Higher Order Asymptotics for the MSE of Robust M-Estimators of Location on Shrinking Total Variation Neighborhoods

Von der Universität Bayreuth
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Abhandlung

von

Dipl.-Math. Matthias Simon Brandl

geboren am 12.04.1978 in Marktredwitz

1. Gutachter:	Prof. Dr. Helmut Rieder
2. Gutachter:	Prof. Dr. Andreas Christmann
Tag der Einreichung:	29.05.2008
Tag des Kolloquiums:	19.12.2008

Einführung und Zusammenfassung

Dieser Dissertation ist eine CD beigefügt, welche die .pdf Version dieses Dokuments sowie die in Anhang E beschriebenen Algorithmen enthält.

Wir beginnen mit einer kurzen Skizze der Problemstellung im Rahmen der zugrunde liegenden bisherigen Ergebnisse hinsichtlich Höherer-Ordnungs-Asymptotik des mittleren quadratischen Fehlers robuster Schätzverfahren. Daran anschließend folgt eine ausführliche deutsche Zusammenfassung dieser in Englisch verfassten Dissertation.

Einleitung

Für im Stichprobenumfang n (mit Rate $1/\sqrt{n}$) schrumpfende Umgebungen eines idealen glatten[1] Zentralmodells finden sich in [Rieder (1994)] die optimalen asymptotisch linearen Schätzer bezüglich des asymptotischen mittleren quadratischen Fehlers (MSE), welcher gleichmäßig auf diesen Umgebungen ausgewertet wird[2]. Analog zu den Untersuchungen in [Ruckdeschel (2005a)],[Ruckdeschel (2005b)] und [Ruckdeschel (2005d)] versuchen wir die Frage zu beantworten, in wie weit sich die asymptotische Optimalität auf endliche Stichproben überträgt. Diese Fragestellung wurde bereits in [Kohl (2005)] aufgegriffen, indem andere Risiken für finite Stichproben aus [Huber (1968)] und [Rieder (1989)] verwendet wurden. Um die Ergebnisse mit den asymptotischen Ergebnissen vergleichen zu können, wurde dann ausgehend von den finiten Aussagen ein Grenzübergang für den Stichprobenumfang $n \to \infty$ gemacht. Im Gegensatz zu diesem Vorgehen geht unser Ansatz vom asymptotisch optimalen Setup aus und versucht dann genaue Rückschlüsse auf das Finite mittels Edgeworth- und Taylor-Entwicklungen zu machen.

In mehreren Arbeiten[3] stellte P. Ruckdeschel einige tiefer gehende Untersuchungen über die Höhere-Ordnungs-Asymptotik des maximalen MSE im Kontext robuster Schätzverfahren auf schrumpfenden Systemen aus Kontaminationsumgebungen an und formulierte das zentrale theoretische Resultat in folgender Form:

$$\sup_{Q_n \in \tilde{\mathcal{Q}}_n(r,F)} n\mathrm{MSE}(S_n, Q_n) = r^2 b^2 + \mathbb{E}_F \psi^2 + \frac{r}{\sqrt{n}} A_1 + \frac{1}{n} A_2 + o\left(\frac{1}{n}\right) \quad (1)$$

[1] glatt im Sinne von L_2-differenzierbar, vgl. Definition 2.8

[2] vgl. hierfür auch Abschnitt 2.4

[3] siehe [Ruckdeschel (2005a)],[Ruckdeschel (2005b)] und [Ruckdeschel (2005d)]

S_n bezeichnet hierbei einen (M-) Schätzer mit (monotoner) Influenzkurve ψ, $\tilde{\mathcal{Q}}_n(r)$ einen (geringfügig ausgedünnten) Ball aus Konvex-Kontaminationen mit Radius $\frac{r}{\sqrt{n}}$ um eine ideale Verteilung F und A_1, A_2 Polynome in Kontaminationsradius r, Bias $b = \sup|\psi|$ und den Momentenfunktionen $t \mapsto \mathbb{E}_F\psi_t^l$, $l = 1, \ldots, 4$ sowie deren Ableitungen, ausgewertet an der Stelle $t = 0$.
P. Ruckdeschel untermauert dieses Ergebnis mit einer Reihe an Cross-Checks und Kommentaren. Die Relevanz dieses Ergebnisses für (kleine) finite Stichprobenumfänge wird in diesen Arbeiten anhand einer Simulationsstudie illustriert. Anhand eines Faltungs-Algorithmus aus [Kohl et al. (2004)] berechnet er außerdem numerisch exakte Werte des MSE. Für endliche Stichprobenlängen schlägt sein zentrales Resultat (1) - wenngleich auch nur geringfügig - Ergebnisse, die sich für im Stichprobenumfang fixe Umgebungen[4] ergeben; allerdings mit dem Vorteil, mit expliziten Ausdrücken statt rein numerischer Lösungen aufwarten zu können.

Für symmetrisches F, d.h. $f(x) = f(-x)$, sind die Erst-Ordnungs-optimalen ICs im Konvex-Kontaminationsfall vom Hampel-Typ, d.h.

$$\eta_c = A(\Lambda_f - a)\min\left\{1, \frac{c}{|\Lambda_f - a|}\right\} \tag{2}$$

mit Scores-Funktion Λ_f, Stutzhöhe c und Lagrange-Multiplikatoren a und A so[5], dass η_c eine Influenzkurve ist. Beim Übergang in die Zweit-Ordnungs-Asymptotik, also bei Berücksichtigung des A_1-Terms, bleibt unter der Symmetrie von F die Optimalität der Klasse der Hampel-Typ-ICs erhalten, nur die Stutzhöhe c muss gegenüber der Erst-Ordnungs-Lösung angepasst, genauer um $O(1/\sqrt{n})$ gesenkt werden. In diesem Sinne gilt [Pfanzagl (1979)]s Schlagwort, dass Erst-Ordnungs-Effizienz Zweit-Ordnungs-Effizienz impliziert, auch dann noch, wenn man zu Umgebungen des idealen (symmetrischen) Modells übergeht.

Das Ergebnis der vorliegenden Arbeit besteht unter anderem in der Übertragung von P. Ruckdeschels Resultaten in [Ruckdeschel (2005b)] für das ein-dimensionale Lokationsmodell auf Systeme von Totalvariations-Umgebungen. In diesem Zusammenhang zeigt sich auch das Verschwinden des A_1-Terms für symmetrisches F.

Ausführliche Zusammenfassung

Im Rahmen des **Vorwort**s gehen wir auf eine potentielle Anwendungsmöglichkeit von Totalvariations-Umgebungen ein. Dabei handelt es sich um die robuste Schätzung von operationalen Risiken, die konkret durch einen Besuch beim Operational Risk Management der WestLB in Düsseldorf motiviert wurde. Da nur Verluste ab einem gewissen Betrag aufwärts von Interesse sind (bzw. gemeldet werden) und nur positive Ausreißer gefährlich sind, bietet sich dieses Problem aufgrund der „Tabu"-Regionen und der dadurch entstehenden Asymmetrien eventuell als geeigneter Kandidat für robuste Schätzverfahren

[4] vgl. [Fraiman et al. (2001)]

[5] Für exakte Definitionen von Λ_f, c, a und A verweisen wir auf Abschnitt 2.4.

auf Totalvariations-Umgebungen an.

Kapitel 1 beschreibt den Aufbau der Arbeit und die erzielten Resultate. Dabei wird eine kurze Hinführung zum Thema der Höheren-Ordnungs-Asymptotik vorangestellt, in der vor allem auf die bisherigen Ergebnisse von P. Ruckdeschel und M. Kohl eingegangen wird.

In **Kapitel 2** stellen wir einen in sich abgeschlossenen theoretischen Rahmen aus Robuster Statistik sowie deren Asymptotik dar, der die Grundlage dieser Dissertation bildet. Wir beginnen in Abschnitt 2.1 mit der Beantwortung der Frage „Was ist Robuste Statistik?". Hierfür wird das Problem von Ausreißern kurz von einem naiven („manuelles Screening") und einem subtileren (vgl. „Cniper" in Abschnitt 2.1.1) Blickwinkel aus skizziert. Unter Zuhilfenahme einfacher Beispiele in Abschnitt 2.1.2 führen wir das Konzept der Influenzkurve (IC) ein, indem wir zunächst von ihrer Interpretation als einer bestimmten Ableitung[6] eines Funktionals ausgehen und schließlich zur Einbettung in den Kontext der L_2-differenzierbaren[7] Modelle gelangen. Das Ziel, optimale ICs zu finden, führt in Abschnitt 2.2 zu asymptotischen Betrachtungen, die auf der Klasse der asymptotisch linearen Schätzer[8] (ALE) basieren. Schließlich definieren wir in Abschnitt 2.3 den infinitesimal robusten Setup, indem wir verschiedene Umgebungssysteme vorstellen, die durch einfache Perturbationen[9] des idealen Modells entstehen. Anschließend wird in Abschnitt 2.4 die Theorie der optimal robusten Influenzkurven in Bezug auf den MSE und die dazu gehörigen (eindeutigen) Lösungen in Theorem 2.33 vorgestellt.

In **Kapitel 3** beschäftigen wir uns noch einmal mit den bereits in Kapitel 2 eingeführten Umgebungssystemen. Da die zentrale Fragestellung dieser Abhandlung in der Untersuchung des Verhaltens eines maximalen Risikos auf einer speziellen Art von Umgebungen, nämlich Totalvariations-Umgebungen, besteht, betrachten wir in Abschnitt 3.1 die beiden in der robusten Statistik hauptsächlich verwendeten Typen, die Konvex-Kontaminations- und die Totalvariations-Umgebung, erneut und stellen sie in Abschnitt 3.1.3 zur Abgrenzung dem Umgebungssystem, das durch die Hellinger-Metrik erzeugt wird, gegenüber.

Für spätere Zwecke interpretieren wir in Abschnitt 3.1.4 wie in der Robusten Statistik üblich eine schrumpfende Kontaminations-Umgebung Q_n als Menge von Verteilungen eines Vektors $(X_i)_{i\le n}$, der entsteht als

$$X_i := (1-U_i)X_i^{\mathrm{id}} + U_i X_i^{\mathrm{di}}, \qquad i = 1,\dots,n \tag{3}$$

mit X_i^{id}, U_i, X_i^{di} stochastisch unabhängig, $X_i^{\mathrm{id}} \overset{\text{i.i.d.}}{\sim} F$, $U_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bin}(1, r/\sqrt{n})$, und $X_i^{\mathrm{di}} \sim P^{\mathrm{di}}$ mit einem beliebigen $P^{\mathrm{di}} \in \mathcal{M}_1(\mathbb{B})$. Danach leiten wir eine Zerlegung von $Q_n = \otimes_{i=1}^n Q_{n,i}$ im Totalvariationsfall ab, indem wir ein signiertes Maß $\Delta_i \in \mathcal{M}_1(\mathbb{B})$ einführen:

$$dQ_{n,i} = dF + r_n d\Delta_i \tag{4}$$

Nachdem somit die Basis für das zentrale Kapitel 6 gelegt wurde, kommen wir in Abschnitt 3.2 zur Motivation, die zu dieser Arbeit führte, und auf ein Ergebnis in [Kohl (2005)]

[6] Konkret gehen wir in Definition 2.1 und 2.4 auf Frechét- und Gâteaux-Differenzierbarkeit ein und erklären damit die IC in Definition 2.6.

[7] vgl. Definition 2.8.

[8] vgl. Definition 2.13 in Abschnitt 2.2.1.

[9] vgl. Gleichung (2.35)

zurück geht. Im Zusammenhang mit der Bestimmung eines anderen exakten Risikos für endliche Stichproben der Länge $n \geq 3$ verwendete M. Kohl Edgeworth-Entwicklungen, um eine Approximation zu berechnen, da es nicht möglich zu sein scheint, die erwarteten Ergebnisse analytisch zu erhalten[10]. Basierend auf diesen Erkenntnissen über die Höhere-Ordnungs-Asymptotik auf Totalvariationsumgebungen, lautet die Vermutung, dass sich in diesem Fall das Risiko in der Form

$$\sup_{Q_n \in \tilde{\mathfrak{Q}}_n(r)} n\mathrm{MSE}(S_n, Q_n) = r^2 b^2 + \mathbb{E}\psi^2 + \frac{1}{n} A_2 + o\left(\frac{1}{n}\right) \tag{5}$$

darstellen lässt, was eine schnellere Konvergenzrate indizieren würde. Allerdings könnte der Grund für das Verschwinden des $n^{-1/2}$-Terms ebenso gut in der Symmetrie von F, liegen, da Kohl in seinen Untersuchungen stets $F_\theta = \mathcal{N}(\theta, 1)$ verwendet. Im Konvex-Kontaminationsfall erzwingt diese Symmetriebedingung jedoch kein Verschwinden des $n^{-1/2}$-Terms.
Die Technik, die wir zur Herleitung unserer Resultate verwenden, basiert auf genauen Approximationen der Limesverteilung. Nun impliziert Nachbarschaft im Sinne der Verteilungs-Konvergenz nicht notwendig Nachbarschaft/Mitkonvergenz des Risikos, was hier auch so zunächst nicht der Fall ist, wie ein Argument basierend auf dem Konzept des Bruchpunktes zeigt. Daher stellen wir in Abschnitt 3.3 das Konzept des Bruchpunkts für endliche Stichproben[11] dar und unterziehen in Definition 3.10 das infinitesimale Model einer zweckdienlichen Modifikation, die einerseits asymptotisch vernachlässigbar ist, aber andererseits die Mitkonvergenz des unmodifizierten MSE unter schwacher Konvergenz erzwingt.
In **Kapitel 4** reduzieren wir die Terme aus der allgemeinen Einleitung in Kapitel 2 auf eine Dimension, da explizite, handhabbare Bias-Terme für Totalvariation nur für eine Dimension zur Verfügung stehen. Wir führen in Abschnitt 4.1 das Resultat für die Erst-Ordnungs-Optimalität an, um aufzuzeigen, dass es hierbei unter Symmetrie von F keine Möglichkeit gibt, Unterschiede zwischen dem Konvex-Kontaminations- und dem Totalvariationsfall festzustellen. Danach entwickeln wir in Abschnitt 4.2 explizit den Setup für die eindimensionale Lokation für beide Umgebungstypen. Das Kapitel schließt mit einer Diskussion von Hubers Monotonie-Ansatz[12] für M-Schätzer[13], der sich zwar im Lokations- nicht aber, zum Beispiel, im Skalenmodell als brauchbar erweist. Für letzteres Modell präsentieren wir in Abschnitt 4.3.2 einen alternativen Ansatz mittels expliziter Taylorentwicklung von k-Schritt-Schätzern, der in Kapitel 8 auch für das Lokationsmodell Anwendung findet.
In **Kapitel 5** fassen wir die Resultate einer Simulationsstudie zusammen, die uns zu einer genaueren Untersuchung der Höheren-Ordnungs-Entwicklung des MSE im darauf folgenden Kapitel geführt hat. Um geeignete Beobachtungen zu erzeugen, approximieren wir diese mittels eines naheliegenden („Abhängigkeits-erzeugenden") Algorithmus (vgl. Abschnitt 5.1 und 8.3 bzw. Anhang E.1.1 und E.1.2). Genauer gesagt erzeugen wir Beobachtungen aus dem ungünstigsten Element der Umgebung, indem wir aus jeder Stichprobe die

[10] siehe Bemerkung 3.8 in dieser Arbeit bzw. Abschnitt 11.3.3 "Higher Order Approximations" in [Kohl (2005)]
[11] vgl. Definition 3.9.
[12] vgl. insbesondere Abbildung 4.1.
[13] Das Konzept von M- bzw. Z-Schätzern wird in Abschnitt 4.3.1 dargestellt.

K kleinsten Beobachtungen heraus greifen und deren Vorzeichen umdrehen. Als Schätzer S_n verwenden wir einen Drei-Schritt-Schätzer mit dem Median als Startschätzer und Influenzkurve vom Hampel-Typ. Wir berechnen den empirischen asymptotischen MSE[14] und wenden die Box-Cox-Power-Transformation[15] an, um die Ordnung der Terme des empirischen MSE[16] höherer Ordnung in n zu bestimmen. Im nächsten Schritt passen wir mit Hilfe des Akaike Informations-Kriteriums[17] (AIC) ein lineares Modell an den empirischen MSE an. Um unsere Ergebnisse mit dem Konvex-Kontaminationsfall vergleichen zu können, fügen wir stets die entsprechenden Box-Cox-Plots (vgl. Abb. 5.2, 5.4 bzw. 5.6) und Regressionsresultate an. Tatsächlich stimmen die Ergebnisse mit unserer Vermutung überein, dass im Totalvariationsfall eine Konvergenzordnung von n^{-1} vorliegt. Ein Cross-Check in Abschnitt 5.2.4 mit den numerischen Resultaten in [Kohl (2005)] schließt dieses Kapitel ab.

Im zentralen **Kapitel 6** konzentrieren wir uns auf die Frage nach der Höheren-Ordnungs-Entwicklung des MSE von M-Schätzern im Lokationsmodell auf schrumpfenden Totalvariationsumgebungen. Zum Zwecke der Vergleichbarkeit führen wir in Theorem 6.4 zunächst kurz das Resultat für Konvex-Kontaminationen aus [Ruckdeschel (2005b)] an. Nach einigen vorbereitenden Definitionen, Notationen und Lemmata in Abschnitt 6.2.1 formulieren wir unser zentrales Theorem 6.13. Darin liefern wir die explizite Entwicklung der Form (1) für den Totalvariationsfall. Für allgemeines F gilt dabei im Gegensatz zur Vermutung in [Kohl (2005)] zunächst, dass $A_1 \neq 0$!

Die Kernidee des Ansatzes besteht darin, Zerlegung (4) direkt in den Mittelwerten und Varianzen $L_{\mathrm{re},i}(t) := \mathbb{E}_{\mathrm{re}}\psi(x_i - t)$ und $V^2_{\mathrm{re},i}(t) := \mathrm{Var}\psi(x_i - t)$ anzuwenden, ehe man diese mittels Taylor-Reihen entwickelt und damit Zugang zu den Koeffizienten erhält, welche die Terme A_1 und A_2 festlegen (vgl. Annahme 6.7 bzw. 6.18). Indem wir allerdings dadurch mehr Information über die Struktur der Totalvariations-Umgebung einbringen, erhalten wir im Gegensatz zum Beweis des konvex-kontaminierten Falls in [Ruckdeschel (2005b)] wesentlich komplexere Ausdrücke (z.B. 98 im Vergleich zu 63 Summanden bei einem vergleichbaren Polynom, vgl. Bemerkung 6.16).

Dem Beweis von Theorem 6.13 in Abschnitt 6.2.2 geht eine Gliederung voran, die den ziemlich aufwendigen Charakter des Beweises in 15 Einzelschritte zerlegt. Daran schließt sich die detaillierte Ableitung des zentralen Ergebnisses an: nach einer Partition der reellen Achse nach Werten der Beobachtung x_i zeigen wir die Vernachlässigbarkeit diverser Fälle (mittels der Chebyshev-Ungleichung und einer Hoeffding-Schranke, vgl. Anhang B) und können uns deswegen auf ein schrumpfendes Kompaktum[18] zurück ziehen, innerhalb dessen wir eine Edgeworth-Entwicklung[19] auf die zentrierte und standardisierte IC $\psi_{t,i}$ anwenden. Der umfangreiche Einsatz des Computer-Algebra-Sytems `MAPLE`[20] ermöglicht es uns dabei, verschiedene komplizierte Taylor-Entwicklungen des Integranden zu berechnen und gleichzeitig die Ordnung von hunderten von Termen im Blick zu behalten. Zusätzlich

[14] vgl. zur Definition des empirischen asymptotischen MSE Gleichung (5.4)

[15] Die Box-Cox-Power-Transformation wird durch das MASS package von [Venables and Ripley (1999)] bereitgestellt und geht zurück auf [Box and Cox (1964)].

[16] vgl. zur Definition des empirischen MSE Gleichung (5.3)

[17] vgl. Gleichung (5.6).

[18] vgl. Intervall I in Abbildung 6.1.

[19] vgl. Theorem A.5.

