}$ to guarantee

$$(k-j-1)^{1/2}\Big(\frac{k-j-1}{n}\Big)^2 \to \infty.$$

But this latter condition is equivalent to (5.22).

5.6 The multiscale test

Having guaranteed sufficient power in Section 5.1, shown convenient properties of the local test statistic $T_n(\boldsymbol{X}_n, \theta)$ in Section 5.4 and justified approximation of the original density f on any interval $\mathcal{I}_{jk}$ through local parabolas in Section 5.5, we will now introduce a multiscale test.
Beneath in Dümbgen and Walther (2006) for mode hunting, multiscale testing in a quite general qualitative setting is described in Dümbgen and Spokoiny (2001) and in a more applied regression framework in Dümbgen (2002).
Adopting the notation of the latter paper, define the global test statistic for a sample $\boldsymbol{X}_n$, $3 \le m \le n-l$ and $3 \le l \le m-1$ as

$$T^*_{l,m,n}(\boldsymbol{X}) \;:=\; \max_{1\le j<k\le n,\;\; l\le k-j\le m}\Big(|T_{jkn}(\boldsymbol{X}, \widehat{\theta}^0_{jk})| - c_{k-j}\Big)$$

where $\widehat{\theta}^0_{jk}$ is the estimated log-linearity parameter θ_{jk} based on the local order statistics $X_{j+1:j,k}, \ldots, X_{k-1;j,k}$ where η_{jk} is assumed to be 0, i.e. estimation of θ_{jk} happens under the null hypothesis. The local test statistics are

$$T_{jkn}(\boldsymbol{X}_n, \theta) \;:=\; \frac{\sum_{i=j+1}^{k-1} T_\theta(X_{i:j,k})}{[(k-j-1)\operatorname{Var}_\theta T_\theta(X_{j+1:j,k})]^{1/2}}$$

and the normalizing constants

$$c_d \;:=\; \Big(2 + 2\log(n/d)\Big)^{1/2}.$$

The papers cited above detail why constants of this type are appropriate in such a multiscale setting. Informally, such an additive correction is introduced to prevent the limiting distribution of $T^*_{l,m,n}$ to be dominated by local statistics T_{jkn} for $(k-j)/n$ small, i.e. those on short intervals.
The test function $T_{jkn}(\boldsymbol{X}, \theta)$ can alternatively be written as

$$\begin{aligned} T_{jkn}(\boldsymbol{X}_n, \theta) \;&:=\; \frac{\sum_{i=j+1}^{k-1}\Big([X_{i:j,k} - a(\theta)]^2/2 - a(\theta)^2/2 - b(\theta)\Big)}{[(k-j-1)\operatorname{Var}_\theta T_\theta(X_{j+1:j,k})]^{1/2}} \\ &:=\; \frac{\sum_{i=j+1}^{k-1}\Big(\delta_\theta(X_{i:j,k}) - a(\theta)^2/2 - b(\theta)\Big)}{[(k-j-1)\operatorname{Var}_\theta T_\theta(X_{j+1:j,k})]^{1/2}} \end{aligned} \tag{5.23}$$

where $\delta_\theta(x) := 2^{-1}(x - a(\theta))^2 1_{\{x\in[0,1]\}}$. If we plug in an estimator $\widehat{\theta}_n$ for θ, (5.23) means that our test functions are parabolas with an estimated (and therefore somehow adaptive) vertex

$$\Big(a(\widehat{\theta}_n), -a(\widehat{\theta}_n)^2/2 - b(\widehat{\theta}_n)\Big).$$

Clearly, this estimator $\widehat{\theta}_n$ will be $\widehat{\theta}^0_{jk}$. According to (5.12) the test functions δ_θ are indifferent with regard to linear density functions. However, if the observations come from a local log-density function $\log f_{jk}$ that is convex or concave, then $T_{jkn}(\boldsymbol{X}_n, \widehat{\theta}^0_{jk})$ tends to be highly positive or negative, respectively, by Theorem 5.2.2. It is important to note that other test functions are equally possible, e.g. parabolas with a fixed vertex, immediately raising further possibilities to design tests for (log-) concavity or (log-) convexity.
As in Dümbgen (2002), we confine our attention in the definition of $T^*_{l,m,n}$ to pairs (j,k) such that their maximal lag $k-j$ is smaller than m (typically we will choose $m < n$, e.g. $m = n/2$), for two reasons. First, to reduce computational burden in numerical simulations and calculations of the test statistic and second because we want to increase sensitivity on smaller intervals. Similarly, only lags $l \geq 3$ are considered, because this is the minimal number of observations to assess concavity or convexity meaningfully.

Suppose we somehow get hold of the distribution of $T^*_{l,m,n}$ as $n \to \infty$ (for details see Section 5.7), define $\kappa(\alpha, f, n)$ as the $(1-\alpha)$-quantile of this distribution. As we do not know the precise limiting behavior of the distribution of $T^*_{l,m,n}$ and therefore the quantiles of it, we make the following working assumption.

Working assumption 5.6.1. *Suppose for the quantile $\kappa(\alpha, f, n)$ that as $n \to \infty$*

$$\kappa(\alpha, f, n) = \kappa(\alpha, g_o) + o(1)$$

for some "null density" g_o and that this latter quantile $\kappa(\alpha, g_o)$ is bounded.

Some indications that this working assumption may hold true are given in Section 5.7.

Now fix α, l, m and n. For a given sample $\boldsymbol{X}_n$, generate the distribution of $T^*_{l,m,n}$ and calculate $\kappa(\alpha) = \kappa(\alpha, f, n)$. Then introduce the following collections of intervals:

$$\begin{aligned}
\mathcal{C}^{\cap}_{l,m,n}(\alpha) &:= \{[X_j, X_k] : 0 \le j < k \le n,\ k-j \le m,\ -T_{jkn}(\boldsymbol{X}_n, \widehat{\theta}_{jk}) > c_{k-j} + \kappa(\alpha)\} \\
\mathcal{C}^{\cup}_{l,m,n}(\alpha) &:= \{[X_j, X_k] : 0 \le j < k \le n,\ k-j \le m,\ T_{jkn}(\boldsymbol{X}_n, \widehat{\theta}_{jk}) > c_{k-j} + \kappa(\alpha)\}.
\end{aligned}$$

With probability at least $1-\alpha$ the following statement holds asymptotically as n tends to infinity. The logarithm of the true density function f is neither concave on any interval in $\mathcal{C}^{\cup}_{l,m,n}(\alpha)$ nor convex on any interval in $\mathcal{C}^{\cap}_{l,m,n}(\alpha)$. Even further, the local score tests imply a lower confidence bound for the location and number of these pieces. Define the sets of bump intervals as follows: If both sets $\mathcal{C}^{\cap}_{l,m,n}(\alpha)$ and $\mathcal{C}^{\cup}_{l,m,n}(\alpha)$ are non-empty, then

$$\begin{aligned}
\mathcal{B}^{\cap}_{l,m,n}(\alpha) &:= \{[x, y'] : [x, y] \in \mathcal{C}^{\cap}_{l,m,n}(\alpha),\ [x', y'] \in \mathcal{C}^{\cup}_{l,m,n}(\alpha),\ y \le x'\} \cup \mathcal{C}^{\cap}_{l,m,n}(\alpha) \\
\mathcal{B}^{\cup}_{l,m,n}(\alpha) &:= \{[x, y'] : [x, y] \in \mathcal{C}^{\cup}_{l,m,n}(\alpha),\ [x', y'] \in \mathcal{C}^{\cap}_{l,m,n}(\alpha),\ y \le x'\} \cup \mathcal{C}^{\cup}_{l,m,n}(\alpha),
\end{aligned}$$

if $\mathcal{C}^{\cap}_{l,m,n}(\alpha) = \varnothing$, set $\mathcal{B}^{\cap}_{l,m,n}(\alpha) = \varnothing$ and let $\mathcal{B}^{\cup}_{l,m,n}(\alpha)$ only contain the first element of $\mathcal{C}^{\cup}_{l,m,n}(\alpha)$ and likewise if $\mathcal{C}^{\cup}_{l,m,n}(\alpha)$ is empty. Post-process the sets $\mathcal{B}^{\cap}_{l,m,n}(\alpha)$ and $\mathcal{B}^{\cup}_{l,m,n}(\alpha)$ as follows. Take the left-most interval endpoint X_q in the set, keep only the longest interval $[X_q, X_r]$ with this starting point and skip all other intervals that are not disjoint with $[X_q, X_r]$. Then continue with the left-most interval endpoint right of X_r and do this until no intervals can be kept anymore.

The sets $\mathcal{B}^{\cup}_{l,m,n}(\alpha)$ and $\mathcal{B}^{\cap}_{l,m,n}(\alpha)$ consist of intervals J which do contain separated (in the sense that they are only allowed to adjoin at one point) regions J_1, J_2 where $\log f$ exhibits both a concave and a convex behavior. Assembly above considerations to conclude the following theorem.

Theorem 5.6.2. *Suppose the Working Assumption 5.6.1 holds true. With probability at least* $1-\alpha$ *as* n *tends to infinity* $\log f$ *is neither concave on* $\mathcal{C}^{\cup}_{l,m,n}(\alpha)$ *nor convex on* $\mathcal{C}^{\cap}_{l,m,n}(\alpha)$*. Furthermore, the number of bumps of* $\log f$ *is not smaller than the number of intervals in* $\mathcal{B}^{\cap}_{l,m,n}(\alpha)$*. On the other hand,* $\log f$ *has at least as many dips as there are intervals in* $\mathcal{B}^{\cup}_{l,m,n}(\alpha)$*.*

It is important to note that it is principally not possible to replace the one-sided statement in Theorem 5.6.2 by a two-sided version. This impossibility is a fundamental property of truly nonparametric functionals of a density f, such as the number of bumps and the number of dips in our case and is elaborated in Donoho (1988).

5.7 The limiting distribution of $T^*_{l,m,n}$

To start the section, let us introduce three distributional laws in Table 5.1.

Table 5.1: Distribution laws used to assess $\mathcal{L}(T^*_{l,m,n})$.

Law	Symbol	Density	Range	Parameters
Exponential(λ)	$\mathcal{E}$	$\lambda \exp(-\lambda z)$	$[0,\infty)$	$\lambda > 0$
Log-linear(θ)	$\overline{\mathcal{E}}$	$\theta \exp(\theta z)/(\exp(\theta)-1)$	$[0,1]$	$\theta \in \mathbb{R}$
Uniform	$\mathcal{U}$	1	$[0,1]$	

With $\boldsymbol{\mathcal{E}}_n, \overline{\boldsymbol{\mathcal{E}}}_n$ and $\boldsymbol{\mathcal{U}}_n$ we mean vectors consisting of n i.i.d. random variables of the given type.

For a fixed n, $T^*_{l,m,n}(\boldsymbol{X}_n)$ is constructed as the maximum over all lags greater than 3 and smaller than m minus the correction c_d, therefore it is not evident whether the limiting distribution as $n \to \infty$, denoted by $\mathcal{L}(T^*(\boldsymbol{X}_n))$, exists, if yes whether it is non-degenerate and finally how it depends on $\boldsymbol{X}_n$. However, in view of the results in Dümbgen and Spokoiny (2001, Theorem 2.1.) it would be of great surprise if the answer to the first two questions is not affirmative. This conjecture is further supported by numerical simulations, clearly pointing to the existence of a limiting distribution $\mathcal{L}(T^*(\boldsymbol{X}_n))$, see Figure 5.2. We sampled 9'999 statistics $T^*_{3,m,n}(\boldsymbol{\mathcal{E}}_n)$ for every combination of m and n detailed in the legend of the figure, where $m = n-l-1$ for $n \leq 200$ and $m = \lfloor n/2 \rfloor - l - 1$ for $n \geq 200$.

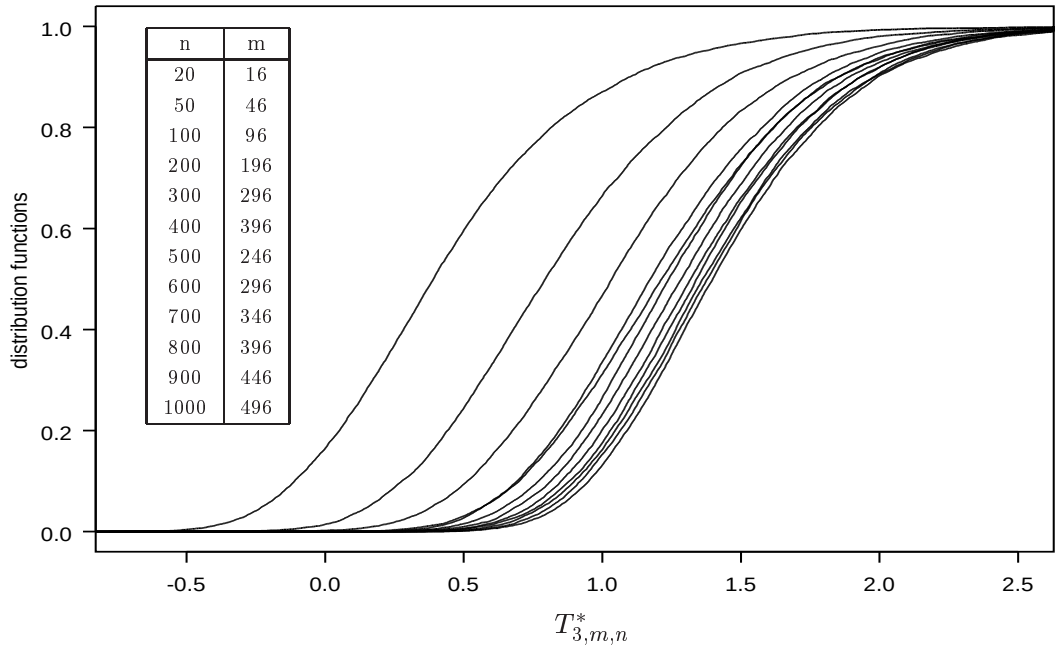

Figure 5.2: Limiting distribution functions for $T^*_{3,m,n}(\boldsymbol{\mathcal{E}}_n)$. The curves are generated from left to right with the parameters in the legend top down.

Having postulated the existence of $\mathcal{L}(T^*(\boldsymbol{X}_n))$, it is however recommendable in applications for a given fixed n to rely on Monte-Carlo simulations to generate the distributions $\mathcal{L}(T^*_{l,m,n}(\boldsymbol{X}_n))$ yielding the quantiles $\kappa(\alpha)$. The problem then is what distribution to choose where $\boldsymbol{X}_n$ is sampled from. We experimented with the distributions detailed in Table 5.1.
Numerical simulations suggest that using vectors $\boldsymbol{\mathcal{E}}_n$ yields test statistics $T^*_{3,m,n}(\boldsymbol{\mathcal{E}}_n)$ whose distributions are stochastically bigger than all other input distributions we tried, i.e.

$$F_{\boldsymbol{\mathcal{E}}_n}(x) \leq F_{\boldsymbol{\mathcal{D}}_n}(x), \quad \text{for all } x \in \mathbb{R}$$

where $F_{\boldsymbol{\mathcal{E}}_n}(x)$ is the distribution function for a sample of $T^*_{3,m,n}(\boldsymbol{\mathcal{E}}_n)$ generated from the entries of $\boldsymbol{\mathcal{E}}_n$ and $F_{\boldsymbol{\mathcal{D}}_n}(x)$ is the distribution function for a sample of $T^*_{3,m,n}(\boldsymbol{\mathcal{D}}_n)$ where $\boldsymbol{\mathcal{D}}_n \in \{\overline{\boldsymbol{\mathcal{E}}}_n, \boldsymbol{\mathcal{U}}_n\}$. Figure 5.3 details the issue. The horizontal lines are drawn at $1-\alpha \in \{0.9, 0.95, 0.99\}$, i.e. where the most widely used quantiles $\kappa(\alpha)$ are calculated from. One hardly sees any difference between the three curves overall and only minor differences in the tails. Per distribution we sampled 9'999 times the statistic $T^*_{3,46,50}$.

