Variability-Reducing Quality Control Methods in Photovoltaics
Final Report (1.9.2010 - 30.12.2012)

Prof. Dr. Ansgar Steland, Dr. Andrey Pepelyshev (RWTH Aachen)

Partners:

Prof. Dr. E. Rafajłowicz (Wrocław UoT)

Dr. Werner Herrmann (TÜV Rheinland)

Dr. G. Kleiss, Fr. C. Goltz (SolarWorld)

Dr. Avellan-Hampe (AVANCIS)

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Umwelt, Naturschutz und Reaktorsicherheit unter dem Förderkennzeichen 0325226 gefördert.

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

12. Februar 2013

0.0

Inhaltsverzeichnis

Teil 1

Zusammenfassung und Kurzdarstellung

Kapitel 1.1

Aufgabenstellung

1.1.1 Motivation

Die Qualitätssicherung von PV-Modulen muss in aller Regel auf Stichproben basieren, die aus Lieferungen oder Lots gezogen werden. Ziel ist es primär, Lieferungen aufzudecken, die zu viele Module enthalten, die nicht den Anforderungen an die Modulleistung entsprechen. Die Anforderungen resultieren hierbei aus Vereinbarungen mit Abnehmern (Kunden) oder, bei interner Anwendung bei Herstellern und Händlern, aus selbstgesteckten Qualitätszielen zur Erreichung der Unternehmensziele. Die Qualitätsanforderungen können aus den technischen Spezifikationen über die Leistungsverteilung in den Anteil nicht-konformer Module (die Qualitätslage) umgerechnet werden. Dieser Anteil ist aus statistischer Sicht der relevante Parameter zur Konstruktion von Prüfverfahren. Er bestimmt insbesondere die Kosten von Ersatzlieferungen im Garantiefall.

Die wissenschaftliche Methode der Wahl ist die *statistische Annahmeprüfung*, die optimale Stichprobenpläne zur Verfügung stellt, bei denen der benötigte Stichprobenumfang der Prüflinge unter Einhaltung vorgegebener Fehlerwahrscheinlichkeiten für Produzent und Abnehmer minimiert wird. Diese Fehlerwahrscheinlichkeiten begrenzen hierbei die Risiken von Fehlentscheidungen, die mit stichprobenbasierten Entscheidungen zwangsläufig verbunden sind. Für den Produzenten besteht die Fehlentscheidung in der Ablehnung einer Lieferung guter Qualität, für den Konsumenten in der Annahme einer Lieferung schlechter Qualität.

Da die Kosten der Qualitätskontrolle i.W. linear im Stichprobenumfang skalieren, werden somit durch die Anwendung statistischer Prüfpläne automatisch die hiermit verbunde-

nen Kosten minimiert. Hierin besteht das substantielle *Kostensenkungspotential*, da somit eine kostenminimale systematische Qualitätskontrolle ausgehender Lieferungen (bzw. definierter Lots) möglich wird.

Bei der Anwendung in der Photovoltaik stoßen die etablierten Verfahren, die auch Eingang in die entsprechenden Normen gefunden haben, auf Grenzen: Photovoltaische Qualitätsdaten verletzen typischerweise die grundlegende Annahme der Variablenprüfung, die Normalverteilungsannahme, so dass dieses Verfahren nur selten anwendbar ist. In diesem Fall erfolgt eine Attributprüfung, die jedoch die in den photovoltaischen Messungen vorhandene statistische Information nur unzureichend ausnutzen: Es wird lediglich die (binäre) Information, ob ein Modul konform oder nicht konform ist, verwendet. Infolgedessen resultieren sehr hohe Stichprobenumfänge, die in der industriellen Praxis häufig nicht realisierbar sind. Dies trifft auch auf die Photovoltaik zu, da hier neben den ohnehin beträchtlichen Kosten einer präzisen Messung in einem spezialisierten Labor hohe Kosten für Demontage, Transport und anschließende Montage anfallen.

In Vorarbeiten zu diesem Forschungsvorhaben wurden an der RWTH spezielle neue Prüfverfahren entwickelt, welche die in der Regel vorliegenden industriellen Flasherdaten ausnutzen, um valide (d.h. gültige) und asymptotisch optimale Verfahren für photovoltaische Leistungsmessungen zu konstruieren.

Da diese Verfahren den benötigten Stichprobenumfang aus der Flasher-Stichprobe berechnen, die als Zufallsstichprobe aus der Produktion angesehen wird, ist dieser Prüfumfang zufälligen Schwankungen unterworfen. Simulationsstudien und reale Datenanalysen haben gezeigt, dass diese Variabilität beträchtlich ist, so dass die Anwendung im Grunde nur bei sehr großen Flasherlisten zu stabilen Prüfumfängen führt. Aus diesem Grund sollten in diesem Forschungsvorhaben varianz-mininmale Verfahren entwickelt, optimiert und durch umfassende Simulationsstudien und Benchmarking an realen Industriedaten analysiert und evaluiert werden.

Ein weiteres wichtiges Anliegen des Vorhabens war die Erweiterung der Verfahren auf Out-of-Spec-Prüfpläne, welche sowohl Lieferungen mit zu vielen Modulen zu geringer Leistung aufdecken als auch Lieferungen mit zu vielen Modulen zu hoher Leistung. Hintergrund ist zum Einen, dass technische Spezifikationen in Pflichtenheften typischerweise in Form von Spezifikationsintervallen fixiert werden. Für eine interne Qualitätssicherung sind somit Out-of-Spec-Verfahren von hohem Interesse, um Abweichungen von Spezifikationen auf Lotebene erkennen zu können.

Diese zweiseitge Spezifikation wird jedoch von Verfahren der Variablenprüfung nicht berücksichtigt, sondern eine einseitige Spezifikation angenommen. Zum Anderen ist es angesichts der Einteilung der produzierten Module in Modulklassen aus Herstellersicht von Interesse, Lieferungen bzw. Lots, in denen zu viele höherwertige Module sind, zu erkennen, um für produzierte Ware angemessene Preise erzielen zu können. Somit kann die Anwendung der Verfahren Bestandteil der internen Qualitätssicherung werden. Out-of-Spec-Verfahren erleichtert auch die Planung von PV-Systemen, da die benötigten Wechselrichter in der Regel aus Kostengründen nicht für theoretisch erreichbare Spitzenleistung ausgelegt werden. Sie schalten aber bei zu hoher Energieproduktion ab, um Ausfälle und eine Reduktion der Lebensdauer der elektronischen Komponenten aufgrund der resultierenden thermischen Belastung zu vermeiden. Schließlich erschließt eine genaue Kontrolle der Leistung ausgehender Lieferungen für die Photovoltaik-Industrie das Potential, bei entsprechender Marktlage für die gelieferte (Mehr-) Leistung einen angemessen Preis erzielen zu könne.

Mit der angestrebten Erweiterung der Variablenprüfung auf zweiseitige Spezifikationsprobleme wurde ein *offenes ungelöstes Problem der Statistik* addressiert. Obwohl die methodischen und theoretischen Grundlagen der Attribut- und Variablenprüfung schon vor vielen Jahrzehnten gelegt wurden, blieb diese insbesondere für die Photovoltaik wichtige Erweiterung bis heute ungelöst.

1.1.2 Problemformulierung

Es ist bekannt, dass der optimale Annahme-Prüfplan für Leistungsmessungen X, die einer Verteilung F, der *Leistungsverteilung* folgen, von ebendieser Verteilung F abhängen, die als unbekannt angenommen werden muss.

Der optimale Prüfplan wird nun aus einer Flasher-Stichprobe vom Umfang m, die einer Verteilung G folgt, geschätzt. Hierzu muss die Flasher-Stichprobe gleichartige Module umfassen, die hinsichtlich ihrer elektrischen Eigenschaften mit dem zu prüfenden Lot übereinstimmen. In der Praxis stellen die zu prüfende Module eine Teilauswahl der Flasher-Stichprobe dar oder stimmen mit den Modulen der Flasher-Stichprobe überein. Letzteres ist der typische und häufigste Fall. Valide statistische Verfahren müssen dies so resultierende Form der Abhängigkeit zwischen Prüfmessung und Flashermessung zulassen.

Im Grundmodell muss zudem abgebildet werden, dass sich die Flasher-Leistungsmessungen

von den Labormessungen (bzw. Nachmessungen) systematisch unterscheiden. Diese in der Praxis bekannten und empirisch auch nachgewiesenen systematischen Messunterschiede erklären sich durch die unterschiedlichen verwendeten Mess-Systeme (Flasher), deren Eichung/Kalibrierung, nicht perfekt kontrollierbaren Einflußgrößen sowie der zeitlichen Differenz zwischen Flashermessung und prüfender Messung.

Hauptziele des Vorhabens waren:

- Die Entwicklung varianz-minimierender Verfahren der Annahmeprüfung bei Ausnutzung von Flasherlisten.
- Die Erweiterung der Methodologie auf Out-of-Spec-Verfahren.

Die Entwicklung neuer Verfahren, welche die Varianz des resultierenden Stichprobenumfangs minimieren. Da die Verfahren typischerweise von Tuning-Konstanten (Parametern) abhängen, sollten diese *datanadaptiv* aus den vorliegenden Stichproben ermittelt werden und in die Optimierung der Verfahren einfließen. Von ausgewählten Verfahren sollte die statistische Konsistenz nachgewiesen werden, um eine praktische Anwendung auf ein solides wissenschaftliches Fundament zu stellen.

Die Bewertung der Verfahren sollte durch drei Ansätze erfolgen:

- Durch Computersimulationen basierend auf idealisierten Leistungsverteilungen erfolgen, die sich an realen Verteilungen orientieren, um das statistische Verhalten der Verfahren an typischen Modellwelten zu studieren.
- Durch Anwendung der entwickelten Verfahren an ausgewählten realen Datensätze unserer Industriepartner, um die Praxisanwendung darzustellen und eine Beurteilung auf Fallebene zu ermöglichen.
- Durch Bootstrap-Verfahren, bei denen reale Datensätze als Verteilungsmodelle für die Leistungsverteilung verwendet werden.

Sofern im Rahmen der Ressourcen möglich, sollte eine prototypische Implementierung für den Praxiseinsatz erfolgen, um eine Aufnahme in das Softwaretool APOS vorzubereiten oder sogar zu ermöglichen (Fast Track Technology Transfer). Neben der Publikationstätigkeit ist APOS unser Mittel, die Forschungsergebnisse des Vorhabens der Allgemeinheit – und insbesondere auch der Industrie – zur Verfügung zu stellen und so eine schnelle Verbreitung zu ermöglichen.

Schließlich sollte ein eindimensionaler Bewertungsmaßstab gewählt bzw. entwickelt werden, der sowohl statistischen Kriterien genügt als auch Anforderungen aus der industriellen Praxis, um die entwickelten Verfahren vergleichen zu können.

Basierend auf den so gewonnen Erkenntnissen sollte nach Möglichkeit ein Ranking der entwickelten Verfahren und Verfahrensvarianten vorgenommen werden, um hieraus Empfehlungen für praktische Anwendung abzuleiten. Dies ist insbesondere in Hinblick auf eine Standardisierung von Prüfprozeduren und ihrer Auswertung von hoher Bedeutung.

Mit dem Problem der Entwicklung von Out-of-Spec-Verfahren stand ferner ein vollkommen ungelöstes Problem der Statistischen Qualitätssicherung auf der Agenda, das angegangen werden sollte. Aus diesem Grund konnnte und wurde hier kein konkretes Forschungsvorgehen definiert.

Kapitel 1.2

Planung und Ablauf des Vorhabens

Entsprechend dem Projektplan wurden verschiedene methodische Ansätze systematisch auf ihr Problemlösungspotential untersucht, entsprechende photovoltaische Prüfverfahren konzipiert und durch erste Computerexperimente untersucht. Die Ergebnisse der Computerexperimente wurden verwendet, um die Spreu vom Weizen zu trennen und nicht-kompetitive Verfahren frühzeitig verwerfen zu können, so dass die Ressourcen auf die aussichtsreichsten Kandidaten konzentriert werden konnten. Die Entscheidung, ob eine Verdichtung des Kandidatenfeldes erfolgen sollte oder nicht, wurde an zwei Milestones getroffen.

Die Verfahren wurden – soweit möglich – auf ihre mathematisch-statistischen Eigenschaften untersucht, wobei hierbei eine Fokussierung auf die für praktische Anwendungen relevanten Eigenschaften erfolgte.

Für alle methodischen Ansätze wurden Verfahrensvarianten entwickelt, welche sich primär durch die verwendeten Strategien für die Wahl der Tuning-Konstanten unterscheiden. Hier gibt es in der Regel verschiedene Möglichkeiten, die auch oftmals einen hohen auf die Performanz haben. Die so gewonnenen Verfahrensvarianten wurde in umfangreichen Simulationsstudien auf ihre Eigenschaften untersucht. Hierbei standen wichtige Verteilungsparameter der Verteilung des Prüfumfangs im Vordergrund. Die Simulationen wurden zu einem großen Teil auf dem lokalen Computing-Cluster unter Ausnutzung paralleler Programmiertechniken durchgeführt. Die wichtigsten und informativsten Ergebnisse dieser Simulationen sind in diesem Bericht tabellarisch reproduziert.

In einem iterativen Prozess wurden die durch die Simulationen gewonnenen Erkenntnisse intern diskutiert und zur kontinuierlichen Verbesserung der Verfahren und Verfahrensvarianten genutzt, bis sich eine Saturierung ergab oder ein weiterer Ressourceneinsatz

im Hinblick auf das zu erwartende Verbesserungspotential unverhältnismäßig erschien.

Da wir schon während der ersten Projektphase ein neues spektralanalytisches Verfahren (SSA-Verfahren) entwickeln konnte, das in den Computerexperimenten sehr gute Ergebnisse zeigt, wurde der ursprünglich verfolgte Ansatz, die Methoden (s.u.) parallel zu entwickeln und zu untersuchen, aufgegeben und auf ein sequentielles Vorgehen gewechselt, bei dem die Methoden der Reihe untersucht wurden. Hierdurch konnte insbesondere das SSA-Verfahren sehr schnell umfassend entwickelt und untersucht werden.

Die Verfahren basierend auf Bernstein-Polynomen und Bernstein-Durrmeyer-Polynomen wurden in enger methodischer Kooperation mit Prof. E. Rafajłowicz von der Wrocław University of Technology entwickelt. Die Initiierung, Vertiefung und Konsolidierung erfolgte im Rahmen von mehreren Forschungsaufenthalten in Wrocław. Hinzu tragen Diskussionen bei Treffen auf mehreren internationalen Konferenzen.

Neben zwei Kick-Off Meetings mit den Projektpartnern (ein gemeinsamer Termin kam leider nicht zustande), telefonischen bilateralen Gesprächen und Emails wurden wesentliche Zwischenergebnisse im Rahmen eines größeren Projektmeetings im April 2012 ausführlich diskutiert. Somit konnten die abschließenden Aktivitäten bis zum Projektende mit i.W. hohem Detaillierungsgrad einvernehmlich festgelegt und zeitlich geplant werden.

Zwischenergebnisse wurden auf mehreren nationalen und internationalen Konferenzen und Workshops vorgestellt und mit internationalen Experten diskutiert:

- Vortrag über Projektergebnisse (Pepelyshev A., Rafajłowicz E. and Steland A., *Functional Change-Point Asymptotics and Applications in Finance and Renewable Energies*) auf der Statistischen Woche, Leipzig.

- Eingeladener Vortrag über Projektergebnisse (Steland A., *Functional Change-Point Asymptotics and Applications*), ISI 2011 World Congress of Statistics, Dublin, UK.

- Eingeladener Vortrag über Projektergebnisse (Steland A., *Functional Asymptotics and Applications*), IWSM 2011, Invited Talk and Session Organizer, Stanford University, CA, U.S.A.

- Vortrag über Projektergebnisse (Pepelyshev, A. Golyandina, N., Steland, A., *New Method of Nonparametric Density Estimation With Applications in Photovoltaics*, Workshop on Stochastic Models and Their Applications 2011, Wismar, Germany.

- Vortrag über Projektergebnisse (Steland, A., *Quality Assessment in the Presence of Additional Data and Robustness Issues*), Workshop on Stochastic Models and Their Applications 2011, Wismar, Germany.
- Vortrag u.A. über Projektergebnisse (Pepelyshev, A., *Change-Point Detection for Monitoring the Quality of Photovoltaic Modules*), Jahrestagung der Deutschen Statistischen Gesellschaft, 2012, Wien, Österreich.
- Vortrag u.A. über Projektergebnisse (Steland, A., *Ermittlung von repräsentativen Modulstichproben zur Leistungsmessung*), 9. Workshop PV-Modultechnik, Köln.

Kapitel 1.3

Wissenschaftlicher Stand

Die statistische Annahmeprüfung stellt ein klassisches Gebiet der Statistik dar, deren Standardverfahren in einschlägigen Normen (z.B. IEC 60410) Eingang gefunden haben. Diese Verfahren führen jedoch in der Photovoltaik entweder – im Fall der Attributprüfung – zu Prüfumfängen, die in der Praxis nicht umsetzbar sind, oder sie sind – im Fall der Verfahren unter Normalverteilungsannahme – für Leistungsmessungen von PV-Modulen typischerweise nicht anwendbar. Letzeres erklärt sich aus dem empirischen Befund, dass diese Messungen nur in seltenen Fällen einer Normalverteilung folgen. Die theoretisch angebbaren optimalen Prüfverfahren bei Abweichungen von der Normalverteilung hängen jedoch von der unbekannten Verteilung der Messungen ab und sind somit in der Praxis nicht anwendbar. In der Photovoltaik liegen oftmals Flasherlisten des Herstellers vor. Diese Flasher-Messungen sind jedoch in aller Regel aus verschiedenen Gründen nicht hinreichend verlässlich, so dass sich Prüfverfahren nicht (ausschließlich) auf diese stützen können. Sie beinhalten jedoch wertvolle Zusatzinformationen, die zur Konstruktion von Prüfverfahren verwendet werden können.

In unseren Vorarbeiten ist es uns gelungen, ein erstes Standardverfahren zu entwickeln, das die Bestimmung von geeigneten und asymptotisch optimalen Prüfplänen für das vorliegende Problem ermöglicht, vgl. [1, 4]. Unsere Ergebnisse umfassen sowohl die Entwicklung einer völlig neuartigen Methode zur Bestimmung des Prüfumfangs im Falle nicht normalverteilter Messungen, als auch mathematische Studien, welche das Verfahren wissenschaftlich begründen. Verfahren für den Vergleich von Flasher- und Labormessungen wurden in [2] diskutiert. In [1] wurde ein mehrstufiges Entscheidungsverfahren konstruiert, welches die neu entwickelten Verfahren mit klassischen Verfahren integriert und alle

relevanten Anwendungssituationen abdeckt. Diese Methodik wurde in einem Softwaretool, [9], in Industriequalität implementiert. Die finiten statistischen Eigenschaften des von uns entwickelten Verfahrens für nicht-normalverteilte Daten wurden durch Computersimulationen in anwendungsrelevanten Szenarien untersucht, s. [1,3,4], deren drei Kernergebnisse wie folgt zusammengefasst werden können

Das entwickelte Standardverfahren ist in vielen Konstellationen praktisch anwendbar, sofern eine hinreichend große Stichprobe von Flasher-Messungen vorliegt. Es führt in vielen Fällen zu deutlich reduzierten Prüfumfängen und somit zu erheblichen Kostensenkungen. Die für den Nachweis der asymptotischen Normalität benötigte Annahme der Existenz des vierten Moments kann nicht abgeschwächt werden. Der (geschätzte) benötigte Stichprobenumfang der Prüfstichprobe weist eine beträchtliche Variabilität auf. Die empirischen Befunde aus Computersimulationen und Datenanalysen haben gezeigt, dass ein substantieller weiterer Forschungsbedarf besteht mit dem Ziel, Verfahren mit reduzierter Variabilität zu konstruieren. Wir haben in [3] hierzu einen ersten Vorschlag basierend auf einem Kernschätzer mit kreuzvalidierter Bandbreitenwahl gemacht, vgl. auch [8]. Kernschätzer wurde von uns auch für andere Fragestellungen intensiv studiert, s. [5, 6, 7]. Diese Vorarbeiten bildeten zusammen mit der Erkenntnislage aus der Literatur der Expertise unserer Partner, insbesondere der Industriepartner, die Ausgangsbasis des Vorhabens.

Zitierte Publikationen und Vorträge:

1. Herrmann W., Althaus J., Steland A. und Zähle H. Statistical and experimental methods for assessing the power output speci?cation of PV modules. Proceedings of the 21st European Photovoltaic Solar Energy Conference, 2416-2420, 2006.

2. Herrmann W., Herff, W., und Steland, A. Sampling procedures for the validation of PV module output power specification. Proceedings of the 23rd European Photovoltaic Solar Energy Conference, accepted contributed paper, 2009.

3. Steland A. and Herrmann W. Evaluation of photovoltaic modules based on sampling inspection using smoothed empirical quantiles. Progress in Photovoltaics, 18, 1, 1-19, 2010.

4. Steland A. and Zähle H. Sampling inspection by variables: Nonparametric setting. Statistica Neerlandica, 63 (1), 101-123, 2009.

5. Steland A. Sequential control of time series by functionals of kernel-weighted empirical processes under local alternatives, Metrika, 60, 229-249, 2004.

6. Steland A. Optimal sequential kernel smoothers under local nonparametric alternatives for dependent processes. Journal of Statistical Planning and Inference, 132, 131-147, 2005.

7. Steland A. Weighted Dickey-Fuller processes to detect stationarity. Journal of Statistical Planning and Inference, 137, 12, 4011-4030.

8. Steland A. A note on data-adaptive bandwidth selection for sequential kernel smoothers. Proceedings of the 6th St. Petersburg Workshop on Simulation, 679-684, 2009.

9. Steland A. Photovoltaic Statlab 1.4, Software Tool, 2009.

10. Herrmann W., Steland A. und Zähle H. Statistische Prüflingswahl zum Leistungsnachweis. 3. Workshop ?Photovoltaik-Modultechnik?, TÜV Rheinland, Köln, 2006.

11. Herrmann W., Steland A. und Zähle H. Qualitative Leistungsbewertungen von Produktionschargen von Solarmodulen basierend auf Stichprobenmessungen. Projektbericht, 2007.

12. Herrmann W., Herff W. und Steland A. Methodik und Algorithmik der statistischen Qualitätsbewertung von Photovoltaik-Modulen bei Ausnutzung von Zusatzinformationen durch Flasherlisten, Projektbericht, 2008.

13. Herrmann W. und Steland A. Stichprobenbasierte Labormessungen für die Leistungsüberprüfung von PV-Modulen, 5. Workshop ?Photovoltaik-Modultechnik?, TÜV Rheinland, Köln, 2008.

Teil 2

Eingehende Darstellung

Kapitel 2.1

Verwendung der Zuwendung

Die Zuwendung wurde gemäß des Antrages und der Zeit-, Arbeits- und Ressourcenplanung verwendet. Diese Planung musste im Projektablauf nur geringfügig aktualisiert und nachgeführt werden. Lediglich eine Forschungsaktivität erforderte einen höheren Ressourceneinsatz, so dass nach 1 1/2 Jahren Projektlaufzeit entsprechend Ressourcen fachlich und zeitlich umgeschichtet wurden. Hierdurch wurden jedoch in der Summe andere Arbeitsergebnisse nicht berührt.

In Tabelle 2.1 sind die im Antrag formulierten Ziele sowie die erreichten Zielerreichungsgrade aufgeführt. Die Arbeiten haben gezeigt, dass in einigen Feldern konkreter Bedarf für Anschlussforschungen besteht.

Ziel	Ziel-erreichung	Anmerkung
1. Erforschung innovativer statistischer Verfahren zur Annahmeprüfung von PV-Modulen durch Ausnutzung der in Flasher-daten enthaltenen Zusatzinformationen.	100 %	Anschlussforschung zur Erweiterung auf Modulklassen und Einbezug zus. Faktoren ist angezeigt.
2. Analyse der finiten Eigenschaften der Verfahren und Vergleich der neu entwickelten Verfahren mit etablierten Standardverfahren durch Computersimulationen.	100%	
3. Erweiterung der Verfahren zu Out-of-Spec-Verfahren.	100%	Anschluss-Studie zur Evaluierung der Out-of-Spec-Pläne bei Verwendung der neu entwickelten Verfahren wünschenswert.
4. Evaluierung der Verfahren an konkreten realen Datensätzen unserer Industriepartner.	100 %	Nachgelagertes Sammeln von Praxiserfahrung i.d. industriellen Anwendung sinnvoll, um zu einer umfassenden Datenbasis für eine Evaluierung der Praxistauglichkeit einiger neu entwickelter Verfahren zu gelangen.
5. Prototypische Implementierung ausgewählter Verfahren zur Vorbereitung einer wirtschaftlichen Verwertung der Ergebnisse.	100 %	Implementierung in APOS war möglich (Fast Track Technology Transfer)

Tabelle 2.1.1: Tabelle: Ziele und Zielerreichung

Kapitel 2.2

Zahlenmäßiger Nachweis (wichtigste Positionen)

Der zahlenmäßige Nachweis findet sich in den Anlagen.

2.2 ZAHLENMÄSSIGER NACHWEIS (WICHTIGSTE POSITIONEN)

Kapitel 2.3

Notwendigkeit und Angemessenheit der Zuwendung

2.3.1 Notwendigkeit

Die im Vorhaben erarbeiteten Ergebnisse hätten ohne Gewährung der Zuwendung weder hinsichtlich der theoretischen noch der angewandten Aspekte erarbeitet werden können. Die Zuwendung war in jeder Hinsicht notwendig. Eine Fördermöglichkeit aus anderen Quellen bestand nicht und ergab sich auch nicht während der Projektbearbeitung.

Für die Erzielung der Ergebnisse war ferner notwendig, mit Industriepartnern eng zu kooperieren. Nur so konnten die praxisrelevanten Aspekte Eingang in die Arbeiten finden und der starke Fokus auf konkrete Forschungs- und Innovationsbedarfe in der Photovoltaik-Industrie gelegt und gehalten werden. Wir haben durch die Kooperation mit unseren Industriepartner die dort angesiedelte Expertise mit unserer verbinden können und so Forschungsergebnisse erzielt, die ansonsten nicht möglich gewesen wären. Daher war der gewählte Förderweg hinsichtlich des Forschungsformats notwendig für die erzielten Arbeitsergebnisse.

2.3.2 Angemessenheit

Die beantragten Mittel waren in Höhe und Verteilung auf Unterbudgets auch bei einer nachgelagerten Betrachtung und Analyse i.W. noch angemessen. Die sehr produktive Projektkonstruktion und gelungene Kooperation sowohl vor Ort in Aachen als auch mit den

Partner führte während des Vorhabens zu einem umfangreicheren Arbeitsprogramm als ursprünglich geplant, welches sich in sehr umfangreichen Arbeitsergebnissen niedergeschlagen hat. Diese sind in Form dieses Berichts, der über die Website pvstatlab.rwth-aachen.de zur Verfügung gestellten Informationen, die bereits publizierten Artikel sowie die konkrete Implementierungen von Ergebnissen in das Softwaretool APOS dokumentiert.

Dies alles führte jedoch zu einer sehr hohen Inanspruchnahme des Projektmitarbeiters und auch der Grundausstattung in Aachen. Hier wäre es im Nachhinein wünschenswert gewesen, auf ein höheres Mitarbeiterbudget zurückgreifen zu können.

Die Kooperation mit der Wrocław University of Technology war sehr produktiv und resultierte in der Entwicklung des besten neue Verfahrens zur Konstruktion photovoltaischer Prüfpläne. Die hierbei anfallenden Reisekosten wurden z.T. von unseren Partnern in Wrocław übernommen, wodurch sich eine entsprechende Entlastung unseres Budgets ergab.

Es waren nur geringe Änderungen an der Finanzplanung nötig, sowohl hinsichtlich der Verteilung auf Posten/Unterbudgets als auch hinsichtlich der zeitlichen Finanzplanung. I.W. waren die gewählten Ansätze gut kalkuliert. Hinsichtlich der Aufwendungen für den Projektmitarbeiter ergab sich ein leichter Fehlbetrag, der aus anderen Mitteln finanziert werden musste. Die beantragten Mittel wurden also vollständig abgerufen und ausschließlich für die beantragten Zwecke verausgabt.

Kapitel 2.4

Voraussichtlicher Nutzen und Verwertbarkeit

2.4.1 Voraussichtlicher Nutzen

Der voraussichtliche Nutzen der Projektergebnisse besteht insbesondere in der signifikanten Verbesserung bereits in der Industrie eingesetzter Verfahren zur Bestimmung von optimalen Prüfplänen zur Evaluierung der Leistung von Solarmodulen.

Hierdurch wird eine deutlich erhöhte Verläßlichkeit des aus Flasherlisten berechneten Umfangs der zu prüfenden Modulstichprobe erreicht und somit auch die Planungssicherheit verbessert. Ferner wird der Einsatz der Prüfmethodik auch bei Vorliegen nur kleiner Flasherlisten substantiell verbessert.

2.4.2 Verwertbarkeit und Verwertung

2.4.2.1 Verwertbarkeit

Die Projektergebnisse sind in mehrerer Hinsicht verwertbar:

- Zum Einen stellen sie für die PV-Arbeitsgruppe am ISW der RWTH einen substantiellen Erkenntnisgewinn dar. Die Ergebnisse befruchten unserer theoretischen Forschungen und bereits weiterführende Forschungsarbeiten initiiert.

- Die erarbeiteten Ergebnisse und die resultierenden Publikationen haben zu einer

Festigung der exzellenten internationalen Forschungsposition geführt. Dies zeigt sich auch daran, dass Prof. Steland inzwischen regelmäßig Forschungspublikationen für die international führende Zeitschrift *Progress in Photovoltaics* begutachtet.

- Ferner sind die Ergebnisse direkt verwertbar in der Qualitätssicherung und Zertifizierung von Modulen durch Prüflabore wie das des TÜV Rheinland. Hier verbessern sie substantiell die Bestimmung von benötigten Prüfumfängen von Modulstichproben zur präzisen Bestimmung der elektrischen Leistung von Modulen unter Standardbedingungen. Auf diese Weise ist es möglich, verläßlichere Prüfungen zu ermöglichen und somit auch die hiermit verbundenen Kosten für Hersteller und Betreiber von PV-Parks zu senken.

- Schließlich sind die Ergebnisse von Herstellern von PV-Modulen wie SolarWorld oder AVANCIS anwendbar. Zum Einen für die interne Qualitätssicherung und -überwachung, zum anderen für die Qualitätssicherung ausgehender Lieferungen an Endkunden und Betreiber von PV-Systemen. Die Hersteller werden durch die optimierten Verfahren in die Lage versetzt, mit kostenminimalen Nachprüfungen der produzierten Module die tatsächliche Qualität von Lieferungen zu bewerten und so die aus schlechter Qualität resultierenden Regressforderungen und Garantieleistungen besser zu begrenzen.

2.4.2.2 Verwertung

Der Verwertungsplan wie im Antrag formuliert hat auch am Projektende Gültigkeit.

- Im Rahmen des Projekts wurde ein Vertrag zwischen allen Partnern geschlossen, der den Partnern während der Projektlaufzeit Zugang zu erarbeiteten Ergebnissen ermöglicht hat. Auf diese Weise konnte auch eine Verwertung von Projektergebnisse bei den Projektpartner sicher gestellt. Die Ergebnisse wurden und werden aber auch der Allgemeinheit zugänglich gemacht und können so verwertet werden. Dies erfolgt insbesondere in Form von Publikationen in allgemein zugänglichen Periodika, einer intensive Vortragstätigkeit auf internationalen Konferenzen und Workshops sowie der Pflege der Seiten unserer pvstatlab-Website.

- Es ist insbesondere gelungen, noch im Rahmen des Vorhabens die beiden besten neu entwickelten Verfahren in Rturbo/P zu implementieren und in das Softwaretool

APOS integriert werden oder in anderer Form die Anwendung erleichtert werden. Auf diese Weise können wesentliche Ergebnisse rasch verbreitet werden und Eingang in die industrielle Praxis finden.

- Es ist darüber hinaus geplant, auch weitere Projektergebnisse zu verwerten. Dies kann jedoch nur nachgelagert geschehen. Zum Einen sollen mittelfristig die Out-of-Spec-Methodologie in APOS integriert werden. Zum Anderen stellen die erzielten Ergebnisse wertvolle und auch notwendige Vorergebnisse für weitere Forschungsvorhaben dar, die in Kooperation mit industriellen Partnern und anderern Forschungseinrichtungen in näherer Zukunft angegangen werden sollen.

2.4.3 Während des Projekts bekannt gewordener Fortschritt

Während der Bearbeitung des Projekts wurde durch regelmäßige Literaturrecherchen und Beobachten der Vortragsprogramme von nationalen und internationalen Workshops und Konferenzen überprüft, ob auf den zu erforschenden Gebieten neue Erkenntnisse von anderen Arbeitsgruppen erzielt wurden. Dies war jedoch nicht der Fall. Die in diesem Vorhaben entwickelten Modelle, Methoden und Verfahren und die gewonnenen Erkenntnisse wurden nicht durch Fortschritte oder Arbeiten bei anderen Stellen berührt.

2.4.4 Erfolgte oder geplante Veröffentlichungen der Ergebnisse

Die Kernergebnisse des Vorhabens wurden in den folgenden Artikeln publiziert:

- Meisen S., Pepelyshev A. and Steland A. 2011. Quality assessment in the presence of additional data in photovoltaics. *Frontiers in Statistical Quality Control*, Vol. 10, 251–274, Springer-Verlag.

- Golyandina N., Pepelyshev A., Steland A., 2012. New approaches to nonparametric density estimation and selection of smoothing parameters. *Computational Statistics and Data Analysis*, 56(7), 2206–2218.

- Pepelyshev A., Steland A., Avellan-Hampe A. 2012. Acceptance sampling plans for photovoltaic modules with two-sided specification limits. *Progress in Photovoltaics*, im Druck.

Hinzu tritt die folgende Publikationen, die sich in Revision befindet:

- Rafajlowicz E., Pepelyshev A., Steland A. 2012. Estimation of the quantile function using Bernstein-Durrmeyer polynomials, in Revision.

Eine weitere Publikation befindet sich in der Vorbereitung:

- Rafajlowicz E., Pepelyshev A., Steland A. 2012. Double kernel density estimator with application to sampling plan construction, in Vorbereitung.

Veröffentlichung in Form von Vorträgen auf internationalen Tagungen und Workshops:

- Steland A. (2012): *Ermittlung von repräsentatitiven Modulstichproben zur Leistungsmessung*. Invited talk given at the 9th Workshop Modultechnologie, Cologne, Germany.

- Pepelyshev A. (2012): *Change-point detection using SSA approach and applications to photovoltaics*. Annual Conference of the German Statistical Society 'Statistische Woche', Vienna, Austria.

- Pepelyshev A., Rafajlowicz E. and Steland A. (2011): *Functional Change-Point Asymptotics and Applications in Finance and Renewable Energies*. Statistische Woche, Leipzig, Germany.

- Steland, A. (2011): *Functional Change-Point Asymptotics and Applications Invited Talk*. ISI 2011 World Congress of Statistics, Dublin, UK.

- Steland, A. (2011): *Functional Asymptotics and Applications*. IWSM 2011, Invited Talk and Session Organizer, Stanford University, CA, U.S.A.

- Pepelyshev, A. Golyandina, N., Steland, A. (2011): *New Method of Nonparametric Density Estimation With Applications in Photovoltaics*. Workshop on Stochastic Models and Their Applications 2011, Wismar, Germany.

2.4.4 ERFOLGTE ODER GEPLANTE VERÖFFENTLICHUNGEN DER ERGEBNISSE

- Steland, A. (2011): *Quality Assessment in the Presence of Additional Data and Robustness Issues*. Workshop on Stochastic Models and Their Applications 2011, Wismar, Germany.

- Steland, A. (2010): *Quality Assessment in the Presence of Additional Data*. Invited talk held at the 2010 ISQC Workshop in Seattle, U.S.A.

2.4 VORAUSSICHTLICHER NUTZEN UND VERWERTBARKEIT

Teil 3

Darstellung der Einzelergebnisse

Kapitel 3.1

Zusammenfassung der Ergebnisse

3.1.1 Methoden

3.1.1.1 Varianzminimierende Verfahren

Es wurden die folgenden methodischen Ansätze eingehend für das Problem der Konstruktion photovoltaischer Annahmeprüfpläne bei Vorliegen von Flasher-Stichproben erforscht:

- Nichtparametrische Kernschätzer mit Bandbreitenwahl.
- Spektralanalytische Methoden (SSA-Ansatz).
- Orthogonale Reihenschätzer.
- Bernstein-Polynome und Bernstein-Durrmeyer Polyomial-Operatoren.
- Nichtparametrische doppelte Kernschätzer.

Allen Ansätzen ist gemein, dass sie noch nicht für das betrachtete Problem studiert wurden.

Nichtparametrische Kernschätzer wurden von uns schon vorher untersucht. Hier stellte sich primär die Frage, welcher Ansatz zur Bandbreitenwahl für das gestellte Problem zu den besten Resultaten führt. Aus Vorstudien war bekannt, dass der Einfluß der Bandbreitenwahl erheblich ist.

Der *spektralanalytische SSA-Ansatz* wurde vom PI, Dr. Andrey Pepelyshev, eingebracht. Er basiert auf einer Anwendung einer Singulärwertzerlegung (ähnlich der bekannten Hauptkomponentenanalyse) auf eine Matrix, welche in geeigneter From die beobachteten

Daten der zu approximierende Funktion zusammenstellt. Die SSA-Methode wurde bereits mit sehr guten Ergebnissen auf vielfältige Probleme insbesondere aus Technik und Physik angewendet, auch wenn der theoretische Unterbau verglichen mit anderen Ansätzen eher spärlich ist.

Orthogonale Reihenschätzer sind ein mathematisch sehr gut verstandenes Standardinstrument zur Approximation und Schätzung unbekannter funktionaler Zusammenhänge und in vielen Anwendungsfeldern etabliert, insbesondere der Verarbeitung, Rekonstruktion und Analyse technischer und physikalischer Signale.

Bernstein-Polynome und *Bernstein-Durrmeyer Polynomialoperatoren* sind ebenfalls sehr gut verstanden und umfassend in Anwendungen wie den Ingenieurwissenschaften, der Informatik (Stichwort: Bezier-Kurven) sowie Mathematik und Physik etabliert.

Nichtparametrische doppelte Kernschätzer sind eng verwandt sowohl mit nichtparametrischen Kernschätzern als auch Reihenschätzern. Sie wurden jedoch bisher kaum in der Literatur studiert.

3.1.1.2 Out-of-Spec-Verfahren

Die Entwicklung von Out-of-Spec-Verfahren der Variablenprüfung sollte nach Möglichkeit im etablierten Rahmen der AQL-RQL-Methodologie erfolgen, also den Anteil nichtkonfomer Module als Kernparameter verstehen, und somit diese substantiell erweitern.

Das Problem galt lange Zeit als unlösbar, da ein zu enges Anlehnen an die klassische Herleitung der Variablenprüfung unter Normalverteilungsannahme sehr schnell auf eine nichtlineare Gleichung führt, die nicht explizit lösbar ist, auch nicht in Spezialfällen.

3.1.2 Ergebnisse

Die Kernergebnisse des Forschungsvorhabens können wie folgt zusammen gefasst werden:

3.1.2.1 Nichtparametrische Kernschätzer

Für diese Klasse von Verfahren war zunächst zu untersuchen, welche der sehr großen Zahl von bekannten Verfahren zur (datenadaptiven) Bandbreitenwahl für die Konstruktion von Prüfplänen bei Vorliegen von Zusatzinformation in Form von Flasherlisten optimal ist.

Hierzu wurde bei Projektbeginn die relevante und neueste Forschungsliteratur aufgebereitet, um die vielversprechendsten Verfahren zu identifizieren. Hier wurde insbesondere die ICV-Methode (erst kurz vor Projektbeginn im Sommer 2010 publiziert) mit in die Untersuchungen aufgenommen.

Unsere umfangreichen Simulationen haben gezeigt, dass die ICV-Methode für Kernschätzer zu empfehlen ist und das bisherige Benchmark-Verfahren, das *modifizierte Verfahren*, dominiert. Dies ist in Teilen überraschend, insbesondere da die Verbesserung z.T. deutlich ist.

3.1.2.2 Spektralanalytische SSA-Verfahren

Das SSA-Verfahren ist eng verwandt mit der bekannten Hauptkomponentenanalyse. Es wird eine Singulärwertzerlegung einer Trajektorienmatrix mit L Zeilen betrachtet, in welcher die empirische Leistungsverteilung in diskretisierter Form eingeht und zeilenweise verschoben wird. Wir haben mehrere Varianten entwickelt, die sich in der Wahl der verwendeten Anzahl r der Komponenten sowie in der Wahl der Zeilenzahl der Trajektorienzahl unterscheiden. Für das betrachtete Probleme zeigten Simulationen schnell, dass $r = 1$ oder $r = 2$ gute Wahlen sind.

Ferner haben wir einen völlig neuen Ansatz zur Wahl des Parameters L entwickelt, der i.W. die Anzahl der lokalen Maxima/Minima der zugehörigen Dichteschätzung betrachtet. Ein zu hoher Wert deutet auf statistische Artifakte hin und zeigt an, dass die Schätzung ungenügend glättet. Wir konnten ein hoch interessantes und praktisch relevantes Ergebnis über die Konsistenz des Verfahrens mit automatischer Wahl des Tuningparameters zeigen.

Dieses Verfahren zur automatischen Wahl eines Tuningparameters konnte auch erfolgreich auf das Problem der Bandbreitenwahl bei nichtparametrischen Kernschätzern übertragen werden.

Abschließende umfangreiche Computersimulationen haben gezeigt, dass die beste SSA-Verfahrensvariante kompetitiv ist, zwei andere Verfahren waren jedoch letztendlich noch besser.

3.1.2.3 Orthogonale Reihenschätzer

Diese Klasse von Verfahren wurde in die Untersuchungen aufgenommen, da sie bei vielen Anwendungsproblemen zu sehr guten Lösungen führt. Es zeigte sich jedoch sehr schnell,

dass orthogonale Reihenschätzer für das betrachtete Problem der Schätzung optimaler Prüfpläne nicht geeignet sind. Sowohl bei Anwendung auf die Leistungsdichte als auch bei Anwendung auf die Quantilfunktion führen sie zu erheblichen Oszillationen genau in den Bereichen, die für die Schätzung der Prüfpläne relevant sind. Hierauf ist auch das schlechte Abschneiden in den Simulationsstudien zurückzuführen. Diese Probleme konnten auch durch Verwendung von komplexeren Methoden zur Auswahl relevanter Terme der approximierenden Orthogonalreihe wie dem modifizierten Verfahren nach Kronmal-Tarter nicht behoben werden.

Orthogonal Reihenschätzer wurden daher am Milestone I von den weiteren Untersuchung ausgeschlossen.

3.1.2.4 Bernstein-Polynome und Bernstein-Durrmeyer-Polynome

Bernstein-Polynome und die verwandten, aber etwas weniger bekannten, Bernstein-Durrmeyer-Polynome stellen einen interessanten recht gut verstandenen Ansatz zur Approximation von funktionalen Zusammenhängen dar. Bernstein-Polynome und Bernstein-Durrmeyer-Operatoren wurden auch in der statistischen Literatur zur Schätzung von Verteilungsfunktionen, Dichtefunktionen, Quantilfunktionen und Regressionen untersucht.

Im Rahmen unsere Arbeiten haben wir ganz neue statistische Ergebnisse über die erreichbare Konvergenzrate von Bernstein-Durrmeyer-Schätzern erzielt: Der Schätzer erreicht fast die parametrische Rate, was ein bemerkenswertes Resultat darstellt. Ferner haben wir einen datenadaptiven Ansatz zur Schätzung des Grads des verwendeten Polynoms entwickelt und konnten die Konsistenz des resultierenden datenadaptiven Verfahrens zeigen. Nach unserem Wissensstand gibt es kein vergleichbares Ergebnis.

Unsere Forschungen haben gezeigt, dass Verfahren basierend auf Bernstein- bzw. Bernstein-Durrmeyer-Polynomen zu guten Prüfplänen bei Vorliegen von Flasherdaten führen, jedoch gehören sie nicht zu den besten Verfahren.

3.1.2.5 Nichtparametrische doppelte Kernschätzer

Dieser Ansatz ist eng verwandt mit nichtparametrischen Kernschätzern und bis zu einem gewissen Grad auch mit (Pseudo-) Orthogonalreihenschätzern. Während nichtparametrische Kerndichteschätzer an den beobachteten Datenpunkte das zugehörige Punktmaß durch eine lokalisierende Dichtefunktion ersetzen, um zu einer geglätteten Schätzung

der Leistungsdichte zu gelangen, werden bei diesem Ansatz solche lokalisierenden Modelldichten an den Stützstellen eines dichten Gitters verwendet, die mit Gewichten versehen werden. Diese Gewichte können als statistische Schätzer des inneren Produkts der verwendeten Modelldichte mit der Leistungsverteilung bzw. -dichte verstanden werden, wodurch sich eine Ähnlichkeit zu Pseudo-Orthogonalreihenschätzern ergibt; es wird hier jedoch keine Basis eines vollständigen Funktionenraums definiert. In diesen Ansatz floßen wesentlich methodisch verwandte Vorarbeiten zu einem regressionsanalytischen Problem unseres Partners von der Wrocław University of Technology, Prof. Dr. E. Rafajłowicz, ein. Wesentliche Teile der Entwicklung wurden bei Forschungsaufenthalten in Wrocław initiiert, diskutiert, konsolidiert und finalisiert.

Wir haben insbesondere einen neuen innovativen datenadaptiven Stutzungsansatz zur Behandlung von kleinen Gewichten speziell an den Rändern der Verteilung entwickelt, durch den der resultierende Schätzer noch einmal deutlich verbessert werden konnte. Ferner gelang es, unter sehr allgemeinen Voraussetzungen die Konsistenz des doppelten Kernschätzers mit datenadaptiver Stutzung nachzuweisen.

Für den in diesem Ansatz ebenfalls zu wählenden Bandbreitenparameter wurde eine ebenfalls neue einfache Regel entwickelt, welche für jeden Stützpunkt eine lokale Wahl vornimmt, die sich an dem Gewicht des jeweiligen Stützpunktes orientiert.

Eine umfassende Simulationsstudie hat gezeigt, dass die aus diesem Ansatz resultierenden Prüfpläne für die photovoltaische Qualitätssicherung bei Einbezug von Flasherlisten in der Mehrzahl der betrachteten Simulationsmodelle allen anderen Ansätzen überlegen ist.

3.1.2.6 Ouf-of-Spec-Verfahren

Die von uns gefundene Lösung, die in Kapitel 3.7 im Detail vorgestellt wird, basiert auf einer intelligenten impliziten Kombination von Teststatistiken für die einseitigen Prüfverfahren. Hierdurch ergibt sich eine geschlossene Formel für die OC-Funktion, auf welche Approximationsergebnisse aus unseren Vorarbeiten angewendet werden können.

Relevante Parameter sind die Anteile von Underperforming– und Overperforming–Modulen, aus denen sich der Anteil der nicht-konformen Modulen zusammensetzt. Während die Produzenten- und Konsumentenforderungen an den optimalen Prüfplan bei einseitigen Spezifikationsintervallen auf zwei leicht und explizit lösbare *Gleichungen* führen, besteht die relevante Charakterisierung eines optimalen Prüfplans zunächst aus einem *System nicht-*

linearer Ungleichungen. Dieses kann jedoch vereinfacht werden. Zum Einen konnten wir zeigen, dass die optimalen Prüfplänen durch den zugehörigen Quotienten γ der Over– und Underperforming–Module parametrisiert sind, wodurch sich analytisch und auch in der praktischen Anwendung eine erhebliche Vereinfachung ergibt. Zudem konnten wir Bestimmungsgleichungen herleiten, die im allgemeinen Fall durch numerische Verfahren gelöst werden können. In Spezialfällen ergeben sich explizite Formeln, die dieselbe Gestalt haben wie die bekannten Formeln.

Wir haben die resultierenden Prüfpläne numerisch umfangreich untersucht, um den Einfluss des Parameters γ und den Einfluss der Leistungsverteilung in praktisch relevanten Situationen zu analysieren. Unsere numerischen Untersuchungen zeigen, dass die resultierenden Prüfpläne in vielen Fällen zu deutlich kleineren Stichprobenumfängen führen. Ferner ist der Einfluss des Quotienten γ zwar präsent, scheint aber in praktisch relevanten Situationen oftmals beherrschbar zu sein. Wir haben für die praktische Anwendung ein Bootstrap–Konfidenzintervall vorgeschlagen. Dieses kann in der Praxis verwendet werden, um den Einfluss auf den Prüfumfang abzuschätzen und einen konkreten Prüfplan zu wählen.

Werden jedoch die Flasherdaten zur Klassifizierung der Module in Leistungsklassen verwendet, so kann der Parameter γ nicht aus den relativen Häufigkeiten geschätzt werden. Dieser Punkt konnte im Rahmen des Vorhabens nicht abschließend behandelt werden.

3.1.2.7 Einfluss der Technologie

In dem Projekt waren zwei PV-Hersteller als Partner vertreten, die im kristallinen bzw. Dünnschichtbereich eine herausragende Position haben. Nach den vorliegenden Ergebnissen sind die Verfahren sowohl für kristalline Module anwendbar als auch für Dünnschichtmodule. Die entwickelten Verfahren passen sich sehr flexibel verschiedensten Leistungsverteilungen an und Tuning-Konstanten werden datenadaptiv geschätzt.

Die nahe liegende Frage, ob die Modultechnologie einen Einfluss darauf hat, welche(s) Verfahren optimal ist, konnte nicht umfassend untersucht werden. Zumindest scheint uns eine Aussage zu diesem Punkt in Anbetracht der geringen Datenbasis nicht sinnvoll. Hinzu kommt die Tatsache, dass der Gestalt der Verteilung eine maßgebliche Rolle zukommt und diese sehr häufig von denjenigen Faktoren, die als Ein- und Ausschlusskriterium für die Definition des Lots dienen, maßgeblich bestimmt wird, insbesondere von der Modulklasse.

3.1.2.8 Harmonisierung

Insbesondere aus Sicht eines zertifizierenden Prüflabors, aber auch aus Sicht von Photovoltaik-Herstellern, ist eine Harmonisierung von Prüfverfahren sinnvoll. Durch klare Definitionen von Prüfumfang, Prüfmerkmalen, Erhebungs- bzw. Messmethodik sowie Analytik wird ein transparenter Rahmen gesetzt, der zu validen und reproduzierbaren Ergebnissen führt.

Im Kontext des Forschungsvorhabens wurde insbesondere festgehalten, dass eine möglichst eindeutige Empfehlung für eine Prüfmethodik wünschenswert ist. Diese Thematik wurde insbesondere mit dem TÜV Rheinland, aber auch den anderen Partner mehrfach diskutiert, insbesondere bei den Projekttreffen.

In Anbetracht der gewonnenen Erkenntnisse gibt es zwei Verfahren, die für einen in diesem Sinne harmonisiertes Prüfverfahren in Frage kommen: Zum Einen das Kernverfahren mit ICV-Bandbreitenwahl und der doppelte Kernschätzer.

Das Kernverfahren mit ICV-Bandbreitenwahl stellt ein etabliertes und gut verstandenes Verfahren der Statistik dar, welches auch bei der Konstruktion von photovoltaischen Prüfpläne ausgezeichnete Ergebnisse liefert.

Das Verfahren basierend auf einem doppelten Kernschätzer ist bei einer Mehrzahl der betrachteten Simulationsmodelle dem Kernverfahren mit ICV-Bandbreitenwahl durchaus überlegen und sollte daher das Verfahren der Wahl sein. Allerdings ist es, auch wenn mit diesem methodisch verwandt, in theoretischer Hinsicht weniger gut verstanden.

Aus diesem Grund empfehlen wir, vorläufig das Kernverfahren mit ICV-Bandbreitenwahl als neues Standardverfahren zu etablieren und mit dem doppelten Kernverfahren zunächst praktische Erfahrungen zu sammeln. Vor solch einem längerfristigen Erfahrungshintergrund sollte dann eine Neubewertung vorgenommen werden, um bei positiven Erfahrungen auf das doppelte Kernverfahren zu wechseln. Daher wurden im Rahmen des Fast Track Technologietransfers auch beide Verfahren in APOS implementiert.

3.1.3 Diskussion und Ausblick

3.1.3.1 Bewertung der Ergebnisse

Das im Vorhaben gesteckte Ziel, varianz-minimierende Prüfpläne für photovoltaische Leistungsmessungen zu entwickeln, konnte umfassend erreicht werden. Zugleich haben die umfangreichen Computersimulationen unsere Voreinschätzung erwiesen, dass es keine 1-

zu-1 Übertragung von Literaturergebnissen zur Dichteschätzung, weder aus theoretischen noch aus Simulationsstudien, auf das Problem der Schätzung von Prüfplänen bei Vorliegen von Zusatzinformation gibt. Ferner hat sich gezeigt, dass zwar eine Reihe von Verfahren 'nahezu optimal' ist, aber kein Verfahren gleichmäßig optimal ist, also in *allen* betrachteten Modellen und für *alle* analysierten Industriedatensätze optimal ist.

Zusammenfassend kann festgestellt werden: Die besten in diesem Vorhaben neu entwickelten Verfahren führen zu deutlich verläßlicheren Prüfverfahren:

Das bisherige Benchmark-Verfahren, das sog. *modifizierte Verfahren*, ist um bis zu 50% ungenauer als die neuen Spitzenreiter.

Dies ist ein aus statistischer Sicht exzellenter Wert und auch praxisrelevat. Diese Verbesserung spiegelt sich i.W. in einer entsprechend kleineren Standardabweichung der (geschätzten) Prüfumfänge nieder. Somit steigt die Verläßlichkeit und Planungssicherheit der ermittelten Prüfpläne. Ferner können die Verfahren auf deutlich kleinere Flasher-Stichproben angewendet werden: Eine um 1/3 reduzierte Standardabweichung[1] bedeutet, dass für gleiche Genauigkeit mit Flasher-Stichproben gearbeitet werden kann, die nur $4/9 \approx 0.44$ des ursprünglichen Umfangs betragen. Die Kernergebnisse konnten in hervorragenden statistischen Zeitschriften publiziert werden oder stehen zur Publikation an.

Die im Vorhaben neu entwickelte Methodologie für Out-of-Spec-Plänen erweitert substantiell die statistische Methodenwelt und erweitert die PV-spefizischen Verfahren der statistischen Qualitätssicherung auf eine praktisch relevante Situation. Die Out-of-Spec-Methodologie löst in praktisch umsetzbarer Weise eine Problematik, die von seiten der Industrie an uns herangetragen wurde. Die Ergebnisse konnten in einem führenden interdisziplinären Journal der Photovoltaik-Forschung publiziert werden, welches besonderen Wert auf die wissenschaftliche Qualität und den praktischen Impact legt.

3.1.3.2 Ausblick

Im Rahmen des Vorhabens wurden Datensätze analysiert von unseren Industriepartner analysiert, die wesentlich die methodischen Entwicklungen beeinflußt und vorangetrieben haben. Es wurde im Rahmen der Datenanalysen aber auch deutlich, dass die entwickelten

[1] Das Kernverfahren mit ICV-Bandbreitenwahl reduziert bei normalverteilten Leistungsmessungen die Standardabweichung des Prüfumfangs bei Schätzung aus einer Flasherliste vom Umfang $m = 100$ von 35 auf 22.3. Dies entspricht einer Reduktion um 36.3%

Verfahren nur in eingeschränktem Maße für einzelne Modulklassen mit steilen Verteilungsflanken (Short Tails) und sich anschließendem Plateau anwendbar sind. In diesem Fall resultieren unrealistisch hohe Prüfumfänge. Die im Rahmen dieses Projekts entwickelten Verfahren führen in der Regel zu deutlich besseren Ergebnissen, der Einfluß der Verteilungsflanke erschwert jedoch nicht nur die statistische Schätzung des Prüfplans, sondern ist generell – begründet durch die Theorie – mit höheren Prüfumfängen verbunden, auch wenn die Verteilung bekannt ist.

Diese Problematik konnte nicht im Rahmen des Forschungsvorhabens studiert werden. Datenanalysen haben jedoch gezeigt, dass sich die Leistungsverteilung nach der Auslieferung ändert und eine physikalisch korrekte Modellierung dieser Änderung ein wesentlicher Schlüssel für die Entwicklung von Prüfverfahren für Modulklassen mit steilen Flanken sein sollte. Hier besteht konkreter Forschungsbedarf. Für solche zu entwickelnden Verfahren stellen die Ergebnisse dieses Vorhabens sehr wertvolle Vorarbeiten dar, da auf Ebene von Modulklassen nur deutlich kleinere Flasher-Stichprobenumfängen zur Verfügung stehen, insbesondere in den Rändern der Verteilung.

Die Regressionsanalyse von Daten – im Cross-Section-Design – unserer Industriepartner zeigte, dass ein Reihe von Faktoren Einfluss auf die Differenz von Flashermessungen und Nachmessungen hat. Hierzu gehören insbesondere der Standort, das messende Labor und die Modulkasse. Die Analysen deuten ferner darauf hin, dass ein faktorieller Einfluss auf die Verteilung der Fehlerterme vorliegt, so dass die Standardannahmen des Regressionsmodells nicht erfüllt sind. Für die valide Analyse von industriellen Produktionsdaten sind daher maßgeschneiderte Regressionsverfahren zu entwickeln. Die Tatsache, dass die obigen Faktoren offenkundig oftmals berücksichtigt werden müssen, führt dazu, dass die Verfahren auch für kleine und unbalanzierte Designs verläßliche Ergebnisse liefern müssen. Auch hier ist ein substantieller Forschungsbedarf angezeigt.