[20] Der verwendete `MAPLE`-Algorithmus wird in Abschnitt E.2 beschrieben.

führt uns das Ausweisen einer ungünstigsten Modifizierung der Daten in Hinblick auf den Totalvariations-Bias (vgl. (6.64) bzw. (6.65)) zur Berechnung der endgültigen Terme.
Um unsere Vermutung (5) zu beweisen, beschränken wir uns in Abschnitt 6.3 auf den symmetrischen Fall. Tatsächlich gelingt es uns dann in Corollar 6.19 das Verschwinden des A_1-Terms zu beweisen. Somit ist die Symmetrie von F Voraussetzung für die höhere Konvergenzordnung.
Mittels eines Arguments aus [Feller (1971)] können wir in den Lemmata 6.21 und 6.22 zudem zeigen, dass die im Beweis von Theorem 6.13 vereinfachend angenommene Situation identisch verteilter Variablen tatsächlich keine Einschränkung ist. Im Falle $r = 0$ ergeben sich aus der Höheren-Ordnungs-Entwicklung Konsequenzen für das ideale Model, die wir in Corollar 6.20 aufzeigen und als Cross-Check mit dem konvex-kontaminierten Fall vergleichen. Hier zeigt sich wie zu erwarten das Zusammenfallen beider Fälle. Als Vorbereitung auf das folgende Kapitel berechnen wir in Abschnitt 6.6 die entsprechenden Terme der Entwicklung im Fall $F = \mathcal{N}(0, 1)$ und geben in Proposition 6.23 die Koeffizienten des (symmetrischen) A_2-Terms in Abhängigkeit von der Dichte der Normalverteilung an. In Bemerkung 6.25 stellen wir diesen Resultaten abermals den konvex-kontaminierten Fall zur Seite.
In **Kapitel 7** untersuchen wir das Verhalten des asymptotischen MSE mittels der Koeffizienten für den repräsentativen Setup $F = \mathcal{N}(0, 1)$, wobei wir besonderes Augenmerk auf den A_2-Term richten. Dann vergleichen wir die Ergebnisse der Erst-, Zweit- und Dritt-Ordnungs-Asymptotik. Bereits die numerischen Ergebnisse in Abschnitt 7.2.1 führen zu der Vermutung, dass im Totalvariationsfall (im Gegensatz zum Konvex-Kontaminationsfall) der maximale MSE auf $\tilde{\mathcal{Q}}_n$, unter Symmetrie und ausreichend großem n, für kleine Radien von Erst- (und Zweit-) Ordnungs-Asymptotik stets unterschätzt, für große Radien hingegen aber überschätzt wird!
Eine nähere Untersuchung des A_2-Terms in Abschnitt 7.2.2 zeigt schließlich, dass wir tatsächlich für kleine Radien (in Abhängigkeit von der Stutzhöhe) einen negativen Beitrag zum MSE erhalten. Deswegen ergibt sich eine Überschätzung des MSE. Die Situation verändert sich allerdings sowohl bei Vergrößerung des Radius wie auch der Stutzhöhe (vgl. Bemerkung 7.3 und Abb. 7.1). Wir geben hierfür die heuristische Erklärung, dass in diesen Situationen die ungünstigsten Abweichungen nicht, wie in Kapitel 6 notwendigerweise vollzogen, angewendet bzw. erreicht werden können. Das Ergebnis ist ein MSE, der mit jeder zusätzlichen „schlechten" Beobachtung in der Stichprobe unbeschränkt wächst.
In **Kapitel 8** beschäftigen wir uns in einem finiten Kontext mit der Frage nach der Existenz einer ungünstigsten Verteilung, wie sie im Beweis des zentralen Theorems ausgewiesen wurde. In einem finiten Szenario mit idealer Ausgangsstichprobe $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} P_n^{id}$, die durch das signierte Maß Δ_i gemäß (1.4) bzw. (3.18) manipuliert werden soll, könnte die ungünstigste Verteilung nicht erreichbar sein. Dies bedeutet, dass wir einen passenden Mechanismus finden und beschreiben müssen, der die Auswirkung von Δ_i auf endliche Stichproben unter vorgegebenen Bedingungen erklärt. In Hinblick auf Corollar 6.19 beschränken wir uns auf den symmetrischen Fall, für ein auf der Borel-Menge $\mathbb{B}$ symmetrisches Maß $F = P^{id}$; die Influenzkurve ist monoton und schiefsymmetrisch. Für einen bestimmten Manipulationsmechanismus[21] erhalten wir dann das theoretisch be-

[21] vgl. Abschnitt 8.3.

wiesene Resultat bis auf die gewünschte Ordnung hin exakt.
In diesem Sinne ordnen wir die Stichprobe zunächst nach der Größe der Beobachtungen. Dabei beschränken wir uns auf Influenzkurven vom Hampel-Typ, die ihr Maximum zumindest für $|x| > c_n$ annehmen, wobei c_n zunächst eine allgemeine, wachsende Folge ist (vgl. Abb. 8.1). Die Anzahl k manipulierbarer Beobachtungen wird durch die Zufallsvariable K bestimmt, deren erstes Moment $\mathbb{E}K = r\sqrt{n}$ so gewählt wurde, dass ein Totalvariationsball $B_v(F, r/\sqrt{n})$ nicht verlassen wird, vgl. Lemma 8.5 und Abbildung 8.2. Das zweite Moment $\mathrm{Var}K = \frac{1}{2}r\sqrt{n}$ resultiert aus einer tiefer gehenden Untersuchung aller Terme des MSE im Beweis von Theorem 8.14, die diesmal mittels eines k-Schritt-Ansatzes erhalten werden, vgl. Abschnitt 8.4. Durch die Anordnung der Stichprobe sind die Beobachtungen nun allerdings (schwach) korreliert, vgl. Proposition 8.16 und Theorem 8.17. Schließlich gelingt es uns aber in Theorem 8.20 zu zeigen, dass diese Korrelation unter bestimmten Bedingungen und für hinreichend großes n verschwindet. Ohne Anwendung weiterer Symmetrieargumente werden wir mit der gemeinsamen Verteilung des k- und $n-k+1$-Quantils $X_{[k:n]}$ und $X_{[n-k+1:n]}$ konfrontiert, was zu Fragestellungen aus dem Gebiet der Ordnungsstatistiken führt. Da sich aber die Integrale, die als Folge dieses Ansatzes zu berechnen sind, als schwer handhabbar erweisen, vermitteln wir in Abschnitt 8.5.1 nur einen kurzen Eindruck dieser Situation und machen statt dessen in Abschnitt 8.6 Gebrauch von einem Symmetrieargument, das in weiterem Sinne durch das Spiegelungsprinzip der elementaren Stochastik inspiriert ist: durch gleichzeitige Betrachtung von mehreren Stichproben $\{x_1, \ldots, x_n\}_j \overset{\text{i.i.d.}}{\sim} F,\ j \in \mathbb{N}$, sind wir schließlich in der Lage, den Unterschied zwischen oberem und unterem k-Quantil zu vernachlässigen.
Weiterhin zeigt sich in Abschnitt 8.8, dass wir nur dann im finiten Kontext das Ergebnis aus Corollar 6.19 erhalten, wenn wir von der Stichprobe bzw. der Influenzkurve in Bedingung 8.19 (p) verlangen, dass Minimum und Maximum der gegebenen Influenzkurve ψ mit einer gewissen Wahrscheinlichkeit tatsächlich angenommen werden. In Abhängigkeit von dieser Wahrscheinlichkeit leiten wir in Theorem 8.20 eine untere Schranke an den Stichprobenumfang n ab, nachdem wir die Existenz einer derartigen Bedingung bereits in vorangegangenen Simulationen (vgl. Abschnitt 8.8.1) vermutet haben.
Schließlich formulieren wir in der Annahme 8.21 (PK) eine restriktive Bedingung an die Verteilung von K, die grob gesprochen[22] garantiert, dass $X_{[k:n]}$ unter einer - nun konkreten - Schranke c_n bleibt und wir dadurch stets ausreichend Beobachtungen zur Verfügung haben, um eine ungünstigste Modifikation der Stichprobe zu erzeugen. Die Schranke c_n wird in Proposition 8.24 für $F = \mathcal{N}(0,1)$ explizit berechnet. Abschließend geben wir in Abschnitt 8.9.3 geeignete Vierpunkt-Verteilungen von K an, die allen bis dahin geforderten Bedingungen genügen.
In **Kapitel 9** listen wir einige denkbare Erweiterungen zu dieser Arbeit auf, offene Fragen betreffend.
Der **Anhang** beinhaltet diverse zusätzliche Resultate für bzw. von voraus gegangenen Kapiteln. **Anhang A** enthält einige Hilfsmittel wie Hoeffding-Schranken, Mills' ratio oder ein Theorem über Edgeworth-Entwicklungen. Diese Resultate werden in den Beweisen der Kapitel 6 und 8 benötigt. **Anhang B** beschäftigt sich ausführlich mit der

[22]Theorem 8.22 zeigt, dass die Wahrscheinlichkeit des Überschreitens der Schranke c_n unter der Bedingung (PK) exponentiell vernachlässigbar ist.

Vernachlässigbarkeit der Fälle außerhalb des schrumpfenden Kompaktums im Beweis von Theorem 6.13. Der allgemeine A_2-Term wurde wegen seiner komplizierten und länglichen Darstellung in den **Anhang C** verschoben. Der A_2-Term in Corollar 6.19 lässt sich für den symmetrischen Fall von diesem allgemeinen Term ableiten. **Anhang D** stellt einige allgemeine Resultate über Verteilungen und Dichten von gemeinsamen Verteilungen zweier Quantile zusammen. In diesem Zusammenhang sammeln wir auch weitere eher technische Lemmata, die in Kapitel 8 benötigt werden. **Anhang E** beinhaltet eine kurze Beschreibung der Algorithmen für R und MAPLE. Wir kommen auch kurz auf das SWEAVE-Paket für R und LaTeXzu sprechen. Als Abrundung und Ergänzung von Kapitel 2 enthält **Anhang F** schließlich noch einige weitere klassische Resultate der asymptotischen Statistik. In **Anhang G** sind einige Errata aufgelistet.

Introduction

Along with this dissertation comes a CD which contains the .pdf version of this document as well as the algorithms described in appendix E.

Genesis of the thesis

Although I did my diploma thesis [Brandl (2003)] in the subject of Mathematical Physics, Mathematical Statistics always was an emphasis during my studies at the University of Bayreuth. I took part in courses on Stochastics, Generalized Linear Models, Time Series Analysis, Data Analysis with R and last but not least Asymptotic Robust Statistics. The latter never lost hold on me and so - yearning for higher mathematics during my provisional teaching period[23] as a trainee teacher at German Gymnasium[24] level - I took part once again in the Seminar on Statistics by Prof. Dr. Rieder in the summer of 2005 when working at the Graf Münster Gymnasium in Bayreuth.
With my interests newly arisen I asked Dr. Ruckdeschel for an adequate research project to work on for a PhD thesis in Robust Statistics. At that time Dr. Ruckdeschel himself was working on Higher Order Asymptotics for the MSE of Robust Estimators on Shrinking Convex Contamination Neighborhoods. As I was told by Prof. Dr. Rieder, before the rise of computers and thereby computer algebra systems (CAS), Higher Order Asymptotics had been treated by a heavily use of color pens, marking the different terms of identical order in a chaos of symbols over several pages. Today one can take advantages of a CAS like `MAPLE` or `MATHEMATICA` and that's what Dr. Ruckdeschel used for his work. In this context he told me about his successful results in the case of Convex Contamination neighborhoods and that he did a sketch - sometime, somewhere - showing the total variation case to be feasible by a straight forward method. Well, as time went by, the method showed up to be not as straight forward as suggested. Obstacles like the loss of an identical distribution or the independence of random variables had to be overcome and led into regions of Fourier transformation and Order Statistics. But finally, at the end of 2007, all difficulties had been settled, delivering a satisfying treatise for the Robust Estimation of one-dimensional Location on Shrinking Total Variation Neighborhoods.
Meanwhile I had gained my final degree as a maths and physics teacher, and half a year of work as a financial analyst and fund manager at an investment company in Frankfurt a. M. lay behind me. The chance of an assistant position at the University of Augsburg finally offered me the unpayable opportunity of finishing my thesis in an academic surrounding.

[23] The German term is "Referendariat".
[24] A German Gymnasium might be described as a college preparatory high school.

A potential application

In October 2007, during the work on this thesis, I was invited by Florian Camphausen and Dr. Frank Beekmann to the Quantification Team of the Operational Risk Management of the WestLB in Düsseldorf. There I was confronted with the necessity of robust estimation for operational risks[25].

As mentioned in [Beekmann and Stemper (2006)], for example, the financial sector is busy with the application of new regulatory requirements that are demanded by the international *Basel Committee on Banking Supervision* of the Bank for International Settlements in its general agreement "Basel II", the second of the Basel Accords (confer [BCBS (2004)]). Basel II sets up rigorous risk and capital management requirements designed to ensure that a bank holds capital reserves appropriate to the risk the bank exposes itself to by its lending and investment practices. Within the variety of risk, operational risks belong to the group of miscellaneous risks and is defined by Basel II as *"the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events. This definition includes legal risk, but excludes strategic and reputational risk"* ([BCBS (2004)], Part 2 V. A. §644.). Two examples, taken from [Beekmann and Stemper (2006)], shall give an impression of the impact of operational risk.

- Barings (Unauthorized Trading) 1995: The Barings Bank collapsed after a loss of 827 mio. GBP arisen from unauthorized overdrawing of limits in trading transactions by Nick Leason.
- Mizuho Securities (Fat-Finger-Syndrom) 2005: A Japanese trader sold 610.000 shares at 1 YEN instead of 1 share for 610.000 YEN. The total damage sums up to approximately 334 mio. USD.

In the light of recent events we add one more example:

- Société Générale (Unauthorized Trading) 2008: The French trader Jérôme Kerviel exceeded his authority to engage in unauthorized trades, involving European stock index futures, totaling as much as € 49.9 billion, a figure far higher than the bank's total market capitalization. In the time Société Générale tried to close out positions built up by Kerviel, the European stock markets suffered heavy losses of about 6%.

In order to cover the estimated risk the bank has to hold (so called regulatory) equity, so that the estimation of the operational risk affects business operations indirectly. This problem was tackled in [Beekmann and Stemper (2006)], where a loss distribution approach (LDA) was developed. The aim of LDA is to estimate an operational Value at Risk[26] (OpVaR) as an aggregate total loss from single losses of the past, not exceeding an

[25] A recent analysis of (qualitative) robustness of risk measurement procedures was done in [Cont et. al. (2007)], for example.

[26] For alternative measures of risk and their properties there is plenty of literature. For instance, we refer to [Artzner et. al. (1998)], [Delbaen (2002)] or [Fernandes et. al. (2007)].

a priori probability.

The number of losses is assumed as a random variable $N \sim \text{Poiss}(\lambda)$ with λ the mean of the observed loss frequency in the data of the past years. The losses themselves are assumed to be i.i.d. random variables $X_1, \ldots, X_N$ and the aggregate loss function is given by the arithmetic mean[27]

$$L = \sum_{i=1}^{N} X_i \tag{6}$$

Then the OpVaR is defined as the α-quantile of the aggregate loss distribution P for $\alpha = 99.9\%$ or even $\alpha = 99.95\%$. Now the choice of P is crucial to the estimation of the OpVaR. In [Beekmann and Stemper (2006)] Lognormal, Weibull or composed distributions are used, whose tails are modeled by a generalized Pareto distribution. But [Beekmann and Stemper (2006)] complains that the steadily change of the data by quarterly loss reports lead to variations of the parameters quarter by quarter, especially if some new high losses were reported.
These high losses and the decision whether to reject them or not is a subject which robust statistics is mainly concerned with. They are called outliers, confer section 2.1.

In a talk on March 13th 2007, at the Global Conference on Operational Risk in New York F. Beekmann summarized some ideas of "Using Robust Estimators to Find Parameters". On slide 6 of his talk he assumes a mixture distribution that is due to an infinitesimal convex contamination neighborhood system[28] of the true model.
Now, as only losses as from a certain amount upwards are of interest (and reported or collected, respectively), and only positive outliers are dangerous, there is good reason for treating the problem in an asymmetric way. Considering other problems like the problem of estimation of mortalities for an insurance company or portfolio selection with respect to the fact that only upside or downside risk is seen as dangerous, P. Ruckdeschel investigated the asymmetric case for convex contamination neighborhoods in [Ruckdeschel (2005c)], which lead to unrealistic results, however.

By contrast, in this thesis we look at optimal robust estimators over infinitesimal **total variation** neighborhood systems[29] of the ideal distribution. They have the several advantages:

(1) intuitively accessible[30]

(2) good algorithmic properties under symmetry[31]

[27] For the (un)robust characteristic of the sample mean see example (1) in subsection 2.1.2.

[28] For a detailed definition and interpretation of an infinitesimal convex contamination neighborhood system see sections 2.3, 3.1.1 and subsection 3.1.4, respectively.

[29] For total variation neighborhood systems see section 2.3 and subsections 3.1.2 and 3.1.4, respectively.

[30] See 3.1.4, especially figure 3.2.

[31] For the improved speed of convergence in contrast to convex contamination see Corollary 6.19.

(3) easily asymmetrically modifiable[32], especially with respect to model-based taboo regions (restrictions)

Actually, we stay especially with the symmetric case and show in chapter 6 that then first order optimality of an estimator implies second order optimality w.r.t. the MSE. Furthermore we get an improved speed of convergence. The modification of the ideal distribution is done by a mechanism described in chapter 8 attaining least favorable deviations.

As a full treatment of the sketched problem in finance concerning the robustification of operational risk estimation would go beyond the scope of this thesis we end this discussion here and are content with bringing the flexible total variation neighborhood systems back to the mind of robust statistics by showing (and proving) some beautiful aspects in the context of higher order asymptotics. But we may propose on a solid base that an approach via asymmetric total variation neighborhoods might be the solution to the problem of robust operational risk estimation.

Acknowledgment

There were many people, who helped and supported me during the work on this thesis. First of all I thank my Ph.D. supervisor Prof. Dr. Rieder, who taught me Robust Statistics and always encouraged me with his appreciation of my efforts.
Without Dr. Peter Ruckdeschel this thesis would not exist; neither would it have been started nor would it have been completed. Having attended me during my studies at the University of Bayreuth for several years, already, he encouraged me in the first place to deal with Robust Statistics again, even if there was no possibility of an assistant position at the chair. In the second place he never got tired to answer my uncomprehending questions while I was trying to get the point of his all new research results on Higher Asymptotics of Robust Estimation. In the third place he helped me through the "doctoral blues" that caught me in the summer of 2007, when I got stuck in a wood of quantiles[33]. And after all, besides finishing his postdoctoral lecture qualification he always found lots of time to read and discuss my ideas. Actually, there is no way to say how much I am indebted to Dr. Peter Ruckdeschel for his support.
Furthermore I thank Prof. Dr. Ulm for tolerating my intention to finish my PhD at a different chair and university. I am very grateful for his nonstop effort of supporting my project with the necessary time-frames.
I thank Florian Camphausen and Dr. Frank Beekmann for their fruitful and inspiring invitation to the Quantification Team of the Operational Risk Management of the WestLB in Düsseldorf.
Many thanks to Dr. Matthias Kohl, too, who provided me with detailed explanations and stuff concerning the results of his own PhD-thesis.
In particular, I thank my wife Birgit Brandl for accompanying me with love and understanding in those exhausting years of work, examination and research. She is the second

[32]For a thinkable asymmetric bias weighting get inspired by (2.53).
[33]A fact that is now briefly sketched in subsection 8.5.1.

one, without whom this thesis never would have been finished.
Last but not least I thank my family and all my friends not mentioned here for all their support during the years.