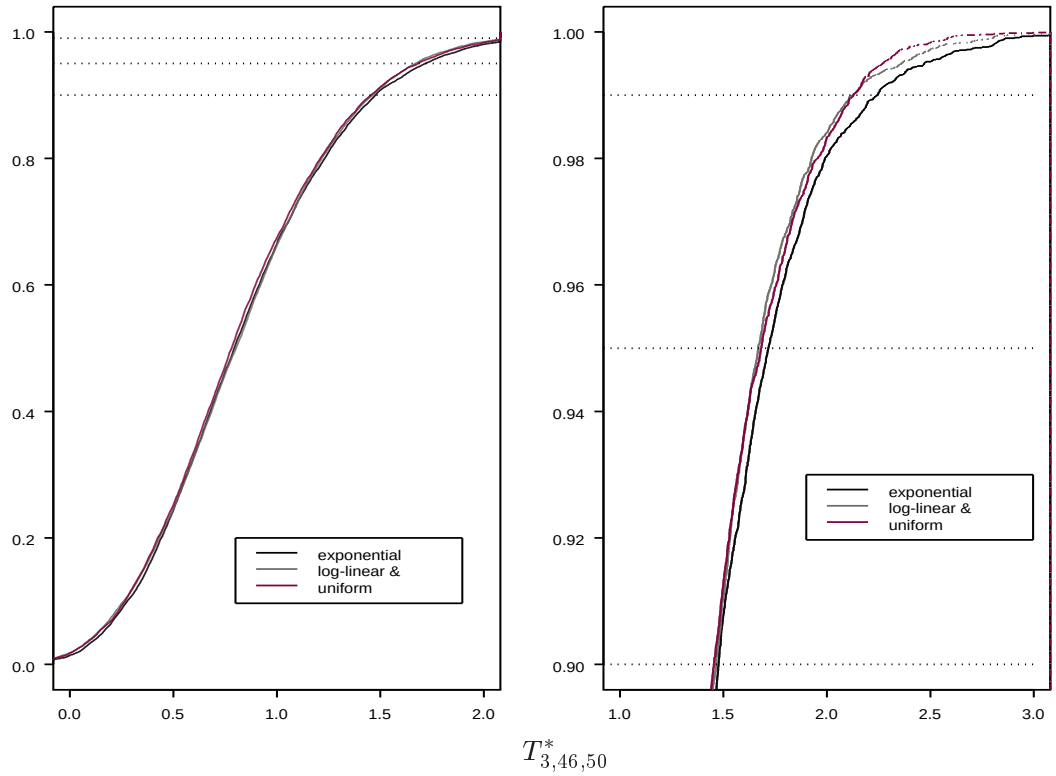

Figure 5.3: Distribution functions for $T^*_{3,46,50}$.

In what follows, we provide a lemma reminiscent of the deterministic inequality of Proposition 1 in Dümbgen and Walther (2006), detailing a vector $\boldsymbol{Y}_n$ having some sort of borderline or worst-case distribution such that the test statistic $T_{jkn}(\boldsymbol{X}_n)$ is bounded from above (log-concave case) or from below (log-convex). However, due to the fact that we have to estimate θ_{jk} we can only provide a weak statement in terms of expectations.

Lemma 5.7.1. *Fix indices j and k where $0 \le j < k \le n$ with $k - j \ge l$. Define the vector $\boldsymbol{Y}_n = (Y_i)_{i=j+1}^{k-1}$ of i.i.d. random variables such that every component Y_i has a log-linear density function g_{jk,θ_1} where*

$$g_{jk,\theta}(x) \;:=\; \frac{\theta}{\exp(\theta) - 1} \exp(\theta x) 1_{\{x \in \mathcal{I}_{jk}\}}.$$

Then:

$$\mathbb{E}_{\theta_{jk},\eta_{jk}} T_{jkn}(\boldsymbol{X}_n, \theta_{jk}) \begin{cases} \le \mathbb{E}_{\theta_{jk},0} T_{jkn}(\boldsymbol{Y}_n, \theta_1) & \text{if } f \text{ is log-concave on } \mathcal{I}_{jk}, \\ \ge \mathbb{E}_{\theta_{jk},0} T_{jkn}(\boldsymbol{Y}_n, \theta_1) & \text{if } f \text{ is log-convex on } \mathcal{I}_{jk} \end{cases}$$

as $n \to \infty$ for all $\theta_1 \le \theta_{jk}$ where θ_{jk} and η_{jk} are the parameters of the density f^*_{jk}.

This lemma suggests an optimal strategy to sample from the distribution $\mathcal{L}(T^*_{l,m,n}(\boldsymbol{X}_n))$. On every interval $\mathcal{I}_{jk}$ estimate θ_{jk}, then generate a random vector with components having density g_{jk,θ_1} for a θ_1 such that $\theta_1 < \widehat{\theta}_{jk}$ and use the distribution of $T^*_{l,m,n}$ generated by M such random vectors to get critical values $\kappa(\alpha)$ of $\mathcal{L}(T^*_{l,m,n}(\boldsymbol{X}_n))$. Note that this procedure provides quantiles depending on the actual data $\boldsymbol{X}_n$. Second, the original condition for θ_1 is to be smaller than the true θ_{jk}. However, θ_{jk} is unknown and replaced by the maximum likelihood estimator $\widehat{\theta}_{jk}$.

Unfortunately, Lemma 5.7.1 is only a limit result as $n \to \infty$. As long as one estimates θ_{jk}, this cannot be improved in the sense to get a result for finite n. However, one can imagine to choose θ differently, e.g. via some "worst θ" or minimax criterion, perhaps yielding results for finite n. The prize to pay when adopting such a procedure is in terms of power. We have no clue how high the power loss is. As described above, to get quantiles generally applicable we sampled vectors $\boldsymbol{\mathcal{E}}_n$ of exponential random variables, which we considered having some sort of general log-linear distribution. At least their parametric shape is justified by Lemma 5.7.1, however, $\theta_{jk} = 1$ is used for all $0 \le j < k \le n + 1$. In Table 5.2 we provide some quantiles $\kappa(\alpha)$, generated from $M = 9'999$ simulations.

To interpolate (or even extrapolate) for values of n not provided in Table 5.2, we recommend to regress $\log \kappa(\alpha)$ on n (among n's where l and m are selected using the same strategy).

Table 5.2: Quantiles $\kappa(\alpha)$ for the multiscale test.

n	l	m	$\kappa(0.90)$	$\kappa(0.95)$	$\kappa(0.99)$
20	3	16	1.0749	1.3335	1.8969
50	3	46	1.4763	1.7007	2.2029
100	3	96	1.6981	1.9253	2.3875
200	3	196	1.8509	2.0702	2.5418
300	3	146	1.8038	2.0098	2.4722
400	3	196	1.8520	2.0699	2.5129
500	3	246	1.8900	2.1052	2.5320
600	4	296	1.9302	2.1346	2.5453
700	5	346	1.9314	2.1270	2.5719
800	6	396	1.9783	2.1729	2.5709
900	7	446	1.9827	2.1908	2.6192
1000	8	496	1.9921	2.2058	2.6391

5.8 Examples in bump hunting

We illustrate the method described above with some examples with simulated data, performed in R, Version 2.1.1. Distributions we used are detailed in Table 5.3. Figures 5.4-5.7 illustrate the results. All figures are to be read as follows: First, we imposed everywhere $\alpha = 0.05$. Two plots always mate vertically. On the upper one, the straight line is the original density we sampled from whereas the dotted line is the standard gaussian kernel estimate. In the lower plot, the sets $\mathcal{C}^{\cap}_{l,m,n}(0.05)$ (above the horizontal dotted line) and $\mathcal{C}^{\cup}_{l,m,n}(0.05)$ (below the dotted line) are displayed. We intentionally omitted plots of the log-density (whereon the method actually works) in order not to overload the figures.

Table 5.3: Distribution laws to illustrate bump hunting method.

Name	Law	Sample Size n
Normal	$\mathcal{N}(0,1)$	50, 200
Contaminated Normal	$0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(6,1)$	200, 500
Two bumps	$0.5\mathcal{N}(0,1) + 0.5\Gamma(5,2)$	200, 500, 700, 1000
Claw density	$0.5\mathcal{N}(0,1) + \sum_{i=0}^{4}(1/10)\mathcal{N}(i/2-1, 1/100)$	200, 500, 700, 1000

In Figure 5.4 we see two standard normal samples of sizes $n = 50$ and $n = 200$. In both cases, only the set $\mathcal{C}^{\cap}_{l,m,n}(0.05)$ is non-empty, so that we conclude by Theorem 5.6.2 that there is at least one bump. Precisely we have:

$$\begin{aligned}\mathcal{C}^{\cap}_{3,46,50}(0.05) &= \{[X_{(4)}, X_{(44)}]\}\\ \mathcal{C}^{\cap}_{3,196,200}(0.05) &= \{[X_{(1)}, X_{(129)}], [X_{(42)}, X_{(135)}], [X_{(44)}, X_{(146)}], [X_{(45)}, X_{(160)}],\\ &\quad [X_{(46)}, X_{(162)}], [X_{(48)}, X_{(163)}], [X_{(54)}, X_{(196)}]\}\end{aligned}$$

and $\mathcal{C}^{\cup}_{3,46,50}(0.05) = \mathcal{C}^{\cup}_{3,196,200}(0.05) = \varnothing$, yielding $\mathcal{B}^{\cap}_{3,46,50}(0.05) = \mathcal{C}^{\cap}_{3,46,50}(0.05)$ and $\mathcal{B}^{\cap}_{3,196,200}(0.05) = \{[X_{(1)}, X_{(129)}]\}$.

Two samples for $n = 200$ and $n = 500$ of a standard normal distribution corrupted by 10% of observations stemming from another normal distribution are displayed in Figure 5.5, see Table 5.4 displaying the number of clearly ascertained bumps and the sets $B^{\cap}_{l,m,n}(0.05)$ and $B^{\cup}_{l,m,n}(0.05)$.

By Theorem 5.6.2 we conclude with the level of the test tending to 0.05 as $n \to \infty$ that we have at least two bumps in the sample of size $n = 500$. Compared to the purely normal distribution we can claim that there must be something different going on here.

A mixture density with two bumps appears in Figure 5.6. Note that the density is constructed such that it has only one mode but two bumps, this being the most specific situation to apply bump hunting compared to mode hunting. The results are given in Table 5.5.

Table 5.4: Results for the contaminated normal density.

n	bumps	dips	$B^{\cap}_{l,m,n}(0.05)$	$B^{\cup}_{l,m,n}(0.05)$
200	1	1	$[X_{(7)}, X_{(182)}]$	$[X_{(154)}, X_{(182)}]$
500	2	1	$[X_{(96)}, X_{(455)}]$	$[X_{(134)}, X_{(447)}]$
			$[X_{(460)}, X_{(499)}]$	

Table 5.5: Results for the two bumps density.

n	bumps	dips	$B^{\cap}_{l,m,n}(0.05)$	$B^{\cup}_{l,m,n}(0.05)$
200	1	1	$[X_{(1)}, X_{(164)}]$	$[X_{(59)}, X_{(151)}]$
500	2	1	$[X_{(3)}, X_{(336)}]$	$[X_{(159)}, X_{(488)}]$
			$[X_{(338)}, X_{(488)}]$	
700	2	1	$[X_{(5)}, X_{(417)}]$	$[X_{(319)}, X_{(671)}]$
			$[X_{(480)}, X_{(671)}]$	
1000	2	1	$[X_{(3)}, X_{(674)}]$	$[X_{(326)}, X_{(970)}]$
			$[X_{(725)}, X_{(970)}]$	

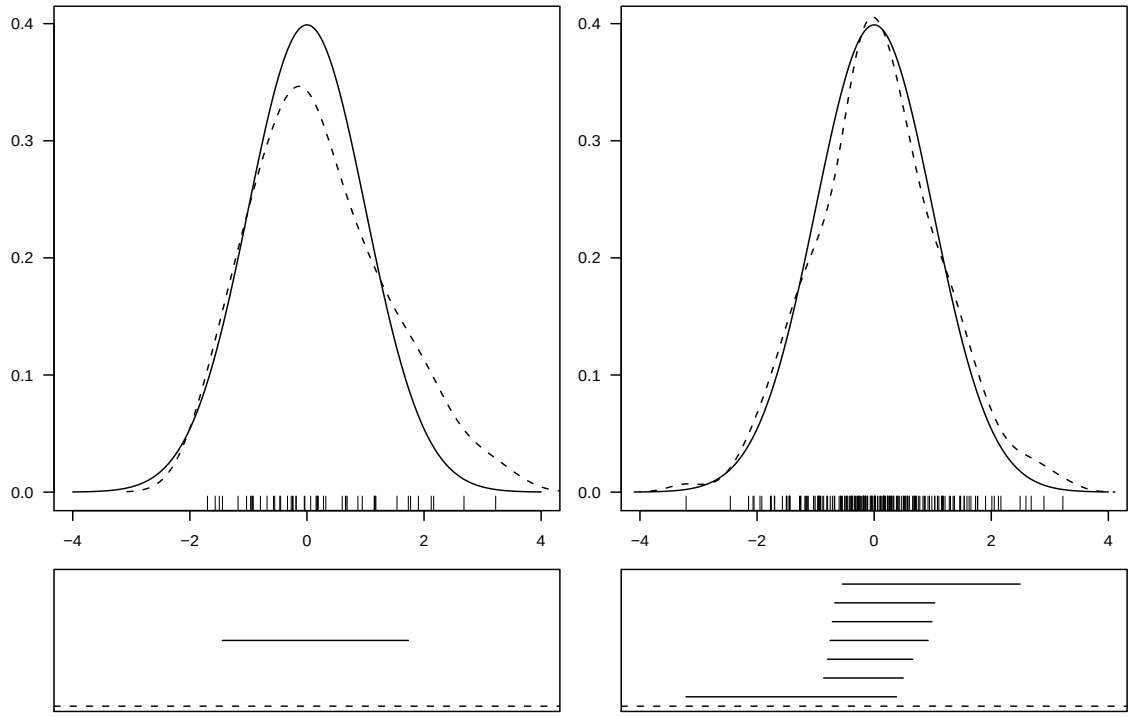

Figure 5.4: Multiscale test results for normal samples for $n = 20$ and $n = 200$.

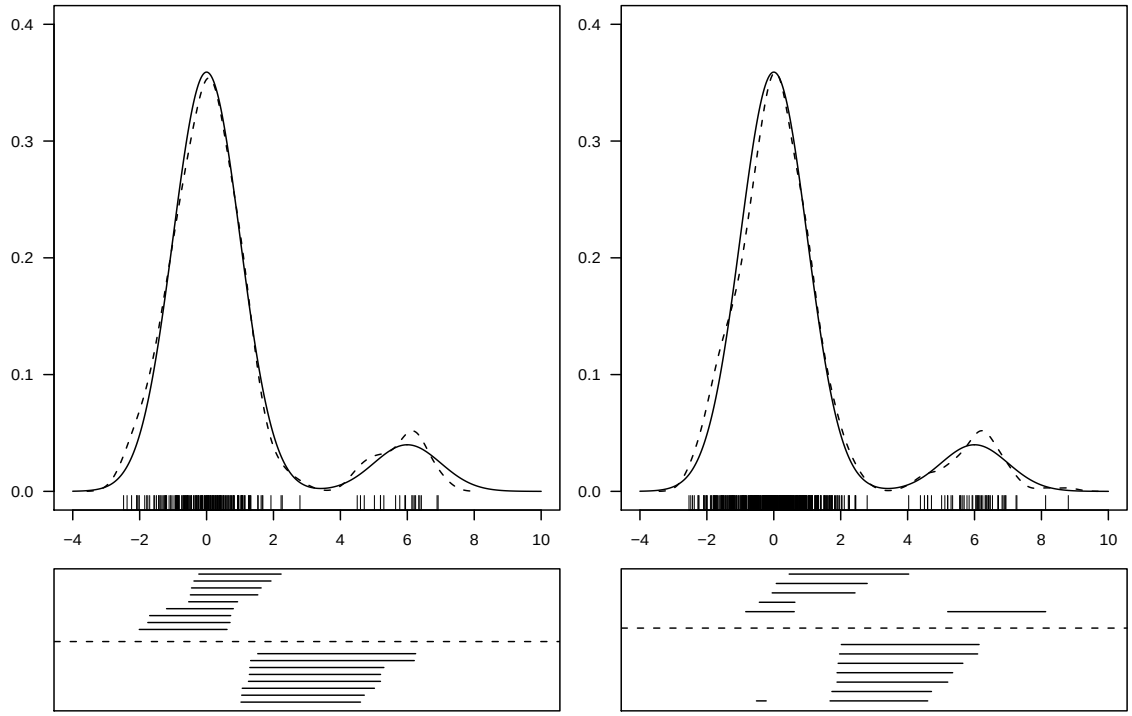

Figure 5.5: Multiscale test results for normal contaminated samples for $n = 200$ and $n = 500$.