Während die im letzten Paragraph erwähnte Problematik die Entwicklung von maßgeschneiderten regressionsanalytischen Verfahren für photovoltaische Qualitätsdaten im Cross-Section-Design addressiert, ist die Entwicklung von PV-spezifischen Verfahren für Zeitreihen-Daten von mindestens ebenso großer Bedeutung. Qualität von PV-Modulen kann nicht nur stichtagsbezogen gesehen werden, sondern sollte den Lebensweg eines produzierten Moduls betrachten. In einem ersten Schritt ist hier insbesondere die Zeitspanne von der Produktion über die Auslieferung bis zur Installation im PV-System bzw. bis zum ersten Inspektionszeitpunkt kurz nach Inbetriebnahme des Systems zu sehen. Insbe-

sondere Dünnschichtmodule verändern während der ersten Betriebszeit ihre elektrischen Eigenschaften, so dass sich die auch die Leistungsverteilung ändert. Wenn auch in geringerem Umfang, so ist doch auch bei kristallinen Modulen hiervon auszugehen. Bis dato liegen keine Erkenntnisse vor, wie lange Flasherlisten valide sind und zur Bewertung der Eigenschaften von PV-Modulen und insbesondere zur Konstruktion von Prüfplänen herangezogen werden können. Bei einer Veränderung der Leistungsverteilung müssen entsprechenden Verfahren entwickelt werden, die diesen zeitlichen Effekt berücksichtigen und praktikable und statistisch valide Verfahren zur Konstruktion von Inspektionsplänen von PV-System liefern. Hier besteht ein signifikanter Forschungsbedarf.

Kapitel 3.2

Acceptance sampling in photovoltaics

3.2.1 Introduction

Statistical quality control of lots or shipments of goods is an important practical problem in industry, particularly when high-quality leadership is a strategic goal. Delivering shipments of bad quality to customers may result in expensive law suits. In the photovoltaic industry, the customer expectations of quality of delivered photovoltaic (PV) modules are very high, presumably due to the fact that the state of the art of semi-conductor technologies are at the core of the business, thus being associated with digital precision. Acceptance sampling, which deals with the problem of determining the minimal sample size which controls the producer's risk as well as the consumer's risk, is therefore an important practical approach to the problem, see the recent monograph (38). The customers are interested in the quality of their shipments of PV modules and not in the average outgoing quality. Thus, manufacturers which are interested in customer satisfaction should control production on the basis of outgoing shipments.

We study the classic acceptance sampling problem under the general assumption that the measurements (the power output in photovoltaics) may follow an arbitrary continuous distribution function. To handle this case, we use additional data in the form of a historic data set, a situation which typically occurs in photovoltaics. We derive sampling plans assuming a less general distributional model compared to a model in our previous work (44), but the results of the present work are valid under less restrictive assumptions. Further, the model in the present work nicely allows us to investigate sensitivity and robustness issues, respectively. We study the effect on the sampling plans, when the distributions of the

shipment and the lab sample are different, that is of substantial interest for applications, where a systematic bias is of primary concern.

The statistical setup on measurements is as follows. We consider a shipment $X_1, \ldots, X_n \sim F$, where measurements have a common distribution function F which is assumed to be continuous and strictly increasing with a finite fourth moment. Here and in what follows, the sign $\sim$ always indicates that the random variables are independent and identically distributed. However, the stochastic relationship *between* samples may be arbitrary, i.e. they are not required to be independent. Let $\mu = E(X_1)$ and $\sigma^2 = E(X_1 - \mu)^2 \in (0, \infty)$. In the photovoltaic problem motivating our work, X_i represents the true but random power output of the ith module, which is classified as non-conforming, i.e. being of low quality, if $X_i \leq \tau$ for some constant τ. In practice, one puts $\tau = \mu(1 - \varepsilon)$ where ϵ is the tolerance. The additional data are provided in the form of a historic sample $Y_1, \ldots, Y_m \sim F$ of size m. Clearly, the expected fraction of non-conforming modules in the shipment is given by

$$p = P(X_1 \leq \tau) = F(\tau).$$

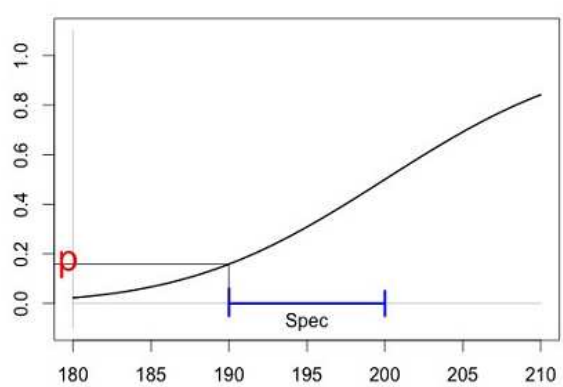


Abbildung 3.2.1: The probability p of PV modules of low quality.

Suppose that we have fixed two numbers $0 < \text{AQL} < \text{RQL} < 1$, namely the *acceptable qualilty level* (AQL) and the *rejectable quality level* (RQL), such that the lot should be accepted if $p \leq \text{AQL}$, whereas it should be rejected when $p \geq \text{RQL}$. Since p is unknown to us and checking all modules is infeasible, our aim is to decide on the basis of a control sample $X'_1, \ldots, X'_n$ of n measurements, n as small as possible, whether or not the shipment has to be accepted with controlled error probabilities of false decisions. Suppose that the decision is based on a statistic $T'_n = T'_n(X'_1, \ldots, X'_n)$ using the decision rule to accept the

shipment if and only if $T'_n > c$. A natural choice to use a standardized sum statistic

$$T'_n = \sqrt{n}\frac{\overline{X}'_n - \tau}{\sigma}.$$

Then a solution (n, c) to the above problem is called *sampling plan*.

3.2.2 Background on and application to photovoltaics

In the present section, we give a brief account of the photovoltaic background which motivated our research and the way how we approached the problem.

Photovoltaics represents one of the key technologies having the potential to provide a substantial contribution to the world's energy problem. Presumably, the main reason why the market share of solar energy is still relatively small compared to its potential and benefits is the fact that the costs per watt are still rather high. Although the costs have been substantially decreased in recent years, research still focuses on further reductions in costs, either by increasing the efficiency of a given solar cell technology or by developing new technologies, e.g. by employing cheaper materials and chemicals.

The economic life time of a photovoltaic system ranges between 20 and 30 years. Thus, the quality in terms of the power output of the PV modules at delivery is a crucial parameter for the profitability of such an investment. Even small departures from the nominal power output accumulate to considerable losses over the years. Assessing the quality of PV modules, which is done under standard conditions (STC) in a lab, is therefore an important issue for quality control.

A PV module is an interconnected assembly of solar cells. To protect the cells from damage during manufacturing, delivery and usage, they are embedded between a tedlar plate on the bottom, a tempered glass on the top, and framed, usually with an aluminium frame. Since a single module can produce only a limited amount of electricity, usually around 200 watts under STC, a PV system consists of many connected PV modules.

There are two common technologies to manufacture PV modules. Crystalline modules use silicon solar cells produced from solid Si wafers, whereas the CIS thin-film technology applies copper (C), indium (I) and selenium (S) in a layer construction of around 2 μm onto a substrate. The electrical properties such as the spectrum of the sun light transformed into electricity, the loss of efficiency when exposed to heat, a serious issue for systems installed in Southern Europe or Africa, or the efficiency when there are clouds as it is often

the case in Northern Europe, heavily depend on the technology and various other physical parameters of the chosen module type.

Calibrating a PV module in a testing laboratory is also a difficult problem since several factors may complicate collecting measurements and may lead to a considerable difference of indoor and outdoor measurements. As reviewed and experimentally analyzed in (45), the following effects matter in practice:

(i) Measurement related to sweep-time effects refereing to the influence of the duration to complete an IV scan. Depending on the selected flash tester, which generates a pulse of calibrated light, such a scan can be based on up to 100 flashes. The duration of each flash is typically 100 ms. For details on accurate testing of PV panels we refer to (32).

(ii) Spectral mismatch arising when using a reference cell with a spectral response is different from that of the device under test; its size depends on the spectral irradiance distribution of the sun simulator with respect to the reference spectrum AM 1.5G.

(iii) Finally, thin-film modules are affected by the effect of light soaking, since the performance (even under standard test conditions) depends on the storage conditions of PV modules (exposure to light or storage in the dark). This phenomenon called *light soaking* is in effect at the time of delivery but disappears when the PV modules are exposed to sun light for several days. The light soaking effect was first reported in (33) and is addressed to the tunneling of electrons trapped in deep states of CdS to holes in the CIS layer valence band under illumination, resulting in an increase in the open-circuit voltage and fill factor. During the light soaking period the performance can increase by $2 - 5\%$, cf. (25). In industrial practice, it can even be larger.

It is common practice in industry to classify the produced PV modules in classes. As a consequence of the above discussion, when analyzing *comparable* modules, i.e. modules of the same technology and power rating satisfying additional criteria for inclusion or exclusion in a study, the true distribution of measurements may have any form. In many cases, measurements are typically non-normal, thus violating the classic assumptions in statistical acceptance sampling.

Relying on ad hoc proposals such as forming subgroups and then applying variables sampling to the subgroup means to ensure approximatively the normal assumption, is not

feasible due to the high costs of taking control measurements, since this procedure leads to enormous sample sizes. However, taking the control measurements is very expensive. For the same reason, applying attribute sampling is no reasonable solution.

PV modules are manufactured in a production line. The performance of each module is measured in a sun simulator using short flashes. These measurements are therefore called flash measurements and form the flash data tables. In present days, they are routinely collected by manufacturers, thus often large samples are available. However, these cheap measurements may differ from the measurements taken in a photovoltaic laboratory. One should check carefully, whether a given flasher report table follows the same distribution as the shipment before applying the methods discussed in the present work. Some standard tests and their application to real photovoltaic data are described in (19). For a new approach to this problem we refer to the recent work (43).

3.2.3 Theory on acceptance sampling

The main issue of quality control is to find an acceptance sampling plan $(n; c)$, i.e. a size n of a control sample and a critical value c, such that

$$P(T_n > c) \geq 1 - \alpha, \quad p \leq \text{AQL}$$

and

$$P(T_n > c) \leq \beta, \quad p \geq \text{RQL}.$$

Here α is an upper bound on the probability that the shipment is rejected when it is of high quality, thus controlling the producers risk, whereas β is the consumers risk that the shipment is accepted although it is of low quality. The operating characteristic $\text{OC}(p) = P(T_n > c)$ is a function of the quality level p. Approximations based on large sample theory can be used to solve the problem of constructing appropriate sampling plans.

Let measurements be distributed according to a distribution $G(x)$ with mean a and variance σ^2 and $F(x)$ be the corresponding standardized distribution, that is, $F(x) = G((x-a)/\sigma)$. Due to the central limit theorem, we obtain the approximation

$$P(T_n > c) \approx 1 - \Phi(c + \sqrt{n}F^{-1}(p)).$$

where $\Phi(x)$ is the distribution function of the standard normal distribution. Notice that this approximation requires n to be large. However, since statistical inference should never

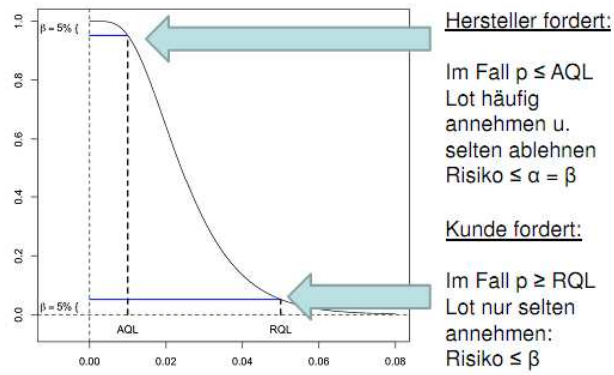


Abbildung 3.2.2: The operating characteristic and parameters.

be based on too few observations, assuming that the central limit theorem provides a sufficiently accurate approximation should not be too restrictive for many distributions F. Further, one may check the accuracy of the above approximation after calculating the sampling plan, such that c is fixed, using historic data which is available by assumption. This could be done, for instance, by estimating the Berry-Esseen upper bound or by means of a simulation study, the latter approach being preferable.

Using the monotonicity of $\mathrm{OC}(p)$, we obtain that the asymptotically optimal sampling plan (n, c) is defined as

$$\begin{aligned} n &= \left\lceil \frac{\left(\Phi^{-1}(\alpha) - \Phi^{-1}(1-\beta)\right)^2}{\left(F^{-1}(\mathrm{AQL}) - F^{-1}(\mathrm{RQL})\right)^2} \right\rceil, \\ c &= -\frac{\sqrt{n}}{2}\left(F^{-1}(\mathrm{AQL}) + F^{-1}(\mathrm{RQL})\right), \end{aligned} \tag{3.2.3.1}$$

where AQL is the acceptable quality level, RQL is the rejectable quality level, α is the producer risk and β is the consumer risk (Schilling, Neubauer, 2009). Typical values are $\alpha = \beta = 0.05$, $\mathrm{AQL} = 0.02$ and $\mathrm{RQL} = 0.05$.

In general, the distribution F is unknown. However we suppose that the additional data are provided in the form of a historic sample $Y_1, \ldots, Y_m \sim G$ of size m or a list of flash measurements. Then we can estimate the sampling plan (n, c) by the formula (3.2.3.1) with replacement F by the empirical distribution function F_m corresponding to the standardized sample $(Y_1 - a)/\sigma, \ldots, (Y_m - a)/\sigma$.

The next result on this estimator is established in (26).

Theorem 3.2.3.1. *Suppose that the historic data set as well as the lot and control measurements after standardization are distributed according to a strictly increasing and continuous*

d.f. F such that $\int x^4 dF(x) < \infty$ and $F'(F^{-1}(\mathrm{AQL})) > 0$ as well as $F'(F^{-1}(\mathrm{RQL})) > 0$. Assume that $n/m \to 0$ as $n, m \to \infty$. Then the sampling plan size n is asymptotically normal and $\sqrt{m}\mathrm{Var}(n) \to \eta^2$ as $m \to \infty$ where η is a constant.

3.2.4 Simulation settings

Sound Monte Carlo simulation studies require carefully chosen simulation models for the distribution of the power output of photovoltaic modules, which cover a sufficiently large amount of distributions which either represent idealized distributional shapes occurring in practice according to expert knowledge and statistical data analysis or representing artificial distribution models which incorporate certain effects or artifacts which may have an impact on the statistical behavior of variability-reducing sampling plans for photovoltaic data.

For this reason, we consider 9 distributional models which will be used as standard choices in all simulation studies. These simulation models are as follows:

$$
\begin{aligned}
&\text{model 1:} && X_i \sim \mathrm{N}(220, 4)\\
&\text{model 2:} && X_i \sim 0.1\mathrm{N}(210, 6) + 0.9\mathrm{N}(230, 4)\\
&\text{model 3:} && X_i \sim 0.9\mathrm{N}(220, 4) + 0.1\mathrm{N}(230, 8)\\
&\text{model 4:} && X_i \sim 0.2\mathrm{N}(210, 8) + 0.6\mathrm{N}(220, 4) + 0.2\mathrm{N}(230, 8)\\
&\text{model 5:} && X_i \sim 0.2\mathrm{N}(200, 8) + 0.6\mathrm{N}(220, 4) + 0.2\mathrm{N}(240, 8)\\
&\text{model 6:} && X_i \sim 0.2\mathrm{N}(210, 4) + 0.6\mathrm{N}(220, 4) + 0.2\mathrm{N}(230, 4)\\
&\text{model 7:} && X_i \sim 0.2\mathrm{N}(200, 4) + 0.6\mathrm{N}(220, 4) + 0.2\mathrm{N}(240, 4)\\
&\text{model 8:} && X_i \sim 0.6\mathrm{N}(220, 12) + 0.4\mathrm{N}(220, 2)\\
&\text{model 9:} && X_i \sim 0.2\mathrm{N}(212, 4) + 0.6\mathrm{N}(220, 8) + 0.2\mathrm{N}(228, 6)
\end{aligned}
$$

We depict densities of distributions for these models in Figures 3.2.3, 3.2.4 and 3.2.5.

Throughout the paper, we impose $\alpha = \beta = 0.05$, AQL $= 0.02$ and RQL $= 0.05$ if it is not stated otherwise. For these values, the true asymptotically optimal sampling plan (n^*, c^*) is given in Table 3.2.1.

Tabelle 3.2.1: True sampling plan for models 1–9.

model	1	2	3	4	5	6	7	8	9
n^*	65	103	209	168	608	324	1196	36	162
c^*	14.9	30.5	18.2	24.5	37.6	32.4	58.2	11.8	23.1

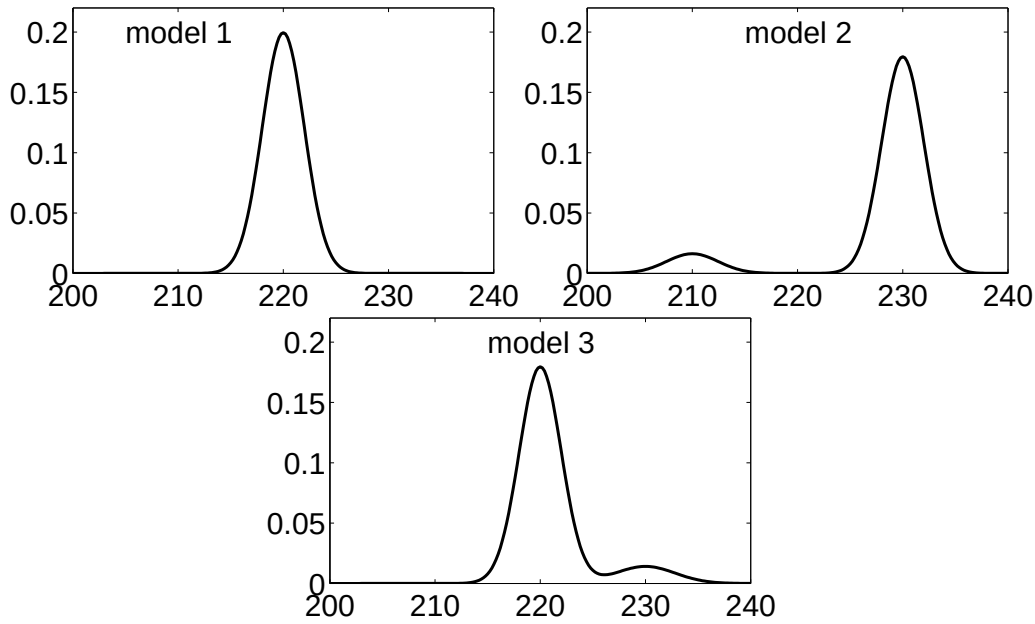


Abbildung 3.2.3: Densities of distributions for models 1–3.

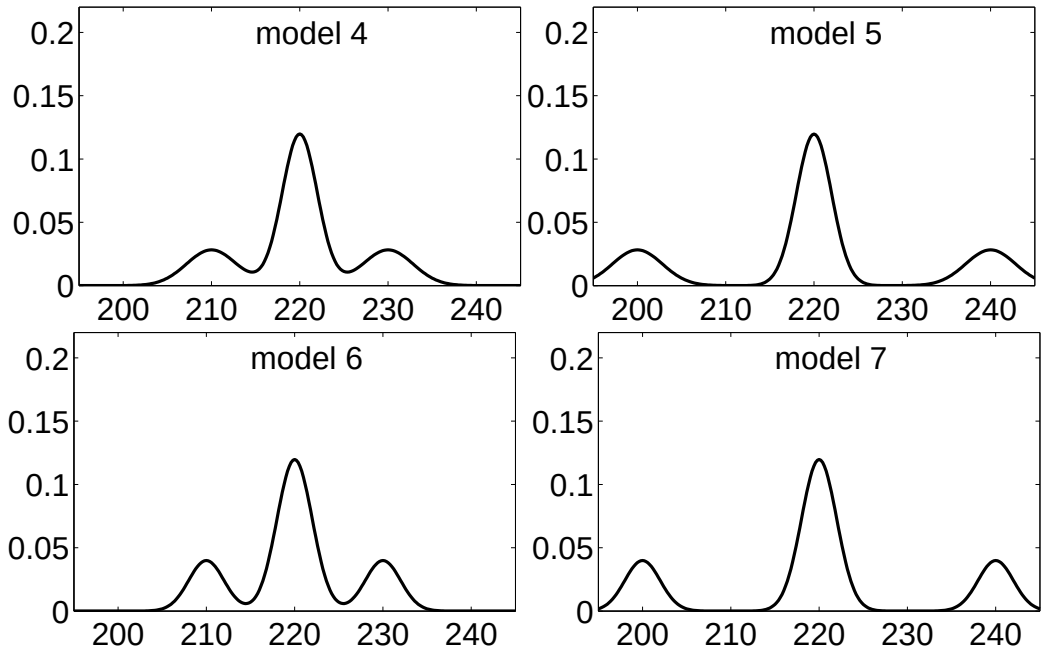


Abbildung 3.2.4: Densities of distributions for models 4–7.

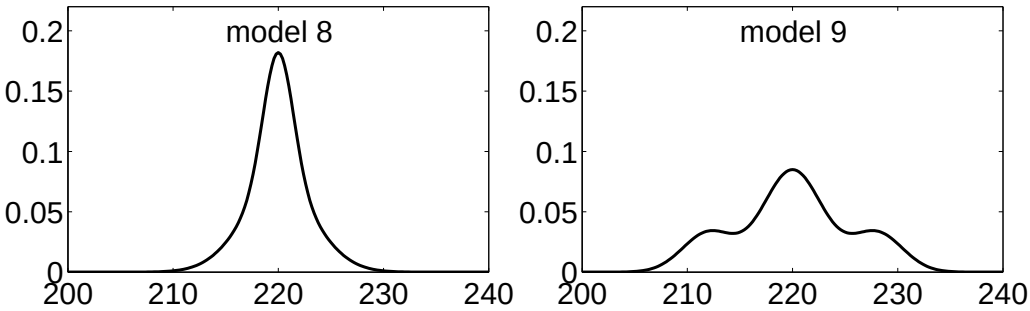


Abbildung 3.2.5: Densities of distributions for models 8 and 9.

3.2.5 Influence of standard quantile algorithms on sampling plans

Statistical software such as R or SAS implements various standard procedures to calculate sample quantiles. R implements nine algorithms discussed in (21). Algorithm 1 corresponds to the left inverse of the empirical distribution function, which is used as the standard procedure in the software tool APOS photovoltaic statlab, namely

$$F_m^{-1}(p) = \inf\{t \in \mathbb{R} : F_m(t) \geq p\}, \qquad p \in (0,1). \tag{3.2.5.1}$$

However, R's default algorithm is Algorithm 7. SAS's PROC UNIVARIATE also provides several methods; the default is to use the average of the npth and $(np+1)$-th order statistic, if np is an integer, and the $(\lfloor np \rfloor + 1)$th order statistic, otherwise, corresponding to Algorithm 2.

We used the parameters $\alpha = \beta = 0.05$, AQL $= 0.02$ and RQL $= 0.05$ and simulated data according to the model

$$X_i \sim \mathrm{N}(220, 4), \quad i = 1, \ldots, m.$$

For each algorithm and sample size $m = 100, 250, 500, 5000$, the same statistical quantities as above were estimated using $50,000$ simulation runs. For better comparison, for each algorithm the same random numbers were used by initializing the random number generator using the statement set.seed(17). The results are presented in Tables 3.2.2 and 3.2.3.

For Algorithm 1, we can see that the median $q_{0.5}$ is close to 65 for small m but the standard deviation of n_m is very large. In general, results for Algorithms 2–6 and 8–9 are rather close to the results for Algorithm 1; in some cases even identical to the results for Algorithm 1. However, the results for Algorithm 7, which is used in R by default, are worse than the results when using Algorithm 1. In particular, the median $q_{0.5}$ is substantially larger than 65. For $m \geq 5000$, there are no notable differences between algorithms.

Tabelle 3.2.2: Characteristics of distributions of n_m and c_m for Algorithms 1–5.

m	$E(n_m)$	$sd(n_m)$	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$E(c_m)$	$sd(c_m)$
			Algorithm 1				
100	195.4	1144.0	30	61	143	19.4	16.9
250	79.1	77.5	37	58	94	15.3	5.6
500	74.9	44.5	46	64	91	15.5	3.9
5000	65.6	10.5	58	65	72	14.9	1.1
			Algorithm 2				
100	162.8	424.6	40	75	156	19.2	12.3
250	95.1	93.3	45	70	112	16.7	6.2
500	78.1	44.8	49	67	95	15.8	3.9
5000	65.9	10.5	59	65	72	15	1.1
			Algorithm 3				
100	195.4	1144.0	30	61	143	19.4	16.9
250	101.6	114.4	44	70	118	17.3	7.0
500	74.9	44.5	46	64	91	15.5	3.9
5000	65.6	10.5	58	65	72	14.9	1.1
			Algorithm 4				
100	195.4	1144.0	30	61	143	19.4	16.9
250	88.2	88.9	40	64	104	16.2	6.1
500	74.9	44.5	46	64	91	15.5	3.9
5000	65.6	10.5	58	65	72	14.9	1.1
			Algorithm 5				
100	162.8	424.6	40	75	156	19.2	12.3
250	95.1	93.3	45	70	112	16.7	6.2
500	78.1	44.8	49	67	95	15.8	3.9
5000	65.9	10.5	59	65	72	15	1.1

Tabelle 3.2.3: Characteristics of distributions of n_m and c_m for Algorithms 6–9.

m	$E(n_m)$	$sd(n_m)$	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	$E(c_m)$	$sd(c_m)$
			Algorithm 6				
100	174	689.3	30	61	140	19.0	15.0
250	87.7	87.9	40	64	103	16.2	6.1
500	74.7	44.2	46	64	91	15.5	3.9
5000	65.5	10.5	58	65	72	14.9	1.1
			Algorithm 7				
100	301.6	1405.4	51	104	239	23.1	19.1
250	108.5	107	50	79	129	17.5	6.6
500	83.2	49.3	51	71	101	16.1	4.1
5000	66.2	10.6	59	65	73	15.0	1.1
			Algorithm 8				
100	157.1	437.8	36	69	147	18.9	12.5
250	92.0	90.1	43	68	109	16.5	6.1
500	76.8	44.3	48	66	93	15.7	3.8
5000	65.8	10.5	58	65	72	15.0	1.1
			Algorithm 9				
100	157.8	431.0	37	71	149	18.9	12.4
250	92.7	90.7	44	68	110	16.5	6.1
500	77.1	44.4	48	66	94	15.7	3.8
5000	65.8	10.5	58	65	72	15.0	1.1

Kapitel 3.3

Kernel estimator

In this chapter we study the approach based on employing kernels for estimating the density of the unknown distribution. We first provide short introduction to kernel density estimation and then describe methods of bandwidth selection. After that, we investigate the performance of the kernel density estimator with different bandwidths numerically for the problem of constructing sampling plans.

3.3.1 Kernel density estimator

Let $x_1, \ldots, x_m$ be independent identically distributed realizations of a random variable with continuous distribution F. For the Gaussian smoothing kernel, the kernel density estimator for the sample $x_1, \ldots, x_m$ is defined by

$$\hat{f}_{m,h}(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{(x - x_i)^2}{2h^2}\right),$$

where h is a smoothing parameter, which controls the degree of smoothness and should be determined from the sample. Using the notation $N(x, h^2) = (\sqrt{2\pi}h)^{-1} \exp(-(x - x_i)^2/(2h^2))$, the kernel density estimator has the form

$$\hat{f}_{m,h}(x) = \frac{1}{m} \sum_{i=1}^{m} N(x - x_i, h^2).$$

The general form of the kernel density estimator with arbitrary kernel K is

$$\hat{f}_{m,h}(x) = \frac{1}{mh} \sum_{i=1}^{m} K((x - x_i)/h).$$

In Figure 3.3.1, we depict the kernel density estimator for models 1 and 2 for different values of m and h, where models are defined as

$$\begin{aligned} &\text{model 1:} \quad X_i \sim \mathrm{N}(220, 4) \\ &\text{model 2:} \quad X_i \sim 0.1\mathrm{N}(210, 6) + 0.9\mathrm{N}(230, 4). \end{aligned}$$

We can see that the kernel density estimator becomes smoother as h increases. Therefore, one should choose an appropriate value of the bandwidth for use in the estimator. In the literature several methods of bandwidth selection have been proposed and they are comprehensively described in the following section.

3.3.2 Methods of bandwidth selection

3.3.2.1 Least-squared cross-validation

For estimating the bandwidth h, Bowman (1984) have proposed to minimize

$$\begin{aligned} \Phi_B(h) = \; & \frac{1}{m-1} N(0, 2h^2) + \frac{m-2}{m(m-1)^2} \sum_{i \neq j} N(x_i - x_j, 2h^2) \\ & - \frac{2}{m(m-1)} \sum_{i \neq j} N(x_i - x_j, h^2). \end{aligned}$$

In (Rudemo, 1982) the function $\Phi_B(h)$ has the form

$$\begin{aligned} \Phi_R(h) &= \frac{1}{m} N(0, 2h^2) + \frac{1}{m^2} \sum_{i \neq j} N(x_i - x_j, 2h^2) - \frac{2}{m(m-1)} \sum_{i \neq j} N(x_i - x_j, h^2) \\ &= \frac{1}{m^2} \sum_{i,j} N(x_i - x_j, 2h^2) - \frac{2}{m(m-1)} \sum_{i \neq j} N(x_i - x_j, h^2), \end{aligned}$$

which is used in (Steland and Herrmann, 2010).

Asymptotic consideration shows that the minimum of $\Phi_R(h)$ is attained at $h \approx h_0$ where $h_0 = 1.096 s_x m^{-1/5}$ and s_x is the standard deviation of the sample $(x_1, \ldots, x_m)$. Thus, h_0 can be used as an initial value for numerical minimization of $\Phi_R(h)$.

It was established (see e.g. Jones, Marron, Sheather, 1996) that the minimum of $\Phi_R(h)$ is sensitive to a sample and the bandwidth can be significantly underestimated. Therefore, we determine the LSCV bandwidth as

$$h_{\mathrm{LSCV}} = \arg \min_{h \in [0.1 h_0, 10 h_0]} \Phi_R(h).$$

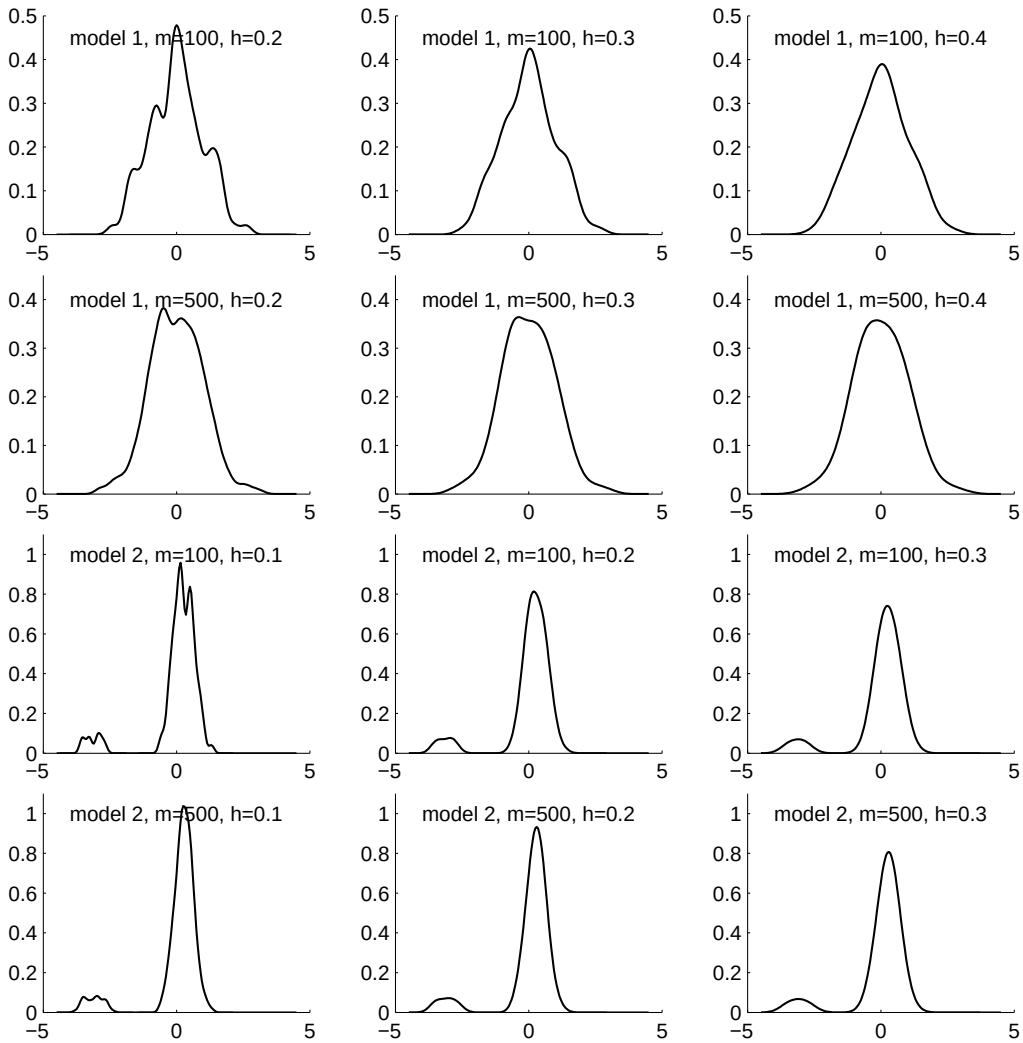


Abbildung 3.3.1: The kernel density estimator with different h for samples of size 100 and 500, when samples are generated from models 1 and 2.

3.3.2.2 Biased cross-validation

Scott and Terrel (1987) have considered the criterion

$$\Phi_{\mathrm{BCV}}(h) = \frac{R(K)}{mh} + h^4 \left(R(\hat{f}''_{m,h}) - \frac{R(K'')}{mh} \right) \left(\int x^2 K(x) dx / 2 \right)^2 ,$$

where $R(g) = \int g^2(x)dx$. Hence, the bandwidth h_{BCV} is defined as the smallest local minimizer of $\Phi_{\mathrm{BCV}}(h)$.

3.3.2.3 Normal reference distribution

The R-function `bw.nrd0` implements a rule-of-thumb for choosing the bandwidth of a Gaussian kernel density estimator. It defaults to 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power (known as Silverman's rule of thumb, Silverman (1986, page 48, eqn (3.31)) unless the quartiles coincide when a positive result will be guaranteed. Specifically,

$$h_{nrd0} = 0.9 \min\{s_x, (Q_m(0.75) - Q_m(0.25))/1.34\} m^{-1/5}$$

if $Q_m(0.75) - Q_m(0.25) > 0$ and $h_{nrd0} = 0.9 s_x m^{-1/5}$ otherwise, where $Q_m(p)$ is the empirical quantile function.

The R-function `bw.nrd` is the more common variation given by Scott (1992), using factor 1.06. Specifically,

$$h_{nrd} = 1.06 \min\{s_x, (Q_m(0.75) - Q_m(0.25))/1.34\} m^{-1/5}.$$

3.3.2.4 Sheather-Jones method

Sheather and Jones (1991) have proposed the following estimator of the bandwidth. Using the plug-in approach, the SJPI bandwidth is determined as

$$h_{SJPI} = \left(\frac{R(K)}{mR(\hat{f}''_{m,g(h)}) \left(\int x^2 K(x) dx \right)^2} \right)^{1/5},$$

where $g(h)$ is a preliminary estimator of the bandwidth, which is good for estimating the second derivative $f''(x)$ rather than the density $f(x)$ itself.

Alternatively, using the solve-the-equation approach, in which $g(h)$ is computed iteratively, the bandwidth h_{SJste} is defined.

3.3.2.5 Indirect cross-validation

Recently, Savchuk, Hart and Sheather (2010) have developed the indirect cross-validation (ICV) method. The idea of this method is to compute the bandwidth for a special kernel and then re-scale it for use in the kernel density estimator with the normal kernel.

The algorithm is as follows. First, compute $\alpha = 2.42$ and $\sigma = \max\{5.06, 0.149 m^{3/8}\}$. Then define a special kernel in the form

$$L(x|\alpha, \sigma) = (1 + \alpha) N(x) - \alpha/\sigma N(x/\sigma)$$

where $N(x) = (2\pi)^{1/2}\exp(-x^2/2)$ is the normal kernel. After that, we compute the LSCV bandwidth, say h_L, for a given sample using the special kernel $L(x) = L(x|\alpha, \sigma)$ through minimization of the criterion

$$\begin{aligned} Q_L(h) = & \frac{1}{mh} + \frac{1}{m^2 h}\sum_{i\neq j}\int L(x)L(x + (x_i - x_j)/h) \\ & -\frac{2}{m(m-1)h}\sum_{i\neq j} L((x_i - x_j)/h). \end{aligned}$$

Then, compute

$$h_N = \left(\frac{R_N \mu_L^2}{\mu_N^2 R_L}\right)^{1/5} h_L,$$

where $\mu_K = \int x^2 K(x)dx$, $R_K = \int K^2(x)dx$, K is either N or L, and

$$h_{\mathrm{OS}} = 3(70\sqrt{\pi}m)^{-1/5} s_x$$

where s_x is the standard deviation of the sample. Note that the bandwidth h_{OS} is defined by the maximal smoothing principle developed by Terrell (1990). Finally, taking

$$h_{\mathrm{ICV}} = \min\{h_N, h_{\mathrm{OS}}\},$$

we obtain the kernel density estimator with the normal kernel and the bandwidth h_{ICV}.

3.3.2.6 Method controlling the number of modes

Golyandina, Pepelyshev and Steland (2011) have proposed a method that controls the number of modes of the estimated density.

Specifically, we should initially determine a maximal value of h as

$$\bar{h} = \max\left\{h : \left\|F_m - \hat{F}_m(\hbar)\right\|_\infty \leq 1/R_m \ \forall\, \hbar \in (0, h)\right\}.$$

which exists since, by construction, $\left\|F_m - \hat{F}_m(h)\right\|_\infty = 0$ if $h = 0$, where $R_m = \dfrac{2\sqrt{m}}{\sqrt{2\ln\ln m}}$, $F_m(x)$ is the empirical distribution function and $\hat{F}_m(h)$ is the kernel density estimator with the bandwidth h. Then, the optimal bandwidth h_a belongs to the interval $[0, \bar{h}]$. Let $\{h_1, \ldots, h_n\}$ be a dense set for this interval. To proceed further, let M_h be the number of modes of estimated density for certain $h = h_i$. It is clear that the sequence

$(M_1, M_2, \ldots, M_n)$ has the decreasing tendency. If $M_{j+1} > M_j$ for some j, we say that the estimated density has a spurious mode, which may be ignored. Therefore, we define the sequence $(\check{M}_1, \check{M}_2, \ldots, \check{M}_n)$, where $\check{M}_j = \min\{M_1, M_2, \ldots, M_j\}$, which is monotonic.

Next, we divide the set $\{h_1, h_2, \ldots, h_n\}$ into groups such that the values $\check{M}_j$ are equal to each other within each group. Specifically, we define $a_1, b_1, \ldots, a_k, b_k$ and k such that

$$h_1 = a_1 \leq b_1 < \ldots < a_k \leq b_k = h_n,$$

$a_{i+1} = h_{j_i}$, $b_i = h_{j_i+1}$ and $\check{M}_i = \check{M}_j$ for all $i, j \in \{a_l, \ldots, b_l\}$, $l \in \{1, \ldots, k\}$.

Finally, we compute the optimal bandwidth as an average with weight coefficients, which are proportional to the sizes of these k groups, namely

$$h_{\text{GPS}} = \sum_{i=1}^{k} c_i w_i, \;\; c_i = \gamma_i a_i + (1 - \gamma_i) b_i, \;\; w_i = \frac{b_i - a_i}{\sum_{j=1}^{k} b_j - a_j},$$

where $\gamma_i = 1/2$ if $\check{M}_{a_i} = 1$ and $\gamma_i = 0.9$ otherwise. This means that c_i is the middle of interval $[a_i, b_i]$ provided one non-spurious mode and c_i is a little larger than the left bound of interval $[a_i, b_i]$ otherwise.

3.3.2.7 Local bandwidths

A kernel density estimator with local bandwidth can be found by use of the following two-stage procedure (Abramson, 1982). It is based on the construction of a local bandwidth factor λ_i at each sample point as follows

$$\lambda_i = \sqrt{\frac{G}{\tilde{f}(x_i)}}$$

where $G = (\prod_{i=1}^{m} \tilde{f}(x_i))^{1/m}$ and $\tilde{f}(x)$ is a density estimator with fixed bandwidth h that is found, for example, by LSCV. Then, the kernel density estimator with local bandwidth is defined by

$$\hat{f}(x) = \sum_{i=1}^{m} \frac{1}{\lambda_i h} K\left(\frac{x - x_i}{\lambda_i h}\right).$$

3.3.3 Simulation study for the kernel density estimator

Let us now study characteristics of the sampling plan obtained by use of the kernel density estimator with different bandwidths. We show results for models 1–9 in Tables 3.3.1–3.3.9, respectively.

We can see that the LSCV bandwidth does not perform well for all models while the BCV bandwidth is not quite good for model 9. The nrd0 bandwidth is not good for models 1,2,3,7,8 while the nrd bandwidth is not preferable for models 6,7,9. The SJPI bandwidth and the SJste bandwidth are not good for models 2,3,4,5,8. The ICV bandwidth is poor for model 7 while the local bandwidth does not give good results for models 1,3,4,5,6,7,9. The GPS bandwidth is not good for models 2,3.

It is very interesting that the BCV bandwidth is, on the average, the best among other bandwidths. The GPS and ICV bandwidths are the strong competitives to the BCV bandwidth. Specifically, the GPS bandwidth is slightly better than the BCV bandwidth for models 4,5,6,7,9 and slightly worse for models 1,2,8. We note that the BCV bandwidth typically provides smaller values of the standard deviation and RMSD but larger values of the bias.

Tabelle 3.3.1: Characteristics of distributions of n_m and c_m using different methods of bandwidth selection for model 1.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	LSCV	33	45	58	73	95	63.7	35.0	-1.3	35.0	15.6	3.0
100	BCV	34	46	60	75	91	61.4	21.7	-3.6	22.0	15.5	2.4
100	nrd0	32	47	67	89	112	70.1	31.1	5.1	31.5	15.8	3.2
100	nrd	33	47	63	81	99	64.8	25.0	-0.2	25.0	15.6	2.7
100	SJPI	33	47	63	81	102	66.3	29.0	1.3	29.0	15.7	2.9
100	SJste	33	47	63	80	100	65.2	26.6	0.2	26.6	15.6	2.8
100	ICV	34	46	60	76	92	62.0	22.3	-3.0	22.5	15.5	2.4
100	local	21	28	37	50	69	44.8	34.2	-20.2	39.7	12.8	3.0
100	GPS	34	46	59	73	89	60.9	22.0	-4.1	22.4	15.5	2.4
250	LSCV	40	49	60	72	86	62.5	21.1	-2.5	21.2	15.3	2.1
250	BCV	40	49	60	72	84	61.5	17.2	-3.5	17.5	15.3	1.9
250	nrd0	40	51	64	79	96	66.3	22.2	1.3	22.3	15.4	2.3
250	nrd	40	50	61	75	88	63.1	18.8	-1.9	18.9	15.3	2.0
250	SJPI	40	50	61	75	89	63.6	19.6	-1.4	19.6	15.3	2.1
250	SJste	40	50	61	75	88	63.3	19.1	-1.7	19.2	15.3	2.0
250	ICV	40	50	60	73	85	61.8	17.5	-3.2	17.8	15.3	1.9
250	local	27	33	40	50	63	44.1	19.6	-20.9	28.7	12.8	2.0
250	GPS	40	49	59	71	83	60.9	16.8	-4.1	17.3	15.3	1.8
500	LSCV	44	52	61	71	82	62.5	16.2	-2.5	16.4	15.2	1.7
500	BCV	45	52	61	71	81	61.8	13.9	-3.2	14.3	15.1	1.5
500	nrd0	44	53	63	75	88	64.8	17.1	-0.2	17.1	15.2	1.8
500	nrd	45	52	61	72	83	62.7	14.8	-2.3	15.0	15.1	1.6
500	SJPI	45	52	62	72	83	63.0	15.2	-2.0	15.4	15.2	1.6
500	SJste	45	52	62	72	83	62.8	15.1	-2.2	15.2	15.2	1.6
500	ICV	45	52	61	71	81	61.9	14.1	-3.1	14.5	15.1	1.5
500	local	32	37	43	51	61	45.7	14.2	-19.3	24.0	12.9	1.6
500	GPS	45	52	60	70	79	61.3	13.7	-3.7	14.2	15.1	1.5
5000	LSCV	57	58	63	66	72	63.1	6.0	-1.9	6.3	14.9	0.6
5000	BCV	57	58	63	66	69	62.8	5.4	-2.2	5.8	14.9	0.6
5000	nrd0	56	59	63	67	70	63.5	6.0	-1.5	6.2	14.9	0.6
5000	nrd	57	58	63	66	69	62.9	5.5	-2.1	5.8	14.9	0.6
5000	SJPI	57	59	63	66	69	63.0	5.5	-2.0	5.9	14.9	0.6
5000	SJste	57	59	63	66	69	63.0	5.5	-2.0	5.9	14.9	0.6
5000	ICV	56	58	62	65	68	61.5	5.0	-3.5	6.1	15.0	0.5
5000	local	48	50	54	58	65	55.3	6.8	-9.7	11.8	13.9	0.7
5000	GPS	56	59	63	67	70	63.3	6.0	-1.7	6.3	15.0	0.6

Compare tables 3.3.1, 3.4.2, 3.5.1, 3.5.19, 3.6.4.

Tabelle 3.3.2: Characteristics of distributions of n_m and c_m using different methods of bandwidth selection for model 2.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	LSCV	22	49	97	167	265	126.8	118.7	23.8	121.0	29.7	12.4
100	BCV	21	49	93	158	238	115.4	93.5	12.4	94.3	28.7	10.9
100	nrd0	22	51	102	179	284	131.5	113.5	28.5	117.0	30.3	12.5
100	nrd	22	50	98	166	251	120.4	97.0	17.4	98.6	29.3	11.4
100	SJPI	22	50	100	175	279	130.4	115.9	27.4	119.0	30.1	12.4
100	SJste	22	50	99	173	272	128.1	111.3	25.1	114.1	29.9	12.2
100	ICV	22	49	91	155	230	113.0	91.1	10.0	91.7	28.5	10.7
100	local	12	29	57	100	158	77.2	77.7	-25.8	81.8	23.2	9.8
100	GPS	22	50	100	175	276	130.3	115.0	27.3	118.2	30.2	12.4
250	LSCV	43	64	97	142	197	111.5	67.9	8.5	68.4	30.1	7.8
250	BCV	43	64	96	139	188	107.9	60.6	4.9	60.8	29.7	7.3
250	nrd0	43	66	100	148	204	113.7	66.9	10.7	67.7	30.4	7.9
250	nrd	43	65	96	140	188	108.1	59.9	5.1	60.1	29.8	7.3
250	SJPI	43	65	98	145	201	112.6	66.4	9.6	67.1	30.2	7.8
250	SJste	43	65	98	145	198	111.9	65.4	8.9	66.0	30.2	7.7
250	ICV	43	64	95	138	186	106.9	59.6	3.9	59.7	29.6	7.2
250	local	32	47	70	101	138	80.0	49.7	-23.0	54.8	25.6	6.5
250	GPS	43	65	97	143	196	110.6	64.2	7.6	64.6	30.0	7.6
500	LSCV	54	72	98	130	167	106.1	47.4	3.1	47.5	30.1	5.7
500	BCV	54	72	98	128	163	104.6	45.2	1.6	45.2	29.9	5.5
500	nrd0	54	73	99	132	170	107.4	47.9	4.4	48.1	30.2	5.8
500	nrd	54	72	98	128	162	103.9	44.0	0.9	44.0	29.8	5.4
500	SJPI	54	72	99	131	168	106.4	47.1	3.4	47.2	30.1	5.7
500	SJste	54	72	99	131	167	106.2	46.8	3.2	46.9	30.1	5.7
500	ICV	54	72	97	128	162	104.1	44.7	1.1	44.7	29.8	5.4
500	local	46	60	80	104	133	85.9	36.9	-17.1	40.6	27.2	4.8
500	GPS	54	72	98	129	164	105.0	45.7	2.0	45.7	29.9	5.6
5000	LSCV	84	91	101	111	121	102.7	16.3	-0.3	16.2	30.3	2.1
5000	BCV	83	90	101	111	120	102.2	15.7	-0.8	15.6	30.2	2.0
5000	nrd0	84	90	101	111	121	102.8	16.0	-0.2	15.9	30.3	2.1
5000	nrd	83	91	101	110	119	101.6	15.4	-1.4	15.3	30.2	2.0
5000	SJPI	83	90	101	111	121	102.4	15.8	-0.6	15.7	30.3	2.1
5000	SJste	83	90	101	111	121	102.4	15.8	-0.6	15.7	30.3	2.1
5000	ICV	82	89	97	106	116	98.2	13.6	-4.8	14.4	29.7	1.8
5000	local	84	91	102	109	120	102.4	15.0	-0.6	15.0	30.3	1.9
5000	GPS	83	90	99	109	118	100.0	14.5	-3.0	14.8	30.0	1.9

Compare tables 3.3.2, 3.4.3, 3.5.2, 3.5.20, 3.6.5.

Tabelle 3.3.3: Characteristics of distributions of n_m and c_m using different methods of bandwidth selection for model 3.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	LSCV	96	130	176	236	308	197.4	114.4	-11.6	115.0	18.5	3.6
100	BCV	97	128	169	217	268	177.3	68.1	-31.7	75.1	18.1	2.6
100	nrd0	99	138	193	259	330	205.6	91.3	-3.4	91.3	18.7	3.4
100	nrd	98	134	179	233	286	187.9	74.2	-21.1	77.1	18.4	2.8
100	SJPI	97	137	188	253	321	202.1	92.7	-6.9	92.9	18.6	3.4
100	SJste	97	136	186	248	314	198.6	87.1	-10.4	87.7	18.5	3.2
100	ICV	97	126	164	211	259	173.4	65.9	-35.6	74.9	18.1	2.5
100	local	63	84	118	165	232	141.8	109.3	-67.2	128.3	15.4	3.6
100	GPS	97	136	189	255	326	203.4	93.4	-5.6	93.5	18.6	3.4
250	LSCV	120	150	188	233	286	197.9	70.2	-11.1	71.1	18.3	2.6
250	BCV	122	150	184	226	265	190.2	57.0	-18.8	60.0	18.2	2.2
250	nrd0	122	154	195	245	293	202.8	67.5	-6.2	67.7	18.4	2.6
250	nrd	122	151	186	228	268	191.4	57.0	-17.6	59.6	18.2	2.2
250	SJPI	122	152	192	240	287	199.8	65.8	-9.2	66.4	18.3	2.5
250	SJste	122	152	192	238	284	198.5	64.2	-10.5	65.1	18.3	2.5
250	ICV	121	149	182	222	262	188.3	55.9	-20.7	59.6	18.2	2.2
250	local	86	106	134	171	219	146.6	61.9	-62.4	87.9	15.6	2.5
250	GPS	122	152	191	238	286	198.6	65.1	-10.4	65.9	18.3	2.5
500	LSCV	137	162	192	228	266	198.4	54.3	-10.6	55.3	18.2	2.1
500	BCV	138	162	190	223	258	194.6	46.7	-14.4	48.9	18.2	1.8
500	nrd0	138	165	196	234	273	201.3	52.4	-7.7	52.9	18.3	2.0
500	nrd	138	162	189	221	254	193.3	45.2	-15.7	47.9	18.2	1.8
500	SJPI	138	164	194	230	267	199.2	51.0	-9.8	51.9	18.2	2.0
500	SJste	138	164	193	230	266	198.6	50.3	-10.4	51.3	18.2	2.0
500	ICV	138	161	189	222	255	193.4	45.9	-15.6	48.5	18.2	1.8
500	local	105	122	144	173	211	153.4	49.0	-55.6	74.1	15.9	2.0
500	GPS	138	163	193	228	264	197.6	49.6	-11.4	50.9	18.2	1.9
5000	LSCV	173	182	199	215	230	200.3	21.5	-8.7	23.1	18.0	0.8
5000	BCV	173	182	199	212	227	199.0	19.9	-10.0	22.2	18.0	0.8
5000	nrd0	174	182	200	214	229	200.2	20.6	-8.8	22.3	18.0	0.8
5000	nrd	173	181	199	210	223	197.7	19.2	-11.3	22.2	18.0	0.8
5000	SJPI	174	182	200	212	227	199.6	20.3	-9.4	22.3	18.0	0.8
5000	SJste	174	182	200	212	227	199.6	20.2	-9.4	22.2	18.0	0.8
5000	ICV	173	184	195	207	220	195.8	18.0	-13.2	22.4	18.1	0.7
5000	local	156	169	185	199	212	185.8	22.7	-23.2	32.4	17.2	0.9
5000	GPS	176	187	201	216	230	201.9	21.4	-7.1	22.5	18.1	0.9

Compare tables 3.3.3, 3.4.4, 3.5.3, 3.5.21, 3.6.6.

Tabelle 3.3.4: Characteristics of distributions of n_m and c_m using different methods of bandwidth selection for model 4.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	LSCV	70	107	162	234	324	185.6	117.8	17.6	119.1	24.8	6.8
100	BCV	66	90	132	195	260	150.1	80.2	-17.9	82.1	22.9	5.3
100	nrd0	71	101	140	182	227	146.0	62.5	-22.0	66.2	22.9	4.6
100	nrd	69	94	125	158	194	129.4	49.9	-38.6	63.1	21.9	3.9
100	SJPI	71	109	165	237	314	182.1	99.1	14.1	100.1	24.7	6.2
100	SJste	71	108	158	221	286	171.4	86.4	3.4	86.4	24.2	5.7
100	ICV	68	95	136	190	249	149.2	74.8	-18.8	77.1	22.9	5.0
100	local	61	86	126	183	263	150.3	106.8	-17.7	108.3	22.4	6.2
100	GPS	72	104	150	211	282	166.5	87.0	-1.5	87.0	23.9	5.6
250	LSCV	91	121	161	215	267	173.8	74.0	5.8	74.3	24.5	4.8
250	BCV	93	120	159	207	252	167.1	63.9	-0.9	63.9	24.2	4.2
250	nrd0	92	117	147	183	216	151.8	48.8	-16.2	51.4	23.3	3.5
250	nrd	91	112	137	166	193	140.2	40.2	-27.8	48.8	22.7	3.0
250	SJPI	92	121	163	214	267	172.7	69.7	4.7	69.8	24.5	4.6
250	SJste	92	121	161	210	257	169.3	65.6	1.3	65.6	24.3	4.4
250	ICV	92	119	156	202	247	164.6	61.6	-3.4	61.6	24.0	4.1
250	local	84	107	141	184	234	152.6	64.9	-15.4	66.7	23.0	4.3
250	GPS	92	120	157	203	248	165.6	63.2	-2.4	63.2	24.1	4.2
500	LSCV	106	129	162	201	243	169.5	55.3	1.5	55.3	24.4	3.6
500	BCV	106	129	160	196	235	166.2	50.6	-1.8	50.6	24.2	3.4
500	nrd0	106	126	151	179	206	154.2	39.2	-13.8	41.6	23.6	2.7
500	nrd	105	122	143	167	190	145.5	33.1	-22.5	40.0	23.1	2.4
500	SJPI	107	130	162	199	240	168.7	53.0	0.7	53.0	24.4	3.5
500	SJste	107	130	161	197	236	167.1	51.1	-0.9	51.2	24.3	3.4
500	ICV	107	129	159	194	232	165.0	49.4	-3.0	49.5	24.2	3.3
500	local	101	121	149	183	222	156.6	50.8	-11.4	52.1	23.4	3.4
500	GPS	106	129	158	192	229	163.6	48.0	-4.4	48.2	24.1	3.2
5000	LSCV	140	152	164	174	187	163.3	18.1	-4.7	18.6	24.2	1.3
5000	BCV	140	152	164	175	188	162.9	17.8	-5.1	18.4	24.1	1.3
5000	nrd0	138	149	160	170	180	158.7	15.6	-9.3	18.1	23.9	1.1
5000	nrd	136	146	157	165	174	155.5	14.4	-12.5	19.0	23.7	1.1
5000	SJPI	140	152	165	175	188	163.1	17.8	-4.9	18.4	24.2	1.3
5000	SJste	140	152	164	175	187	163.0	17.7	-5.0	18.3	24.1	1.3
5000	ICV	140	149	162	172	182	161.2	16.4	-6.8	17.7	24.1	1.2
5000	local	142	155	168	176	188	166.2	18.2	-1.8	18.2	24.3	1.3
5000	GPS	141	151	164	176	188	163.9	17.8	-4.1	18.2	24.3	1.3

Compare tables 3.3.4, 3.4.5, 3.5.4, 3.5.22, 3.6.7.