Contents

List of Figures

Notation

Abbreviations

a.e.	almost everywhere, almost surely
c.d.f.	cumulative distribution function
eventually	for all sufficiently large sequence indices
ibid.	ibidem, in the same place; confer the book, chapter, article, or page cited just before
i.i.d.	stochastically independent, identically distributed
iff	if and only if
s.t.	subject to
se	standard error (of estimated regression coefficients)
w.r.t.	with respect to, relative to
AIC	Akaike information criterium
ALE	asymptotically linear estimator
CAS	computer algebra system
CLT	central limit theorem
GBP	Great Britain Pound
IC	influence curve
IF	influence function
LDA	loss distribution approach
M, L,R	maximum likelihood type, linear function of order statistics, and rank based, respectively
MLE	maximum likelihood estimator
MSE	mean square error
maxMSE	minimax asymptotic MSE
$\overline{empMSE}_n$	empirical MSE
$\overline{asyempMSE}_n$	empirical asymptotic MSE
OpVaR	operational value at risk
RHS, LHS	right/left-hand side
RSS	residual sum of squares
SSY	squared sum of deviation of sample values from the sample mean in the context of variance analysis
USD	United States Dollar
$\square$	QED

Sets and functions

$\mathbb{N}$	the natural numbers $1, 2, ...$
$\mathbb{Z}$	the integers $..., -1, 0, 1, ...$
$\mathbb{R}$	the real numbers $(-\infty, \infty)$
$\bar{\mathbb{R}}$	the extended real numbers $[-\infty, \infty]$, homeomorphic to $[-1, 1] \subset R$ via the isometry $z \mapsto z/(1 + \|z\|)$
$\mathbb{C}$	the complex numbers
$\times$	Cartesian product of sets; $A^m = A \times \cdots \times A$ (m times)
$\mathbb{I}_A$, $\mathbb{I}(A)$	indicator function of a set or statement A; thus, for any set A, we may write $\mathbb{I}_A(x) = \mathbb{I}(x \in A)$
id_Ω	identity function on the set Ω
med	the median
sign	$\mathrm{sign}(x) = -1,\ 0,\ 1$ for x negative/zero/positive
$f(x \pm 0)$	left/right-hand limit at x of a function f
Λ_f, Λ_θ, Λ	L_2 derivative; parametric tangent
$\mathcal{I}_\theta$, $\mathcal{I}$	Fisher information
$B(P_\theta, r)$	ball about P_θ with radius r
$\bar{X}_n$	arithmetic mean od the (random) variables $X_1, ..., X_n$
Ω	sample space
ψ, ψ_θ	influence curve
ψ_h	classical scores $\mathcal{I}^{-1}\Lambda \in \Psi_2$
η_h	classical partial scores $D\psi_h$ with $\mathbb{E}_\theta \Lambda_\theta^\tau = D$
Θ	parameter space
$\mathcal{X}_S$	characteristic function of an estimator S

σ-Algebras

$\mathcal{A}$	σ-Algebras
$\mathbb{B}$, $\bar{\mathbb{B}}$	Borel σ -algebras on $\mathbb{R}$ and $\bar{\mathbb{R}}$, respectively
$\sigma(\mathfrak{E})$	smallest σ-algebra (on Ω) including a system $\mathfrak{E} \subset 2^\Omega$
$\bigotimes$	product of σ -algebras; $\mathcal{A}^m = A \otimes \cdots \otimes A$ (m times)

Measures

P, P_θ	distribution
F, F_θ	ideal distribution
$\mathcal{M}_b(\mathcal{A})$	the finite (or $b$ounded) measures on a σ-algebra $\mathcal{A}$
$\mathcal{M}_1(\mathcal{A})$	the probability measures (mass 1) on $\mathcal{A}$
Δ_i, Δ	signed measure $\Delta \in \mathcal{M}_1(\mathbb{B})$
H	arbitrary probability measure $H \in \mathcal{M}_1(\mathbb{B})$
$\mathcal{P}$	family of probability mesaures
P^{di}	disturbing measure

support P	smallest closed subset A of Ω (separable, metric) such that $P(\Omega \backslash A) = 0$; cf. II Definition 2.1 of [Parthasarathy (1967)]
$\ll$	domination of measures
$\bigotimes$	product of measures
$*$	convolution of measures
$\overrightarrow{w}$	weak convergence of (bounded) measures
$w_h(A)$	upper probability of $B_h(P)$; $w_h(A) = \sup_{Q \in B_h(P)} Q(A)$ with $A \in \mathcal{A}$

Random Variables and Expectation

$\sim$	distributed according to
$X_1, ...X_n$	sample of random variables
$(X_i)_{i \leq n}$	vector of random variables
$\mathcal{X}$	real valued sample space, $\mathcal{X} \subset \mathbb{R}$
U_i	switching random variable $U_i \overset{\text{i.i.d.}}{\sim} \text{Bin}(1, r/\sqrt{n})$
X_i^{id}	random variable $X_i \overset{\text{i.i.d.}}{\sim} F$
X_i^{di}	random variable $X_i \overset{\text{i.i.d.}}{\sim} P^{\text{di}}$
$\cdot_{\text{re}}$	evaluation under Q_n
$\cdot_{\text{id}}$	evaluation under F
$\cdot_{\text{di}}$	evaluation under P^{di}
$\mathcal{L}_P(X)$	law of X under P
$\mathbb{E}X$	expectation of X
$\text{Var}X$	variance of X
$\text{Cov}X$	covariance of X
$\xrightarrow{P_n}$	stochastic convergence, convergence in P_n probability
o, O	stochastic Landau symbols; that is, $o(r_n)/r_n \xrightarrow{P_n} 0$, respectively, the sequence $\lvert O(r_n)/rn \rvert (P_n)$ tight on $\mathbb{R}$
$X_{[k:n]}$, $x_{(k)}$	k-quantile $F^{-1}(k/n)$

Laws

$\mathbb{I}_{\{a\}}$, δ_a, $\mathbb{I}(a)$	(Dirac) one-point measure in a
$\text{Bin}(m, p)$	binomial distribution with size $m \in \mathbb{N}$ and probability of success $p \in [0, 1]$
$\mathcal{N}_k(\mu, \sigma^2)$	k-dimensional normal law on $(\mathbb{R}^m, \mathbb{B}^m)$ with mean $\mu \in \mathbb{R}^m$ and standard deviation σ
$F_{n,(k)}$	the F distribution with n and k degrees of freedom
$t_{(k)}$	Student's t-distribution with k degrees of freedom
φ, Φ	standard normal density and distribution function on $\mathbb{R}$
$\text{Poiss}(\lambda)$	Poisson distribution with mean $\lambda \in (0, 1)$

Mathematical Symbols

$\sharp A$	cardinality of a set A
A^c	complement of A
$\subset$, $\supset$	subset/supset, or equal
$\leq$	less or equal, coordinatewise on $\mathbb{R}^m$
$\|.\|$	Euclidean norm on $\mathbb{R}^m$
x^+ , x^-, $(.)_+$, $(.)_-$	positive, negative parts
$\wedge$, min	minimum
$\vee$, max	maximum
inf, sup	pointwise infimum/supremum
$\inf_P$, $\sup_P$	P essential infimum/supremum
$\uparrow$, $\downarrow$	monotone convergence from below/above of numbers, functions (their values), and sets (their indicators)
$a \rightsquigarrow b$	a replaced by b
$\mathrm{lin}(x_1, ..., x_k)$	linear space generated by $x_1, ..., x_k$
$d_*(Q, F)$	the total variation ($* = v$), Prokhorov ($* = \pi$), Lévy ($* = \lambda$), Kolmogorov ($* = K$) and Hellinger ($* = h$) distance, respectively, between the measures Q and F

Matrices

$\mathbb{I}_k$	the unit $k \times k$ matrix
$A \in \mathbb{R}^{p \times k}$	a real matrix with p rows and k columns
A^τ	transpose of a matrix A
$\mathrm{rk}A$	rank of A
$\mathrm{tr}A$	trace of A
$A \succ B$	$A - B$ positive definite
$A \succeq B$	$A - B$ positive semidefinite

Function Spaces

C_c^1	functions: $\mathbb{R} \to \mathbb{R}$ which are continuously differentiable functions and have compact support
C_c^1	functions: $\mathbb{R} \to \mathbb{R}$ which are infinitely differentiable and have compact support
$L_2^k(P)$	the Hilbert space of (equivalence classes of) $\mathbb{R}^k$-valued functions f such that $\int \|f\|^2 dP < \infty$; $L_2(P) = L_2^1(P)$
$\mathcal{L}_2^k(\mathcal{A})$	the Hilbert space of (equivalence classes of) $\xi\sqrt{dP}$ with any $P \in L_2^k(P)$, $P \in \mathcal{M}_b(\mathcal{A})$

$L_\infty^k(P)$	the space of (equivalence classes of) $\mathbb{R}^k$-valued functions f such that $\sup_p \lvert f\rvert dP < \infty$; $L_\infty(P) = L_\infty^1(P)$
$Z_\alpha(\theta)$	$L_\alpha^p(P_\theta) \cap \{E_\theta = 0\}$; space of square integrable ($\alpha = 2$), and bounded ($\alpha = \infty$) tangents at P_θ
$\Psi_\alpha(\theta)$, $\Psi_\alpha^D(\theta)$	set of square integrable ($\alpha = 2$), and bounded ($\alpha = \infty$), influence curves at P_θ; respectively, partial influence curves at P_θ, with some matrix $D \in \mathbb{R}^{p\times k}$ such that $\mathrm{rk}D = p \leq k$

Neighborhoods and Bias Terms

$* = c, v$	type of balls and metric: contamination, total variation
A^ε	closed ε-neighborhood of A
$U_*(\theta)$	neighborhood system about P_θ
$U_*(\theta, r)$	such a neighborhood about P_θ of radius $r \in (0, \infty)$; in the infinitesimal robust setup, usually $r = O(1/\sqrt{n})$
$\mathcal{G}_*(\theta)$	corresponding tangent classes
Q_n	shrinking infinitesimal neighborhoods
$\mathcal{Q}_n$	ball of shrinking infinitesimal neighborhoods Q_n
$\omega_{*,\theta}$, ω_*	standardized (infinitesimal) bias terms
ε_0	finite sample breakdown point

Variables

a, A	Lagrangian multipliers
b	bias bound $b \in (0, \infty)$
c	clipping height
q_i, q	tangent $q \in \mathcal{G}_*(\theta)$
r	deviation radius
θ	the (true) parameter to be estimated; $\theta \in \Theta$
$\theta_n^{(k)}$	k-step estimator
n	sample length
A_1, A_2	polynomials appearing in the higher order expansion of the MSE in the moment functions $\mathbb{E}_F\psi_t^l$ or $\mathbb{E}_{Q_n}\psi_t^l$ and $\mathbb{E}_\Delta\psi_t^l$, $l = 1...4$, respectively, and their derivatives evaluated in $t = 0$
D_n	rest term in the one-term Edgeworth expansion
K, k	random variable and actual realization, respectively, for number of modified observations
S_n	estimator

Chapter 1

Organization of the thesis and description of the results

In the setup of shrinking neighborhoods in sample size n (at rate $1/\sqrt{n}$) about an ideal (L_2-differentiable) central model, [Rieder (1994)] determines the optimal asymptotic linear estimator w.r.t. the asymptotic MSE evaluated uniformly on these neighborhoods. Standing in line with results attained by P. Ruckdeschel, we try to answer the question to which degree the asymptotic optimality carries over to finite sample size. This problem already was tackled in [Kohl (2005)] by taking over finite sample risks from [Huber (1968)] and [Rieder (1989)], starting from small sample sizes to be increased afterwards. Contrary, our approach stays with the asymptotically optimal setup and steps "backwards" from the infinite to the finite by application of Edgeworth and Taylor expansions.

In a number of papers[1], P. Ruckdeschel did some deeper investigations on higher-order asymptotics of the maximal mean squared error in the context of robust estimation on shrinking contamination neighborhood systems and formulated the central theoretical result, which is of the following form:

$$\sup_{Q_n \in \tilde{\mathcal{Q}}_n(r,F)} n\mathbf{MSE}(S_n, Q_n) = r^2 b^2 + \mathbb{E}_F \psi^2 + \frac{r}{\sqrt{n}} A_1 + \frac{1}{n} A_2 + o\left(\frac{1}{n}\right) \tag{1.1}$$

Here S_n is an (M-) estimator with (monotone) influence curve (IC) ψ, $\tilde{\mathcal{Q}}_n(r)$ is a (slightly thinned out) ball of convex contaminations of radius $\frac{r}{\sqrt{n}}$ about the ideal distribution F and A_1, A_2 are polynomials in the contamination radius r, in bias $b = \sup|\psi|$, and in the moment functions $t \mapsto \mathbb{E}_F \psi_t^l$, $l = 1, \ldots, 4$ and their derivatives evaluated in $t = 0$.
P. Ruckdeschel gives a number of cross checks and comments on this result. The relevance of his results for (small) finite sample sizes is shown by a simulation study. By means of an adopted convolution algorithm taken from [Kohl et al. (2004)], he also computes numerically exact values of the MSE. Measured at a finite sample context, his main result in most cases beats —albeit only by a minor amount— results obtainable in the fixed-neighborhood setup, compare [Fraiman et al. (2001)], with the advantage of explicit

[1] We refer to [Ruckdeschel (2005a)],[Ruckdeschel (2005b)] and [Ruckdeschel (2005d)]

expressions instead of numerical solutions.

For F symmetric, i.e. $f(x) = f(-x)$, one achieves first-order optimality by Hampel-type ICs, i.e.

$$\eta_c = A(\Lambda_f - a) \min\left\{1, \frac{c}{|\Lambda_f - a|}\right\} \tag{1.2}$$

with scores function Λ_f, clipping height c and Lagrange multipliers a and A such[2] that η_c is an IC. The first-order optimality persists if we account for the A_1 term in (1.1). Hence, in this sense, [Pfanzagl (1979)]'s catchword "*First order efficiency implies second order efficiency*" survives (at least partially) when passing to neighborhoods around the ideal (symmetric) model.

It is the achievement of this thesis to transfer P. Ruckdeschel's results in [Ruckdeschel (2005b)] for the one–dimensional location model to the case of total variation neighborhood systems and thereby to prove the vanishing of the term A_1 in (1.1) for F symmetric.

In **Chapter 2: Robust Statistics and its Asymptotic Theory** we give a sufficiently comprehensive framework of robust statistics and its asymptotics. In section 2.1 we start by answering the question "What is Robust Statistics?". Therefore the problem of outliers is briefly sketched from a naive ("manual screening") and a more subtle (conf. "Cniper" in subsection 2.1.1) point of view. Accompanied by simple examples we introduce the concept of influence curves (IC) in subsection 2.1.2 starting from the interpretation as a special derivative[3] of a functional and leading to the embedding in the context of L_2-differentiable[4] models. The aim to detect optimal ICs leads to asymptotic considerations in section 2.2 mainly based on the class of asymptotically linear estimators[5] (ALE). We finally define the infinitesimal robust setup in section 2.3 considering several neighborhood systems derived by simple perturbations of the ideal model. Subsequently the theory of optimal robust influence curves with respect to mean squared error (MSE) and its (unique) solution is presented in section 2.4.
In **Chapter 3: Motivation** we come back once again to the neighborhood systems already introduced in chapter 2. But as the main concern of this thesis is the investigation of the behavior of a maximal risk on a special kind of neighborhoods, i.e., total variation neighborhoods, we want to lay sufficient emphasis on this subject. Therefore in subsection 3.1 the two mainly used types in robust statistics, convex contamination and total variation neighborhoods, are reconsidered and as contrast the neighborhood system generated by the Hellinger distance is discussed in subsection 3.1.3.
Additionally, in section 3.1.4 we repeat the interpretation of Q_n, being a shrinking contamination neighborhood, as the distribution of the vector $(X_i)_{i\le n}$ with components

$$X_i := (1 - U_i)X_i^{\mathrm{id}} + U_i X_i^{\mathrm{di}}, \qquad i = 1, \ldots, n \tag{1.3}$$

[2]For detailed definitions of Λ_f, c, a and A we refer to section 2.4.

[3]We define Frechét- and Gâteaux-differentiability in Definition 2.1 and 2.4 give the declaration for the IC in Definition 2.6.

[4]conf. Definition 2.8.

[5]conf. Definition 2.13 in subsection 2.2.1.

for X_i^{id}, U_i, X_i^{di} stochastically independent, $X_i^{\mathrm{id}} \overset{\text{i.i.d.}}{\sim} F$, $U_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bin}(1, r/\sqrt{n})$, and $X_i^{\mathrm{di}} \sim P^{\mathrm{di}}$ for some arbitrary $P^{\mathrm{di}} \in \mathcal{M}_1(\mathbb{B})$. Then we derive a decomposition of $Q_n = \otimes_{i=1}^n Q_{n,i}$ in the total variation case by introducing a signed measure $\Delta_i \in \mathcal{M}_1(\mathbb{B})$:

$$dQ_{n,i} = dF + r_n d\Delta_i \tag{1.4}$$

After having laid the basis for the main chapter 6 we come to the motivation for this thesis in section 3.2 that originates from a result in [Kohl (2005)]. In the context of determining the exact finite sample risk for sample size $n \geq 3$ M. Kohl uses Edgeworth expansions to compute an approximation as it seems to be impossible to achieve the expected results analytically[6]. Based on these insights on the higher order asymptotics on total variation neighborhoods the conjecture is that in this case the risk reads as

$$\sup_{\boldsymbol{Q_n} \in \tilde{\boldsymbol{\mathcal{Q}}}_n(\boldsymbol{r})} \boldsymbol{n}\mathbf{MSE}(\boldsymbol{S_n}, \boldsymbol{Q_n}) = \boldsymbol{r^2 b^2} + \mathbb{E}\boldsymbol{\psi^2} + \frac{\mathbf{1}}{\boldsymbol{n}} \boldsymbol{A_2} + \boldsymbol{o}\left(\frac{\mathbf{1}}{\boldsymbol{n}}\right) \tag{1.5}$$

which would indicate a faster rate of convergence. But the reason for the vanishing of the $n^{-1/2}$-term could as well be found in the symmetry of F, i.e. $f(x) = f(-x)$, which is used by M. Kohl throughout his investigations as there is $F_\theta = \mathcal{N}(\theta, 1)$. In case of convex contamination this symmetry condition indicates no vanishing of the $n^{-1/2}$-term, however.

The techniques we use to derive our results are based on exact approximations of the limit distribution. However, contiguity in the sense of convergence in distribution does not implicate contiguity of the risk necessarily. An argument based on the breakdown point illustrates this fact. So in section 3.3 we recall the concept of the finite sample breakdown point[7] and employ a convenient modification of the infinitesimal models in Definition 3.10 that on the one hand is asymptotically negligible, but on the other hand forces the unmodified MSE to converge along with weak convergence.

In **Chapter 4: First Order Optimality for Robust Estimation of Location in one dimension**, we specialize the terms of the general introduction in chapter 2 for the one dimensional case, as explicit and manageable bias terms for total variation only are available for one dimension. In section 4.1 we give the first order optimality result to show that under symmetry of F there is no possibility to see any differences between the convex contamination and the total variation case. Then in section 4.2 the setup for one-dimensional location is given explicitly in both types of neighborhoods. The chapter closes by discussing Huber's monotony approach[8] for M-estimators[9] that turns out to be useful for the location but not for the scale model, for example. In the latter case, an alternative approach by Taylor expansions of k-step-estimators is presented in subsection 4.3.2 that is used for the location model in chapter 8, too.

In **Chapter 5: A first simulation study**, we summarize the results of a simulation study that lead us to the closer examination of higher order expansions of the MSE in

[6] confer Remark 3.8 in this thesis or section 11.3.3 "Higher Order Approximations" of [Kohl (2005)], respectively.

[7] conf. Definition 3.9.

[8] conf. figure 4.1, especially.

[9] The concept of M- or Z-estimators, respectively, is sketched in subsection 4.3.1.

the following chapter. In order to produce appropriate observations we approximate them by a "dependence-creating" but straight forward algorithm (conf. section 5.1 and 8.3 or Appendix E.1.1 and E.1.2, respectively.). In detail, for each sample we generate observations from the least favorable element of the neighborhood system by picking up the K smallest observations and changing their sign. As estimator S_n we considered a three-step-estimator with the median as a starting estimate with Hampel-type IC. We compute the empirical asymptotic MSE[10] and apply the Box-Cox Power Transformation[11]. In the next step, we take a closer look and carry out a linear model on the empirical MSE[12] by application of the Akaike Information Criterium[13] (AIC) to indicate an appropriate structure of the linear model. In order to contrast our result to the convex-contamination case we add a short look at the corresponding Box-Cox-plots (conf. Fig. 5.2, 5.4 and 5.6) and regression results. Indeed, the results are in accordance with our conjecture that we have a convergence of order n^{-1}. A cross-check in subsection 5.2.4 against the numerical results in [Kohl (2005)] closes this chapter.

In the main **Chapter 6: Higher Order Asymptotics for the MSE in the One–dimensional location model**, we focus on the question of a higher order expansion for the MSE of robust M-estimators of location on shrinking total variation neighborhoods. For reasons of comparison, in Theorem 6.4 we first repeat the result for convex contamination from [Ruckdeschel (2005b)]. Then we go on with some preparative Definitions, Notations and Lemmata in subsection 6.2.1 until we can state our Main Theorem 6.13. Therein we give the explicit expansion of form (1.1) for the total variation case. The key idea of the approach consists of transferring decomposition (1.4) to the expectation and variance terms $L_{\mathrm{re},i}(t) := \mathbb{E}_{\mathrm{re}}\psi(x_i - t)$ and $V^2_{\mathrm{re},i}(t) := \mathrm{Var}\psi(x_i - t)$ directly, before expanding them by Taylor series to get access to the coefficients defining the terms A_1 and A_2 (conf. Assumption 6.7 and 6.18, respectively). Thereby putting in more structure of / information about the basic total variation neighborhoods we get expressions more complex than in the proof for the convex contamination case in [Ruckdeschel (2005b)] (98 summands compared to 63 summands for a certain comparable polynomial, for instance, confer Remark 6.16).