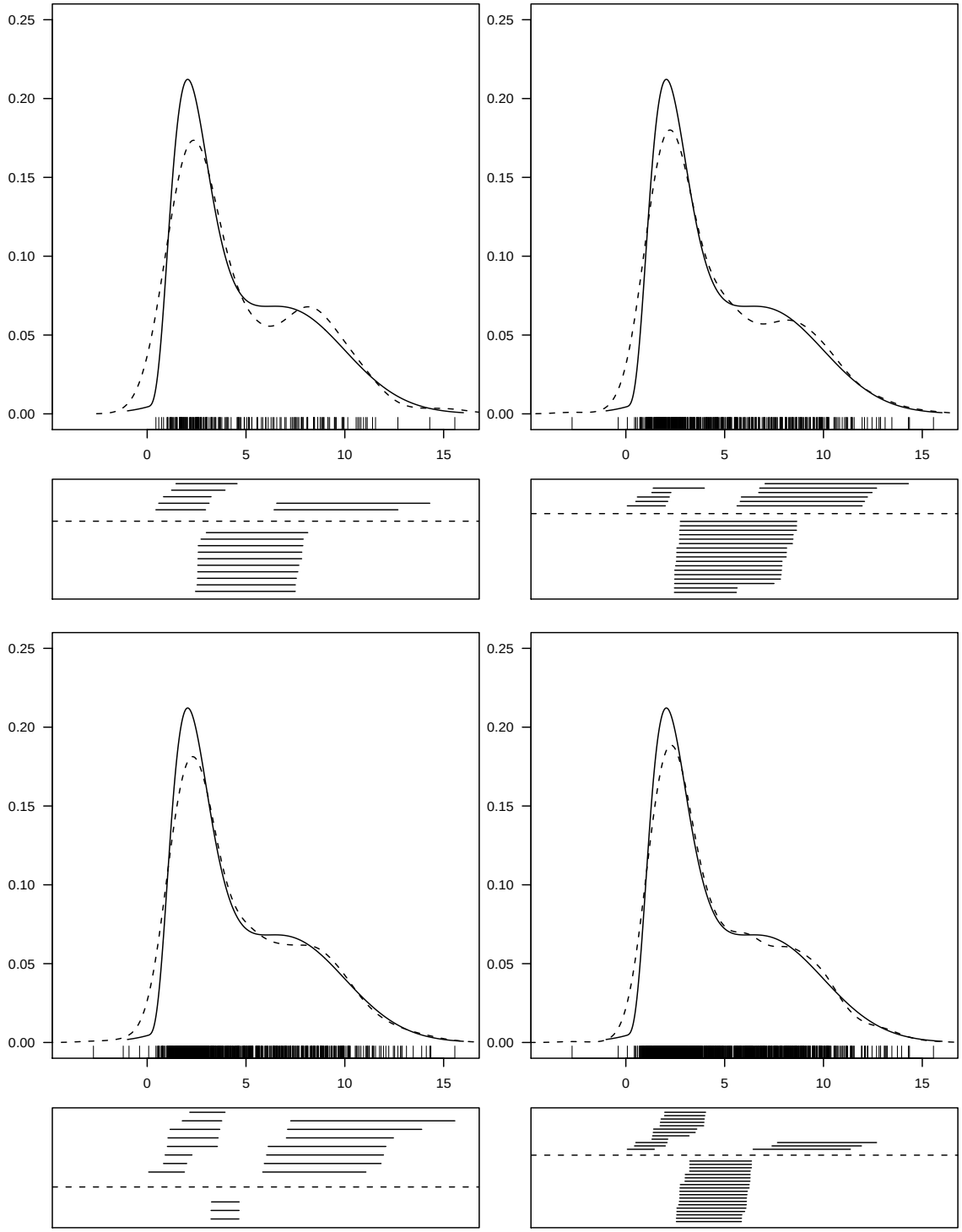

Figure 5.6: Test results for a two bumps sample for $n = 200, 500, 700$ and $n = 1000$.

As an illustration how the power of our multiscale test evolves when n increases, Figure 5.7 displays a mixture of six normal distributions, the so-called Claw density, introduced by Marron and Wand (1992). Here modes and bumps are the same. Table 5.6 details the results. Clearly, the method is at a sample size of $n = 1000$ not able to detect the bumps in the statistically strict sense of Theorem 5.6.2. However, looking at Figure 5.7 in a more explorative manner, one already has clear indications at a sample size of $n = 500$ that there might be five bumps present, because we have alternating intervals whereon we claim log-concavity and log-convexity, but the intervals still overlap.

Table 5.6: Results for the Claw density.

n	bumps	dips	$B^{\cap}_{l,m,n}(0.05)$	$B^{\cup}_{l,m,n}(0.05)$
200	1	1	$[X_{(9)}, X_{(185)}]$	$[X_{(136)}, X_{(185)}]$
500	2	2	$[X_{(33)}, X_{(134)}]$	$[X_{(94)}, X_{(299)}]$
			$[X_{(209)}, X_{(439)}]$	$[X_{(350)}, X_{(439)}]$
			$[X_{(48)}, X_{(182)}]$	$[X_{(4)}, X_{(244)}]$
700	3	3	$[X_{(291)}, X_{(476)}]$	$[X_{(252)}, X_{(521)}]$
			$[X_{(566)}, X_{(671)}]$	$[X_{(521)}, X_{(626)}]$
			$[X_{(67)}, X_{(256)}]$	$[X_{(5)}, X_{(349)}]$
1000	3	3	$[X_{(420)}, X_{(635)}]$	$[X_{(356)}, X_{(605)}]$
			$[X_{(799)}, X_{(980)}]$	$[X_{(705)}, X_{(855)}]$

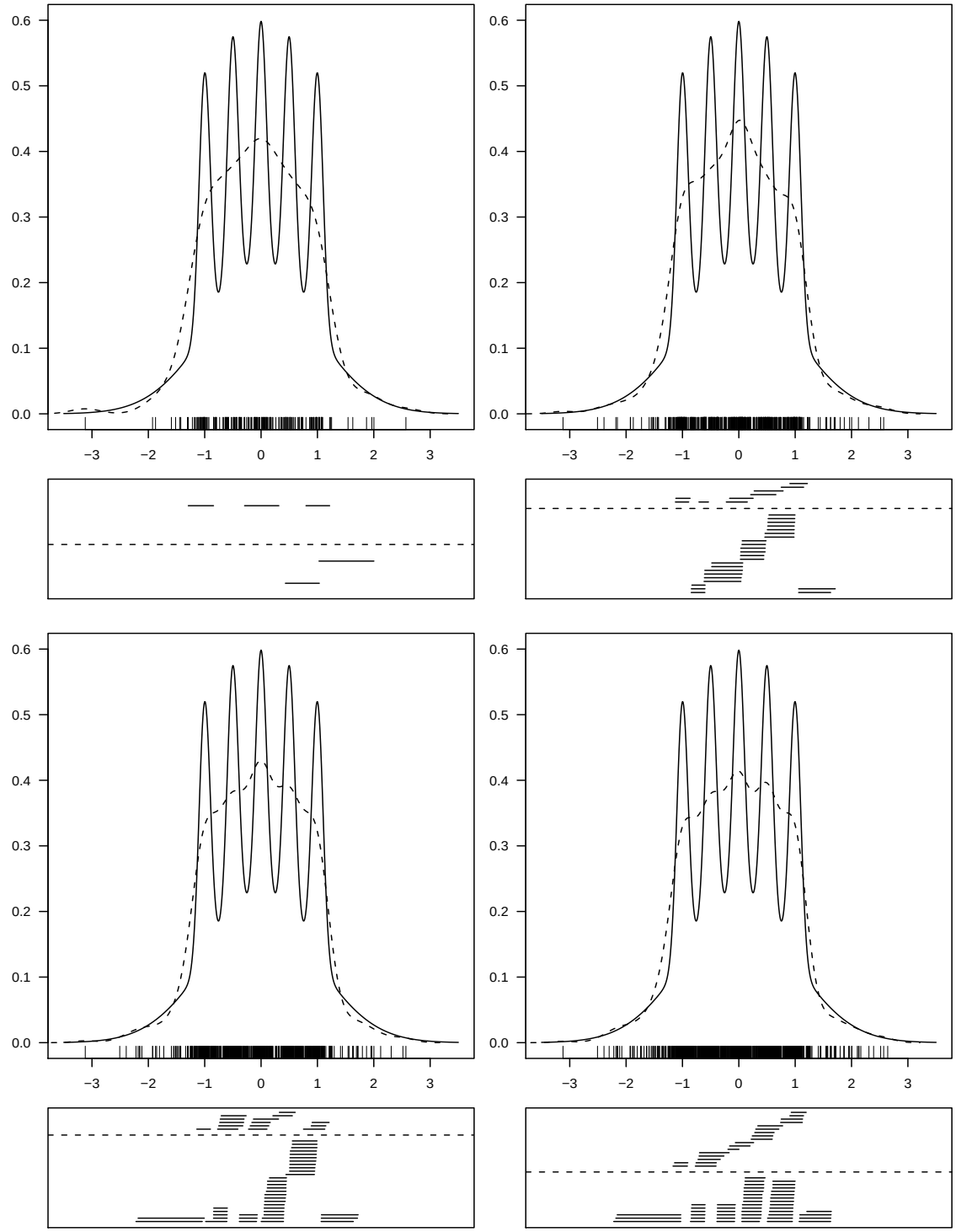

Figure 5.7: Multiscale test results for a Claw sample for $n = 200, 500, 700$ and $n = 1000$.

5.9 PROOFS

Proof of Lemma 5.1.1: Lehmann (1986, p. 59) gives the first statement of the lemma in an even more general form, including a proof. As for the Laplace transform,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}} \exp\Big(\boldsymbol{u}^\top \boldsymbol{t}(X)\Big) &= \int_{\mathcal{X}} \exp[\boldsymbol{u}^\top \boldsymbol{t}(x)] c(\boldsymbol{\theta}) h(x) \exp[\boldsymbol{\theta}^\top \boldsymbol{t}(x)] \,\mathrm{d}x \\
&= c(\boldsymbol{\theta}) \int_{\mathcal{X}} h(x) \exp[(\boldsymbol{\theta}+\boldsymbol{u})^\top \boldsymbol{t}(x)] \,\mathrm{d}x \\
&= \frac{c(\boldsymbol{\theta})}{c(\boldsymbol{\theta}+\boldsymbol{u})}. \qquad \square
\end{aligned}$$

Proof of Theorem 5.1.2: Before attacking directly the difference in (5.6), some preliminary considerations have to be made. First, note that for a random variable X having density function $p_{\boldsymbol{\theta}}$ and vectors $\boldsymbol{\alpha}, \boldsymbol{\delta} \in \mathbb{R}^p$ we have for the function $\boldsymbol{t}$, by Lemma 5.1.1:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\alpha}+\boldsymbol{\delta}} \boldsymbol{t}(X) &= \frac{\int_{\mathcal{X}} h(x) \boldsymbol{t}(x) \exp[(\boldsymbol{\alpha}+\boldsymbol{\delta})^\top \boldsymbol{t}(x)] \,\mathrm{d}x}{\int_{\mathcal{X}} h(x) \exp[(\boldsymbol{\alpha}+\boldsymbol{\delta})^\top \boldsymbol{t}(x)] \,\mathrm{d}x} \\
&= \frac{\mathbb{E}_{\boldsymbol{\alpha}} \boldsymbol{t}(X) + \boldsymbol{\delta}^\top \mathbb{E}_{\boldsymbol{\alpha}}[\boldsymbol{t}(X)\boldsymbol{t}(X)^\top] + O(\|\boldsymbol{\delta}\|_2^2)}{1 + \boldsymbol{\delta}^\top \mathbb{E}_{\boldsymbol{\alpha}} \boldsymbol{t}(X) + O(\|\boldsymbol{\delta}\|_2^2)} \\
&= \Big(\mathbb{E}_{\boldsymbol{\alpha}} \boldsymbol{t}(X) + \boldsymbol{\delta}^\top \mathbb{E}_{\boldsymbol{\alpha}}[\boldsymbol{t}(X)X^\top]\Big)[1 - \boldsymbol{\delta}^\top \mathbb{E}_{\boldsymbol{\alpha}}(X)] + O(\|\boldsymbol{\delta}\|_2^2) \\
&= \mathbb{E}_{\boldsymbol{\alpha}} \boldsymbol{t}(X) + \boldsymbol{\delta}^\top \mathrm{Cov}_{\boldsymbol{\alpha}}(X, \boldsymbol{t}(X)) + O(\|\boldsymbol{\delta}\|_2^2) \qquad (5.24)
\end{aligned}$$

using

$$(1+x)^{-1} = 1 - x + O(x^2)$$

which holds for any $x \neq -1$. Similarly one can derive

$$\mathrm{Var}_{\boldsymbol{\alpha}+\boldsymbol{\delta}} \boldsymbol{t}(X) = \mathrm{Var}_{\boldsymbol{\alpha}} \boldsymbol{t}(X) + O(\|\boldsymbol{\delta}\|_2). \qquad (5.25)$$

Now define the random vectors $\boldsymbol{Z}_{in}$ as

$$\boldsymbol{Z}_{in} := n^{-1/2}\Big(\boldsymbol{t}(X_{in}) - \overline{\boldsymbol{t}(\boldsymbol{X}_n)}\Big)$$

where

$$\overline{\boldsymbol{t}(\boldsymbol{X}_n)} := (1/n) \sum_{i=1}^{n} \boldsymbol{t}(X_{in})$$

is the empirical counterpart of (5.1). By the law of large numbers in Section A.9 and Lemma 5.1.1 we have that

$$\sum_{i=1}^{n} \boldsymbol{Z}_{in} \boldsymbol{Z}_{in}^{\top} \quad \to_{\mathrm{p}} \quad \mathrm{Cov}_{\boldsymbol{\theta}_o} \boldsymbol{t}(X).$$

The assertion follows because

$$\begin{aligned} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}_n} \min\Big(\|\boldsymbol{Z}_{in}\|_2, \|\boldsymbol{Z}_{in}\|_2^2 \Big) &= n \, \mathbb{E}_{\boldsymbol{\theta}_n}(\|\boldsymbol{Z}_{1n}\|_2) \\ &\leq p \, \mathbb{E}_{\boldsymbol{\theta}_o} \Big(t_1(X_{1n}) - \overline{t_1(\boldsymbol{X}_n)} \Big) + o(1) \\ &\to 0, \end{aligned}$$

by the row-wise identical distribution of the X_{in} together with (5.24) and assuming without loss of generality that the maximal difference appears in the first component of $\boldsymbol{t}(X)$. To be able to apply the Lindeberg-Feller central limit theorem of Section A.9 to $\sqrt{n}(\boldsymbol{t}(X_{in}) - \mathbb{E}_{\boldsymbol{\theta}_n} \boldsymbol{t}(X_{1n}))$, the corresponding condition (A.11) remains to be verified. For all $\varepsilon > 0$,

$$\begin{aligned} &\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}_n} \Big(\|\boldsymbol{Z}_{in}^2\|_2 \Big) 1_{\{\|\boldsymbol{Z}_{in}\|_2 > \varepsilon\}} \\ &\quad \leq \quad p \, \mathbb{E}_{\boldsymbol{\theta}_o} \Big(t_1(X_{1n}) - \overline{t_1(\boldsymbol{X}_n)} \Big)^2 \sum_{i=1}^{n} 1_{\{\boldsymbol{t}(X_{in}) - \overline{\boldsymbol{t}(\boldsymbol{X}_n)} > (\varepsilon/p)\sqrt{n}\}} + o(1) \\ &\quad = \quad O\Big(\sum_{i=1}^{n} 1_{\{\sqrt{n}(\boldsymbol{t}(X_{in}) - \overline{\boldsymbol{t}(\boldsymbol{X}_n)}) > (\varepsilon/p)n\}} \Big) \\ &\quad \to_{\mathrm{p}} \quad 0 \end{aligned}$$

as $n \to \infty$ because the difference in the indicator function is a.s. bounded. Now apply Lindeberg's central limit Theorem A.9.2 to conclude:

$$n^{-1/2} \sum_{i=1}^{n} \Big(\boldsymbol{t}(X_{in}) - \mathbb{E}_{\boldsymbol{\theta}_n} \boldsymbol{t}(X_{1n}) \Big) \quad \to_{\mathcal{D}} \quad \mathcal{N}_{\mathrm{p}}\Big(\mathbf{0}, \mathrm{Cov}_{\boldsymbol{\theta}_o} \boldsymbol{t}(X) \Big). \tag{5.26}$$

Together with the moment condition

$$\overline{\boldsymbol{t}(\boldsymbol{X}_n)} \quad = \quad \mathbb{E}_{\widehat{\boldsymbol{\theta}}_n} \boldsymbol{t}(X_{1n}) \tag{5.27}$$

for the maximum likelihood estimator, (5.26) implies

$$\begin{aligned}
&\mathbb{E}_{\widehat{\boldsymbol{\theta}}_n}\, \boldsymbol{t}(X_{1n}) - \mathbb{E}_{\boldsymbol{\theta}_n}\, \boldsymbol{t}(X_{1n}) \\
&= (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n)\Big(\mathrm{Cov}_{\boldsymbol{\theta}_o}\, \boldsymbol{t}(X) + O_{\mathrm{p}}(\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\|_2)\Big) + O_{\mathrm{p}}\Big(\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\|_2^2\Big) \\
&= (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n)\,\mathrm{Cov}_{\boldsymbol{\theta}_o}\, \boldsymbol{t}(X) + o_{\mathrm{p}}(n^{-1/2})
\end{aligned}$$

wherefrom we deduce, using again (5.26) as well as (5.27):

$$\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) &= \sqrt{n}\Big(\overline{\boldsymbol{t}(\boldsymbol{X}_n)} - \mathbb{E}_{\boldsymbol{\theta}_n}\, \boldsymbol{t}(X_n)\Big)\Big(\mathrm{Cov}_{\boldsymbol{\theta}_o}\, \boldsymbol{t}(X)\Big)^{-1} + o_{\mathrm{p}}(1) \\
&\to_{\mathcal{D}} \mathcal{N}_{\mathrm{p}}(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_o)^{-1}) \text{ as } n \to \infty
\end{aligned}$$

by (5.2) and (5.26). □

Proof of Theorem 5.2.1: First let us derive a Taylor expansion for the log-likelihood function $\widehat{L}_n$ and two vectors $\boldsymbol{\alpha}, \boldsymbol{\alpha}_o \in \Theta$:

$$\begin{aligned}
\widehat{L}_n(\boldsymbol{\alpha}) &= \widehat{L}_n(\boldsymbol{\alpha}_o) + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_o)^\top \sum_{i=1}^n \dot{\boldsymbol{\ell}}_{\boldsymbol{\alpha}_o}(X_{in}) + \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_o)^\top \sum_{i=1}^n \ddot{\boldsymbol{\ell}}_{\boldsymbol{\alpha}_o}(X_{in})(\boldsymbol{\alpha} - \boldsymbol{\alpha}_o) + \\
&\quad + \frac{1}{6}\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n (\alpha_j - \alpha_{0,j})(\alpha_k - \alpha_{0,k})(\alpha_l - \alpha_{0,l}) \sum_{i=1}^n \gamma_{jkl} M_{jkl}(X_{in}) \qquad (5.28)
\end{aligned}$$

where $|\gamma_{jkl}| = 1$ and $M_{jkl}(x)$ is such that

$$\Big|\frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} \boldsymbol{\ell}_{\boldsymbol{\theta}}(x)\Big| \leq M_{jkl}(x)$$

for all $j, k, l = 1, \ldots, n$. Write R_n for the fourth summand in (5.28). It is not difficult but tedious to verify that all the above third derivatives of $\boldsymbol{\ell}_{\boldsymbol{\theta}}$ are linear combinations of moments of $\boldsymbol{t}(X)$ and therefore, by Lemma 5.1.1, bounded. This implies by Theorem A.9.1:

$$\begin{aligned}
\Big|\frac{1}{n}\sum_{i=1}^n M_{jkl}(X_{in})\Big| &\leq \mathbb{E}_{\boldsymbol{\theta}_o}\, M_{jkl}(X_{1n}) \\
&\leq C(p_{\boldsymbol{\theta}}, \boldsymbol{\theta}_o)
\end{aligned}$$

with probability tending to one for all $j, k, l = 1, \ldots, n$, a constant $C = C(p_{\boldsymbol{\theta}}, \boldsymbol{\theta}_o)$ only depending on the exponential family under consideration and $\boldsymbol{\theta}_o$. Consequently, R_n can be written as

$$|R_n| = O_{\mathrm{p}}\Big(\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n \sqrt{n}(\alpha_j - \alpha_{0,j})\sqrt{n}(\alpha_k - \alpha_{0,k})(\alpha_l - \alpha_{0,l})\Big).$$