Tabelle 3.3.5: Characteristics of distributions of n_m and c_m using different methods of bandwidth selection for model 5.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	LSCV	254	389	584	844	1165	665.9	403.8	57.9	407.9	43.8	11.8
100	BCV	246	374	540	747	965	581.4	286.4	-26.6	287.6	41.6	9.4
100	nrd0	185	250	432	610	795	462.2	243.4	-145.8	283.7	37.7	9.4
100	nrd	152	232	390	539	683	406.8	206.4	-201.2	288.2	35.7	8.6
100	SJPI	257	393	589	830	1090	640.3	335.4	32.3	337.0	43.4	10.6
100	SJste	257	381	546	741	964	588.8	290.7	-19.2	291.3	42.0	9.7
100	ICV	257	374	532	724	929	570.7	270.0	-37.3	272.5	41.4	8.8
100	local	224	318	465	672	958	546.1	349.6	-61.9	354.9	39.8	10.5
100	GPS	259	381	547	762	992	596.4	296.2	-11.6	296.4	42.1	9.4
250	LSCV	330	436	586	777	970	627.6	263.8	19.6	264.5	43.4	8.4
250	BCV	334	434	572	744	908	603.7	230.6	-4.3	230.6	42.8	7.5
250	nrd0	303	396	515	648	775	530.9	183.0	-77.1	198.5	40.6	6.6
250	nrd	289	381	481	588	692	488.7	155.8	-119.3	196.2	39.2	5.9
250	SJPI	332	438	587	769	953	621.3	248.0	13.3	248.3	43.3	8.1
250	SJste	333	436	578	747	915	606.8	232.5	-1.2	232.5	42.9	7.7
250	ICV	333	432	565	730	887	594.7	221.9	-13.3	222.3	42.5	7.3
250	local	305	391	516	678	865	558.8	232.5	-49.2	237.6	41.0	7.7
250	GPS	333	426	554	709	857	581.3	211.9	-26.7	213.6	42.1	7.0
500	LSCV	384	469	587	728	879	613.9	199.4	5.9	199.5	43.4	6.5
500	BCV	385	467	578	709	849	601.1	182.4	-6.9	182.5	43.0	6.0
500	nrd0	382	453	544	644	744	555.0	141.4	-53.0	151.0	41.6	4.9
500	nrd	375	439	515	600	682	522.8	120.0	-85.2	147.2	40.6	4.3
500	SJPI	385	470	586	722	869	609.9	190.8	1.9	190.8	43.3	6.3
500	SJste	385	469	582	714	853	603.9	184.1	-4.1	184.1	43.1	6.1
500	ICV	385	466	574	703	837	596.2	177.9	-11.8	178.2	42.9	5.9
500	local	367	443	545	673	817	574.5	185.3	-33.5	188.3	41.9	6.1
500	GPS	384	460	560	677	798	578.7	162.7	-29.3	165.3	42.3	5.4
5000	LSCV	507	549	594	638	680	593.3	66.5	-14.7	67.8	43.1	2.4
5000	BCV	505	549	592	633	680	590.4	64.7	-17.6	66.7	43.0	2.3
5000	nrd0	500	539	578	612	654	574.7	56.8	-33.3	65.6	42.5	2.0
5000	nrd	493	528	566	599	634	562.9	52.1	-45.1	68.7	42.1	1.9
5000	SJPI	505	549	593	638	682	591.1	65.1	-16.9	66.9	43.0	2.3
5000	SJste	505	549	593	636	680	590.6	64.7	-17.4	66.6	43.0	2.3
5000	ICV	505	539	585	624	662	583.5	59.4	-24.5	64.2	42.8	2.1
5000	local	518	563	611	646	694	605.9	68.4	-2.1	68.1	43.4	2.4
5000	GPS	509	545	592	633	671	590.3	62.7	-17.7	65.0	43.0	2.2

Compare tables 3.3.5, 3.4.6, 3.5.5, 3.5.23, 3.6.8.

Tabelle 3.3.6: Characteristics of distributions of n_m and c_m using different methods of bandwidth selection for model 6.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	LSCV	137	192	269	369	495	301.8	167.4	-22.2	168.9	30.6	7.1
100	BCV	123	178	246	326	410	259.3	111.9	-64.7	129.3	29.0	5.5
100	nrd0	122	157	199	250	305	207.6	72.9	-116.4	137.3	26.7	4.3
100	nrd	108	138	170	212	255	177.1	57.9	-146.9	157.9	25.1	3.7
100	SJPI	139	197	276	372	472	294.2	133.1	-29.8	136.4	30.5	6.2
100	SJste	138	191	260	338	417	270.9	111.4	-53.1	123.4	29.6	5.6
100	ICV	124	174	236	309	387	249.2	105.7	-74.8	129.5	28.5	5.2
100	local	109	149	212	298	421	248.0	159.6	-76.0	176.8	27.7	6.9
100	GPS	134	188	264	360	463	285.8	134.4	-38.2	139.7	30.1	6.3
250	LSCV	175	223	285	363	442	301.3	112.1	-22.7	114.4	31.0	5.2
250	BCV	177	219	277	346	411	287.5	93.0	-36.5	99.9	30.5	4.4
250	nrd0	166	198	235	277	319	239.3	60.1	-84.7	103.8	28.4	3.2
250	nrd	154	179	209	243	276	212.7	47.8	-111.3	121.2	27.1	2.7
250	SJPI	177	225	287	362	431	298.1	100.8	-25.9	104.0	31.0	4.8
250	SJste	177	222	281	351	414	290.4	93.6	-33.6	99.5	30.6	4.5
250	ICV	176	217	273	340	402	282.9	89.9	-41.1	98.9	30.3	4.3
250	local	154	192	244	312	392	263.2	105.2	-60.8	121.5	29.0	5.0
250	GPS	176	222	283	359	432	297.0	102.3	-27.0	105.8	30.9	4.8
500	LSCV	203	242	292	352	415	302.6	86.3	-21.4	89.0	31.3	4.0
500	BCV	203	241	287	343	396	295.3	76.5	-28.7	81.7	31.0	3.6
500	nrd0	193	221	254	290	323	256.7	50.6	-67.3	84.2	29.3	2.6
500	nrd	182	206	233	261	287	234.0	40.6	-90.0	98.7	28.3	2.2
500	SJPI	204	243	292	350	408	300.6	80.1	-23.4	83.5	31.2	3.8
500	SJste	204	242	289	345	399	296.8	76.9	-27.2	81.5	31.1	3.6
500	ICV	203	240	285	340	393	292.7	74.7	-31.3	81.0	30.9	3.5
500	local	185	218	262	317	379	275.3	83.0	-48.7	96.3	29.8	3.9
500	GPS	204	242	291	348	407	299.8	80.3	-24.2	83.8	31.2	3.8
5000	LSCV	266	287	308	328	348	307.9	31.6	-16.1	35.3	31.7	1.6
5000	BCV	265	288	309	328	350	306.8	30.8	-17.2	35.2	31.6	1.5
5000	nrd0	257	272	291	308	320	289.8	24.5	-34.2	42.0	30.9	1.2
5000	nrd	250	263	279	292	305	278.5	21.4	-45.5	50.2	30.4	1.1
5000	SJPI	266	288	310	328	351	307.3	30.9	-16.7	35.0	31.7	1.5
5000	SJste	266	288	309	328	350	306.9	30.8	-17.1	35.1	31.6	1.5
5000	ICV	262	279	299	316	332	297.9	26.6	-26.1	37.2	31.3	1.3
5000	local	270	290	315	331	348	311.5	32.0	-12.5	34.2	31.8	1.6
5000	GPS	269	287	312	333	354	311.1	32.0	-12.9	34.4	31.9	1.6

Compare tables 3.3.6, 3.4.7, 3.5.6, 3.5.24, 3.6.9.

Tabelle 3.3.7: Characteristics of distributions of n_m and c_m using different methods of bandwidth selection for model 7.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	LSCV	501	707	982	1355	1808	1098.5	578.9	-106.5	588.5	55.3	12.5
100	BCV	472	662	898	1168	1459	935.9	390.5	-269.1	474.3	51.8	9.6
100	nrd0	226	425	676	915	1130	682.6	340.6	-522.4	623.6	44.9	11.3
100	nrd	173	380	595	782	943	581.5	285.5	-623.5	685.8	41.8	10.5
100	SJPI	508	720	992	1308	1629	1042.7	452.0	-162.3	480.2	54.5	10.8
100	SJste	485	662	877	1150	1446	931.0	388.2	-274.0	475.1	51.9	10.0
100	ICV	483	656	872	1123	1398	915.4	368.9	-289.6	468.9	51.4	8.8
100	local	401	547	774	1101	1537	900.5	538.6	-304.5	618.6	49.9	12.1
100	GPS	504	708	976	1321	1685	1048.9	477.7	-156.1	502.5	54.4	10.9
250	LSCV	648	825	1054	1345	1638	1111.6	403.8	-93.4	414.4	56.5	9.3
250	BCV	655	806	1016	1269	1505	1055.2	338.0	-149.8	369.7	55.3	7.9
250	nrd0	549	710	865	1029	1189	864.3	257.9	-340.7	427.3	50.7	7.4
250	nrd	515	652	775	903	1025	767.6	215.6	-437.4	487.6	48.1	6.7
250	SJPI	657	829	1053	1325	1580	1094.1	364.1	-110.9	380.6	56.2	8.5
250	SJste	653	813	1024	1266	1503	1057.2	335.7	-147.8	366.7	55.4	8.1
250	ICV	648	798	999	1242	1464	1036.0	325.6	-169.0	366.8	54.9	7.7
250	local	566	711	903	1158	1461	972.7	373.8	-232.3	440.1	52.8	8.9
250	GPS	651	811	1024	1285	1529	1069.8	356.3	-135.2	381.1	55.6	8.2
500	LSCV	754	899	1085	1306	1544	1123.3	317.6	-81.7	327.9	57.1	7.3
500	BCV	755	892	1062	1267	1460	1091.3	280.5	-113.7	302.7	56.4	6.6
500	nrd0	711	821	945	1078	1205	952.1	193.9	-252.9	318.7	53.2	5.0
500	nrd	674	763	865	970	1072	867.7	157.7	-337.3	372.4	51.1	4.3
500	SJPI	759	901	1084	1297	1508	1112.5	294.4	-92.5	308.6	56.9	6.9
500	SJste	757	896	1070	1275	1471	1096.2	280.8	-108.8	301.1	56.6	6.6
500	ICV	752	887	1052	1250	1443	1079.2	272.5	-125.8	300.1	56.2	6.4
500	local	686	815	976	1184	1422	1025.7	304.4	-179.3	353.3	54.5	7.2
500	GPS	751	888	1058	1257	1458	1086.6	279.2	-118.4	303.2	56.3	6.5
5000	LSCV	991	1074	1150	1229	1305	1150.5	119.5	-54.5	130.8	58.2	3.0
5000	BCV	990	1075	1147	1210	1295	1141.4	114.8	-63.6	130.7	58.0	2.9
5000	nrd0	958	1016	1080	1144	1196	1079.0	91.4	-126.0	155.4	56.5	2.3
5000	nrd	930	979	1036	1088	1139	1037.0	80.2	-168.0	185.9	55.5	2.1
5000	SJPI	990	1073	1148	1218	1304	1143.7	115.2	-61.3	130.0	58.0	2.9
5000	SJste	991	1072	1147	1213	1302	1142.1	114.6	-62.9	130.3	58.0	2.9
5000	ICV	975	1033	1108	1173	1231	1105.3	98.2	-99.7	139.8	57.2	2.4
5000	local	1003	1089	1183	1242	1329	1171.2	124.7	-33.8	128.6	58.6	3.1
5000	GPS	995	1064	1155	1234	1305	1153.0	119.5	-52.0	130.2	58.3	2.9

Compare tables 3.3.7, 3.4.8, 3.5.7, 3.5.25, 3.6.10.

Tabelle 3.3.8: Characteristics of distributions of n_m and c_m using different methods of bandwidth selection for model 8.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	LSCV	18	27	40	59	79	46.4	29.2	10.4	31.0	13.3	3.5
100	BCV	19	27	39	54	70	42.3	19.9	6.3	20.9	13.1	2.8
100	nrd0	17	27	42	63	88	47.8	28.3	11.8	30.7	13.3	3.6
100	nrd	17	27	41	59	79	45.1	23.9	9.1	25.6	13.1	3.2
100	SJPI	17	27	42	62	85	47.4	28.1	11.4	30.3	13.3	3.5
100	SJste	17	27	42	61	83	46.6	26.5	10.6	28.6	13.2	3.4
100	ICV	19	28	39	54	70	42.3	20.0	6.3	21.0	13.1	2.8
100	local	13	18	25	36	53	31.1	24.6	-4.9	25.1	11.0	2.9
100	GPS	18	27	40	57	74	44.1	22.9	8.1	24.2	13.1	3.1
250	LSCV	21	28	38	50	63	40.8	17.9	4.8	18.5	12.5	2.4
250	BCV	22	28	38	49	60	39.5	15.1	3.5	15.5	12.5	2.2
250	nrd0	21	28	38	51	65	41.1	17.6	5.1	18.3	12.5	2.5
250	nrd	21	28	38	49	61	39.8	15.6	3.8	16.0	12.5	2.2
250	SJPI	21	28	38	51	64	40.9	17.4	4.9	18.1	12.5	2.4
250	SJste	21	28	38	50	64	40.7	17.0	4.7	17.6	12.5	2.4
250	ICV	22	28	37	48	59	39.3	14.7	3.3	15.0	12.5	2.1
250	local	17	21	28	36	45	30.1	13.2	-5.9	14.4	10.9	2.0
250	GPS	22	28	38	49	61	39.7	15.4	3.7	15.9	12.5	2.2
500	LSCV	24	29	37	46	57	38.9	13.4	2.9	13.8	12.3	1.9
500	BCV	24	30	37	45	54	38.3	12.1	2.3	12.3	12.3	1.7
500	nrd0	24	29	37	47	56	39.0	13.2	3.0	13.5	12.3	1.9
500	nrd	24	30	37	45	54	38.2	11.9	2.2	12.1	12.3	1.7
500	SJPI	24	29	37	46	56	38.9	13.1	2.9	13.5	12.3	1.9
500	SJste	24	29	37	46	56	38.8	12.9	2.8	13.2	12.3	1.8
500	ICV	24	30	37	45	54	38.2	11.8	2.2	12.0	12.3	1.7
500	local	21	24	30	36	44	31.3	10.4	-4.7	11.4	11.1	1.6
500	GPS	24	30	37	45	54	38.2	11.9	2.2	12.1	12.3	1.7
5000	LSCV	31	32	35	38	40	35.6	4.3	-0.4	4.3	11.8	0.6
5000	BCV	30	32	35	38	40	35.7	4.2	-0.3	4.2	11.8	0.6
5000	nrd0	31	32	35	38	40	35.7	4.2	-0.3	4.2	11.8	0.6
5000	nrd	30	32	35	38	40	35.6	4.1	-0.4	4.1	11.8	0.6
5000	SJPI	31	32	35	38	40	35.7	4.3	-0.3	4.2	11.8	0.6
5000	SJste	31	32	35	38	40	35.7	4.2	-0.3	4.2	11.8	0.6
5000	ICV	31	33	36	39	40	35.8	4.1	-0.2	4.1	11.9	0.6
5000	local	30	32	35	38	40	35.1	4.2	-0.9	4.2	11.7	0.6
5000	GPS	31	33	36	39	41	36.1	4.4	0.1	4.4	11.9	0.7

Compare tables 3.3.8, 3.4.9, 3.5.8, 3.5.26, 3.6.11.

Tabelle 3.3.9: Characteristics of distributions of n_m and c_m using different methods of bandwidth selection for model 9.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	LSCV	63	76	104	147	205	123.6	75.7	-38.4	84.9	20.6	4.5
100	BCV	60	72	86	99	111	86.0	20.2	-76.0	78.6	18.4	1.8
100	nrd0	69	89	112	135	157	112.4	33.8	-49.6	60.1	20.1	2.8
100	nrd	65	80	98	115	129	97.5	25.1	-64.5	69.2	19.2	2.2
100	SJPI	68	86	113	148	189	122.5	51.9	-39.5	65.3	20.6	3.6
100	SJste	67	84	107	134	164	112.3	40.1	-49.7	63.9	20.1	3.1
100	ICV	61	73	88	102	114	88.0	21.4	-74.0	77.0	18.6	1.9
100	local	39	50	71	105	166	91.4	72.0	-70.6	100.8	17.5	4.7
100	GPS	60	71	84	102	125	89.9	29.6	-72.1	77.9	18.7	2.4
250	LSCV	84	102	127	160	197	136.1	49.1	-25.9	55.5	21.5	3.3
250	BCV	78	89	102	118	137	106.3	27.4	-55.7	62.1	19.8	2.0
250	nrd0	85	101	119	140	158	120.6	27.9	-41.4	49.9	20.6	2.1
250	nrd	81	93	107	123	136	107.8	21.4	-54.2	58.3	19.9	1.7
250	SJPI	87	104	128	157	187	133.6	40.4	-28.4	49.3	21.4	2.8
250	SJste	86	102	123	149	172	127.0	34.8	-35.0	49.4	21.0	2.5
250	ICV	79	90	103	120	137	107.0	26.4	-55.0	61.0	19.8	2.0
250	local	59	73	96	125	162	105.0	46.8	-57.0	73.8	18.8	3.4
250	GPS	77	89	102	120	141	106.4	27.0	-55.6	61.8	19.8	2.0
500	LSCV	99	116	136	162	190	141.8	38.0	-20.2	43.0	21.9	2.5
500	BCV	95	109	125	147	169	129.4	29.5	-32.6	44.0	21.2	2.1
500	nrd0	97	110	125	142	158	126.6	23.5	-35.4	42.5	21.1	1.8
500	nrd	92	102	115	128	140	115.4	18.6	-46.6	50.2	20.4	1.5
500	SJPI	100	117	136	160	184	140.0	33.3	-22.0	39.9	21.8	2.3
500	SJste	99	114	133	154	176	135.6	30.2	-26.4	40.1	21.6	2.1
500	ICV	94	107	122	142	163	126.2	27.8	-35.8	45.3	21.0	2.0
500	local	77	91	109	132	161	115.6	37.0	-46.4	59.3	19.7	2.6
500	GPS	91	103	117	133	151	119.6	24.4	-42.4	48.9	20.6	1.8
5000	LSCV	133	142	152	162	168	152.0	14.9	-10.0	17.9	22.5	1.1
5000	BCV	132	142	152	160	171	151.0	14.2	-11.0	17.9	22.5	1.0
5000	nrd0	128	136	144	151	158	143.4	11.5	-18.6	21.8	22.0	0.8
5000	nrd	124	130	137	145	150	137.2	9.9	-24.8	26.6	21.7	0.7
5000	SJPI	132	142	152	160	171	151.5	14.4	-10.5	17.7	22.5	1.0
5000	SJste	132	141	152	159	169	151.0	14.1	-11.0	17.8	22.5	1.0
5000	ICV	128	136	144	151	158	143.5	11.4	-18.5	21.7	22.1	0.8
5000	local	129	137	148	156	164	147.2	15.7	-14.8	21.5	22.0	1.1
5000	GPS	131	140	149	158	166	148.9	13.3	-13.1	18.7	22.4	0.9

Compare tables 3.3.9, 3.4.10, 3.5.9, 3.5.27, 3.6.12.

3.3.4 Simulation study for the two-stage procedure

Let us describe a two-stage procedure for computing plans, which is implemented in the software tool APOS photovoltaic statlab. At the first stage, the hypothesis on the normality of a sample is tested using, for example, the Shapiro criterion. If the hypothesis is accepted, then the sampling plan is computed using the normal distribution. If the hypothesis is rejected, then the sampling plan is evaluated by using the kernel density estimator with some bandwidth for the given sample. From this point of view, the procedure studied in previous section can be called the one-stage procedure.

We present the characteristics of distributions of n_m and c_m for models 8 and 9, which are close to the normal distribution, using the two-stage procedure in Tables 3.3.10 and 3.3.11 and the one-stage procedure in Tables 3.3.8 and 3.3.9. We can see that $\mathrm{E}\, n_m$ has more bias for the two-stage procedure than for the one-stage-procedure.

Tabelle 3.3.10: Characteristics of distributions of n_m and c_m using the two-stage procedure and different methods of bandwidth selection for model 8.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	LSCV	19	39	65	65	65	55.2	24.9	19.2	31.4	13.8	2.9
100	BCV	20	37	65	65	65	53.2	19.9	17.2	26.3	13.7	2.4
100	nrd0	19	40	65	65	65	55.2	22.9	19.2	29.9	13.8	2.9
100	nrd	19	38	65	65	65	54.2	21.1	18.2	27.9	13.8	2.6
100	SJPI	19	40	65	65	65	55.3	23.4	19.3	30.3	13.8	2.9
100	SJste	19	39	65	65	65	55.0	22.5	19.0	29.4	13.8	2.8
100	ICV	19	37	65	65	65	55.2	28.5	19.2	34.4	13.8	3.0
100	local	15	25	65	65	65	49.6	25.7	13.6	29.1	13.0	3.0
100	GPS	19	38	65	65	65	54.2	21.3	18.2	28.0	13.8	2.6
250	LSCV	21	28	41	63	65	43.8	19.1	7.8	20.6	12.8	2.5
250	BCV	22	29	40	60	65	42.7	17.2	6.7	18.5	12.7	2.2
250	nrd0	21	28	41	64	65	43.8	18.7	7.8	20.3	12.7	2.5
250	nrd	21	28	40	60	65	42.9	17.4	6.9	18.7	12.7	2.3
250	SJPI	21	28	41	64	65	43.8	18.7	7.8	20.3	12.7	2.5
250	SJste	21	28	41	63	65	43.6	18.4	7.6	19.9	12.7	2.4
250	ICV	22	29	39	56	65	41.9	16.6	5.9	17.6	12.7	2.1
250	local	17	22	29	45	65	35.4	18.1	-0.6	18.1	11.4	2.4
250	GPS	22	29	40	60	65	42.9	17.5	6.9	18.8	12.7	2.3
500	LSCV	24	29	37	46	58	39.1	13.7	3.1	14.0	12.3	1.9
500	BCV	24	30	37	45	55	38.6	12.4	2.6	12.6	12.3	1.8
500	nrd0	24	29	37	47	58	39.2	13.4	3.2	13.7	12.3	1.9
500	nrd	24	30	37	45	55	38.4	12.2	2.4	12.4	12.3	1.7
500	SJPI	24	29	37	47	58	39.1	13.3	3.1	13.7	12.3	1.9
500	SJste	24	29	37	46	57	39.0	13.2	3.0	13.5	12.3	1.9
500	ICV	25	30	36	43	51	37.2	10.5	1.2	10.5	12.3	1.5
500	local	21	24	30	36	45	31.7	11.0	-4.3	11.8	11.1	1.6
500	GPS	24	30	37	45	55	38.4	12.2	2.4	12.4	12.3	1.7
5000	LSCV	31	33	36	40	42	36.2	4.6	0.2	4.6	11.9	0.7
5000	BCV	31	33	36	39	41	36.1	4.5	0.1	4.4	11.9	0.7
5000	nrd0	31	33	36	39	42	36.2	4.5	0.2	4.5	11.9	0.7
5000	nrd	31	33	36	39	41	36.0	4.3	0.0	4.3	11.9	0.7
5000	SJPI	31	33	36	39	42	36.2	4.5	0.2	4.5	11.9	0.7
5000	SJste	31	33	36	39	42	36.2	4.5	0.2	4.5	11.9	0.7
5000	ICV	30	33	36	41	44	36.9	5.6	0.9	5.7	11.9	0.8
5000	local	30	32	35	39	41	35.5	4.4	-0.5	4.4	11.7	0.7
5000	GPS	31	33	36	39	41	36.1	4.4	0.1	4.4	11.9	0.7

Compare tables 3.3.8, 3.4.9, 3.5.8, 3.5.26, 3.6.11.

Tabelle 3.3.11: Characteristics of distributions of n_m and c_m using the two-stage procedure and different methods of bandwidth selection for model 9.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	LSCV	65	65	65	65	134	83.3	58.0	-78.7	97.8	16.3	3.9
100	BCV	65	65	65	65	94	70.6	16.3	-91.4	92.9	15.6	1.9
100	nrd0	65	65	65	65	125	76.2	30.3	-85.8	91.0	16.0	2.8
100	nrd	65	65	65	65	107	72.8	21.4	-89.2	91.7	15.8	2.3
100	SJPI	65	65	65	65	141	81.6	47.6	-80.4	93.4	16.3	3.6
100	SJste	65	65	65	65	130	78.1	36.6	-83.9	91.6	16.1	3.1
100	ICV	65	65	65	65	107	78.9	52.8	-83.1	98.4	16.1	3.5
100	local	65	65	65	65	94	77.4	47.8	-84.6	97.1	15.9	3.2
100	GPS	65	65	65	65	100	73.3	25.1	-88.7	92.2	15.8	2.4
250	LSCV	65	65	118	158	195	123.0	57.4	-39.0	69.4	20.0	4.4
250	BCV	65	65	98	118	137	98.7	32.7	-63.3	71.2	18.7	3.0
250	nrd0	65	65	112	139	157	108.7	37.2	-53.3	65.0	19.3	3.4
250	nrd	65	65	102	122	136	98.7	28.5	-63.3	69.4	18.7	2.9
250	SJPI	65	65	120	156	187	120.6	50.3	-41.4	65.2	19.9	4.1
250	SJste	65	65	116	148	172	114.9	44.3	-47.1	64.6	19.6	3.8
250	ICV	65	65	96	117	136	98.9	35.8	-63.1	72.6	18.6	3.1
250	local	65	65	87	121	159	100.6	47.0	-61.4	77.3	18.3	3.6
250	GPS	65	65	98	119	140	98.9	32.3	-63.1	70.8	18.7	3.0
500	LSCV	98	116	136	162	190	141.4	38.6	-20.6	43.8	21.8	2.6
500	BCV	95	108	125	147	169	129.0	30.1	-33.0	44.6	21.1	2.2
500	nrd0	96	110	125	142	158	126.2	24.3	-35.8	43.3	21.0	1.9
500	nrd	91	102	115	128	140	115.0	19.2	-47.0	50.7	20.3	1.6
500	SJPI	99	116	136	160	184	139.5	34.0	-22.5	40.8	21.8	2.4
500	SJste	98	114	133	154	176	135.2	30.9	-26.8	40.9	21.5	2.2
500	ICV	87	97	108	120	132	109.2	18.7	-52.8	56.1	20.0	1.5
500	local	76	91	109	132	161	115.5	37.1	-46.5	59.5	19.7	2.7
500	GPS	91	102	117	133	151	119.4	24.9	-42.6	49.4	20.6	1.9
5000	LSCV	133	142	153	163	172	153.2	14.9	-8.8	17.3	22.6	1.0
5000	BCV	133	142	153	162	170	151.9	14.0	-10.1	17.3	22.6	1.0
5000	nrd0	129	136	145	152	158	144.2	11.2	-17.8	21.0	22.1	0.8
5000	nrd	125	131	138	145	150	138.0	9.6	-24.0	25.9	21.8	0.7
5000	SJPI	133	142	153	162	171	152.6	14.2	-9.4	17.0	22.6	1.0
5000	SJste	133	142	153	161	170	152.0	14.0	-10.0	17.2	22.6	1.0
5000	ICV	139	150	163	178	193	164.3	21.6	2.3	21.7	23.2	1.5
5000	local	129	137	147	157	166	147.9	15.6	-14.1	21.0	22.1	1.1
5000	GPS	131	140	149	158	166	148.9	13.3	-13.1	18.7	22.4	0.9

Compare tables 3.3.9, 3.4.10, 3.5.9, 3.5.27, 3.6.12.

3.3.5 The double kernel density estimator

In the present section we introduce the new estimator which is called the double kernel estimator. This estimator essentially has the form of the weighted average of certain kernels with weights constructed on the base of the classical kernel estimator.

Let $x_1, \ldots, x_n$ be a sample from a distribution with unknown density $f(x)$. We define the double kernel estimator as

$$\hat{f}_2(x) = \gamma \sum_{j \in J} \frac{\widehat{b}_j}{h_j} K\left(\frac{x - \widetilde{x}_j}{h_j}\right) \tag{3.3.5.1}$$

where $\gamma = (\sum_{j \in J} \widehat{b}_j)^{-1}$ is the normalization constant, $\widetilde{x}_j = jh$ are equidistant points (which can be called support points), $\widehat{b}_j = \widehat{f}_1(\widetilde{x}_j)$ are weight coefficients, the function $\widehat{f}_1(x) = (nh)^{-1} \sum_{i=1}^n K((x - x_i)/h)$ serves as a preliminary density estimator, $J = \{j \in \mathbb{Z} : \widehat{b}_j \geq c\}$ is a finite set for some positive threshold c.

The introduction of the double kernel estimator is motivated by the form of the Bernstein-Durrmeyer estimator. Specifically, the weights in (3.3.5.1) are estimated from the sample in an analogous way as coefficients are determined in the Bernstein-Durrmeyer approach. Indeed, in this approach a function $f(x)$ is approximated by

$$f_N(x) = \sum_{k=0}^{N} a_k B_k^{(N)}(x)$$

where coefficients are defined by

$$a_k = \int_0^1 f(x) B_k^{(N)}(x)\, dx.$$

For large N the Bernstein polynomials almost have the form of a Gaussian density. Thus, the polynomial $B_k^{(N)}(x)$ looks similarly to $K((x - k/N)/h)$ with appropriate $h > 0$ and the standard Gaussian kernel K.

More generally, for an arbitrary but appropriate kernel K and a bandwidth $h > 0$ let us consider the estimator

$$\widetilde{f}(x) = \sum_{j \in \mathbb{Z}} \frac{b_j}{h} K\left(\frac{x - \widetilde{x}_j}{h}\right), \tag{3.3.5.2}$$

for given support points $\widetilde{x}_j = jh$ and the coefficients are defined as

$$b_j = \int_{-\infty}^{\infty} f(x) \frac{1}{h} K\left(\frac{x - \widetilde{x}_j}{h}\right) dx$$

in analogy to the Bernstein approach. Note that for the uniform kernel $K(z) = \frac{1}{2}1_{[-1,1]}(z)$, we have

$$b_j = \frac{1}{h}EK\left(\frac{\xi - \widetilde{x}_j}{h}\right) = \frac{P(\xi \in [\widetilde{x}_j - h, \widetilde{x}_j + h])}{\lambda(I_j)}$$

where ξ is a random variable with density $f(x)$, $I_j = [\widetilde{x}_j - h, \widetilde{x}_j + h]$ and λ denotes the Lebesgue measure.

3.3.5.1 Properties of the deterministic approximation

The following lemma shows that under quite general conditions $\widetilde{f}(x)$ provides an approximation to

$$\overline{f}(x) = \sum_{j\in\mathbb{Z}} \frac{f(\widetilde{x}_j)}{h} K\left(\frac{x - \widetilde{x}_j}{h}\right), \tag{3.3.5.3}$$

which is approximately proportional to $f(x)$, $\overline{f}(x) \approx f(x)\frac{1}{\widetilde{x}_1 - \widetilde{x}_0}\int_{\widetilde{x}_0}^{\widetilde{x}_1} \sum_{j\in\mathbb{Z}} \frac{1}{h}K\left(\frac{z-\widetilde{x}_j}{h}\right) dz$. Note that for the uniform kernel and $\widetilde{x}_j = 2jh$, $\overline{f}(x)$ is the step function having values $f(\widetilde{x}_j)$ on $I_j = [\widetilde{x}_j - h, \widetilde{x}_j + h]$.

Lemma 1. *Let $K(x)$ be a symmetric probability density with bounded support and suppose that $f \in C_b^2([a,b])$. Then the following assertions hold true,*

(i) $\max_{j\in\mathbb{Z}} |b_j - f(\widetilde{x}_j)| = O(h^2)$,

(ii) $\|\widetilde{f} - \overline{f}\|_\infty = O(h^2)$.

Proof. Due to assumptions we can replace the sum over $\mathbb{Z}$ in (3.3.5.2) and (3.3.5.3) by the sum over a finite set J. Using the Taylor expansion we obtain

$$\begin{aligned} b_j &= \int \frac{1}{h}K\left(\frac{x - \widetilde{x}_j}{h}\right) f(x)\, dx \\ &= \int K(z) f(\widetilde{x}_j + hz)\, dz \\ &= \int K(z)[f(\widetilde{x}_j) + f'(\widetilde{x}_j)hz + f''(\beta_j)h^2z^2]\, dz \\ &= f(\widetilde{x}_j) + O(h^2), \end{aligned}$$

where $\beta_j \in [\widetilde{x}_j - h, \widetilde{x}_j + h]$, since $f \in C_b^2([a,b])$ and $\|K''\|_\infty < \infty$ by assumption. Thus, we proved (i). The statement (ii) now follows from

$$\begin{aligned}
\left|\widetilde{f}(x) - \overline{f}(x)\right| &= \left|\sum_{j\in J}[b_j - f(\widetilde{x}_j)]\frac{1}{h}K\left(\frac{x-\widetilde{x}_j}{h}\right)\right| \\
&\le \max_{j\in J}\left|b_j - f(\widetilde{x}_j)\right| \cdot \frac{1}{h}\sum_{j\in J} K\left(\frac{x-\widetilde{x}_j}{h}\right) \\
&\le \max_{j\in J}\left|b_j - f(\widetilde{x}_j)\right| \cdot \|K_s\|_\infty \\
&= O(h^2).
\end{aligned}$$

□

Let us now reformulate the new estimator more rigorously in terms of orthogonal expansion. Let $K(x)$ be a symmetric probability density with $\int K^2(x)\,dx < \infty$. Suppose that the support points are selected as $\widetilde{x}_j = 2jh$. If $K(x)$ has support $(-1,1]$, then the functions

$$\varphi_j(x) = \varphi_j(x;h) = \frac{1}{h}K\left(\frac{x-\widetilde{x}_j}{h}\right), \qquad x \in \mathbb{R}, j \in \mathbb{Z},$$

are supported on intervals $I_j = ((2j-1)h, (2j+1)h]$. Moreover, we have

$$\int h^{-1}K([x-2jh]/h)h^{-1}K([x-2ih]/h)\,dx = 0$$

for $i,j \in \mathbb{Z}$ with $i \neq j$. Hence, the system $\{\varphi_j : j \in \mathbb{Z}\}$ forms an orthogonal family in $L_2(\mathbb{R};\mathbb{R})$ and the function

$$f_\infty(x) = \sum_{j\in\mathbb{Z}} b_j\varphi_j(x) = \sum_{j\in\mathbb{Z}} b_j\frac{1}{h}K\left(\frac{x-\widetilde{x}_j}{h}\right),$$

with coefficients

$$b_j = \int f(x)\varphi_j(x)\,dx = \int f(x)\frac{1}{h}K\left(\frac{x-\widetilde{x}_j}{h}\right)\,dx$$

is an approximation of $f(x)$. If $\{\varphi_j\}$ spans the subset of all L_2-functions which are constant on the intervals I_j, the function $f_\infty(x)$ is the orthogonal projection onto that subspace.

Remark 3.3.5.1. *If $K(x) = 1_{(0,1]}(x)$ is the one-sided uniform kernel, then*

$$f_\infty(x) = \sum_{j\in\mathbb{Z}} b_j\varphi_j(x) = \sum_{j\in 2\mathbb{Z}} b_j\frac{1}{h}K\left((x-\widetilde{x}_j)/h\right),$$

with $b_j = \int f(x)\varphi_j(x)\,dx = \int f(x)h^{-1}K\left((x-\widetilde{x}_j)/h\right)\,dx$ *is the orthogonal projection onto the space of functions which are constant on* $(j, j+1]$, $j \in \mathbb{Z}$. *In this case, and also for certain other kernels* K, *one can define a multiresolution analysis. Let* V_0 *denote the subspace of* $L_2(\mathbb{R})$ *of those functions which are constant on* $(j, j+1]$, $j \in \mathbb{Z}$. *Then* $\{\varphi_j(\bullet; 1) : j \in \mathbb{Z}\}$, *i.e. the family of functions corresponding to* $h = 1$, *is an ONS of* V_0. V_0 *is a subspace of* $V_1 = \{h \in L_2(\mathbb{R}) : h(x) = f(2x), f \in V_0\}$ *of those functions which are constant on* $(j/2, (j+1)/2]$. *An ONS of* V_1 *is given by the system of functions* $\varphi_{1j} = \sqrt{2}K(2x - j)$, $j \in \mathbb{Z}$. *With* $h = 1$ *we can write* $f_\infty = \sum_j b_j K(x - j)$. *We have* $V_1 = V_0 \oplus W_0$ *where the subspace* W_0 *is the orthogonal complement of* V_0 *in* V_1. W_0 *has the basis* $\{\psi_{0j}\}$, *where* $\psi_{0j}(x) = \psi(x - j)$ *and*

$$\psi(x) = -1_{[0,1/2]}(x) + 1_{(1/2,1]}(x).$$

More generally, $V_0 \subset V_1 \subset V_2 \subset \cdots$, *if* $V_k = \{h(x) = f(2^k x) : f \in V_0\}$, $k \in \mathbb{Z}$. *A basis of* V_k *is given by* $\{\varphi_{kj} : j \in \mathbb{Z}\}$, *where* $\varphi_{kj}(x) = 2^{k/2}K(2^k - j)$, $j \in \mathbb{Z}$.

3.3.5.2 Consistency

Given a sample $x_1, \ldots, x_n$ from a distribution $F(x)$, we can estimate the coefficients

$$b_j = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x - \widetilde{x}_j}{h}\right)\,dF(x)$$

by

$$\widehat{b}_j = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - \widetilde{x}_j}{h}\right).$$

Therefore, the double kernel estimator has the form

$$\widehat{f}_2(x) = \gamma \sum_{j \in J} \frac{\widehat{b}_j}{h} K\left(\frac{x - \widetilde{x}_j}{h}\right)$$

where $\gamma = 1/\sum_{j\in J}\widehat{b}_j$, $J \subset \mathbb{Z}$. This choice of γ implies that $\int \widehat{f}_2(x)dx = 1$. Note that $\widehat{b}_j$ is the kernel density estimator $\widehat{f}_1(x)$ evaluated at the support point $\widetilde{x}_j$, $\widehat{b}_j = \widehat{f}_1(\widetilde{x}_j)$. As a consequence,

$$\widehat{f}_2(x) = \gamma \sum_{j \in J} \widehat{f}_1(\widetilde{x}_j) w_h(x; \widetilde{x}_j)$$

is a weighted average of the kernel density estimator subsampled at the points $\widetilde{x}_j$ with kernel weights

$$w_h(x; \widetilde{x}_j) = \frac{1}{h} K\left(\frac{x - \widetilde{x}_j}{h}\right).$$

This double-smoothing operation should avoid that the density estimator becomes less wiggly. The following result states the consistency of the double kernel estimator.

Lemma 2. *Let $x_1, \ldots, x_n$ be a sample from a distribution with density $f(x)$ and $K(x)$ be a symmetric probability density with bounded support. The double kernel estimator $\widehat{f}_2(x)$ is consistent for estimating the density $f(x)$ as $n \to \infty$.*

Proof. It is easy to see that

$$\left|\widehat{f}_2(x) - f(x)\right| \leq B_1 + B_2,$$

where

$$B_1 = \gamma \sum_{j \in J} \left|\widehat{f}_1(\widetilde{x}_j) - f(\widetilde{x}_j)\right| w_h(x; \widetilde{x}_j)$$

and

$$B_2 = \left|\gamma \sum_{j \in J} f(\widetilde{x}_j) w_h(x; \widetilde{x}_j) - f(x)\right|.$$

Note that $B_1 \leq \gamma \max_{j \in J} |\widehat{f}_1(\widetilde{x}_j) - f(\widetilde{x}_j)| \sum_{j \in J} w_h(x; \widetilde{x}_j) \leq C \max_{j \in J} |\widehat{f}_1(\widetilde{x}_j) - f(\widetilde{x}_j)|$ with some constant C. Therefore, $B_1 \to 0$ since $\widehat{b}_j = \widehat{f}_1(\widetilde{x}_j) \to_P f(\widetilde{x}_j)$ uniformly with appropriate selection of the bandwidth h, see (46).

We will now show that $B_2 \to 0$ that means the convergence of the deterministic approximation. For given x, there is a set J_x such that $w_h(x; \widetilde{x}_j) > 0$ for $j \in J_x$ and $w_h(x; \widetilde{x}_j) = 0$ for $j \notin J_x$ since K has a bounded support. Consequently, we can write $B_2 = \left|\gamma \sum_{j \in J_x} f(\widetilde{x}_j) w_h(x; \widetilde{x}_j) - f(x)\right|$. Then $B_2 \leq B_{21} + B_{22}$, where

$$B_{21} = f(x) \left|\gamma \sum_{j \in J_x} w_h(x; \widetilde{x}_j) - 1\right|$$

and

$$B_{22} = \left|\gamma \sum_{j \in J_x} (f(\widetilde{x}_j) - f(x)) w_h(x; \widetilde{x}_j)\right|.$$

Note that $B_{22} \leq C \max_{j \in J_x} |f(\widetilde{x}_j) - f(x)| \to 0$ since $f(\widetilde{x}_j) \to f(x)$ for $j \in J_x$ as $n \to \infty$. We further obtain that $h\gamma^{-1} = h \sum_{j \in J} f(\widetilde{x}_j) = \sum_{j \in J} h f(jh) \to \int f(x) dx = 1$ and $h \sum_{j \in J} w_h(x; \widetilde{x}_j) = h \sum_{j \in J} w_h(x; jh) \to \int w_h(x; t) dt = 1$. Therefore, we obtain $B_{21} \to 0$ as $n \to \infty$. $\square$

Note that the coefficients $\hat{b}_j$ are positive for the infinite number of indexes if K has a bounded support. Therefore, for convenience of numerical computation we have to introduce the threshold c and define the set J as follows

$$J = \{j \in \mathbb{Z} : \widehat{b}_j > c\}.$$

Thus, we neglect those summands in $\widehat{f}_2(x)$ whose coefficients are small. We determine the threshold c from asymptotic properties of $\widehat{f}_1(\widetilde{x}_j)$. Recall that for the classical kernel density estimator we have $Var(\widehat{b}_j) = Var(\widehat{f}_1(\widetilde{x}_j)) = \frac{f(\widetilde{x}_j)}{nh} \int K^2(x)dx + o_P(\frac{1}{nh})$, see (36, p. 130). Thus, we can define the threshold as

$$c = \alpha \sqrt{\frac{\max_j f(\widetilde{x}_j) \int K^2(x)dx}{nh}}.$$

where the parameter $\alpha \in (0, 1)$ controls the degree of truncation. We set $\alpha = 0.2$ that leads to a small degree of truncation.

To improve the flexibility of the double kernel density estimator, we can use the adaptive bandwidth in the following form

$$\hat{f}_2(x) = \gamma \sum_{j \in J} \frac{\widehat{b}_j}{h_j} K\left(\frac{x - \widetilde{x}_j}{h_j}\right)$$

where

$$h_j = (\beta \sqrt{c/\widehat{b}_j} + 0.5)h.$$

In the following section we investigate the accuracy of this estimator in dependence on the choice of β and h by examples.

In Table3.3.12, we present simulated root mean squared deviations (RMSD) of the sampling plan size using the double kernel estimator with different bandwidth selection for several models of the distribution function of measurements. We can see in Table 3.3.12 that the RMSD depends on β non-homogeneously and the best choice is $\beta = 1$.

Tabelle 3.3.12: RMSD of the distribution of the sampling plan size N using the double kernel estimator and adaptive ICV-based bandwidth for models 1-9, for different β.

model	n	$\beta = 0$	$\beta = 0.5$	$\beta = 1$	$\beta = 1.5$	$\beta = 2$
1	100	20.0	17.7	16.9	19.0	22.6
1	250	16.6	15.6	14.6	14.7	16.2
1	500	13.8	13.3	12.8	12.4	12.8
2	100	90.7	86.0	79.6	72.8	67.3
2	250	60.4	58.5	56.3	53.6	50.6
2	500	44.4	43.4	42.4	41.3	39.9
3	100	67.9	68.8	75.5	86.2	97.7
3	250	54.8	54.1	54.7	58.7	65.7
3	500	44.9	44.1	43.5	44.1	46.9
4	100	71.9	70.0	69.1	70.6	74.7
4	250	58.6	56.6	54.3	51.7	49.9
4	500	49.0	47.8	46.5	45.1	43.4
5	100	245.5	237.0	229.1	227.6	235.4
5	250	206.9	200.7	194.4	187.5	182.2
5	500	174.7	170.4	166.7	162.6	157.9
6	100	125.0	128.4	135.8	147.5	160.9
6	250	93.4	93.6	94.5	97.7	104.2
6	500	78.3	78.1	78.1	78.7	80.7
7	100	459.6	473.0	497.6	535.5	580.2
7	250	353.4	356.2	361.2	372.8	393.7
7	500	294.0	294.0	295.3	298.0	304.8
8	100	24.6	21.9	18.0	14.3	11.7
8	250	18.1	17.2	15.9	14.2	12.3
8	500	14.2	13.7	13.2	12.4	11.5
9	100	78.5	84.5	92.8	101.2	108.5
9	250	63.0	66.5	71.8	78.4	85.3
9	500	46.5	48.4	50.9	54.7	59.5

3.3.6 Simulation study for the double kernel estimator

Let us now study characteristics of the sampling plan obtained by use of the double kernel density estimator with different bandwidths. We show results for models 1–9 in Tables 3.3.13–3.3.16. We can see that the ICV bandwidth is typically better than other bandwidths. The BCV and GPS bandwidths are slightly worse than the ICV bandwidth.

Tabelle 3.3.13: Characteristics of distributions of n_m and c_m using the double kernel estimator with threshold $c = 0.2\sqrt{\frac{\max_j f(\widetilde{x}_j)\int K^2(x)dx}{nh}}$ and adaptive LSCV-based bandwidth for models 1-9.

model	m	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
1	100	36	44	52	63	82	58.7	31.1	-6.3	31.7	15.4	2.4
1	250	43	50	59	68	82	61.4	18.9	-3.6	19.3	15.4	1.8
1	500	47	53	61	70	80	63.0	15.7	-2.0	15.8	15.3	1.5
2	100	17	47	94	160	241	118.2	104.4	15.2	105.5	28.3	11.9
2	250	43	66	99	143	193	111.0	66.6	8.0	67.1	29.7	7.7
2	500	56	74	99	128	163	105.8	45.2	2.8	45.3	29.9	5.4
3	100	100	124	158	210	270	179.5	101.0	-29.5	105.2	18.1	3.0
3	250	123	148	181	221	265	191.0	65.0	-18.0	67.4	18.3	2.3
3	500	142	165	193	223	260	198.7	50.8	-10.3	51.8	18.4	1.8
4	100	78	109	152	214	284	174.3	100.4	6.3	100.6	24.2	5.8
4	250	95	125	163	208	255	171.9	67.2	3.9	67.3	24.4	4.3
4	500	108	132	163	201	241	170.8	53.0	2.8	53.1	24.5	3.4
5	100	274	390	551	776	1026	621.2	342.6	13.2	342.8	42.7	10.2
5	250	337	441	578	744	914	610.5	237.3	2.5	237.2	42.9	7.7
5	500	392	473	586	721	864	610.7	188.6	2.7	188.6	43.3	6.1
6	100	138	184	244	328	430	273.8	141.8	-50.2	150.4	29.4	6.2
6	250	177	220	275	340	406	287.5	99.2	-36.5	105.7	30.4	4.6
6	500	204	239	287	340	397	295.6	79.9	-28.4	84.8	31.0	3.7
7	100	499	672	894	1203	1583	990.9	476.6	-214.1	522.4	52.9	10.7
7	250	649	808	1007	1251	1500	1051.0	348.5	-154.0	380.9	55.1	8.2
7	500	750	880	1058	1251	1465	1086.1	289.2	-118.9	312.6	56.3	6.7
8	100	23	32	44	59	76	48.3	26.2	12.3	29.0	13.6	2.8
8	250	24	32	43	54	68	44.8	18.3	8.8	20.3	13.0	2.3
8	500	27	32	41	50	61	42.5	14.3	6.5	15.7	12.7	1.8
9	100	54	65	91	129	185	108.7	66.7	-53.3	85.3	19.6	4.1
9	250	82	97	119	148	181	127.7	45.2	-34.3	56.8	20.9	3.0
9	500	97	112	132	155	181	136.6	36.0	-25.4	44.0	21.6	2.4

Tabelle 3.3.14: Characteristics of distributions of n_m and c_m using the double kernel estimator with threshold $c = 0.2\sqrt{\frac{\max_j f(\widetilde{x}_j)\int K^2(x)dx}{nh}}$ and adaptive BCV-based bandwidth for models 1-9.

model	m	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
1	100	37	46	55	65	75	55.7	14.6	-9.3	17.3	15.2	1.7
1	250	43	50	59	69	78	60.1	13.8	-4.9	14.6	15.3	1.5
1	500	47	53	61	70	78	62.0	12.3	-3.0	12.7	15.3	1.3
2	100	18	46	90	145	214	105.5	81.5	2.5	81.5	27.2	10.3
2	250	43	65	97	137	181	106.2	56.9	3.2	57.0	29.3	7.0
2	500	57	74	98	125	160	103.9	42.9	0.9	42.9	29.7	5.2
3	100	97	120	150	185	219	155.3	49.6	-53.7	73.1	17.6	1.8
3	250	124	147	177	210	240	180.2	46.2	-28.8	54.5	18.1	1.8
3	500	143	163	191	219	250	193.4	41.1	-15.6	44.0	18.3	1.6
4	100	63	82	127	175	225	136.1	65.8	-31.9	73.1	22.0	4.5
4	250	97	123	159	199	236	163.7	55.9	-4.3	56.0	23.9	3.7
4	500	109	132	162	195	229	166.2	47.6	-1.8	47.7	24.2	3.1
5	100	259	363	503	661	828	527.1	228.4	-80.9	242.3	40.0	7.8
5	250	341	436	562	707	841	582.5	200.1	-25.5	201.7	42.2	6.6
5	500	392	469	577	698	825	595.4	170.6	-12.6	171.0	42.9	5.6
6	100	123	166	219	278	340	226.7	86.6	-97.3	130.2	27.4	4.4
6	250	177	214	263	320	369	270.9	77.3	-53.1	93.8	29.7	3.7
6	500	204	236	280	330	376	286.3	69.0	-37.7	78.7	30.6	3.3
7	100	467	607	791	1005	1218	818.1	305.3	-386.9	492.8	48.8	8.0
7	250	646	778	957	1168	1348	987.3	283.2	-217.7	357.2	53.7	6.7
7	500	749	872	1027	1207	1384	1050.1	251.6	-154.9	295.4	55.6	5.9
8	100	23	32	42	53	64	43.2	15.9	7.2	17.5	13.3	2.1
8	250	25	32	41	52	63	42.8	14.8	6.8	16.3	12.9	2.0
8	500	27	33	40	48	58	41.4	12.2	5.4	13.3	12.6	1.6
9	100	53	60	69	78	86	69.5	13.5	-92.5	93.5	17.1	1.3
9	250	71	79	89	101	122	93.7	23.8	-68.3	72.3	18.9	1.8
9	500	90	101	117	137	158	121.3	26.7	-40.7	48.7	20.7	1.9

Tabelle 3.3.15: Characteristics of distributions of n_m and c_m using the double kernel estimator with threshold $c = 0.2\sqrt{\frac{\max_j f(\widetilde{x}_j)\int K^2(x)dx}{nh}}$ and adaptive ICV-based bandwidth for models 1-9.

model	m	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
1	100	37	46	56	66	75	56.3	14.7	-8.7	17.1	15.3	1.7
1	250	43	50	60	70	79	60.5	14.1	-4.5	14.8	15.3	1.5
1	500	47	53	61	70	78	62.1	12.5	-2.9	12.8	15.3	1.3
2	100	18	45	87	142	207	103.0	79.4	0.0	79.4	26.9	10.1
2	250	43	65	97	135	178	105.1	55.9	2.1	56.0	29.1	6.9
2	500	57	74	97	125	159	103.3	42.2	0.3	42.2	29.7	5.1
3	100	95	116	143	179	213	150.7	48.3	-58.3	75.7	17.5	1.7
3	250	123	146	174	205	237	177.8	45.6	-31.2	55.2	18.0	1.7
3	500	141	164	189	217	246	191.8	40.2	-17.2	43.7	18.3	1.5
4	100	66	87	125	170	213	134.2	60.5	-33.8	69.3	22.0	4.2
4	250	97	121	155	192	228	160.5	53.7	-7.5	54.2	23.8	3.6
4	500	109	131	160	193	226	164.7	46.3	-3.3	46.4	24.2	3.1
5	100	271	366	491	646	794	517.5	211.4	-90.5	229.9	39.8	7.1
5	250	342	431	556	689	818	571.9	191.3	-36.1	194.6	41.9	6.4
5	500	390	466	574	689	809	589.3	165.5	-18.7	166.5	42.7	5.4
6	100	120	158	207	263	322	215.4	81.7	-108.6	135.9	26.8	4.2
6	250	174	211	258	312	362	265.8	75.0	-58.2	94.9	29.5	3.6
6	500	203	234	277	325	372	283.1	66.9	-40.9	78.4	30.5	3.2
7	100	472	596	762	956	1166	797.0	284.2	-408.0	497.2	48.4	7.0
7	250	638	764	943	1131	1307	964.5	271.0	-240.5	362.2	53.1	6.5
7	500	747	861	1014	1191	1356	1036.5	243.6	-168.5	296.1	55.2	5.8
8	100	24	32	42	53	65	43.4	16.0	7.4	17.6	13.4	2.1
8	250	25	32	41	51	62	42.4	14.3	6.4	15.7	12.9	1.9
8	500	27	33	40	48	56	41.1	11.9	5.1	12.9	12.6	1.6
9	100	54	62	70	79	87	70.6	14.1	-91.4	92.5	17.2	1.3
9	250	72	81	90	102	118	93.8	21.9	-68.2	71.7	18.9	1.7
9	500	89	99	114	132	151	117.6	25.1	-44.4	51.0	20.5	1.8

Tabelle 3.3.16: Characteristics of distributions of n_m and c_m using the double kernel estimator with threshold $c = 0.2\sqrt{\frac{\max_j f(\widetilde{x}_j)\int K^2(x)dx}{nh}}$ and adaptive GPS-based bandwidth for models 1-9.

model	m	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
1	100	37	45	54	64	74	55.3	16.0	-9.7	18.7	15.2	1.7
1	250	43	50	58	68	76	59.3	13.9	-5.7	15.0	15.3	1.5
1	500	47	53	60	68	77	61.2	12.2	-3.8	12.7	15.3	1.3
2	100	18	49	99	168	254	122.3	102.4	19.3	104.2	29.0	11.8
2	250	43	65	99	141	189	109.3	60.5	6.3	60.8	29.6	7.4
2	500	56	74	98	126	161	104.3	43.3	1.3	43.3	29.8	5.2
3	100	105	137	176	227	279	186.7	70.9	-22.3	74.3	18.3	2.5
3	250	126	153	186	226	263	191.7	55.1	-17.3	57.8	18.3	2.1
3	500	144	167	194	225	257	198.0	44.3	-11.0	45.7	18.4	1.7
4	100	75	102	141	191	245	152.9	69.9	-15.1	71.5	23.1	4.6
4	250	96	121	156	195	235	162.1	55.6	-5.9	55.9	23.8	3.7
4	500	108	130	159	189	223	162.8	44.7	-5.2	45.0	24.0	3.0
5	100	271	373	505	687	867	545.6	234.6	-62.4	242.7	40.6	7.7
5	250	341	424	538	667	795	556.9	182.3	-51.1	189.3	41.4	6.1
5	500	390	457	553	658	767	566.7	149.2	-41.3	154.8	42.0	4.9
6	100	134	177	241	319	397	255.8	106.2	-68.2	126.2	28.8	5.1
6	250	179	220	271	336	397	282.6	87.3	-41.4	96.6	30.2	4.1
6	500	206	239	286	336	391	292.1	72.9	-31.9	79.6	30.9	3.4
7	100	503	668	881	1156	1457	938.8	375.5	-266.2	460.2	51.9	8.8
7	250	645	782	975	1192	1397	1004.6	303.3	-200.4	363.5	54.0	7.1
7	500	751	864	1018	1191	1372	1042.8	248.8	-162.2	297.0	55.4	5.8
8	100	23	32	43	56	70	45.4	19.2	9.4	21.4	13.4	2.4
8	250	25	32	41	52	64	43.1	15.5	7.1	17.0	12.9	2.0
8	500	27	33	40	48	57	41.2	12.2	5.2	13.2	12.6	1.6
9	100	52	59	68	81	102	73.7	24.0	-88.3	91.5	17.4	1.9
9	250	70	78	89	104	122	93.4	22.5	-68.6	72.2	18.8	1.7
9	500	86	95	106	122	137	109.6	21.2	-52.4	56.5	20.0	1.6

3.3.7 Summary

In Table 3.3.17, we present simulated root mean squared deviations (RMSD) of the sampling plan size using the classical kernel estimator and the double kernel estimator with different bandwidth selection for several models of the distribution of measurements. We can observe in Table 3.3.17 that the double kernel estimator is typically better than the classical kernel estimator with corresponding bandwidth. We also see that the BCV, ICV and GPS bandwidths are more preferable for use in practice since they typically provide smaller values of RMSD.