The proof of Theorem 6.13 in subsection 6.2.2 is prefixed by an outline packing the rather laborious character of the proof into 15 steps, before going on to the detailed deduction of the main result: after a partition of the real line, we show the negligibility of several cases (via the Chebyshev inequality and a Hoeffding bound, conf. Appendix B) and hence can confine ourselves to a shrinking compactum[14], wherein we apply an Edgeworth expansion[15] to the centered and standardized influence curve $\psi_{t,i}$. The massive use of the CAS `MAPLE`[16] enables us to compute several complex Taylor expansions of the integrand by keeping hold on the order of hundreds of terms. Additionally, the detection of a least favorable modification of the data with respect to the total variation bias term (conf.

[10] For the definition of the empirical asymptotic MSE equation (5.4).

[11] The Box-Cox Power Transformation is provided by the MASS package of [Venables and Ripley (1999)] and originates from [Box and Cox (1964)].

[12] For the definition of the empirical MSE equation (5.3)

[13] conf. equation (5.6).

[14] conf. interval I in Figure 6.1.

[15] conf. Theorem A.5.

[16] The `MAPLE`-algorithm used is described in section E.2.

(6.64) and (6.65), respectively) leads us to the calculation of the final terms.
In order to prove our conjecture (1.5), in section 6.3 we confine ourselves to the symmetric case. Actually, we are able to confirm the vanishing of the A_1-term in Corollary 6.19. By an argument taken from [Feller (1971)], we show in the Lemmata 6.21 and 6.22 that the simplific assumption of identically distributed variables in the proof of Theorem 6.13 actually is no limitation. By the higher order expansion there are consequences for the ideal model, i.e. in the case $r = 0$, as we show in Corollary 6.20 and compare to the convex contaminated case as a cross-check. We see the expected consistency here. As a preparation for the following chapter, in section 6.6 we provide the corresponding terms in case of $F = \mathcal{N}(0, 1)$ and state in Proposition 6.23 the coefficients from the (symmetric) A_2-term in dependence on the normal density. This result is accompanied by Remark 6.25 concerning the convex contaminated case.
In **Chapter 7: Numerical investigation of the Higher Order MSE**, by application of the coefficients for the representative symmetric setup $F = \mathcal{N}(0, 1)$ we examine the behavior of the exact asymptotic MSE paying special attention to the A_2-term and compare the results of first, second and third order asymptotics. Just looking at the numerical results in subsection 7.2.1 it seems that in contrast to the convex contamination case, the total variation setup leads to the conjecture that under symmetry and for large enough n, the maximal MSE on $\tilde{\mathcal{Q}}_n$ is always overestimated for small radius but underestimated for large radius by first-order (and second-order) asymptotics!.
A closer investigation of the A_2-term in subsection 7.2.2 then shows that for small radii (depending on the clipping height) we always get a negative contribution to the MSE, indeed. Hence we get an overestimation of the MSE. But the situation changes by increasing (both) radius and the clipping height (conf. Remark 7.3 and Fig. 7.1). We offer the heuristic explanation that in these situations we cannot apply, or achieve, respectively, the least favorable deviation as we do in chapter 6, in order to get the asymptotic expansion of the MSE. The result is a MSE that increases beyond every bound with each additional "bad" member of the sample.
In **Chapter 8: Generation of least favorable deviations in total variation for finite sample** we deal with the question of an actual realization of a least favorable deviation in a finite context as detected in the proof of the main theorem in chapter 6. In a finite scenario with original sample $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} P_n^{id}$ to be manipulated by the signed measure Δ_i as defined in (1.4) or (3.18), respectively, the least favorable deviation may not be possible. This means that we have to find and declare a suitable mechanism explaining the effect of Δ_i for every finite sample according to previous given conditions. In the face of Corollary 6.19 we settle on the symmetric case for the measure $F = P^{id}$ symmetric on the Borel set $\mathbb{B}$; the influence curve ψ is seen as monotone and as odd. We show that for a certain kind of manipulation mechanism[17] we then gain the theoretically proven result up to suitable order. In this sense we first carry out a reordering of the sample by conditioning with respect to the arrangement. Actually, we confine ourselves to influence curves of Hampel-type form, at least attaining their maximum for $|x| > c_n$ with a general increasing sequence c_n initially (conf. Fig. 8.1). The amount k of manipulable observations is given by a random variable K with first moment $\mathbb{E}K = r\sqrt{n}$ chosen

[17] conf. section 8.3.

to satisfy the requirement of staying in a total variation ball $B_v(F, r/\sqrt{n})$, conf. Lemma 8.5 and figure 8.2. The second moment $\mathrm{Var}K = \frac{1}{2}r\sqrt{n}$ results from a deeper investigation of all terms in the expansion of the MSE in the proof of Theorem 8.14 given by a k-step-approach. By ordering the sample the observations become (weakly) correlated, however, confer Proposition 8.16 and Theorem 8.17. But in in Theorem 8.20 we can show that under certain assumptions to choose the correlation vanishes. Without application of further symmetry arguments we are confronted with the common law of the k- and $n-k+1$-quantiles $X_{[k:n]}$ and $X_{[n-k+1:n]}$, which leads to questions concerning order statistics. But the integrals to be evaluated in this setup show up to be very hard to handle, so we just give a short impression of these circumstances in subsection 8.5.1 and make use of a symmetry argument loosely inspired by the reflection principle known from elementary stochastic in section 8.6: by consideration of several samples $\{x_1, \ldots, x_n\}_j \overset{\text{i.i.d.}}{\sim} F,\ j \in \mathbb{N}$, at once we are able to neglect the difference between the lower and upper k-quantile.
Furthermore it shows up in section 8.8 that we only get access to the result of Corollary 6.19 in the finite context if we require in Assumption 8.19 (p) the finite sample and the IC, respectively, to attain the minimum and maximum of the given influence curve ψ with a certain probability already. Depending on this probability we derive a lower bound on the sample length n in Theorem 8.20, after having conjectured the existence of a such condition by preceding numerical investigations (conf. subsection 8.8.1). Finally, we give a restrictive condition on the distribution of K in assumption 8.21 (PK) that guarantees[18] the desired realization of $X_{[k:n]}$ beyond a now concrete bound c_n. The bound c_n is explicitly calculated for $F = \mathcal{N}(0,1)$ in Proposition 8.24 and at last suitable four-point distributions for K are given in subsection 8.9.3 that satisfy all the previous claimed conditions.
In **Chapter 9: Outlook** we list some imaginable extensions to this thesis concerning open questions.
The **Appendix** provides several additional results for or from the previous chapters, respectively. Appendix A contains some tools used in the proofs of chapter 6 and 8 like Hoeffding bounds, Mills's ratio or a theorem concerning Edgeworth expansions. Appendix B spells out the negligibility of the cases residing outside the shrinking compactum in the proof of theorem 6.13. The general A_2-term, calculated in the same Theorem, was shifted to Appendix C because of its complicated and longish form. The A_2-term in Corollary 6.19 is gained as a special case of this general one. Appendix D provides some general results concerning distributions and densities of the common law of two quantiles. In this context we also collect further rather technical Lemmata that are needed in chapter 8. Appendix E gives a short description of the algorithms used both for R and `MAPLE`. Some words are spent on the SWEAVE package for R and LaTeX, too. Eventually, to support chapter 2, Appendix F contains some further classical results from asymptotic statistics like local asymptotic normality, the convolution representation and the asymptotic minimax bound. In Appendix G we list some errata.

[18] Theorem 8.22 shows the probability of exceeding the bound c_n negligible exponentially by assumption (PK).

Chapter 2

Robust Statistics and its Asymptotic Theory

2.1 What is Robust Statistics?

Instead of giving references in literature at once, we like to proceed in a more fashioned way as it seems to be en vogue nowadays when gaining information on a new subject. Therefore we turn on our computer and google the words "Robust Statistics". The first hit is what we expected: a link to the free encyclopedia Wikipedia, more precisely the website [Wikipedia (2008)]. We follow the link and are welcomed by an introductory paragraph:

> **Robust statistics** provides an alternative approach to classical statistical methods. The motivation is to produce estimators that are not unduly affected by small departures from model assumptions.

Now, although it seems to be a modern agreement to believe in anonymous texts without reference to sources or authors we like to give some reliable foundation to the probably most read information about robust statistics.
First, the Wikipedia introduction strongly reminds on the introductory words of Huber in [Huber (1997)], p.1, or [Huber (1981)], p.1 , respectively:

> The word "robust" is loaded with many - sometimes inconsistent - connotations. We shall use it in a relatively narrow sense: for our purposes, *"robustness" signifies insensitivity against small deviations from the assumptions.*
> Primarily, we shall be concerned with distributional robustness: the shape of the true underlying distribution deviates from the assumed model (usually the Gaussian law).

Another characterization of robust statistic is given in the introduction of [Hampel et al. (1986)], p.1:

> Robust statistics, in a loose, nontechnical sense, is concerned with the fact that many assumptions commonly made in statistics (such as normality, linearity, independence) are at most approximations to reality. One reason is the

> occurrence of gross errors, such as copying or keypunch errors. They usually show up as outliers, which are far away from the bulk of the data, and are dangerous for many classical statistical procedures. The outlier problem is well known and probably as old as statistics, and any method for dealing with it, such as subjective rejection or any formal rejection rule, belongs to robust statistics in this broad sense.

Here the "outlier problem" is mentioned. The Wikipedia website also delivers a small paragraph "Manual screening for outliers" on this subject. The encyclopedia tells us that although statisticians traditionally remove obvious outliers in a data set by manual screening methods, data sets in modern times on the one hand consist of large numbers of variables being measured on large numbers of experimental units and therefore turn out to be infeasible by manual screening. On the other hand it is pointed out that outliers often can interact in such ways that they mask each other out:

> As a simple example, consider a small univariate data set containing one modest and one large outlier. The estimated standard deviation will be grossly inflated by the large outlier. The result is that the modest outlier looks relatively normal. As soon as the large outlier is removed, the estimated standard deviation shrinks, and the modest outlier now looks unusual.
> This problem of masking gets worse as the complexity of the data increases. For example, in regression problems, diagnostic plots are used to identify outliers. However, it is common that once a few outliers have been removed, others become visible. The problem is even worse in higher dimensions.

To go back to our tangible sources we like to quote on p.11 of [Hampel et al. (1986)], where "The Aims of Robust Statistics" are formulated:

(i) To describe the structure best fitting the bulk of the data.

(ii) To identify deviating data points (outliers) or deviating substructures for further treatment, if desired.

(iii) To identify and give a warning about highly influential data points ("leverage points").

(iv) To deal with unsuspected serial correlations, or more generally, with deviations from the assumed correlation structures.

For more detailed explanations of the single points we refer the reader to chapter 1 of [Hampel et al. (1986)]. But we would like to end up now with those general, non technical statements on robust statistics and go on to a more mathematical treatise.

2.1.1 Cniper: a most innocent least favorable contamination

First, we would like to widen the scope of the Wikipedia formulation on the problem of outliers. Actually, there is one sort of contamination of a data set that is hardly recognizable. Picking up the qualitative statement of "deviations from the assumptions that are near

or below the limits of detectability" from [Huber (1997)], p.61 , [Ruckdeschel (2005a)], for instance, quantifies this "most innocent least favorable contamination", calling it "Cniper" contamination, confer section 5 (ibid.). In this context we look at a distribution Q_n of the data, resulting from a convex contamination of a ideal measure F_θ and contaminations by Dirac measures at $a \in \mathbb{R}$; i.e.

$$Q_n(r,a) = \left[\left(1 - \frac{r}{\sqrt{n}}\right) F_\theta + \frac{r}{\sqrt{n}} \mathbb{I}_{\{a\}}\right]^n$$

with a contamination radius $r \in [0, \infty]$ in the sense of shrinking[1] infinitesimal neighborhoods at the rate of $\sqrt{n}$. Furthermore, we restrict the value of a in a way that under $Q_n(r,a)$ the arithmetic mean $\bar{X}_n$ of the sample does not perform better than a robust estimator S_n^c with influence curve of Hampel-type form given a priori; i.e.,

$$a := \inf\left\{t > 0 | \mathrm{MSE}_{Q_n(r,a)}(\bar{X}_n) > \mathrm{MSE}_{Q_n(r,a)}(S_n^c)\right\}$$

It is shown in section 5.3 of [Ruckdeschel (2005a)] that under $Q_n(r,a)$ we obtain

$$n\mathrm{MSE}_{Q_n(r,a)}(\bar{X}_n) = \left(1 - \frac{r}{\sqrt{n}}\right) + a^2\left(r^2 + \sqrt{r}\sqrt{n} - \frac{r^2}{n}\right)$$

which leads for $M_c := n \max \mathrm{MSE}(S_n^c)$ to

$$a = \sqrt{\frac{M_c - \left(1 - \frac{r}{\sqrt{n}}\right)}{r^2 + \frac{r}{\sqrt{n}} - \frac{r^2}{n}}} \tag{2.1}$$

In the introduction of [Kohl (2005)] M. Kohl did an exemplary computation for the finite-sample maximum MSE of asymptotically optimal robust estimators for sample size $n = 16$, $r = 0$ and $r = 0.2$. The results are summarized in table 1 (ibid.). For these estimators the value for a is computed by equation (2.1) between 2.345 and 2.427. To quote M. Kohl: *"Such small contaminations, which lie well within 3 standard deviations, will hardly be detected by outlier rejection rules."* M. Kohl additionally performed several tests for normality (Anderson-Darling, Cramer-von Mises, Kolmogorov-Smirnov, Shapiro-Wilk) on this sample using the R package fBasics. He gained type II errors (the null hypothesis is not rejected when it is false) in the range between 93.3% and 94.2%. To quote M. Kohl again: *"... the results for the chosen tests are very similar and indicate that the power (ability to reject the null hypothesis when it is actually false) of goodness-of-fit tests is very small in the case of such innocent contaminations."* Thus, we conclude together with M. Kohl and according to the superficial paragraph on Wikipedia that small deviations have nontrivial effects on statistical procedures and cannot be detected surely by goodness-of-fit tests.

2.1.2 Simple examples

To go on with the mathematical concepts of robust statistics we look again at the Wikipedia page and read:

[1] A motivation for this shrinkage is given in [Ruckdeschel (2004)], for example.

> Robust methods provide automatic ways of detecting, downweighting (or removing), and flagging outliers, largely removing the need for manual screening. [...] In order to quantify the robustness of a method, it is necessary to define some measures of robustness. Perhaps the most common of these are the breakdown point and the influence function ...

In order to provide some basis for this statement we give a standard example of robust statistics, which is the setup of this thesis, too: the estimation of location in the one-dimensional case. On Wikipedia there are some diagrams illustrating the behavior of the mean and, as an example for a kind of ad hoc robust estimator, the 10% trimmed mean on a data set[2] relating to speed of light measurements made by the Canadian-American astronomer and mathematician Simon Newcomb (1835-1909). But as it lacks mathematical details we will shortly provide some in form of two simple examples. Therefore we look at the location model $F_\theta = \Phi(x-\theta)$ and choose the mean and the median as the two extreme counterparts of estimating the location parameter, introducing the concepts of breakdown point and influence curve by application and defer the theoretical background to the next subsection.

(1) The mean given by $T_n = \frac{1}{n}\sum_{i=1}^{\infty} X_i$ has an influence curve $IC(x;T,\Phi) = x$, which means that every observation is weighted by its actual value. As $\mathrm{Var}(T_n,\Phi) = 1$ and Fisher information $\mathcal{I}(\Phi) = 1$ we have $\mathrm{Var}T_n = \mathcal{I}^{-1}$, which means that the mean has best efficiency 1, attaining the Cramér-Rao bound. But besides this variance optimality result the mean possesses bad robustness features, as both its asymptotically infinitesimal bias $\sup_x |IC| = \infty$ is unbounded and its rejection point $\inf(r > 0 : IC(x) = 0,\ |x| > r) = \infty$. Furthermore the breakdown point (the proportion of incorrect observations an estimator can handle before giving an arbitrarily large result) is 0. All in all this means that only one (arbitrarily large) outlier suffices to tamper the result.

(2) The median given by $T(F) = F^{-1}\left(\frac{1}{2}\right)$ has influence curve $IC(x;T,\Phi) = \frac{sign(x)}{2\Phi(0)}$, which means that every positive (negative) observation is weighted equally the same by $\frac{1}{2\Phi(0)} = \sqrt{\frac{\pi}{2}}$. As $\mathrm{Var}(T_n,\Phi) = (2\Phi(0))^{-2} = \frac{\pi}{2} \approx 1.571$ we have an efficiency of $\frac{1}{\mathrm{Var}T_n} = \frac{2}{\pi} \approx 0.637$, which is far less than the efficiency of the mean. But besides this the median appears to have good robustness features, as its bias $\sup_x |IC| = (2\Phi(0))^{-1} = \sqrt{\frac{\pi}{2}} \approx 1.253$ is bounded. Indeed, 1.253 is the minimal value as it is shown in subsection 2.5c of [Hampel et al. (1986)], for example, which makes the median the most robust estimate of location. But although with rejection point ∞ outliers do have influence on the median, the breakdown point calculates to its highest level of 0.5, saying that the median will not be driven to infinity unless 50% of the data is (badly) contaminated - a situation worst, when it is not possible any longer to distinguish between the underlying and the contaminating distribution.

[2]This data set is taken from [Stigler (1977)] and also available at http://www.stat.columbia.edu/g̃elman/book/data/light.asc.

2.1.3 The concept of influence curves

But what exactly is an influence curve IC as it is mentioned on the Wikipedia page or used in the two examples above? It is part of the so called "differential approach" going back to Hampel in [Hampel (1968)] and several subsequent surveys by Hampel, Krasker, Ronchetti, Rousseeuw et al.. The differential approach is based on three central concepts: qualitative robustness, influence function and breakdown point. They correspond in some sense to continuity, first derivative of a function and nearest singularity. Without stressing on these concepts in too much detail we stay with the influence curve (IC) or influence function (IF). Providing the richest quantitative robustness information, it describes the (approximate and standardized) effect of an additional observation in any point x on a statistic T, given a (large) sample with distribution F. Roughly speaking, the influence function is the first derivative of a statistic T at an underlying distribution F, where the point x plays the role of the coordinate in the infinite-dimensional space of probability distributions. But by the interpretation of derivation as a linearization of function, we can use an one-step Taylor-expansion to replace our statistic T by a linear statistic or functional, respectively. In the infinite dimensional setup there are different notions of a derivative. The following notion is taken from [Huber (1981)], p.35.

Definition 2.1. *We say that a statistical functional T is Frechét differentiable at F if it can be approximated by a linear functional L (defined on the space of finite signed measures) such that, for all distributions G,*

$$|T(G) - T(F) - L(G - F)| = o(d_*(F, G)) \tag{2.2}$$

with d_* a metric in the space $\mathcal{M}$ of probability measures, that:

(1) d_* is compatible with the weak topology in the sense that $\{F|d_*(G, F) < \varepsilon\}$ is open for all $\varepsilon > 0$.

(2) d_* is compatible with the affine structure of $\mathcal{M}$: let $F_t = (1 - t)F_0 + tF_1$, then $d_*(F_t, F_s) = O(|t - s|)$.

Remark 2.2. *a) The usual distance functions metrizing the weak topology satisfy both conditions, i.e. the Levy metric[3] d_λ, the Prokhorov metric[4] d_π and the bounded Lipschitz metric d_{BL}. For a short proof with respect to Condition (2) in the previous definition we refer to [Huber (1981)], p. 35.*

b) Another (strong) concept of differentiability is compact differentiability (confer [Rieder (1994)], [Reeds (1976)], [Serfling (1980)] or [Fernholz (1983)]).

The next proposition is taken from [Huber (1981)], p. 37, called Proposition 5.1 (ibid.).

Proposition 2.3. *If T is weakly continuous in a neighborhood of F and Frechét differentiable at F, then its Frechét derivative at F is a weakly continuous linear functional, and it is representable as*

$$L(G - F) = \int \psi_F dG \tag{2.3}$$

[3] confer (3.7)
[4] confer (3.5)

with ψ_F bounded and continuous, and $\int \psi_F dF = 0$.