Now if we have $\sqrt{n}(\alpha_i - \alpha_{0,i}) = O(1)$ for all $i = 1, \ldots, n$ then

$$|R_n| \;=\; o(1).$$

The expansion in (5.28) will now be used to derive the limit distribution of Λ_n. By the assumption of the theorem,

$$\widehat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n \;=\; \widehat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^0 - \boldsymbol{e}_p \frac{h}{\sqrt{n}}. \tag{5.29}$$

Setting $\boldsymbol{\alpha} = \widehat{\boldsymbol{\theta}}_n^0$ and $\boldsymbol{\alpha}_o = \boldsymbol{\theta}_n$ in (5.28) we get a first approximation as follows:

$$\begin{aligned}\widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^0) - \widehat{L}_n(\boldsymbol{\theta}_n) = \sqrt{n}(\widehat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n)^\top n^{-1/2} \sum_{i=1}^n \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_n}(X_{1n}) + \\ + \frac{1}{2}\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n)^\top \frac{1}{n}\sum_{i=1}^n \ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_n}(X_{1n})\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n) + o_{\mathrm{p}}(1).\end{aligned}$$

Combining (5.3) and again Theorem A.9.1 one has

$$\frac{1}{n}\sum_{i=1}^n \ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_n}(X_{1n}) \;=\; -\mathbf{I}(\boldsymbol{\theta}_o) + o_{\mathrm{p}}(1)$$

what together with (5.29) yields:

$$\begin{aligned}&\widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^0) - \widehat{L}_n(\boldsymbol{\theta}_n) = \\ &= \sqrt{n}(\widehat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^0)^\top n^{-1/2} \sum_{i=1}^n \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_n}(X_{1n}) \\ &\quad -\frac{1}{2}\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^0)^\top \mathbf{I}(\boldsymbol{\theta}_o)\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^0) - \boldsymbol{e}_p^\top \frac{h}{\sqrt{n}}\sum_{i=1}^n \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_n}(X_{1n}) + \\ &\quad +\boldsymbol{e}_p^\top h\mathbf{I}(\boldsymbol{\theta}_o)\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^0) - (h^2/2)\mathbf{I}^{22}(\boldsymbol{\theta}_o) + o_{\mathrm{p}}(1) \\ &= \tilde{\boldsymbol{Y}}_n^\top \Big(\tilde{\boldsymbol{V}}_n + \mathbf{I}^{12}(\boldsymbol{\theta}_o)h\Big) - hV_{n,p} - \frac{1}{2}\tilde{\boldsymbol{Y}}_n^\top \mathbf{I}^{11}(\boldsymbol{\theta}_o)\tilde{\boldsymbol{Y}}_n - \frac{1}{2}\mathbf{I}^{22}(\boldsymbol{\theta}_o)h^2 + o_{\mathrm{p}}(1)\end{aligned} \tag{5.30}$$

where we introduced

$$\begin{aligned}\boldsymbol{Y}_n &:= \sqrt{n}(\widehat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^0) \\ \boldsymbol{V}_n &:= \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_n}(X_{1n}).\end{aligned}$$

In order to get $\tilde{\boldsymbol{Y}}_n$, minimize the difference $\widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^0) - \widehat{L}_n(\boldsymbol{\theta}_n)$ over $\tilde{\boldsymbol{Y}}_n$. Therefore, set the derivative of the expression in (5.30) equal to 0, yielding:

$$\tilde{\boldsymbol{Y}}_n^{\min} = \mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\Big(\tilde{\boldsymbol{V}}_n + \mathbf{I}^{12}(\boldsymbol{\theta}_o)h\Big). \tag{5.31}$$

Reinserting $\tilde{\boldsymbol{Y}}_n^{\min}$ in (5.30) we finally get

$$\widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^0) - \widehat{L}_n(\boldsymbol{\theta}_n) = \frac{1}{2}\Big(\tilde{\boldsymbol{V}}_n + \mathbf{I}(\boldsymbol{\theta}_o)^{12}h\Big)^\top \mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\Big(\tilde{\boldsymbol{V}}_n + \mathbf{I}(\boldsymbol{\theta}_o)^{12}h\Big) - V_{n,p}h - \frac{1}{2}\mathbf{I}^{22}(\boldsymbol{\theta}_o)h^2. \tag{5.32}$$

Using again the approximation (5.28) with $\boldsymbol{\alpha} = \widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\alpha}_o = \boldsymbol{\theta}_n$ and taking into account that

$$\sum_{i=1}^{n} \dot{\boldsymbol{\ell}}_{\widehat{\boldsymbol{\theta}}_n}(X_{1n}) = \mathbf{0}$$

one can derive in a similar fashion as above:

$$\begin{aligned}\widehat{L}_n(\widehat{\boldsymbol{\theta}}_n) - \widehat{L}_n(\boldsymbol{\theta}_n) &= \frac{1}{2}\boldsymbol{V}_n^\top \mathbf{I}(\boldsymbol{\theta}_o)\boldsymbol{V}_n + o_{\mathrm{p}}(1) \\ &= \frac{1}{2}\tilde{\boldsymbol{V}}_n^\top \mathbf{I}(\boldsymbol{\theta}_o)\tilde{\boldsymbol{V}}_n + \frac{1}{2}V_{n,p\cdot 1}^\top \mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)^{-1}V_{n,p\cdot 1} + o_{\mathrm{p}}(1)\end{aligned} \tag{5.33}$$

defining

$$\begin{aligned}V_{n,p\cdot 1}(\boldsymbol{\theta}) &= V_{n,p} - \mathbf{I}^{21}(\boldsymbol{\theta})\mathbf{I}^{11}(\boldsymbol{\theta})^{-1}\tilde{\boldsymbol{V}}_n \\ \mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}) &= \mathbf{I}^{22}(\boldsymbol{\theta}) - \mathbf{I}^{21}(\boldsymbol{\theta})\mathbf{I}^{11}(\boldsymbol{\theta})^{-1}\mathbf{I}^{12}(\boldsymbol{\theta})\end{aligned}$$

and applying Lemma A.10.2. Now again by Lindeberg's Central Limit Theorem (Theorem A.9.2) we have for the vector of scores $\boldsymbol{V}_n$, as $n \to \infty$,

$$\boldsymbol{V}_n \to_{\mathcal{D}} \mathcal{N}_{\mathrm{p}}\Big(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_o)\Big) \tag{5.34}$$

(see also Section 5.3 in van der Vaart, 1998). Consequently,

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{\theta}_o} V_{n,p\cdot 1} &= \mathbb{E}_{\boldsymbol{\theta}_o} V_{n,p}^2 - 2\mathbf{I}^{21}(\boldsymbol{\theta}_o)\mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\,\mathbb{E}_{\boldsymbol{\theta}_o}(V_{n,p}\tilde{\boldsymbol{V}}_n) + \\ &\qquad \mathbb{E}_{\boldsymbol{\theta}_o}\Big(\mathbf{I}^{21}(\boldsymbol{\theta}_o)\mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\tilde{\boldsymbol{V}}_n\tilde{\boldsymbol{V}}_n^\top \mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\mathbf{I}^{12}(\boldsymbol{\theta}_o)\Big) + o_{\mathrm{p}}(1) \\ &= \mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o) + o_{\mathrm{p}}(1).\end{aligned}$$

This together with Lemma A.10.1 implies that

$$Z \;=\; \mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)^{-1/2} V_{n,p\cdot 1} \tag{5.35}$$

converges in distribution to a standard normal distribution.
All ingredients to tackle Λ_n are now made available. Subtracting (5.32) from (5.33) and multiplied by 2 results in

$$\begin{aligned}
\Lambda_n &= 2\widehat{L}_n(\widehat{\boldsymbol{\theta}}_n) - 2\widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^0) + o_{\mathrm{p}}(1) \\
&= \mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)h^2 + 2V_{n,p\cdot 1}h + V_{n,p\cdot 1}^2 \mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)^{-1} + o_{\mathrm{p}}(1) \\
&= \Big(Z + \mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)^{1/2}h\Big)^2 + o_{\mathrm{p}}(1).
\end{aligned}$$

Due to (5.35), Λ_n converges in distribution to a χ^2-distribution with one degree of freedom and non-centrality parameter $\mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)h^2$. The above representation also details that $\Lambda_n \to_{\mathrm{p}} \infty$ whenever $h = \sqrt{n}\theta_{n,p} \to \infty$. □

Proof of Theorem 5.2.2: Generalizing (5.29) one has

$$\widehat{\boldsymbol{\theta}}_n^{\eta} - \boldsymbol{\theta}_n \;=\; \widehat{\boldsymbol{\theta}}_n^{\eta} - \boldsymbol{\theta}_n^{\eta} + \Big(\eta - \frac{h}{\sqrt{n}}\Big)\boldsymbol{e}_p.$$

Similarly to (5.30) one can derive the following Taylor approximation:

$$\begin{aligned}
&\widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^{\eta}) - \widehat{L}_n(\boldsymbol{\theta}_n) = \\
&= (\widehat{\boldsymbol{\theta}}_n^{\eta} - \boldsymbol{\theta}_n^{\eta})^{\top} \sum_{i=1}^{n} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_n}(X_{1n}) + \Big(\eta - \frac{h}{\sqrt{n}}\Big)\boldsymbol{e}_p^{\top} \sum_{i=1}^{n} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_n}(X_{1n}) \\
&\quad -\frac{1}{2} n (\widehat{\boldsymbol{\theta}}_n^{\eta} - \boldsymbol{\theta}_n^{\eta})^{\top} \mathbf{I}(\boldsymbol{\theta}_o)(\widehat{\boldsymbol{\theta}}_n^{\eta} - \boldsymbol{\theta}_n^{\eta}) - \\
&\quad -n(\widehat{\boldsymbol{\theta}}_n^{\eta} - \boldsymbol{\theta}_n^{\eta})^{\top} \mathbf{I}(\boldsymbol{\theta}_o)\Big(\eta - \frac{h}{\sqrt{n}}\Big)\boldsymbol{e}_p - \frac{1}{2} n \Big(\eta - \frac{h}{\sqrt{n}}\Big)^2 \mathbf{I}^{22}(\boldsymbol{\theta}_o) + o_{\mathrm{p}}(1).
\end{aligned}$$

Taking the derivative w.r.t. to η yields:

$$\frac{\partial}{\partial \eta}\widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^{\eta}) \;=\; \boldsymbol{e}_p^{\top} \sum_{i=1}^{n} \dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}_n}(X_{1n}) - n(\widehat{\boldsymbol{\theta}}_n^{\eta} - \boldsymbol{\theta}_n^{\eta})^{\top} \mathbf{I}(\boldsymbol{\theta}_o)\boldsymbol{e}_p - n\Big(\eta - \frac{h}{\sqrt{n}}\Big)^2 \mathbf{I}^{22}(\boldsymbol{\theta}_o) + o_{\mathrm{p}}(1).$$

Dividing by $\sqrt{n}$ and setting $\eta = 0$ finally gives for the score statistic:

$$\begin{aligned}
S_n &= n^{-1/2} \frac{\partial}{\partial \eta}\widehat{L}_n(\widehat{\boldsymbol{\theta}}_n^{\eta})\Big|_{\eta=0} \\
&= V_{n,p} - \tilde{\boldsymbol{Y}}_n^{\top} \mathbf{I}^{12}(\boldsymbol{\theta}_o) + \mathbf{I}^{22}(\boldsymbol{\theta}_o)h + o_{\mathrm{p}}(1).
\end{aligned}$$

To derive the limiting distribution for the LRT we already figured out the form of $\tilde{\boldsymbol{Y}}$, see equation (5.31). Therewith,

$$\begin{aligned} S_n &= V_{n,p} - \Big(\tilde{\boldsymbol{V}}_n + \mathbf{I}^{12}(\boldsymbol{\theta}_o)h\Big)^\top \mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\mathbf{I}^{12}(\boldsymbol{\theta}_o) + \mathbf{I}^{22}(\boldsymbol{\theta}_o)h + o_{\mathrm{p}}(1) \\ &= V_{n,p} + \mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)h - \tilde{\boldsymbol{V}}_n^\top \mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\mathbf{I}^{12}(\boldsymbol{\theta}_o) + o_{\mathrm{p}}(1). \end{aligned} \tag{5.36}$$

Using (5.34) the variance of (5.36) is

$$\begin{aligned} &\mathrm{Var}_{\boldsymbol{\theta}_o}\Big(V_{n,p} - \tilde{\boldsymbol{V}}_n^\top \mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\mathbf{I}^{12}(\boldsymbol{\theta}_o)\Big) \\ &= \mathbb{E}_{\boldsymbol{\theta}_o}\Big([V_{n,p} - \tilde{\boldsymbol{V}}_n^\top \mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\mathbf{I}^{12}(\boldsymbol{\theta}_o)]^2\Big) + o_{\mathrm{p}}(1) \\ &= \mathbb{E}_{\boldsymbol{\theta}_o} V_{n,p}^2 - 2\,\mathbb{E}_{\boldsymbol{\theta}_o}(V_{n,p}\tilde{\boldsymbol{V}}_n^\top)\mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\mathbf{I}^{12}(\boldsymbol{\theta}_o) + \mathbb{E}_{\boldsymbol{\theta}_o}\Big([\tilde{\boldsymbol{V}}_n^\top \mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\mathbf{I}^{12}(\boldsymbol{\theta}_o)]^2\Big) + o_{\mathrm{p}}(1) \\ &= \mathbf{I}^{22}(\boldsymbol{\theta}_o) - 2\mathbf{I}^{21}(\boldsymbol{\theta}_o)\mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\mathbf{I}^{12}(\boldsymbol{\theta}_o) + \\ &\qquad \mathbf{I}^{21}(\boldsymbol{\theta}_o)\mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\,\mathbb{E}_{\boldsymbol{\theta}_o}(\tilde{\boldsymbol{V}}_n^\top \tilde{\boldsymbol{V}}_n)\mathbf{I}^{11}(\boldsymbol{\theta}_o)^{-1}\mathbf{I}^{12}(\boldsymbol{\theta}_o) + o_{\mathrm{p}}(1) \\ &= \mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o) + o_{\mathrm{p}}(1). \end{aligned}$$

This together with (5.36) finally entails

$$S_n \;\to_{\mathcal{D}}\; \mathcal{N}\Big(\mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)h, \mathbf{I}^{22\cdot 1}(\boldsymbol{\theta}_o)\Big)$$

wherefrom we easily deduce the latter two statements in Theorem 5.2.2. From (5.36) it follows that $S_n \to_{\mathrm{p}} \infty$ if $h \to \infty$. □

Proof of Theorem 5.3.1: The proof of this lemma consists of elementary, tedious and little instructive manipulations and is therefore omitted. We only point out that the following recursion formula helps:

$$H_k(\theta) \;=\; \exp(\theta)/\theta - (k/\theta)H_{k-1}(\theta)$$

for $k = 1, 2, ...$ and any $\theta \in \mathbb{R}$. □

Proof of Theorem 5.4.1: Using (5.25) one has

$$\begin{aligned} \mathrm{Var}_{\widehat{\theta}_n^0}\, T_{\widehat{\theta}_n^0}(X_{1n}) &= \mathrm{Var}_{\theta_o}\, T_{\theta_o}(X_{1n}) + O(|\widehat{\theta}_n^0 - \theta_o|) \\ &= \mathrm{Var}_{\theta_o}\, T_{\theta_o}(X_{1n}) + o_{\mathrm{p}}(1) \end{aligned} \tag{5.37}$$

by assumption (5.16), because $\eta_n \to 0$ entails that $\widehat{\theta}_n^0$ consistently estimates θ_o. This continuity property of the variance together with Theorem 5.2.2 already entails the statement of the present theorem if $\sqrt{n}|\eta_n| \to h$, where $h \geq 0$.