Tabelle 3.3.17: RMSD of the distribution of the sampling plan size using the classical kernel estimator with different bandwidths and the double kernel estimator with adaptive LSCV-, BCV-, ICV-, GPS-based bandwidths for models 1-9.

		classical kernel estimator					double kernel estimator				
model	m	LSCV	BCV	SJPI	ICV	GPS	LSCV	BCV	SJPI	ICV	GPS
1	100	35.0	21.8	26.8	22.6	22.3	31.7	17.3	19.8	17.1	18.7
1	250	21.3	17.6	19.3	17.8	17.5	19.3	14.6	16.1	14.8	15.0
1	500	16.4	14.3	15.3	14.6	14.1	15.8	12.7	13.6	12.8	12.7
2	100	122.8	96.9	115.6	93.5	121.8	105.5	81.5	99.5	79.4	104.2
2	250	69.9	61.5	65.9	59.7	65.7	67.1	57.0	62.9	56.0	60.8
2	500	47.5	44.9	46.3	44.2	45.5	45.3	42.9	44.9	42.2	43.3
3	100	118.2	75.4	87.2	74.4	94.0	105.2	73.1	70.5	75.7	74.3
3	250	72.3	60.2	65.1	60.4	66.3	67.4	54.5	56.4	55.2	57.8
3	500	54.9	48.7	51.4	48.3	50.7	51.8	44.0	46.0	43.7	45.7
4	100	121.2	82.0	85.1	76.7	88.3	100.6	73.1	67.6	69.3	71.5
4	250	75.1	63.9	65.2	61.3	63.4	67.3	56.0	57.7	54.2	55.9
4	500	56.9	51.5	51.9	50.4	49.1	53.1	47.7	48.2	46.4	45.0
5	100	411.6	288.5	284.7	269.2	297.1	342.8	242.3	240.6	229.9	242.7
5	250	266.3	230.8	232.4	221.4	213.9	237.2	201.7	204.1	194.6	189.3
5	500	203.9	185.9	186.5	181.4	168.0	188.6	171.0	172.0	166.5	154.8
6	100	170.7	128.5	121.9	128.8	140.3	150.4	130.2	118.9	135.9	126.2
6	250	115.9	99.6	99.5	99.0	106.1	105.7	93.8	91.6	94.9	96.6
6	500	91.0	82.8	83.0	82.5	85.2	84.8	78.7	77.5	78.4	79.6
7	100	586.1	470.0	470.0	464.3	500.5	522.4	492.8	497.6	497.2	460.2
7	250	416.7	368.4	369.3	367.7	380.3	380.9	357.2	353.5	362.2	363.5
7	500	333.2	306.3	306.1	305.6	307.6	312.6	295.4	290.2	296.1	297.0
8	100	30.8	20.8	27.9	20.8	24.3	29.0	17.5	26.3	17.6	21.4
8	250	18.6	15.6	17.5	15.0	16.0	20.3	16.3	19.6	15.7	17.0
8	500	13.9	12.4	13.5	12.2	12.2	15.7	13.3	15.0	12.9	13.2
9	100	85.8	78.1	63.7	76.7	77.3	85.3	93.5	73.2	92.5	91.5
9	250	56.2	61.9	49.2	60.9	61.4	56.8	72.3	53.4	71.7	72.2
9	500	43.7	44.0	40.5	45.6	48.8	44.0	48.7	42.3	51.0	56.5

Kapitel 3.4

Estimators based on series expansion

This chapter deals with approaches to quantile estimation which are based on orthogonal series estimators. These approaches use the fact that any sufficiently smooth function can be approximated by the truncated series expansion. Thus, the orthogonal series estimator is smoother than an empirical function.

Indeed, consider a function $g(x)$ defined on an interval. Without loss of generality, we assume that this interval is $[0, 1]$. Consider also an orthogonal system $\{\phi_0(x), \phi_1(x), \ldots\}$. For example, the trigonometric orthogonal system is constituted by $\phi_0(x) \equiv 1$ and $\phi_j(x) = \sqrt{2/\pi}\cos(\pi j)$, $j = 1, 2, \ldots$. Then we can write the expansion

$$g(x) = \sum_{j=0}^{\infty} \theta_j \phi_j(x),$$

where

$$\theta_j = \int_0^1 g(x)\phi_j(x)dx$$

is called the j-th coefficient. We note that θ_j tends to zero as j increases if $g(x)$ is sufficiently smooth. Therefore, we define the function

$$g_J(x) = \sum_{j=0}^{J} \theta_j \phi_j(x),$$

where J is the cutoff parameter, which is used to approximate $g(x)$.

We now note that there are two ways for estimating quantiles using series expansions. One way is to estimate the density function using the orthogonal series estimator and then compute quantiles using the estimated density, which is studied in Section 3.4.1.

Another way is to construct the orthogonal series estimator of the quantile function, which is discussed in Section 3.4.2.

3.4.1 Orthogonal series estimator of the density function

Let $(x_1, \ldots, x_m)$ be a sample from a distribution with density $p(x)$. Without loss of generality, we assume that the density is defined on the interval $[0, 1]$, otherwise we can apply the linear transformation from the interval $[\min_i x_i - 0.5s_x, \max_i x_i + 0.5s_x]$ to the interval $[0, 1]$, where s_x is the standard deviation of the sample.

The orthogonal series estimator of the density function is given by $\hat{p}_J(x) = (\bar{p}_J(x) - c)_+$, where

$$\bar{p}_J(x) = \sum_{j=0}^{J} \hat{\theta}_j \phi_j(x),$$

the coefficient $\hat{\theta}_j$ is computed as

$$\hat{\theta}_j = \frac{1}{m} \sum_{i=1}^{m} \phi_j(x_i),$$

the symbol $(w(x))_+$ means $\max\{0, w(x)\}$ and c is such that $\int \hat{p}_J(x)dx = 1$. Then the estimator of the quantile x_γ is defined by

$$\gamma = \int_0^{\hat{x}_\gamma} \hat{p}_J(x)dx.$$

In Figure 3.4.1 we depict the orthogonal series estimator of the density function for models 1 and 2 for different values of m and J. We can see that for model 1 the estimator $J = 16$ is oscillating in a larger degree compared to the estimator with $J = 8$. Meanwhile, for model 2 the estimator with $J = 8$ is too smooth and not accurate enough compared to the estimator with $J = 16$.

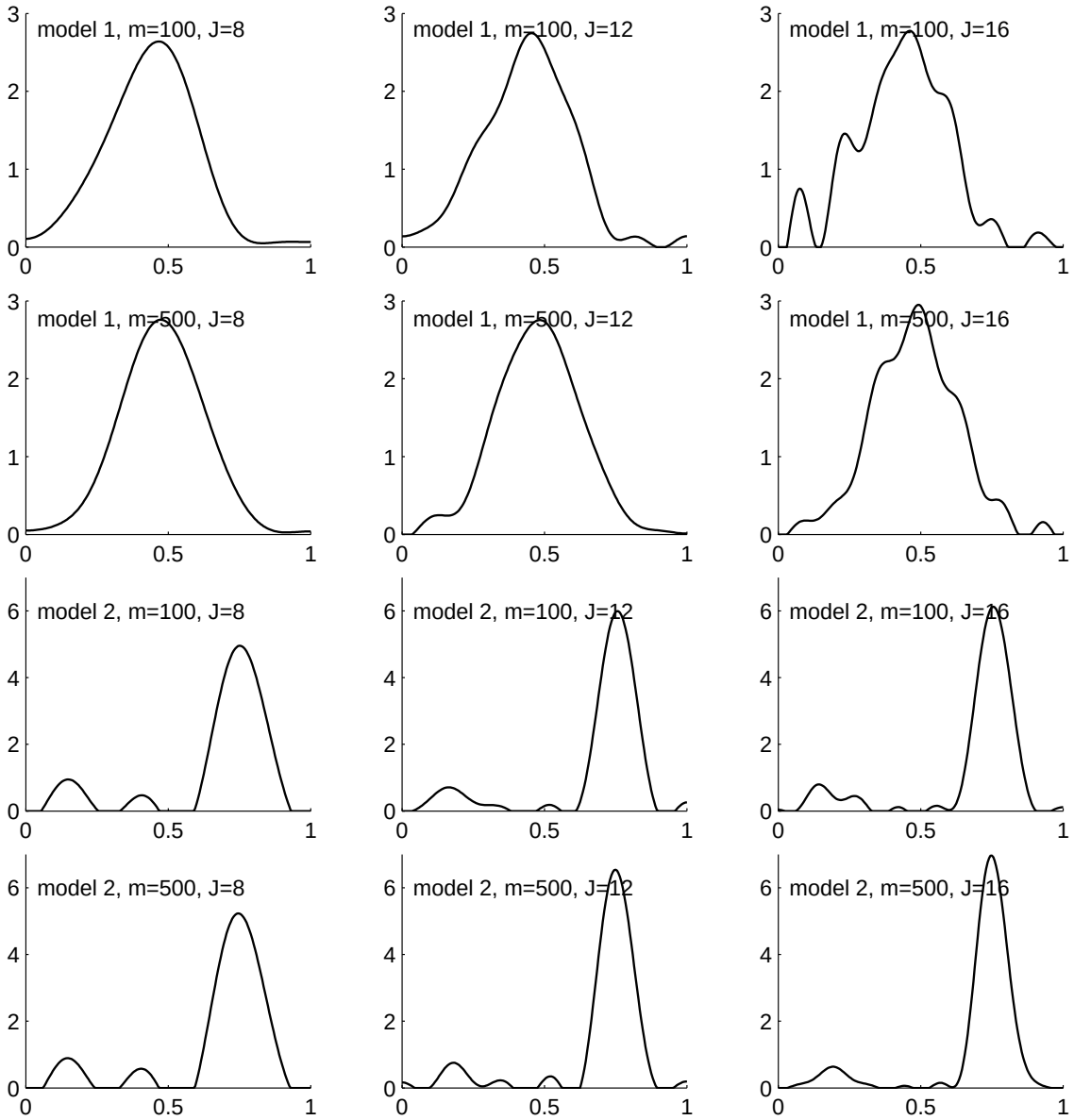


Abbildung 3.4.1: The orthogonal series estimator of the density function with different J for samples of size 100 and 500 given by models 1 and 2.

3.4.2 Orthogonal series estimator of the quantile function

Let $(x_1, \ldots, x_m)$ be a sample from a distribution with distribution function $F(x)$. The empirical quantile function $Q_m(x)$ is the inverse of the empirical distribution function

$$F_m(x) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{[x_i,\infty)}(x),$$

that is $Q_m(x) = F_m^{-1}(x)$. Then the orthogonal series estimator of the quantile function is

$$Q_J(x) = \sum_{j=0}^{J} \hat{c}_j \phi_j(x)$$

where

$$\hat{c}_j = \frac{1}{m} \sum_{i=1}^{m} x_{(i)} \phi_j(i/m)$$

where $x_{(1)}, \ldots, x_{(m)}$ are the sample order statistics. The formula for $\hat{c}_j$ follows from the fact that

$$\begin{aligned}
\hat{c}_j &= \int_0^1 Q_m(x)\phi_j(x)dx \\
&= \int_{-\infty}^{\infty} Q_m(F_m(z))\phi_j(F_m(z))dF_m(z) \\
&= \int_{-\infty}^{\infty} z\phi_j(F_m(z))dF_m(z) \\
&= \frac{1}{m} \sum_{i=1}^{m} x_i \phi_j(F_m(x_i)) \\
&= \frac{1}{m} \sum_{i=1}^{m} x_{(i)} \phi_j(i/m).
\end{aligned}$$

In Figure 3.4.2 we depict the empirical quantile function and the orthogonal series estimator for models 1 and 2 for different values of m and J. We can see that the estimator $Q_J(x)$ is not necessarily monotonically increased. Moreover, the estimator behaves poorly near the bounds since the true quantile function is infinite at the bounds but the derivative of $Q_J(x)$ is zero at the bounds. We can also see the Gibbs effect (i.e. oscillation) of the estimator for model 2. Therefore, we discard the orthogonal series estimator of the quantile function from further consideration.

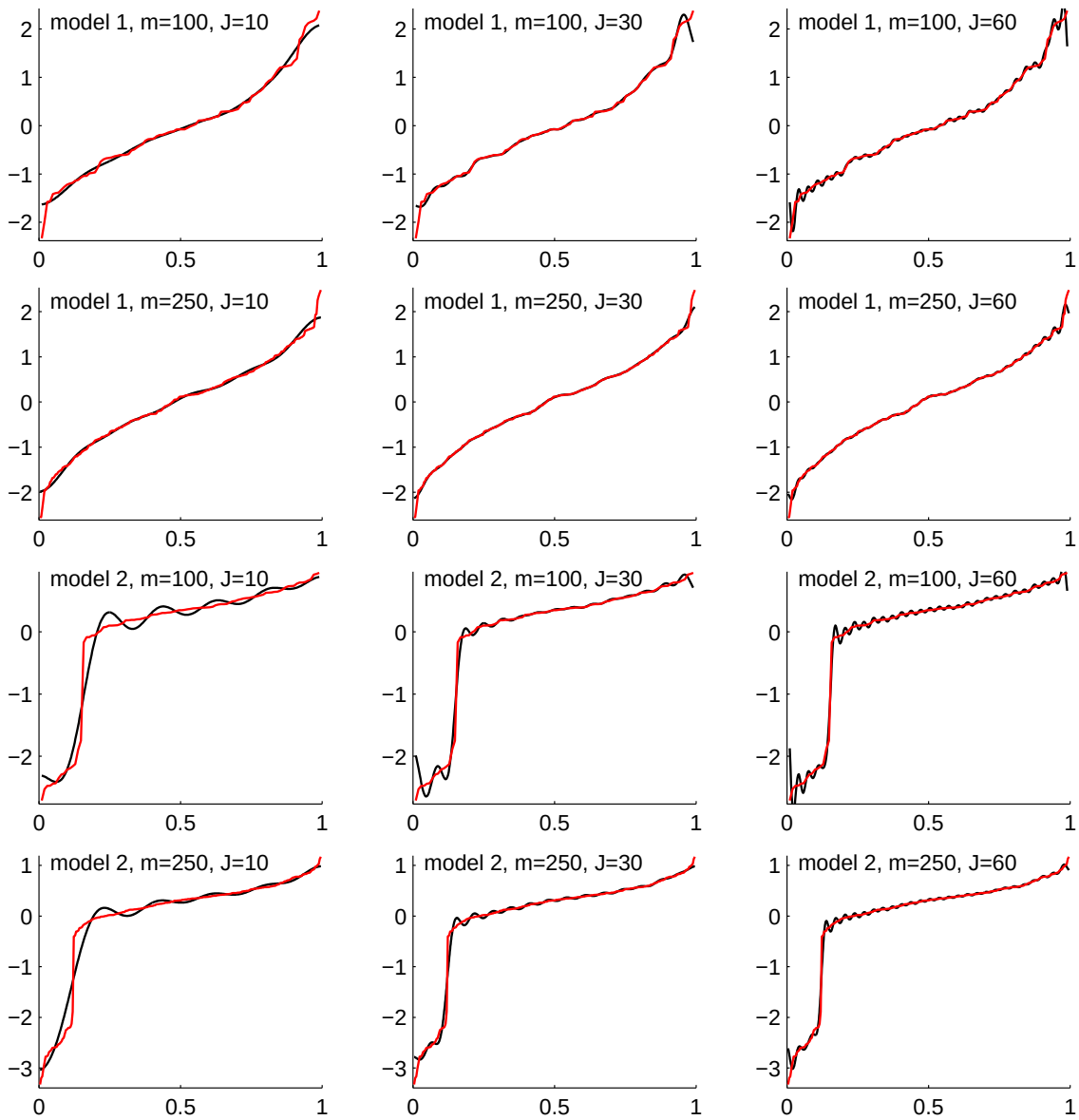


Abbildung 3.4.2: The empirical quantile function (red) and the orthogonal series estimator (black) with different J for samples of size 100 and 250 given by models 1 and 2.

3.4.3 Influence of the cutoff parameter

In Table 3.4.1 we show the influence of the cutoff J on the characteristics of sampling plans. We can see that for model 1, the mean $\mathrm{E}\, n_m$ is more close to the true value (that is 65) for $J \in \{8, \ldots, 12\}$. Meanwhile, for model 2, the mean $\mathrm{E}\, n_m$ is more close to the true value (that is 103) for $J = 14, 15$. We can also that the standard deviation of n_m is larger

using the orthogonal series estimator compared to using the kernel density estimator.

Tabelle 3.4.1: Characteristics of distributions of n_m and c_m using the orthogonal series estimator of the density function with the cutoff J for models 1 and 2.

m	J	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
model 1										
500	6	49	60	72	86	97.0	72.8	18.6	15.4	1.7
500	8	42	54	68	83	99.0	69.3	22.2	15.2	2.2
500	9	41	52	68	85	102.0	69.9	24.0	15.2	2.4
500	10	40	52	67	84	103.0	69.7	24.5	15.2	2.5
500	11	39	51	67	87	105.0	70.3	26.2	15.2	2.6
500	12	39	51	67	86	106.5	70.3	26.8	15.2	2.7
500	14	38	51	67	88	109.0	71.1	28.5	15.3	2.8
model 2										
500	12	51	74	109	150	192.0	115.8	55.5	30.9	7.1
500	13	58	78	107	142	175.0	113.2	46.8	30.9	5.5
500	14	50	68	93	125	159.0	100.2	43.9	29.1	5.5
500	15	56	75	100	133	166.0	106.9	44.7	30.2	5.4
500	16	59	79	107	141	180.0	114.2	47.9	31.2	5.7
500	17	55	76	105	139	177.0	111.3	48.5	30.7	5.8
500	18	56	75	103	139	179.0	111.7	50.4	30.7	6.0

3.4.4 Method of cutoff selection

Kronmal and Tarter (1968) have considered the mean integrated squared error

$$\Psi(J) = \int \mathbf{E}(\hat{p}_J(x) - p(x))^2 dx = \frac{1}{m} \sum_{j=0}^{J} (d_j^2 - c_j^2) + \sum_{j=J+1}^{\infty} c_j^2,$$

where $d_j^2 = \mathbf{E}_p \phi_j^2(\xi)$. Therefore, the optimal value of J is such that minimizes $\Psi(J)$. Unfortunately, the criterion $\Psi(J)$ cannot be computed since the number of summands is infinite and the estimators $\hat{c}_j = \sum_{i=1}^m \phi_j(x_i)/m$ and $\hat{d}_j^2 = \sum_{i=1}^m \phi_j^2(x_i)/m$ of c_j^2 and d_j^2 is not accurate for large j. Consequently, adaptive rules for cutoff selection are needed.

The Kronmal-Tarter method (with parameter t, KT(t)) defines J as a minimal positive integer such that

$$\hat{c}_j^2 < \frac{2}{m+1}\hat{d}_j^2 \tag{3.4.4.1}$$

for $j = J+1, \ldots, J+t$.

The modified Kronmal-Tarter (mKT) method determines the orthogonal series estimator of the density function by including the jth summand $\hat{c}_j\phi_j(x)$ if and only if inequality (3.4.4.1) is satisfied. This method potentially involves examining the infinite number of inequalities.

Diggle and Hall (1986) have proposed to do not handle inequalities separately and employed an asymptotic formula for the infinite number of summands. Specifically, in the DH method the criterion $\Psi(J)$ is approximated by

$$\Phi(J) = \frac{J}{m\pi} + \frac{1}{1-1/m}\sum_{j=J}^{3J^{3/2}}(c_j^2 - d_j^2/m)_+,$$

whose minimization yields the optimal cutoff.

3.4.5 Simulation study

In Tables 3.4.2–3.4.10 we show the characteristics of sampling plans using different methods of optimal cutoff selection. We can see that the standard deviation of n_m using the mKT method is not small. Moreover, the standard deviation of n_m using the KT(t) and DH methods is larger compared to using the kernel density approach. Also, the performance of all orthogonal series estimators is not good for the sample size $m = 100$.

Tabelle 3.4.2: Characteristics of distributions of n_m and c_m using the orthogonal series estimator of the density function with optimal cutoff selection for model 1.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	KT(1)	23	47	83	126	165	89.4	53.4	24.4	58.7	16.2	4.6
100	KT(2)	25	50	87	128	165	92.7	55.1	27.7	61.7	16.6	4.8
100	KT(3)	24	49	87	130	170	94.5	60.0	29.5	66.9	16.7	5.1
100	KT(4)	22	47	87	133	176	97.3	68.4	32.3	75.6	16.9	5.6
100	mKT	14	28	60	128	252	103.3	125.4	38.3	131.1	16.9	8.6
100	DH	26	61	106	141	167	102.6	54.7	37.6	66.4	17.5	4.7
250	KT(1)	37	53	73	94	115	75.8	32.3	10.8	34.1	15.5	3.0
250	KT(2)	38	53	74	96	118	76.9	32.3	11.9	34.4	15.7	3.0
250	KT(3)	36	52	74	97	121	77.2	34.7	12.2	36.8	15.7	3.2
250	KT(4)	35	50	73	97	125	77.2	37.1	12.2	39.1	15.7	3.4
250	mKT	19	30	49	81	127	64.9	56.0	-0.1	56.0	14.4	5.0
250	DH	38	53	73	102	134	79.8	36.2	14.8	39.1	15.9	3.3
500	KT(1)	46	59	73	88	98	73.1	20.7	8.1	22.2	15.4	1.9
500	KT(2)	45	56	71	87	100	72.1	21.5	7.1	22.6	15.4	2.0
500	KT(3)	44	56	71	87	101	72.3	23.0	7.3	24.1	15.4	2.2
500	KT(4)	42	54	70	88	102	72.1	24.9	7.1	25.9	15.4	2.4
500	mKT	27	38	55	77	104	61.8	33.8	-3.2	33.9	14.3	3.4
500	DH	45	58	72	85	98	72.1	21.3	7.1	22.4	15.3	2.0
5000	KT(1)	54	60	66	76	91	69.1	13.2	4.1	13.8	15.2	1.2
5000	KT(2)	56	60	64	70	75	64.8	7.4	-0.2	7.4	14.9	0.7
5000	KT(3)	56	60	64	70	75	64.9	7.5	-0.1	7.5	14.9	0.8
5000	KT(4)	55	59	64	70	75	64.8	7.7	-0.2	7.7	14.9	0.8
5000	mKT	50	55	61	68	76	62.0	10.2	-3.0	10.7	14.6	1.1
5000	DH	53	58	63	69	73	63.4	7.7	-1.6	7.8	14.7	0.8

Compare tables 3.3.1, 3.4.2, 3.5.1, 3.5.19, 3.6.4.

Tabelle 3.4.3: Characteristics of distributions of n_m and c_m using the orthogonal series estimator of the density function with optimal cutoff selection for model 2.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	KT(1)	3	24	76	142	222	97.4	92.6	-5.6	92.7	23.6	13.3
100	KT(2)	15	50	106	188	282	134.0	114.9	31.0	119.0	30.1	12.7
100	KT(3)	16	49	108	196	310	143.1	134.7	40.1	140.6	30.9	13.8
100	KT(4)	17	49	109	199	321	148.6	152.0	45.6	158.7	31.3	14.4
100	mKT	14	34	85	207	404	161.1	209.5	58.1	217.3	31.3	18.1
100	DH	9	49	101	170	252	120.9	98.3	17.9	99.9	28.3	12.1
250	KT(1)	14	45	78	123	175	89.0	61.8	-14.0	63.4	24.9	10.0
250	KT(2)	41	66	103	153	207	116.2	68.4	13.2	69.7	30.4	8.1
250	KT(3)	41	66	103	157	217	119.0	73.8	16.0	75.5	30.8	8.6
250	KT(4)	41	65	103	158	218	120.7	77.6	17.7	79.6	30.9	8.9
250	mKT	31	50	88	149	235	116.8	102.0	13.8	102.9	29.8	11.1
250	DH	41	65	98	142	191	109.0	60.6	6.0	60.9	29.5	7.4
500	KT(1)	33	57	84	117	154	89.7	48.6	-13.3	50.4	26.0	8.2
500	KT(2)	54	75	102	137	173	109.4	48.1	6.4	48.5	30.4	5.8
500	KT(3)	54	74	102	138	176	110.5	50.7	7.5	51.3	30.5	6.1
500	KT(4)	54	74	102	139	178	111.5	53.1	8.5	53.8	30.6	6.3
500	mKT	44	63	91	132	179	105.2	61.7	2.2	61.7	29.5	7.5
500	DH	55	74	100	131	163	106.0	44.3	3.0	44.4	29.9	5.4
5000	KT(1)	77	90	101	112	123	100.3	17.7	-2.7	17.9	29.4	3.4
5000	KT(2)	86	92	101	113	121	102.7	14.6	-0.3	14.6	30.3	1.9
5000	KT(3)	85	92	101	113	121	102.6	14.7	-0.4	14.7	30.3	1.9
5000	KT(4)	85	92	101	113	121	102.6	14.9	-0.4	14.9	30.3	2.0
5000	mKT	82	91	101	112	122	101.6	16.1	-1.4	16.1	30.2	2.1
5000	DH	87	93	102	112	121	102.7	14.0	-0.3	14.0	30.3	1.8

Compare tables 3.3.2, 3.4.3, 3.5.2, 3.5.20, 3.6.5.

Tabelle 3.4.4: Characteristics of distributions of n_m and c_m using the orthogonal series estimator of the density function with optimal cutoff selection for model 3.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	KT(1)	142	244	365	467	554	356.1	159.6	147.1	217.0	23.4	5.6
100	KT(2)	105	173	266	369	477	281.3	145.4	72.3	162.3	20.3	4.9
100	KT(3)	98	167	264	365	472	277.9	149.1	68.9	164.3	20.2	5.0
100	KT(4)	92	163	263	371	480	280.9	160.2	71.9	175.5	20.2	5.3
100	mKT	52	103	228	409	636	298.6	266.0	89.6	280.7	20.1	8.2
100	DH	91	183	318	445	546	319.6	174.7	110.6	206.8	21.5	6.3
250	KT(1)	169	227	300	385	444	305.5	108.6	96.5	145.3	22.1	4.4
250	KT(2)	121	179	246	307	360	244.7	93.4	35.7	100.0	19.2	3.3
250	KT(3)	119	174	241	305	362	243.1	96.6	34.1	102.5	19.1	3.4
250	KT(4)	115	170	238	304	368	242.5	100.9	33.5	106.3	19.1	3.6
250	mKT	79	121	189	295	419	226.9	150.5	17.9	151.5	18.2	5.4
250	DH	136	195	255	316	381	257.3	97.6	48.3	108.9	19.5	3.5
500	KT(1)	169	229	279	338	390	280.0	85.1	71.0	110.8	21.2	4.0
500	KT(2)	133	175	229	285	326	230.1	74.5	21.1	77.4	18.7	2.7
500	KT(3)	135	175	226	283	326	229.7	74.5	20.7	77.3	18.8	2.7
500	KT(4)	134	174	225	283	328	229.8	76.1	20.8	78.9	18.8	2.8
500	mKT	105	142	198	266	342	213.5	98.8	4.5	98.9	18.1	3.7
500	DH	115	191	256	297	333	241.6	81.0	32.6	87.3	19.0	3.0
5000	KT(1)	166	185	208	239	282	215.5	44.8	6.5	45.2	18.8	2.8
5000	KT(2)	176	191	205	222	238	206.9	24.4	-2.1	24.5	18.1	0.9
5000	KT(3)	179	192	207	224	241	208.5	24.9	-0.5	24.9	18.1	0.9
5000	KT(4)	176	192	206	224	242	208.4	25.5	-0.6	25.4	18.1	1.0
5000	mKT	163	178	198	216	238	198.8	29.2	-10.2	30.9	17.8	1.2
5000	DH	163	177	193	212	230	195.1	26.9	-13.9	30.3	17.6	1.0

Compare tables 3.3.3, 3.4.4, 3.5.3, 3.5.21, 3.6.6.

Tabelle 3.4.5: Characteristics of distributions of n_m and c_m using the orthogonal series estimator of the density function with optimal cutoff selection for model 4.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	KT(1)	20	44	104	207	319	141.9	125.0	-26.1	127.7	21.3	9.2
100	KT(2)	59	106	186	279	372	204.7	127.5	36.7	132.6	25.6	7.8
100	KT(3)	57	111	197	294	399	217.8	141.0	49.8	149.5	26.3	8.2
100	KT(4)	55	107	199	301	415	225.1	158.4	57.1	168.4	26.5	8.9
100	mKT	38	72	155	336	554	242.6	246.0	74.6	257.0	26.6	12.3
100	DH	74	145	222	305	395	231.7	124.4	63.7	139.8	27.4	7.5
250	KT(1)	15	28	86	193	273	118.6	103.6	-49.4	114.8	19.1	8.7
250	KT(2)	81	119	171	234	300	182.7	85.6	14.7	86.8	24.8	5.5
250	KT(3)	81	118	171	237	306	185.1	89.7	17.1	91.3	25.0	5.7
250	KT(4)	78	117	171	241	315	187.1	94.5	19.1	96.4	25.0	6.0
250	mKT	52	82	135	213	326	167.8	123.7	-0.2	123.7	23.5	7.8
250	DH	63	96	153	232	297	169.2	91.0	1.2	91.0	23.7	6.1
500	KT(1)	15	25	117	180	240	118.0	87.8	-50.0	101.0	19.2	8.1
500	KT(2)	104	132	169	215	258	176.9	61.8	8.9	62.4	24.8	4.1
500	KT(3)	102	131	169	216	261	177.3	63.8	9.3	64.5	24.8	4.2
500	KT(4)	100	130	169	218	265	177.8	65.6	9.8	66.3	24.8	4.3
500	mKT	78	105	146	198	265	161.7	80.2	-6.3	80.4	23.7	5.4
500	DH	83	108	145	191	242	154.9	62.6	-13.1	64.0	23.2	4.4
5000	KT(1)	17	122	158	178	192	134.9	64.0	-33.1	71.9	21.2	6.6
5000	KT(2)	143	154	169	181	194	168.0	19.6	0.0	19.5	24.5	1.4
5000	KT(3)	142	154	167	181	194	168.0	19.9	0.0	19.8	24.5	1.4
5000	KT(4)	142	154	167	181	194	167.9	19.9	-0.1	19.9	24.5	1.4
5000	mKT	134	144	162	177	190	162.1	22.5	-5.9	23.2	24.1	1.6
5000	DH	143	156	169	184	195	169.4	20.5	1.4	20.5	24.6	1.4

Compare tables 3.3.4, 3.4.5, 3.5.4, 3.5.22, 3.6.7.

Tabelle 3.4.6: Characteristics of distributions of n_m and c_m using the orthogonal series estimator of the density function with optimal cutoff selection for model 5.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	KT(1)	153	246	409	628	863	466.6	287.0	-141.4	319.9	38.9	9.8
100	KT(2)	248	429	688	973	1273	735.3	406.6	127.3	426.1	45.6	12.2
100	KT(3)	244	426	700	1003	1339	759.7	443.2	151.7	468.4	46.2	12.9
100	KT(4)	228	413	696	1041	1424	780.7	493.7	172.7	523.0	46.6	14.1
100	mKT	142	315	651	1099	1674	795.6	630.1	187.6	657.4	45.8	17.6
100	DH	242	437	704	988	1241	731.2	383.4	123.2	402.7	45.3	12.0
250	KT(1)	124	199	464	680	856	470.3	289.8	-137.7	320.8	38.2	10.0
250	KT(2)	327	459	647	864	1067	676.7	286.8	68.7	294.9	44.7	9.0
250	KT(3)	318	452	648	871	1082	679.9	301.9	71.9	310.4	44.8	9.5
250	KT(4)	303	439	643	881	1107	683.0	322.9	75.0	331.4	44.8	10.1
250	mKT	219	340	548	858	1220	650.4	420.9	42.4	423.0	43.0	13.0
250	DH	324	462	641	842	1035	665.3	275.8	57.3	281.7	44.2	8.8
500	KT(1)	103	180	517	673	809	476.4	274.8	-131.6	304.7	38.0	10.3
500	KT(2)	386	484	613	773	921	637.7	210.3	29.7	212.4	44.0	6.8
500	KT(3)	378	479	613	776	930	638.8	218.6	30.8	220.7	44.0	7.1
500	KT(4)	371	474	612	778	940	639.4	227.9	31.4	230.0	44.0	7.4
500	mKT	299	405	558	752	962	603.8	272.6	-4.2	272.6	42.5	9.0
500	DH	371	488	636	794	932	647.6	217.3	39.6	220.9	44.2	7.2
5000	KT(1)	49	492	580	631	679	495.9	220.3	-112.1	247.0	38.2	10.9
5000	KT(2)	520	554	606	653	696	607.0	67.9	-1.0	67.8	43.5	2.3
5000	KT(3)	520	554	607	654	696	606.8	69.1	-1.2	69.0	43.5	2.4
5000	KT(4)	518	551	607	654	696	605.5	69.4	-2.5	69.3	43.5	2.4
5000	mKT	495	532	591	642	690	592.2	76.4	-15.8	77.9	43.0	2.6
5000	DH	599	645	696	742	780	689.4	79.6	81.4	113.7	46.2	2.6

Compare tables 3.3.5, 3.4.6, 3.5.5, 3.5.23, 3.6.8.

Tabelle 3.4.7: Characteristics of distributions of n_m and c_m using the orthogonal series estimator of the density function with optimal cutoff selection for model 6.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	KT(1)	47	83	136	293	424	197.9	157.1	-126.1	201.5	25.2	9.2
100	KT(2)	109	206	315	411	522	320.8	159.3	-3.2	159.3	31.4	7.6
100	KT(3)	157	247	344	452	574	361.7	173.8	37.7	177.9	33.2	7.3
100	KT(4)	148	244	348	471	634	378.7	208.8	54.7	215.9	33.7	8.4
100	mKT	77	145	304	553	891	407.9	358.9	83.9	368.6	33.5	13.8
100	DH	202	272	359	462	572	375.3	151.0	51.3	159.5	34.3	6.4
250	KT(1)	33	52	137	330	406	193.4	152.2	-130.6	200.5	24.1	9.6
250	KT(2)	188	256	327	394	456	325.8	106.6	1.8	106.6	32.1	4.9
250	KT(3)	192	258	332	404	478	335.6	115.4	11.6	116.0	32.5	5.1
250	KT(4)	185	255	334	421	514	346.1	135.4	22.1	137.2	32.9	5.9
250	mKT	113	170	268	416	608	322.5	214.8	-1.5	214.8	31.2	9.4
250	DH	216	280	341	397	448	338.0	97.8	14.0	98.8	32.7	4.3
500	KT(1)	28	43	235	332	389	205.1	146.8	-118.9	188.9	24.4	9.8
500	KT(2)	216	265	322	376	429	322.7	82.9	-1.3	82.9	32.2	3.8
500	KT(3)	212	266	325	385	443	328.8	91.9	4.8	92.0	32.4	4.2
500	KT(4)	209	263	327	397	464	334.5	101.7	10.5	102.3	32.6	4.6
500	mKT	151	203	277	379	493	304.4	140.7	-19.6	142.0	31.0	6.6
500	DH	204	261	328	380	424	321.9	86.7	-2.1	86.7	32.1	4.0
5000	KT(1)	26	34	293	320	342	232.4	127.2	-91.6	156.6	26.1	9.5
5000	KT(2)	283	301	326	343	364	323.5	31.3	-0.5	31.3	32.4	1.5
5000	KT(3)	281	301	326	344	366	323.7	32.6	-0.3	32.6	32.4	1.5
5000	KT(4)	279	300	326	345	366	324.0	33.7	0.0	33.7	32.4	1.6
5000	mKT	261	283	309	339	362	311.5	39.7	-12.5	41.6	31.9	1.9
5000	DH	284	302	327	351	369	326.7	33.2	2.7	33.3	32.6	1.6

Compare tables 3.3.6, 3.4.7, 3.5.6, 3.5.24, 3.6.9.

Tabelle 3.4.8: Characteristics of distributions of n_m and c_m using the orthogonal series estimator of the density function with optimal cutoff selection for model 7.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	KT(1)	290	410	562	745	1009	612.4	294.4	-592.6	661.7	45.5	7.9
100	KT(2)	610	819	1109	1431	1763	1151.1	466.6	-53.9	469.7	56.8	10.4
100	KT(3)	607	863	1181	1540	1914	1243.3	559.5	38.3	560.7	58.7	11.8
100	KT(4)	560	889	1273	1704	2218	1348.8	676.3	143.8	691.4	60.7	14.0
100	mKT	356	704	1254	1878	2617	1389.0	894.1	184.0	912.7	60.1	18.9
100	DH	680	940	1210	1500	1806	1233.9	456.4	28.9	457.3	58.9	9.3
250	KT(1)	252	347	566	811	1119	624.0	340.7	-581.0	673.5	44.8	8.7
250	KT(2)	764	965	1196	1426	1653	1209.8	358.6	4.8	358.6	58.8	8.0
250	KT(3)	757	977	1229	1490	1752	1250.8	406.7	45.8	409.3	59.7	8.9
250	KT(4)	711	965	1267	1576	1884	1293.6	474.2	88.6	482.3	60.5	10.3
250	mKT	492	758	1146	1638	2161	1263.5	680.8	58.5	683.2	58.7	15.1
250	DH	814	1014	1241	1500	1756	1263.5	369.8	58.5	374.4	60.1	7.9
500	KT(1)	221	309	629	1005	1222	683.4	381.8	-521.6	646.4	45.6	10.2
500	KT(2)	861	1027	1206	1379	1560	1208.7	276.0	3.7	276.0	59.2	6.1
500	KT(3)	846	1022	1222	1417	1624	1232.6	310.0	27.6	311.2	59.6	6.9
500	KT(4)	812	1005	1234	1467	1690	1250.8	354.4	45.8	357.3	60.0	7.8
500	mKT	650	858	1137	1474	1824	1201.4	472.3	-3.6	472.3	58.3	10.8
500	DH	869	1053	1261	1444	1616	1250.3	286.7	45.3	290.2	60.2	6.3
5000	KT(1)	138	618	1098	1230	1287	892.3	432.4	-312.7	533.1	50.2	13.3
5000	KT(2)	1061	1130	1210	1278	1346	1208.9	108.6	3.9	108.5	59.6	2.5
5000	KT(3)	1057	1129	1201	1278	1353	1205.5	113.4	0.5	113.3	59.5	2.6
5000	KT(4)	1044	1120	1201	1275	1353	1201.6	118.3	-3.4	118.2	59.4	2.8
5000	mKT	988	1072	1175	1262	1344	1171.7	140.0	-33.3	143.8	58.7	3.4
5000	DH	1120	1179	1237	1293	1337	1232.7	87.4	27.7	91.6	60.2	2.0

Compare tables 3.3.7, 3.4.8, 3.5.7, 3.5.25, 3.6.10.

Tabelle 3.4.9: Characteristics of distributions of n_m and c_m using the orthogonal series estimator of the density function with optimal cutoff selection for model 8.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	KT(1)	5	10	29	63	111	46.6	51.0	10.6	52.1	11.5	5.5
100	KT(2)	11	21	44	76	115	55.1	44.2	19.1	48.1	13.4	5.1
100	KT(3)	11	21	44	78	120	56.8	47.6	20.8	52.0	13.6	5.4
100	KT(4)	11	20	43	78	125	57.9	51.2	21.9	55.7	13.7	5.6
100	mKT	8	14	30	69	148	60.2	84.4	24.2	87.8	13.6	7.6
100	DH	6	12	31	73	132	51.4	52.5	15.4	54.8	12.3	5.9
250	KT(1)	6	16	32	53	73	37.4	27.6	1.4	27.6	11.2	4.0
250	KT(2)	17	25	39	57	76	43.5	24.2	7.5	25.3	12.6	3.3
250	KT(3)	16	25	39	58	77	43.7	25.1	7.7	26.2	12.6	3.4
250	KT(4)	16	25	39	58	78	44.0	25.9	8.0	27.1	12.6	3.5
250	mKT	11	17	27	46	73	36.4	31.6	0.4	31.6	11.7	4.2
250	DH	12	20	33	52	72	38.2	23.9	2.2	24.0	11.8	3.7
500	KT(1)	12	23	37	54	68	39.1	21.8	3.1	22.0	11.7	3.2
500	KT(2)	21	29	39	50	63	40.9	16.9	4.9	17.6	12.4	2.4
500	KT(3)	21	28	39	51	64	41.1	17.6	5.1	18.3	12.4	2.4
500	KT(4)	21	28	38	51	65	41.2	17.9	5.2	18.6	12.4	2.5
500	mKT	15	21	31	45	62	35.7	20.5	-0.3	20.5	11.7	2.9
500	DH	18	25	36	50	63	38.8	18.2	2.8	18.4	12.2	2.8
5000	KT(1)	23	32	38	42	52	39.0	15.0	3.0	15.2	12.1	2.1
5000	KT(2)	30	33	37	40	44	36.8	5.4	0.8	5.5	11.9	0.8
5000	KT(3)	30	33	37	40	44	36.8	5.5	0.8	5.5	11.9	0.8
5000	KT(4)	30	33	37	40	44	36.8	5.5	0.8	5.5	11.9	0.8
5000	mKT	28	32	35	39	43	35.7	6.0	-0.3	6.0	11.8	0.9
5000	DH	31	35	39	42	46	38.5	5.8	2.5	6.4	12.2	0.9

Compare tables 3.3.8, 3.4.9, 3.5.8, 3.5.26, 3.6.11.

Tabelle 3.4.10: Characteristics of distributions of n_m and c_m using the orthogonal series estimator of the density function with optimal cutoff selection for model 9.

m	method	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	KT(1)	37	63	94	128	161	99.1	53.1	-62.9	82.3	17.7	4.2
100	KT(2)	67	89	117	165	243	138.1	78.1	-23.9	81.6	20.8	5.2
100	KT(3)	73	110	169	249	330	190.1	110.1	28.1	113.6	24.1	6.5
100	KT(4)	68	112	179	262	351	200.7	125.3	38.7	131.1	24.6	7.1
100	mKT	40	75	149	305	519	226.8	223.5	64.8	232.7	25.1	11.4
100	DH	67	92	127	181	291	153.8	96.9	-8.2	97.2	21.7	6.3
250	KT(1)	41	57	92	124	152	95.6	47.6	-66.4	81.7	17.5	4.1
250	KT(2)	87	108	137	181	225	147.9	55.5	-14.1	57.2	21.9	3.9
250	KT(3)	94	129	173	221	267	178.3	69.5	16.3	71.4	23.8	4.4
250	KT(4)	92	129	176	225	277	181.6	74.8	19.6	77.3	24.0	4.7
250	mKT	58	91	146	224	335	176.7	123.4	14.7	124.2	23.3	7.3
250	DH	96	124	163	217	264	173.0	67.5	11.0	68.3	23.7	4.5
500	KT(1)	48	61	81	120	171	96.1	48.9	-65.9	82.1	17.7	4.1
500	KT(2)	106	130	163	197	225	164.7	46.0	2.7	46.1	23.2	3.1
500	KT(3)	111	140	173	207	239	175.1	50.3	13.1	52.0	23.8	3.2
500	KT(4)	109	137	172	207	242	174.8	53.7	12.8	55.2	23.8	3.4
500	mKT	80	108	148	203	270	164.7	80.4	2.7	80.4	23.0	5.1
500	DH	120	149	182	214	241	181.5	48.8	19.5	52.5	24.3	3.1
5000	KT(1)	72	77	86	163	175	115.5	44.5	-46.5	64.4	19.4	3.6
5000	KT(2)	145	155	165	175	185	165.3	15.2	3.3	15.5	23.3	1.0
5000	KT(3)	143	153	165	175	185	164.8	16.0	2.8	16.3	23.3	1.1
5000	KT(4)	143	153	165	175	185	164.4	16.4	2.4	16.5	23.3	1.1
5000	mKT	132	144	159	171	181	157.9	20.0	-4.1	20.4	22.9	1.4
5000	DH	149	157	166	175	185	166.1	14.3	4.1	14.9	23.4	0.9

Compare tables 3.3.9, 3.4.10, 3.5.9, 3.5.27, 3.6.12.

3.4.6 Summary

In Table 3.4.11, we present simulated root mean squared deviations (RMSD) of the sampling plan size using the OSE estimators for several models of the distribution of measurements. We can observe that the OSE estimator with DH cut-off is typically better than other OSE estimators.

Tabelle 3.4.11: RMSD of the distribution of the sampling plan size using the OSE estimators for models 1–9.

model	m	KT(1)	KT(2)	KT(3)	KT(4)	mKT	DH
1	100	58.7	61.7	66.9	75.6	131.1	66.4
1	250	34.1	34.4	36.8	39.1	56.0	39.1
1	500	22.2	22.6	24.1	25.9	33.9	22.4
2	100	92.7	119.0	140.6	158.7	217.3	99.9
2	250	63.4	69.7	75.5	79.6	102.9	60.9
2	500	50.4	48.5	51.3	53.8	61.7	44.4
3	100	217.0	162.3	164.3	175.5	280.7	206.8
3	250	145.3	100.0	102.5	106.3	151.5	108.9
3	500	110.8	77.4	77.3	78.9	98.9	87.3
4	100	127.7	132.6	149.5	168.4	257.0	139.8
4	250	114.8	86.8	91.3	96.4	123.7	91.0
4	500	101.0	62.4	64.5	66.3	80.4	64.0
5	100	319.9	426.1	468.4	523.0	657.4	402.7
5	250	320.8	294.9	310.4	331.4	423.0	281.7
5	500	304.7	212.4	220.7	230.0	272.6	220.9
6	100	201.5	159.3	177.9	215.9	368.6	159.5
6	250	200.5	106.6	116.0	137.2	214.8	98.8
6	500	188.9	82.9	92.0	102.3	142.0	86.7
7	100	661.7	469.7	560.7	691.4	912.7	457.3
7	250	673.5	358.6	409.3	482.3	683.2	374.4
7	500	646.4	276.0	311.2	357.3	472.3	290.2
8	100	52.1	48.1	52.0	55.7	87.8	54.8
8	250	27.6	25.3	26.2	27.1	31.6	24.0
8	500	22.0	17.6	18.3	18.6	20.5	18.4
9	100	82.3	81.6	113.6	131.1	232.7	97.2
9	250	81.7	57.2	71.4	77.3	124.2	68.3
9	500	82.1	46.1	52.0	55.2	80.4	52.5

Kapitel 3.5

Estimators based on Bernstein polynomials

3.5.1 Introduction

Bernstein and Bernstein-Durrmeyer polynomials are convenient tools for smoothing functions defined on a finite interval. For estimation of probability distributions, Bernstein polynomials have been studied in a number of papers Cheng (1995) and Perez, Palacin (1987) investigated the estimation of the quantile function. while Babu, Canty and Chaubey (2002) obtained results for estimating the distribution function. Density estimation based on Bernstein-Durrmeyer polynomials dating back to the seminal work of Durrmeyer (1967) has been studied by Ciesielski (1978). Rafajłowicz and Skubalska-Rafajłowicz (1999) have investigated related estimators for nonparametric regression and obtained results on optimal convergence rates.

It is well-known that Bernstein polynomials have good approximation properties, which are given in the following theorem.

Theorem. (Feller, 1965, Theorem 1, Section VII.2). If $u(x)$ is a bounded and continuous function on the interval $[0,1]$, then as $N \to \infty$

$$\hat{u}_N(x) = \sum_{i=1}^{N} u((i-1)/(N-1))b_i(x|N) \to u(x)$$

uniformly for $x \in [0,1]$, where

$$b_i(x|N) = C_{N-1}^{i-1} x^{i-1}(1-x)^{N-i}$$

and $C_n^j = n!/(j!(n-j)!)$.

Another approximation is based on Bernstein-Durrmeyer polynomials which differ by another definition of coefficients proposed by Durrmeyer (1967). Specifically, the approximation of a function $u(x)$ using Bernstein-Durrmeyer polynomials is defined by

$$D_N(u(x)) = (N+1)\sum_{i=0}^{N} a_i B_i^{(N)}(x)$$

where $B_i^{(N)}(x)$, $i = 0, \ldots, N$, are Bernstein polynomials of degree N and

$$a_i = \int_0^1 u(x) B_i^{(N)}(x) dx.$$

Theoretical properties of this approximation for deterministic functions were particularly established by Chen and Ditzian (1991), where it is shown that $\|u(x) - D_N(u(x))\|_p = O(N^{-\delta/2})$, if and only if

$$\epsilon(u) = \inf_{a_0,\ldots,a_N} \{\|u(x) - a_0 - a_1 x - \cdots - a_N x^N\|_p\} = O(N^{-\delta}),$$

as $N \to \infty$ for some $\delta \in (0,2)$ and $p = 2, \infty$, that is, if the function $u(x)$ can be approximated by a polynomial of degree less or equal to N at the L_p-rate $N^{-\delta}$.

However, in the context of quantile estimation we have to apply Bernstein-Durrmeyer polynomials to a random function given by the empirical quantile function. Therefore, we will establish the consistency and the rate of convergence of the Bernstein-Durrmeyer estimator. We also introduce an error-corrected estimator which improves the Bernstein-Durrmeyer estimator. Moreover, we propose a novel procedure of adaptive selection of the degree of the Bernstein-Durrmeyer polynomial.

3.5.2 The BP estimator of the distribution function

To apply the Bernstein polynomial approach, we transform the original interval $(-\infty, \infty)$ to the interval $(0,1)$, for example, using the transformation $1/2 + \pi^{-1}\arctan(x)$. Define $u(x) = F(t(x))$, where $t(x) = \tan(\pi(x - 0.5))$. Thus, $u(x)$ is a distribution function defined on the interval $(0,1)$ and $F_m(t(x))$ is the empirical distribution function. We can now apply the Bernstein polynomial approach to estimating $u(x)$.

In Figure 3.5.1, we depict the empirical distribution function and the Bernstein polynomial estimator for models 1 and 2 for different values of m and $N = k$.

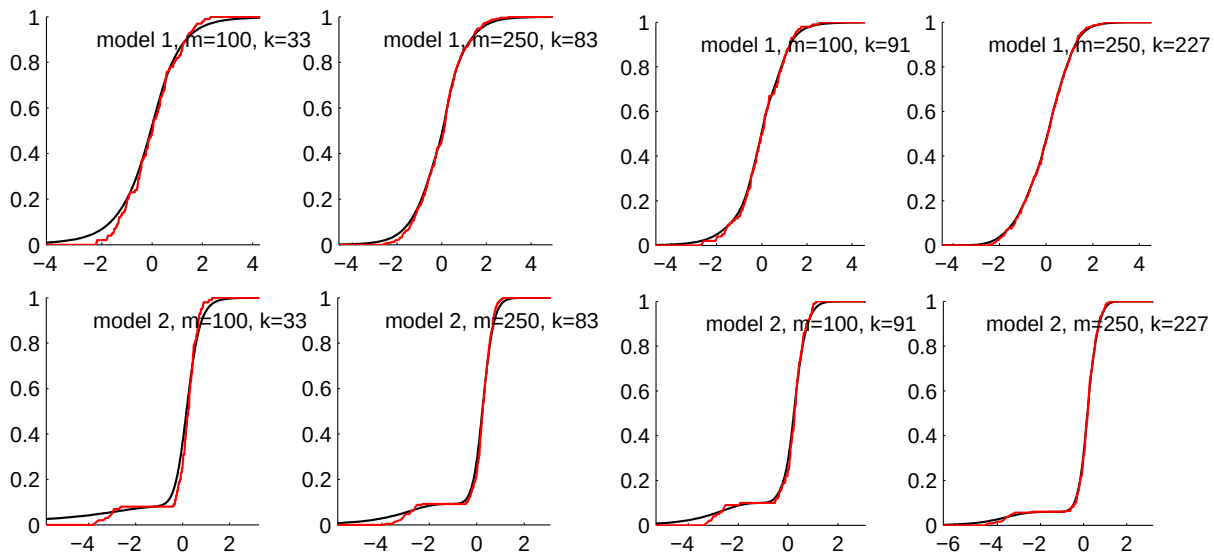


Abbildung 3.5.1: The empirical distribution function (red) and the Bernstein polynomial estimator (black) for models 1 and 2.

3.5.3 The BP estimator of the quantile function

Let $F(x)$ be a (cumulative) distribution function. The Bernstein polynomial (of degree k) estimator of the quantile function $F^{-1}(x)$ is given by

$$\hat{F}^{-1}_{m,k}(x) = \sum_{i=1}^{k} F^{-1}_m((i-1)/(k-1))C^{i-1}_{k-1}x^{i-1}(1-x)^{k-i},$$

where $C^j_n = n!/(j!(n-j)!)$ and $F^{-1}_m(x)$ is the empirical quantile function for a sample $(x_1, \ldots, x_m)$. In the case of $k = m$, we have that the BP estimator has the form

$$\hat{F}^{-1}_{m,m}(x) = \sum_{i=1}^{m} x_{(i)}C^{i-1}_{m-1}x^{i-1}(1-x)^{m-i}$$

where $x_{(i)}$ is the i-th order statistics.

In Figure 3.5.2, we depict the empirical quantile function and the Bernstein polynomial estimator for models 1 and 2 for different values of m and k.

3.5.4 The BDP estimator of the distribution function

To apply the Bernstein-Durrmeyer polynomial approach, we transform the original interval $(-\infty, \infty)$ to the interval $(0, 1)$. Define $u(x) = F(t(x))$, where $t(x) = \tan(\pi(x - 0.5))$. Thus,

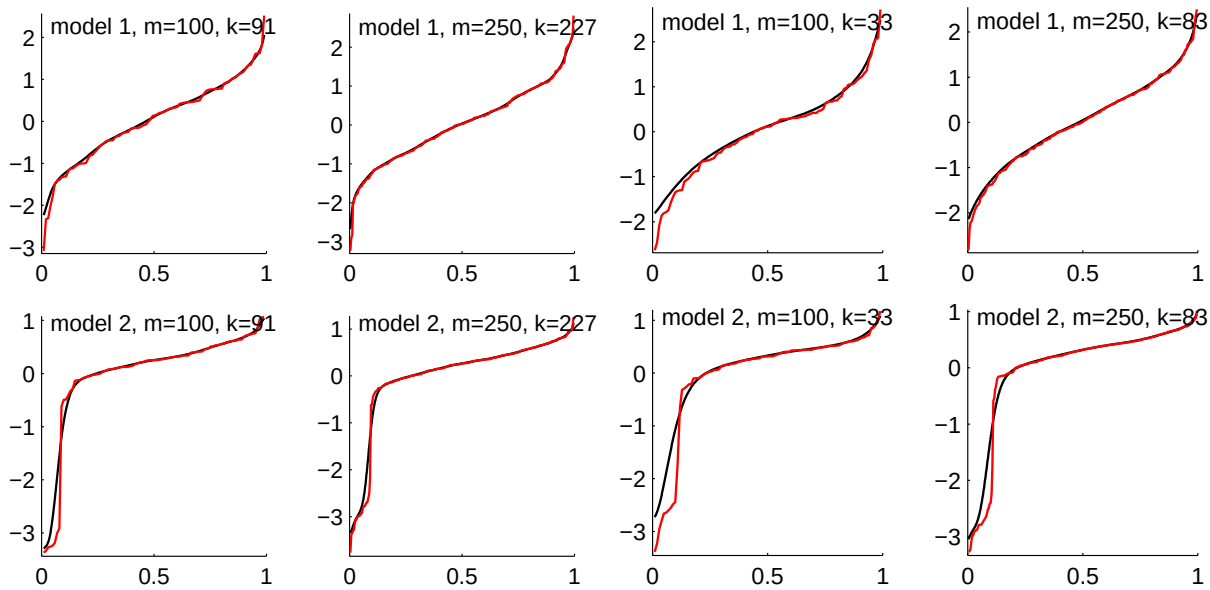


Abbildung 3.5.2: The empirical quantile function (red) and the Bernstein polynomial estimator (black) for models 1 and 2.

$u(x)$ is a distribution function defined on the interval $(0,1)$ and $F_m(t(x))$ is the empirical distribution function. Then the estimator of the distribution function using the Bernstein-Durrmeyer polynomial approach has the following form

$$\hat{F}(x) = \hat{u}(t^{-1}(x)),$$

where

$$\hat{u}(x) = \sum_{i=1}^{k} a_i b_i(x|k)$$

and

$$a_i = \int_0^1 F_m(t(x)) b_i(x|k) dx \approx \sum_{j=1}^{m} F_m(t(\frac{j-1}{m-1})) b_i(\frac{j-1}{m-1}|k).$$

In Figure 3.5.3, we depict the empirical distribution function and the Bernstein-Durrmeyer polynomial estimator for models 1 and 2 for different values of m and k.