Proof. Proposition 5.1 of [Huber (1981)] □

This function ψ_F will be called influence function. But first, we have to weaken up our concept of differentiability, because unfortunately the concept of Frechét differentiability appears to be too strong or elaborate, respectively. A way out is offered by the weakest concept of differentiability, the *Gâteaux derivative*. The following definition is taken from [Huber (1997)] and mixed up with the definition in [Hampel et al. (1986)] p. 83:

Definition 2.4. *A functional T is called Gâteaux differentiable at F, if there is a function ψ such that for all $G \in \mathcal{M}$,*

$$\lim_{t\to 0} \frac{T((1-t)F + tG) - T(F)}{t} = \int \psi(x) G(dx).$$

which may also be written as

$$\frac{\partial}{\partial t}[T((1-t)F + tG)]_{t=0} = \int \psi(x) dG(x). \tag{2.4}$$

We take the next remarks from [Huber (1981)] and [Hampel et al. (1986)], p. 84:

Remark 2.5. *a) Clearly, if T is Frechét differentiable, it is also Gâteaux differentiable, and the two derivatives agree.*

b) The basic idea of differentiation of statistical functionals goes back to von Mises ([von Mises (1937)], [von Mises (1947)]) and Filippova ([Filippova (1961)]); one says that T is a von Mises functional, with first kernel function ψ.

At this point the practical meaning of $\psi(x)$ is not yet evident, as it appears only implicitly in (2.4). Following Hampel an explicit expression may be obtained by putting $G = \delta_x$ in (2.4). We then get Definition 1 of [Hampel et al. (1986)] p. 84, for an influence function $\psi(x)$:

Definition 2.6. *The influence function $\psi(x)$ of T at F is given by*

$$\psi(x; T, F) = \lim_{t\to 0} \frac{T((1-t)F + t\delta_x) - T(F)}{t} \tag{2.5}$$

in those x where the limit exists.

The heuristically important interpretation of this form was first pointed out by Hampel in [Hampel (1968)]: $\psi(x)$ gives the suitably scaled differential influence of one additional observation with value x, if the sample size $n \to \infty$. Therefore, Hampel called it the *influence curve* (IC). Huber notes in this context that there indeed are pathological cases where the influence curve exists, but not the Gâteaux derivative, confer [Huber (1997)]. Therefore we use a more elegant definition of the influence function given by [Rieder (1994)]. To be able to give the definition we need the concept of L_2 differentiability. To avoid domination assumptions in the definition of L_2 differentiability, we employ the following square root calculus that was introduced by Le Cam. The following definition is taken from [Rieder (1994)]; for more details confer Subsection 2.3.1 of [Rieder (1994)].

Definition 2.7. *For any measurable space $(\Omega, \mathcal{A})$ and $k \in \mathbb{N}$ we define the following real Hilbert space that includes the ordinary $L_2^k(P)$*

$$\mathcal{L}_2^k(\mathcal{A}) = \{\xi\sqrt{dP} \mid \xi \in L_2^k(P), P \in \mathcal{M}_b(\mathcal{A})\} \tag{2.6}$$

On this space, an equivalence relation is given by

$$\xi\sqrt{dP} \equiv \eta\sqrt{dQ} \qquad \Longleftrightarrow \qquad \int |\xi\sqrt{p} - \eta\sqrt{q}\,|^2 d\mu = 0 \tag{2.7}$$

where $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^k$ and $\mu \in \mathcal{M}_b(\mathcal{A})$ may be any measure, depending on P and Q, so that $dP = p\,d\mu$, $dQ = q\,d\mu$. We define linear combinations with real coefficients and a scalar product by

$$\alpha\xi\sqrt{dP} + \beta\eta\sqrt{dQ} = (\alpha\xi\sqrt{p} + \beta\eta\sqrt{q}\,)\sqrt{d\mu} \tag{2.8}$$

$$\langle \xi\sqrt{dP} \mid \eta\sqrt{dQ}\,\rangle = \int \xi^\tau \eta \sqrt{pq}\,d\mu \tag{2.9}$$

Then for fixed $\theta \in \Theta$ we define L_2 differentiability of the family $\mathcal{P}$ at θ using the square root calculus; confer Definition 2.3.6 of [Rieder (1994)].

Definition 2.8. *Model $\mathcal{P}$ is called L_2 differentiable at θ if there exists some function $\Lambda_\theta \in L_2^k(P_\theta)$ such that, as $t \to 0$,*

$$\|\sqrt{dP_{\theta+t}} - \sqrt{dP_\theta}\,(1 + \tfrac{1}{2}t^\tau\Lambda_\theta)\|_{\mathcal{L}_2^k} = o(|t|) \tag{2.10}$$

$$\mathcal{I}_\theta = \mathbb{E}_\theta \Lambda_\theta \Lambda_\theta^\tau \succ 0 \tag{2.11}$$

The function Λ_θ is called the L_2 derivative and the $k \times k$ matrix $\mathcal{I}_\theta$ Fisher Information of $\mathcal{P}$ at θ.

The following definition corresponds to Definition 4.2.10 of [Rieder (1994)].

Definition 2.9. *Suppose $\mathcal{P}$ is L_2 differentiable at θ, and assume some matrix $D \in \mathbb{R}^{p\times k}$ of full rank $p \le k$. Let $\alpha = 2, \infty$, respectively.*
(a) *Then the set $\Psi_2(\theta)$ of all* square integrable *and the subset $\Psi_\infty(\theta)$ of all* bounded influence curves at P_θ *, respectively, are*

$$\Psi_\alpha(\theta) = \left\{\psi_\theta \in L_2^k(P_\theta) \mid \mathbb{E}_\theta\psi_\theta = 0,\ \mathbb{E}_\theta\Lambda_\theta^\tau = \mathbb{I}_k\right\} \tag{2.12}$$

(b) *The set $\Psi_2^D(\theta)$ of all* square integrable *and the subset $\Psi_\infty^D(\theta)$ of all* bounded, partial influence curves at P_θ, *respectively, are*

$$\Psi_\alpha^D(\theta) = \left\{\psi_\theta \in L_2^p(P_\theta) \mid \mathbb{E}_\theta\psi_\theta = 0,\ \mathbb{E}_\theta\Lambda_\theta^\tau = D\right\} \tag{2.13}$$

For the sake of completeness we add Remark 4.2.11 of [Rieder (1994)] parts (a) to (c):

Remark 2.10. **(a)** The attribute *square integrable* will usually be omitted.
(b) The *classical scores* and the *classical partial scores*,

$$\psi_{h,\theta} = \mathcal{I}_\theta^{-1}\Lambda_\theta \in \Psi_2(\theta) \tag{2.14}$$

$$\eta_{h,\theta} = D\psi_{h,\theta} = D\mathcal{I}_\theta^{-1}\Lambda_\theta \in \Psi_2^D(\theta) \tag{2.15}$$

are always ICs, respectively, partial ICs, at P_θ.
(c) The definition of $\Psi_2(\theta)$ and $\Psi_\infty(\theta)$ requires $\mathcal{I}_\theta \succ 0$, and Λ_θ nondegenerate in the sense that, for all $t \in \mathbb{R}^k$,

$$t^\tau \Lambda_\theta = 0 \quad \text{a.e. } P_\theta \qquad \Longrightarrow \qquad t = 0 \tag{2.16}$$

For questions of existence of (square integrable) partial ICs we cite Lemma 1.1.3 from [Kohl (2005)], which gives a necessary and sufficient condition.

Lemma 2.11. *It holds*

$$\Psi_2^D(\theta) \neq \emptyset \Leftrightarrow \exists A \in \mathbb{R}^{p\times k} : D = A\mathcal{I}_\theta \Leftrightarrow \ker\mathcal{I}_\theta \subset \ker D \tag{2.17}$$

Proof. Lemma 1.1.3 in [Kohl (2005)]. □

M. Kohl adds two further remarks on this lemma, from which we only cite part a):

Remark 2.12. *(a) The previous lemma shows that we do not necessarily need $\mathcal{I}_\theta \succ 0$ for the existence of partial ICs. But, since* $\operatorname{rank}(D) = p$*, it has to hold* $\operatorname{rank}(A\mathcal{I}_\theta) = p$ *where*

$$\operatorname{rank}(A\mathcal{I}_\theta) = \operatorname{rank}(\mathcal{I}_\theta) - \dim(\mathcal{C}(\mathcal{I}_\theta) \cap \mathcal{N}(A)) \tag{2.18}$$

with $\mathcal{C}(\mathcal{I}_\theta)$ the column space of $\mathcal{I}_\theta$ and $\mathcal{N}(A)$ the null space of A; confer Theorem 17.5.4 of [Harville (1997)]. Consequentially, the Fisher information $\mathcal{I}_\theta$ at least has to have rank p.

Definition 2.9 turns out be be very useful in the context of robust asymptotic statistics as most proofs of asymptotic normality in the i.i.d. case head for an estimator expansion, in which ICs canonically occur as summands. This leads to the framework of Asymptotic Statistics in the next section.

2.2 Asymptotic Theory of Robustness

The aim now is to detect optimal influence functions in a certain sense of optimality. But only for a relatively small number of statistical problems there exists an exact, optimal solution. For instance, the Neyman-Pearson theory leads to optimal (uniformly most powerful) tests in certain exponential family models and the Rao-Blackwell theory allows to conclude that certain estimators are of minimum variance among unbiased estimators[5]. If exact optimality theory does not give results, then asymptotic optimality theory can help, although it is what it is: an approximation.
Van der Vaart gives several advantages of the asymptotic approach in [van der Vaart (1998)], p. 3:

- The maximum likelihood estimators are asymptotically consistent: The sequence of estimators converges in probability to the true value of the parameter.

[5] In particular the unbiasedness may be questioned, confer problem 8 in [Bickel and Doksum (2001)], section 1.3 (BLUE, MLE and MSE-optimal estimator for the sample variance).

- The rate at which maximum likelihood estimators converge to the true value is the fastest possible, typically $1/\sqrt{n}$.
- Their asymptotic variance, the variance of the limit distribution of $\sqrt{n}(S_n - \theta)$, is minimal; in fact, maximum likelihood estimators "asymptotically attain" the Cramér-Rao bound.
- Even though the method of maximum likelihood often leads to reasonable estimators and has great intuitive appeal, in general it does not lead to best estimators for finite sample. Thus the use of an asymptotic criterion simplifies optimality theory considerably.

Now, our wish to construct asymptotically efficient estimators leads us to the concept of asymptotic linear estimators mentioned in chapter 25.9 of [van der Vaart (1998)] and used as a basis in [Rieder (1994)].

2.2.1 Asymptotically Linear Estimators

We give the definition of asymptotically linear estimators (ALEs) from Definition 4.2.16 of [Rieder (1994)].

Definition 2.13. *An asymptotic estimator*

$$S = (S_n) \qquad S_n : (\Omega^n, \mathcal{A}^n) \to (\mathbb{R}^k, \mathbb{B}^k) \tag{2.19}$$

is called asymptotically linear at P_θ *if there is an IC* $\psi_\theta \in \Psi_2(\theta)$ *such that*

$$R_n = \sqrt{n}\,(S_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\theta(y_i) + o_{P_\theta^n}(n^0) \tag{2.20}$$

We call $R = (R_n)$ standardization, *and* ψ_θ *the* IC, *of* S *at* P_θ.

We state Remark 4.2.17 of [Rieder (1994)] without part (c) on L_1 differentiability and part (f) on the nonparametric convolution and asymptotic minimax theorems.

Remark 2.14. *(a) The expansion (2.20) determines the IC* ψ_θ *uniquely, because* $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta(y_i)$ *with* $\eta \in L_2^k(P_\theta)$, $\mathbb{E}_\theta \eta = 0$, *can tend to* 0 *in* P_θ^n *probability only if* $\mathbb{E}_\theta |\eta|^2 = 0$; *that is,* $\eta = 0$ *a.e.* P_θ.
(b) If S *is asymptotically linear at* P_θ *with IC* $\psi_\theta \in \Psi_2(\theta)$, *then*

$$\sqrt{n}\,(S_n - \theta)(P_\theta^n) \xrightarrow{w} \mathcal{N}\,(0, Cov_\theta(\psi_\theta)) \tag{2.21}$$

because of $\psi_\theta \in L_2^k(P_\theta)$, $\mathbb{E}_\theta \psi_\theta = 0$, *and the Lindeberg-Lévy theorem. The third condition* $\mathbb{E}_\theta \psi_\theta \Lambda_\theta^\tau = \mathbb{I}_k$, *as already noted in the remarks of [Rieder (1980)] (p. 108), is equivalent to the locally uniform extension of this asymptotic normality; see Lemma F.8.*
(d) *Extending general M estimates, the class of ALEs has in the case* $k = 1$ *been introduced by [Rieder (1980)]. [Bickel (1981)] defined the related notion CULAN, employing however compact subsets of* Θ *instead of compacts in the local parameter space.*

(e) The class of ALEs contains the common asymptotically normal M, L, R and many MD (minimum distance) estimates; confer chapters 1 and 6 of [Rieder (1994)]. In fact, most proofs of asymptotic normality in the i.i.d. case end up with an extension (2.20); the corresponding conditions need to be verified only under the ideal model.
(g) The previous robustness theories of [Huber (1964)], [Hampel (1974)], [Rieder (1980)] and [Bickel (1981)] have been formulated but for ALEs or, even more specialized, for M estimates.

For the Cramér-Rao bound of AL estimators we repeat Proposition 1.1.7 from [Kohl (2005)].

Proposition 2.15. *Consider an estimator $S = (S_n)$ that is asymptotically linear at P_θ with IC $\psi_\theta \in \Psi_2(\theta)$. Then*

$$\mathrm{Cov}_\theta(\psi_\theta) \succeq \mathcal{I}_\theta^{-1} = \mathrm{Cov}_\theta(\psi_{h,\theta}) \tag{2.22}$$

in the positive definite sense, with equality iff $\psi_\theta = \psi_{h,\theta}$.

Proof. Proposition 1.1.7 in [Kohl (2005)] following sections 3.2, 3.3 of [Rieder (1994)] □

As already mentioned in Remark 1.1.8 of [Kohl (2005)] this optimality result can be extended to arbitrary, measurable estimators. Joining Kohl we refer for more details to Sections 3.2, 3.3 of [Rieder (1994)], Sections 8.5, 8.7 of [van der Vaart (1998)] or Section 2.3 of [Bickel et al. (1998)].

For further classical results of asymptotic statistics we refer to Appendix E.

2.3 The Infinitesimal Robust Setup

We introduce the infinitesimal robust setup according to Subsection 4.2.1 of [Rieder (1994)]. A more detailed introduction to this topic is given in [Bickel (1981)], for example.

In this sense we look at a neighborhood system

$$\mathcal{U}(\theta) = \big\{U(\theta, r) \,\big|\, r \in [0, \infty)\big\} \tag{2.23}$$

with $U(\theta, r)$ of radius $r \in [0, \infty)$ about P_θ such that

$$P_\theta \in U(\theta, r_1) \subset U(\theta, r_2) \subset \mathcal{M}_1(\mathcal{A}) \qquad 0 \le r_1 < r_2 < \infty \tag{2.24}$$

To anticipate Remark 2.19, we point out that these neighborhoods are considered as shrinking balls at the rate of $1/\sqrt{n}$. This approach was first utilized by [Huber-Carol (1970)] and an explicit motivation for this shrinkage is given in [Ruckdeschel (2004)], for example, where a more general remark of [Huber (1997)], p.62, is worked out in detail.

Within this thesis we restrict ourselves to (convex) contamination ($* = c$) and total variation ($* = v$) neighborhood systems $\mathcal{U}_*(\theta)$.

Remark 2.16. *[Rieder (1994)] also considers Hellinger ($* = h$), Kolmogorov ($* = \kappa$), Cramér-von Mises ($* = \mu$), Prokhorov ($* = \pi$) and Lévy ($* = \lambda$) neighborhood systems.*

For $* = c, v$ the system $\mathcal{U}_*(\theta)$ consists of closed balls about P_θ defined for an arbitrary sample space,

$$U_*(\theta, r) = B_*(P_\theta, r) \qquad r \in [0, \infty) \tag{2.25}$$

where

$$B_c(P_\theta, r) = \big\{(1-r)_+ P_\theta + (1 \wedge r)\, Q \,\big|\, Q \in \mathcal{M}_1(\mathcal{A})\big\} \tag{2.26}$$

$$B_v(P_\theta, r) = \big\{Q \in \mathcal{M}_1(\mathcal{A}) \,\big|\, d_v(Q, P_\theta) \le r\big\} \tag{2.27}$$

with metric

$$d_v(Q, P_\theta) = \frac{1}{2}\int |dQ - dP_\theta| = \sup_{A \in \mathcal{A}} |Q(A) - P_\theta(A)| \tag{2.28}$$

and it holds $B_c(P_\theta, r) \subset B_v(P_\theta, r)$.

Most of robust optimality theory in the infinitesimal setup may already be derived by means of the smaller subclass of *simple perturbations*. They are introduced in [Rieder (1994)] pp. 125, and similar but with an extra approximation, in [Bickel (1981)], pp.16. For the introduction of simple perturbations of P_θ we first define p-dimensional tangents at P_θ.

Definition 2.17. *For any dimension $p \in \mathbb{N}$ and exponent $\alpha = 2, \infty$, respectively, we define*

$$Z_\alpha^p(\theta) = \big\{\zeta \in L_\alpha^p(P_\theta) \,\big|\, \mathbb{E}_\theta \zeta = 0\big\} \tag{2.29}$$

The elements of $Z_2^p(\theta)$ respectively, $Z_\infty^p(\theta)$ are called square integrable *respectively,* bounded *p-dimensional tangents at P_θ. If a parametric model $\mathcal{P}$ is L_2 differentiable at θ, the L_2 derivative Λ_θ is called* parametric tangent.

Definition 2.18. *A sequence $Q_n(\zeta, .\,)$ of* simple perturbations *of P_θ along $\zeta \in Z_2^k(\theta)$ is given by*

$$dQ_n(\zeta, t) = \Big(1 + \frac{1}{\sqrt{n}} t^\tau \zeta_n\Big) dP_\theta \qquad |t| \le \frac{\sqrt{n}}{\sup_{P_\theta} |\zeta_n|} \tag{2.30}$$

where the approximating bounded tangents $\zeta_n \in Z_\infty^k(\theta)$ are chosen such that

$$\lim_{n\to\infty} \mathbb{E}_\theta |\zeta_n - \zeta|^2 = 0 \qquad \sup_{P_\theta} |\zeta_n| = o(\sqrt{n}\,) \tag{2.31}$$

Every $t \in \mathbb{R}^k$ is eventually admitted as a parameter value. In case of $\zeta \in Z_\infty^k(\theta)$ we may choose $\zeta_n = \zeta$; confer Remark 4.2.3 of [Rieder (1994)].

The contamination and total variation neighborhood systems cover simple perturbations along $Z_\infty^k(\theta)$ ($* = c$) respectively, $Z_2^k(\theta)$ ($* = v$), on the $1/\sqrt{n}$ scale, confer [Rieder (1994)], parts (c) and (v) of Lemma 4.2.8. In order to explain the terminus *infinitesimal* in connection with neighborhoods we note Remark 4.2.7 of [Rieder (1994)].

Remark 2.19. *With the $1/\sqrt{n}$ scaling, a neighborhood system is also called infinitesimal. For sample size $n \to \infty$, neighborhoods and simple perturbations are scaled down so, because, on the one hand, such deviations from the ideal model have nontrivial effects on statistical procedures, while, on the other hand, they cannot be detected surely by goodness-of-fit tests.*

For $\theta \in \Theta$ fix we specialize on the one-dimensional case and get the following subclasses of bounded tangents

$$\mathcal{G}_c(\theta) = \big\{q \in Z_\infty(\theta) \,\big|\, \inf_{P_\theta} q \geq -1\big\} \tag{2.32}$$
$$\mathcal{G}_v(\theta) = \big\{q \in Z_\infty(\theta) \,\big|\, \mathbb{E}_\theta|q| \leq 2\big\} \tag{2.33}$$

where

$$Z_\infty(\theta) = \big\{q \in L_\infty(P_\theta) \,\big|\, \mathbb{E}_\theta q = 0\big\} \tag{2.34}$$

By formally identifying $t^\tau\zeta = rq$, the simple perturbations along $\zeta \in Z_\infty^k(\theta)$ are, for $\sqrt{n} \geq -r \inf_{P_\theta} q$,

$$dQ_n(q,r) = dQ_n(\zeta,t) = \left(1 + \frac{r}{\sqrt{n}}q\right)dP_\theta \tag{2.35}$$

Lemma 2.20. *Given $q \in Z_\infty(\theta)$ and $r \in (0,\infty)$. Then, in the cases $* = c, v$, for every $n \in \mathbb{N}$ such that $\sqrt{n} \geq -r \inf_{P_\theta} q$,*

$$Q_n(q,r) \in B_*(P_\theta, r/\sqrt{n}\,) \quad \Longleftrightarrow \quad q \in \mathcal{G}_*(\theta) \tag{2.36}$$

Proof. On identifying $t^\tau\zeta = rq$ this may be read off the parts (c) and (v) of the proof of Lemma 4.2.8 in [Rieder (1994)]. □

In the next Proposition and Remark, appearing in this form as Proposition 1.2.5 and Remark 1.2.6 in [Kohl (2005)], we see that asymptotic linear estimators are asymptotically normal distributed under simple perturbations.