Next, rewrite $T_n(\boldsymbol{X}_n, \widehat{\theta}_n^0)$ as:

$$\begin{aligned}
&T(\boldsymbol{X}_n, \widehat{\theta}_n^0) = \\
&= n^{-1/2} \sum_{i=1}^{n} \frac{X_{in}^2/2 - a(\widehat{\theta}_n^0) X_{in} - b(\widehat{\theta}_n^0)}{[\mathrm{Var}_{\widehat{\theta}_n^0} T_{\widehat{\theta}_n^0}(X_{1n})]^{1/2}} \\
&= \Big(C + o_{\mathrm{p}}(1)\Big) \sqrt{n} \Big(2^{-1} (\overline{X_n^2} - \mathbb{E}_{\theta_n, \eta_n} X_{1n}^2) - a(\widehat{\theta}_n^0)(\overline{X_n} - \mathbb{E}_{\theta_n, \eta_n} X_{1n}) \Big) + \\
&\qquad \Big(C + o_{\mathrm{p}}(1)\Big) \sqrt{n} \Big(2^{-1} \, \mathbb{E}_{\theta_n, \eta_n} X_{1n}^2 - a(\widehat{\theta}_n^0) \, \mathbb{E}_{\theta_n, \eta_n} X_{1n} - b(\widehat{\theta}_n^0) \Big) \quad \text{by (5.37)} \\
&= \Big(C + o_{\mathrm{p}}(1)\Big) \sqrt{n} \Big(2^{-1} \, \mathbb{E}_{\theta_n, \eta_n} X_{1n}^2 - a(\widehat{\theta}_n^0) \, \mathbb{E}_{\theta_n, \eta_n} X_{1n} - b(\widehat{\theta}_n^0) \Big) \qquad (5.38) \\
&= \sqrt{n} \Big(O(|\theta_n - \widehat{\theta}_n^0|) + O(|\eta_n|) \Big) \quad \text{by (5.12) and (5.24)} \\
&= \sqrt{n} \Big(o_{\mathrm{p}}(n^{-1/2}) + O_{\mathrm{p}}(|\eta_n|) \Big) \\
&= o_{\mathrm{p}}(1) + O_{\mathrm{p}}(n^{1/2} |\eta_n|)
\end{aligned}$$

for a generic positive constant C independent of n where (5.38) is received via (5.26). From these derivations we see that indeed

$$T_n(\boldsymbol{X}_n, \widehat{\theta}_n^0) \quad \to_{\mathrm{p}} \quad \infty$$

as $n \to \infty$ if ever $|\eta_n|$ diminishes at a slower rate than $n^{-1/2}$. □

Proof of Lemma 5.5.1: We start the proof with a generally applicable result for spacings when the underlying density f is differentiable and j and k are fulfilling (5.17):

$$X_k - X_j = O_{\mathrm{p}}\Big(\frac{k-j}{n+1}\Big). \qquad (5.39)$$

To proof (5.39), introduce a random vector $\boldsymbol{U}_n := (U_1, \ldots, U_n)$ containing the order statistics of an i.i.d. sample of uniformly on $[0,1]$ distributed random variables U_i, $i = 1, \ldots, n$. Denote the distribution function corresponding to f by F. First use Lemma A.7.1 to receive for all $l = 1, \ldots, n$,

$$U_l \quad = \quad \frac{l}{n+1} + O_{\mathrm{p}}\Big(\sqrt{\Big(\frac{l}{n+1}\Big)\Big(\frac{1}{n+2}\Big)\Big(1 - \frac{l}{n+1}\Big)} \, \Big).$$

Then, using this and applying the mean value theorem for a $z \in]U_j, U_k[$:

$$\begin{aligned}
X_k - X_j &= F^{-1}(U_k) - F^{-1}(U_j) \\
&= (U_k - U_j)(F^{-1})'(z) \\
&= \frac{U_k - U_j}{f(F^{-1}(z))} \\
&= \frac{U_k - U_j}{f(x_\gamma) + f'(x_\gamma)(F^{-1}(z) - x_\gamma) + o(F^{-1}(z) - x_\gamma)} \\
&= \frac{k-j}{n+1}\Big(\frac{1}{f(x_\gamma)} + o_{\mathrm{p}}(1)\Big) + O_{\mathrm{p}}\Big(\sqrt{\frac{k-j}{n^2}\Big(1 - \frac{k-j}{n+1}\Big)}\,\Big) \\
&= O_{\mathrm{p}}\Big(\frac{k-j}{n+1}\Big) + O_{\mathrm{p}}\Big(\frac{\sqrt{k-j}}{n}\Big(1 - \frac{k-j}{n+1}\Big)^{1/2}\Big) \\
&= O_{\mathrm{p}}\Big(\frac{k-j}{n+1}\Big)
\end{aligned}$$

by Assumptions (5.17). To proof the lemma as $n \to \infty$, note that verifying the limit $\mathrm{TV}(\boldsymbol{f}_n(\boldsymbol{X}), \boldsymbol{g}_n(\boldsymbol{X})) \to_{\mathrm{p}} 0$ is equivalent to

$$H^2\Big(\boldsymbol{f}_n(\boldsymbol{X}), \boldsymbol{g}_n(\boldsymbol{X})\Big) \to_{\mathrm{p}} 0 \tag{5.40}$$

by (A.9), where H is the Hellinger distance between two density functions, see Section A.8. The limit in (5.40) holds if

$$\Big(1 - \frac{1}{2}H^2(f_{jk}, h_{jk})\Big)^{k-j-1} \to_{\mathrm{p}} 1$$

using (A.10). Finally, with another simple manipulation, we arrive at the key condition to be verified:

$$(k-j-1)H^2(f_{jk}, h_{jk}) \to_{\mathrm{p}} 0.$$

First, use that as $n \to \infty$,

$$\begin{aligned}
&\int_0^1 \exp(h_{jk}(x) + r_j(x\delta_{jk})\delta_{jk}^2)\,\mathrm{d}x = \\
&= \int_0^1 \Big(1 + \varphi'(X_j)\delta_{jk}x + \varphi''(X_j)\delta_{jk}^2x^2/2 + r_j(x\delta_{jk})\delta_{jk}^2 + O_{\mathrm{p}}(\delta_{jk}^2)\Big)\,\mathrm{d}x \\
&= 1 + O_{\mathrm{p}}(\delta_{jk}).
\end{aligned}$$

Similarly,

$$\int_0^1 \exp h_{jk}(x)\,\mathrm{d}x = 1 + O_{\mathrm{p}}(\delta_{jk}).$$

Now, inserting the definitions of f_{jk} and h_{jk} into the total variation distance and using (5.19) we get as $n \to \infty$,

$$\begin{aligned}
&(k-j-1)H^2(f_{jk}, h_{jk}) = \\
&= (k-j-1)\int_0^1 \Big(\frac{\exp[h_{jk}(x)/2 + r_j(x\delta_{jk})\delta_{jk}^2/2]}{(\int_0^1 \exp(h_{jk}(v) + r_j(v\delta_{jk}))\,\mathrm{d}v)^{1/2}} - \frac{\exp(h_{jk}(x)/2)}{(\int_0^1 \exp h_{jk}(v)\,\mathrm{d}v)^{1/2}} \Big)^2 \mathrm{d}x \\
&= (k-j-1)\int_0^1 \exp h_{jk}(x) \Big([\exp(r_j(x\delta_{jk})\delta_{jk}^2/2) - 1][1 + O_{\mathrm{p}}(\delta_{jk})]^{-1/2} \Big)^2 \mathrm{d}x \\
&\leq \frac{k-j-1}{(1 + O_{\mathrm{p}}(\delta_{jk}))} \Big(\sup_{x\in[0,1]} |r_j(x\delta_{jk})|\delta_{jk}^2/2 + o_{\mathrm{p}}(\delta_{jk}^4) \Big)^2 \Big(\int_0^1 \exp h_{jk}(x)\,\mathrm{d}x \Big) \\
&= (k-j-1)o_{\mathrm{p}}(\delta_{jk}^4)(1 + o_{\mathrm{p}}(1)) \to_{\mathrm{p}} 0. \qquad \square
\end{aligned}$$

Proof of Theorem 5.5.2: Since Lemma 5.5.1 holds, we can restrict our attention to the parabolic density g_{jk}. Generalize the notation for this density to

$$g_{jk}(u, \theta, \eta) = \frac{\exp(\theta u + \eta u^2/2)}{\int_0^1 \exp(\theta v + \eta v^2)\,\mathrm{d}v} 1_{u\in[0,1]}.$$

Generalizing Lemma 14.31 in van der Vaart (1998) to composite hypotheses, we have to verify, in order to proof the theorem,

$$(k-j-1)H^2\Big(g_{jk}(u, \widehat{\theta}_{jk}, \eta_n), g_{jk}(u, \theta_{jk}^0, 0) \Big) \to_{\mathrm{p}} \infty,$$

where the sequences θ_n and η_n are as introduced in (5.20) and (5.21). Finally, $\widehat{\theta}_{jk}$ is a $(k-j-1)^{1/2}-$consistent estimator of θ_n and θ_{jk}^0 is the true parameter of g_{jk} on $\mathcal{I}_{jk}$. Similarly to the calculations in Lemma 5.5.1 one can derive, as $n \to \infty$,

$$(k-j-1)H^2\Big(g_{jk}(u, \widehat{\theta}_{jk}, \eta_n), g_{jk}(u, \theta_{jk}^0, 0) \Big) = O_{\mathrm{p}}(\eta_n (k-j-1)^{1/2}).$$

But thanks to the assumption given by (5.22) this latter expression is unbounded as $n \to \infty$. $\square$

Proof of Lemma 5.7.1: Suppose that $\log f_{jk}$ is concave on $\mathcal{I}_{jk}$. Let F_{jk} and $G_{jk,\theta}$ be the distribution functions corresponding to the densities f_{jk} and $g_{jk,\theta}$. Choose $\theta_1 \le \theta_{jk}$ such that

$$G_{jk,\theta_1}\Big(\frac{a(\widehat{\theta}_{jk})}{2}\Big) \;=\; F_{jk}\Big(\frac{a(\widehat{\theta}_{jk})}{2}\Big).$$

Both F_{jk} and $G_{jk,\theta}$ are distribution functions, what means

$$G_{jk,\theta_1} \;=\; F_{jk} \quad \text{on } \{X_j, a(\widehat{\theta}_{jk})/2, X_k\}.$$

Hence the densities satisfy:

$$\int_{X_j}^{a(\widehat{\theta}_{jk})/2} (g_{jk,\theta_1} - f_{jk}) \;=\; \int_{a(\widehat{\theta}_{jk})/2}^{X_k} (g_{jk,\theta_1} - f_{jk}) = 0. \tag{5.41}$$

Because $\log g_{jk,\theta_1}$ is linear and $\log f_{jk}$ is concave on $\mathcal{I}_{jk}$, (5.41) entails that either $g_{jk,\theta_1} \equiv f_{jk}$ on $\mathcal{I}_{jk}$ or the difference $g_{jk,\theta_1} - f_{jk}$ has exactly two changes of sign, namely at

$$c_1 \in \Big(X_j, a(\widehat{\theta}_{jk})/2\Big) \text{ and } c_2 \in \Big(a(\widehat{\theta}_{jk})/2, X_k\Big)$$

such that

$$g_{jk,\theta_1} - f_{jk} \begin{cases} \ge 0 & \text{on } (X_j, c_1) \cup (c_2, X_k) \\ \le 0 & \text{on } (c_1, c_2). \end{cases}$$

Using Lemma 9 in Dümbgen and Walther (2006) together with (5.41) then yields:

$$F_{jk}^{-1} - G_{jk,\theta_1}^{-1} \begin{cases} \ge 0 & \text{on } \Big(0, F_{jk}\Big(a(\widehat{\theta}_{jk})/2\Big)\Big) \\ \le 0 & \text{on } \Big(F_{jk}\Big(a(\widehat{\theta}_{jk})/2\Big), 1\Big). \end{cases}$$

Consequently,

$$\begin{aligned} X_i = F_{jk}^{-1}(U_{i;j,k}) \;&\gtreqless\; G_{jk,\theta_1}^{-1}(U_{i;j,k}) \\ &=\; X_j + (X_k - X_j) G_{jk,\theta_1}^{-1}(U_{i;j,k}) \end{aligned} \tag{5.42}$$

depending whether $U_{i;j,k} \lesseqgtr F_{jk}(a(\widehat{\theta}_{jk})/2)$, this condition being equivalent to $X_{i;j,k} \lesseqgtr a(\widehat{\theta}_{jk})/2$. The uniform local order statistics $U_{i;j,k}$ are defined similarly

to $X_{i;j,k}$ but for uniform order statistics $U_0, \ldots, U_{n+1}$ instead of the $X_0, \ldots, X_{n+1}$ having density function f. Equation (5.42) entails:

$$\sum_{i=1}^{n} \delta_{\widehat{\theta}_{jk}}(X_{i;j,k}) \;\leq\; \sum_{i=1}^{n} \delta_{\widehat{\theta}_{jk}}(Y_{i;j,k}).$$

where δ_θ was defined in Section 5.6. Tedious calculations reveal that $-a(\theta)^2/4 - b(\theta)$ is a non-decreasing function on $\mathbb{R}$. Hence:

$$\begin{aligned}
0 \;&\leq\; \frac{1}{n} \sum_{i=1}^{n} \Big(\delta_{\widehat{\theta}_{jk}}(Y_{i;j,k}) - \delta_{\widehat{\theta}_{jk}}(X_{i;j,k}) \Big) \\
&=\; \frac{1}{n} \sum_{i=1}^{n} \Big(\delta_{\widehat{\theta}_{jk}}(Y_{i;j,k}) - \delta_{\theta_1}(X_{i;j,k}) - \delta'_{\theta_1}(X_{i;j,k})(\widehat{\theta}_{jk} - \theta_1) \Big) + o_{\mathrm{p}}(n^{-1}) \\
&=\; \frac{1}{n} \sum_{i=1}^{n} \Big(T_{jkn}(\boldsymbol{Y}, \theta_1) - T_{jkn}(\boldsymbol{X}, \widehat{\theta}_{jk}) \Big) + o_{\mathrm{p}}(n^{-1/2}) \\
&\to_{\mathrm{p}}\; \mathbb{E}_{\theta_{jk},0}\, T_{jkn}(\boldsymbol{Y}, \theta_1) - \mathbb{E}_{\theta_{jk},\eta_{jk}}\, T_{jkn}(\boldsymbol{X}, \theta_{jk})
\end{aligned}$$

as $n \to \infty$ by the law of large numbers. The case where f_{jk} is log-convex can be treated analogously. □

Chapter 6

Outlook and open problems

6.1 Estimation based on censored observations

Log-concavity could offer a compromise between fully nonparametric methods such as Kaplan-Meier (or Grenander) and fully parametric models in the estimation of a survival function (via its log-concave density) from censored data as it is smooth, compared to the former two which are step functions with possible high jumps. Compared to the unimodal distribution function estimator of Dümbgen, Freitag, and Jongbloed (2006), the assumption of log-concavity could possibly yield more powerful procedures. However, censored observations complicate the situation compared to the i.i.d. case. One of the obstacles is that the log-likelihood function corresponding to Ψ_n in (4.2) is convex with respect to the density f, but not with respect to the log-density. A first algorithmic approach to tackle this task was taken by Hüsler (2005).
This reasoning also applies to functions derived from a log-concave density, such as the hazard function λ in Section 3.6.