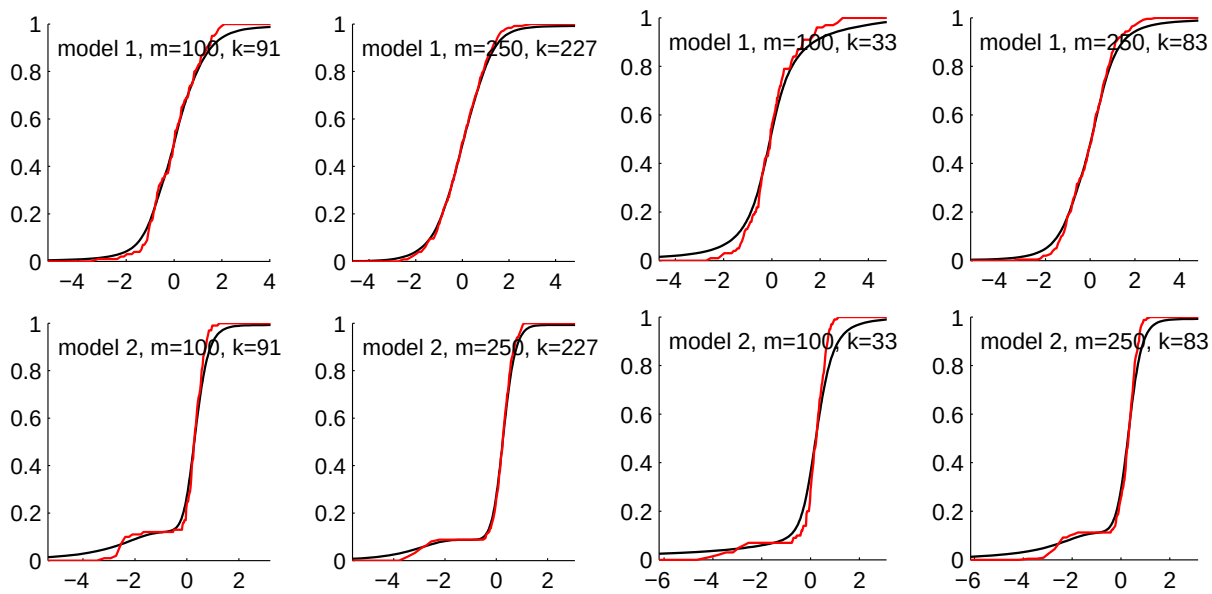


Abbildung 3.5.3: The empirical distribution function (red) and the Bernstein-Durrmeyer polynomial estimator (black) for models 1 and 2.

3.5.5 The BDP estimator of the quantile function

Let $F(x)$ be a (cumulative) distribution function. The Bernstein-Durrmeyer polynomial (of degree k) estimator of the quantile function $F^{-1}(x)$ is given by

$$\hat{F}^{-1}_{m,k}(x) = \sum_{i=1}^{k} a_i b_i(x|k) = \sum_{i=1}^{k} a_i C^{i-1}_{k-1} x^{i-1}(1-x)^{k-i},$$

where $C^j_n = n!/(j!(n-j)!)$,

$$a_i = \int_0^1 F^{-1}_m(x) b_i(x|k)dx \approx \frac{k}{m}\sum_{j=1}^{m} x_{(j)} b_i((j-1)/(m-1)|k),$$

and $x_{(j)}$ is the j-th order statistics.

In Figure 3.5.4, we depict the empirical quantile function and the Bernstein-Durrmeyer polynomial estimator for models 1 and 2 for different values of m and k.

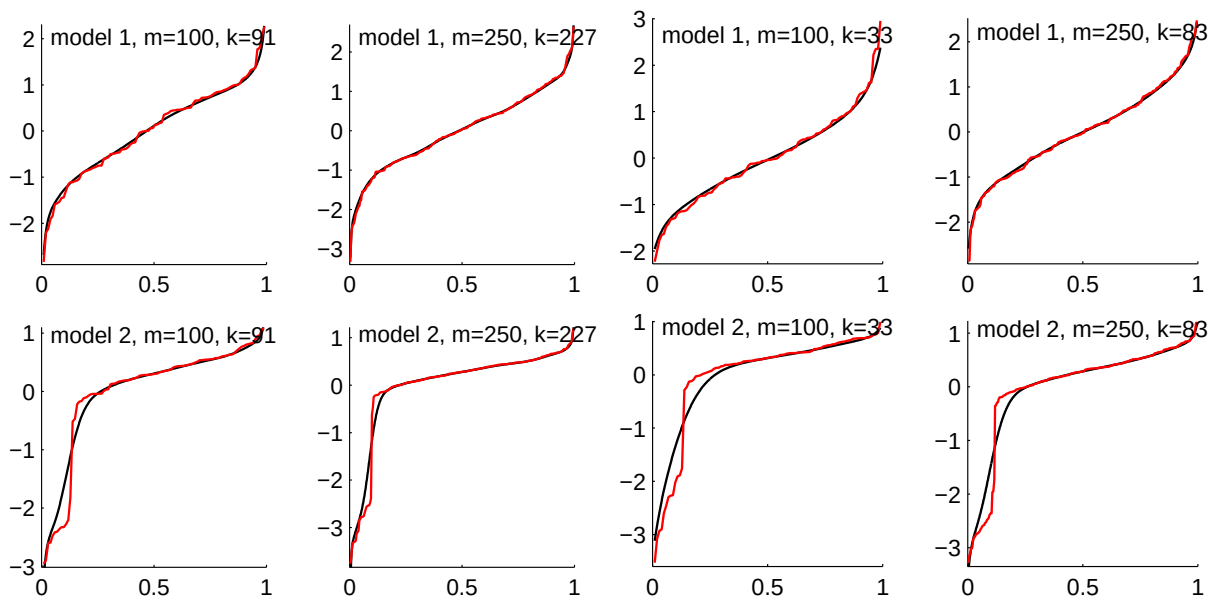


Abbildung 3.5.4: The empirical quantile function (red) and the Bernstein-Durrmeyer polynomial estimator (black) for models 1 and 2.

3.5.6 Consistency of the Bernstein-Durrmeyer estimator and error-correction

Throughout let $Q_m(x)$ be the empirical quantile function for a sample $X_1, \ldots, X_m$ from a distribution with the quantile function $Q(x)$. Then the function $D_N(Q_m(x))$ obtained by smoothing the sample quantiles by the Bernstein-Durrmeyer smoothing operator is is called the Bernstein-Durrmeyer estimator of the quantile function.

We provide a general result on the MSE- and MISE-consistency as well as the L_p-consistency. It turns out that the MISE is of the order $O(1/m + N^{-\delta})$ where δ is the L_p-rate at which the true quantile function $Q(x)$ can be approximated by a polynomial of degree N, cf. our discussion in the introduction.

Then we discuss a new error-corrected estimator and show its consistency.

3.5.6.1 MISE and L_p consistency

As a preparation, we need the following result that holds true for *any* sequence $X_1, \ldots, X_m$ of random variables, regardless of their joint distribution.

Lemma 3. *Let $X_1, \dots, X_m$ be m random variables. For the Bernstein-Durrmeyer estimator of the quantile function $D_N(Q_m(x)) = (N+1)\sum_{i=0}^N a_i B_i^{(N)}(x)$ with coefficients $a_i = \frac{1}{m}\sum_{j=1}^m X_{(j)} B_i^{(N)}(\frac{j-1}{m-1})$ we have*

$$\mathrm{Var}(a_i) \le \max_{j=1,\dots,m} \mathrm{Var}(X_{(j)}) \Big(\int_0^1 B_i^{(N)}(x)dx \Big)^2.$$

Proof. To get a bound for the variance of a_i, we write

$$\begin{aligned}\mathrm{Var}(a_i) &= m^{-2}\mathbf{E}\Big(\sum_{j=1}^m (X_{(j)} - \mathbf{E}X_{(j)}) B_i^{(N)}(\frac{j-1}{m-1})\Big)^2 \\ &= m^{-2}\Big(\sum_{l=1}^m B_i^{(N)}(\frac{l-1}{m-1})\Big)^2 \mathbf{E}\Big(\sum_{j=1}^m (X_{(j)} - \mathbf{E}X_{(j)}) w_j\Big)^2,\end{aligned}$$

where

$$w_j = \frac{B_i^{(N)}(\frac{j-1}{m-1})}{\sum_{l=1}^m B_i^{(N)}(\frac{l-1}{m-1})}.$$

Note that $w_1, \dots, w_N$ can be considered as weight coefficients since $w_j \ge 0$ and $\sum_{j=1}^N w_j = 1$. Therefore, the sum $\sum_{j=1}^N (X_{(j)} - \mathbf{E}X_{(j)}) w_j$ can be treated as the expectation $\mathbf{E}_\zeta(a_\zeta - \mathbf{E}a_\zeta)$ where ζ is a random variable such that $\mathbb{P}(\zeta = j) = w_j$, $j = 1, \dots, N$. By the Jensen inequality we have $(\mathbf{E}_\zeta(a_\zeta - \mathbf{E}a_\zeta))^2 \le \mathbf{E}_\zeta(a_\zeta - \mathbf{E}a_\zeta)^2$. Therefore, we obtain

$$\begin{aligned}\mathrm{Var}(a_i) &\le m^{-2}\Big(\sum_{l=1}^m B_i^{(N)}(\frac{l-1}{m-1})\Big)^2 \mathbf{E}\sum_{j=1}^m (X_{(j)} - \mathbf{E}X_{(j)})^2 \frac{B_i^{(N)}(\frac{j-1}{m-1})}{\sum_{l=1}^m B_i^{(N)}(\frac{l-1}{m-1})} \\ &= m^{-2}\sum_{l=1}^m B_i^{(N)}(\frac{l-1}{m-1}) \sum_{j=1}^m \mathrm{Var}(X_{(j)}) B_i^{(N)}(\frac{j-1}{m-1}) \\ &\le \max_{j=1,\dots,m} \mathrm{Var}(X_{(j)}) \Big(\int_0^1 B_i^{(N)}(x)dx \Big)^2,\end{aligned}$$

that completes the proof. □

In the following theorem we present an upper bound for the variance of $D_N(Q_m(x))$.

Theorem 3.5.6.1. *For the Bernstein-Durrmeyer estimator $D_N(Q_m(x)) = (N+1)\sum_{i=0}^N a_i B_i^{(N)}(x)$ of the quantile function $Q(x)$ of a random sample $X_1, \dots, X_m$, we have*

$$\mathrm{Var}(D_N(Q_m(x))) \le C_m$$

and

$$\int_0^1 \mathrm{Var}(D_N(Q_m(x)))dx \le C_m,$$

where $C_m = \max_{j=1,\dots,m} \mathrm{Var}(X_{(j)})$. *Moreover,* $C_m = O(1/m)$ *if the derivatives of* $Q(x)$ *are bounded.*

Proof. It is easy to see that

$$\mathrm{Var}(D_N(Q_m(x))) = (N+1)^2 \mathbf{E}\Big(\sum_{i=0}^N (a_i - \mathbf{E}a_i) B_i^{(N)}(x)\Big)^2.$$

Note that $B_0^{(N)}(x), \dots, B_N^{(N)}(x)$ for fixed x can be considered as weight coefficients since $B_i^{(N)}(x) \ge 0$ and $\sum_{i=0}^N B_i^{(N)}(x) = 1$. By the Jensen inequality we have

$$\mathrm{Var}(D_N(Q_m(x))) \le (N+1)^2 \sum_{i=0}^N \mathrm{Var}(a_i) B_i^{(N)}(x).$$

Since $\int_0^1 B_i^{(N)}(x)dx \le 1/(N+1)$, we obtain $\mathrm{Var}(a_i) \le C_m/(N+1)^2$. Therefore, $\mathrm{Var}(D_N(Q_m(x))) \le C_m$. We also have

$$\int \mathrm{Var}(D_N(Q_m(x)))dx \le C_m.$$

Let us now study the behavior of C_m. First, consider the sample $X_1, \dots, X_m$ taken from the uniform distribution. It is well-know that in this case the order statistics have the beta distribution, $X_{(j)} \sim \beta(j, m-j+1)$. Note that the variance of the beta distribution $\beta(a,b)$ is $\frac{ab}{(a+b)^2(a+b+1)}$. Therefore, it is easy to see that $C_m = O(1/m)$.

Consider now the general case. Since $Q(x)$ is continuous, the order statistics $X_{(j)}$ can be written as the transformation of the order statistics $U_{(j)}$ for a sample from the uniform distribution, namely $X_{(j)} = Q(U_{(j)})$. Expanding $Q(U_{(j)})$ in a Taylor series at $\mathrm{E}U_{(j)} = j/(m+1) = p_j$, we obtain

$$X_{(j)} = Q(p_j) + (U_{(j)} - p_j)Q'(p_j) + \frac{1}{2!}(U_{(j)} - p_j)^2 Q''(p_j) + \frac{1}{3!}(U_{(j)} - p_j)^3 Q'''(p_j) + \dots$$

Then as shown in (David, 1981, Sec. 4.5) we have

$$\mathrm{Var}X_{(j)} = \frac{p_j(1-p_j)}{m+2} Q'^2(p_j) + \frac{p_j(1-p_j)}{(m+2)^2} v_j + o((m+2)^{-2}).$$

where $v_j = 2(1-2p_j)Q'(p_j)Q''(p_j) + p_j(1-p_j)(Q'(p_j)Q'''(p_j) + Q''^2(p_j)/2)$. Thus we can see that $C_m = O(1/m)$ if the derivatives of $Q(x)$ are bounded. □

The following result provides the MSE and MISE consistency of $D_N(Q_m(x))$. We show that the rate of convergence of the Bernstein-Durrmeyer estimator to the true quantile function is bounded by the rate of convergence of the empirical quantile function to the true quantile function and the rate of convergence of the Bernstein-Durrmeyer approximation of the true quantile function.

Theorem 3.5.6.2. *Let $Q(x)$ be a quantile function that is continuous on $[a,b] \subset [0,1]$. Then the Bernstein-Durrmeyer estimator $D_N(Q_m(x))$ satisfies*

$$\max_{x\in[a,b]} \mathbf{E}(D_N(Q_m(x)) - Q(x))^2 = O(1/m + N^{-\delta/2})$$

and

$$\mathbf{E}\int_a^b (D_N(Q_m(x)) - Q(x))^2 dx = O(1/m + N^{-\delta/2})$$

as $m \to \infty$ and $N \to \infty$.

Proof. We can easily see that $\mathbf{E}(D_N(Q_m(x)) - Q(x))^2 = \mathrm{Var}(D_N(Q_m(x))) + (D_N(Q(x)) - Q(x))^2 + (D_N(Q(x)) - \mathbf{E}D_N(Q_m(x)))^2$. The first term converges uniformly to zero with the rate $1/m$ as $m \to \infty$ by Theorem 3.5.6.1. The second term converges uniformly to zero with the rate $N^{-\delta/2}$ as $N \to \infty$ by the approximation properties of Bernstein-Durrmeyer polynomials. Finally, we have

$$\begin{aligned} |D_N(Q(x)) - \mathbf{E}D_N(Q_m(x))| &\le \sum_{i=0}^{N} B_i^{(N)}(x) \int_0^1 |Q(u) - \bar{Q}_m(u)| B_i^{(N)}(u) du \\ &\le \frac{1}{N+1} \max_u |Q(u) - \bar{Q}_m(u)| \le O(N^{-1}) \end{aligned}$$

where $\bar{Q}_m(u) = \sum_{i=1}^m Q(i/m)\mathbf{1}_{[(i-1)/m, i/m)}(u)$ is a stepwise function. Now the statement of theorem obviously follows by combining bounds for these three terms. □

In the following theorem we prove the L_p-consistency of the Bernstein-Durrmeyer estimator.

Theorem 3.5.6.3. *For the Bernstein-Durrmeyer estimator of the quantile function $D_N(Q_m(x)) = (N+1)\sum_{i=0}^N a_i B_i^{(N)}(x)$ with $a_i = \frac{1}{m}\sum_{j=1}^m X_{(j)} B_i^{(N)}(\frac{j-1}{m-1})$ we have*

$$\|D_N(Q_m(x)) - Q(x)\|_p \le \|Q_m(x) - Q(x)\|_p + \|D_N(Q(x)) - Q(x)\|_p \to 0 \quad (3.5.6.1)$$

for any $p \ge 1$.

Proof. Due to the triangular inequality and the linearity of $D_N(\cdot)$, we obtain

$$\|D_N(Q_m(u)) - Q(u)\|_p \leq \|D_N(Q_m(x) - Q(x))\|_p + \|D_N(Q(x)) - Q(x)\|_p.$$

Now the statement follows from the fact that $\|D_N(u(x))\|_p \leq \|u(x)\|_p$. □

3.5.6.2 Bernstein-Durrmeyer approach with correction term

According to the above results the Bernstein-Durrmeyer estimator $D_{Q_m,N}(x) = D_N(Q_m(x))$ yields a consistent estimator for the true quantile function $Q(x)$. Adding the true but unobservable correction term

$$e_N(x) = Q(x) - D_{Q_m,N}(x),$$

to $D_{Q_m,N}(x)$ recovers the true quantile function $Q(a)$. This fact suggests to add a smoothed estimator of $e_N(x)$ in order to obtain a corrected estimator with higher precision. Thus, let us study the estimator

$$\widehat{e}_l(x) = D_l(Q_m(x) - D_{Q_m,N}(x))$$

for some degree $l \in \mathbb{N}_0$. Here $Q_m(x)$ is the quantile estimator based on the sample $X_1, \ldots, X_m$, usually again the classical empirical sample quantile function as indicated by our notation. However, our general consistency result below shows that the latter can be replaced by *any* uniformly strong consistent estimator. The resulting *error-corrected Bernstein-Durrmeyer estimator* is now defined as

$$\widehat{D}_{Q_m,N,l}(x) = D_{Q_m,N}(x) + \widehat{e}_l(x).$$

The following result shows that $\widehat{D}_{Q_m,N,l}(x)$ is consistent for $Q(x)$ in the pth mean under very general conditions.

Theorem 3.5.6.4. *Suppose $X_1, X_2, \ldots$ are i.i.d. with common distribution function $F(x)$ satisfying $\int |x|^p dF(x) < \infty$ for some $1 \leq p \leq \infty$. Let $Q_m(x)$ be some quantile estimator based on $X_1, \ldots, X_m$ such that*

$$\sup_{0<x<1} |Q_m(x) - Q(x)| \to 0, \qquad \textit{almost surely}, \tag{3.5.6.2}$$

as $m \to \infty$. Then the error-corrected estimator $\widehat{D}_{Q_m,N,l}(x)$ is consistent in the pth mean for the true quantile function $Q(x)$,

$$\|\widehat{D}_{Q_m,N,l} - Q\|_p \to 0,$$

as $m \to \infty$ *and* $N, l \to \infty$.

Proof. By virtue of (3.5.6.1) and the estimator

$$\|\widehat{D}_{Q_m,N,l} - Q\|_p \leq \|D_{Q_m,N} - Q\|_p + \|D_l(Q_m - D_{Q_m,N})\|_p,$$

it suffices to show that the second term converges to zero, as $N, l, m \to \infty$. Notice that

$$\|D_{Q_m,N}\|_p = \|D_N(Q_m)\|_p \leq \|Q_m\|_p < \infty,$$

a.s., since $\|Q_m\|_p \leq \|Q_m - Q\|_p + \|Q\|_p$, where

$$\|Q_m - Q\|_p \leq \|Q_m - Q\|_\infty \to 0,$$

a.s. as $m \to \infty$, by (3.5.6.2), and $\|Q\|_p \leq \int |x| dF(x) < \infty$. Hence $Q_m - Q \in L_p$ and we obtain

$$\begin{aligned}
\|D_l(Q_m - D_{Q_m,N})\|_p &\leq \|Q_m - D_N(Q_m)\|_p \\
&\leq \|Q_m - Q\|_p + \|D_N(Q_m) - Q\|_p \\
&\leq \|Q_m - Q\|_\infty + \|D_N(Q_m) - Q\|_p
\end{aligned}$$

and a further application of (3.5.6.1) yields the result. □

3.5.7 Adaptive selection of N

In Figure 3.5.5 we depict the Bernstein-Durrmeyer estimator for different values of N. We can observe that the estimator becomes less smoother as N increases. Therefore, for practical use of the Bernstein-Durrmeyer estimator we need to have a procedure of adaptive selection of the parameter N.

Let us adapt the method proposed in (Golyandina et al., 2012) for selection of the bandwidth and SSA estimation to the case of selection of N for the Bernstein-Durrmeyer estimator of the quantile function. The basic idea of this method is to control the number of modes appropriately.

Thus, we need to find the number of modes of the density of the distribution $\hat{F}_N(x)$ corresponding to $\hat{Q}_N(q)$. To do this, we differentiate the identity $\hat{F}_N(\hat{Q}_N(q)) \equiv q$ that gives

$$\hat{p}_N(\hat{Q}_N(q)) = 1/\hat{Q}'_N(q), \tag{3.5.7.1}$$

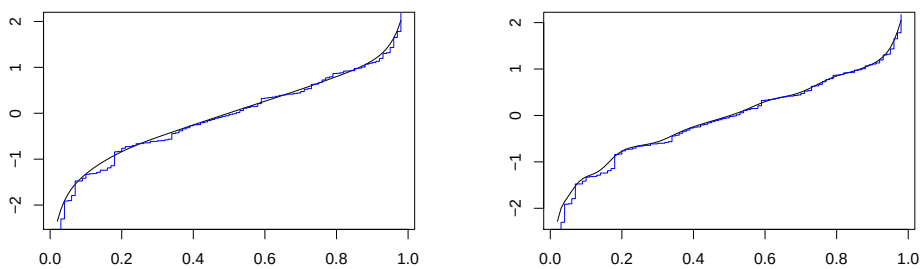


Abbildung 3.5.5: The empirical quantile function for a sample of size $m = 100$ from the normal distribution and the Bernstein-Durrmeyer estimator with $N = 50$ (left) and $N = 500$ (right).

where $\hat{p}_N(x) = \hat{F}'_N(x)$. Note that the condition of the existence of the mode of the density $\hat{p}_N(x)$ at a point $x = \hat{Q}_N(q)$ can be written as

$$\hat{p}_N(\hat{Q}_N(q-\varepsilon)) < \hat{p}_N(\hat{Q}_N(q)) > \hat{p}_N(\hat{Q}_N(q+\varepsilon))$$

for some small $\varepsilon > 0$. Using (3.5.7.1) we obtain the condition

$$1/\hat{Q}'_N(q-\varepsilon) < 1/\hat{Q}'_N(q) > 1/\hat{Q}'_N(q+\varepsilon).$$

Since $\hat{Q}'_N(q) > 0$ for all $q \in (0,1)$, we get

$$\hat{Q}'_N(q-\varepsilon) > Q'_N(q) < \hat{Q}'_N(q+\varepsilon).$$

This implies that the number of modes of the estimated density $p_N(x)$ is equal to the number of local minimums of the function $\hat{Q}'_N(q)$.

Let us now define a quantity depending on N in such a way that it can be used as a bandwidth in common smoothing procedures such as quantile estimation based on inverting the integrated kernel density estimator. To this end, notice that, as a consequence of the Moivre-Laplace theorem, the Bernstein polynomial $B_i^{(N)}(x)$ can be approximated by the density of the normal distribution $\phi((i/N - x)/s)$, where $s = \sqrt{x(1-x)/N}$ and $\phi(x) = (2\pi)^{-0.5}e^{-x^2/2}$. Thus, we can introduce the quantity $h = 1/\sqrt{N}$ which plays a role of the bandwidth in what follows.

We are now ready to formulate the algorithm of adaptive selection of the parameter N as follows.

Algorithm:

1. Compute

$$\bar{h} = \max\left\{h \in (0,1] : \max_q \left|F_m(Q_{1/h^2}(q)) - q\right|_\infty \leq 1/R_m \ \forall\, \hbar \in (0,h)\right\}$$

   where $F_m(x)$ is the empirical distribution function and

$$R_m = 2\sqrt{m}/\sqrt{2\log\log m}. \tag{3.5.7.2}$$

2. Define a dense set $\{h_1, \ldots, h_n\} \in (0, \bar{h})$ and compute the sequence $M_1, \ldots, M_n$, where M_j be the number of local minimums of the Bernstein-Durrmeyer estimator with $N = 1/h_j^2$.

3. Compute $\check{M}_j = \min\{M_1, M_2, \ldots, M_j\}$, $j = 1, \ldots, n$.

4. Divide the set $\{h_1, \ldots, h_n\}$ into groups as follows. Define a_i and b_i such that $a_1 \leq b_1 < a_2 \leq b_2 < \ldots < a_k \leq b_k$ and $\check{M}_i = \check{M}_j$ for all $h_i, h_j \in [a_l, b_l]$ for $l \in \{1, \ldots, k\}$.

5. Compute $h_a = \sum_{i=1}^k a_i w_i$, where $w_i = (b_i - a_i)/\sum_{j=1}^k (b_j - a_j)$, and then set $N = 1/h_a^2$.

Proposition 3.5.7.1. *Suppose that $Q(q)$ is a continuous quantile function. The Bernstein-Durrmeyer estimator $\hat{Q}_m(q) = D_N(Q_m(q))$ with adaptive selection of N is consistent as $m \to \infty$. Moreover, we have the uniform error bound*

$$\sup_{-\infty<x<\infty} |\hat{Q}_m^{-1}(x) - F(x)| \leq \sqrt{2\ln\ln m}/\sqrt{m}.$$

Proof. Denote the underlying distribution function by $F(x)$ and let $\hat{Q}(q) = D_N(Q_m(q))$. Using the definition of R_m we note that

$$\begin{aligned}
|\hat{Q}_m^{-1}(x) - F(x)| &\leq |\hat{Q}_m^{-1}(x) - F_m(x)| + |F_m(x) - F(x)| \\
&\leq \frac{\sqrt{2\ln\ln m}}{2\sqrt{m}} + \|F_m - F\|_\infty \\
&\leq 2\frac{\sqrt{2\ln\ln m}}{2\sqrt{m}} + R_m^{-1}\frac{2\sqrt{m}}{\sqrt{2\ln\ln m}}\|F_m - F\|_\infty \\
&\leq 2\frac{\sqrt{2\ln\ln m}}{2\sqrt{m}},
\end{aligned}$$

where the upper bound for the first term follows from the first step of the procedure of adaptive selection of N and an application of the law of iterated logarithms. This establishes

the uniform error bound and in turn the consistency of $\hat{Q}_m(q)$ for estimating the quantile function $F^{-1}(x)$ by continuity. $\square$

Putting a tighter bound on the condition that constraints the selection of $\bar{h}$ allows us to establish a weak convergence result for the empirical process

$$\sqrt{m}\,[\hat{Q}_m^{-1}(x) - F(x)]$$

taking values in the Skorohod space $D(\mathbb{R};\mathbb{R})$. So let us now replace (3.5.7.2) by

$$R_m^{-1} = o(m^{-1/2}). \tag{3.5.7.3}$$

Let $B_t, t \in [0,1]$, be a standard Wiener process, i.e. a Gaussian process with mean 0 and covariance function $\mathrm{Cov}(B_s, B_t) = \min(s,t)$, $s,t \in [0,1]$. Then $B^0(t) = W_t - tW_1$, $t \in [0,1]$, is a Brownian bridge. Its covariance function is given by $\mathrm{Cov}(B^0(s), B^0(t)) = s(1-t)$, $s,t \in [0,1]$. It appears as the limit process of the empirical process $U_n(t) = n^{-1/2}\sum_{i\le nt}[1(U_i \le t) - t]$, $t \in [0,1]$, of an i.i.d. sample $U_1,\dots,U_n$ with uniform distribution, i.e. $U_n(t) \Rightarrow B^0(t)$, as $n \to \infty$, where $\Rightarrow$ stands for weak convergence in the Skorohod space $D([0,1];\mathbb{R})$ of right-continuous functions $f : [0,1] \to \mathbb{R}$ with existing left-hand limits. This implies that $\sqrt{m}[F_m(x) - F(x)]$, $x \in \mathbb{R}$, converges weakly to $B^0(F)$. It is worth mentioning that on richer probability space it holds with probability 1

$$\max_{n\le m}\sup_{-\infty<x<\infty}|\sqrt{n}[F_n(x) - F(x)] - B^0(F)(x)| = O(m^{1/2}(\log m)^{-\lambda})$$

for some $\lambda > 0$, cf. Berkes and Philipp (1977) and Philipp and Pinzur (1980).

For the Bernstein-Durrmeyer polynomial estimator with adaptive selection of the degree N we have the following invariance principle.

Theorem 3.5.7.1. *If R_m is selected according to (3.5.7.3), then we have*

$$\{\sqrt{m}[\hat{Q}_m^{-1}(x) - F(x)] : \infty < x < \infty\} \Rightarrow \{B^0(F(x)) : -\infty < x < \infty\},$$

as $m \to \infty$, in the Skorohod space $D(\mathbb{R};\mathbb{R})$.

Proof. By continuity of the limit process, it suffices to show the weak convergence of $\sqrt{m}[\hat{Q}_m^{-1}(x) - F(x)]$, $a < x < b$, in the Skorohod space $D([a,b];\mathbb{R})$ for $-\infty < a < b < \infty$. Consider

$$\sqrt{m}[\hat{Q}_m^{-1}(x) - F(x)] = \sqrt{m}[F_m(x) - F(x)] + G_m(x),$$

where $F_m(x)$ is the empirical distribution function. As well known,

$$\sqrt{m}[F_m(x) - F(x)] \Rightarrow B^0(F(x)),$$

as $m \to \infty$, and, denoting by $\| \bullet \|_\infty$ the supnorm over $[a, b]$, we have

$$\|G_m\|_\infty = \|\sqrt{m}[\hat{Q}_m^{-1} - F_m]\|_\infty = o(1),$$

as $m \to \infty$, by Assumption (3.5.7.3), which completes the proof. $\square$

3.5.8 Simulation study

3.5.8.1 Results for Bernstein polynomial approach

We can see that the Bernstein polynomial estimator for the quantile function (Tables 3.5.1–3.5.9) gives better results than the Bernstein polynomial estimator for the empirical distribution function (Tables 3.5.19–3.5.27) but yields worsen results than the kernel density estimator.

Tabelle 3.5.1: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the quantile function for model 1.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	34	48	75	122	194	104.3	108.5	39.3	115.4	17.0	6.1
100	150	34	52	85	146	249	126.2	152.0	61.2	163.8	18.3	7.5
100	200	29	44	72	125	212	109.7	139.6	44.7	146.6	17.2	7.3
100	250	30	46	78	142	255	127.0	178.3	62.0	188.8	18.1	8.4
100	350	28	44	77	144	263	131.3	202.9	66.3	213.5	18.2	9.1
250	100	56	72	95	128	174	107.4	55.6	42.4	69.9	17.7	3.8
250	150	47	62	83	114	158	95.6	52.9	30.6	61.1	17.0	3.9
250	200	43	57	78	109	153	91.3	53.9	26.3	60.0	16.6	4.1
250	250	38	50	70	97	140	81.8	49.8	16.8	52.6	16.0	4.0
250	350	39	52	73	105	153	88.5	58.8	23.5	63.3	16.4	4.5
500	100	63	75	92	114	139	97.4	31.7	32.4	45.3	17.2	2.5
500	150	55	67	83	105	130	88.5	31.1	23.5	39.0	16.6	2.6
500	200	50	61	77	97	122	82.4	30.1	17.4	34.8	16.1	2.6
500	250	46	57	72	92	117	77.7	29.6	12.7	32.2	15.8	2.7
500	350	46	57	73	95	122	79.2	32.0	14.2	35.0	15.9	2.9
5000	100	84	88	95	100	106	94.9	8.8	29.9	31.2	17.1	0.7
5000	150	73	77	82	88	93	82.9	8.2	17.9	19.6	16.3	0.7
5000	200	68	71	77	82	88	77.2	8.0	12.2	14.6	15.8	0.8
5000	250	65	69	74	80	85	74.5	8.1	9.5	12.5	15.6	0.8
5000	350	62	66	72	77	82	72.1	8.3	7.1	10.9	15.5	0.8
100	adapt	31	45	72	121	194	102.9	110.5	37.9	116.8	17.6	6.3
250	adapt	30	39	52	71	98	59.4	31.9	-5.6	32.3	14.4	3.1
500	adapt	32	38	48	61	76	51.8	18.5	-13.2	22.8	13.6	2.1

Compare tables 3.3.1, 3.4.2, 3.5.1, 3.5.19, 3.6.4.

Tabelle 3.5.2: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the quantile function for model 2.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	6	20	62	139	248	107.7	162.8	4.7	162.9	23.9	13.9
100	150	7	29	80	173	320	142.6	232.4	39.6	235.7	27.6	16.4
100	200	8	31	83	167	308	138.8	220.5	35.8	223.4	28.0	16.0
100	250	9	36	92	187	354	162.9	281.9	59.9	288.2	30.1	17.7
100	350	11	41	97	196	381	176.1	323.7	73.1	331.8	31.4	18.7
250	100	14	30	67	122	188	88.4	78.1	-14.6	79.4	24.0	9.9
250	150	20	45	84	142	208	104.0	83.0	1.0	82.9	26.9	10.0
250	200	26	53	92	148	218	111.8	85.7	8.8	86.2	28.4	9.9
250	250	32	58	94	146	214	113.1	82.9	10.1	83.6	29.3	9.5
250	350	36	62	100	156	234	122.6	93.3	19.6	95.3	30.4	10.1
500	100	25	42	68	104	147	78.5	51.1	-24.5	56.7	23.8	7.2
500	150	34	56	87	123	167	95.6	54.9	-7.4	55.4	27.0	7.1
500	200	42	65	95	131	172	103.4	55.1	0.4	55.1	28.6	6.9
500	250	48	69	98	133	175	106.4	54.3	3.4	54.4	29.4	6.6
500	350	52	73	102	139	185	112.1	57.2	9.1	57.9	30.3	6.7
5000	100	49	54	64	80	87	66.8	15.6	-36.2	39.4	23.2	2.6
5000	150	68	77	87	98	108	87.5	16.1	-15.5	22.3	27.2	2.2
5000	200	76	86	96	106	115	96.2	15.5	-6.8	16.9	28.8	2.0
5000	250	81	90	99	109	118	99.8	15.0	-3.2	15.3	29.5	1.9
5000	350	84	92	102	111	121	102.2	14.8	-0.8	14.8	30.0	1.9
100	adapt	8	24	69	146	269	127.5	319.8	24.5	320.7	26.3	15.9
250	adapt	30	53	88	134	197	105.6	78.0	2.6	78.1	28.4	9.2
500	adapt	51	69	93	125	165	103.1	51.5	0.1	51.5	29.5	6.3

Compare tables 3.3.2, 3.4.3, 3.5.2, 3.5.20, 3.6.5.

Tabelle 3.5.3: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the quantile function for model 3.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	100	148	235	395	651	330.2	323.0	121.2	345.0	20.8	7.4
100	150	102	161	267	459	820	401.1	458.4	192.1	497.0	22.3	9.1
100	200	86	137	230	402	710	346.9	400.5	137.9	423.5	21.0	8.7
100	250	89	145	250	453	840	404.0	537.0	195.0	571.3	22.2	10.2
100	350	84	138	244	457	872	419.2	629.7	210.2	663.8	22.3	11.1
250	100	169	222	301	416	574	344.1	180.7	135.1	225.6	21.8	4.8
250	150	143	192	266	372	522	306.8	171.1	97.8	197.1	20.7	4.9
250	200	132	179	249	355	501	292.6	172.4	83.6	191.6	20.3	5.0
250	250	115	157	221	319	454	261.2	157.6	52.2	166.0	19.4	4.9
250	350	118	163	233	348	495	282.6	184.0	73.6	198.1	20.0	5.4
500	100	195	236	294	367	449	312.3	107.9	103.3	149.4	21.1	3.2
500	150	172	210	265	336	417	284.0	104.9	75.0	128.9	20.3	3.3
500	200	156	193	245	314	396	264.5	101.8	55.5	115.9	19.7	3.3
500	250	144	180	230	299	377	249.3	99.2	40.3	107.1	19.3	3.4
500	350	142	179	232	306	392	254.2	107.4	45.2	116.5	19.4	3.6
5000	100	264	282	300	325	346	303.6	31.8	94.6	99.8	21.0	1.0
5000	150	227	244	262	282	303	264.3	29.6	55.3	62.7	19.9	1.0
5000	200	211	227	244	266	285	246.6	28.5	37.6	47.2	19.3	1.0
5000	250	202	219	236	256	275	237.7	28.6	28.7	40.5	19.1	1.1
5000	350	195	210	228	249	267	229.8	28.9	20.8	35.6	18.8	1.1
100	adapt	94	138	225	386	650	327.9	351.2	118.9	370.7	21.2	7.9
250	adapt	93	122	169	229	317	193.4	109.9	-15.6	111.0	17.4	4.0
500	adapt	112	136	171	216	267	182.7	66.5	-26.3	71.5	17.1	2.6

Compare tables 3.3.3, 3.4.4, 3.5.3, 3.5.21, 3.6.6.

Tabelle 3.5.4: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the quantile function for model 4.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	80	116	181	290	457	241.4	225.7	73.4	237.3	26.7	9.3
100	150	82	126	206	347	587	293.5	310.5	125.5	334.9	28.9	11.5
100	200	71	111	181	311	521	262.3	286.7	94.3	301.8	27.5	11.1
100	250	73	115	194	349	619	302.0	366.9	134.0	390.5	29.0	12.9
100	350	69	112	190	355	654	314.9	428.2	146.9	452.6	29.3	14.0
250	100	127	164	221	299	397	246.6	118.3	78.6	142.0	27.9	5.9
250	150	113	147	201	278	380	229.6	120.6	61.6	135.4	27.1	6.2
250	200	106	139	193	270	375	222.4	123.8	54.4	135.2	26.7	6.5
250	250	95	126	177	250	348	205.1	119.3	37.1	124.9	25.9	6.5
250	350	96	130	185	267	381	219.1	138.0	51.1	147.1	26.5	7.1
500	100	146	174	213	265	324	226.9	73.1	58.9	93.9	27.2	3.9
500	150	133	160	199	251	311	213.1	73.6	45.1	86.3	26.6	4.1
500	200	124	151	189	239	301	203.1	73.5	35.1	81.4	26.1	4.3
500	250	117	143	181	231	293	194.8	73.0	26.8	77.8	25.7	4.3
500	350	116	142	183	237	305	198.6	79.0	30.6	84.7	25.9	4.7
5000	100	194	205	220	234	243	219.9	19.8	51.9	55.6	27.1	1.1
5000	150	175	186	201	213	224	200.1	19.2	32.1	37.4	26.1	1.2
5000	200	165	176	191	203	215	190.2	19.1	22.2	29.3	25.6	1.2
5000	250	160	171	186	198	211	185.4	19.3	17.4	26.0	25.4	1.3
5000	350	155	166	182	194	208	181.1	19.8	13.1	23.8	25.2	1.3
100	adapt	79	113	176	286	465	242.0	249.3	74.0	260.0	27.3	9.7
250	adapt	84	108	150	211	278	170.7	91.4	2.7	91.4	24.0	5.5
500	adapt	99	120	153	196	243	164.6	61.6	-3.4	61.7	24.0	4.0

Compare tables 3.3.4, 3.4.5, 3.5.4, 3.5.22, 3.6.7.

Tabelle 3.5.5: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the quantile function for model 5.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	282	420	655	1055	1673	877.1	823.9	269.1	866.6	48.0	16.8
100	150	291	454	744	1268	2143	1067.8	1134.1	459.8	1223.6	52.1	20.7
100	200	257	400	656	1131	1907	954.1	1047.9	346.1	1103.5	49.3	20.1
100	250	262	419	707	1267	2256	1099.0	1340.8	491.0	1427.8	52.1	23.2
100	350	247	405	691	1289	2377	1146.0	1565.8	538.0	1655.5	52.6	25.3
250	100	462	591	801	1084	1439	894.7	432.1	286.7	518.5	50.4	10.7
250	150	406	532	730	1009	1382	833.4	440.0	225.4	494.3	48.8	11.3
250	200	381	504	695	979	1367	807.1	451.8	199.1	493.6	48.0	11.7
250	250	345	457	639	907	1258	744.0	435.3	136.0	456.0	46.3	11.7
250	350	346	470	670	969	1377	795.0	503.0	187.0	536.6	47.5	12.9
500	100	528	632	774	964	1180	823.0	266.4	215.0	342.3	49.2	7.1
500	150	481	583	723	910	1132	773.1	267.7	165.1	314.5	47.8	7.5
500	200	448	545	685	868	1093	736.8	267.2	128.8	296.6	46.8	7.7
500	250	423	517	654	834	1065	706.6	265.6	98.6	283.3	45.9	7.8
500	350	418	516	663	858	1108	720.3	287.3	112.3	308.5	46.3	8.4
5000	100	702	742	799	849	884	797.4	72.1	189.4	202.6	49.0	2.1
5000	150	634	673	727	771	811	725.2	69.9	117.2	136.4	47.0	2.1
5000	200	596	640	693	739	780	689.6	69.7	81.6	107.2	46.0	2.2
5000	250	580	619	673	719	765	672.0	70.1	64.0	94.9	45.5	2.3
5000	350	561	601	659	703	751	656.3	72.3	48.3	86.9	45.0	2.4
100	adapt	246	372	606	1057	1830	961.6	1269.3	353.6	1317.4	49.0	22.5
250	adapt	308	418	591	847	1179	693.6	426.9	85.6	435.3	44.7	11.8
500	adapt	366	458	595	772	1020	651.4	274.7	43.4	278.0	44.3	8.4

Compare tables 3.3.5, 3.4.6, 3.5.5, 3.5.23, 3.6.8.

Tabelle 3.5.6: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the quantile function for model 6.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	152	224	349	560	883	466.0	435.4	142.0	458.0	35.5	12.4
100	150	156	242	397	671	1141	566.9	599.6	242.9	646.8	38.5	15.3
100	200	137	213	349	600	1008	506.5	553.9	182.5	583.2	36.5	14.8
100	250	140	224	376	673	1195	583.3	709.0	259.3	754.8	38.6	17.2
100	350	132	216	367	685	1264	608.2	827.9	284.2	875.3	38.9	18.7
250	100	246	315	425	575	762	475.5	228.6	151.5	274.2	37.2	7.9
250	150	217	283	388	535	734	442.9	233.0	118.9	261.5	36.1	8.3
250	200	204	269	370	520	724	428.9	239.3	104.9	261.2	35.5	8.6
250	250	184	244	340	481	669	395.4	230.5	71.4	241.3	34.3	8.6
250	350	185	250	356	515	733	422.5	266.5	98.5	284.1	35.2	9.5
500	100	281	336	411	511	626	437.5	141.1	113.5	181.0	36.3	5.2
500	150	256	309	385	484	600	410.9	141.9	86.9	166.4	35.4	5.5
500	200	239	291	364	462	580	391.6	141.7	67.6	157.0	34.7	5.7
500	250	225	275	348	444	564	375.6	140.9	51.6	150.0	34.1	5.8
500	350	222	274	353	458	588	382.9	152.4	58.9	163.4	34.3	6.2
5000	100	373	395	425	451	470	423.9	38.2	99.9	107.0	36.1	1.5
5000	150	337	358	387	410	431	385.6	37.1	61.6	71.9	34.8	1.6
5000	200	317	340	368	392	415	366.7	36.9	42.7	56.4	34.0	1.6
5000	250	308	329	359	382	406	357.3	37.2	33.3	49.9	33.7	1.7
5000	350	298	320	350	374	399	349.0	38.3	25.0	45.7	33.4	1.8
100	adapt	145	213	328	541	858	470.7	541.1	146.7	560.5	35.8	14.0
250	adapt	166	217	302	421	567	345.7	192.8	21.7	194.0	32.4	7.8
500	adapt	190	234	302	383	482	322.7	123.6	-1.3	123.6	31.9	5.5

Compare tables 3.3.6, 3.4.7, 3.5.6, 3.5.24, 3.6.9.

Tabelle 3.5.7: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the quantile function for model 7.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	557	830	1298	2090	3319	1736.2	1631.2	531.2	1715.3	65.6	23.1
100	150	577	896	1470	2509	4263	2115.1	2246.0	910.1	2423.2	71.2	28.5
100	200	507	793	1299	2243	3775	1890.1	2075.9	685.1	2185.9	67.4	27.5
100	250	520	833	1403	2503	4492	2177.3	2657.0	972.3	2829.1	71.2	31.9
100	350	491	807	1373	2553	4744	2270.4	3104.1	1065.4	3281.5	71.8	34.7
250	100	914	1171	1586	2147	2842	1771.5	856.2	566.5	1026.5	69.1	14.7
250	150	803	1054	1446	1999	2730	1650.1	871.9	445.1	978.9	66.8	15.5
250	200	755	998	1377	1942	2707	1598.1	895.3	393.1	977.7	65.7	16.1
250	250	683	906	1266	1796	2489	1473.1	862.6	268.1	903.2	63.2	16.0
250	350	688	930	1328	1915	2731	1574.2	996.7	369.2	1062.8	64.9	17.7
500	100	1047	1252	1533	1910	2340	1629.8	527.5	424.8	677.2	67.4	9.8
500	150	951	1152	1432	1806	2246	1530.9	530.0	325.9	622.2	65.4	10.2
500	200	886	1079	1357	1720	2165	1458.9	529.0	253.9	586.7	64.0	10.5
500	250	837	1023	1297	1654	2112	1399.1	525.9	194.1	560.6	62.7	10.7
500	350	828	1021	1312	1697	2192	1426.4	568.9	221.4	610.4	63.2	11.5
5000	100	1391	1469	1581	1681	1751	1578.9	142.8	373.9	400.1	67.1	2.8
5000	150	1254	1331	1442	1528	1608	1436.0	138.6	231.0	269.3	64.3	2.9
5000	200	1181	1267	1372	1461	1546	1365.3	138.1	160.3	211.4	62.8	3.0
5000	250	1147	1227	1333	1424	1515	1330.6	138.9	125.6	187.2	62.1	3.1
5000	350	1112	1190	1306	1390	1489	1299.5	143.3	94.5	171.5	61.5	3.3
100	adapt	480	732	1237	2194	4080	2055.5	3002.2	850.5	3119.7	68.4	34.1
250	adapt	609	829	1168	1683	2370	1380.6	853.9	175.6	871.5	61.1	16.3
500	adapt	722	906	1176	1535	2027	1293.6	551.5	88.6	558.5	60.4	11.7

Compare tables 3.3.7, 3.4.8, 3.5.7, 3.5.25, 3.6.10.

Tabelle 3.5.8: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the quantile function for model 8.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	20	28	42	68	111	58.2	61.5	22.2	65.4	13.5	4.7
100	150	20	30	48	81	144	70.9	87.4	34.9	94.1	14.5	5.8
100	200	17	25	41	70	123	61.5	80.8	25.5	84.8	13.7	5.6
100	250	17	27	44	80	148	71.5	103.7	35.5	109.6	14.5	6.5
100	350	16	26	43	81	154	74.0	118.6	38.0	124.5	14.6	7.0
250	100	31	40	53	72	95	59.2	27.7	23.2	36.1	13.9	2.8
250	150	27	34	46	65	86	52.8	26.6	16.8	31.4	13.3	2.9
250	200	24	32	44	62	84	50.2	26.8	14.2	30.3	13.1	3.0
250	250	21	28	39	55	76	44.9	24.6	8.9	26.2	12.6	3.0
250	350	22	29	41	59	85	48.4	28.7	12.4	31.3	12.9	3.3
500	100	35	42	51	63	77	54.4	17.8	18.4	25.6	13.5	1.9
500	150	31	37	46	58	71	49.4	17.4	13.4	21.9	13.1	2.0
500	200	28	34	43	54	67	46.0	16.9	10.0	19.6	12.8	2.1
500	250	26	32	40	51	64	43.4	16.6	7.4	18.2	12.5	2.1
500	350	26	32	41	52	67	44.2	18.0	8.2	19.8	12.6	2.3
5000	100	46	49	52	56	60	52.8	5.3	16.8	17.6	13.4	0.6
5000	150	40	43	46	49	52	45.9	4.9	9.9	11.1	12.8	0.6
5000	200	37	40	43	46	49	42.8	4.8	6.8	8.3	12.5	0.6
5000	250	35	38	41	44	47	41.3	4.8	5.3	7.1	12.4	0.7
5000	350	34	37	40	43	46	39.9	4.8	3.9	6.2	12.2	0.7
100	adapt	18	25	40	66	108	57.4	75.6	21.4	78.5	14.0	5.0
250	adapt	17	22	30	41	55	33.8	17.1	-2.2	17.2	11.6	2.4
500	adapt	20	24	30	38	47	31.9	11.6	-4.1	12.3	11.3	1.8

Compare tables 3.3.8, 3.4.9, 3.5.8, 3.5.26, 3.6.11.

Tabelle 3.5.9: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the quantile function for model 9.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	79	115	178	283	440	236.7	227.3	74.7	239.2	25.4	8.8
100	150	81	123	202	337	568	287.6	313.3	125.6	337.5	27.5	10.9
100	200	71	109	176	300	501	256.3	289.6	94.3	304.5	26.1	10.6
100	250	72	114	190	339	605	295.5	369.3	133.5	392.7	27.5	12.3
100	350	68	109	186	345	637	307.9	428.3	145.9	452.5	27.8	13.4
250	100	126	160	215	290	384	240.6	114.4	78.6	138.8	26.4	5.5
250	150	110	144	197	269	369	223.4	116.8	61.4	131.9	25.7	5.9
250	200	103	136	187	261	363	215.9	119.7	53.9	131.3	25.3	6.1
250	250	93	124	171	240	335	198.6	114.9	36.6	120.6	24.5	6.1
250	350	94	127	180	257	365	212.3	133.1	50.3	142.3	25.1	6.7
500	100	143	171	208	259	315	221.1	70.3	59.1	91.8	25.8	3.7
500	150	130	156	194	244	302	207.0	70.8	45.0	83.9	25.1	3.9
500	200	121	146	184	233	291	197.0	70.6	35.0	78.8	24.7	4.0
500	250	114	138	175	223	282	188.7	70.3	26.7	75.2	24.3	4.1
500	350	112	138	178	230	294	192.3	76.0	30.3	81.9	24.4	4.4
5000	100	189	200	215	227	238	214.5	19.1	52.5	55.8	25.6	1.1
5000	150	170	181	195	206	217	194.3	18.4	32.3	37.2	24.7	1.1
5000	200	160	171	185	197	208	184.5	18.5	22.5	29.1	24.2	1.2
5000	250	155	165	180	192	204	179.6	18.5	17.6	25.5	24.0	1.2
5000	350	150	161	175	187	201	175.3	19.1	13.3	23.3	23.8	1.2
100	adapt	81	114	178	279	428	236.6	233.9	74.6	245.5	26.1	9.0
250	adapt	80	102	134	183	235	150.5	70.3	-11.5	71.3	22.0	4.5
500	adapt	88	105	128	158	191	135.5	44.1	-26.5	51.5	21.2	3.0

Compare tables 3.3.9, 3.4.10, 3.5.9, 3.5.27, 3.6.12.

3.5.8.2 Results for Bernstein-Durrmeyer polynomial approach

We can see that the Bernstein-Durrmeyer polynomial estimator for the quantile function (Tables 3.5.10–3.5.18) is better than the Bernstein polynomial estimator for the quantile function but is not better than the kernel density estimator for the purpose of estimating sampling plans.

Tabelle 3.5.10: Characteristics of distributions of n_m and c_m using Bernstein-Durrmeyer polynomial for the quantile function for model 1.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	31	40	56	77	105	63.5	34.7	-1.5	34.7	13.7	3.0
100	150	40	56	83	125	188	103.3	77.4	38.3	86.3	16.8	4.9
100	200	46	68	107	175	287	149.4	153.2	84.4	174.9	19.5	7.3
100	250	47	70	115	200	345	175.4	220.4	110.4	246.5	20.6	8.8
100	350	42	64	107	191	345	171.3	230.8	106.3	254.1	20.2	9.2
250	100	43	52	65	82	102	69.8	25.4	4.8	25.8	14.8	2.3
250	150	45	56	71	93	120	78.5	33.2	13.5	35.8	15.6	2.8
250	200	46	59	77	103	138	86.4	41.2	21.4	46.4	16.3	3.3
250	250	46	60	80	108	147	90.5	47.0	25.5	53.4	16.6	3.6
250	350	44	59	80	110	155	92.6	53.1	27.6	59.8	16.7	4.0
500	100	51	59	69	83	97	72.4	18.7	7.4	20.1	15.3	1.7
500	150	50	59	70	85	102	73.5	21.1	8.5	22.7	15.4	1.9
500	200	50	59	72	88	108	75.9	23.6	10.9	26.0	15.6	2.1
500	250	49	59	73	90	112	77.3	25.4	12.3	28.2	15.8	2.3
500	350	48	58	73	92	115	78.0	27.8	13.0	30.7	15.8	2.5
5000	100	68	71	75	79	83	75.3	6.0	10.3	11.9	15.8	0.6
5000	150	63	66	70	74	78	70.2	6.0	5.2	7.9	15.3	0.6
5000	200	61	64	68	72	77	68.7	6.2	3.7	7.2	15.2	0.6
5000	250	60	63	68	72	76	68.1	6.5	3.1	7.2	15.1	0.7
5000	350	60	63	67	71	76	67.5	6.9	2.5	7.4	15.1	0.7
100	adapt	31	40	53	71	93	58.7	27.2	-6.3	27.9	13.4	2.4
250	adapt	45	54	66	83	104	71.6	25.3	6.6	26.2	15.0	2.2
500	adapt	51	59	70	85	99	73.3	19.6	8.3	21.3	15.4	1.8

Compare tables 3.3.1, 3.4.2, 3.5.1, 3.5.19, 3.6.4.

Tabelle 3.5.11: Characteristics of distributions of n_m and c_m using Bernstein-Durrmeyer polynomial for the quantile function for model 2.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	6	10	21	42	70	31.0	30.0	-72.0	78.0	13.5	5.7
100	150	6	15	39	93	159	68.2	84.2	-34.8	91.1	19.0	10.1
100	200	6	20	65	164	305	131.4	223.0	28.4	224.8	24.9	15.8
100	250	7	25	84	211	410	185.0	413.2	82.0	421.2	28.7	19.9
100	350	7	30	94	213	419	187.9	378.7	84.9	388.1	29.9	19.8
250	100	11	18	30	48	71	36.4	25.7	-66.6	71.4	15.7	4.9
250	150	13	24	45	76	114	56.2	43.3	-46.8	63.7	19.6	6.9
250	200	16	32	62	106	155	76.7	60.6	-26.3	66.1	22.9	8.3
250	250	19	39	74	126	185	92.1	73.0	-10.9	73.8	25.3	9.2
250	350	24	50	89	145	213	108.6	84.4	5.6	84.6	27.8	9.9
500	100	17	23	34	49	66	38.6	21.2	-64.4	67.8	16.7	4.0
500	150	21	31	48	71	98	55.1	32.1	-47.9	57.7	20.3	5.3
500	200	26	40	63	92	125	70.7	40.8	-32.3	52.1	23.2	6.1
500	250	31	49	75	106	143	82.7	46.5	-20.3	50.7	25.3	6.4
500	350	39	61	89	122	162	96.9	51.7	-6.1	52.1	27.7	6.6
5000	100	31	34	39	43	48	38.9	6.8	-64.1	64.4	17.5	1.4
5000	150	40	46	52	59	65	52.5	9.9	-50.5	51.5	20.8	1.8
5000	200	49	57	65	73	80	65.1	12.0	-37.9	39.7	23.5	1.9
5000	250	58	66	75	84	91	75.2	13.1	-27.8	30.7	25.5	2.0
5000	350	70	79	88	96	104	87.7	13.8	-15.3	20.6	27.7	1.9
100	adapt	6	10	22	72	162	71.5	208.3	-31.5	210.6	17.1	13.0
250	adapt	17	34	68	119	180	87.3	73.9	-15.7	75.5	24.3	9.5
500	adapt	38	60	88	121	164	96.0	52.9	-7.0	53.3	27.6	6.9

Compare tables 3.3.2, 3.4.3, 3.5.2, 3.5.20, 3.6.5.

Tabelle 3.5.12: Characteristics of distributions of n_m and c_m using Bernstein-Durrmeyer polynomial for the quantile function for model 3.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	83	111	157	221	306	180.0	98.3	-29.0	102.5	15.9	3.4
100	150	115	164	249	383	586	313.3	232.8	104.3	255.1	20.2	5.8
100	200	137	209	338	566	966	481.0	497.8	272.0	567.3	24.0	9.1
100	250	140	219	372	656	1191	581.1	760.8	372.1	846.9	25.7	11.3
100	350	125	199	342	624	1148	559.3	794.9	350.3	868.6	25.0	11.6
250	100	125	156	197	252	318	212.3	81.1	3.3	81.1	17.7	2.8
250	150	134	171	223	295	380	245.1	107.3	36.1	113.2	18.8	3.4
250	200	140	182	243	333	444	274.1	133.8	65.1	148.8	19.8	4.1
250	250	141	186	253	351	483	289.2	152.1	80.2	171.9	20.2	4.5
250	350	136	184	254	360	507	296.7	170.0	87.7	191.3	20.4	4.9
500	100	156	182	217	260	311	226.4	62.4	17.4	64.8	18.5	2.1
500	150	154	183	221	270	327	232.9	70.9	23.9	74.8	18.7	2.4
500	200	155	185	229	284	346	242.3	79.7	33.3	86.3	19.0	2.7
500	250	154	186	232	291	359	247.5	86.1	38.5	94.3	19.2	2.9
500	350	150	184	233	297	371	250.4	93.8	41.4	102.5	19.3	3.2
5000	100	214	224	240	255	269	240.8	22.6	31.8	39.0	19.2	0.8
5000	150	196	208	223	239	254	224.4	22.4	15.4	27.2	18.7	0.8
5000	200	192	203	219	234	250	219.8	23.0	10.8	25.4	18.5	0.9
5000	250	189	201	217	233	249	217.8	23.7	8.8	25.2	18.4	0.9
5000	350	186	199	215	231	249	215.9	24.8	6.9	25.7	18.4	0.9
100	adapt	81	105	143	197	262	162.7	122.3	-46.3	130.8	15.3	3.1
250	adapt	125	155	199	257	324	215.6	87.4	6.6	87.7	17.8	2.9
500	adapt	154	184	223	274	328	234.1	71.7	25.1	76.0	18.7	2.4

Compare tables 3.3.3, 3.4.4, 3.5.3, 3.5.21, 3.6.6.