Proposition 2.21. *Let S be an estimator that is asymptotically linear at P_θ with IC $\psi_\theta \in \Psi_2(\theta)$ and given $q \in Z_\infty(\theta)$ and $r \in (0,\infty)$ consider the simple perturbations $Q_n(q,r)$. Then*

$$\sqrt{n}\,(S_n - \theta)(Q_n^n(q,r)) \xrightarrow{w} \mathcal{N}_k(\mathbb{E}_\theta\psi_\theta q,\, Cov_\theta(\psi_\theta)) \tag{2.37}$$

for all convergent sequences $t_n \to t$ in $\mathbb{R}^k$.

Proof. Consequence of Lemma 4.2.4 in [Rieder (1994)] together with Slutzky's lemma, the Cramér-Wold device and Le Cam's third lemma (Theorem F.2) □

Remark 2.22. Assume transforms $\tau\colon \mathbb{R}^k \to \mathbb{R}^p$ $(p \leq k)$ which are differentiable at θ with bounded derivative $D = d\tau(\theta)$ of full rank p,

$$\tau(\theta + t) = \tau(\theta) + Dt + o(|t|) \qquad rkD = p \tag{2.38}$$

Then, one gets by the finite-dimensional delta method, setting $\eta_\theta = D\psi_\theta$

$$\sqrt{n}\,\big(\tau \circ S_n - \tau(\theta)\big)(Q_n^n(q,r)) \xrightarrow{w} \mathcal{N}_p(\mathbb{E}_\theta\eta_\theta q,\, \mathrm{Cov}_\theta(\eta_\theta)) \tag{2.39}$$

2.4 Optimal Influence Curves

In Lemma 5 of [Hampel (1968)] and, more generally, in [Hampel et al. (1986)], optimal robust influence curves are determined such that a corresponding M-estimator minimizes the asymptotic variance subject to a bound on the asymptotic bias, a context originally called *gross error sensitivity* there. Similarly to the setup of influence curves these problems arise in the infinitesimal robust setup as well and are solved in [Bickel (1981)], [Bickel (1984)], and, more generally, [Rieder (1994)] for several neighborhood systems and corresponding bias terms. The most common criterium combining bias and variance is the mean squared error.

2.4.1 Risk and MSE problems

With Proposition 2.21 and Remark 2.22 we get the following result stated in this form as Proposition 1.3.1 in [Kohl (2005)], for example, where the clipping of the loss function l by M in part (b) is only necessary for attaining the lower bound $\rho_0(q)$.

Proposition 2.23. *Let S be an estimator that is asymptotically linear at P_θ with IC $\psi_\theta \in \Psi_2(\theta)$ and given $q \in Z_\infty(\theta)$ and $r \in (0,\infty)$ consider the simple perturbations $Q_n(q,r)$. Moreover assume transforms $\tau\colon \mathbb{R}^k \to \mathbb{R}^p$ $(p \le k)$ of form (2.38) and let $\eta_\theta = D\psi_\theta$ and*

$$\rho_0 = \int \ell \, d\mathcal{N}_k(r\mathbb{E}_\theta \eta_\theta q,\ Cov_\theta(\eta_\theta)) \tag{2.40}$$

(a) *If $\ell\colon \mathbb{R}^p \to [0,\infty]$ is lower semicontinuous then for all $r \in (0,\infty)$,*

$$\liminf_{n\to\infty} \int \ell\big(\sqrt{n}\,(\tau \circ S_n - \tau(\theta))\big)\, dQ_n^n(q,r) \ge \rho_0(q) \tag{2.41}$$

(b) *If $\ell\colon \mathbb{R}^p \to [0,\infty]$ is continuous a.e. λ^p then for all $r \in (0,\infty)$,*

$$\lim_{M\to\infty} \lim_{n\to\infty} \int M \wedge \ell\big(\sqrt{n}\,(\tau \circ S_n - \tau(\theta))\big)\, dQ_n^n(q,r) = \rho_0(q) \tag{2.42}$$

Proof. Consequence of Proposition 2.21, Remark 2.22 together with the Lemma of Fatou in the version of Lemma A.2.1 of [Rieder (1994)] and the continuous mapping theorem. □

Being interested in the behavior of the mean squared error or its maximum, respectively, we choose

- $l(z) = |z|^2$ for quadratic loss and
- the supremum over all tangents $q \in \mathcal{G}_*(\theta)$ in the risk 2.42.

Thus we obtain the subsequent asymptotic mean square error (MSE) problems from the moments of the limit distribution in (2.37):

$$\mathrm{maxMSE}_\theta(\eta_\theta, r) := E_\theta|\eta_\theta|^2 + r^2\omega_{*,\theta}(\eta_\theta)^2 = \min\,! \qquad \eta_\theta \in \Psi_2^D(\theta) \tag{2.43}$$

$$\omega_{*,\theta}(\eta_\theta) \;=\; \sup\big\{|\mathbb{E}_\theta \eta_\theta q| \,\big|\, q \in \mathcal{G}_*(\theta)\big\} \tag{2.44}$$

with fixed radius $r \in (0,\infty)$ of the simple perturbations (2.35). In [Rieder (1994)] this leads to the so called Hampel type problem[6], with bound $b \in (0,\infty)$ fixed,

$$E_\theta|\eta_\theta|^2 = \min\,! \qquad \eta_\theta \in \Psi_2^D(\theta),\ \omega_{*,\theta}(\eta_\theta) \le b \tag{2.45}$$

The determination of the solutions is based on Langrange multiplier theorems derived in Appendix B of [Rieder (1994)]. But whereas [Hampel et al. (1986)] must assume the existence of Lagrange multipliers, [Rieder (1994)] (as well as [Bickel (1981)] and [Bickel (1984)]) proves their existence.

In terms of statistical risk, the following result may be interpreted as an extension of the classical Cramér-Rao bound under quadratic loss, with $\mathrm{tr}A = \mathrm{tr}\mathcal{I}^{-1}$.

Lemma 2.24. *It holds for the solution $\tilde{\eta}$ to the MSE problem* (2.43) *that*

$$\mathrm{maxMSE}\,(\tilde{\eta}, r) = trAD^\tau \tag{2.46}$$

Proof. [Kohl (2005)], Proposition 2.1.1, p. 19 □

We add Remark 2.1.2 from [Kohl (2005)].

Remark 2.25. *This correspondence for the asymptotic minimax MSE holds more generally and can be verified for the cases $* = c, v$, $t = 0$, ε, α, $s = 0, e, 2$ considered in [Rieder (1994)]. Exceptions are the cases $* = h$, $t = 0$, $s = 0, e$ and $* = h$, $t = \alpha = 2$, $s = e$, where the optimal robust ICs are identical to η_h and* $\mathrm{maxMSE}(\eta_h, r) = trD\mathcal{I}^{-1}D^\tau + r^2b^2$. □

Before turning to the solutions to these Hampel type problems we add Remark 1.3.3 from [Kohl (2005)], where we abbreviate part (a).

Remark 2.26. **(a)** *Actually, we are interested in the following limiting risk*

$$\lim_{M\to\infty}\lim_{n\to\infty}\sup_{Q\in U_*(\theta,r/\sqrt{n})}\int M\wedge\ell\big(\sqrt{n}\,(\tau\circ S_n-\tau(\theta))\big)\,dQ^n \tag{2.47}$$

Thus, it must be made sure, that at least for the optimal ICs, the interchanging of $\lim_M \lim_n$ *and* $\sup_Q$ *and the passage from the neighborhood submodel to full neighborhoods does not increase the asymptotic risk. Under additional assumptions on the optimal ICs, this goal can be achieved by suitable estimator constructions described in Chapter 6 of [Rieder (1994)].*

(b) *Since the normal distribution in* (2.40) *is fully specified by its first two moments, one might, analogously to pp. 197 of [Fraiman et al. (2001)], think of the following general optimality problem*

$$\sup_{q\in\mathcal{G}_*(\theta)} g\big(r\mathbb{E}_\theta\eta_\theta q, Cov_\theta(\eta_\theta)\big) = \min\,! \qquad \eta_\theta \in \Psi_2^D(\theta) \tag{2.48}$$

[6] in allusion to the problem solved in Lemma 5 of [Hampel (1968)]

for suitable functions g. By choosing $g(x_1, x_2) = |x_1|^2 + tr(x_2)$ *and* $g(x_1, x_2) = \infty \mathbb{I}_{\{|x_1|>b\}}(x_1) + tr(x_2)$*, respectively, this problem also covers the MSE and the Hampel type problem stated above. [Ruckdeschel and Rieder (2004)] consider the similar problem*

$$G(r\omega_*(\eta_\theta), \sqrt{\mathbb{E}_\theta|\eta_\theta|^2}) = \min!, \qquad \eta_\theta \in \Psi_2^D(\theta) \tag{2.49}$$

where G is some positive and convex function which is strictly isotone in both arguments. They show that the solution to (2.49) also solves the corresponding Hampel type problem (2.45), respectively the corresponding MSE problem (2.43) where one only has to transform the bias weight according to the given risk; confer Section 8.1 (ibid.). Using this fact, they derive necessary and sufficient conditions for the optimally robust ICs including an additional equation for the determination of the optimal bias bound b; confer Theorem 3.1 (ibid.).

2.4.2 Bias Terms

To lighten the notation we drop the fixed parameter θ. Hence we write $\omega_* = \omega_{*,\theta}$, $\eta = \eta_\theta$. $\mathcal{G}_* = \mathcal{G}_*(\theta)$ and $\Psi_2^D = \Psi_2^D(\theta)$. Moreover let $\mathbb{E} = \mathbb{E}_\theta$ denote expectation, $\mathrm{Cov} = \mathrm{Cov}_\theta$ covariance, and $\inf_P$, $\sup_P$ the essential extrema, under $P = P_\theta$. Furthermore, in view of the following chapters, we specialize on the one-dimensional case $p = k = 1$.
Following [Rieder (1994)] chapter 5, the standardized bias terms ω_* for $* = c, v$ have the following general properties.

Lemma 2.27. *Let* $* = c, v$ *and* $\eta \in L_1(P)$*. Then*

$$\omega_*(\eta) = \omega_*(\eta - \mathbb{E}\eta) \tag{2.50}$$
$$\omega_c(\eta) \le \omega_v(\eta) \le 2\omega_c(\eta) \tag{2.51}$$

The terms ω_* *are positively homogeneous, subadditive, hence convex on* $L_1(P)$*, and weakly lower semicontinuous on* $L_2(P)$*.*

Proof. [Rieder (1994)], Lemma 5.3.2. □

One gets the following explicit expressions for ω_*.

Proposition 2.28. *Let* $\eta \in L_1(P)$ *with* $\mathbb{E}\eta = 0$*. Then*

$$\omega_c(\eta) = \sup_P |\eta| \tag{2.52}$$
$$\omega_v(\eta) = \sup_P \eta - \inf_P \eta \tag{2.53}$$

Proof. [Rieder (1994)], Proposition 5.3.3 (a). □

2.4.3 Unique Solutions to the Hampel problem

We now give the unique solutions to the Hampel type problems (2.45) in one dimension ($p = k = 1$) and start with the case $* = c$.

Theorem 2.29. (a) *In case* $\omega_c^{\min} < b \leq \omega_c(\eta_h)$*, there exist some* $a \in \mathbb{R}$ *and* $A \in \mathbb{R}$ *such that the solution is of the form*

$$\tilde{\eta} = (A\Lambda - a)w \qquad w = \min\left\{1, \frac{b}{|A\Lambda - a|}\right\} \tag{2.54}$$

Conversely, if some $\tilde{\eta} \in \Psi_2^D$ *is of form (2.54) for any* $b \in (0, \infty)$*,* $a \in \mathbb{R}$*, and* $A \in \mathbb{R}$*, then* $\tilde{\eta}$ *is the solution, and the following representations hold,*

$$a = Az \qquad 0 = \mathbb{E}(\Lambda - z)w \qquad D = A\mathbb{E}(\Lambda - z)^2 w \tag{2.55}$$

where $AD > 0$*.*
(b) *It holds that*

$$\omega_c^{\min} = \min\left\{\frac{\mathbb{E}|A\Lambda - a|}{AD} \;\middle|\; a \in \mathbb{R},\ A \in \mathbb{R} \setminus \{0\}\right\} \tag{2.56}$$

There exist $a,\ A \in \mathbb{R}$ *and* $\bar{\eta} \in \Psi_2^D$ *achieving* $\omega_c^{\min} = b$*, respectively. And then necessarily*

$$\bar{\eta} = b\frac{A\Lambda - a}{|A\Lambda - a|} \qquad \text{on } \{A\Lambda \neq a\} \tag{2.57}$$

Moreover, $a = Az$ *for some* $z \in \mathbb{R}$*, and* $AD \geq 0$*.*
If $\bar{\eta}$ *in addition is constant on* $\{A\Lambda = a\}$*, then it is the solution.*

Proof. [Rieder (1994)], Theorem 5.5.1. □

We add Remark 1.3.8 from [Kohl (2005)], where we state the result of part (b) and omit the proof.

Remark 2.30.

(a) We obtain

$$\omega_c^{min} = \frac{|D|}{\mathbb{E}|\Lambda - m|} \tag{2.58}$$

with any $m = \text{med}(\Lambda)$ *and the solution* $\bar{\eta}$ *reads*

$$\bar{\eta} = \omega_c^{min}\text{sign}(D)\left[\mathbb{I}(\Lambda > m) - \mathbb{I}(\Lambda < m) + \beta\mathbb{I}(\Lambda = m)\right] \tag{2.59}$$

with

$$\beta = \Big[P(\Lambda < m) - P(\Lambda > m)\Big/P(\Lambda = m)\Big] \tag{2.60}$$

where $|\beta| \leq 1$*.*

(b) It is

$$\omega_c(\eta) \geq \frac{1}{\mathcal{I}^{1/2}} \tag{2.61}$$

Now we look at the case $* = v$.

Theorem 2.31. (a) *In case $\omega_v^{\min} < b \le \omega_v(\eta_h)$, there exist some $c \in (-b, 0)$ and $A \in \mathbb{R}$ such that*

$$\tilde{\eta} = c \vee A\Lambda \wedge (c + b) \tag{2.62}$$

is the solution, and

$$\omega_v(\tilde{\eta}) = b \tag{2.63}$$

Conversely, if some $\tilde{\eta} \in \Psi_2^D$ is of form (2.62) for any $b \in (0, \infty)$, $c \in \mathbb{R}$, and $A \in \mathbb{R}$, then $\tilde{\eta}$ is the solution, and the following representations hold,

$$\mathbb{E}(c - A\Lambda)_+ = \mathbb{E}(A\Lambda - (c + b))_+ \qquad D = \mathbb{E}\big[c \vee A\Lambda \wedge (c + b)\big]\Lambda \tag{2.64}$$

(b) *It holds that*

$$\omega_v^{\min} = \min\left\{ \frac{\mathbb{E}(A\Lambda)_+}{AD} \;\middle|\; A \in \mathbb{R} \setminus \{0\} \right\} \tag{2.65}$$

There exist $A \in \mathbb{R}$ and $\bar{\eta} \in \Psi_2^D$ achieving $\omega_v^{\min} = b$, respectively. And then necessarily

$$\bar{\eta}\mathbb{I}(A\Lambda \neq 0) = c\mathbb{I}(A\Lambda < 0) + (c + b)\mathbb{I}(A\Lambda > 0) \tag{2.66}$$

for some $c \in (-b, 0)$. The solution is

$$\bar{\eta} = b \operatorname{sign}(D)\left(\frac{P(\Lambda < 0)}{P(\Lambda \neq 0)}\mathbb{I}(\Lambda > 0) - \frac{P(\Lambda > 0)}{P(\Lambda \neq 0)}\mathbb{I}(\Lambda < 0) \right) \tag{2.67}$$

Proof. [Rieder (1994)], Theorem 5.5.5. □

We add Remark 1.3.10 from [Kohl (2005)], where we again state the result of part (b) and omit the proof.

Remark 2.32.

(a) We obtain

$$\omega_v^{min} = \frac{|D|}{\mathbb{E}(\Lambda - m)_+} \tag{2.68}$$

(b) It is

$$\omega_v(\eta) \ge \frac{1}{\mathcal{I}^{1/2}} \tag{2.69}$$

2.4.4 Unique Solution to the MSE problems

Finally we give the solutions to the MSE problems (2.43).

Theorem 2.33. (a) *The solutions to problem (2.43) for $* = c$ and $(* = v,\ p = 1)$, respectively, are unique.*

(b) *The solution to problem (2.43) and $* = c$ coincides with the solution of problem (2.45) and $* = c$, with $b \in (0, \infty)$ and $r \in (0, \infty)$ related by*

$$r^2 b = \mathbb{E}\big(|A\Lambda - a| - b\big)_+ \tag{2.70}$$

(c) *The solution to problem (2.43) and $(* = v,\ p = 1)$ coincides with the solution of problem (2.45) and $(* = v,\ p = 1)$, with $b \in (0, \infty)$ and $r \in (0, \infty)$ related by*

$$r^2 b = \mathbb{E}\big(c - A\Lambda\big)_+ \tag{2.71}$$

Proof. [Rieder (1994)], Theorem 5.5.7. □

Chapter 3

Motivation

As the main concern of this thesis is the investigation of the behavior of maximal risk on a special kind of neighborhoods, i.e. total variation neighborhoods, we want to lay sufficient emphasis on this subject. Therefore the two mainly used types in Robust Statistics, convex contamination and total variation neighborhoods, are reconsidered and as contrast the neighborhood system generated by the Hellinger distance is discussed. In view of the calculations in chapter 6 we give an appropriate interpretation of the convex contamination and total variation neighborhoods.
After having laid this basis for the main chapter 6 we come to the motivation for this thesis that originates from a result in [Kohl (2005)].
The techniques we use to derive our results are based on exact approximations of the limit distribution. However, contiguity in the sense of convergence in distribution does not implicate contiguity of the risk necessarily. An argument based on the breakdown point illustrates this fact. So in section 3.3 we recall the concept of the finite sample breakdown point and employ a convenient modification of the infinitesimal models that on the one hand is asymptotically negligible, but on the other hand forces the unmodified MSE to converge along with weak convergence.

3.1 Neighborhood systems reconsidered

Robust statistics allows the real distribution to be any member of some suitably full neighborhood of F_θ. According to (2.23) we denote by

$$\mathcal{U}(\theta) = \{U(\theta, r) | r \in [0, \infty)\} \tag{3.1}$$

any system of 'neighborhoods' $U(\theta, r)$ of 'radius' $r \in [0, \infty)$ about F_θ such that

$$F_\theta \in U(\theta, r) \subset \mathcal{M}_1(\mathcal{A}). \tag{3.2}$$

3.1.1 Gross Error Model (Convex Contamination)

As noted in Subsection 1.2c of [Hampel et al. (1986)], 1 - 10% "wrong values" (gross errors, outliers) are typical in routine data. Often, such real data sets are well modeled by the

well-known gross error model (convex contamination)

$$Q = (1 - \varepsilon)F_\theta + \varepsilon H$$

where H is some arbitrary probability measure and $\varepsilon \in [0, 1]$ is the amount of gross errors (contamination); confer [Tukey (1960)]. These neighborhoods are intuitively very appealing, which is the cause why they are used in the majority of treatments confering to contaminated data. Maybe a prize to pay for intuition is a lack of symmetry that inherits the definition. But there is another method of outlier modeling that cannot be explained by the Gross Error Model as it offers a more symmetric approach: total variation.