6.2 Tests for distribution functions

Theorem 3.5.1 suggests that the estimator $\widehat{F}_n$ is essentially equivalent to the empirical distribution function $\mathbb{F}_n$. It can therefore be looked at as a smoother for $\mathbb{F}_n$. One should expect that every procedure where somehow the jump function $\mathbb{F}_n$ is plugged into can be improved in terms of accuracy (estimators) or power (tests) when plugging in the smooth function $\widehat{F}_n$ instead, at least if the underlying density function is indeed log-concave. We sketch an example. Consider two i.i.d. samples

$(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$ of equal size (for ease of simplicity) and each component having distribution functions F^X and F^Y, respectively. To test whether $H_o : F^X = F^Y$ versus $H_1 : F^X \neq F^Y$ a common used two-sample test statistic is Kolmogorov-Smirnov, relying on the empirical distribution functions $\mathbb{F}_n^X$ and $\mathbb{F}_n^Y$ of the samples:

$$\begin{aligned} \mathbb{K} &:= \mathbb{K}(\mathbb{F}_n^X, \mathbb{F}_n^Y) \\ &= \sqrt{n}\|\mathbb{F}_n^X - \mathbb{F}_n^Y\|_\infty^{[0,1]}. \end{aligned}$$

The limiting distribution of $\mathbb{K}$ and the therefrom derived asymptotic test can be found e.g. in Durbin (1973). If one imposes that F^X and F^Y both have log-concave density functions, we instead propose to use the following modified test statistic:

$$\begin{aligned} \widehat{K} &:= \widehat{K}(\widehat{F}_n^X, \widehat{F}_n^Y) \\ &= \sqrt{n}\|\widehat{F}_n^X - \widehat{F}_n^Y\|_\infty^{[0,1]} \end{aligned}$$

where $\widehat{F}_n^X$ and $\widehat{F}_n^Y$ are the log-concave distribution function estimators introduced in Section 3.1. Deriving the limiting distribution of this statistic is presumably a difficult task, but if one assumes that under H_o our pooled data $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ has the same distribution as $(X_{\Pi_1}, \ldots, X_{\Pi_n}, Y_{\Pi_{n+1}}, \ldots, Y_{\Pi_{2n}})$ where Π is a random permutation of $\{1, \ldots, 2n\}$ (that does not depend on the data), one can attack the distribution of $\widehat{K}$ under H_o via a Monte Carlo permutation test. Generate M samples of Π and calculate the corresponding values of the test statistic $\widehat{K}_1, \ldots, \widehat{K}_M$. A nonparametric p-value $\widehat{p}$ is then given by:

$$\widehat{p} = \frac{1 + \#\{i \leq M : \widehat{K}_i \geq \widehat{K}_o\}}{1 + M}$$

where $\widehat{K}_o$ is the test statistic for the original samples. It could be exciting to compare the power of this test to that of well established alternatives, such as the above described Kolmogorov-Smirnov or χ^2-tests.

6.3 Tail index estimation

An example for accuracy improvement of an estimator using $\widehat{F}_n$ is given in Müller and Rufibach (2006). We show that both parametric distributions appearing in extreme value theory, the generalized Pareto and the generalized extreme value distribution, have a log-concave density function if the tail index parameter γ lies in $[-1, 0]$.

Suppose we are given an ordered sample $X_1 < \ldots < X_n$ from one of the above two limiting distributions having distribution function G_γ. The most widely used estimators for this tail index γ, such as Pickand's or Falk's, are defined as weighted averages of log-spacings. In order to improve the accuracy of these estimators, the idea is to replace the order statistics used to calculate them by quantiles received via inversion of $\widehat{F}_n$. This smoothing technique substantially reduces variance in estimation not only of γ but already in estimation of the quantiles. We intend to compare this new approach to existing tail index estimation methods and to deduce recommendations when to use which tail index estimator and whether smoothed or not.

Furthermore, we have shown in the above paper that all distribution functions F having a log-concave density function belong to the max-domain of attraction of the generalized extreme value distribution, for some $\gamma \in [-1, 0]$. This result relies on the continuity, unimodality, and the non-decreasing hazard property (see Lemma 2.3.1) of log-concave density functions. It seems clear that the max-domain of attraction should be obtainable for function classes that assume less than log-concavity, as in fact only the tail (i.e. local) behavior of a distribution matters in determining its max-domain of attraction. But log-concavity is a global property of the density.

6.4 Deconvolution with log-concave densities

Groeneboom and Jongbloed (2003) consider the following setting. Suppose we observe random variables $Z_1, \ldots, Z_n$ having density

$$g_F(z) \quad = \quad \int_{\mathbb{R}} k(z - x)\, \mathrm{d}F(x), \quad z \in \mathbb{R}.$$

Here k is a known probability density on $\mathbb{R}$ and F is an arbitrary distribution function. The question is: how can one estimate F or quantities related to it, e.g. moments? Equivalently, one could think of observing

$$Z_i = X_i + Y_i, \quad i = 1, \ldots, n$$

where the X_i are distributed according to F and Y_i have density k. The authors then simplify the task assuming that k is the uniform density on $[0, 1)$, yielding a uniform (or boxcar) deconvolution problem. The nonparametric maximum likelihood estimator $\widehat{F}$ of F is not continuous.

However, the authors introduce a smoothed density estimator $\widehat{f}_{n,h}$:

$$\widehat{f}_{n,h}(t) \;=\; \int_{\mathbb{R}} K_h(t-y)\,\mathrm{d}\widehat{F}_n(y)$$

for some kernel function K_h, $t \in \mathbb{R}$, and a specific bandwidth $h = h(n)$. Beneath the fact that this two-stage kernel estimator $\widehat{f}_{n,h}$ has some undesired boundary properties, it could be fruitful to calculate the estimate $\widehat{F}_n$ directly assuming that F has a log-concave density, i.e. no additional smoothing via K is then necessary.

6.5 Rates for different norms

In this thesis we only considered consistency and rate of convergence in the uniform norm $\|.\|_\infty^T$ on compact intervals T. First, the results in Chapter 3 should somehow be generalized to the whole real line. Then, other norms could be considered, e.g. the limiting behavior (consistency, rate of convergence, limiting distribution) of

$$\|\widehat{f}_n - f\|_p^T \;:=\; \Big(\int_T |\widehat{f}_n(x) - f(x)|^p\,\mathrm{d}x\Big)^{1/p}$$

for any $p \in \mathbb{N}$. This work has already been accomplished for the Grenander estimator by Kulikov and Lopuhaä (2005a, 2006).

Another open problem is a proof that the uniform rate of convergence for the convex decreasing density estimator of Groeneboom, Jongbloed, and Wellner (2001b) has, under their assumptions, uniform rate of convergence of $(\log(n)/n)^{2/5}$, and the generalization of their whole work to density functions belonging to Hölder smoothness classes.

One could also think of a maximum likelihood version of the uniform rate of convergence result in the current status data regression setting of Dümbgen, Freitag, and Jongbloed (2004). Finally, least squares log-concave density estimation could also be tackled.

6.6 Limiting distribution at fixed point

Preliminary considerations suggest that the limiting distribution of

$$n^{\beta/(2\beta+1)}(\widehat{f}_n - f)(x_o)$$

at a fixed point $x_o \in \mathbb{R}$ can possibly be derived in a similar way like in the convex case in Groeneboom, Jongbloed, and Wellner (2001b). One has to consider suitable Taylor approximations to

$$\int_{\mathbb{R}} \Delta(x)(\widehat{f}_n - f)(x)\,\mathrm{d}x,$$

choose the perturbation function Δ such that the first two terms in the series disappear and make suitable application of (3.3). The remaining terms are then approximated through suitable local empirical processes. Since the constant appearing in the limiting distribution for the convex density estimator depends on $f''(x_o)^{-1}$, Groeneboom, Jongbloed, and Wellner (2001b) simply assume $f''(x_o) > 0$. However, for the log-concave density estimator such an assumption would be much too restrictive (if f e.g. stands for the normal density function we have $f''(\pm 1) = 0$), whence presumably an even more involved limiting behavior will outcrop.

6.7 Log-concavity and total positivity

As described in the introduction, monotonicity and convexity are special cases for $k = 1, 2$ in the class of k-monotone densities. These classes were treated by Balabdaoui and Wellner (2004a-d) as a step to the solution of the case $k = \infty$ (complete monotonicity). The relevance of the latter case comes from the fact that the class of completely monotone densities is equivalent to that of scale mixtures of exponentials. Unimodality and log-concavity on the other side are equally special cases for $k = 1, 2$ in the notion of total positivity, see Karlin (1968). Perhaps it could be fruitful to similarly consider the estimation of total positive density functions of order $k = 3, \ldots, \infty$. However, as log-concavity covers many parametric models, imposing further constraints on the density possibly narrows the window too much for statistical applications.

6.8 Multivariate context

Polonik (1995, 1998) pioneered multivariate density estimation under shape constraints. Log-concavity could be another option to be studied in this context, e.g. imposed univariately in some dimensions or globally.

6.9 Bump hunting

The method we propose in Part 2 still relies on the Working Assumption 5.6.1 that a limiting distribution for $T^*_{l,m,n}$ as $n \to \infty$ exists (and is non-degenerate and at best independent of f). A thorough analysis of this limiting distribution is still lacking. Furthermore, our approach estimates the nuisance parameter θ, implying that the size of the test is only guaranteed asymptotically. Minimax approaches (i.e. taking the "worst" θ with respect to a certain criterion) possibly yield procedures that hold the significance level also for finite n. However, presumably an improvement in this sense has to be paid by a loss of power.
As already pointed out in Section 5.6, test functions ϱ instead of T_θ are equally possible, as long as they wipe out linear functions in the sense that

$$\int_{\mathbb{R}} x\varrho(x)\,\mathrm{d}x = 0.$$

Alternative test functions possibly offer a way to directly test convexity or concavity of the underlying density. Probably not all approaches perform equally on all types of underlying densities. These different performances could be assessed empirically and theoretically. Furthermore, (theoretical) power considerations for the method described in Part 2 as well as different assumptions for the alternatives could facilitate the decision for a method in a specific problem. Existing approaches like Silverman's approach (Silverman 1981), the Dip test of Hartigan and Hartigan (1985) or SiZer of Chaudhuri and Marron (1999) could be incorporated in these comparisons.

Appendix A

Standard results

We state here several well known theorems, in the order they appear in Chapters 3 to 5.

A.1 Lebesgue's dominated convergence Theorem

We borrow the formulation and the proof from Pollard (2002).

Theorem A.1.1. *Let f_n be a sequence of μ-integrable functions (i.e. $\int f \, \mathrm{d}\mu < \infty$) for which $\lim_{n\to\infty} f_n(x)$ exists for all x. Suppose there exists a $\mu-$integrable function F, independent of n, such that $|f_n(x)| \leq F(x)$ for all x and all n. Then the limit function $f := \lim_{n\to\infty} f_n$ is integrable and*

$$\lim_{n\to\infty} \int f_n = \int \lim_{n\to\infty} f_n = \int f.$$

A.2 Modulus of continuity of a uniform empirical process

First, define the uniform empirical process. Let $\xi_1, \ldots, \xi_n$ denote independent uniform random variables supported on $[0, 1]$. Introduce for $t \in [0, 1]$

$$\mathbb{G}_n(t) \;:=\; \frac{1}{n} \sum_{i=1}^{n} 1_{\{\xi_i \leq t\}}$$

the empirical distribution function of the sample. Let $(\mathbb{U}_n(t))_{t\in[0,1]}$ denote the uniform empirical process where

$$\mathbb{U}_n(t) \quad := \quad \sqrt{n}(\mathbb{G}_n(t) - t)$$

for $t \in [0,1]$. Our function of interest, the modulus of continuity, is then:

$$\omega(g,d) \quad := \quad \sup_{x\in[A,B-d]} \sup_{|h|\leq d} |g(x+h) - g(x)|$$

for $d > 0$ and functions g bounded on $[A,B]$. From Donsker's Theorem we know that the sequence of processes $(\mathbb{U}_n)_n$ converges weakly to a Brownian Bridge $\mathbb{B}$. Since $\mathbb{B}$ is continuous one can except that $\omega(\mathbb{B},d) \to 0$ a.s. and a famous result by Lévy (1937) specifies the rate of convergence to 0. Stute (1982) carried this result from $\mathbb{B}$ over to $\mathbb{U}_n$, and this is exactly what fits our purposes:

Theorem A.2.1. *Let r_n satisfy the regularity conditions:*

$$\begin{aligned} r_n &\to 0 \\ nr_n &\nearrow \infty \\ \log(r_n^{-1})/\log\log n &\to \infty \\ \log(r_n^{-1})/(nr_n) &\to 0. \end{aligned}$$

The modulus of continuity $\omega(\mathbb{U}_n, r_n)$ of the uniform empirical process then almost surely satisfies:

$$\lim_{n\to\infty} \frac{\omega(\mathbb{U}_n, r_n)}{\sqrt{2r_n \log(r_n^{-1})}} = 1.$$

Sequences r_n complying to the above four conditions are named "bandsequences". A proof for this theorem can be found in the original paper or in Shorack and Wellner (1986).

A.3 The Massart - Dvoretzky - Kiefer - Wolfowitz inequality

In 1956, Dvoretzky, Kiefer and Wolfowitz gave a bound on the tail probability of $\|\mathbb{F}_n - F\|_\infty^{[0,\infty)}$.

Theorem A.3.1. *Let $\mathbb{F}_n$ be the empirical and F the true distribution function for an i.i.d. sample $X_1, \ldots, X_n$. Then there exists a constant $C > 0$ such that for every $x > 0$*

$$P\Big(\sqrt{n}\|\mathbb{F}_n - F\|_\infty^{[0,\infty)} > x\Big) \;\leq\; Ce^{-2x^2}.$$

The constant C was decreased several times until Massart (1990) showed that $C = 2$ holds and that no further improvement is possible. For proofs we refer to the original papers. The expression on the left is the tail probability of the Kolmogorov-Smirnov statistic, see e.g. van der Vaart (1998), Section 19.3.

A.4 Some results from optimization

Suppose we would like to optimize a differentiable convex functional $\Psi_n(\boldsymbol{\eta})$ over vectors $\boldsymbol{\eta} \in \mathbb{R}^n$ under the linear constraint $\mathbf{B}\boldsymbol{\eta} \leq \mathbf{0}$ where $\mathbf{B}$ is a $m \times n$-dimensional matrix, implying that m constraints are present. It would be convenient to know whether an actual candidate $\widehat{\boldsymbol{\eta}}$ already solves the problem. The following theorem delivers exactly what the doctor ordered.

Theorem A.4.1. *Let $\widehat{\boldsymbol{\eta}}$ be a vector in $\mathbb{R}^n$ such that $\Psi_n(\widehat{\boldsymbol{\eta}}) < \infty$. Then $\widehat{\boldsymbol{\eta}}$ minimizes $\Psi_n(\boldsymbol{\eta})$ over the set of vectors $\boldsymbol{\eta}$ such that $\mathbf{B}\boldsymbol{\eta} \leq \mathbf{0}$, if, and only if, the following conditions hold for some vectors $\boldsymbol{v}, \boldsymbol{s} \in \mathbb{R}^m$:*

$$\nabla_{\boldsymbol{\eta}}\Psi_n + \mathbf{B}^\top \boldsymbol{v} = \mathbf{0} \tag{A.1}$$

$$\mathbf{B}\widehat{\boldsymbol{\eta}} + \boldsymbol{s} = \mathbf{0} \tag{A.2}$$

$$v_i s_i = 0 \;\; \textit{for all } i = 1, \ldots, m \tag{A.3}$$

$$\boldsymbol{v} \geq \mathbf{0} \tag{A.4}$$

$$\boldsymbol{s} \geq \mathbf{0}. \tag{A.5}$$

Conditions (A.1)-(A.5) are referred to as the Karush-Kuhn-Tucker conditions. Note that we consider here a special formulation of the problem. Generalizations also include equality constraints and non-linear constraint functions. For a formulation of such a much more general version of the theorem, consider e.g. Wright (1997, Appendix A). This book also provides a proof of Theorem A.4.1.
In Chapter 4 we introduce a primal log-barrier algorithm. The theorem below ensures that an algorithm based on this method indeed finds the solution $\widehat{\boldsymbol{\eta}}$.

Theorem A.4.2. *Suppose that there exists a point $\widehat{\boldsymbol{\eta}} \in \mathcal{F}$, where $\mathcal{F}$ is the feasible set introduced in Section 4.2. Let the level sets $\{\boldsymbol{\eta} : \mathbf{B}\boldsymbol{\eta} \leq \mathbf{0}, \Psi_n(\boldsymbol{\eta}) \leq c\}$ be bounded for every $c > 0$. Assume further that the functional Ψ_n is differentiable and convex. Then the optimization problem*

$$\min_{\boldsymbol{\eta} \in \mathbb{R}^n} \Big\{ \Psi_n(\boldsymbol{\eta}) - \mu \sum_{i=1}^{m} \log\Big(-(\mathbf{B}\boldsymbol{\eta})_i\Big) \Big\}$$

has a solution for all $\mu > 0$ and this solution is unique. Furthermore, $\boldsymbol{\eta}(\mu)$ tends to the optimal solution $\widehat{\boldsymbol{\eta}}$ as μ is driven down to 0.

A proof is given e.g. by Fiacco and McCormick (1968) who in fact introduced this method.