Tabelle 3.5.13: Characteristics of distributions of n_m and c_m using Bernstein-Durrmeyer polynomial for the quantile function for model 4.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	55	71	97	130	169	106.2	50.0	-61.8	79.5	18.4	3.6
100	150	81	114	168	251	360	202.4	134.9	34.4	139.2	24.5	6.7
100	200	103	154	244	403	646	334.4	320.7	166.4	361.2	30.3	11.1
100	250	108	166	276	481	835	417.2	503.4	249.2	561.7	33.0	14.1
100	350	96	151	255	456	810	396.2	500.0	228.2	549.6	32.1	14.3
250	100	89	108	133	168	205	141.3	47.2	-26.7	54.2	21.7	3.1
250	150	101	126	161	210	262	174.0	68.2	6.0	68.5	23.9	4.1
250	200	109	138	182	244	316	201.8	90.2	33.8	96.3	25.6	5.0
250	250	111	144	193	263	350	217.0	105.9	49.0	116.7	26.4	5.6
250	350	109	143	196	275	376	225.5	121.7	57.5	134.6	26.8	6.3
500	100	113	130	153	182	211	158.7	39.7	-9.3	40.8	23.2	2.5
500	150	117	138	164	199	236	172.0	48.1	4.0	48.3	24.1	3.0
500	200	121	144	174	214	257	183.6	56.0	15.6	58.1	24.9	3.4
500	250	122	146	179	223	272	190.0	61.6	22.0	65.4	25.3	3.7
500	350	121	146	181	229	286	194.4	68.1	26.4	73.0	25.6	4.0
5000	100	161	169	178	187	196	178.2	13.6	10.2	17.0	24.8	0.9
5000	150	155	163	173	182	191	172.7	14.2	4.7	14.9	24.5	0.9
5000	200	153	161	172	182	190	171.7	14.9	3.7	15.4	24.5	1.0
5000	250	151	160	172	182	191	171.4	15.6	3.4	16.0	24.6	1.0
5000	350	150	159	171	183	192	171.2	16.7	3.2	17.0	24.6	1.1
100	adapt	48	61	82	112	163	108.0	249.7	-60.0	256.7	17.7	6.1
250	adapt	95	122	160	217	283	179.4	84.5	11.4	85.2	24.1	4.9
500	adapt	121	144	176	219	269	187.6	62.3	19.6	65.3	25.2	3.7

Compare tables 3.3.4, 3.4.5, 3.5.4, 3.5.22, 3.6.7.

Tabelle 3.5.14: Characteristics of distributions of n_m and c_m using Bernstein-Durrmeyer polynomial for the quantile function for model 5.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	129	169	224	288	361	236.2	96.2	-371.8	384.0	26.3	4.4
100	150	246	346	498	724	999	576.9	342.2	-31.1	343.6	39.6	9.8
100	200	368	562	902	1504	2426	1249.6	1233.6	641.6	1390.4	55.6	20.8
100	250	412	657	1113	2018	3645	1825.4	2717.0	1217.4	2977.0	64.3	30.7
100	350	358	569	966	1776	3186	1568.4	2200.1	960.4	2400.4	60.2	28.2
250	100	254	309	378	466	558	394.8	121.9	-213.2	245.6	34.4	4.4
250	150	327	404	510	657	813	547.9	203.0	-60.1	211.7	40.2	6.5
250	200	379	478	628	839	1077	691.8	302.1	83.8	313.5	44.8	8.6
250	250	398	514	691	941	1250	776.0	378.0	168.0	413.6	47.2	10.1
250	350	395	520	713	1005	1374	821.4	446.7	213.4	495.0	48.3	11.5
500	100	356	411	482	566	654	496.4	119.3	-111.6	163.3	38.7	4.0
500	150	398	466	553	666	783	576.6	156.3	-31.4	159.4	41.7	5.0
500	200	426	506	612	749	901	642.6	193.3	34.6	196.4	43.9	5.9
500	250	438	523	643	797	972	679.7	219.6	71.7	231.0	45.1	6.5
500	350	435	527	658	829	1035	704.2	247.3	96.2	265.3	45.8	7.3
5000	100	573	596	631	663	697	631.9	48.9	23.9	54.3	44.0	1.5
5000	150	557	586	621	654	689	620.5	51.2	12.5	52.6	43.8	1.7
5000	200	552	581	622	655	687	619.4	54.2	11.4	55.3	43.8	1.8
5000	250	548	578	620	660	690	620.0	56.8	12.0	58.0	43.8	1.9
5000	350	542	576	621	663	697	620.3	60.8	12.3	61.9	43.9	2.0
100	adapt	286	429	681	1182	2137	1055.2	1321.9	447.2	1395.2	50.1	21.7
250	adapt	387	509	696	975	1307	796.3	428.9	188.3	468.3	47.6	11.1
500	adapt	417	514	655	836	1064	707.3	269.4	99.3	287.0	45.9	7.9

Compare tables 3.3.5, 3.4.6, 3.5.5, 3.5.23, 3.6.8.

Tabelle 3.5.15: Characteristics of distributions of n_m and c_m using Bernstein-Durrmeyer polynomial for the quantile function for model 6.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	88	115	153	202	258	165.1	71.7	-158.9	174.3	22.2	4.0
100	150	146	205	297	439	619	352.1	221.6	28.1	223.4	31.2	8.2
100	200	198	298	474	787	1258	653.1	632.0	329.1	712.5	40.7	15.0
100	250	213	332	555	984	1733	866.4	1126.4	542.4	1250.1	45.4	20.3
100	350	188	296	501	905	1616	791.7	1038.3	467.7	1138.7	43.5	19.8
250	100	155	188	232	288	349	243.8	78.2	-80.2	112.0	27.3	3.7
250	150	186	230	293	379	474	316.0	120.7	-8.0	120.9	30.9	5.1
250	200	207	262	344	460	593	380.0	168.0	56.0	177.1	33.6	6.5
250	250	214	275	370	504	669	415.9	202.7	91.9	222.5	35.0	7.5
250	350	210	276	378	533	728	435.6	235.8	111.6	260.8	35.7	8.4
500	100	205	237	278	329	381	287.7	70.5	-36.3	79.2	29.9	3.2
500	150	220	258	307	371	437	320.9	88.4	-3.1	88.5	31.6	3.8
500	200	231	274	332	407	489	348.7	105.6	24.7	108.5	32.9	4.4
500	250	235	280	344	427	522	364.2	117.8	40.2	124.5	33.6	4.9
500	350	232	281	350	441	550	374.6	131.2	50.6	140.6	34.0	5.4
5000	100	309	322	340	357	375	340.7	26.1	16.7	30.9	32.8	1.1
5000	150	298	313	332	350	367	331.7	27.2	7.7	28.3	32.5	1.2
5000	200	294	310	331	349	366	330.2	28.8	6.2	29.4	32.5	1.3
5000	250	292	308	330	351	368	330.1	30.1	6.1	30.7	32.5	1.4
5000	350	288	306	330	352	371	330.0	32.2	6.0	32.7	32.6	1.5
100	adapt	76	102	156	269	477	279.4	607.2	-44.6	608.7	25.5	12.9
250	adapt	199	252	332	455	588	374.0	176.7	50.0	183.6	33.3	6.8
500	adapt	231	277	342	429	527	364.5	123.6	40.5	130.0	33.6	5.1

Compare tables 3.3.6, 3.4.7, 3.5.6, 3.5.24, 3.6.9.

Tabelle 3.5.16: Characteristics of distributions of n_m and c_m using Bernstein-Durrmeyer polynomial for the quantile function for model 7.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	192	252	326	412	501	338.3	126.4	-866.7	875.9	30.8	4.6
100	150	428	599	846	1203	1613	957.1	523.4	-247.9	579.1	50.0	11.5
100	200	732	1128	1813	3041	4956	2533.9	2555.2	1328.9	2879.9	76.9	29.3
100	250	860	1398	2425	4520	8527	4352.4	8493.7	3147.4	9057.3	94.2	50.6
100	350	728	1162	2002	3724	6802	3350.3	5242.6	2145.3	5664.1	84.9	41.8
250	100	425	514	624	760	897	646.2	187.8	-558.8	589.5	42.8	5.1
250	150	595	728	913	1162	1425	972.7	343.2	-232.3	414.4	52.1	8.0
250	200	724	912	1194	1585	2027	1308.8	558.6	103.8	568.1	59.9	11.3
250	250	781	1008	1352	1839	2441	1518.3	735.1	313.3	799.0	64.1	13.7
250	350	783	1032	1414	1992	2719	1631.1	889.5	426.1	986.3	66.1	15.8
500	100	637	733	856	998	1146	878.2	203.4	-326.8	384.9	50.1	4.9
500	150	747	872	1030	1237	1448	1072.4	282.5	-132.6	312.1	55.3	6.4
500	200	823	976	1181	1441	1728	1236.1	366.6	31.1	367.9	59.2	7.8
500	250	859	1024	1261	1560	1905	1330.8	427.4	125.8	445.5	61.3	8.8
500	350	860	1043	1305	1641	2047	1393.1	488.9	188.1	523.8	62.6	9.9
5000	100	1118	1160	1230	1292	1357	1230.4	95.4	25.4	98.6	59.7	2.1
5000	150	1097	1152	1224	1286	1356	1219.6	100.4	14.6	101.3	59.5	2.3
5000	200	1091	1145	1227	1292	1358	1222.2	106.9	17.2	108.1	59.7	2.4
5000	250	1083	1144	1228	1304	1363	1225.7	112.4	20.7	114.1	59.8	2.6
5000	350	1074	1137	1230	1310	1378	1227.9	120.4	22.9	122.4	59.9	2.8
100	adapt	579	873	1481	2676	4725	2253.0	2529.6	1048.0	2737.5	71.6	30.9
250	adapt	767	1007	1385	1948	2637	1590.8	875.3	385.8	956.4	65.3	15.6
500	adapt	824	1014	1299	1657	2123	1401.2	539.7	196.2	574.1	62.7	10.9

Compare tables 3.3.7, 3.4.8, 3.5.7, 3.5.25, 3.6.10.

Tabelle 3.5.17: Characteristics of distributions of n_m and c_m using Bernstein-Durrmeyer polynomial for the quantile function for model 8.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	19	25	34	48	69	40.1	22.8	4.1	23.2	11.4	2.5
100	150	24	33	48	73	113	61.0	46.7	25.0	53.0	13.6	4.0
100	200	26	39	60	98	165	83.8	86.4	47.8	98.7	15.3	5.6
100	250	26	40	64	109	194	96.4	121.7	60.4	135.9	16.1	6.7
100	350	24	36	60	106	195	95.7	132.8	59.7	145.6	15.9	7.1
250	100	26	31	38	48	60	41.0	14.4	5.0	15.2	12.0	1.8
250	150	26	32	41	54	68	44.7	17.9	8.7	19.9	12.5	2.2
250	200	26	33	44	58	76	48.2	21.3	12.2	24.6	12.9	2.5
250	250	26	33	45	61	80	50.0	23.8	14.0	27.6	13.1	2.7
250	350	25	33	45	62	84	50.9	26.4	14.9	30.3	13.2	3.0
500	100	30	34	41	48	56	42.1	10.8	6.1	12.4	12.3	1.4
500	150	29	34	40	48	58	42.1	12.0	6.1	13.4	12.4	1.5
500	200	28	34	41	50	60	42.9	13.3	6.9	15.0	12.5	1.7
500	250	28	33	41	50	62	43.4	14.3	7.4	16.1	12.5	1.8
500	350	27	33	41	51	64	43.7	15.6	7.7	17.4	12.6	2.0
5000	100	38	40	42	45	47	42.7	3.7	6.7	7.6	12.6	0.5
5000	150	35	37	39	42	44	39.5	3.7	3.5	5.1	12.2	0.5
5000	200	34	36	38	41	43	38.5	3.7	2.5	4.5	12.1	0.5
5000	250	33	35	38	41	43	38.0	3.8	2.0	4.3	12.0	0.6
5000	350	33	35	37	40	42	37.6	4.0	1.6	4.3	12.0	0.6
100	adapt	20	25	34	46	63	38.6	19.2	2.6	19.4	11.3	2.2
250	adapt	26	31	39	50	63	42.3	15.6	6.3	16.8	12.2	2.0
500	adapt	29	34	40	48	58	42.3	12.3	6.3	13.8	12.4	1.6

Compare tables 3.3.8, 3.4.9, 3.5.8, 3.5.26, 3.6.11.

Tabelle 3.5.18: Characteristics of distributions of n_m and c_m using Bernstein-Durrmeyer polynomial for the quantile function for model 9.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	58	74	99	133	174	109.6	50.4	-52.4	72.8	18.0	3.5
100	150	84	117	169	251	358	203.3	134.8	41.3	141.0	23.6	6.4
100	200	103	153	240	393	626	328.8	318.5	166.8	359.5	28.9	10.5
100	250	107	165	271	466	801	406.4	498.8	244.4	555.4	31.3	13.4
100	350	95	149	251	442	783	387.3	493.0	225.3	542.0	30.5	13.6
250	100	91	109	134	168	204	142.0	46.4	-20.0	50.5	20.9	3.0
250	150	101	125	159	206	258	171.7	66.6	9.7	67.3	22.8	3.9
250	200	108	136	178	238	308	197.4	87.4	35.4	94.2	24.3	4.7
250	250	109	141	188	255	340	211.2	102.3	49.2	113.5	25.0	5.3
250	350	107	140	191	266	365	219.0	117.5	57.0	130.5	25.4	6.0
500	100	113	130	152	180	209	157.8	38.7	-4.2	38.9	22.2	2.4
500	150	116	136	162	195	230	168.7	46.5	6.7	47.0	23.0	2.8
500	200	119	140	170	208	250	178.9	54.0	16.9	56.6	23.6	3.2
500	250	119	142	175	217	263	184.6	59.3	22.6	63.5	24.0	3.5
500	350	117	141	176	221	275	188.5	65.5	26.5	70.6	24.2	3.8
5000	100	159	167	175	184	192	175.4	13.1	13.4	18.7	23.6	0.8
5000	150	151	159	168	177	185	168.4	13.6	6.4	15.0	23.3	0.9
5000	200	149	157	167	177	184	166.8	14.3	4.8	15.1	23.2	0.9
5000	250	147	156	166	176	185	166.3	15.0	4.3	15.6	23.2	1.0
5000	350	145	154	166	177	187	165.9	16.0	3.9	16.5	23.2	1.1
100	adapt	53	65	84	108	133	97.9	222.7	-64.1	231.7	16.7	4.7
250	adapt	92	109	133	164	195	140.5	45.7	-21.5	50.5	20.7	2.9
500	adapt	114	131	156	184	214	161.2	42.1	-0.8	42.1	22.4	2.5

Compare tables 3.3.9, 3.4.10, 3.5.9, 3.5.27, 3.6.12.

3.5.8.3 Results for Bernstein polynomial approach (distribution function)

Tabelle 3.5.19: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the distribution function for model 1.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	19	24	31	40	51	33.4	13.0	-31.6	34.2	11.9	1.8
100	150	23	31	41	55	71	44.5	19.6	-20.5	28.4	13.0	2.3
100	200	25	34	48	65	85	52.3	25.2	-12.7	28.2	13.7	2.7
100	250	26	37	53	74	98	59.0	30.1	-6.0	30.7	14.3	3.1
100	350	27	39	58	84	115	66.2	37.9	1.2	37.9	14.8	3.6
250	100	22	25	30	35	41	30.6	7.6	-34.4	35.2	11.6	1.1
250	150	27	32	39	46	55	40.1	11.2	-24.9	27.3	12.6	1.4
250	200	30	36	44	54	65	46.4	14.2	-18.6	23.4	13.2	1.7
250	250	32	40	49	61	74	51.8	17.0	-13.2	21.5	13.7	1.9
250	350	33	42	53	68	84	56.5	20.6	-8.5	22.3	14.1	2.2
500	100	24	26	29	33	37	29.8	5.2	-35.2	35.6	11.5	0.8
500	150	29	33	38	43	49	38.7	7.7	-26.3	27.4	12.5	1.0
500	200	33	38	43	50	58	44.5	9.7	-20.5	22.7	13.1	1.2
500	250	36	41	48	57	65	49.4	11.5	-15.6	19.4	13.6	1.4
500	350	37	44	52	61	72	53.4	13.7	-11.6	18.0	13.9	1.6
5000	100	27	28	29	30	31	28.9	1.6	-36.1	36.2	11.4	0.2
5000	150	35	36	37	39	40	37.5	2.3	-27.5	27.6	12.4	0.3
5000	200	40	41	43	45	46	42.9	2.8	-22.1	22.3	12.9	0.4
5000	250	43	45	47	49	51	47.4	3.3	-17.6	17.9	13.4	0.4
5000	350	46	48	51	53	56	50.7	3.9	-14.3	14.8	13.7	0.5

Compare tables 3.3.1, 3.4.2, 3.5.1, 3.5.19, 3.6.4.

Tabelle 3.5.20: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the distribution function for model 2.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	3	6	9	14	18	10.2	5.8	-92.8	92.9	10.5	2.8
100	150	5	10	15	22	30	16.5	9.6	-86.5	87.1	12.6	3.4
100	200	7	13	20	29	40	22.1	13.4	-80.9	82.0	14.1	3.9
100	250	8	15	25	36	50	27.1	16.7	-75.9	77.7	15.3	4.3
100	350	11	20	32	48	67	36.1	23.0	-66.9	70.7	17.3	5.0
250	100	5	7	9	12	14	9.7	3.5	-93.3	93.3	10.8	1.6
250	150	9	11	15	19	23	15.5	5.7	-87.5	87.7	12.9	1.9
250	200	11	15	20	26	31	20.7	7.8	-82.3	82.7	14.5	2.2
250	250	14	18	24	31	38	25.3	9.7	-77.7	78.3	15.7	2.4
250	350	18	24	32	42	51	33.5	13.2	-69.5	70.8	17.7	2.8
500	100	7	8	9	11	13	9.6	2.5	-93.4	93.4	10.9	1.1
500	150	10	13	15	18	21	15.3	4.0	-87.7	87.8	13.0	1.3
500	200	14	17	20	24	28	20.3	5.4	-82.7	82.8	14.6	1.5
500	250	17	20	24	29	34	24.9	6.8	-78.1	78.4	15.8	1.6
500	350	22	26	32	38	45	32.8	9.2	-70.2	70.8	17.8	1.9
5000	100	8	9	9	10	10	9.4	0.8	-93.6	93.6	10.9	0.4
5000	150	13	14	15	16	16	14.8	1.2	-88.2	88.2	13.0	0.4
5000	200	18	19	20	21	22	19.8	1.6	-83.2	83.2	14.6	0.4
5000	250	21	23	24	26	27	24.2	2.1	-78.8	78.9	15.8	0.5
5000	350	28	30	32	34	35	31.8	2.8	-71.2	71.3	17.8	0.6

Compare tables 3.3.2, 3.4.3, 3.5.2, 3.5.20, 3.6.5.

Tabelle 3.5.21: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the distribution function for model 3.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	61	79	105	133	163	109.0	40.0	-100.0	107.7	14.8	2.0
100	150	72	97	132	173	219	139.8	58.3	-69.2	90.4	15.9	2.5
100	200	77	108	151	203	265	162.4	74.3	-46.6	87.6	16.7	3.0
100	250	80	114	163	225	297	177.7	86.6	-31.3	92.1	17.1	3.4
100	350	83	122	182	261	352	203.4	110.2	-5.6	110.4	17.9	4.0
250	100	73	86	103	120	138	104.4	25.2	-104.6	107.6	14.6	1.3
250	150	88	105	128	154	179	131.2	35.8	-77.8	85.7	15.6	1.6
250	200	96	117	145	178	210	149.9	44.9	-59.1	74.2	16.3	1.9
250	250	101	124	155	193	230	161.5	51.5	-47.5	70.1	16.6	2.2
250	350	106	134	172	218	266	180.4	64.0	-28.6	70.1	17.2	2.6
500	100	81	90	101	114	126	102.3	17.5	-106.7	108.1	14.6	0.9
500	150	97	111	126	143	161	127.7	24.7	-81.3	85.0	15.5	1.2
500	200	107	124	142	164	187	145.1	30.8	-63.9	71.0	16.1	1.4
500	250	113	131	152	177	203	155.5	35.3	-53.5	64.1	16.5	1.5
500	350	120	141	167	199	231	172.1	43.4	-36.9	57.0	17.0	1.8
5000	100	93	97	101	105	109	100.9	6.1	-108.1	108.2	14.5	0.3
5000	150	114	119	125	130	135	124.6	8.0	-84.4	84.7	15.4	0.4
5000	200	128	134	141	147	154	140.7	9.9	-68.3	69.0	16.0	0.5
5000	250	135	143	150	158	165	150.2	11.2	-58.8	59.8	16.3	0.5
5000	350	147	156	164	174	183	164.9	13.7	-44.1	46.2	16.8	0.6

Compare tables 3.3.3, 3.4.4, 3.5.3, 3.5.21, 3.6.6.

Tabelle 3.5.22: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the distribution function for model 4.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	26	32	39	48	58	40.6	12.8	-127.4	128.0	13.9	1.6
100	150	37	46	58	72	87	60.4	20.1	-107.6	109.5	16.0	2.1
100	200	45	57	72	92	112	76.2	27.2	-91.8	95.7	17.5	2.5
100	250	51	66	87	110	135	90.5	33.5	-77.5	84.4	18.7	2.9
100	350	57	76	102	133	167	108.0	44.6	-60.0	74.7	20.0	3.5
250	100	30	34	38	44	50	39.1	7.8	-128.9	129.2	13.8	1.0
250	150	43	49	57	65	74	57.8	12.2	-110.2	110.9	15.9	1.3
250	200	52	61	71	82	94	72.4	16.4	-95.6	97.0	17.3	1.6
250	250	61	72	84	98	113	85.7	20.2	-82.3	84.7	18.5	1.8
250	350	69	82	98	117	136	100.4	26.3	-67.6	72.5	19.6	2.2
500	100	32	35	38	41	45	38.3	5.3	-129.7	129.8	13.8	0.7
500	150	46	51	56	62	68	56.6	8.5	-111.4	111.7	15.8	0.9
500	200	57	63	70	78	86	70.7	11.3	-97.3	97.9	17.2	1.1
500	250	66	74	83	92	102	83.6	14.0	-84.4	85.6	18.4	1.2
500	350	75	84	96	109	121	97.3	18.1	-70.7	72.9	19.5	1.5
5000	100	36	37	38	39	39	37.7	1.4	-130.3	130.3	13.7	0.2
5000	150	53	54	56	57	58	55.5	2.3	-112.5	112.5	15.8	0.3
5000	200	65	67	69	72	73	69.3	3.3	-98.7	98.7	17.1	0.3
5000	250	77	79	82	84	87	81.9	4.0	-86.1	86.2	18.3	0.4
5000	350	88	91	95	98	101	94.7	5.2	-73.3	73.4	19.3	0.5

Compare tables 3.3.4, 3.4.5, 3.5.4, 3.5.22, 3.6.7.

Tabelle 3.5.23: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the distribution function for model 5.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	39	49	62	76	88	63.1	19.4	-544.9	545.2	17.0	1.7
100	150	64	80	99	120	141	101.4	30.6	-506.6	507.5	20.3	2.0
100	200	86	107	133	161	189	135.6	41.3	-472.4	474.2	22.7	2.3
100	250	107	132	163	197	234	167.5	50.7	-440.5	443.4	24.7	2.5
100	350	133	167	210	260	312	217.6	71.6	-390.4	396.9	27.4	3.2
250	100	47	53	61	71	78	62.1	12.3	-545.9	546.1	17.0	1.1
250	150	75	87	99	112	125	99.7	19.1	-508.3	508.7	20.3	1.3
250	200	101	115	132	149	167	133.1	25.7	-474.9	475.6	22.7	1.4
250	250	125	142	162	183	205	163.7	31.2	-444.3	445.4	24.7	1.6
250	350	156	180	208	238	271	211.1	44.1	-396.9	399.4	27.4	2.0
500	100	50	54	61	69	73	61.5	9.0	-546.5	546.6	17.0	0.8
500	150	81	90	98	107	116	98.6	13.4	-509.4	509.6	20.3	0.9
500	200	108	119	131	143	155	131.6	17.9	-476.4	476.8	22.7	1.0
500	250	134	146	161	176	190	161.7	21.6	-446.3	446.8	24.7	1.1
500	350	168	185	207	228	248	207.7	30.8	-400.3	401.5	27.3	1.4
5000	100	56	58	60	63	65	60.4	3.2	-547.6	547.6	17.0	0.3
5000	150	93	95	97	100	102	97.6	3.6	-510.4	510.5	20.3	0.2
5000	200	123	128	131	134	136	130.5	5.1	-477.5	477.5	22.7	0.3
5000	250	151	155	160	165	168	159.9	6.7	-448.1	448.1	24.6	0.3
5000	350	193	199	204	211	216	204.7	9.0	-403.3	403.4	27.2	0.4

Compare tables 3.3.5, 3.4.6, 3.5.5, 3.5.23, 3.6.8.

Tabelle 3.5.24: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the distribution function for model 6.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	35	42	52	65	76	54.0	16.6	-270.0	270.5	15.6	1.7
100	150	52	65	81	99	118	83.4	25.8	-240.6	242.0	18.3	2.1
100	200	67	84	105	129	154	108.1	34.5	-215.9	218.7	20.2	2.4
100	250	80	101	126	155	186	130.4	42.3	-193.6	198.1	21.8	2.7
100	350	94	120	155	196	240	161.9	58.4	-162.1	172.3	23.7	3.4
250	100	40	45	52	60	66	52.6	10.4	-271.4	271.6	15.6	1.1
250	150	61	70	80	91	103	81.0	16.0	-243.0	243.5	18.3	1.3
250	200	78	90	103	118	133	104.6	21.3	-219.4	220.4	20.2	1.5
250	250	94	108	124	142	160	126.0	25.8	-198.0	199.7	21.7	1.7
250	350	110	129	151	176	203	154.0	35.5	-170.0	173.7	23.5	2.1
500	100	43	46	51	57	62	51.8	7.5	-272.2	272.3	15.6	0.8
500	150	65	72	79	87	94	79.8	11.2	-244.2	244.5	18.2	0.9
500	200	84	92	102	112	122	102.8	14.9	-221.2	221.7	20.1	1.1
500	250	101	111	123	135	147	123.7	17.9	-200.3	201.1	21.6	1.2
500	350	120	133	149	166	183	150.6	24.6	-173.4	175.2	23.4	1.5
5000	100	48	49	51	53	54	50.9	2.4	-273.1	273.1	15.5	0.3
5000	150	75	77	79	81	82	78.8	2.9	-245.2	245.2	18.2	0.2
5000	200	95	98	101	105	107	101.4	4.6	-222.6	222.7	20.0	0.3
5000	250	115	118	122	125	128	121.8	5.2	-202.2	202.2	21.5	0.3
5000	350	139	143	148	153	156	147.6	7.1	-176.4	176.5	23.3	0.4

Compare tables 3.3.6, 3.4.7, 3.5.6, 3.5.24, 3.6.9.

Tabelle 3.5.25: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the distribution function for model 7.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	46	58	73	85	103	72.8	21.8	-1132.2	1132.4	18.0	1.8
100	150	77	97	117	141	167	120.5	35.6	-1084.5	1085.1	21.8	2.0
100	200	106	133	162	195	229	165.8	49.0	-1039.2	1040.4	24.7	2.3
100	250	135	167	203	246	288	208.7	61.5	-996.3	998.2	27.2	2.5
100	350	179	224	277	337	399	284.2	87.5	-920.8	925.0	30.8	3.0
250	100	55	62	73	81	88	72.1	13.6	-1132.9	1133.0	18.2	1.1
250	150	91	104	118	133	147	119.1	22.0	-1085.9	1086.1	21.9	1.2
250	200	126	143	162	183	203	163.6	30.3	-1041.4	1041.9	24.9	1.4
250	250	158	179	204	230	255	205.5	37.9	-999.5	1000.3	27.3	1.5
250	350	210	241	277	313	351	278.8	54.3	-926.2	927.8	30.9	1.9
500	100	59	64	73	79	83	71.8	9.6	-1133.2	1133.3	18.2	0.8
500	150	99	107	118	129	137	118.3	15.3	-1086.7	1086.8	22.0	0.9
500	200	134	148	161	177	189	162.3	21.0	-1042.7	1042.9	24.9	1.0
500	250	170	185	203	221	238	203.7	26.4	-1001.3	1001.7	27.3	1.1
500	350	227	249	275	301	325	276.0	38.0	-929.0	929.8	30.9	1.3
5000	100	67	70	72	75	76	72.0	3.5	-1133.0	1133.0	18.3	0.3
5000	150	111	114	117	121	124	117.3	5.2	-1087.7	1087.7	22.0	0.3
5000	200	153	156	160	164	170	160.5	6.4	-1044.5	1044.5	24.8	0.3
5000	250	192	197	201	207	212	201.8	7.9	-1003.2	1003.2	27.3	0.3
5000	350	259	266	273	282	289	273.4	11.3	-931.6	931.7	30.9	0.4

Compare tables 3.3.7, 3.4.8, 3.5.7, 3.5.25, 3.6.10.

Tabelle 3.5.26: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the distribution function for model 8.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	12	15	21	27	36	22.4	10.0	-13.6	16.9	10.1	1.7
100	150	14	19	26	36	48	29.0	14.6	-7.0	16.2	11.0	2.2
100	200	14	21	29	42	58	33.6	18.4	-2.4	18.5	11.5	2.6
100	250	15	22	32	48	67	37.5	21.9	1.5	22.0	11.9	2.9
100	350	15	23	35	54	78	41.8	26.8	5.8	27.5	12.3	3.4
250	100	14	16	19	23	27	20.0	5.4	-16.0	16.9	9.8	1.1
250	150	16	20	24	30	35	25.3	7.7	-10.7	13.2	10.5	1.4
250	200	18	22	27	34	41	28.6	9.6	-7.4	12.1	10.9	1.6
250	250	18	23	30	38	47	31.5	11.4	-4.5	12.2	11.3	1.8
250	350	19	24	32	42	52	34.0	13.4	-2.0	13.5	11.6	2.0
500	100	15	17	19	22	24	19.4	3.7	-16.6	17.0	9.7	0.8
500	150	18	21	24	27	31	24.3	5.4	-11.7	12.9	10.4	1.0
500	200	19	23	27	31	36	27.4	6.7	-8.6	10.9	10.8	1.1
500	250	21	24	29	34	40	29.9	7.9	-6.1	9.9	11.1	1.3
500	350	21	26	31	37	44	32.0	9.1	-4.0	10.0	11.3	1.5
5000	100	17	18	19	19	20	18.6	1.3	-17.4	17.4	9.6	0.3
5000	150	21	22	23	24	25	23.2	1.7	-12.8	12.9	10.3	0.3
5000	200	24	25	26	27	29	25.9	2.1	-10.1	10.3	10.6	0.4
5000	250	25	27	28	30	31	28.2	2.4	-7.8	8.2	10.9	0.4
5000	350	27	28	30	32	33	29.9	2.8	-6.1	6.7	11.1	0.5

Compare tables 3.3.8, 3.4.9, 3.5.8, 3.5.26, 3.6.11.

Tabelle 3.5.27: Characteristics of distributions of n_m and c_m using Bernstein polynomial for the distribution function for model 9.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	100	29	35	43	54	64	45.3	13.9	-116.7	117.5	14.0	1.6
100	150	40	50	63	78	95	65.6	21.7	-96.4	98.8	15.9	2.1
100	200	47	61	78	99	120	81.5	28.8	-80.5	85.5	17.3	2.5
100	250	54	70	91	116	142	95.4	35.2	-66.6	75.3	18.3	2.9
100	350	59	79	106	139	175	112.7	47.1	-49.3	68.2	19.5	3.5
250	100	33	37	43	49	55	43.3	8.6	-118.7	119.0	13.9	1.0
250	150	46	53	61	71	80	62.4	13.3	-99.6	100.5	15.8	1.3
250	200	55	64	75	88	100	76.8	17.5	-85.2	87.0	17.0	1.6
250	250	63	75	88	103	118	89.5	21.3	-72.5	75.5	18.1	1.8
250	350	70	84	101	121	141	103.6	27.9	-58.4	64.7	19.1	2.2
500	100	35	38	42	46	50	42.4	6.0	-119.6	119.7	13.8	0.7
500	150	50	54	60	67	73	60.9	9.2	-101.1	101.5	15.7	0.9
500	200	60	66	74	82	91	74.8	12.1	-87.2	88.0	16.9	1.1
500	250	69	77	86	96	107	87.1	14.7	-74.9	76.4	17.9	1.3
500	350	76	86	98	112	126	100.1	19.2	-61.9	64.8	18.9	1.5
5000	100	39	40	42	43	44	41.7	1.9	-120.3	120.3	13.8	0.2
5000	150	57	58	60	61	63	59.8	2.6	-102.2	102.2	15.6	0.3
5000	200	69	71	73	76	78	73.2	3.6	-88.8	88.8	16.8	0.3
5000	250	80	82	85	88	90	85.1	4.2	-76.9	77.1	17.8	0.4
5000	350	90	93	97	101	104	97.2	5.5	-64.8	65.1	18.8	0.5

Compare tables 3.3.9, 3.4.10, 3.5.9, 3.5.27, 3.6.12.

3.5.9 Summary

In Table 3.5.28, we present simulated root mean squared deviations (RMSD) of the sampling plan size using the Bernstein estimator and the Bernstein-Durrmeyer estimator for the quintile function with adaptive selection of the degree for several models of the distribution of measurements. We can see that the Bernstein-Durrmeyer estimator is typically better than the Bernstein estimator, especially for the sample size $m = 100$.

Tabelle 3.5.28: RMSD of the distribution of the sampling plan size using the Bernstein estimator and the Bernstein-Durrmeyer estimator for the quintile function with adaptive selection of the degree for models 1-9.

model	m	B est	BD est
1	100	116.8	27.9
1	250	32.3	26.2
1	500	22.8	21.3
2	100	320.7	210.6
2	250	78.1	75.5
2	500	51.5	53.3
3	100	370.7	130.8
3	250	111	87.7
3	500	71.5	76.0
4	100	260.0	256.7
4	250	91.4	85.2
4	500	61.7	65.3
5	100	1317.4	1395.2
5	250	435.3	468.3
5	500	278.0	287.0
6	100	560.5	608.7
6	250	194.0	183.6
6	500	123.6	130.0
7	100	3119.7	2737.5
7	250	871.5	956.4
7	500	558.5	574.1
8	100	78.5	19.4
8	250	17.2	16.8
8	500	12.3	13.8
9	100	245.5	231.7
9	250	71.3	50.5
9	500	51.5	42.1

Kapitel 3.6

Estimators based on Singular Spectrum Analysis

The basic idea of an approach studied in the present chapter is to construct an artificial time series by evaluating the empirical distribution function at equidistant points, which is then smoothed by applying Singular Spectrum Analysis (SSA), developed in (Golyandina, Nekrutkin, Zhigljavsky, 2001), yielding a density estimator by taking the differences of the smoothed series. Roughly speaking, the SSA procedure extracts a smooth distribution function treating the empirical distribution function as a noisy function. In fact, the SSA procedure creates a data-adaptive filter depending on two parameters L and r, which jointly regulate filter properties. The parameter L is half of the length of the filter and corresponds to the number of rows of the trajectory matrix constructed from a series. The parameter r is the number of leading components of a singular value decomposition of the trajectory matrix and therefore equals the dimension of the matrix subspace used to determine the estimator. Consequently, r has exactly the same interpretation as the number of leading components in PCA and SVD and therefore can be regarded as the complexity of the filter (with fixed L).

3.6.1 A new approach to density estimation

Let $x_1, \ldots, x_m$ be a sample from a distribution having a density $p(x)$ and let $F(x)$ be the corresponding cumulative distribution function. The empirical distribution function

$$F_m(x) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{[x_i,\infty)}(x), \quad x \in \mathbb{R},$$

is the minimum-variance unbiased nonparametric estimator of $F(x)$. However, it cannot be used directly to estimate the density $p(x)$ by taking the derivative since $F_m(x)$ is a step function, and it is known that estimation can often be improved substantially by using smoothed but biased estimators of $F(x)$. To smooth the empirical distribution function, we propose to consider the values of $F_m(x)$ on an equidistant dense grid as a series and then apply the SSA procedure to smooth this series.

3.6.1.1 Outline of the approach

Let $t_1, \ldots, t_N$ be equidistant points such that $t_j - t_{j-1} = \delta$, $j = 2, \ldots, N$, $t_1 < \min_i x_i$ and $t_N > \max_i x_i$, where δ is a small positive constant. We then define the series

$$\mathcal{F}_m = (f_1, \ldots, f_N), \quad f_j = F_m(t_j),$$

$j = 1, \ldots, N$, which is nondecreasing and serves as a discrete version of $F_m(x)$, and the series

$$\mathcal{F}^* = (f_1^*, \ldots, f_N^*), \quad f_j^* = F(t_j),$$

$j = 1, \ldots, N$, which has to be estimated from the sample.

We now choose a certain positive integer L satisfying $1 \le L \le (N+1)/2$, which will regulate the smoothness of the proposed estimator and is a parameter of the following procedure adapted from SSA theory, cf. (10). If we consider the smoothing procedure from the viewpoint of filtration, L controls the filter bandwidth. Also, the parameter L has the sense of resolution, that is, the smaller L, the more refined and less stable smoothing is performed.

Since the empirical distribution function is nontrivial on the interval $[\min_i x_i, \max_i x_i]$, the natural upper bound for selection of the smoothing parameter L is

$$L_{\max} = \left\lfloor \frac{\max_i x_i - \min_i x_i}{2\delta} \right\rfloor.$$

Note that the series $\mathcal{F}_m$ may have any number of zeroes to the left and any number of ones to the right. Extension of the series $\mathcal{F}_m$ enables us to remove the boundary effects of the filters constructed in what follows.

The SSA algorithm for processing the series $\mathcal{F}_m$ now works as follows. Recall that a matrix $\mathbb{A} = (a_{i,j})$ is called *Hankel matrix*, if $a_{i,j}$ depends only on $i + j$, such that all diagonals are constant. The first step is to construct the $L \times K$ Hankel matrix $\mathbb{X} = (x_{i,j})$ with entries

$$x_{i,j} = f_{i+j-1}$$

corresponding to $\mathcal{F}_m$, where $K = N - L + 1$. That is, we embed the series $\mathcal{F}_m$ into the space of $L \times K$ matrices. The Singular Value Decomposition (SVD) of $\mathbb{X}$ yields the representation

$$\mathbb{X} = \sum_{i=1}^{L} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}, \tag{3.6.1.1}$$

where the eigenvalues λ_i are in nonincreasing order, U_i and V_i are the left and right singular vectors, respectively, $U_i^{\mathrm{T}} U_i = 1$, $i = 1, \ldots, L$. According to SSA theory, a few leading terms $\sqrt{\lambda_1} U_1 V_1^{\mathrm{T}}, \ldots, \sqrt{\lambda_r} U_r V_r^{\mathrm{T}}$ are of interest for the problem of smoothing (10, Section 1.3.2). Therefore, we define the matrix

$$\mathbb{X}^{(r)} = \left(x_{i,j}^{(r)}\right) = \sum_{i=1}^{r} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}. \tag{3.6.1.2}$$

The last step of the SSA algorithm is the Hankelization (averaging along anti-diagonals) of $\mathbb{X}^{(r)}$ with a subsequent operation that is opposite to embedding. Taking into consideration the specifics of cumulative distribution functions, we define

$$\hat{f}_j = \hat{f}_j(L, r) = \begin{cases} 0 & 1 \le j < L, \\ \dfrac{1}{L} \displaystyle\sum_{k=1}^{L} x_{k,j-k+1}^{(r)} & L \le j \le K, \\ 1 & K < j \le N, \end{cases}$$

Thus, we obtain the series

$$\hat{\mathcal{F}}_m = \hat{\mathcal{F}}_m(L, r) = \Big(\hat{f}_1(L, r), \ldots, \hat{f}_N(L, r)\Big),$$

which will be called the SSA estimator of $\mathcal{F}^*$. By construction, we have $\hat{\mathcal{F}}_m(L, r) = \mathcal{F}_m$ if $L = r$.

The series $\hat{\mathcal{F}}_m$ can be transformed to the SSA estimator of $F(x)$ using the linear interpolation as follows

$$\hat{F}_m(x) = \sum_{j=2}^{N} \left(\hat{f}_{j-1} + (\hat{f}_j - \hat{f}_{j-1}) \frac{x - t_{j-1}}{t_j - t_{j-1}} \right) \mathbf{1}_{[t_{j-1}, t_j)}(x) + \mathbf{1}_{[t_N, \infty)}(x), \qquad (3.6.1.3)$$

$x \in \mathbb{R}$. Note that $\hat{F}_m(x)$ is actually a linear spline on $[t_1, t_N]$ with knots $(t_1, \hat{f}_1), \ldots, (t_N, \hat{f}_N)$. Moreover, we can consider the derivative of $\hat{F}_m(x)$ from the left as an estimator of the density. In particular,

$$\hat{p} = \big(\hat{p}_1, \ldots, \hat{p}_N\big), \quad \hat{p}_j = \big(\hat{f}_j - \hat{f}_{j-1}\big)/\delta,$$

$j = 2, \ldots, N$, gives an estimator of the density $p(x)$ at the equidistant points.

Let us now describe how to select the points $t_1, \ldots, t_N$ used to transform the empirical distribution function to a series. To do this, we first notice that $V_i = \mathbb{X}^{\mathrm{T}} U_i / \sqrt{\lambda_i}$ and $\mathbb{X}^{(r)} = (U_1 U_1^{\mathrm{T}} + \ldots + U_r U_r^{\mathrm{T}})\mathbb{X}$, and, consequently, $\hat{f}_j$ enables the representation

$$\hat{f}_j = \sum_{i=1}^{L} \sum_{l=1}^{L} (u_{1,i} u_{1,l} + \ldots + u_{r,i} u_{r,l}) f_{j+i-l}/L \qquad (3.6.1.4)$$

for $j = L, \ldots, K$, where $(u_{l,1}, \ldots, u_{l,L})^{\mathrm{T}} = U_l$. Thus, formula (3.6.1.4) means that the SSA procedure creates a data-adaptive filter of size $2L-1$, where r can be interpreted as the complexity of the SSA filter and L controls the smoothness of filtered series. Therefore, we should determine the points $t_1, \ldots, t_N$ such that the series $\mathcal{F}_m$ contains $2L$ zeroes at the beginning and $2L$ ones at the end.

We suggest to choose δ to satisfy the inequality

$$\mathbf{P}(\xi \in [x, x+\delta]) \approx p(x)\delta < 1/m,$$

where the density $p(x)$ is assumed to be smooth enough. Then

$$\delta \le \frac{1}{m \max_x p(x)} = \frac{\sigma}{m \max_x p_{\mathrm{st}}(x)}, \qquad (3.6.1.5)$$

where $p_{\mathrm{st}}(x)$ is the standardized density having zero mean and unit variance, $p_{\mathrm{st}}(x) = p((x-a)/\sigma)$. This inequality provides the reasonable correspondence between the sample size and the interpolation step δ. In following examples we consider $\delta = 0.01\sigma$ that approximately satisfies inequality (3.6.1.5).

Summarizing above arguments, we define points $t_1, \ldots, t_N$ by

$$t_j = \delta(j - 2L) + \min_{i=1,\ldots,m} x_i,$$

$j = 1, \ldots, N$, $N = 2L_{\max} + 4L$.

Note that the specifics of analyzed series (namely, monotonicity and slow variation) allow us to consider only a few leading components in (3.6.1.1) for smoothing. Contrary to the problems of signal extraction with fixed L and estimated r, we fix r and adjust L to control the smoothness.

For clarity, we now demonstrate the influence of the parameter L on the SSA estimator $\hat{\mathcal{F}}_m(L, r)$ with $r = 1$ visually.

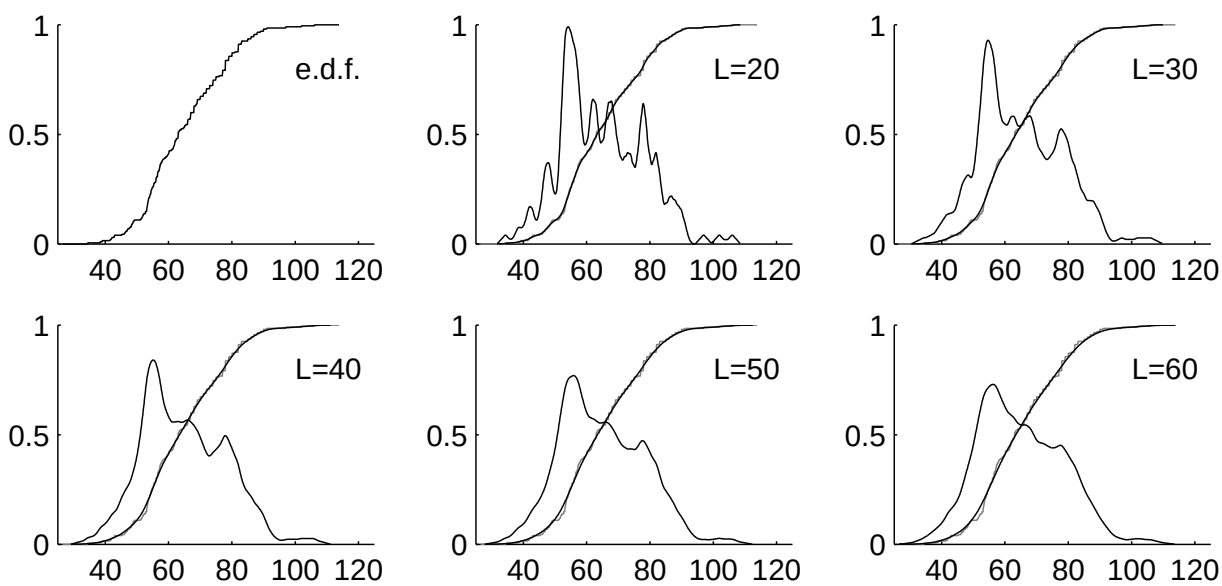


Abbildung 3.6.1: The empirical distribution function for measurements of Lean Body Mass from the Australian Institute of Sport data and the SSA estimators of the distribution function and density for $L = 20, 30, \ldots, 60$.

Example 3.6.1.1. *Let us consider 202 measurements of Lean Body Mass from the Australian Institute of Sport, which were studied by (22), amongst others. Figure 3.6.1 depicts the empirical distribution function for these measurements (which give $\delta = 0.13$ and $L_{\max} = 274$) and the SSA estimators for $L = 20, 30, \ldots, 60$. The densities in the figure are re-scaled for convenience of visual impression. It can be seen that the SSA estimator $\hat{\mathcal{F}}_m(L, 1)$ becomes smoother and the number of modes of the corresponding density decreases*

as the parameter L increases. Comparing to kernel smoothing, small values of L correspond to small bandwidths, and large values of L to large bandwidths, and both methods lead to relatively similar results for this data set.

Example 3.6.1.2. *As a second example, we consider 22 observations of silica in chondrite meteors (12). In Figure 3.6.2, we depict the empirical distribution function for these observations (which give $\delta = 0.043$ and $L_{\max} = 163$) and the SSA estimators for $L = 20, 40, \ldots, 100$. Notice that the density has three modes for a large range of values of L. In general, the SSA estimator may be used to investigate multimodality in a way as discussed in (39).*

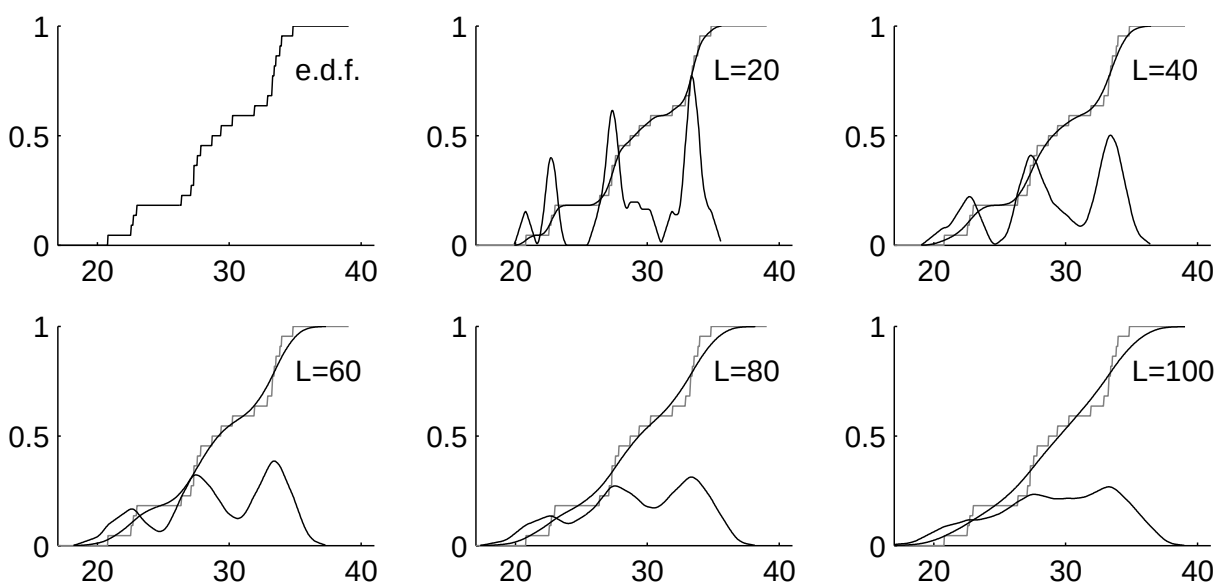


Abbildung 3.6.2: The empirical distribution function for observations of chondrite meteors and the SSA estimators of the distribution function and density for $L = 20, 40, \ldots, 100$.

3.6.1.2 Specific SSA estimators and bias correction

It is worth to study the SSA estimator for selected values of r, the number of leading components used in (3.6.1.2), in greater detail and to develop a procedure to ensure that the final estimator is monotone.

The SSA estimator, as any smoothing procedure, suffers from estimation bias. Therefore, we propose an effective approach for bias correction, where the bias is estimated by

repeated applications of the SSA filter. A simulation study evaluates the performance of the SSA estimators with $r = 1$, $r = 2$, and the bias-corrected version, in terms of three standard measures including the integrated mean squared error and the Kolmogorov-Smirnov distance.

The SSA estimator with $r = 1$

The case of $r = 1$ provides the simplest estimator, which will be denoted as the SSA^{1c} estimator, since one component of the SVD decomposition (3.6.1.1) is used. In the following lemma, we establish that the SSA algorithm with $r = 1$ yields a valid result.

Lemma 4. *The* SSA^{1c} *estimator* (3.6.1.3) *is a valid distribution function.*

Proof. Since the elements of the series $\mathcal{F}_m$ are positive, the elements of $\mathbb{X}\mathbb{X}^{\mathrm{T}}$ are also positive. This implies that the elements of the leading eigenvector U_1 have the same sign, which may be assumed to be positive. Therefore, the monotonicity of the SSA^{1c} estimator follows from (3.6.1.4).

Applying Jensen's inequality, we obtain

$$\sum_{i=1}^{L}\sum_{l=1}^{L} u_{1,i}u_{1,l}/L = \left(\sum_{i=1}^{L} u_{1,i}\right)^2 \Big/ L \le \sum_{i=1}^{L} u_{1,i}^2 = 1.$$

Therefore, using (3.6.1.4) we have $0 \le \hat{f}_j \le 1$, which completes the proof. □

The SSA estimator with $r = 2$

We note that the estimator $\hat{\mathcal{F}}_m(L, 2)$ (i.e., the SSA estimator with $r = 2$) may be a series that is not necessarily a discrete version of a distribution function. In particular, the series $\hat{\mathcal{F}}_m(L, 2)$ can be non-monotonic. Let us introduce an operator $M(\mathcal{H})$ for a given non-decreasing series $\mathcal{H} = (H_1, \ldots, H_N)$, $H_i \in \mathbb{R}$, $1 \le i \le N$, such that $M(\mathcal{H})$ is a discrete version of a distribution function and close to $\mathcal{H}$. This can be achieved by replacing all negative values by zeroes and all values exceeding one by ones, and then taking the average of left and right monotonic functions. Thus, we define SSA^{2c} estimator as $\hat{\mathcal{F}}_m^{2c}(L) = M(\hat{\mathcal{F}}_m(L, 2))$.

Our numerical results (see Table 3.6.2) indicate that the SSA estimator with $r > 2$ is, on average, not really better. This can be explained as follows. In SSA theory, one of the

aims of SSA is to extract a series of finite rank τ from a noisy series, where a series of finite rank τ is a series whose trajectory matrix $\mathbb{X}$ has exactly τ nonzero eigenvalues. Since distribution functions are generally not of finite rank, the problem of approximating the series $\mathcal{F}^*$ by a series of finite rank arises. The accuracy of this approximation depends on L and r, which jointly give too much freedom. Therefore, the SSA estimator with $r = 2$ is typically sufficient for practical purposes.

The SSA estimator with $r = 1$ and bias correction

Recall that the SSA^{1c} estimator is a linear filter of size $2L$–1. Moreover, the proof of Lemma 1 shows that this filter is a weighted moving average with positive coefficients. Therefore, the SSA^{1c} estimator generally has a bias, which is desirable to remove.

Therefore we introduce the operator S defined as the SSA^{1c} estimator, that is,

$$S(\mathcal{F}_m|L) = \hat{\mathcal{F}}_m(L, 1).$$

We also consider the operator S^k which means repeated applications of the operator S, that is, $S^k = S(S^{k-1})$. In other words, $S^k(\mathcal{F}|L)$ means that the SSA^{1c} filter is consequently applied k times. To compute $S^k(\mathcal{F}|L)$, a series $\mathcal{F}$ should have at least $(1 + k)L$ zeroes at the beginning and $(1 + k)L$ ones at the end. For further considerations, we suppose that m is large enough such that $S(\mathcal{F}^*|L) \approx S(\mathcal{F}_m|L)$ and L is relatively small.

Let us define the operator

$$S^b(\mathcal{F}_m|L) = 3S(\mathcal{F}_m|L) - 3S^2(\mathcal{F}_m|L) + S^3(\mathcal{F}_m|L),$$

which has a smaller bias than the operator S as proved in Section 3.6.4.2. We notice that the series $S^b(\mathcal{F}|L)$ is not necessarily monotonic but it changes from 0 to 1. Therefore, the series $\hat{\mathcal{F}}^b_m(L) = M(S^b(\mathcal{F}_m|L))$ gives an estimator which will be called the SSA^b estimator of $\mathcal{F}^*$.

3.6.1.3 Performance of SSA estimators

To gain some insight into the finite sample properties, we investigate the SSA estimators by simulation and compare them with the kernel density estimator with the least-squares cross-validation (LSCV) bandwidth.

To do this, we consider the integrated squared error (ISE)

$$D_{\mathrm{ISE}}(\hat{p}) = \int \big(\hat{p}(x) - p(x)\big)^2 dx,$$

the Kolmogorov-Smirnov distance

$$D_{\mathrm{KS}}(\hat{F}) = \big\|\hat{F} - F\big\|_\infty = \max_x \big|\hat{F}(x) - F(x)\big|$$

and the Hellinger distance

$$D_{\mathrm{H}}(\hat{p}) = \int \Big(\sqrt{\hat{p}(x)} - \sqrt{p(x)}\Big)^2 dx,$$

where $p(x)$ is the density and $F(x)$ is the distribution function for the assumed model.

We consider two models: the normal distribution $N(0,1)$ and the mixture $0.4N(0,1) + 0.6N(5,2^2)$, and compute the average values of the above distances using 10000 simulated samples of size 100. These settings give $\delta \approx 0.01$ and $L_{\max} \approx 240$ for the first model and $\delta \approx 0.03$ and $L_{\max} \approx 200$ for the second model.

In Figure 3.6.3, we present the kernel density estimators and the SSA estimators with $L = 100$ and $L = 130$ for five samples of size 100 from $N(0,1)$, while these estimators with $L = 60$ and $L = 70$ for the mixture $0.4N(0,1) + 0.6N(5,2^2)$ are depicted in Figure 3.6.4. We can observe that the SSA estimators are slightly smoother than the kernel density estimators with the LSCV bandwidth.

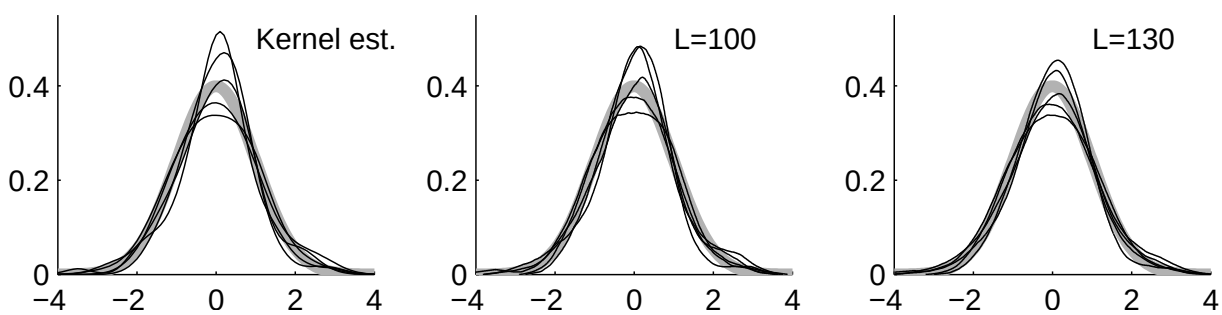


Abbildung 3.6.3: The kernel density estimators and the SSA^{1c} estimators with $L = 100$ and $L = 130$ for samples of size 100 from $N(0,1)$, whose density is given in grey.