3.1.2 Total Variation

An alternative way of describing or generating, respectively, "dirty" data is the model of total variation. In robust statistics first mentioned by Huber and Hampel, both 1968, one looks at the maximum distance between two measures F and Q as to write

$$d_v(F, Q) = \sup_{A \in \mathcal{A}} |Q(A) - F(A)| \tag{3.3}$$

and on the real line this is the Kolmogorov distance

$$d_K(F, Q) = \sup |Q(x) - F(x)|. \tag{3.4}$$

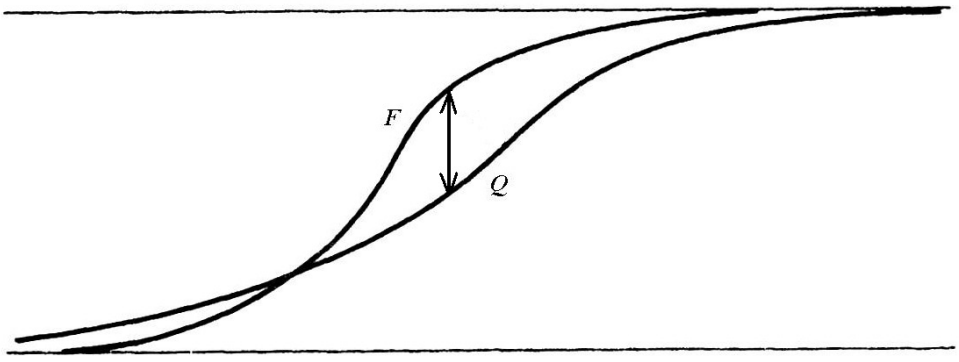

Figure 3.1: Illustration of the Kolmogorov metric d_K.

These distances do not generate the weak topology. But one can think of the Lévy distance (on the real line $\mathbb{R}$) and conceptually attractive Prokhorov distance (on a general polish sample space Ω), which measurizes the weak convergence, as a kind of generalization of the model of total variation: the defining equation for the Prokhorov distance

$$d_\pi(F, Q) = \inf\{\varepsilon | \forall A \in \mathbb{B}, F(A) \leq Q(A^\varepsilon) + \varepsilon\}, \tag{3.5}$$

where $A^\varepsilon = \{y \in \Omega | \inf_{x \in A} d(y, x) \leq \varepsilon\}$ is a closed ε-neighborhood of A, is turned into a definition of the Lévy distance d_λ if we decrease the range of conditions to sets A of the form $(-\infty, x]$ and $[x, \infty)$. It is turned into a definition of the total variation distance d_v

if we replace A^ε by A and thus make the condition harder to fulfill. This again can be converted into a definition of the Kolmogorov distance if we restrict the range of A to sets $(-\infty, x]$ and $[x, \infty)$. As a result it holds[1] that

$$d_\lambda \leq d_\pi \leq d_v \geq d_K. \tag{3.6}$$

The Lévy-metric, which reads most generally

$$d_\lambda(F, Q) = \inf\{\varepsilon | \forall x \; F(x - \varepsilon) - \varepsilon \leq Q(x) \leq F(x + \varepsilon) + \varepsilon\}, \tag{3.7}$$

can be illustrated in a good way as it is done in [Huber (1981)]: $\sqrt{2}d_\lambda(F, Q)$ is the maximum distance between the graphs F and Q, measured along a 45°-direction:

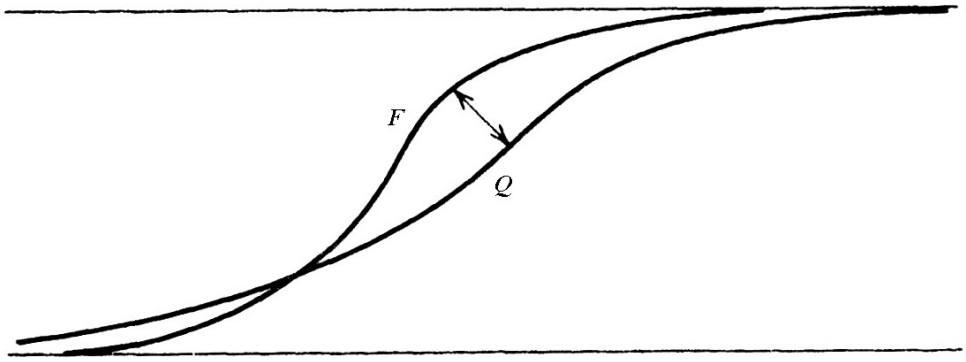

Figure 3.2: Modified exhibit 2.3.1 from [Huber (1981)], illustrating the Lévy metric.

Besides this illustration the Lévy metric unfortunately does not possess an intuitive interpretation in the style of the Prokhorov metric.
The distances d_π and d_v can be seen in the light of Strassen's theorem in [Strassen (1965)], as discusses Huber in section 2.3 of [Huber (1981)]: if Q is the idealized model and F is the true underlying distribution, such that $d_\pi(F, Q) \leq \varepsilon$, i.e. the two laws are close to each other in the Prokhorov metric, then Strassen's theorem shows that we can always assume that there is an ideal (but unobservable) random variable Y with $\mathcal{L}(Y) = Q$, and an observable X with $\mathcal{L}(X) = F$, such that $P\{d(X, Y) \leq \varepsilon\} \geq 1 - \varepsilon$, that is, the Prokhorov distance provides both for small errors occurring with large probability, and for large errors occurring with low probability, in a very explicit, quantitative fashion. The joint distribution (X, Y) will be concentrated near the diagonal $X = Y$, so that X and Y will be far from independent.
We add the exact theorem as stated in [Huber (1981)] for a Polish[2] space Ω. A similar formulation that is proved in a more general context based on a finite combinatorial fact called a pairing theorem, is given as Theorem 11.6.2 in [Dudley (1989)].

Theorem 3.1. (Strassen)
The following two statements are equivalent:

[1] confer Lemma 3.5

[2] A Polish space is a separable completely metrizable topological space. Common examples are the real line and the Cantor space, the topological abstraction of the classical Cantor set.

(1) $F\{A\} \leq Q\{A^\delta\} + \varepsilon$ *for all* $A \in \mathbb{B}$.

(2) There are (dependent) random variables X *and* Y *with values in* Ω*, such that* $\mathcal{L}(X) = F$*,* $\mathcal{L}(Y) = Q$*, and* $P\{d(X,Y) \leq \delta\} \geq 1 - \varepsilon$.

Proof. As $\{X \in A\} \subset \{Y \in A^\delta\} \cup \{d(X,Y) > \delta\}$, (1) is an immediate consequence of (2). The proof of the converse is contained in a famous paper of Strassen, confer [Strassen (1965)] p. 436 ff. □

Remark 3.2. *For the Prokhorov distance* d_π *put* $\delta = \varepsilon$ *in Theorem 3.1.*

Another description for the total variation distance is the L_1-distance of two measures, which allows a more practical application of the induced neighborhood system:

$$d_v(Q,F) = \frac{1}{2}\int |dQ - dF| = \sup_{A\in\mathcal{A}} |Q(A) - F(A)| \tag{3.8}$$

3.1.3 Hellinger

In the sense of exposing the convex contamination and the total variation neighborhoods as our favorite pair of opponent models we spend some words on another kind of neighborhood system using the Hellinger distance

$$d_h^2(Q,F_\theta) = \frac{1}{2}\int |\sqrt{dQ} - \sqrt{dF_\theta}|^2. \tag{3.9}$$

In contrast to the convex contamination and the total variation neighborhoods the Hellinger balls are two small to be characterized by capacities as it is done by [Huber (1981)], for example. The following example was given by L. Birgé and can be found in [Huber-Carol (1986)], p. 108:

Theorem 3.3. *Let* $P = \lambda|_{[0,1]}$ *and* $\mathcal{B} = \mathbb{B}|_{[0,1]}$ *and define*

$$Q_r(dx) := \frac{r^2}{2}\delta_{\{0\}}(dx) + \left[(1 + \frac{r}{\sqrt{2x}})\mathbb{I}_{(0;1/2]}(x) + (1 - 2r - r^2)\mathbb{I}_{(1/2;1]}(x)\right]\lambda(dx). \tag{3.10}$$

Then for $r < \frac{1}{3}$ *it holds that*

$$Q_r(B) \leq w_h(B) \quad \forall\ B \in \mathcal{B}, \tag{3.11}$$

but

$$d_h(Q_r, P) > r. \tag{3.12}$$

Proof. [Huber-Carol (1986)], pp. 108 □

Remark 3.4. *a) Simply speaking the cause for the paradox in Theorem 3.3 is the use of the square root in the definition of the Hellinger distance* d_h *on the one hand and in the definition of* Q_r *on the other hand. In the latter case* Q_r *results to be a probability measure, as the hyperbola* $1/\sqrt{x}$ *falls fast enough to be measureable in the area* $(0, 1/2]$*. Contrary, the square root used in the distance definition pushes the values of* $Q_r(dx)$ *near zero much higher than they already are and so extends the integral to a final value bigger than* r.

b) [Bickel (1981)], pp. 36-38, shows in Theorem 8 (ibid.) that the Hellinger neighborhoods are too small to allow for identifiability in the sense of [Hoeffding (1956)] and [Hoeffding and Wolfowitz (1958)] at shrinking rate $1/\sqrt{n}$.

c) [Ruckdeschel (2005c)] constructs the detailed minimax test for equal vs. higher outlier probability and shows that there is no problem of decision with shrinking contamination and total variation balls of rate $1/\sqrt{n}$ indeed, but considering general probabilities of exact Hellinger distance r_n to the ideal measure F, a shrinking factor of $1/\sqrt[4]{n}$ must be required and in this case the bias will dominate variance eventually. This means that considering a decision criterium as the MSE we have to standardize by $\sqrt[4]{n}$ instead of $\sqrt{n}$ and get the same optimality theory results if Hellinger bias is considered alone (that is, without variance). Nevertheless this leads to the same optimality theory, i.e. the classically optimal scores, as in the corresponding $1/\sqrt{n}$ setup that can be justified by summarizing the neighborhoods to their upper probability.

It holds the following hierarchy of metrics and balls:

Lemma 3.5.

$$B_c \subset B_v, \qquad d_h^2 \le d_v \le \sqrt{2} d_h, \qquad d_K \le d_v \ge d_\pi \ge d_\lambda. \tag{3.13}$$

Proof. [Rieder (1994)], Lemma 4.2.8. □

3.1.4 Interpretation of the neighborhoods

For our purpose we consider two types of infinitesimal neighborhood systems $\mathcal{U}_*(\theta) = \{U(\theta, r) | r \in [0, \infty[\}$: contamination ($* = c$) and total variation ($* = v$). The system $\mathcal{U}_*(\theta)$ then consists of closed balls $B_*(F, r)$, $r \in [0, \infty[$, about F. Setting $r_n = \frac{r}{\sqrt{n}}$, $r > 0$, we can derive both neighborhood systems from the following set $\mathcal{Q}_n^{(*)}(r)$:

$$\begin{aligned} \mathcal{Q}_n^{(*)}(r) &:= U_*(\theta, r_n^{(c)}, r_n^{(v)}) && (3.14)\\ &= \{Q_n \in \mathcal{M}_1(\mathbb{B}) | Q_n(dy) \ge (1 - r_n^{(c)})_+ F(dy) - (r_n^{(v)} \wedge 1)\} && (3.15) \end{aligned}$$

(* = c):

By speaking of shrinking contamination neighborhoods we define the set $\mathcal{Q}_n^{(c)}(r) := B_c(F, r_n) = U_c(\theta, r_n) = \{Q_n \in \mathcal{M}_1(\mathbb{B}) | Q_n(dy) \ge (1 - r_n)_+ F(dy)\}$ of distributions

$$Q_n(r) = \mathcal{L}_\theta^{\mathrm{re}}(X_1, \ldots, X_n) = \bigotimes_{i=1}^{n} [(1 - r_n)_+ F + (r_n \wedge 1)\, P_{n,i}^{\mathrm{di}}] \tag{3.16}$$

with $r_n = \frac{r}{\sqrt{n}}$, $r > 0$ the contamination radius and $P_{n,i}^{\mathrm{di}} \in \mathcal{M}_1(\mathbb{B})$ arbitrary, uncontrollable contaminating distributions.

We may interpret Q_n as the distribution of the vector $(X_i)_{i \le n}$ with components

$$X_i := (1 - U_i)X_i^{\mathrm{id}} + U_i X_i^{\mathrm{di}}, \qquad i = 1, \ldots, n \tag{3.17}$$

for X_i^{id}, U_i, X_i^{di} stochastically independent, $X_i^{\mathrm{id}} \overset{\text{i.i.d.}}{\sim} F$, $U_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bin}(1, r/\sqrt{n})$, and $X_i^{\mathrm{di}} \sim P^{\mathrm{di}}$ for some arbitrary $P^{\mathrm{di}} \in \mathcal{M}_1(\mathbb{B})$.

(* = v):

In the context of total variation neighborhoods we have sequences of shrinking balls $\mathcal{Q}_n^{(v)}(r) := B_v(F, r_n) = U_v(\theta, r_n) = \{Q_n \in \mathcal{M}_1(\mathbb{B}) | Q_n(dy) \geq F(dy) - (r_n \wedge 1)\}$ about F with radius $r_n = \frac{r}{\sqrt{n}}$, sample size n, given by the metric

$$d_v(Q_n, F) = \frac{1}{2} \int |dQ - dF| = \sup_A |Q(A) - F(A)| \leq r_n \wedge 1$$

with $Q_n = \mathcal{L}_\theta^{\mathrm{re}}(X_1, \ldots, X_n)$.

By confining ourselves to the model of simple perturbations introduced in (2.30) and (2.35), respectively, we interpret $Q_n = \otimes_{i=1}^n Q_{n,i}$ as generated by some density $1 + r_n q_i$ for some tangent $q_i \in \mathcal{G}_v(\theta)$:

$$dQ_{n,i} = (1 + r_n q_i)dF = dF + r_n q_i dF = dF + r_n d\Delta_i \tag{3.18}$$

and

$$Q_n = \bigotimes_{i=1}^n Q_{n,i} = \bigotimes_{i=1}^n (F + r_n \Delta_i) = P_n^{\mathrm{id}} + r_n P_n^{\mathrm{di}} \tag{3.19}$$

Remark 3.6. *a) By Lemma 3.5 it holds that* $\mathcal{Q}_n^{(c)}(r) = B_c(F, r_n) \subset B_v(F, r_n) = \mathcal{Q}_n^{(v)}(r)$.

b) Relation (2.51) offers a decomposition of its RHS:

$$\begin{aligned} & \omega_v(\eta) \leq 2\omega_c(\eta) =: \omega_{c,1}(\eta) + \omega_{c,2}(\eta) \\ \Rightarrow \quad & \max \omega_v(\eta) = \omega_{c,1}(\eta) + \omega_{c,2}(\eta) \end{aligned} \tag{3.20}$$

This means, choosing a least favorable deviation within a total variation neighborhood, it can asymptotically always be expressed by two convex contaminated balls.

3.2 Conjecture out of M. Kohl's and P. Ruckdeschel's work

In the context of determining the exact finite sample risk (different to the asymptotic MSE) for sample size $n \geq 3$ M. Kohl used Edgeworth expansions to compute an approximation as it seems to be impossible to achieve the expected results analytically (conf. section 11.3.3 "Higher Order Approximations" of [Kohl (2005)]). We first prefix some notation used in [Kohl (2005)]:

Notation 3.7. *We fix $n \in \mathbb{N}$ and radius $r_n^c \in [0,1)$, respectively radius $r_n^v \in [0,1)$. Given some clipping bound $b \in (0,1)$, we then want to determine the finite-sample risk of an M estimator S satisfying*

$$\sum_{i=1}^{n} \Lambda_0(y_i - S) = 0 \qquad \Lambda_0(u) = u \min\{1, \frac{b}{|u|}\}.$$

with equal randomization between the smallest and the largest solutions.
In case of contamination neighborhoods and a given number (width) $\tau_n = \tau/\sqrt{n} \in (0,\infty)$ we get $Q'_{-\tau_n} \in U_c(-\tau_n)$ and $Q''_{\tau_n} \in U_c(\tau_n)$ with

$$Q'_{-\tau_n} = (1 - r_n^c)\mathcal{N}(-\tau_n, 1) + r_n^c H'_{-\tau_n}$$

and

$$Q''_{\tau_n} = (1 - r_n^c)\mathcal{N}(\tau_n, 1) + r_n^c H''_{\tau_n}$$

where $H'_{-\tau_n}$ and H''_{τ_n} are concentrated on $[\tau_n + b, \infty)$ and $(-\infty, -\tau_n - b]$, respectively. In case of total variation neighborhoods this leads us to $Q'_{-\tau_n} \in U_v(-\tau_n)$ and $Q''_{\tau_n} \in U_v(\tau_n)$ having cumulative distribution functions (confer also [Rieder (1994)], pp.174)

$$Q'_{-\tau_n}(t) = (\Phi(t + \tau_n) - r_n^v)_+ + r_n^v H'_{-\tau_n}(t) \tag{3.21}$$

and

$$Q''_{\tau_n}(t) = \min\{[\Phi(t - \tau_n) + r_n^v H''_{\tau_n}(t)], 1\} \tag{3.22}$$

for all $t \in \mathbb{R}$ where $H'_{-\tau_n}$ and H''_{τ_n} are concentrated on $[\tau_n + b, \infty)$ and $(-\infty, -\tau_n - b]$, respectively. That is, in case of $Q'_{-\tau_n}$ mass r_n^v is moved from the left tail to $[\tau_n + b, \infty)$, whereas in case of Q''_{τ_n} it is moved from the right tail to $(-\infty, -\tau_n - b]$.

Instead of spelling out the important Remarks 11.2.4, 11.3.7 and 11.3.8 (a) of [Kohl (2005)] literally, we give a summary of these in the next remark in order to show the motivation for our investigations:

Remark 3.8. *It is shown in section 11.2 of [Kohl (2005)] that there is a clear difference between contamination and total variation neighborhoods concerning the speed of convergence of the optimal clipping bounds (**$\mathbf{O(n^{-1/2})}$ vs. $\mathbf{O(n^{-1})}$**). Apparently, the speed of convergence towards the asymptotic risk is faster by an order in case of total variation neighborhoods. M. Kohl conjectures that this is caused by the higher symmetry of total variation neighborhoods, which also shows up by calculating the Edgeworth expansions in section 11.3 (ibid): if k is even, calculating $E_R \Lambda_0^k$ ($k \in \mathbb{N}$) ($R = Q'_{-\tau_n}, Q''_{\tau_n}$ absolutely continuous) for total variation neighborhoods gives $E_{Q'_{-\tau_n}} \Lambda_0^k = \int \Lambda_0^k d\mathcal{N}(-\tau_n, 1)$ and $E_{Q''_{\tau_n}} \Lambda_0^k = \int \Lambda_0^k d\mathcal{N}(\tau_n, 1)$, respectively.*
Furthermore, it holds for the bias that $b_c^{finite} = b_c^{asympt} + O(n^{-1/2})$ whereas $b_v^{finite} = b_v^{asympt} + O(n^{-1})$. All results are confirmed by numerical results in section 11.4 (ibid), also showing that the same holds for the corresponding finite-sample risks of the finite-sample and also the asymptotic minimax estimator.

Based on these insights on the higher asymptotics on total variation neighborhoods the conjecture is that in this case the risk reads as

$$\sup_{Q_n \in \tilde{\mathcal{Q}}_n(r)} n\,\mathbf{MSE}(S_n, Q_n) = r^2 b^2 + \mathbb{E}\psi^2 + \frac{1}{n} A_2 + o\left(\frac{1}{n}\right) \tag{3.23}$$

which would indicate a faster rate of convergence. But the reason for the vanishing of the $n^{-1/2}$-term could as well be found in the symmetry of F, i.e. $f(x) = f(-x)$, which is used by M. Kohl throughout his investigations as there is $F_\theta = \mathcal{N}(\theta, 1)$. In the case of convex contamination this symmetry condition indicates no vanishing of the $n^{-1/2}$-term, however. It only gets some easier, algebraically speaking (conf. [Ruckdeschel (2005b)] Remark 3.2 and Remark 3.4).

3.3 Finite Sample Breakdown Point

Concerning the finite sample breakdown point we work with the definition of [Donoho and Huber (1983)], p. 161. As therein the definition for the finite sample breakdown point ε_0 is not restricted for contamination neighborhoods it is applicable for total variation neighborhood systems, too.
Anticipating the mechanism of modification in section 8.3 we use the concept of "ε-replacement" on p. 160 (ibid.) for the description of the actual impact of a modification of a sample $X = (x_1, \ldots, x_n)$ by means of total variation:

> ε-replacement: we replace an arbitrary subset of size m of the sample by arbitrary values $y_1, \ldots, y_m$. The fraction of bad values in the corrupted sample X' is $\varepsilon = m/n$.