A.5 Isotonic regression

Suppose a real-valued bivariate random vector (X, Y) is given. Let $F(\cdot|x)$ denote the conditional distribution function of Y given $X = x$, i.e. for $x, y \in \mathbb{R}$:

$$F(y|x) \;=\; P(Y \leq y | X = x).$$

In linear regression, one now assumes that the unknown mean function

$$\begin{aligned} m(x) \;&:=\; \mathbb{E}(Y|X = x) \\ &=\; \int y \, \mathrm{d}F(y|x), \;\; x \in \mathbb{R} \end{aligned}$$

is affine linear and lies in a given d-dimensional space of functions, denoted by $\mathcal{L}^d$, where $d \in \mathbb{N}$ is known and fixed. An example for $\mathcal{L}^d$ is

$$\{f \;:\; x \mapsto f(x) = \sum_{i=0}^{d} a_i x^i\},$$

i.e. the $(d+1)$-dimensional vector space of all polynomials of at most dimension d. Given a sample of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $(X_i, Y_i) =_{\mathcal{D}} (X, Y)$ for all $i = 1, \ldots, n$, a possible way to define an estimator $\widehat{m}$ is via weighted least squares:

$$\widehat{m}(x) \quad := \quad \underset{m \in \mathcal{L}^d}{\arg\min} \sum_{i=1}^{n} w_i (Y_i - m(X_i))^2 \tag{A.6}$$

where the $w_i, i = 1, \ldots, n$ are specifying the weight that each observation is given to. Sometimes it is plausible to assume that the function m is isotonic rather than linear, i.e. monotone non-decreasing in x implying that for any $x_1, x_2 \in \mathbb{R}$ such that $x_1 \leq x_2$ and $y \in \mathbb{R}$ one has

$$F(y|x_1) \quad \geq \quad F(y|x_2).$$

Problem (A.6) then transforms to

$$\widehat{m}_{\uparrow} \quad = \quad \underset{m(X_1) \leq \ldots \leq m(X_n)}{\arg\min} \sum_{i=1}^{n} w_i (Y_i - m(X_i))^2 \tag{A.7}$$

where we focus our attention on estimation of m on the set of observations $\mathcal{X} := \{X_1, \ldots, X_n\}$. Lower and upper bounds for $\widehat{m}_{\uparrow}(x)$ for $x \notin \mathcal{X}$ can then be found via the isotonic property, e.g. trough linear step functions. Now the PAVA comes into play. The crucial point is to introduce the cumulative sum diagram (CSD), i.e. to plot the points $p_j = (W_j, G_j)$ for $j = 0, \ldots, n$ where

$$W_j := \sum_{k=1}^{j} w(X_k) \qquad G_j := \sum_{k=1}^{j} w(X_k) Y_i.$$

Define the greatest convex minorant (GCM) at a place $t \in \mathbb{R}$ as the supremum of the values at t of all convex functions that lie entirely below the CSD. Theorem 1.2.1 in Robertson, Wright, and Dykstra (1988) then guarantees that the left derivative of the GCM solves problem (A.7). The key is that if we have two violators of the monotonicity constraint, i.e. there exist a $i_o \in \{2, \ldots, n\}$ such that $Y_{i_o - 1} > Y_{i_o}$, we can connect the points $P_{i_o - 2}$ and P_{i_o} in the CSD via a straight line, a modification that leaves the GCM unchanged but the above points do not violate the monotonicity constraint for the left derivative anymore. The same theorem ensures that a solution found by this procedure indeed minimizes the weighted sum of squares in (A.7).
Finally, the aforementioned book details an algorithm to find $\widehat{m}_{\uparrow}$ via an iterative algorithm. It can be shown that this algorithm in this specific least square case needs at most $O(n)$ operations to find $\widehat{m}_{\uparrow}$.

A.6 A convergence theorem for iterative algorithms

Dümbgen, Freitag, and Jongbloed (2006) present a framework to compute MLEs iteratively, well applicable to many known iterative algorithms. For ease of completeness we summarize their theorem on convergence of these algorithms. We make use of this theorem in Sections 4.5 and 4.6.

Suppose we want to maximize a functional $L : \Theta \to [-\infty, \infty)$ over some metric space (Θ, ρ). The following regularity conditions are imposed on L.

(A.1) The functional L is continuous on Θ, and the set $\{L > -\infty\}$ is nonvoid.

(A.2) For any $r \in \mathbb{R}$ the set $\{L \geq r\}$ is compact (or empty).

The second condition implies that the set

$$\widehat{\Theta} := \arg\max_{x \in \Theta} L(x)$$

is nonvoid and compact. Note that if $\Theta = \mathbb{R}$ and L is concave, Conditions (**A.1**) and (**A.2**) are easily guaranteed. To perform the maximiziation, introduce an algorithmic mapping Π from $\Theta_o := \Theta \cap \{L > -\infty\}$ onto itself. This algorithmic mapping Π should satisfy the following conditions:

(B.1) All iterates lie in $\widehat{\Theta}$: $\Pi(x) \in \widehat{\Theta}$ for all $x \in \widehat{\Theta}$.

(B.2) Improve the iterates in every step: For any $x \in \Theta_o \setminus \widehat{\Theta}$,

$$\liminf_{y \to x} L\Big(\Pi(y)\Big) \quad > \quad L(x).$$

Note that only requesting $L(\Pi(x)) > L(x)$ for any $x \in \Theta_o \setminus \widehat{\Theta}$ is not strict enough to guarantee Theorem A.6.1.

Theorem A.6.1. *Suppose that L and Π satisfy Conditions (**A.1-2**) and (**B.1-2**). For an arbitrary starting point $x_o \in \Theta_o$ define inductively new iterates $x_n := \Pi(x_{n-1})$ for $n \geq 1$. Then*

$$\lim_{n \to \infty} \min_{\widehat{x} \in \widehat{\Theta}_o} \rho(x_n, \widehat{x}) \quad = \quad 0.$$

This theorem is Proposition 3.1 in Dümbgen, Freitag, and Jongbloed (2006). The proof can be found there.

A.7 Some results about order statistics

Here we give some fundamental properties of order statistics. Let $U_1 < \ldots < U_n$ be an i.i.d. ordered random sample of uniformly distributed random variables. For a distribution function F define another sample $X_1 < \ldots < X_n$ via

$$X_i \;:=\; F^{-1}(U_i), \qquad i = 1, \ldots, n.$$

It is well known that then all the X_i are distributed according to F. Introduce another ordered i.i.d. exponentially distributed sample $Y_1 < \ldots < Y_n$. We summarize the facts used in the proofs in Section 5.5 in the following lemma:

Lemma A.7.1. *For the ordered random variables we have:*

$$(U_k)_{k=1}^n \;=_{\mathcal{D}}\; \left(\frac{\sum_{i=1}^k Y_i}{\sum_{j=1}^{n+1} Y_j}\right)_{k=1}^n \tag{A.8}$$

whereas for the spacings

$$\Big(U_k - U_{k-1}\Big)_{k=1}^{n+1} \;=_{\mathcal{D}}\; \left(\frac{Y_k}{\sum_{j=1}^{n+1} Y_j}\right)_{k=1}^{n+1}.$$

Finally, one single order statistic U_j has a $Beta(j, n+1-j)$-distribution where

$$\mathbb{E}(U_j) = j/(n+1) \qquad \mathrm{Var}(U_j) = \Big(\frac{j}{n+1}\Big)\Big(\frac{1}{n+2}\Big)\Big(1 - \frac{j}{n+1}\Big).$$

The proof of this lemma is elementary and can e.g. be found in Arnold et. al (1992). Through application of (A.8) one can further deduce that $U_k - U_j =_{\mathcal{D}} U_{k-j}$.

A.8 Total variation and Hellinger distance

When replacing a density function by local parabolas in Section 5.5 we argue that by doing this the total variation distance between the original density and the approximation is asymptotically negligible. The proof relies among other things on Lemma A.8.1. Usually, the following definitions are given with (probability) measures as arguments, (e.g. in van der Vaart (1998), Chapter 14). However, our arguments will be the densities directly. For two probability densities $p : \mathbb{R}^k \to \mathbb{R}$ and $q : \mathbb{R}^k \to \mathbb{R}$ define the total variation distance as

$$\mathrm{TV}(p, q) \;:=\; \int_{\mathbb{R}^k} |p(\boldsymbol{x}) - q(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x}$$

and the Hellinger distance

$$H(p,q) \;:=\; \Big(\int_{\mathbb{R}^k}\Big(\sqrt{p(\boldsymbol{x})}-\sqrt{q(\boldsymbol{x})}\Big)^2\,\mathrm{d}\boldsymbol{x}\Big)^{1/2}.$$

The following lemma delivers the critical (in-)equalities.

Lemma A.8.1. *For two probability densities $p, q \in \mathrm{L}_1(\mathbb{R}^k)$ we have*

$$H^2(p,q) \;\le \mathrm{TV}(p,q) \le\; 2H(p,q). \tag{A.9}$$

If $\boldsymbol{u}$ and $\boldsymbol{v}$ are the densities corresponding to the joint distributions received from n i.i.d. random variables having densities u and v respectively one has:

$$H^2(\boldsymbol{u},\boldsymbol{v}) \;=\; 2-2\Big(1-\frac{1}{2}H^2(u,v)\Big)^n. \tag{A.10}$$

The proof of Lemma A.8.1 relies on fundamental manipulations with minima and integrals plus the Cauchy-Schwarz inequality and can e.g. be found in the proof of Lemma 14.31 of van der Vaart (1998). The Hellinger distance is especially convenient when considering product measures, as it is, by (A.10), easily expressible in terms of Hellinger distance of the individual measures. This is much more difficult (if not even impossible) for the total variation distance, therefore (A.9) is used as a detour.

A.9 Limit theorems for triangular arrays

Now to the law of large numbers and the classical Lindeberg-Feller central limit theorem for triangular arrays. A triangular array of random vectors is a row-wise independent sequence $\boldsymbol{X}_{n,k_n}$. The generalization compared to the standard central limit theorem is that the distributions of $\boldsymbol{X}_{n,k_n}$ may depend on n. For such an array, a law of large numbers can be stated as follows.

Theorem A.9.1. *For each n let $\boldsymbol{X}_{n,1},\ldots,\boldsymbol{X}_{n,k_n}$ be independent random vectors such that, as $n\to\infty$,*

$$\sum_{i=1}^{k_n}\mathbb{E}\min\Big(\|\boldsymbol{X}_{n,i}\|,\|\boldsymbol{X}_{n,i}\|^2\Big) \;\to\; 0.$$

Then

$$\sum_{i=1}^{k_n}\Big(\boldsymbol{X}_{n,i}-\mathbb{E}\,\boldsymbol{X}_{n,i}\Big) \;\to_{\mathrm{p}}\; \mathbf{0}.$$

The next theorem gives the corresponding central limit theorem.

Theorem A.9.2. *For each n let $\boldsymbol{X}_{n,1}, \ldots, \boldsymbol{X}_{n,k_n}$ be independent random vectors with finite variances such that the Lindeberg condition*

$$\sum_{i=1}^{k_n} \mathbb{E}\Big(\|\boldsymbol{X}_{n,i}^2\| \Big) 1_{\{\|\boldsymbol{X}_{n,i}\| > \varepsilon\}} \quad \to \quad 0 \tag{A.11}$$

holds for every $\varepsilon > 0$ and

$$\sum_{i=1}^{k_n} \operatorname{Var} \boldsymbol{X}_{n,i} \quad \to \quad \Sigma.$$

Then

$$\sum_{i=1}^{k_n} \Big(\boldsymbol{X}_{n,i} - \mathbb{E}\, \boldsymbol{X}_{n,i} \Big) \quad \to_{\mathcal{D}} \quad \mathcal{N}_{\mathrm{p}}(\mathbf{0}, \Sigma).$$

In applications, as in the proof of Theorem 5.4.1, often $k_n = n$. Proofs can e.g. be found in Borovkov (1998), or for the latter Theorem in van der Vaart (1998), Proposition 2.27.

A.10 Some formulas from multivariate statistics

Here we give two lemmas that are used in matrix manipulations in Section 5.2. The first result is about inversion of block matrices.

Lemma A.10.1. *Let $\mathbf{A}$ be a $r \times r$ non-singular matrix, $\mathbf{B}$ a $r \times s$ matrix, $\mathbf{C}$ a $s \times r$ matrix and $\mathbf{D}$ a non-singular $s \times s$ matrix such that $\mathbf{T} := \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ is non-singular. The inverse of the $(r+s) \times (r+s)$ matrix*

$$\mathbf{M} \quad := \quad \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$$

is then:

$$\mathbf{M}^{-1} \quad = \quad \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{T}^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{T}^{-1} \\ -\mathbf{T}^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{T}^{-1} \end{pmatrix}.$$

This lemma can be proven explicitly showing that $\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$. Using the notation of Lemma A.10.1 the next lemma provides another shortcut useful in manipulations of block matrices.

Lemma A.10.2. *Let* $\boldsymbol{v} \in \mathbb{R}^{r+s}$, $\boldsymbol{v}_1 = (v_i)_{i=1}^{r}$ *and* $\boldsymbol{v}_2 = (v_i)_{i=r+1}^{s}$. *Then:*

$$\boldsymbol{v}^\top \mathbf{M}^{-1}\boldsymbol{v} \;=\; \boldsymbol{v}_1^\top \mathbf{A}^{-1}\boldsymbol{v}_1 + (\boldsymbol{v}_2 - \mathbf{C}\mathbf{A}^{-1}\boldsymbol{v}_1)^\top \mathbf{T}^{-1}(\boldsymbol{v}_2 - \mathbf{C}\mathbf{A}^{-1}\boldsymbol{v}_1).$$

Again, this result can be verified through brute force calculation, at best not without taking advantage of Lemma A.10.1.

Appendix B

List of special symbols

Part I: Log-concave density estimation

$\mathrm{L}_1(\mathbb{R})$	real-valued and on $\mathbb{R}$ Lebesgue-integrable functions, p. 2
$\widehat{f}_G$	Grenander density estimator, p. 4
$\widehat{F}_G$	Grenander distribution function estimator, p. 5
X	log-concave random variable, having distribution function F with log-concave Lebesgue density function f, p. 15
F	distribution function $F : \mathbb{R} \to [0, 1]$ on the real line, having log-concave Lebesgue-density f, p. 15
f	density function of F with respect to Lebesgue measure, p. 15
φ	logarithm of f, p. 15
$d_1 * d_2$	convolution for two density functions $d_1, d_2 \in \mathrm{L}_1(\mathbb{R})$, p. 15
λ	hazard rate function derived from f and F, p. 17
n	number of order statistics under consideration (sample size), p. 21
X_i	$i = 1, \ldots, n$, $X_1 < \ldots < X_n$ i.i.d. order statistics, all having distribution function F, p. 21
L_n	general maximum log-likelihood functional, p. 21
$\mathbb{F}_n$	empirical distribution function for a sample $X_1 < \ldots < X_n$, p. 21
1_A	indicator function for a condition A
$\Psi_n(\varphi)$	maximum log-likelihood functional, depending on φ, such that its exponentiated minimizer is a probability density, p. 22
$\widehat{\varphi}_n$	maximum likelihood estimator of φ, p. 22
$\widehat{f}_n$	maximum likelihood estimator of f, p. 22

$\widehat{F}_n$	maximum likelihood estimator of F, p. 22
$\boldsymbol{v}$	general vector notation, $\boldsymbol{v} := (v_1, \ldots, v_n)$, p. 23
$\widehat{\boldsymbol{\varphi}}$	the piecewise linear function $\widehat{\varphi}_n$, viewed as a vector of its knot points, p. 23
$\mathcal{S}(h_n)$	set of knots of a piecewise linear continuous function h_n, p. 24
$\mu(G)$	mean of a distribution function G, p. 25
$\mathrm{Var}(G)$	variance of a distribution function G, p. 25
ρ_n	$\rho_n = \log(n)/n$, p. 27
$\|g\|_\infty^I$	uniform norm of a function g on an interval I, p. 27
T	fixed compact interval $[A, B]$ with endpoints $A < B$, p. 27
$\mathcal{H}^{\beta,L}(T)$	Hölder class of functions for an exponent β and a constant L on a compact interval T, p. 27
$\to_{\mathrm{p}}$	convergence in probability, p. 27. Applied to vectors this operator is to be understood componentwise.
$\to_{\mathcal{D}}$	convergence in law, p. 27. Applied to vectors this operator is to be understood componentwise.
$=_{\mathcal{D}}$	equality in law, p. 27. Applied to vectors this operator is to be understood componentwise.
$\widehat{F}_{n,h}$	estimator of F based on a kernel with bandwidth h, p. 32
$\widehat{\lambda}_n$	estimator of λ based on $\widehat{f}_n$ and $\widehat{F}_n$, p. 32
$\|\boldsymbol{x}\|_2$	L_2-norm for a vector $\boldsymbol{x}$, p. 34
$\mathcal{D}^1(g)$	class of all functions Δ such that $g + t\Delta$ is concave for some $t > 0$ and a concave function g, p. 36
$\mathcal{D}^2(g)$	all piecewise linear functions Δ such that any knot q of Δ fulfills either (3.14) or (3.15), for a concave function g, p. 36
$\mathcal{D}^3(g)$	continuous and piecewise linear functions in $\mathcal{D}^2(g)$ with knots only in $\mathcal{S}(g)$, p. 36
$\mathbb{E}(X)$	expectation for a random variable $X \in \mathrm{L}_1(\mathbb{R})$, p. 49
$(g)_+$	positive part of a real-valued function g: $(g)_+ := \max\{0, g\}$, p. 56
$\#A$	number of elements of a set A, p. 49
Δv_i	difference of two successive elements of a vector: $\Delta v_i := v_i - v_{i-1}$ for $\boldsymbol{v} \in \mathbb{R}^n$ and $i = 2, \ldots, n$, p. 64

$\mathbf{A}$ general notation for a $m \times n$ matrix where the elements are

$$\mathbf{A} = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{m1} & \cdots & A_{mn} \end{pmatrix},$$

p. 65

$\boldsymbol{x} \leq \boldsymbol{y}$ for two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we say that $\boldsymbol{x} \leq \boldsymbol{y}$ holds if $x_i \leq y_i$ for all $i = 1, \ldots, n$. Equality is likewise, p. 65