Table 3.6.1 shows the means of the ISE, the Kolmogorov-Smirnov and Hellinger distances of the estimators to the true distribution functions. We can see that the SSA^{1c}, SSA^{2c} and SSA^{b} estimators with L from a wide range perform better in terms of all three criteria than the kernel density estimators.

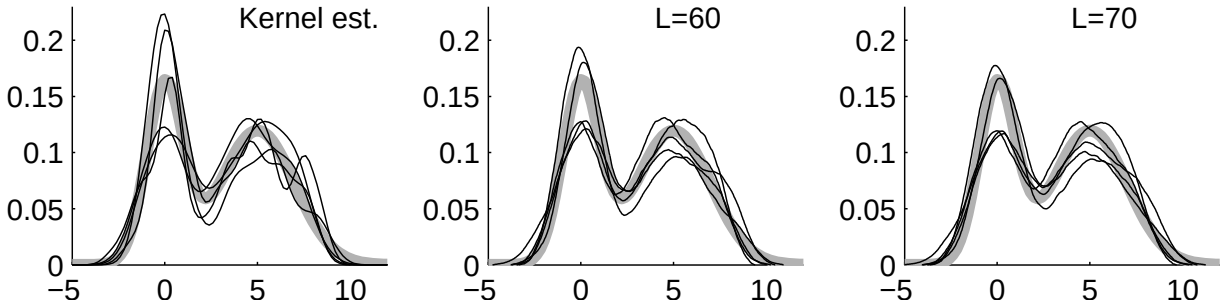


Abbildung 3.6.4: The kernel density estimators and the SSA^{1c} estimators with $L = 60$ and $L = 70$ for samples of size 100 from $0.4N(0,1)+0.6N(5,2^2)$, whose density is given in grey.

We notice that larger values of L are used for the model given by the normal distribution and smaller values of L are used for the model with two modes. This is similar to the behavior of the LSCV bandwidth, which decreases as a distribution becomes less normal.

We can observe that each SSA estimator has its own range of favorable values of L. The SSA^b estimator should be used with larger values of L since this gives greater smoothness while the SSA^{1c} estimator for such L has a noticeable bias. Meanwhile, the ranges for the SSA^{2c} and SSA^b estimators are relatively similar.

3.6.2 An automatic procedure to select a smoothing parameter

One of the crucial issues is how to automatically select the smoothing parameter L of the SSA estimator. Therefore, we propose a data-adaptive procedure, which selects L by considering the number of modes of the estimated density. It turns out that this procedure can be easily generalized to other density estimators such as the kernel density estimator.

3.6.2.1 Outline of the procedure

The probabilistic starting point for our automatic procedure is the following result, whose proof can be found in (Shorack, Wellner, 1986).

Theorem 3.6.2.1. *The empirical distribution function satisfies*

$$\limsup_{m\to\infty} R_m \left\| F_m - F \right\|_\infty \leq 1$$

Tabelle 3.6.1: Means of the ISE, the Kolmogorov-Smirnov and Hellinger distances for the kernel, SSA^{1c}, SSA^{2c} and SSA^{b} density estimators for samples of size 100.

	$\mathbf{E}D_{\mathrm{ISE}}$	$\mathbf{E}D_{\mathrm{KS}}$	$\mathbf{E}D_{\mathrm{H}}$
model $N(0,1)$			
Kernel est. with h_{LSCV}	0.0071	0.0551	0.0143
SSA^{1c} est. with $L=80$	0.0065	0.0521	0.0122
SSA^{1c} est. with $L=100$	0.0055	0.0509	0.0113
SSA^{1c} est. with $L=130$	0.0061	0.0550	0.0139
SSA^{2c} est. with $L=160$	0.0054	0.0496	0.0135
SSA^{2c} est. with $L=190$	0.0048	0.0486	0.0122
SSA^{2c} est. with $L=220$	0.0048	0.0493	0.0117
SSA^{b} est. with $L=130$	0.0054	0.0497	0.0144
SSA^{b} est. with $L=160$	0.0046	0.0479	0.0135
SSA^{b} est. with $L=190$	0.0048	0.0492	0.0137
model $0.4N(0,1)+0.6N(5,2^2)$			
Kernel est. with h_{LSCV}	0.0058	0.0617	0.0231
SSA^{1c} est. with $L=50$	0.0044	0.0595	0.0172
SSA^{1c} est. with $L=60$	0.0044	0.0593	0.0179
SSA^{1c} est. with $L=70$	0.0048	0.0603	0.0204
SSA^{2c} est. with $L=80$	0.0045	0.0598	0.0190
SSA^{2c} est. with $L=90$	0.0042	0.0591	0.0179
SSA^{2c} est. with $L=100$	0.0042	0.0588	0.0174
SSA^{b} est. with $L=70$	0.0044	0.0608	0.0193
SSA^{b} est. with $L=80$	0.0042	0.0602	0.0182
SSA^{b} est. with $L=90$	0.0044	0.0607	0.0185

almost surely, where $R_m = \dfrac{2\sqrt{m}}{\sqrt{2\ln\ln m}}$.

In the statement of Theorem 1, let us now substitute $F(x)$ by the estimator $\hat{\mathcal{F}}_m(L)$. Thus, we are interested in values of L such that $\|\mathcal{F}_m - \hat{\mathcal{F}}_m(L)\|_\infty \leq 1/R_m$. We now determine a maximal value of L that satisfies this approximation as follows

$$\bar{L} = \max\Big\{L \leq L_{\max} \,:\, \big\|\mathcal{F}_m - \hat{\mathcal{F}}_m(l)\big\|_\infty \leq 1/R_m \;\forall\, l \in \{1,\ldots,L\}\Big\},$$

which exists since $\left\|\mathcal{F}_m - \hat{\mathcal{F}}_m(L)\right\|_\infty = 0$ if $L = r$. Thus, we have to find an optimal value L_a from the set $\{1, \ldots, \bar{L}\}$. To proceed further, let M_L be the number of modes of the estimated density for certain L and consider the sequence

$$(M_1, M_2, \ldots, M_{\bar{L}}),$$

which has a decreasing tendency. If, however, $M_{j+1} > M_j$ for some j, we say that the estimated density has a *spurious mode*, which should be ignored. Therefore, we define the sequence

$$(\check{M}_1, \check{M}_2, \ldots, \check{M}_{\bar{L}}),$$

where $\check{M}_j = \min\{M_1, M_2, \ldots, M_j\}$, which is monotonic.

Next, we divide the set $\{1, 2, \ldots, \bar{L}\}$ into groups such that the values $\check{M}_j$ are equal to each other within each group. Specifically, we define $a_1, b_1, \ldots, a_k, b_k$ and k such that

$$1 = a_1 \leq b_1 < \ldots < a_k \leq b_k = \bar{L},$$

$a_{i+1} = b_i + 1$ and $\check{M}_i = \check{M}_j$ for all $i, j \in \{a_l, \ldots, b_l\}$, $l \in \{1, \ldots, k\}$.

For clarity, in Figure 3.6.5 we demonstrate the definition of k, a_l and b_l which are uniquely determined by the sequence $\check{M}_1, \check{M}_2, \ldots, \check{M}_{\bar{L}}$. For the Lean Body Mass data set and $r = 1$, we obtain $k = 20$, $a_{15} = 37$, $a_{16} = 46$, $a_{17} = 54$, $a_{18} = 67$, $a_{19} = 73$, $a_{20} = 88$ and $b_{20} = \bar{L} = 104$. We can observe that the plotted sequences for $r = 1$ and $r = 2$ are quite similar.

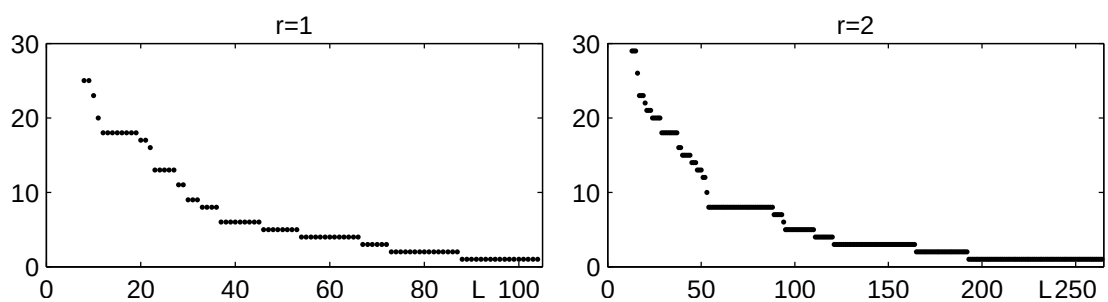


Abbildung 3.6.5: The sequence $\check{M}_1, \check{M}_2, \ldots, \check{M}_{\bar{L}}$ for $r = 1$ (left) and $r = 2$ (right). The sample is measurements of Lean Body Mass from the Australian Institute of Sport data.

Finally, we compute the optimal value L_a as an average with weight coefficients, which are proportional to the sizes of these k groups, namely

$$L_a = \left\lfloor \sum_{i=1}^{k} c_i w_i \right\rfloor, \; c_i = \gamma_i a_i + (1 - \gamma_i) b_i, \; w_i = \frac{b_i - a_i}{\sum_{j=1}^{k} b_j - a_j},$$

where $\gamma_i = 1/2$ if $\check{M}_{a_i} = 1$ and $\gamma_i = 0.9$ otherwise. This means that c_i is the middle of the interval $[a_i, b_i]$ provided there is one non-spurious mode, and c_i is a little larger than the left bound of the interval $[a_i, b_i]$ otherwise.

To adapt this procedure for the automatic choice of the bandwidth of the kernel density estimator, we should replace the set $\{1, 2, \ldots, L_{\max}\}$ by a dense set $\{h_1, h_2, \ldots\}$ and define the optimal bandwidth as $h_a = \sum_{i=1}^{k} c_i w_i$.

3.6.2.2 Consistency of SSA estimators

Let us now study the question whether the SSA estimator $\hat{F}_m(x)$ is consistent for $F(x)$, if the automatic data-adaptive procedure for the selection of the smoothing parameter L is used. We give an answer by the following Glivenko-Cantelli result.

Theorem 3.6.2.2. *Let $F(x)$ be a continuous distribution function. Then the SSA estimator $\hat{F}_m(x) = \hat{F}_m(x|L_a)$ is consistent and*

$$\left\|\hat{F}_m - F\right\|_\infty = \sup_{-\infty<x<\infty} |\hat{F}_m(x) - F(x)| \overset{a.s.}{\to} 0$$

as $m \to \infty$ and $\delta \to 0$.

This theorem follows from the following general result, which provides an explicit uniform upper bound on the estimation error $|\hat{F}_m(x) - F(x)|$. To proceed, we need some notation. Let

$$\varepsilon_N(H) = \max_{j=1,\ldots,N+1} \sup_{x\in[t_{j-1},t_j)} \left|H(x) - H(t_{j-1})\right|$$

denote the sampling error when sampling a distribution function $H(x)$ on a grid $\{t_1, \ldots, t_N\}$, $t_1 < \cdots < t_N$, where $t_0 = -\infty$ and $t_{N+1} = +\infty$.

Let the modulus of continuity be defined as

$$\omega(H, I) = \sup_{x,y\in I} \left|H(x) - H(y)\right|$$

for an interval $I \subset \mathbb{R}$. Notice that $\omega(H, I)$ is bounded for a differentiable distribution function H with bounded density h,

$$\omega(H, I) \leq \|h\|_\infty |I|,$$

where $|I|$ denotes the length of I.

We are now in a position to formulate the following Glivenko-Cantelli result for the SSA estimator in the case of a general distribution function.

Theorem 3.6.2.3. *Let $F(x)$ be an arbitrary distribution function. Then there exists a sequence of partitions such that*

$$\left\|\hat{F}_m - F\right\|_\infty \stackrel{a.s.}{\to} 0$$

along these partitions as $m \to \infty$. For a fixed partition $\{t_j\}$, we have the bounds

$$\begin{aligned}\left|\hat{F}_m(x) - F(x)\right| \le\ & 3\left\|F_m - F\right\|_\infty + \frac{3\sqrt{2\ln\ln m}}{2\sqrt{m}} \\ & +\varepsilon_N(F) + \max_{j=1,\dots,N+1} \omega\Big(F, [t_{j-1}, t_j)\Big)\end{aligned}$$

uniformly in $x \in \mathbb{R}$. If, additionally, $F(x)$ is differentiable with bounded derivative $p(x) = F'(x)$, then

$$\begin{aligned}\left|\hat{F}_m(x) - F(x)\right| \le\ & 3\left\|F_m - F\right\|_\infty + \frac{3\sqrt{2\ln\ln m}}{2\sqrt{m}} \\ & +\varepsilon_N(F) + \|p\|_\infty \max_{j=1,\dots,N+1} |t_j - t_{j-1}|,\end{aligned}$$

uniformly in $x \in \mathbb{R}$.

The proof of the theorem is deferred to the Appendix.

Note that the upper bound consists of four additive components. The first term is determined by the classical Kolmogorov-Smirnov distance, the second one is $o(1)$, as $m \to \infty$, the third term measures the sampling error controlled by the choice of the sampling points $\{t_i\}$ and the last term is the modulus of continuity of I with respect to the grid. This shows that, in addition to the probabilistic behavior of the Kolmogorov-Smirnov distance, the smoothness properties of the underlying distribution function matters.

3.6.3 Numerical examples

In Table 3.6.2 we present the performance of the automatic procedure for the SSA and kernel density estimators. We can see that the kernel density estimator with the bandwidth h_a is better than the kernel density estimator with other bandwidths. We see that the Sheather-Jones plug-in (SJPI) bandwidth is better than the LSCV bandwidth and the ICV bandwidth is not quite good for the model of normal distribution. However, the SJPI bandwidth is worse than the LSCV bandwidth if the model has two modes of significantly distinct width. It is worth to notice that the density with the LSCV bandwidth gives a

small ISE for the model $0.3N(0,1) + 0.7N(15, 4^2)$ but it yields a lot of modes while the density with other bandwidths and the SSA estimators have typically 2 or 3 modes.

We can observe that the SSA^{1c}, SSA^{b} and SSA^{2c} estimators with an automatic choice of the smoothing parameter provide a small ISE compared to the kernel density estimator, while results for the Kolmogorov-Smirnov and Hellinger distances are similar or slightly better. The SSA^{3c} estimator is useful only for 'hard' distributions, which can be identified by distinctness of values L_a for $r = 1, 2, 3$. Overall, the SSA^{2c} estimator has the better performance.

Note that the computational costs of SSA-based procedures can be considerable. Our results show that the computational time for the SSA estimator is moderately larger than the computational time for the LSCV and ICV estimators using standard routines in Matlab. In addition, there are very fast realizations of the SSA procedure (23).

In Table 3.6.3, we present simulated means, standard deviations and mean squared deviations (MSD) of the sampling plan size using several estimators of the distribution function for the model $0.4N(0,1) + 0.6N(5, 2^2)$. We can see that the straightforward use of the empirical distribution function provides a huge variance of the estimated sampling plan size. The kernel density estimator yields a bias for estimating the sampling plan, since the LSCV bandwidth tends to be large for estimating extreme quantiles for the considered model. At the same time, the SSA^{1c} estimator gives a similar bias but a smaller standard deviation than the kernel density estimator. We can observe that the SSA^{2c} estimator is the best one, since it provides a very small bias for estimating the sampling plan and standard deviation is rather small while the SSA^{b} estimator has similar features.

3.6.4 Proofs of statements

3.6.4.1 Proof of Theorem 2

We note that by right continuity

$$\varepsilon_N(H) \to 0$$

as $N \to \infty$, if $\|\{t_j\}\| \to 0$, where $\|\{t_j\}\| = \max_j |t_j - t_{j-1}|$ is the size of partition.

Recall now that $f_j = F_m(t_j)$ and that by construction

$$\max_{j=1,\ldots,N} \left| \hat{f}_j - f_j \right| \leq \frac{\sqrt{2 \ln \ln m}}{2\sqrt{m}},$$

Tabelle 3.6.2: Means of the ISE, the Kolmogorov-Smirnov and Hellinger distances for the kernel density estimators with different bandwidths and the SSA^{1c}, SSA^{b} and SSA^{2c} estimators using the automatic procedure for samples of size 100. Last column shows the mean of the parameter L_a, which is found by the automatic procedure.

	$\mathbf{E}D_{\mathrm{ISE}}$	$\mathbf{E}D_{\mathrm{KS}}$	$\mathbf{E}D_{\mathrm{H}}$	$\mathbf{E}L_a$
model $N(0,1)$				
Kernel est. with h_{LSCV}	0.0071	0.0551	0.0143	
Kernel est. with h_{SJPI}	0.0066	0.0536	0.0131	
Kernel est. with h_{ICV}	0.0075	0.0546	0.0146	
Kernel est. with h_a	0.0063	0.0546	0.0131	
SSA^{1c} est.	0.0061	0.0537	0.0128	107.6
SSA^{2c} est.	0.0060	0.0503	0.0142	145.7
SSA^{3c} est.	0.0090	0.0556	0.0189	143.4
SSA^{b} est.	0.0052	0.0488	0.0141	141.7
model $0.4N(0,1) + 0.6N(5,2^2)$				
Kernel est. with h_{LSCV}	0.0058	0.0617	0.0231	
Kernel est. with h_{SJPI}	0.0052	0.0609	0.0210	
Kernel est. with h_{ICV}	0.0055	0.0614	0.0229	
Kernel est. with h_a	0.0053	0.0611	0.0236	
SSA^{1c} est.	0.0051	0.0607	0.0215	67.6
SSA^{2c} est.	0.0047	0.0610	0.0195	86.4
SSA^{3c} est.	0.0054	0.0623	0.0219	90.3
SSA^{b} est.	0.0052	0.0617	0.0206	89.7
model $0.3N(0,1) + 0.7N(15,4^2)$				
Kernel est. with h_{LSCV}	0.0048	0.0672	0.0394	
Kernel est. with h_{SJPI}	0.0069	0.0733	0.0602	
Kernel est. with h_{ICV}	0.0049	0.0670	0.0396	
Kernel est. with h_a	0.0050	0.0674	0.0440	
SSA^{1c} est.	0.0053	0.0679	0.0451	40.9
SSA^{2c} est.	0.0046	0.0660	0.0380	60.0
SSA^{3c} est.	0.0043	0.0643	0.0349	73.8
SSA^{b} est.	0.0047	0.0670	0.0391	48.9

Tabelle 3.6.3: Mean, standard deviation and MSD of the sampling plan size using the empirical distribution function, the kernel density estimator and the SSA^{1c}, SSA^{b} and SSA^{2c} estimators for samples of size m from $0.4N(0,1) + 0.6N(5, 2^2)$. The true sampling plan size is 392.

	$m = 250$			$m = 500$			$m = 1000$		
	mean	s.d.	MSD	mean	s.d.	MSD	mean	s.d.	MSD
EDF	469.7	468.9	475.3	462.7	278.9	287.7	422.0	160.1	162.9
Kernel est.	322.3	102.8	124.2	324.0	79.7	104.8	337.8	59.2	80.3
SSA^{1c} est.	301.4	71.8	115.6	316.3	61.3	97.4	329.1	46.9	78.5
SSA^{b} est.	409.6	112.9	114.3	408.9	88.8	90.4	403.6	67.7	68.7
SSA^{2c} est.	390.4	107.8	107.8	393.1	81.4	81.4	392.8	62.1	62.1

yielding

$$\left|\hat{F}_m(x) - Z(x|f_1, \ldots, f_N)\right| \leq \frac{3\sqrt{2\ln\ln m}}{2\sqrt{m}},$$

where

$$Z(x|g_1, \ldots, g_N) = \sum_{j=2}^{N} \left(g_{j-1} + (g_j - g_{j-1}) \frac{x - t_{j-1}}{t_j - t_{j-1}} \right) \mathbf{1}_{[t_{j-1}, t_j)}(x) + \mathbf{1}_{[t_N, \infty)}(x)$$

for $g_1, \ldots, g_N \in [0, 1]$. Further, we get

$$\left|\hat{F}_m(x) - Z\big(x|F(t_1), \ldots, F(t_N)\big)\right| \leq \frac{3\sqrt{2\ln\ln m}}{2\sqrt{m}} + 3\|F_m - F\|_\infty$$

uniformly in $x \in \mathbb{R}$. Clearly, each interpolation term can be estimated by $\omega\big(F, [t_{j-1}, t_j)\big)$ such that

$$\begin{aligned} \Big|\hat{F}_m(x) \; - \; & \sum_{j=2}^{N} F(t_{j-1}) \mathbf{1}_{[t_{j-1}, t_j)}(x) - \mathbf{1}_{[t_N, \infty)}(x)\Big| \leq \\ & \leq \frac{3\sqrt{2\ln\ln m}}{2\sqrt{m}} + 3\|F_m - F\|_\infty + \max_{j=1,\ldots,N+1} \omega\Big(F, [t_{j-1}, t_j)\Big). \end{aligned}$$

Hence, we obtain

$$\begin{aligned} \left|\hat{F}_m(x) - F(x)\right| \leq \; & \frac{3\sqrt{2\ln\ln m}}{2\sqrt{m}} + 3\|F_m - F\|_\infty \\ & + \max_{j=1,\ldots,N+1} \omega\Big(F, [t_{j-1,t_j})\Big) + \varepsilon_N(F). \end{aligned}$$

Now let $\epsilon > 0$. By (4, p. 122), there exists a partition $\|\{t_j^\epsilon\}\|$ such that the third term is less than $\epsilon/3$. By use of a subpartition, if necessary, we can ensure that $\varepsilon_N(F) < \epsilon/3$. Now choose m large enough such that $\frac{3\sqrt{2\ln\ln m}}{2\sqrt{m}} < \epsilon/3$. Then $\|\hat{F}_m - F\|_\infty \leq 3\|F_m - F\|_\infty + \epsilon$. Choosing $\epsilon = 1/m$, the statement of the theorem follows from the Glivenko-Cantelli theorem. □

3.6.4.2 Justification of the bias approximation

Consider the operator $C(F) = \mu_\varepsilon * F$, where μ_ε is a measure with finite moments,

$$C(F)(x) = \int F(x-y)\mu_\varepsilon(dy),$$

and we interpret μ_ε as a kernel with width ε. If μ_ε is a discrete measure, then $C(F)$ is a weighted moving-average of F.

Let our problem be estimating F by means of linear combinations of $C^j(F)$ for $j = 1, 2, \ldots$, where C^j is the j-fold composition of the operator C, $C^0(F) = F$. Therefore, we consider an estimator of F in the form $\widehat{F}_P(\mu_\varepsilon) = P(\mu_\varepsilon) * F$, where P is a polynomial with no intercept. We note that this estimator has the required form since $P(0) = 0$. Besides, the error of the estimator is equal to $P(\mu_\varepsilon) * F - F = P^{\mathrm{err}}(\mu_\varepsilon) * F$, where $P^{\mathrm{err}} = P - 1$.

Let the estimator $\widehat{F}_P(\mu_\varepsilon)$ be generated by $P_k(z) = 1-(1-z)^k$. Then we have $P_k^{\mathrm{err}}(z) = -(1-z)^k$. To study this estimator, we define $w_{k,\varepsilon} = \int y^k \mu_\varepsilon(dy)$ for $k \in \mathbb{N}$, $w_{0,\varepsilon} = \int \mu_\varepsilon(dy) - 1$ and consider the formal Taylor expansion of a smooth function F at the point x:

$$F(x-y) = \sum_{i=0}^{\infty} (-1)^i \frac{F^{(i)}(x)}{i!} y^i$$

and its derivatives

$$F^{(j)}(x-y) = \sum_{i=0}^{\infty} (-1)^i \frac{F^{(i+j)}(x)}{i!} y^i.$$

Then we have the following formal expansion of the error term:

$$\widehat{F}_{P_k}(\mu_\varepsilon) - F = P_k^{\mathrm{err}}(\mu_\varepsilon) * F = \sum_{i_1,\ldots,i_k=0}^{\infty} (-1)^{k+1+\sum i_j} \frac{F^{\left(\sum_{j=1}^k i_j\right)}}{\prod_{j=1}^k i_j!} \prod_{j=1}^k w_{i_j,\varepsilon}. \qquad (3.6.4.1)$$

Note that the SSA^{1c} estimator corresponds to the case $k = 1$, where $P_1(z) = 1-(1-z) = z$ implies $\widehat{F}_{P_1}(\mu_\varepsilon) = \mu_\varepsilon * F = C(F)$. Meanwhile, the SSA^b estimator corresponds to the case $k = 3$, where $P_3(z) = 1-(1-z)^3 = z^3 - 3z^2 + 3z$ entails $\widehat{F}_{P_1}(\mu_\varepsilon) = C^3(F) - 3C^2(F) + C(F)$.

The formula (3.6.4.1) shows that if $w_{j,\varepsilon} \to 0$ as $\varepsilon \to 0$ with sufficient rates of convergence (for example, $w_{j,\varepsilon} = O(\varepsilon^j)$), then the error $\widehat{F}_{P_k}(\mu_\varepsilon) - F$ decreases as k increases for small ε.

Recall that the operator S is an operator C with the measure μ_ε concentrated at the points $\pm i\delta$, $i = 0, \ldots, L$. Therefore, μ_ε is close to the Dirac delta function concentrated at 0 if L is small. Thus, we can expect that the SSA^b estimator of a smooth enough distribution function F has the smaller bias than the SSA^{1c} estimator that is confirmed by extensive numerical studies.

3.6.5 Simulation study for SSA estimators

In Tables 3.6.4–3.6.12 we can observe that the SSA1c estimators is better for small sample sizes ($m \leq 500$). For large sample sizes ($m = 5000$) the use of the SSA2c and SSAb estimators is more preferable.

Tabelle 3.6.4: Characteristics of distributions of n_m and c_m using the SSA estimator for model 1.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	SSA1c	35	46	60	74	90	61.6	22.3	-3.4	22.6	15.6	2.4
100	SSA2c	37	54	77	104	133	82.5	39.7	17.5	43.4	16.5	3.4
100	SSA3c	31	48	74	107	148	83.8	50.2	18.8	53.7	16.4	4.3
100	SSAb	41	59	82	108	133	85.8	37.1	20.8	42.5	16.7	3.1
250	SSA1c	41	49	59	71	82	60.7	16.2	-4.3	16.8	15.3	1.8
250	SSA2c	48	58	72	87	102	73.6	21.6	8.6	23.2	16.0	2.0
250	SSA3c	41	51	66	83	101	69.1	24.5	4.1	24.9	15.5	2.4
250	SSAb	49	62	77	93	108	78.3	23.0	13.3	26.6	16.3	2.1
500	SSA1c	45	52	60	69	78	60.9	13.2	-4.1	13.8	15.1	1.4
500	SSA2c	52	60	70	80	91	70.6	15.5	5.6	16.5	15.8	1.5
500	SSA3c	46	54	64	75	86	65.1	16.0	0.1	16.0	15.2	1.6
500	SSAb	52	61	72	85	96	73.5	17.1	8.5	19.1	15.9	1.6
5000	SSA1c	56	59	63	66	70	62.8	5.7	-2.2	6.1	15.0	0.6
5000	SSA2c	59	62	66	70	73	65.8	5.7	0.8	5.7	15.1	0.6
5000	SSA3c	57	60	64	67	71	63.6	5.6	-1.4	5.8	15.0	0.6
5000	SSAb	59	62	66	70	73	66.0	5.8	1.0	5.8	15.1	0.6

Compare tables 3.3.1, 3.4.2, 3.5.1, 3.5.19, 3.6.4.

Tabelle 3.6.5: Characteristics of distributions of n_m and c_m using the SSA estimator for model 2.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	SSA1c	21	50	98	166	255	122.5	103.1	19.5	104.9	29.4	11.8
100	SSA2c	24	54	106	184	284	136.5	119.0	33.5	123.6	30.9	12.5
100	SSA3c	23	53	105	183	287	137.1	121.3	34.1	126.0	30.9	12.6
100	SSAb	24	56	108	186	289	138.9	120.8	35.9	126.0	31.2	12.5
250	SSA1c	43	64	95	137	184	106.0	58.8	3.0	58.9	29.5	7.1
250	SSA2c	45	68	100	147	199	113.5	64.7	10.5	65.5	30.4	7.6
250	SSA3c	44	67	99	147	202	113.7	66.4	10.7	67.2	30.4	7.8
250	SSAb	46	69	102	150	202	115.5	65.7	12.5	66.9	30.7	7.6
500	SSA1c	54	71	96	125	157	101.7	42.5	-1.3	42.5	29.5	5.2
500	SSA2c	56	74	100	131	166	107.0	45.4	4.0	45.6	30.2	5.5
500	SSA3c	55	74	99	131	168	106.9	46.6	3.9	46.8	30.2	5.6
500	SSAb	57	76	102	134	169	109.1	46.8	6.1	47.2	30.5	5.6
5000	SSA1c	82	88	98	106	115	98.5	13.7	-4.5	14.5	29.8	1.8
5000	SSA2c	84	91	99	108	118	100.4	14.2	-2.6	14.4	30.0	1.9
5000	SSA3c	84	90	99	108	118	99.9	14.3	-3.1	14.6	30.0	1.9
5000	SSAb	84	91	100	109	119	101.3	14.5	-1.7	14.6	30.2	1.9

Compare tables 3.3.2, 3.4.3, 3.5.2, 3.5.20, 3.6.5.

Tabelle 3.6.6: Characteristics of distributions of n_m and c_m using the SSA estimator for model 3.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	SSA1c	100	137	187	248	312	199.2	86.7	-9.8	87.3	18.6	3.1
100	SSA2c	115	166	236	316	400	250.5	115.0	41.5	122.3	19.9	3.7
100	SSA3c	102	150	218	295	381	232.2	111.9	23.2	114.3	19.2	3.8
100	SSAb	115	172	249	343	443	268.5	130.4	59.5	143.3	20.2	4.0
250	SSA1c	123	152	188	231	272	194.4	60.0	-14.6	61.7	18.3	2.3
250	SSA2c	139	176	224	280	334	232.6	77.3	23.6	80.8	19.3	2.6
250	SSA3c	128	161	205	258	308	213.3	72.1	4.3	72.2	18.7	2.6
250	SSAb	141	181	233	294	356	242.7	84.9	33.7	91.3	19.4	2.8
500	SSA1c	139	163	190	220	255	193.8	45.6	-15.2	48.1	18.2	1.8
500	SSA2c	155	184	218	256	298	222.7	56.1	13.7	57.8	18.9	2.0
500	SSA3c	144	170	202	239	276	207.0	52.9	-2.0	52.9	18.4	2.0
500	SSAb	155	187	223	265	311	229.0	60.6	20.0	63.8	19.0	2.1
5000	SSA1c	175	185	198	210	225	198.8	19.6	-10.2	22.1	18.1	0.8
5000	SSA2c	183	196	209	223	237	209.6	20.9	0.6	20.9	18.3	0.8
5000	SSA3c	177	189	203	216	231	203.2	20.7	-5.8	21.5	18.2	0.8
5000	SSAb	184	196	209	223	238	210.2	21.2	1.2	21.2	18.3	0.8

Compare tables 3.3.3, 3.4.4, 3.5.3, 3.5.21, 3.6.6.

Tabelle 3.6.7: Characteristics of distributions of n_m and c_m using the SSA estimator for model 4.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	SSA1c	74	104	146	203	262	159.8	77.2	-8.2	77.7	23.6	5.1
100	SSA2c	80	116	167	239	316	186.8	99.1	18.8	100.9	25.0	5.9
100	SSA3c	76	114	170	246	333	190.8	105.0	22.8	107.5	25.2	6.2
100	SSAb	83	120	176	248	333	195.5	105.0	27.5	108.5	25.5	6.0
250	SSA1c	93	118	152	193	233	159.0	56.5	-9.0	57.2	23.7	3.8
250	SSA2c	98	127	168	219	271	178.0	69.1	10.0	69.8	24.8	4.4
250	SSA3c	96	124	165	215	267	174.8	68.6	6.8	68.9	24.6	4.4
250	SSAb	101	130	173	226	277	182.7	71.2	14.7	72.7	25.0	4.4
500	SSA1c	106	127	153	183	215	157.8	43.0	-10.2	44.2	23.8	2.9
500	SSA2c	111	135	165	202	238	171.5	50.8	3.5	50.9	24.6	3.3
500	SSA3c	108	132	163	199	235	168.7	50.5	0.7	50.5	24.4	3.3
500	SSAb	112	137	168	206	242	174.3	51.8	6.3	52.2	24.7	3.3
5000	SSA1c	140	150	163	173	185	161.9	16.7	-6.1	17.8	24.1	1.2
5000	SSA2c	142	153	167	178	190	166.2	17.8	-1.8	17.9	24.4	1.2
5000	SSA3c	141	152	165	177	189	164.8	17.8	-3.2	18.0	24.3	1.2
5000	SSAb	143	154	168	179	192	167.3	18.2	-0.7	18.2	24.5	1.3

Compare tables 3.3.4, 3.4.5, 3.5.4, 3.5.22, 3.6.7.

Tabelle 3.6.8: Characteristics of distributions of n_m and c_m using the SSA estimator for model 5.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	SSA1c	263	375	520	711	921	566.1	268.7	-41.9	271.9	41.2	8.6
100	SSA2c	286	420	612	855	1127	667.8	340.4	59.8	345.6	44.2	10.0
100	SSA3c	274	411	598	833	1099	652.1	330.7	44.1	333.6	43.7	9.9
100	SSAb	299	436	632	880	1163	691.2	348.5	83.2	358.2	44.9	9.9
250	SSA1c	336	418	533	671	805	556.5	189.3	-51.5	196.2	41.4	6.3
250	SSA2c	359	457	593	765	934	625.8	230.2	17.8	230.8	43.5	7.2
250	SSA3c	348	448	580	753	917	613.2	229.9	5.2	230.0	43.1	7.3
250	SSAb	372	473	610	781	946	640.7	230.1	32.7	232.4	44.0	7.1
500	SSA1c	382	452	542	642	749	556.1	145.0	-51.9	153.9	41.6	4.9
500	SSA2c	404	485	588	713	831	607.4	170.3	-0.6	170.3	43.3	5.5
500	SSA3c	394	475	581	702	828	599.8	173.2	-8.2	173.3	43.0	5.6
500	SSAb	416	498	600	723	841	618.7	170.8	10.7	171.1	43.6	5.4
5000	SSA1c	502	535	576	614	652	576.4	57.5	-31.6	65.6	42.6	2.0
5000	SSA2c	515	552	598	638	679	596.3	62.1	-11.7	63.2	43.2	2.1
5000	SSA3c	510	545	595	636	676	592.5	63.2	-15.5	65.0	43.1	2.2
5000	SSAb	516	556	604	641	679	601.0	63.1	-7.0	63.4	43.4	2.2

Compare tables 3.3.5, 3.4.6, 3.5.5, 3.5.23, 3.6.8.

Tabelle 3.6.9: Characteristics of distributions of n_m and c_m using the SSA estimator for model 6.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	SSA1c	136	184	250	333	421	268.0	117.4	-56.0	130.1	29.4	5.5
100	SSA2c	154	216	306	421	542	332.2	159.9	8.2	160.1	32.1	6.7
100	SSA3c	150	216	311	431	570	339.9	170.2	15.9	170.9	32.3	7.1
100	SSAb	161	224	322	436	570	347.5	168.5	23.5	170.2	32.7	6.8
250	SSA1c	176	216	269	335	402	281.5	91.5	-42.5	100.9	30.2	4.3
250	SSA2c	193	243	312	399	487	328.7	116.8	4.7	116.9	32.2	5.1
250	SSA3c	187	237	306	392	478	321.5	115.3	-2.5	115.3	31.9	5.1
250	SSAb	199	251	325	413	506	340.8	122.1	16.8	123.2	32.7	5.2
500	SSA1c	201	237	280	331	382	287.6	72.1	-36.4	80.8	30.7	3.4
500	SSA2c	216	259	312	374	437	321.1	87.8	-2.9	87.8	32.1	3.9
500	SSA3c	211	253	306	367	431	315.2	87.4	-8.8	87.8	31.8	4.0
500	SSAb	220	265	319	384	448	329.4	91.2	5.4	91.4	32.5	4.0
5000	SSA1c	266	283	305	325	346	305.4	30.2	-18.6	35.4	31.6	1.5
5000	SSA2c	274	293	320	339	364	318.0	33.0	-6.0	33.5	32.2	1.6
5000	SSA3c	271	290	316	336	359	314.4	32.8	-9.6	34.2	32.0	1.6
5000	SSAb	277	296	323	341	367	320.9	33.8	-3.1	33.9	32.3	1.6

Compare tables 3.3.6, 3.4.7, 3.5.6, 3.5.24, 3.6.9.

Tabelle 3.6.10: Characteristics of distributions of n_m and c_m using the SSA estimator for model 7.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	SSA1c	500	684	929	1234	1578	995.2	433.3	-209.8	481.4	53.3	9.9
100	SSA2c	566	801	1119	1529	1975	1212.1	564.9	7.1	564.9	58.1	11.7
100	SSA3c	554	781	1086	1482	1911	1173.0	546.3	-32.0	547.2	57.2	11.6
100	SSAb	599	833	1163	1572	2028	1253.6	580.9	48.6	582.8	59.1	11.6
250	SSA1c	645	791	988	1216	1457	1028.8	332.4	-176.2	376.2	54.7	7.7
250	SSA2c	717	891	1133	1425	1717	1184.1	405.4	-20.9	405.9	58.2	8.8
250	SSA3c	695	867	1103	1395	1684	1155.8	400.3	-49.2	403.2	57.6	8.9
250	SSAb	744	925	1161	1447	1752	1211.9	401.8	6.9	401.8	58.9	8.5
500	SSA1c	738	861	1009	1189	1376	1038.8	252.7	-166.2	302.4	55.2	5.9
500	SSA2c	801	949	1132	1350	1569	1166.8	307.5	-38.2	309.8	58.2	6.9
500	SSA3c	780	932	1117	1337	1562	1149.5	307.9	-55.5	312.8	57.8	7.0
500	SSAb	830	977	1156	1368	1580	1186.4	298.7	-18.6	299.3	58.7	6.6
5000	SSA1c	970	1024	1097	1165	1217	1096.3	98.2	-108.7	146.4	57.0	2.4
5000	SSA2c	1019	1080	1169	1243	1322	1168.7	117.8	-36.3	123.2	58.6	2.8
5000	SSA3c	1006	1072	1161	1235	1310	1157.1	119.3	-47.9	128.4	58.4	2.8
5000	SSAb	1030	1097	1185	1261	1321	1182.0	114.3	-23.0	116.5	59.0	2.7

Compare tables 3.3.7, 3.4.8, 3.5.7, 3.5.25, 3.6.10.

Tabelle 3.6.11: Characteristics of distributions of n_m and c_m using the SSA estimator for model 8.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	SSA1c	18	28	40	56	73	43.6	22.1	7.6	23.4	13.1	3.0
100	SSA2c	18	28	43	63	87	48.3	28.0	12.3	30.6	13.0	3.3
100	SSA3c	16	26	42	65	93	49.8	32.9	13.8	35.7	13.3	3.9
100	SSAb	16	26	42	63	89	48.2	29.5	12.2	31.9	12.8	3.4
250	SSA1c	22	29	37	48	59	39.2	14.6	3.2	14.9	12.5	2.1
250	SSA2c	22	29	38	50	62	40.3	15.8	4.3	16.4	12.3	2.2
250	SSA3c	21	28	37	49	60	39.2	15.7	3.2	16.0	12.3	2.2
250	SSAb	20	27	36	48	60	38.3	15.9	2.3	16.0	11.9	2.2
500	SSA1c	25	30	36	44	52	37.7	11.2	1.7	11.4	12.2	1.6
500	SSA2c	24	30	37	45	53	38.0	11.7	2.0	11.8	12.0	1.6
500	SSA3c	24	29	36	44	53	37.6	11.6	1.6	11.7	12.1	1.7
500	SSAb	23	28	35	43	51	36.4	11.5	0.4	11.5	11.7	1.7
5000	SSA1c	31	33	36	39	41	35.9	4.2	-0.1	4.2	11.9	0.6
5000	SSA2c	31	33	36	39	41	36.0	4.3	0.0	4.3	11.9	0.6
5000	SSA3c	31	33	36	39	41	36.1	4.3	0.1	4.3	11.9	0.7
5000	SSAb	30	33	36	39	41	36.0	4.3	0.0	4.3	11.9	0.7

Compare tables 3.3.8, 3.4.9, 3.5.8, 3.5.26, 3.6.11.

Tabelle 3.6.12: Characteristics of distributions of n_m and c_m using the SSA estimator for model 9.

m	k	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$\mathrm{E}\, n_m$	$\mathrm{sd}\, n_m$	bias	RMSD	$\mathrm{E}\, c_m$	$\mathrm{sd}\, c_m$
100	SSA1c	60	71	85	102	124	89.5	28.4	-72.5	77.9	18.7	2.3
100	SSA2c	75	98	131	170	213	138.9	56.1	-23.1	60.6	21.4	3.7
100	SSA3c	78	108	152	209	275	167.1	83.1	5.1	83.3	23.0	5.0
100	SSAb	65	84	110	141	173	115.8	44.4	-46.2	64.0	19.8	3.3
250	SSA1c	76	87	100	115	135	103.2	25.3	-58.8	64.0	19.6	1.9
250	SSA2c	87	103	122	145	167	125.5	32.4	-36.5	48.7	20.8	2.3
250	SSA3c	93	113	138	169	202	143.7	43.5	-18.3	47.2	22.0	2.8
250	SSAb	78	91	108	128	149	112.2	31.0	-49.8	58.6	19.8	2.3
500	SSA1c	89	99	112	128	145	115.5	23.0	-46.5	51.9	20.4	1.7
500	SSA2c	98	111	126	143	162	128.9	26.1	-33.1	42.2	21.1	1.8
500	SSA3c	102	116	134	154	174	136.8	29.1	-25.2	38.5	21.7	2.0
500	SSAb	94	107	123	143	167	127.2	29.5	-34.8	45.6	20.9	2.1
5000	SSA1c	129	137	145	155	163	146.0	13.3	-16.0	20.8	22.2	0.9
5000	SSA2c	137	144	155	163	170	153.8	13.1	-8.2	15.4	22.7	0.9
5000	SSA3c	134	142	153	161	168	151.9	13.2	-10.1	16.6	22.6	0.9
5000	SSAb	140	148	158	166	175	157.7	13.4	-4.3	14.1	22.9	0.9

Compare tables 3.3.9, 3.4.10, 3.5.9, 3.5.27, 3.6.12.

3.6.6 Summary

In Table 3.6.13, we present simulated root mean squared deviations (RMSD) of the sampling plan size using the SSA estimators for several models of the distribution function of measurements. We observe that the SSA1c estimator is typically better than SSA2c, SSA3c and SSAb estimators.

Tabelle 3.6.13: RMSD of the distribution of the sampling plan size using the SSA estimators for models 1–9.

model	m	SSA1c	SSA2c	SSA3c	SSAb
1	100	22.6	43.4	53.7	42.5
1	250	16.8	23.2	24.9	26.6
1	500	13.8	16.5	16.0	19.1
2	100	104.9	123.6	126.0	126.0
2	250	58.9	65.5	67.2	66.9
2	500	42.5	45.6	46.8	47.2
3	100	87.3	122.3	114.3	143.3
3	250	61.7	80.8	72.2	91.3
3	500	48.1	57.8	52.9	63.8
4	100	77.7	100.9	107.5	108.5
4	250	57.2	69.8	68.9	72.7
4	500	44.2	50.9	50.5	52.2
5	100	271.9	345.6	333.6	358.2
5	250	196.2	230.8	230.0	232.4
5	500	153.9	170.3	173.3	171.1
6	100	130.1	160.1	170.9	170.2
6	250	100.9	116.9	115.3	123.2
6	500	80.8	87.8	87.8	91.4
7	100	481.4	564.9	547.2	582.8
7	250	376.2	405.9	403.2	401.8
7	500	302.4	309.8	312.8	299.3
8	100	23.4	30.6	35.7	31.9
8	250	14.9	16.4	16.0	16.0
8	500	11.4	11.8	11.7	11.5
9	100	77.9	60.6	83.3	64.0
9	250	64.0	48.7	47.2	58.6
9	500	51.9	42.2	38.5	45.6

Kapitel 3.7

Acceptance sampling plans for two-sided specification limits

3.7.1 Introduction

Previous works (44; 20; 26) have shown that flash data from the production line can be used to infer successfully the optimal sampling plan, and the proposed procedures are already in active use by photovoltaic laboratories and manufacturers. However, the established methods aim at revealing *underperforming* lots of modules and cannot be used to detect lots of PV modules being out-of-spec. Note that overperforming modules can compensate underperforming modules and, therefore, the optimization of a photovoltaic system is mitigated, especially if the modules are as similar as possible. Further, in general, distributing the production capacity by reconciliation of specifications and demand leads to lower costs and a more economically optimal resource allocation.

In this chapter we extend the methodology to deal with the out-of-spec formulation of the photovoltaic acceptance sampling problem. The goal of the approach is also to find a method accounting for both manufacturer's and customer's interests. The shipment should be as close as possible to the agreed specification which gives both parties a sound base for energy yield forecasts and predictable return on invest. System design and operation is facilitated if a tight distribution of the lot within the specified limits is ensured. For instance, solar power inverters cut the system's output to avoid thermal problems and ensure high efficiency, improved lifetime and reliability, cf. (18).

The proposed methodology turns out to be far from being straightforward. Namely,

the formulation of the problem requires some care, in order to match the technical and physical needs and lead to a well defined solvable mathematical problem. It turns out that the two-sided out-of-spec formulation leads to nonlinear equations, which allow for explicit solutions only in special cases. Moreover, the solutions of those equations depend on both the fraction of underperforming modules and the fraction of overperforming modules. A practical proposal to handle that issue will be made, but the question arises how strongly the optimal sampling plan depends on the subdivision of the out-of-spec modules into these two classes. We investigated this issue by numerical computations for various models. The results are informative in many respects and show a notable dependence, which therefore has to be taken into account in practice. Also, numerical studies illustrate to some extent which output power distributions lead to large control samples.

3.7.2 Out-of-spec acceptance sampling framework

Acceptance sampling deals with the determination of the minimal sample size needed to assess the quality of a *lot* of items with a specified degree of confidence in terms of error probabilities. A lot may consist of those modules produced during a given period, represent a shipment for a customer or may be a collection of modules of a specific module class. In order to apply the acceptance sampling methodology, one needs to define the quality requirements, particularly the nominal specification of a module. In previous works (44; 20; 26), a PV module is regarded as non-conforming, if its output power X, as measured by a laboratory under STC, is less or equal than $\mu(1-\varepsilon)$, where $\mu = E(X)$ denotes the expectation and ε denotes the tolerance. Present day typical values for ε are 3% and 5%. This one-sided formulation aims at the detection of underperforming shipments. As argued above, a two-sided definition intending at the assessment of the formal specification is often more appropriate for the quality control. Thus we call a module *conforming*, if

$$|X - \mu| \leq \mu\varepsilon,$$

and *non-conforming* or *out-of-spec* otherwise. In other words, a PV module with output power X is conforming if

$$X \in [\tau_1, \tau_2],$$

$[\tau_1, \tau_2]$ being the *specification interval* defined by the bounds

$$\tau_1 = \mu(1 - \varepsilon),$$
$$\tau_2 = \mu(1 + \varepsilon)$$

according to the specification sheet. Note that in practice the interval $[\tau_1, \tau_2]$ typically stands for a module class. Usually, the parameter μ has to be chosen as the midpoint, $\mu = (\tau_1 + \tau_2)/2$. For example, for nominal power output 200W and tolerance 5%, we have $\mu = 205$, $\varepsilon = 210/205 - 1$, and $[\tau_1, \tau_2] = [200, 210]$.

Let N be the number of modules of the lot of interest and K be the number of out-of-spec modules. Then the fraction $p = K/N$ of non-conforming modules allows for two basic interpretations. First, p coincides with the expectation

$$p = \mathbb{E}(C/n)$$

where C denotes the number of out-of-spec modules in a control sample of size n, where $n \leq N$. This gives a sample based interpretation for p. Second, p equals the probability that a module is out-of-spec, i.e.

$$p = \mathbb{P}(X \notin [\tau_1, \tau_2]).$$

This probability is also called the *quality level*. For the two-sided out-of-spec formulation, the fraction of non-conforming modules is the sum of the fraction p_1 of underperforming modules and the fraction p_2 of overperforming modules, i.e.

$$p = p_1 + p_2$$

with $p_1 = \mathbb{P}(X < \tau_1)$ and $p_2 = \mathbb{P}(X > \tau_2)$. We follow the classic approach of statistical quality control and define the quality requirements by referring to these probabilities, which determine the expected number of modules which may be subject to misclassifications or warranty claims. One fixes the *acceptable quality level* (AQL) and the *rejectable quality level* (RQL). The lot is of low quality and should be rejected, if $p >$ RQL, and of high quality and should be accepted, if $p <$ AQL. The parameters AQL and RQL have to be selected in advance; either by the parties of a contract, often by a manufacturer and a consumer, or by a manufacturer only in order to define the desired quality.

Since p is unknown and unobservable except the case when total sampling is feasible, one has to rely on a control sample of randomly selected PV modules. If $X_1, \ldots, X_n$ denote the measurements of such a control sample, one applies a statistical decision function

$$T_n = T_n(X_1, \ldots, X_n)$$

to those observations in order to make a decision on acceptance or rejection of the lot. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote the sample mean of the control sample. Given a threshold τ, it is common to express the difference $\overline{X} - \tau$ as a multiple of the standard deviation σ of X. This leads to the standardized statistic

$$T_{n,\tau} = \sqrt{n}\frac{\bar{X} - \tau}{\sigma}.$$

Large values of $T_{n,\tau}$ indicate that the PV modules perform better than the bound τ, whereas small (negative) values indicate underperforming modules. Having a lower bound τ_1 as well as an upper bound τ_2, we consider the following decision rule. The lot is accepted, if

$$T_{n,\tau_1} < c \qquad \text{and} \qquad -T_{n,\tau_2} > c$$

for *some critical value* c. A sampling plan now consists of the pair (n, c), where n is minimal such that PV modules are inferred as high quality when $p <$ AQL with high probability and inferred as low quality when $p >$ RQL with low probability.

To proceed further, we need to introduce the *operating characteristic* (OC) curve of a sampling plan. In general, the OC curve is defined as the probability to accept the lot and considered as a function of the fraction p of non-conforming modules. In the case of the out-of-spec formulation, it turns out that the OC curve becomes a function of the two parameters p_1 and p_2,

$$\mathrm{OC}(p_1, p_2) = \mathbb{P}\Big(T_{n,\tau_1} < c, T_{n,\tau_2} < -c\Big).$$

Note that $\mathrm{OC}(p_1, p_2)$ depends on n and c although this is not reflected in notation. We can see that $\mathrm{OC}(p_1, p_2)$ extends the concept of the classic OC curve from the one-dimensional case to the two-dimensional case. Since false decisions may occur when one has to rely on control samples of randomly chosen PV modules, the best one can do is to design the decision procedure in such a way that both the probabilities of false acceptance and false rejection are controlled. Note that these probabilities are given by $\mathrm{OC}(p_1, p_2)$ and $1 - \mathrm{OC}(p_1, p_2)$ respectively.

Let α be the maximal probability of false acceptance, that the procedure should guarantee, and β be the maximal probability to reject, although the lot is of high quality. If p_1 and p_2 are unknown or unspecified, a statistically valid sampling plan (n, c) should ensure that the conditions

$$\begin{cases} \mathrm{OC}(p_1, p_2) \geq 1 - \alpha \text{ for all } p_1, p_2 \text{ such that } p_1 + p_2 \leq \mathrm{AQL}, \\ \mathrm{OC}(p_1, p_2) \leq \beta \text{ for all } p_1, p_2 \text{ such that } p_1 + p_2 \geq \mathrm{RQL} \end{cases} \tag{3.7.2.1}$$

hold true. A sampling plan (n, c) is called optimal if n is minimal among all solutions of the system of nonlinear inequalities (3.7.2.1).

3.7.3 Sampling plans for photovoltaic data

The approach introduced in the previous section, which is an extension the statistical quality control where the standard notions and measures are preserved, can be further integrated with the physical real data observed in photovoltaics, whose distribution can arbitrary without any shape restrictions. Note that the derivation of a mathematically sound and practically meaningful solution of (3.7.2.1) is a bit involved. Therefore, we describe in some detail the mathematical pitfalls and how they can be overcome. We start with a discussion of an approximation of the OC curve, which is vital for solving (3.7.2.1). To obtain a solution which is applicable in photovoltaics, we assume that flash data can used to estimate the unknown distribution of control observations, since flash data are usually available in practice. Monotonicity arguments allow to simplify the crucial nonlinear *inequalities* (3.7.2.1) to a system of nonlinear *equations*, which still depends on both probabilities p_1 and p_2 corresponding to the two classes of out-of-spec modules. We provide a recommendation to handle that issue in practice. Further, under certain symmetry assumptions we derive an explicit solution, which is in agreement with the existing methodology.

3.7.3.1 Approximation of the OC curve

Since the OC curve can only be calculated explicitly in special cases, one cannot determine practical sampling plans from (3.7.2.1). One needs to derive appropriate approximations.

Straightforward calculus shows that

$$\mathrm{OC}(p_1, p_2) = \mathbb{P}(T_{n,\tau_1} > c) - \mathbb{P}(T_{n,\tau_2} > -c),$$

where the right hand side is indeed a function of p_1 and p_2. This expression shows that the OC curve splits in two terms, which only depend on either T_{n,τ_1} or T_{n,τ_2}. This property facilitates considerably the construction of a valid solution. In (44) and (26), normal approximations of tail probabilities of the form $\mathbb{P}(T_{n,\tau} > c)$ have been established under the weak assumption that the measurements are distributed according to a distribution function $F(x) = \mathbb{P}(X \leq x)$, $x \in \mathbb{R}$, which has a finite second moment. This approximation asserts that

$$\mathbb{P}(T_{n,\tau} > c) \approx 1 - \Phi\left(c + \sqrt{n} G^{-1}(\mathbb{P}(X < \tau))\right), \tag{3.7.3.1}$$

where $G^{-1}(q)$ denotes the quantile function associated to the distribution function $G(x)$ of the standardized output power $(X - \mathbb{E}(X))/\sqrt{\mathrm{Var}(X)}$. Here $\Phi(x)$ is the distribution function of the standard normal distribution.

The fundamental approximation (3.7.3.1) still requires the knowledge of $G(x)$. Thus it provides a valuable tool for the theoretical and numerical analysis, but it cannot be used in photovoltaic applications, since $G(x)$ is known only in rare cases. Therefore, we need to establish a more involved approximation which takes into account an additional sample $Y_1, \ldots, Y_m$ of m independent and identically distributed flash measurements. It is worth mentioning that the observations $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ may be dependent, that frequently occurs in practice. In view of systematic effects due to different measurement systems, we assume that those flash data follow the location-scale *measurement model*

$$Y = \sigma X + \delta$$

for a location shift δ and a scale parameter $\lambda > 0$, where Y denotes a generic flash measurement and X is a generic control measurement. Both δ and σ are not required to be known in what follows. Let

$$Z_i = \frac{Y_i - \overline{Y}_m}{S_m}, \qquad i = 1, \ldots, m,$$

be the standardized flash data, where $\overline{Y}_m = \frac{1}{m}\sum_{i=1}^m Y_i$ and $S_m^2 = \frac{1}{m-1}\sum_{i=1}^m (Y_i - \overline{Y}_m)^2$. Thus, the function $G^{-1}(q)$ in (3.7.3.1) should be replaced by an estimate, for example, the empirical quantile function for the sample $Z_1, \ldots, Z_m$ or a smoothed quantile estimate, see (20) for details.

Using (3.7.3.1) we obtain

$$\mathrm{OC}(p_1, p_2) \cong \Phi\big(-c + \sqrt{n}G^{-1}(1-p_2)\big) - \Phi\big(c + \sqrt{n}G^{-1}(p_1)\big). \tag{3.7.3.2}$$

A sampling plan (n, c) solving (3.7.2.1) with $\mathrm{OC}(p_1, p_2)$ replaced by the approximation (3.7.3.2) and minimal n is called an *(asymptotically) optimal sampling plan*. To simplify notation, we shall denote the approximation (3.7.3.2) by $\mathrm{OC}(p_1, p_2)$ in the sequel.

In Figure 3.7.1 the OC curve is depicted for the two sampling plans $(n, c) = (27, 11)$ and $(n, c) = (65, 15)$ assuming normally distributed measurements. Note that $\mathrm{OC}(p_1, p_2)$ becomes sharper as n increases and yields negative values for some p_1 and p_2 since the inequality $-c + \sqrt{n}G^{-1}(1-p_2) > c + \sqrt{n}G^{-1}(p_1)$ is violated for large enough p_1 and p_2. Therefore, we restrict $\mathrm{OC}(p_1, p_2)$ in (3.7.3.2) by 0 from below that is shown in Figure 3.7.1.

3.7.3.2 Nonlinear equations for optimal sampling plans

An optimal sampling plan exists due to the monotonicity of $\mathrm{OC}(p_1, p_2)$ as a function of (p_1, p_2) and the smoothness of $\mathrm{OC}(p_1, p_2)$ with respect to n (regarded as a real variable) and c. Moreover, since $\mathrm{OC}(p_1, p_2)$ decreases w.r.t. p_1 for fixed p_2 and decreases w.r.t. p_2 for fixed p_1, a minimal value of n is attained when the inequalities for $p_1 + p_2$ hold as equalities. Thus, we need to solve the system of nonlinear inequalities with equality constraints,

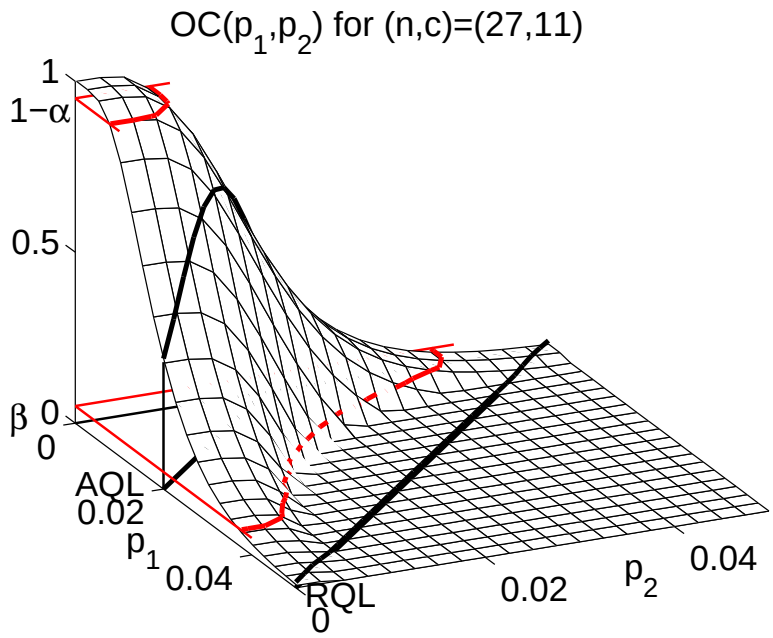
$$\begin{cases} \Phi(-c + \sqrt{n}G^{-1}(1-p_2)) - \Phi(c + \sqrt{n}G^{-1}(p_1)) \geq 1 - \alpha \text{ for } p_1 + p_2 = \mathrm{AQL}, \\ \Phi(-c + \sqrt{n}G^{-1}(1-p_2)) - \Phi(c + \sqrt{n}G^{-1}(p_1)) \leq \beta \text{ for } p_1 + p_2 = \mathrm{RQL}. \end{cases} \tag{3.7.3.3}$$

Unfortunately, these equations can not be solved explicitly. For given parameters AQL, RQL, α and β, the system (3.7.3.3) have to be solved numerically leading to a sampling plan

$$\big(n(p_1, p_2), c(p_1, p_2)\big).$$

In Figure 3.7.1 the line segment corresponding to the constraints $p_1 + p_2 = \mathrm{AQL}$ and $p_1 + p_2 = \mathrm{RQL}$ have been added and it is shown how the requirements to control the consumer's and producer's risk correspond to nonlinear submanifolds of the OC surface.

The system (3.7.3.3) where all possible p_1 and p_2 are considered is too general for describing reality. In practice, we can expect that the ratio $\gamma = p_2/p_1$ is fixed. For instance, one may estimate this ratio from flash data or historical lab samples or use expert's



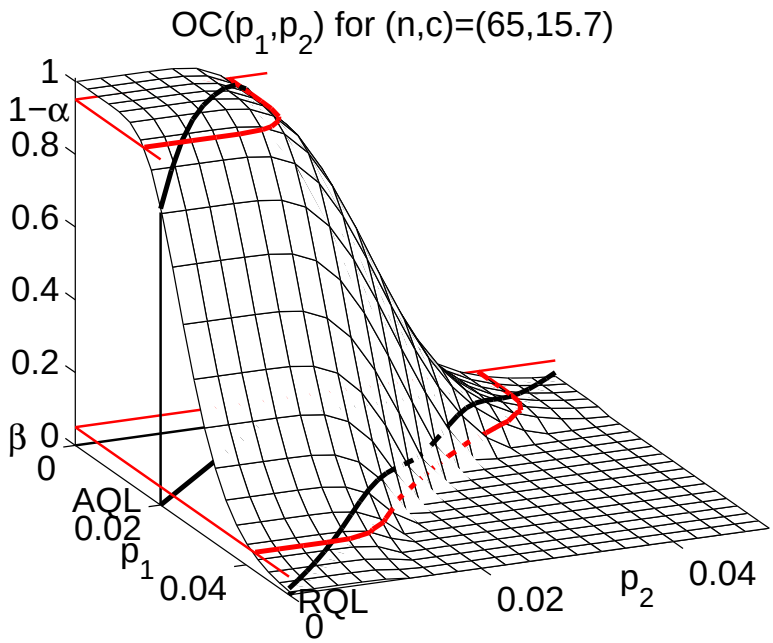


Abbildung 3.7.1: The surface of the operating characteristic, $\mathrm{OC}(p_1, p_2)$, for the out-of-spec formulation and the sampling plans $(n, c) = (27, 11)$ and $(n, c) = (65, 15.7)$, $(p_1, p_2) \in [0, 0.05]^2$ and observations are normally distributed.

judgements. Thus, for fixed γ we have the system of the nonlinear equations

$$\begin{cases} \Phi\left(-c+\sqrt{n}G^{-1}\left(1-\frac{\gamma\mathrm{AQL}}{1+\gamma}\right)\right)-\Phi\left(c+\sqrt{n}G^{-1}\left(\frac{\mathrm{AQL}}{1+\gamma}\right)\right) &= 1-\alpha \\ \Phi\left(-c+\sqrt{n}G^{-1}\left(1-\frac{\gamma\mathrm{RQL}}{1+\gamma}\right)\right)-\Phi\left(c+\sqrt{n}G^{-1}\left(\frac{\mathrm{RQL}}{1+\gamma}\right)\right) &= \beta, \end{cases} \tag{3.7.3.4}$$

whose solution gives the sampling plan

$$(n(\gamma), c(\gamma))\,.$$

The question arises whether one-sided acceptance sampling plans appear as special cases. By standard properties of quantile functions, we have $G_m^{-1}(1-\frac{\gamma\mathrm{AQL}}{1+\gamma}) \to \infty$ and $G^{-1}(1-\frac{\gamma\mathrm{RQL}}{1+\gamma}) \to \infty$ as $\gamma \to 0$. Therefore, the sampling plan $(n(0), c(0))$ equals the sampling plan for the left-sided specification limit when PV modules are regarded as non-conforming if $X < \tau_1$. Similarly, the sampling plan $(n(\infty), c(\infty))$ equals the sampling plan for the right-sided specification limit when PV modules are regarded as non-conforming if $X > \tau_2$.

It can be proved that the following result holds for the sampling plan $(n(\gamma), c(\gamma))$.

Proposition 3.7.3.1. *If $G(x)$ is symmetric, then*

$$(n(\gamma), c(\gamma)) = (n(1/\gamma), c(1/\gamma))$$

for any $\gamma > 0$.

3.7.3.3 Sampling plans under symmetry

The methodology described above is valid under very general assumptions. Particularly, there is no need for restrictive assumptions on the shape of the distribution of the output power. However, it turns out that the nonlinear equations, which have to be solved numerically in the general case, become substantially simpler under symmetry and allow for explicit formulas.

Let us make two symmetry assumptions. First, we assume that the distribution function is symmetric,

$$G^{-1}(1-p) = 1 - G^{-1}(p), \qquad \text{for all } p\ .$$

When G has a density function g, i.e. $G(x) = \int_{-\infty}^{x} g(u)\,du$, the above assumption can be rewritten in the more classical way as

$$g(x) = g(-x), \qquad \text{for all } x.$$

Second, we assume that the fraction of underperforming modules is equal to the fraction of overperforming modules, i.e.

$$p_1 = p_2,$$

such that the equations (3.7.3.3) are parameterized by p_1. Under these two assumptions, the system (3.7.3.3) can be simplified to the form

$$\begin{cases} 1 - 2\Phi(c + \sqrt{n}G^{-1}(p_1)) = 1 - \alpha \text{ for } p_1 = \text{AQL}/2, \\ 1 - 2\Phi(c + \sqrt{n}G^{-1}(p_1)) = \beta \text{ for } p_1 = \text{RQL}/2. \end{cases}$$

Collecting terms and applying the inversion Φ^{-1} to both sides, we obtain the system

$$\begin{cases} c + \sqrt{n}G^{-1}(\text{AQL}/2) = \Phi^{-1}(\alpha/2), \\ c + \sqrt{n}G^{-1}(\text{RQL}/2) = \Phi^{-1}((1-\beta)/2), \end{cases}$$

whose solution is

$$\begin{aligned} n &= \left\lceil \frac{\left(\Phi^{-1}(\alpha/2) - \Phi^{-1}((1-\beta)/2)\right)^2}{\left(G^{-1}(\text{AQL}/2) - G^{-1}(\text{RQL}/2)\right)^2} \right\rceil, \\ c &= -\frac{\sqrt{n}}{2}\left(G^{-1}(\text{AQL}/2) + G^{-1}(\text{RQL}/2)\right). \end{aligned}$$

These formulas are analogous to the expression of sampling plans which was found in (20).

3.7.4 Numerical results

To illustrate the proposed methodology, its scope and limits, and make a comparison with previous approaches, we study various sampling plans numerically. First, we were interested in the true sampling plans in the general case and particularly how strongly the sampling plan depends on γ, the guess for the ratio of the fractions p_1 and p_2 of under- and overperforming modules, respectively. Second, we illustrate how the sampling plans under symmetry for the out-of-spec formulation are related to the sampling plans for the one-sided formulation.

3.7.4.1 General case

In our simulation study, we consider the following models of the distribution of observations.

$$
\begin{aligned}
\text{model 11:} \quad & X_i \sim \mathrm{N}(220, 4) \\
\text{model 12:} \quad & X_i \sim 0.1\mathrm{N}(212, 6) + 0.9\mathrm{N}(220, 4) \\
\text{model 13:} \quad & X_i \sim 0.9\mathrm{N}(220, 4) + 0.1\mathrm{N}(228, 8) \\
\text{model 14:} \quad & X_i \sim 0.6\mathrm{N}(220, 12) + 0.4\mathrm{N}(220, 2) \\
\text{model 15:} \quad & X_i \sim 0.2\mathrm{N}(212, 4) + 0.6\mathrm{N}(220, 8) + 0.2\mathrm{N}(228, 6)
\end{aligned}
$$

The densities of distributions for these models are depicted Figures 3.7.2 and 3.7.3. Note that models 12 and 13 have asymmetric distributions, whereas distributions for other models are symmetric.

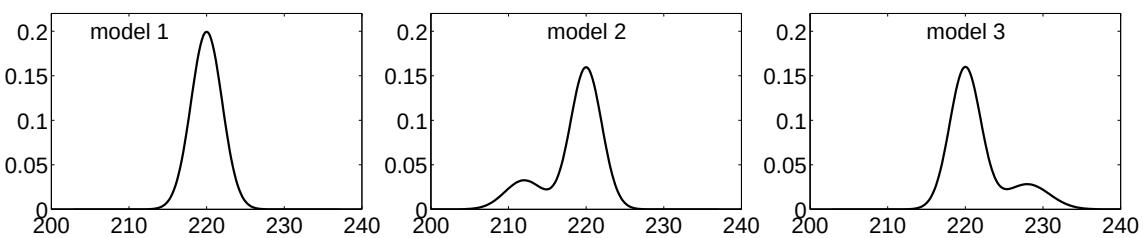


Abbildung 3.7.2: Densities of distributions for models 11–13.

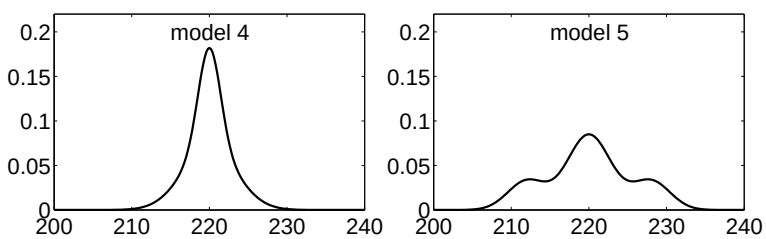


Abbildung 3.7.3: Densities of distributions for models 14 and 15.

Also we consider values of error probabilities given by $\alpha = \beta = 5\%$ and quality parameters given by AQL $= 2\%$ and RQL $= 5\%$, which are frequently used in practice. Due to the constraint $0 \leq \gamma < \infty$, we consider values of $\gamma \in \{1/6, 1/4, 1/2, 1, 2, 4, 6\}$ to study the effect of symmetric and extreme asymmetric fractions of under- or overperforming modules.

In Table 3.7.1 we represent the optimal sampling plans $(n(\gamma), c(\gamma))$ for models 11–15. We can see in Table 3.7.1 that the optimal sampling plans essentially depends on γ. For

example, for model 11 the sampling plans with $\gamma = 1$ is about 2.5 times than the sampling plan with $\gamma \leq 1/4$ and $\gamma \geq 4$. This is also held for models 14 and 15, which corresponds to symmetrical distributions like model 11. For models with asymmetrical distributions, the sampling plan is typically large for many γ.

Tabelle 3.7.1: The optimal sampling plans $(n(\gamma), c(\gamma))$ for models 11–15, $\alpha = \beta = 5\%$, AQL = 2% and RQL = 5%. Last columns contains optimal sampling plans for the left-sided formulation.

	out-of-spec							left-sided
γ	1/6	1/4	1/2	1	2	4	6	
model 11								
$n(\gamma)$	68	69	43	27	43	69	68	65
$c(\gamma)$	15.8	16.2	12.9	10.1	12.9	16.2	15.8	14.9
model 12								
$n(\gamma)$	383	354	311	276	252	237	231	72
$c(\gamma)$	34.1	31.6	27.8	24.8	22.7	21.3	20.8	19.5
model 13								
$n(\gamma)$	328	337	358	392	441	503	544	311
$c(\gamma)$	22.4	22.9	24.3	26.5	29.6	33.5	36.1	21.2
model 14								
$n(\gamma)$	38	39	24	15	24	39	38	36
$c(\gamma)$	12.7	13.1	10.5	8.2	10.5	13.1	12.7	11.8
model 15								
$n(\gamma)$	179	186	124	81	124	186	179	162
$c(\gamma)$	24.9	25.6	21.0	16.9	21.0	25.6	24.9	23.1

Let us now compare sampling plans for the one-sided and out-of-spec formulation. In Table 3.7.1 we show these sampling plans for symmetric models and $\alpha = \beta = 5\%$, AQL = 2% and RQL = 5%. We can observe the sampling plan size for the out-of-spec formulation with $\gamma = 1$ and symmetrical distributions is smaller than the sampling plan size for the left-sided formulation. This happens because the true probabilities of non-conforming PV modules are different for the same left control limit but AQL and RQL for the two-sided specification limits are the same as for the one-sided specification limit.

In Figure 3.7.4 the sampling plan $(n(\gamma), c(\gamma))$ is depicted as a function of $\gamma \in [0, 1]$ for model 11. We can observe that the sampling plan slightly increases for small γ that is due to a wiggle shape of the contour line of $\mathrm{OC}(p_1, p_2)$ at the level β as shown in Figure 3.7.1. Starting at $\gamma = 0.35$, the sampling plan faster decreases to a minimal value. For $\gamma > 1$ the sampling plan for model 11 can be found by the relation from Proposition since model 11 is symmetrical.

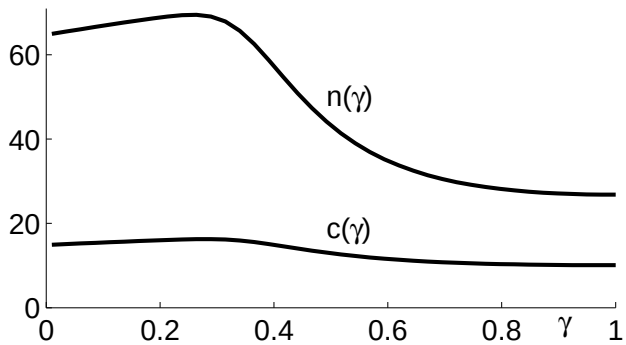


Abbildung 3.7.4: The sampling plan $(n(\gamma), c(\gamma))$ as a function of γ for model 11.

At first glance, the fact that Figure 3.7.4 shows that $n(\gamma)$ is not monotone in γ is puzzling. This effect can be explained as follows. In general, the OC surface is relatively wavy as can be seen from Figure 3.7.1. The OC surface gets steeper, if n increases, and for large enough n, the corresponding point $\mathrm{OC}(p_1, p_2)$ on the OC surface satisfies the constraint, say $\mathrm{OC}(p_1, p_2) \geq 1 - \alpha$ for $p_1 + p_2 = \mathrm{AQL}$. Note that γ fixes a point on that line $p_1 + p_2 = \mathrm{AQL}$, thus fixing a point on the OC surface. If we change γ, that point changes as well, and, depending on how the gradient of the OC surface changes, the minimal value of n ensuring the constraints may get smaller or larger.

3.7.5 Examples

Let us consider two sets of flash observations corresponding to two lots. These sets are provided by TÜV Immissionsschutz und Energiesysteme GmbH. Let $\alpha = \beta = 5\%$, AQL = 2%, and RQL = 5%, and the method of the smoothed empirical quantiles is used to estimate $G^{-1}(\cdot)$, see (20).

The histogram for the lot 1 consisting of 386 flash observations is shown in Figure 3.7.5. For purpose of illustration suppose that the specification interval is $[\tau_1, \tau_2] = [176, 180]$. Now

we should determine γ to compute the sampling plan. We estimate γ as the ratio of the number of flash measurements above τ_2 to the number of flash measurements below τ_1, that gives $\hat{\gamma} = 70/32 \approx 2.19$. Then solving the system (3.7.3.4), we obtain that the sampling plan for the out-of-spec formulation is

$$(n, c) = (11, 6.99).$$

For sake of comparison, the sampling plan for the left-sided formulation is

$$(n, c) = (47, 12.97),$$

which requires more PV modules to be taken for the control sample.

Let us now study the sensitivity of the sampling plan with respect to γ. We need this because inaccuracy of estimation of γ from flash data. Using bootstrap we obtain that the confidence interval for γ is $[2.01, 3.14]$. In Table 3.7.2 we present the sampling plans for γ varying around 2.19. We can observe that the sampling plan slightly depends on γ from the confidence interval and even for the larger interval $[1.5, 4]$.

Tabelle 3.7.2: The sampling plans $(n(\gamma), c(\gamma))$ for the lot 1, $\alpha = \beta = 5\%$, AQL = 2% and RQL = 5%.

γ	1	1.5	2	3	4
$n(\gamma)$	20	13	11	12	14
$c(\gamma)$	9.5	7.5	7.0	7.3	8.2

In the second example, we consider the lot 2 consisting of 126 flash observations, whose histogram is shown in Figure 3.7.6. Suppose that the specification interval is $[\tau_1, \tau_2] = [180, 183]$. Acting as above, we obtain that the estimate of γ is $\hat{\gamma} = 7/29 \approx 0.24$ with the confidence interval $[0.20, 0.57]$ and the sampling plan for the out-of-spec formulation for this value of γ is

$$(n, c) = (44, 14.05).$$

Note that the sampling plan for the left-sided formulation is

$$(n, c) = (80, 20.28),$$

which also gives a larger size for the control sample.

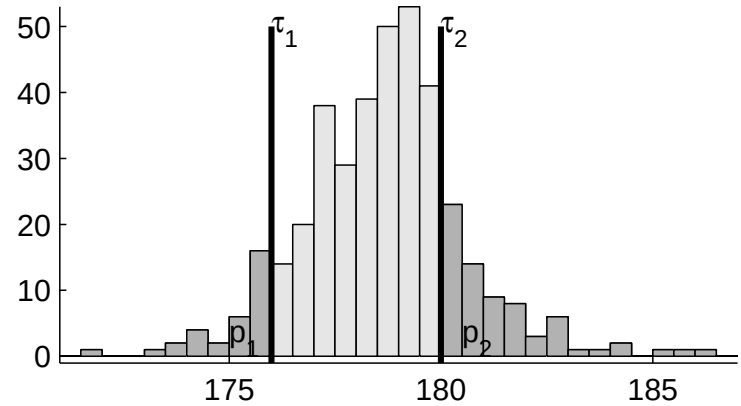


Abbildung 3.7.5: Histogram of flash observations for the lot 1.

To verify the robustness of the sampling plans, we show plans for γ varying around 0.24 in Table 3.7.3. We can see that the sampling plan slightly depends on γ from the confidence interval.

Tabelle 3.7.3: The sampling plans $(n(\gamma), c(\gamma))$ for the lot 2, $\alpha = \beta = 5\%$, AQL = 2% and RQL = 5%.

γ	0.1	0.15	0.2	0.3	0.4	0.5	0.6
$n(\gamma)$	80	56	47	44	49	59	72
$c(\gamma)$	19.5	16.2	14.6	13.9	14.5	15.8	17.4

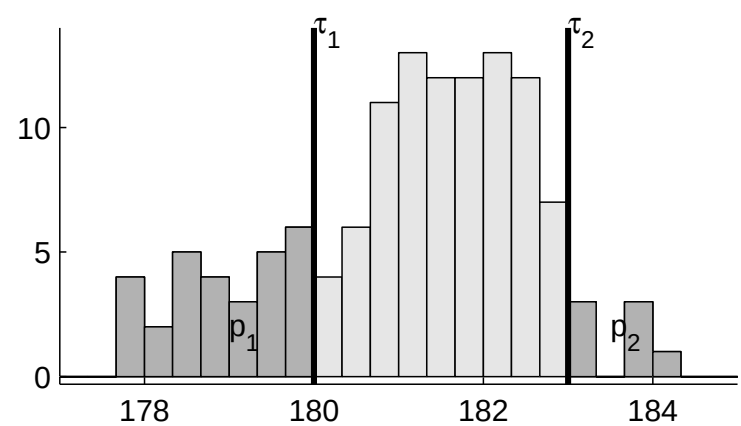


Abbildung 3.7.6: Histogram of flash observations for the lot 2.

Kapitel 3.8

Benchmarking

In the present chapter we perform benchmarking, that is, we investigate the efficiency of different estimators of construction of sampling plans for real-life distributions rather than nine models considered in previous chapters. In our study we consider several real flasher lists as models of distributions of power measurements. Histograms of these flasher lists are shown in Figure 3.8.1. Distributions of flasher lists are typically unimodal except the flasher list 2, which has a bimodal distribution. Note that flasher lists 1, 5 and 7 are rather close to the normal distribution and flasher lists 2, 4 and 6 have heavy tails compared to tails of the normal distribution.

To proceed with benchmarking, we generate samples of size $m = 100, 250, 500$ using the technique of smoothed bootstrap. Specifically, we simulate samples from a distribution given by the kernel density estimator with BCV bandwidth for a specified flasher list. In Table 3.8.1 we present the accuracy of different estimators measured in terms of the root mean square deviations. We can see that the double kernel estimator is typically better than the kernel estimator especially for small sample sizes. The double kernel estimator is worse than the kernel estimator for the flasher list 5 only. Note that the BCV, ICV and GPS bandwidth are always better than the LSCV bandwidth and typically outperform the SJPI bandwidth. We can also see that the SSA1c estimator behaves similarly to the kernel estimator with GPS bandwidth.

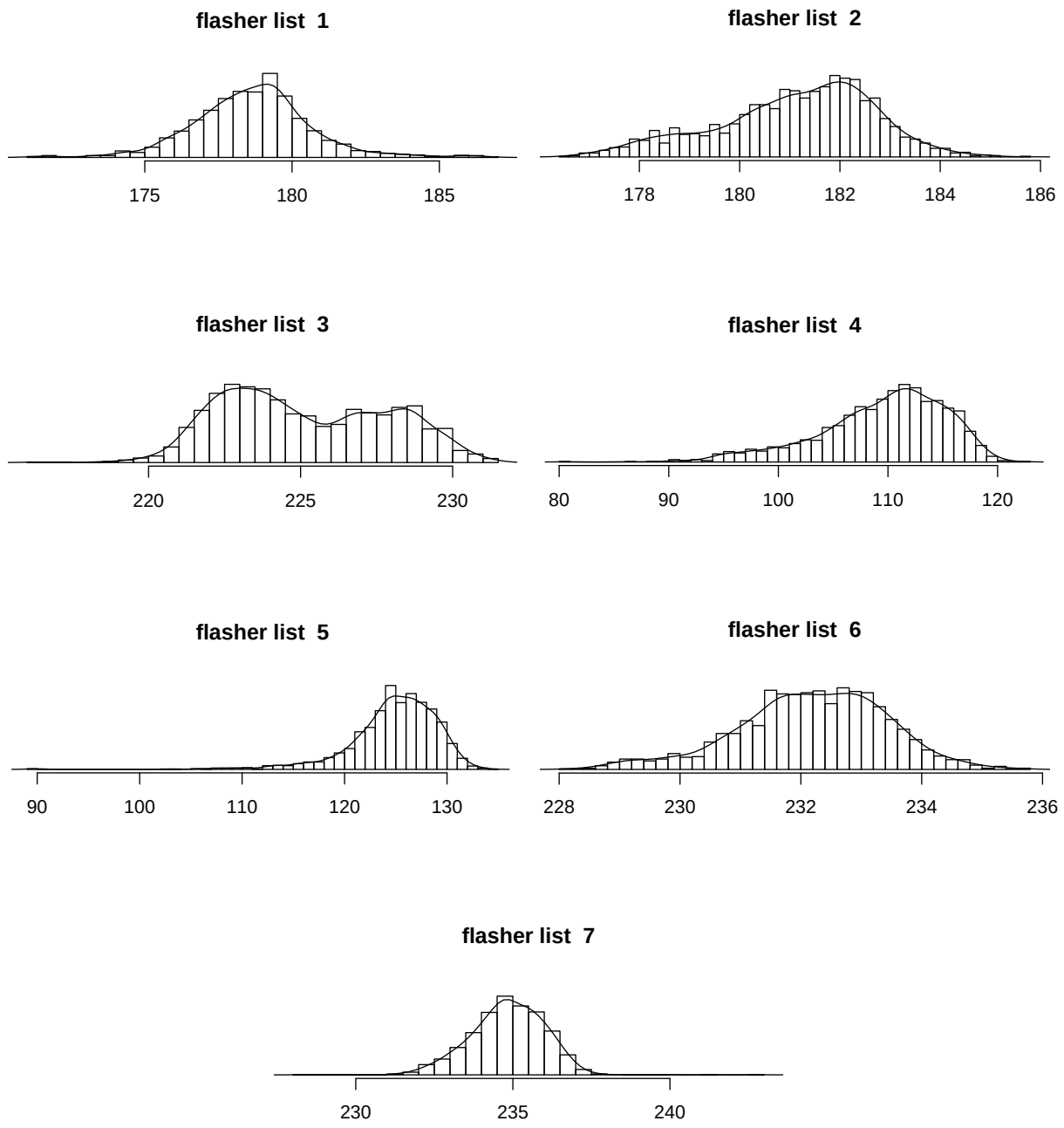


Abbildung 3.8.1: Histograms of real flasher lists.

Tabelle 3.8.1: RMSD of the distribution of the sampling plan size using the classical kernel estimator with different bandwidths and the double kernel estimator with LSCV, BCV, SJPI, ICV, GPS bandwidths and the SSA1c estimator for models 1-9.

		classical kernel estimator					double kernel estimator					SSA1c
flasher list	m	LSCV	BCV	SJPI	ICV	GPS	LSCV	BCV	SJPI	ICV	GPS	
1	100	38.1	23.9	32.7	24.6	29.5	35.9	18.3	28.5	18.7	26.0	29.6
1	250	24.0	19.8	23.0	19.6	21.9	25.9	20.0	24.6	19.4	23.4	21.0
1	500	18.2	15.8	17.2	15.5	16.8	21.1	17.6	19.7	17.1	19.2	16.0
2	100	47.5	38.3	36.9	38.1	39.1	47.4	45.5	38.4	45.3	43.7	39.3
2	250	35.0	31.2	30.0	31.5	31.5	35.6	34.5	30.8	35.2	34.5	32.3
2	500	27.6	25.3	24.6	25.6	25.7	28.3	27.0	25.0	27.6	27.7	26.8
3	100	93.4	52.3	60.9	50.9	74.0	78.0	52.4	45.5	50.8	59.7	68.1
3	250	58.3	45.4	48.3	44.1	55.6	54.7	39.4	41.7	38.1	51.1	51.3
3	500	45.3	39.2	40.5	38.4	44.4	44.8	35.7	37.5	34.8	42.3	41.5
4	100	25.9	18.2	22.9	18.0	21.4	24.5	15.6	20.8	15.4	19.3	20.7
4	250	17.1	12.9	14.7	12.8	14.2	18.6	13.5	15.8	13.2	15.4	13.5
4	500	11.2	9.7	10.5	9.6	10.3	12.9	10.8	11.9	10.7	11.6	9.7
5	100	18.0	14.7	17.3	14.7	17.7	26.8	24.9	26.4	24.4	25.4	17.2
5	250	10.3	9.2	10.0	9.1	10.2	17.9	15.7	16.7	15.6	17.2	9.8
5	500	6.8	6.2	6.6	6.2	6.8	12.0	11.1	11.5	11.2	11.7	6.5
6	100	27.2	18.9	22.4	19.1	19.7	23.6	15.0	19.0	15.3	16.1	19.6
6	250	14.9	12.4	14.0	12.5	12.7	14.6	11.6	13.6	11.8	12.1	12.0
6	500	10.9	9.4	10.2	9.4	9.5	10.8	9.2	10.1	9.2	9.3	8.9
7	100	30.8	21.0	24.9	21.2	22.5	28.8	16.0	19.1	15.9	18.3	23.2
7	250	20.2	16.1	17.5	16.2	17.2	18.4	13.3	15.0	13.3	15.7	17.2
7	500	15.0	13.1	13.8	13.1	14.0	13.8	11.0	12.2	11.0	13.0	14.0

3.8.1 Combined comparison of estimators

For sake of convenience of comparison, we compute relative inaccuracies defined as the ratio of the RMSD of estimators to the RMSD of the best estimator for each model of the distribution, which are shown in Table 3.8.2. We also define the combined accuracy of estimators by taking the geometric mean of relative inaccuracies across different models including flasher lists. In Figures 3.8.2–3.8.5 we depict the paired comparison of relative inaccuracies where the kernel estimator with LSCV bandwidth is used as the principal method which was developed and implemented in APOS software before the beginning of the present work. .

Tabelle 3.8.2: Relative inaccuracies of sampling plan size estimation using the classical kernel estimator with different bandwidths and the double kernel estimator with LSCV, BCV, SJPI, ICV, GPS bandwidths and the SSA1c estimator for samples of size $m = 250$.

		classical kernel estimator					double kernel estimator					SSA1c
	m	LSCV	BCV	SJPI	ICV	GPS	LSCV	BCV	SJPI	ICV	GPS	
model 1	250	1.46	1.21	1.32	1.22	1.20	1.32	1	1.10	1.01	1.03	1.15
model 2	250	1.25	1.10	1.18	1.07	1.17	1.20	1.02	1.12	1	1.09	1.05
model 3	250	1.33	1.10	1.19	1.11	1.22	1.24	1	1.03	1.01	1.06	1.13
model 4	250	1.39	1.18	1.20	1.13	1.17	1.24	1.03	1.06	1	1.03	1.06
model 5	250	1.41	1.22	1.23	1.17	1.13	1.25	1.07	1.08	1.03	1	1.04
model 6	250	1.27	1.09	1.09	1.08	1.16	1.15	1.02	1	1.04	1.05	1.10
model 7	250	1.18	1.04	1.04	1.04	1.08	1.08	1.01	1	1.02	1.03	1.06
model 8	250	1.25	1.05	1.17	1.01	1.07	1.36	1.09	1.32	1.05	1.14	1
model 9	250	1.14	1.26	1	1.24	1.25	1.15	1.47	1.09	1.46	1.47	1.30
flasher list 1	250	1.24	1.02	1.19	1.01	1.13	1.34	1.03	1.27	1	1.21	1.08
flasher list 2	250	1.17	1.04	1	1.05	1.05	1.19	1.15	1.03	1.17	1.15	1.08
flasher list 3	250	1.53	1.19	1.27	1.16	1.46	1.44	1.03	1.09	1	1.34	1.35
flasher list 4	250	1.34	1.01	1.15	1	1.11	1.45	1.05	1.23	1.03	1.20	1.05
flasher list 5	250	1.13	1.01	1.10	1	1.12	1.97	1.73	1.84	1.71	1.89	1.08
flasher list 6	250	1.28	1.07	1.21	1.08	1.09	1.26	1	1.17	1.02	1.04	1.03
flasher list 7	250	1.52	1.21	1.32	1.22	1.29	1.38	1	1.13	1	1.18	1.29
geometric mean		1.30	1.11	1.16	1.10	1.16	1.30	1.09	1.15	1.08	1.17	1.11

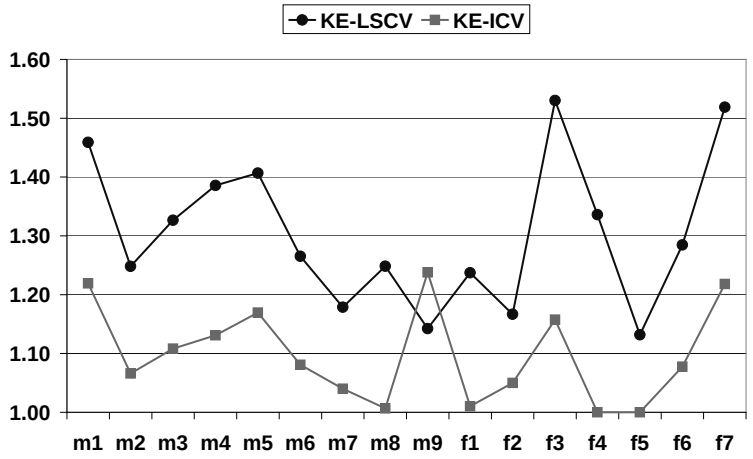


Abbildung 3.8.2: Relative inaccuracies of sampling plan size estimation for the kernel estimator with LSCV bandwidth (shortly, KE-LSCV, marked as the modified procedure in APOS software) and the kernel estimator with ICV bandwidth (shortly, KE-ICV, marked as the kernel ICV method in APOS software).

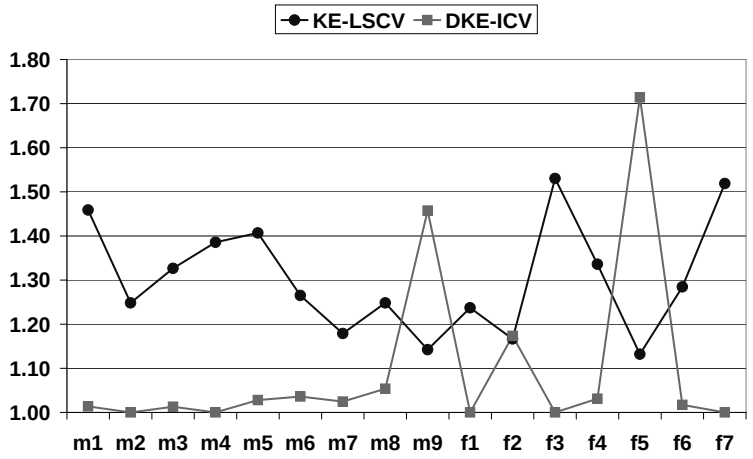


Abbildung 3.8.3: Relative inaccuracies of sampling plan size estimation for the kernel estimator with LSCV bandwidth (shortly, KE-LSCV, marked as the modified procedure in APOS software) and the double kernel estimator with ICV bandwidth (shortly, DKE-ICV, marked as the double kernel method in APOS software).

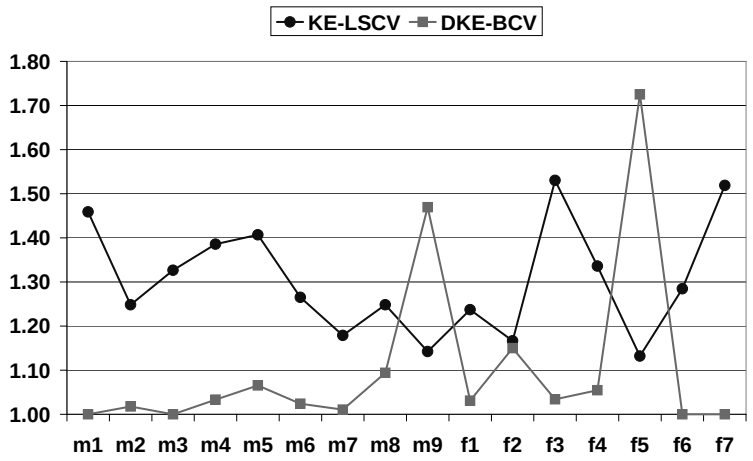


Abbildung 3.8.4: Relative inaccuracies of sampling plan size estimation for the kernel estimator with LSCV bandwidth (shortly, KE-LSCV, marked as the modified procedure in APOS software) and the double kernel estimator with BCV bandwidth (shortly, DKE-BCV).

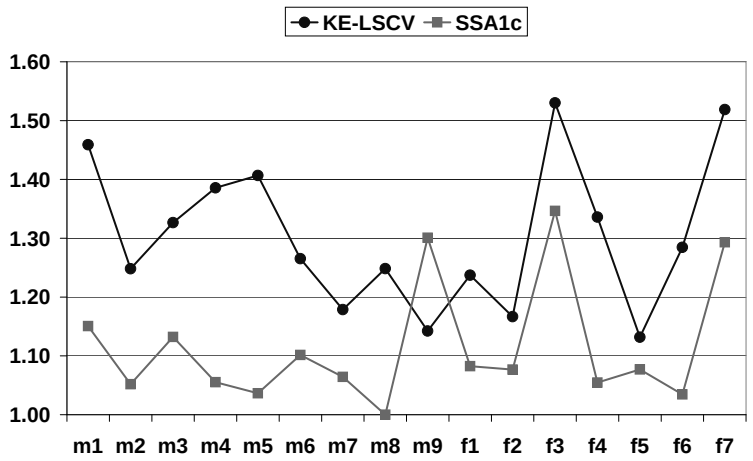


Abbildung 3.8.5: Relative inaccuracies of sampling plan size estimation for the kernel estimator with LSCV bandwidth (shortly, KE-LSCV, marked as the modified procedure in APOS software) and the SSA1c estimator.

We can see that the double kernel estimator is more accurate than the kernel estimator except the model 9 and the flasher list 5. We also observe that the LSCV bandwidth is

worse than other bandwidths for all models and flasher list. Meanwhile, the ICV, BCV and GPS bandwidths are typically better that the SJPI bandwidth except the model 9 and the flasher list 2. The SSA1c estimator is rather good except the model 9 and flasher lists 3 and 7.

In Figure 3.8.2 we can see that the kernel estimator with ICV bandwidth (KE-ICV) is uniformly better than the kernel estimator with LSCV bandwidth (shortly, KE-LSCV) except the model 9. However, the KE-ICV is not the best possible estimator since it's relative inaccuracy varies around 1.1 (not close to 1).

In Figure 3.8.3 we observe that the double kernel estimator with ICV bandwidth (DKE-ICV) has the smallest inaccuracy for models 1–8 and flasher lists 1, 3, 4, 6, and 7. The DKE-ICV does not perform well for the flasher list 5 because it's histogram has a very narrow long left tail which is truncated in the double kernel estimator. In Figure 3.8.4 we can see that the performance of the double kernel estimator with BCV bandwidth is similar to the DKE-ICV.

In Figure 3.8.5 we observe that the SSA1c estimator is uniformly better than the KE-LSCV except model 9 and the relative inaccuracy of the SSA1c estimator varies around 1.1 that is similar to the behavior of the KE-ICV.

For convenience, in Figure 3.8.6 we depict relative inaccuracies of all estimators including the estimator based on the empirical distribution function which is also implemented in APOS software. Using the combined accuracy, we note that estimators can be sorted in term of their performance as follows.

1. The double kernel estimator with ICV bandwidth.
2. The double kernel estimator with BCV bandwidth.
3. The kernel estimator with ICV bandwidth.
4. The kernel estimator with BCV bandwidth and the SSA1c estimator.
5. The double kernel estimator with SJPI bandwidth.
6. The kernel estimator with SJPI and GPS bandwidth.
7. The double kernel estimator with GPS bandwidth.
8. The double kernel estimator with LSCV bandwidth and the kernel estimator with LSCV bandwidth.

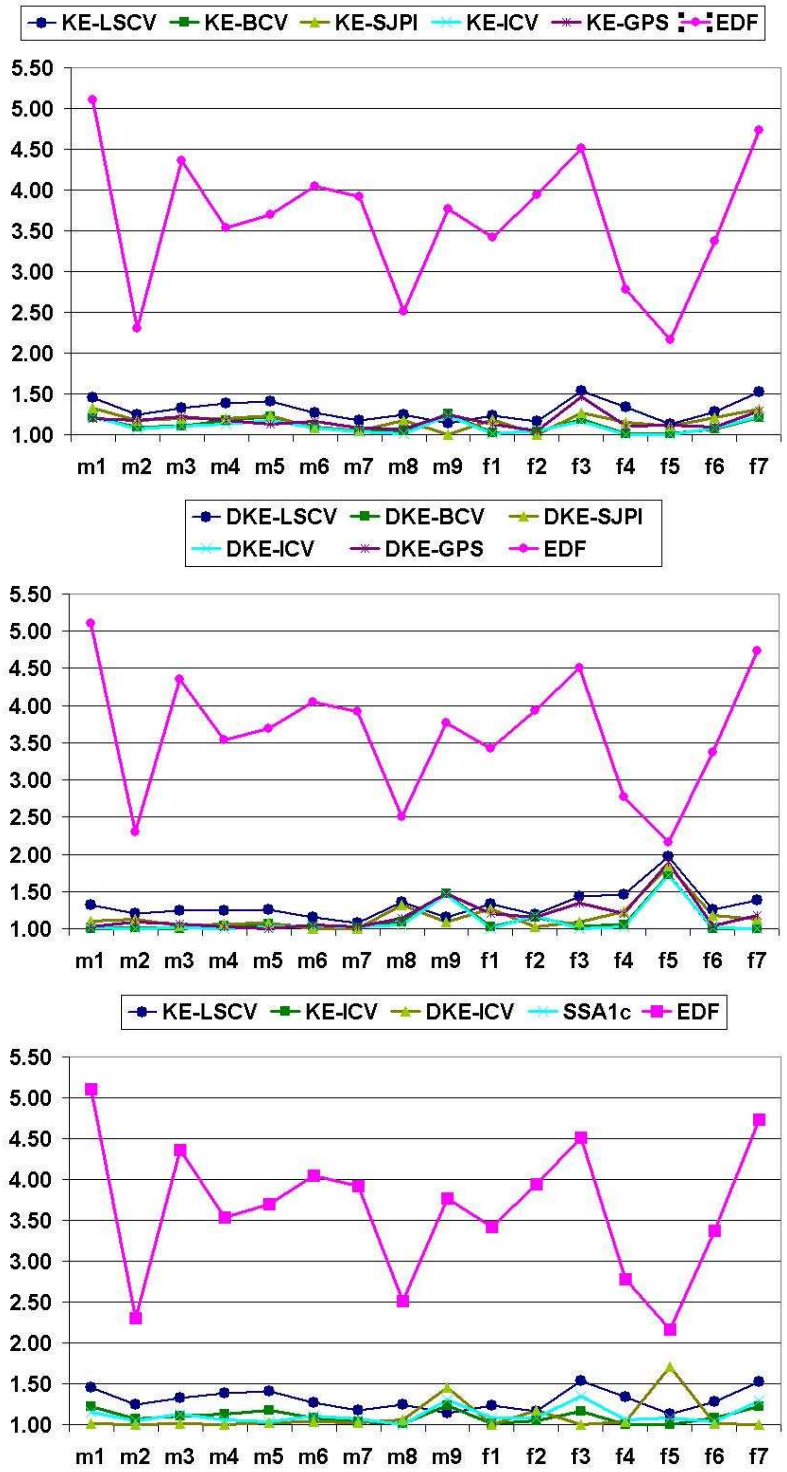


Abbildung 3.8.6: Relative inaccuracies of sampling plan size estimation for all main estimators.

3.8.2 Sampling plans in APOS software

We prepared a new version of APOS software by implementing the double kernel estimator with ICV bandwidth and the kernel estimator with ICV bandwidth for computing sampling plans. In Table 3.8.3 we show the sampling plan size computed in APOS software for 7 flasher lists.

We can see that the essential difference between methods exist for flasher lists 2, 3 and 5. The double kernel estimator with ICV bandwidth gives a smaller sampling plan size for flasher list 2 and does not give too small sampling plan size for flasher list 5. However, the double kernel estimator with ICV bandwidth yields a bigger sampling plan size for flasher list 3.

Tabelle 3.8.3: The sampling plan size computed in APOS software for 7 flasher lists, m is the number of measurements in a flasher list.

	m	standard	modified	kernel ICV	double kernel ICV
AQL = 1%, RQL = 5%					
flasher list 1	386	15	17	17	21
flasher list 2	126	82	37	34	28
flasher list 3	518	28	36	36	50
flasher list 4	12767	13	12	12	13
flasher list 5	713	7	5	5	11
flasher list 6	5140	14	14	14	14
flasher list 7	5140	23	22	21	23
AQL = 2%, RQL = 5%					
flasher list 1	386	30	47	47	56
flasher list 2	126	100	89	82	69
flasher list 3	518	160	126	127	152
flasher list 4	12767	29	30	30	31
flasher list 5	713	16	15	15	23
flasher list 6	5140	31	33	32	33
flasher list 7	5140	59	53	54	57

Kapitel 3.9

Conclusion

The present work is devoted to variability-reducing quality control methods in photovoltaics. The primary instrument in lot based quality control are acceptance sampling plans. In practice the appropriate sampling plans are highly important because the sample size control the producers risk of rejecting PV modules of high quality and consumers risk of receiving PV modules of low quality.

The main attention in this work is paid to the case when the measurements follow an arbitrary continuous distribution function rather than the normal distribution which is the traditional assumption in IEC standards. For constructing the acceptance sampling plans in this general situation, we have to use additional data in the form of a historic data set or flasher lists. The accuracy of sampling plan construction depends on how the quantiles for these data are estimated. In our work we considered several classical methods of quantile estimation such as the kernel estimator, the Bernstein polynomial estimator and the orthogonal series estimator, and proposed new estimators.

To be precise, the theoretical results consist of developing several new estimation procedures including the double kernel estimator, the SSA estimator and the Bernstein-Durrmeyer polynomial estimator.

The double kernel estimator essentially has the form of the weighted average of certain kernels with weights constructed on the base of the classical kernel estimator. We proved that the double kernel estimator is consistent for estimating a density and more robust than the classical kernel estimator.

The SSA estimator is a result of the application of Singular Spectrum Analysis to the empirical distribution function sampled at a grid of points spanning the range of the sample.

SSA yields a data-adaptive filter, whose length is a parameter that controls the smoothness of the filtered series. We introduced a data-adaptive algorithm for the automatic selection of a general smoothing parameter, which controls the number of modes of the estimated density. A general uniform error bound is proved for the proposed SSA estimator of the distribution function, which ensures its uniform consistency.

We studied quantile estimation using Bernstein-Durrmeyer polynomials in terms of its MSE and IMSE and proposed an improved estimator based on an error-correction. A crucial issue of this estimation is to select the degree of Bernstein-Durrmeyer polynomials. We developed a data-adaptive approach to choosing the polynomial degree that controls the number of modes of the corresponding density estimator and show its consistency as well as its limiting distribution.

Moreover, we derived the acceptance sampling plans for the out-of-spec formulation. For such setting the sampling plans are solutions of rather involved nonlinear equations. Explicit formulas, which resemble known sampling plans, can only be obtained under symmetry assumptions. Further, the solution depends on the ratio of overperforming modules to underperforming modules. We investigated by numerical studies to which extent the required sample size depends on that ratio and the shape of the underlying output power distribution. The application to real examples indicates that in practice the new approach often results in substantially smaller control samples than sampling plans for the left-sided formulation.

We performed the large simulation study where the classical and new estimators are compared in terms of estimating sampling plans. This study shows that, in general, there is no estimator that is the best for all types of distributions. However, we can order all estimators in the sense of the average accuracy for estimating sampling plans as follows. The first place in such competition is taken by the double kernel estimator with ICV bandwidth, which is typically better than all other estimators. The second place is occupied by the double kernel estimator with GPS and BCV bandwidth. The third place is taken by the SSA1c estimator and the kernel estimator with ICV, GPS and BCV bandwidth.

The developed estimators are implemented in the new version of APOS software. Using this program practitioners can easily compute acceptance sampling plans and analyze samples with laboratory measurements to make a decision whether accept or reject a lot of PV modules.

The obtained results are disseminated in the following papers.

- Meisen S., Pepelyshev A. and Steland A. 2011. Quality assessment in the presence of additional data in photovoltaics. Frontiers in Statistical Quality Control, Vol. 10, 251–274.

- Golyandina N., Pepelyshev A., Steland A., 2012. New approaches to nonparametric density estimation and selection of smoothing parameters. Computational Statistics and Data Analysis, 56(7), 2206–2218.

- Pepelyshev A., Steland A., Avellan-Hampe A. 2012. Acceptance sampling plans for photovoltaic modules with two-sided specification limits. Progress in Photovoltaics. in press.

- Rafajlowicz E., Pepelyshev A., Steland A. 2012. Estimation of the quantile function using Bernstein-Durrmeyer polynomials. submitted.

Literaturverzeichnis

[1] Babu, G. J., Canty, A. J., Chaubey, Y. P. 2002. Application of Bernstein polynomials for smooth estimation of a distribution and density function. J. Statist. Plann. Inference 105, no. 2, 377–392.

[2] Bahadur, R. R. 1966. A note on quantiles in large samples. Ann. Math. Statist. 37, 577–580.

[3] Berkes, I. and Philipp, W., 1977. An almost sure invariance principle for the empirical distribution function of mixing random variables, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 41, 115–137.

[4] Billingsley, P., 1999. Convergence of probability measures. John Wiley & Sons, Inc., New York.

[5] Bowman, A. W. 1984. An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71, 353–360.

[6] Chan N.-H., Lee T. C. M., Peng L. 2010. On nonparametric local inference for density estimation. Computational Statistics & Data Analysis, Volume 54, 509–515.

[7] Ciesielski, Z., 1988. Nonparametric polynomial density estimation. Probab. Math. Statist. 9(1), 1–10.

[8] Cheng, C., 1995. The Bernstein polynomial estimator of a smooth quantile function. Statist. Probab. Lett. 24, no. 4, 321–330.

[9] Durrmeyer, J. L., 1967. Une formule d'inversion de la transformee de Laplace: Application a la theorie des des moments., PhD thesis, These de 3e cycle, Faculte des Sciences de l'Universite de Paris.

[10] Golyandina, N., Nekrutkin, V., Zhigljavsky, A., 2001. Analysis of Time Series Structure: SSA and Related Techniques. London: Chapman & Hall/CRC.

[11] Golyandina N., Pepelyshev A., Steland A., 2012. New approaches to nonparametric density estimation and selection of smoothing parameters. Computational Statistics and Data Analysis, 56(7), 2206–2218.

[12] Good, I.J., Gaskins, R.A., 1980. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. J. Amer. Statist. Assoc. 75, 42–73.

[13] Efromovich, S. 1999. Nonparametric curve estimation. Methods, theory, and applications. Springer-Verlag, New York.

[14] Hart, J. D., Yi, S. 1998. One-sided cross-validation. Journal of the American Statistical Association 93, 620–631.

[15] Heitjan, D. F., Rubin, D. B. 1990. Inference from coarse data via multiple imputation with application to age heaping. Journal of the American Statistical Association 85, 304–314.

[16] Herrmann, W., Althaus, J., Steland, A. and Zaehle, H. 2006. Statistical and experimental methods for assessing the power output specification of PV modules. Proceedings of the 21st European Photovoltaic Solar Energy Conference, 2416–2420.

[17] Herrmann, W., Steland, A. 2010. Evaluation of photovoltaic modules based on sampling inspection using smoothed empirical quantiles. Progress in Photovoltaics, 18, 1–9.

[18] Herrmann W. 2005. Analyses of array losses caused by electrical mismatch of PV modules. 20th European Photovoltaic Solar Energy Conference and Exhibition, June 6–10, 2005, Barcelona, Spain.

[19] Herrmann, W., Steland, A. and Herff, W. 2010. Sampling procedures for the validation of PV module output specification. Proceedings of the 24th European Photovoltaic Solar Energy Conference, Hamburg, Germany, ISBN 3-936338-25-6, 3540-3547, DOI: 10.4229/24thEUPVSEC2009-4AV.3.70

[20] Herrmann, W. and Steland, A. 2010. Evaluation of photovoltaic modules based on sampling inspection using smoothed empirical quantiles. Progress in Photovoltaics 18(1), 1–9.

[21] Hyndman R. J. and Fan Y. 1996. Sample quantiles in statistical packages. The American Statistician 50(4), 361–364.

[22] Jones, M.C., Marron, J.S., Sheather, S.J., 1996. A brief survey of bandwidth selection for density estimation. J. Amer. Statist. Ass. 91, 401–407.

[23] Korobeynikov, A., 2010. Computation- and space-efficient implementation of SSA. Stat. Interface 3, 357–368.

[24] Kiefer, J. 1967. On Bahadurs representation of sample quantiles. Ann. Math. Statist. 38, 1323–1342.

[25] Kuurne, J., Tolvanen A., Hyvärinen J. and Oy, E. 2008. Sweep time, spectral mismatch and ligh soaking in thin film module measurements. PVSC '08. 33rd IEEE, 1–3, ISSN 0160-8371.

[26] Meisen S., Pepelyshev A. and Steland A. 2011. Quality assessment in the presence of additional data in photovoltaics. Frontiers in Statistical Quality Control, Vol. 10, 251–274.

[27] Parzen, E. 1962, On the estimation of probability density function and mode, The Annals of Mathematical Statistics, 33, 1065–107.

[28] Perez, M. J., Palacin, F., 1987. Estimating the quantile function by Bernstein polynomials. Comput. Statist. Data Anal 5, 391–397.

[29] Philipp, W., Pinzur, L., 1980. Almost sure approximation theorems for the multivariate empirical process, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 54, 1–13.

[30] Rafajlowicz, E., Skubalska-Rafajlowicz, E., 1999. Nonparametric regression estimation by Bernstein-Durrmeyer polynomials. PROBASTAT '98 (Smolenice Castle). Tatra Mt. Math. Publ. 17, 227–239.

[31] Rosenblatt, M. 1956. Remarks on some non-parametric estimates of a density function. The Annals of Mathematical Statistics, 27, 832–837.

[32] Roy, J.N., Gariki, G.R. and Nagalakhsmi, V. 2010. Reference module selection criteria for accurate testing of photovoltaic (PV) panels. Solar Energy 84, 32–36.

[33] Ruberto M.N. and Rothwarf, J. 1987. Time-dependent open-circuit voltage in $CuInSe_2$/CdS solar cells: Theory and experiment. J. Appl. Phys. 61(9), 4662–4669.

[34] Rudemo, M. 1982. Empirical choice of histograms and kernel density estimators. Scandinavian Journal of Statistics 9, 65–78.

[35] Savchuk, O. Y., Hart, J. D., Sheather, S. J. 2010. Indirect cross-validation for density estimation. J. Amer. Statist. Assoc. 105, no. 489, 415–423.

[36] Scott, D.W. 1992. Multivariate Density Estimation: Theory,Practice, and Visualization. John Wiley, New York.

[37] Scott, D.W., Terrell, G.R., 1987. Biased and unbiased cross-validation in density estimation. J. Amer. Statist. Assoc. 82, 1131–1146.

[38] Schilling, D.G. and Neubauer, D.V. 2009. Acceptance Sampling in Quality Control, Chapman & Hall/CRC, Boca Raton.

[39] Silverman, B.W., 1981. Using Kernel Density Estimates to Investigate Multimodality J. Royal Statist. Soc. B 43, 97–99.

[40] Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.

[41] Sheather, S. J., Jones, M. C. 1991. A Reliable Data-Based Bandwidth SelectionMethod for Kernel Density Estimation. Journal of the Royal Statistical Society, Ser. B, 53, 683–690.

[42] Shorack, G.R., and Wellner, J.A. 1986. Empirical processes with applications to statistics. Wiley, New York.

[43] Steland, A., Padmanabhan, A.R. and Akram, M. 2012. Resampling methods for the nonparametric and generalized Behrens-Fisher problems. Sankhya A, tentatively accepted.

[44] Steland, A., Zähle, H., 2009. Sampling inspection by variables: nonparametric setting. Stat. Neerl. 63, 101–123.

[45] Virtuani, A., Muellejans, H., Ponti, F. and Dunlop, E. 2010. Comparison of indoor and outdoor performance measurements of recent commercially available solar modules. Proceedings of the 23th European Photovoltaic Solar Energy Conference, Hamburg, Germany, ISBN 3-936338-25-6, DOI: 10.4229/24thEUPVSEC2009-3CO.12.1, 2379–2385.

[46] Wied, D., Weissbach, R., 2012. Consistency of the kernel density estimator: a survey. Statistical Papers 53, 1–21, DOI: 10.1007/s00362-010-0338-1 http://www.springerlink.com/content/u14u4j245um0t805/?MUD=MP