$\boldsymbol{x}^\top$ transposed vector $\boldsymbol{x}$, p. 66

$\|\boldsymbol{x}\|_{\mathbf{A}}$ norm of the vector $\boldsymbol{x}$ with respect to the matrix A: $\|\boldsymbol{x}\|_{\mathbf{A}} := \sqrt{\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}}$, p. 67

$\operatorname{diag}(\boldsymbol{x})$ diagonal matrix with the vector $\boldsymbol{x}$ on its diagonal, p. 69

$\operatorname{diag}(\mathbf{A})$ vector consisting of the diagonal of the matrix $\mathbf{A}$, p. 77

$\mathcal{N}(\mu, \sigma)$ Univariate Normal distribution with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$, p. 78

$\Gamma(\alpha, \beta)$ Gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$, p. 78

Part II: Bump hunting

$p_{\boldsymbol{\theta}}$ parametric density function, with parameter $\boldsymbol{\theta} \in \Theta \in \mathbb{R}^p$, p. 89

$\mathbb{E}_{\boldsymbol{\theta}}\, u(X)$ expectation of a function u such that $up_{\boldsymbol{\theta}} \in \mathrm{L}_1(\mathbb{R})$ for a random variable X, where X has density function $p_{\boldsymbol{\theta}}$, p. 89

$\overline{u(X_i)}$ sample mean of the random variables $u(X_i)$, $i = 1, \ldots, n$, p. 120

$\ell_{\boldsymbol{\theta}}$ log of $p_{\boldsymbol{\theta}}$, p. 90

$\dot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}$ score function of $p_{\boldsymbol{\theta}}$, p. 90

$\mathbf{I}_{\boldsymbol{\theta}}$ Fisher information matrix of $p_{\boldsymbol{\theta}}$, p. 90

$\tilde{\boldsymbol{x}}$ For a given vector in $\mathbb{R}^k$, $\tilde{\boldsymbol{x}} \in \mathbb{R}^{k-1}$ is the vector omitting the last component, p. 92

$\mathcal{N}_{\mathrm{p}}(\boldsymbol{\mu}, \Sigma)$ p-variate Normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$, p. 91

$\chi^2_1(p)$	χ^2-distribution with one degree of freedom and a non-centrality parameter $p \geq 0$, p. 93
$\chi^2_{1;\alpha}$	α-quantile of a χ^2-distribution with one degree of freedom, $\alpha \in (0,1)$, p. 94
$\chi^2_1(p,.)$	χ^2-distribution function with one degree of freedom and non-centrality parameter $p \geq 0$, p. 94
z_α	α-quantile of a standard normal distribution, $\alpha \in (0,1)$, p. 95
Φ_1	distribution function of a standard normal distribution, p. 96
$f_{\theta,\eta}$	specific two-parameter density used to define bump hunting test statistic, p. 97
f	twice continuously differentiable density function, p. 103
X_i	order statistics $X_0 < \ldots < X_n$, having distribution function F and density function f, p. 103
f_{jk}	"local" density function, p. 103
$\mathcal{I}_{jk}$	intervals spanned by two order statistics: $\mathcal{I}_{jk} = (X_j, X_k)$, p. 103
$X_{i;j,k}$	local order statistics, p. 103
$TV(f,g)$	total variation distance between two densities f and g, p. 105
$\mathcal{C}^{\cap}_{l,m,n}$	set of intervals whereon multiscale test claims that f is convex, p. 108
$\mathcal{C}^{\cup}_{l,m,n}$	set of intervals whereon multiscale test claims that f is concave, p. 108
$\mathcal{B}^{\cap}_{l,m,n}(\alpha)$	set of intervals whereon multiscale test claims that f has a bump, p. 108
$\mathcal{B}^{\cup}_{l,m,n}(\alpha)$	set of intervals whereon multiscale test claims that f has a antibump, p. 108
$\mathcal{L}(X)$	distribution of the random variable X, p. 109

Bibliography

An, M. Y. (1995). Log-concave probability distributions: Theory and statistical testing. *Preprint, Economics Department, Duke University, Durham, N.C.*

An, M. Y. (1998). Logconcavity versus logconvexity: A complete characterization. *J. Econ. Theory*, **80**, 350–369.

Anevski, D. (1994). Estimating the derivative of a convex density. *Preprint, Department of Mathematical Statistics, University of Lund.*

Anevski, D. (2003). Estimating the derivative of a convex density. *Stat. Neerl.*, **57**, 245–257.

Arnold, B. C., Balakrishnan, N., Nagaraja, H. N. (1992). *A First Course in Order Statistics*, Wiley, New York.

Bagnoli, M., Bergstrom, T. (1989). Log-concave probability and its applications. *Preprint, University of Michigan.*

Bagnoli, M., Bergstrom, T. (2005). Log-concave probability and its applications. *Econ. Theory*, **26**, 445–469.

Balabdaoui, F., Wellner, J. A. (2004a). Estimation of a k-monotone density, part 1: characterizations, consistency, and minimax lower bounds. *Technical report 459, Department of Statistics, University of Washington.*

Balabdaoui, F., Wellner, J. A. (2004b). Estimation of a k-monotone density, part 2: algorithms for computation and numerical results. *Technical report 460, Department of Statistics, University of Washington.*

Balabdaoui, F., Wellner, J. A. (2004c). Estimation of a k-monotone density, part 3: limiting Gaussian versions of the problem. *Technical report 461, Department of Statistics, University of Washington.*

Balabdaoui, F., Wellner, J. A. (2004d). Estimation of a k-monotone density, part 4: limit distribution theory and the spline connection. *Technical report 462, Department of Statistics, University of Washington.*

BARLOW, E. B., BARTHOLOMEW, D. J., BREMNER, J. M., BRUNK, H. D. (1972). *Statistical Inference under Order Restrictions*, Wiley, New York.

BARLOW, E. B., PROSCHAN, F. (1975). *Statistical Theory of Reliability and Life Testing*, To begin with, Silver.

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*, Wiley, New York.

BICKEL, P., FAN, J. (1996). Some problems on the estimation of unimodal densities. *Stat. Sin.*, **6**, 23–45.

BOROVKOV, A. A. (1998). *Mathematical Statistics*, Gordon and Breach Science Publishers, Amsterdam.

CAPLIN, A., NALEBUFF, B. (1991a). Aggregation and social choice: A mean voter theorem. *Econometrica*, **59**, 1–24.

CAPLIN, A., NALEBUFF, B. (1991b). Aggregation and imperfect competition: On the existence of equilibrium. *Econometrica*, **59**, 26–60.

CHAUDHURI, P., MARRON, J. S. (1999). SiZer for the exploration of structures in curves. *J. Amer. Statist. Assoc.*, **94**, 807–823.

CHAUDHURI, P., MARRON, J. S. (2000). Scale space view of curve estimation. *Ann. Statist.*, **28**, 408–428.

CHENG, M. Y., HALL, P. (1999). Mode testing in difficult cases. *Ann. Statist.*, **27**, 1294–1315.

DEVROYE, L. (1986). *Non-Uniform Random Variate Generation*, Springer, New York.

DONOHO, D. (1988). One-sided inference about functionals of a density. *Ann. Statist.*, **16**, 1390–1420.

DÜMBGEN, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *Ann. Statist.*, **26**, 288–314.

DÜMBGEN, L. (2002). Application of local rank tests to nonparametric regression. *J. Nonparametric Stat.*, **14**, 511–537.

DÜMBGEN, L., FREITAG S., JONGBLOED, G. (2004). Consistency of concave regression, with an application to current status data. *Math. Methods Stat.*, **13**, 69–81.

DÜMBGEN, L., FREITAG S., JONGBLOED, G. (2006). Estimating a Unimodal Distribution from Interval-Censored Data. *J. Amer. Statist. Assoc., to appear.*

DÜMBGEN, L., SPOKOINY, V. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.*, **29**, 124–152.

DÜMBGEN, L., WALTHER, G. (2006). Multiscale inference about a density. *Technical report 56, IMSV, University of Bern.*

DURBIN, J. (1973). *Distribution theory for tests based on the sample distribution function*, SIAM, Philadelphia.

DVORETZKY, A, KIEFER, J. C., WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, **33**, 642–669.

EGGERMONT, P. P. B., LARICCIA, V. N. (1999). Optimal convergence rates for Good's nonparametric maximum likelihood density estimation. *Ann. Statist.*, **27**, 1600–1615.

EGGERMONT, P. P. B., LARICCIA, V. N. (2000). Maximum Likelihood Estimation of Smooth Monotone and Unimodal Densities. *Ann. Statist.*, **28**, 922–947.

EGGERMONT, P. P. B., LARICCIA, V. N. (2001). *Maximum Penalized Likelihood Estimation, Volume 1: Density Estimation*, Springer, New York.

FIACCO, A. V., MCCORMICK, G. P. (1968). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York.

VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*, Cambridge University Press, Cambridge.

GOOD, I. J. (1971). A nonparametric roughness penalty for probability densities. *Nature*, **229**, 29–30.

GRENANDER, U. (1956). On the theory of mortality measurement, part II. *Skandinavisk Aktuarietidskrift*, **39**, 125–153.

GROENEBOOM, P. (1985). *Estimating a monotone density*, Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Volume II, Lucien M. LeCam and Richard A. Ohlsen eds.

GROENEBOOM, P. (1988). Brownian motion with a parabolic drift and Airy functions. *Probab. Theory Relat. Fields*, **81**, 79–109.

GROENEBOOM, P., HOOGHIEMSTRA, G., LOPUHAÄ, H. P. (1999). Asymptotic normality of the L_1 error of the Grenander estimator. *Ann. Statist.*, **27**, 1316–1347.

GROENEBOOM, P., JONGBLOED, G., WELLNER, J.A. (2001a). A canonical process for estimation of convex functions: the "invelope" of integrated Brownian motion $+t^4$. *Ann. Statist.*, **29**, 1620–1652.

GROENEBOOM, P., JONGBLOED, G., WELLNER, J.A. (2001b). Estimation of a convex function: characterization and asymptotic theory. *Ann. Statist.*, **29**, 1653–1698.

GROENEBOOM, P., JONGBLOED, G. (2003). Density estimation in the uniform deconvolution model. *Stat. Neerl.*, **57**, 136–157.

GROENEBOOM, P., JONGBLOED, G., AND WELLNER, J. A. (2003). The support reduction algorithm for computing nonparametric function estimates in mixture models. *Preprint, Department of Mathematics, Vrije Universiteit Amsterdam.*

GROENEBOOM, P., WELLNER, J.A. (1992). *Information bounds and nonparametric maximum likelihood estimation.*, DMV Seminar, 19, Birkhuser Verlag, Basel.

HALL, P., HUANG, L. S., GIFFORD, J. A., GIJBELS, I. (2001). Nonparametric estimation of hazard rate under the constraint of monotonicity. *Comput. Graph. Statist.*, **10**, 592–614.

HALL, P., HUANG, L. S. (2002). Unimodal density estimation using kernel methods. *Statist. Sinica*, **12**, 965–990.

HALL, P., VAN KEILEGOM, I. (2005). Testing for monotone increasing hazard rate. *Ann. Statist.*, **33**, 1109–1137.

HAMPEL, F. R. (1987). *Design, modelling and analysis of some biological datasets*, Design, data and analysis, by some friends of Cuthbert Daniel, C.L. Mallows editor, Wiley, New York.

HARTIGAN, J. A., HARTIGAN, P. M. (1985). The dip test of unimodality. *Ann. Statist.*, **13**, 70–84.

HÜSLER, A. (2005). Estimating Log-Concave Densities from Censored Data. *Technical report 53, IMSV, University of Bern.*

IBRAGIMOV, I. A. (1956). On the composition of unimodal distributions. *Theory Probab. Appl.*, **1**, 255–260.

JONGBLOED, G. (1995). *Three Statistical Inverse Problems*, Ph.D. Dissertation, Delft University of Technology.

JONGBLOED, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *J. Comp. Graph. Statist.*, **7**, 310–321.

JONKER, M., VAN DER VAART, A. (2001). A semi-parametric model for censored and passively registered data. *Bernoulli*, **7**, 1–31.

KARLIN, S. (1968). *Total positivity, Volume 1*, Stanford University Press, Stanford.

KIEFER, J., WOLFOWITZ, J. (1976). Asymptotically minimax estimation of concave and convex distribution functions. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **34**, 73–85.

KULIKOV, V. N., LOPUHAÄ, H. P. (2005a). Asymptotic normality of the L_k-error of the Grenander estimator. *Ann. Statist.*, **33**, 2228–2255.

KULIKOV, V. N., LOPUHAÄ, H. P. (2005b). The limit process of the difference between the empirical distribution function and its concave majorant. *Preprint, Delft University of Technology.*

KULIKOV, V. N., LOPUHAÄ, H. P. (2006). Distribution of global measures of deviation between the empirical distribution function and its concave majorant. *Preprint, Delft University of Technology.*

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, Wiley, New York.

LÉVY, P. (1937). *Theorie de l'Addition des Variables Aleatoires*, Gauthier-Villars, Paris.

MARRON, J. S., WAND, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.*, **20**, 712–736.

MARSHALL, A. W. (1970). Discussion of Barlow and van Zwet's papers. In M. L. Puri (ed.), *Nonparametric Techniques in Statistical Inference.* Cambridge University Press, 175–176.

MASSART, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, **18**, 1269–1283.

MEYER, C. M., WOODROOFE, M. (2004). Consistent maximum likelihood estimation of a unimodal density using shape restrictions. *Can. J. Stat.*, **32**, 85–100.

MÜLLER, S., RUFIBACH, K. (2006). Smoothed semi-parametric tail index estimation. *Technical report 58, IMSV, University of Bern.*

POLLARD, D. (2002). *A user's guide to measure theoretic probability*, Cambridge University Press, Cambrigde.

POLONIK, W. (1995). Density estimation under qualitative assumptions in higher dimensions. *J. Multivariate Anal.*, **55**, 61–81.

POLONIK, W. (1998). The silhouette, concentration functions and ML-density estimation under order restrictions. *Ann. Statist.*, **26**, 1857–1877.

R DEVELOPMENT CORE TEAM. (2005). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, ISBN 3–900051–07–0, URL http://www.R-project.org, Vienna, Austria.

RAO, PRAKASA (1969). Estimation of a unimodal density. *Sankhya Ser. A*, **31**, 23–36.

REISS, R. D., THOMAS, M. (2001). *Statistical Analysis of Extreme Values*, Birkhäuser, Basel.

ROBERTSON, T., WRIGHT, F. T., DYKSTRA, R. L. (1988). *Order restricted statistical inference*, Wiley, New York.

SENGUPTA, D., PAUL, D. (2004). Some tests for log-concavity of life distributions. *Preprint, Department of statistics, Stanford University.*

SHAO, J. (2003). *Mathematical Statistics*, Springer, New York.

SHORACK, G. R., WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*, Wiley, New York.

SILVERMAN, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Royal. Statist. Society, Series B*, **43**, 97–99.

STUTE, W. (1982). The oscillation behaviour of empirical processes.. *Ann. Prob.*, **10**, 86–107.

TERLAKY, T. (1996). *Interior Point Methods of Mathematical Programming*, Kluwer Academic Publishers, Dordrecht.

TERLAKY, T., VIAL, J-PH. (1998). Computing maximum likelihood estimators of convex density functions. *SIAM J. Scientific Comp.*, **19**, 675–694.

VAN DER VAART, A. (1998). *Asymptotic Statistics*, Cambridge University Press, Cambrigde.

VAN DER VAART, A., VAN DER LAAN, M. J. (2003). Smooth estimation of a monotone density. *Statistics*, **37**, 189–203.

WALTHER, G. (2000). Detecting the presence of mixing with multiscale maximum likelihood. *J. Am. Stat. Assoc.*, **97**, 508–514.

WALTHER, G. (2001). Multiscale maximum likelihood analysis of a semiparametric model, with applications. *Ann. Statist.*, **29**, 1297–1319.

WEGMAN, E. J. (1970). Maximum likelihood estimation of a unimodal density function. *Ann. Math. Statist.*, **41**, 457–471.

WOODROOFE, M, SUN, J. (1993). A penalized maximum likelihood estimate of $f(0+)$ when f is non-increasing. *Statist. Sinica*, **3**, 501–515.

WRIGHT, J. W. (1997). *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia.

WRIGHT, J. W. (1998). Superlinear Convergence of a Stabilized SQP Method to a Degenerate Solution. *Comput. Optim. Appl.*, **11**, 253-275.

Curriculum Vitae

Name:	Kaspar Rufibach
Date of birth	April 23, 1977
Nationality	Swiss

Education

1984-1993	Schools in Meiringen and Interlaken
1993-1997	Gymnasium Interlaken
1997	Matura Typus C (University Entrance Exam)
1997-2001	University of Bern Studies in Statistics, Mathematics and Economics
2001	Diploma in Statistics "Phasenrekonstruktion von Gleichverteilungen in $\mathbb{R}^d$" (supervised by Prof. Dr. H. Carnal)
2002-2006	PhD in Mathematics "Log-concave Density Estimation and Bump Hunting for i.i.d. Observations" (supervised by Prof. Dr. L. Dümbgen)

Professional Activities

1999-2002	Internships and consultant, UBS Switzerland
2000-today	Assistant, Institute for Mathematical Statistics and Actuarial Science, University of Bern
2002-2006	Biostatistician, Swiss Group for Clinical Cancer Research SAKK, Bern