Then we define the finite sample breakdown point ε_0 as done in [Donoho and Huber (1983)]:

Definition 3.9. *Let $T = \{T_n\}_{n=1,2,\ldots}$ be an estimator with values in the Euclidean space, and $T(X)$ its value at the sample X. Then we define the breakdown point as*

$$\varepsilon_0(X, T) = \inf\{\varepsilon : \ b(\varepsilon; X, T) = \infty\} \tag{3.24}$$

with

$$b(\varepsilon; X, T) = \sup |T(X') - T(X)|$$

the maximum bias that can be caused by ε-corruption.

Simply speaking, the (replacement) breakdown point of T at X is the smallest fraction of the sample for which the estimator, when applied to the ε-corrupted sample X' can take values arbitrarily far from $T(X)$.

Based on this concept we join [Ruckdeschel (2005a)] or [Ruckdeschel (2005b)], respectively, and [Kohl (2005)] by employing an asymptotically negligible modification of the infinitesimal models (3.16) and (3.19): For sample length n and K the number of contaminated or modified observations, we exclude all samples with $K > n/2$. Let

$$K_n := \Big\{ K \leq \frac{n}{2} \Big\} \tag{3.25}$$

then

$(* = c)$ we look at the conditional neighborhoods

$$Q_n(\ .\ |K_n) = \Big\{ \mathcal{L}\left([(1-U_i)X_i + U_i Y_i]_{i=1,...,n} | K = \sum U_i \leq \frac{n}{2} \right) \Big\} \tag{3.26}$$

with random variables $U_1, ..., U_n \overset{\text{i.i.d.}}{\sim} \text{Bin}(1, r/\sqrt{n})$, $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} F, Y_1, ..., Y_n \overset{\text{i.i.d.}}{\sim} H \in \mathcal{M}_1(\mathbb{B})$ and all random variables stochastically independent.

$(* = v)$ we look at the conditional neighborhoods $Q_n(\ .\ |K_n)$ with random variable $K \overset{\text{i.i.d.}}{\sim} P$ such that

$$\mathbb{E}_P K = r\sqrt{n} \tag{3.27}$$

Definition 3.10. *With the conditional neighborhoods*

$$Q_n := Q_n(\ .\ |K_n) \tag{3.28}$$

for $(* = c)$ and $(* = v)$, respectively, we subsequently employ as standard neighborhood systems the (slightly) thinned out balls

$$\tilde{\mathcal{Q}}_n^{(*)}(\ .\ ) = \{ Q_n(\ .\ |K_n) \} \tag{3.29}$$

Remark 3.11. *a) The modification was motivated by a closer inspection of simulations done by M. Kohl, who found out that larger inaccuracies of (first order) asymptotics only occurred when there were extraneous sample situations where more than half the sample size stemmed from a contamination. This instance led him to the conjecture that excluding such samples, asymptotics might then prove useful even for very small samples.*

b) In section 2.3 of [Ruckdeschel (2005a)] or [Ruckdeschel (2005b)], respectively, it is shown by the Hoeffding bound (A.2) that the above modification is asymptotically negligible, as $P(K > n/2)$ decays exponentially fast. Furthermore, it enforces the unmodified MSE, i.e. without clipping, to converge along with weak convergence, confer Proposition 2.2 in section 2.4 of [Ruckdeschel (2005a)]. This is not self-evident, as weak convergence in general is too weak to entail convergence of the risks; the standard way out in Asymptotic Statistics is to clip the unbounded loss function. The clipping of unbounded loss functions is commonly used in asymptotic statistics in order to attain the lower bound in asymptotic minimax theorems, confer Proposition 2.23, for instance. In this context we refer to [Le Cam (1986)], [Rieder (1994)], [Bickel et al. (1998)] or [van der Vaart (1998)], respectively.

c) In Assumption 8.21 of section 8.9 we arrive at condition (PK) on the distribution of K:

$$(PK) \qquad P(K \in [r\sqrt{n}(1-\eta), r\sqrt{n}(1+\eta)] = 1 - O(e^{-n^\delta}) \quad \textit{for some} \quad \delta,\ \eta > 0$$

Obviously, condition (PK) implies that no more than 50% *of the sample is modified. Therefore we could get rid of the modification (3.28). But as the main results in chapter 6 are available with the "weaker" restriction, already, we stay with it in the meantime.*

d) In the sequel we suppress the conditioning w.r.t. K_n *and write* Q_n *meaning* $Q_n = Q_n(\ .\ |K_n)$ *and* K_n *as defined in* (3.25), (3.26) *and* (3.27) *or according to (PK), respectively.*

Chapter 4

First Order Optimality for Robust Estimation of Location in one dimension

As explicit and manageable bias terms for total variation are only available for one dimension, we briefly summarize the results of chapter 2.4 as far as one-dimensional location, MSE and neighborhoods of type $(* = c, v)$ are concerned. We give the first order optimality result to show that under symmetry of F there is no possibility to see any differences between the convex contamination and the total variation case. We add Huber's monotony approach for M-estimators that turns out to be useful for the location but not for the scale model, for example. In this case, an alternative approach by k-step-estimators is presented, too.

4.1 Optimal Influence Curves for one dimension

For a sequence of estimators S_n we consider the asymptotic (modified) maximal mean squared error on $\tilde{\mathcal{Q}}_n$

$$\tilde{R}(S_n, r) := \lim_{M\to\infty} \lim_{n\to\infty} \sup_{Q_n \in \tilde{\mathcal{Q}}_n(r)} \int \min\{M,\, n\,|S_n - \theta|^2\}\, dQ_n \tag{4.1}$$

As summed up in chapter 2 it is shown in [Rieder (1994)] that with scores Λ and Fisher-Information $\mathcal{I}$ a (suitably constructed) asymptotically linear estimator S_n with IC ψ has risk

$$(* = c) \qquad \tilde{R}(S_n, r) = r^2 \sup|\psi|^2 + \mathbb{E}|\psi|^2 \tag{4.2}$$

$$(* = v) \quad \tilde{R}(S_n, r) = r^2(\sup\psi - \inf\psi)^2 + \mathbb{E}|\psi|^2 \tag{4.3}$$

with the expectations evaluated under the law F.

In one dimension $k = p = 1$, for given $r \geq 0$, among all such ALEs, any (suitably constructed) ALE with IC η minimizes $\tilde{R}(\,\cdot\,, r)$ where η is

$(* = c)$: of Hampel form

$$\eta = Y \min\{1, b/|Y|\}, \qquad Y = A\Lambda - a \tag{4.4}$$

for some $A \in \mathbb{R}$ and $a \in \mathbb{R}$ such that η is an IC, i.e.

$$\mathbb{E}\eta = 0, \qquad \mathbb{E}\eta\Lambda = 1, \tag{4.5}$$

and b solving

$$r^2 b = \mathbb{E}(|Y| - b)_+. \tag{4.6}$$

$(* = v)$: of form

$$\eta = c \vee A\Lambda \wedge (c + b) \tag{4.7}$$

for some $A \in \mathbb{R}\backslash\{0\}$ and numbers $c \in (-b, 0)$, $b \in (0, \infty)$, such that η is an IC, i.e.

$$\mathbb{E}\eta = 0, \qquad \mathbb{E}\eta\Lambda = 1, \tag{4.8}$$

and

$$r^2 b = \mathbb{E}(c - A\Lambda)_+. \tag{4.9}$$

Remark 4.1. *The risks* (4.2) *and* (4.3) *only are first-order asymptotic solutions for optimal ALEs w.r.t. the MSE. Up to now it is not clear to which degree the asymptotic optimality carries over to finite sample size. Especially the influence of the radius r, the sample size n and the clipping height b is not visible. Therefore in chapter 6 we investigate the higher order asymptotics for the MSE in the one-dimensional location model. Section 6.1 summarizes briefly the result for the convex contaminated case as it was worked out in [Ruckdeschel (2005b)]. Section 6.2 then delivers the result for infinitesimal total variation neighborhood systems.*

4.2 The one-dimensional location model

We consider the one–dimensional location model, i.e.

$$x_i = y_i + \theta, \qquad y_i \overset{\text{i.i.d.}}{\sim} F \tag{4.10}$$

for some ideal distribution F with finite Fisher-Information of location $\mathcal{I}(F)$, i.e.

$$\Lambda_f = -\dot{f}/f \in L_2(F), \qquad \mathcal{I}(F) = \mathbb{E}[\Lambda_f^2] < \infty \tag{4.11}$$

4.2.1 Illustration for $F = \mathcal{N}(0,1)$

In the case of $y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_y^2)$, scale $\sigma_y \in]0,\infty[$ known, the scores function reads $\Lambda_\theta(x) = \sigma_y^{-2}(x-\theta)$ and $\mathcal{I}_\theta = \sigma_y^{-2}$. By translation equivariance, we may restrict ourselves to $\theta = 0$ which will be suppressed in the notation. With $\theta = 0$ we have

$$\Lambda_0(x) = \sigma_y^{-2}x, \qquad \mathcal{I} = \sigma_y^{-2},$$

and in the simplest case with $\sigma_y = 1$ we get

$$\Lambda_0(x) = x, \qquad \mathcal{I} = 1. \tag{4.12}$$

$(* = c)$:

The equation for the optimal robust IC now is

$$\eta = (Ax - a)\min\{1, \frac{b}{|Ax-a|}\} \tag{4.13}$$

for some $A,\ a \in \mathbb{R}$, $b \in (0,\infty)$ satisfying conditions (4.5) and (4.6).

If F is symmetric ($F = \mathcal{N}(0,\sigma_y^2)$, for example), i.e. the scores Λ_f is odd, $\Lambda_f(-x) = -\Lambda_f(x)$, then it holds that $\mathbb{E}\eta = 0$ if we assume $a = 0$. In this sense we cut (4.13) down to

$$\eta = Ax\min\{1, \frac{c}{|x|}\} = -b \vee Ax \wedge b \tag{4.14}$$

where $c := \frac{b}{|A|}$. With $\mathbb{E}\eta = 0$ the remaining conditions are

$$\begin{aligned} 1 &= A\mathbb{E}|x|\min\{|x|, c\} && (4.15)\\ r^2c &= \mathbb{E}(|x| - c)_+ && (4.16) \end{aligned}$$

$(* = v)$:

Equations (4.7) to (4.9) for the optimal robust IC specialize (with $\theta = 0$) to

$$\eta = c \vee Ax \wedge (c + b) = A \cdot \{g \vee x \wedge (g + c')\} \tag{4.17}$$

for some $A \in \mathbb{R}$, $g = \frac{c}{A}$, $c' = \frac{b}{A}$ and

$$\begin{aligned} 0 &= \mathbb{E}(g - x)_+ - \mathbb{E}(x - g - c')_+ && (4.18)\\ 1 &= A\mathbb{E}x\{g \vee x \wedge (g + c')\} && (4.19)\\ r^2c' &= \mathbb{E}(g - x)_+ && (4.20) \end{aligned}$$

Of course, for $F = \mathcal{N}(0,1)$ the IC is symmetric, too. So we have $c = -b/2$ and get

$$\eta = -\frac{b}{2} \vee Ax \wedge \frac{b}{2} \tag{4.21}$$

which corresponds to (4.14) in the case ($* = c$) except for the clipping heights being half the size. But according to (2.52) and (2.53) or (4.2) and (4.3), respectively, the evaluated bias-terms coincide and so does the MSE.

Remark 4.2. *The optimal robust ICs can be computed in* R *with the packages* ROptEst *and* RobLox*, both developed by M. Kohl.*

4.3 Approach by M- and k-step-estimators

4.3.1 Location

As estimators to achieve (2.20) in definition 2.13 for a given IC ψ, we consider M-estimators (or Z-estimators for *zero*). The name "M-estimator", first mentioned in [Huber (1964)], comes from "generalized *m*aximum likelihood", but as [van der Vaart (1998)] p. 41 points out w.r.t. the way we usually compute these maxima, respectively to cover a broader class of estimators, the name Z-estimator (for zero) is probably a better choice as one is interested in the root of an equation. We shortly describe the general case of an M-/Z-estimator as it is introduced in subsection 2.3a of [Hampel et al. (1986)] and refer to [van der Vaart (1998)] Chapter 5 for more detailed information about this concept of estimators.

Suppose we have one-dimensional independent and identically distributed observations $X_1, \ldots, X_n$ belonging to some sample space $\mathcal{X} \subseteq \mathbb{R}$. We look at the parametric model of a family of probability distributions F_θ on the sample space, where the unknown parameter θ belongs to some parameter space Θ. The well-known maximum likelihood estimator (MLE) is defined as the value $T_n = T_n(X_1, \ldots, X_n)$ which maximizes $\theta \mapsto \prod_{i=1}^n f_\theta(X_i)$, or equivalently by

$$\sum_{i=1}^n [-\log f_\theta(X_i)] = \min_\theta ! \tag{4.22}$$

Huber proposed to generalize this to

$$\sum_{i=1}^n \rho(X_i, \theta) = \min_\theta ! \tag{4.23}$$

where ρ is some function on $\mathcal{X} \times \Theta$. Suppose that ρ has a derivative $\psi(x, \theta)$ so the estimate T_n satisfies the implicit equation

$$\sum_{i=1}^n \psi(X_i, T_n) = 0. \tag{4.24}$$

Definition 4.3. *Any estimator defined by* (4.23) *or* (4.24) *is called an M-estimator.*

Remark 4.4. *If G_n is the empirical c.d.f. generated by the sample, then the solution T_n of* (4.24) *can also be written as $T_n(G_n)$, where T is the functional given by*

$$\int \psi(x, T(G))dG(x) = 0 \tag{4.25}$$

for all distributions G for which the integral is defined.

More specifically we require ψ to be monotone and bounded and write $\psi_t(\,\cdot\,)$ for $\psi(\,\cdot\, - t)$ in the location case.

Following the notation in [Huber (1981)], pp.45,let

$$S_n^* := \sup\Big\{ t \,|\, \sum_{i\le n} \psi_t(x_i) > 0\Big\}, \qquad S_n^{**} := \inf\Big\{ t \,|\, \sum_{i\le n} \psi_t(x_i) < 0\Big\} \tag{4.26}$$

and S_n be any estimator satisfying $S_n^* \le S_n \le S_n^{**}$. By monotonicity of ψ, we get

$$P\{S_n^* < t\} = P\Big\{ \sum_{i\le n} \psi_t(x_i) \le 0\Big\}, \qquad P\{S_n^{**} < t\} = P\Big\{ \sum_{i\le n} \psi_t(x_i) < 0\Big\} \tag{4.27}$$

in the continuity points t of the LHS.

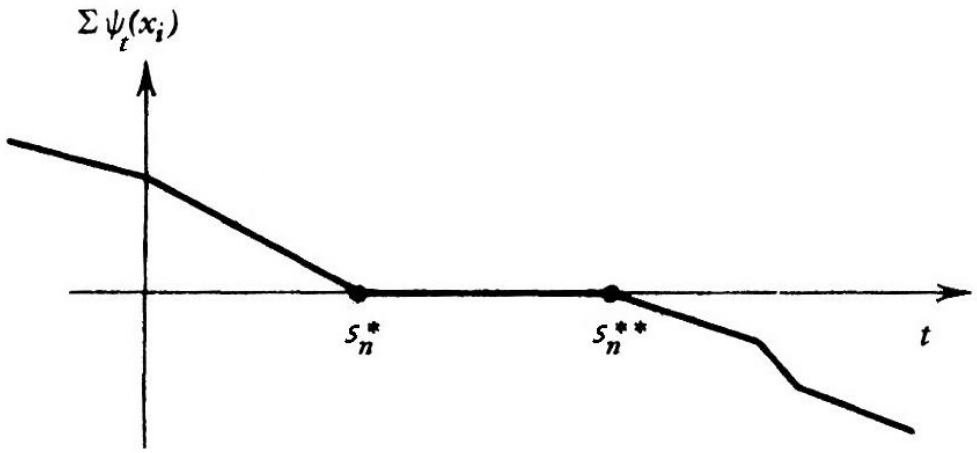

Figure 4.1: Modified Exhibit 3.2.1 from [Huber (1981)].

The next lemma appears together with preceding remarks as Lemma 1.1 in [Ruckdeschel (2005b)] and shows that we may ignore the event $S_n^* \neq S_n^{**}$ if we are interested in statements valid up to $o(1/n)$.

Lemma 4.5. *Let*

$$p_D := \sup_t P(D_t) < 1 \tag{4.28}$$

then

$$P(S_n^* \neq S_n^{**}) = O(\exp(-\gamma n)) \quad \textit{for some} \quad \gamma > 0$$

Proof. Immediate consequence of Theorem 2.3 in [Hall (1992)]. □

We add Remark 1.2 of [Ruckdeschel (2005b)] as part b) in the following remark.

Remark 4.6. *a) Assumption* (4.28) *says that the set D_t of discontinuities of the c.d.f. of $\psi_t(X)$ has to carry less mass than* 1 *uniformly and is motivated by technical reasons in chapter chapter 6, where we assume that the law of $\psi_t(X^{\mathrm{id}})$ has non-trivial absolutely continuous component uniformly in t — compare condition (C) in Assumption 6.3 or 6.9, respectively.*

b) If $\bigcup_t D_t = \{\pm c\}$ for some $c > 0$, $P(S_n^* \neq S_n^{**}) = 0$ for n odd.

4.3.2 Scale

We take a short look at the one–dimensional scale model, i.e.

$$x_i = \theta \cdot y_i, \quad \theta \in]0, \infty[, \quad y_i \overset{\text{i.i.d.}}{\sim} F \tag{4.29}$$

for some ideal distribution F with finite Fisher-Information $\mathcal{I}(F)$. We assume $dF = f d\lambda$ and f absolute continuous. For $P_\theta = \mathcal{L}(\theta y) = \mathcal{L}(x)$ we get $P(x \le t) = P(y \le \frac{t}{\theta}) = F(t/\theta)$, hence $dF(t/\theta) = \frac{1}{\theta} f(t/\theta) dt$. With $\Lambda_f = -\dot{f}/f \in L_2(F)$ we get

$$\begin{aligned} \Lambda_\theta(t) &= \log(dF(t/\theta)) = -\log\theta + \log f\,(t/\theta) = -\frac{1}{\theta} + \frac{f'}{f}\,(t/\theta) \cdot \left(-\frac{t}{\theta^2}\right) \\ &= \frac{1}{\theta}\left[\frac{t}{\theta}\Lambda_f\,(t/\theta) - 1\right] \end{aligned}$$

Therefore it always holds that

$$\Lambda_\theta(x) = \frac{1}{\theta}\Lambda_1\,(x/\theta) \tag{4.30}$$

and

$$\mathcal{I}_\theta = \mathbb{E}_\theta(\Lambda_\theta)^2 = \frac{1}{\theta^2}\mathbb{E}_\theta\,(\Lambda_1\,(x/\theta))^2 = \frac{1}{\theta^2}\mathbb{E}_1(\Lambda_1(x))^2 = \frac{1}{\theta^2}\mathcal{I}_1. \tag{4.31}$$

Then the classical IC $\eta_\theta(x) = \mathcal{I}_\theta^{-1}\Lambda_\theta(x)$ always satisfies the equivariance

$$\eta_\theta(x) = \theta\eta_1\,(x/\theta)\,. \tag{4.32}$$

Illustration for $F = \mathcal{N}(0,1)$

In the case of $F = \mathcal{N}(0,1)$ we obtain

$$\mathcal{I}_\theta = 2\theta^{-2}. \tag{4.33}$$

By $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma)$ we get:

$$\sigma\Lambda_{\mu,\sigma}(x) = (X-\mu)^2\sigma^2 - 1 \tag{4.34}$$

We set $\mu = 0$ and $\sigma = \theta$, which leads to $\theta\Lambda_\theta(x) = \theta^{-2}x^2 - 1$. With $\theta = 1$ by equivariance this reads

$$\Lambda_1(x) = x^2 - 1, \qquad \mathcal{I} = 2.$$

$(* = c)$:

The equation for the optimal robust IC now is

$$\eta = [A(x^2-1) - a]\min\left\{1, \frac{b}{|A(x^2-1)-a|}\right\} = A(x^2-\alpha)\min\left\{1, \frac{c}{|x^2-\alpha|}\right\} \tag{4.35}$$

for some $A, a \in \mathbb{R}$, $b \in (0,\infty)$ and $c := \frac{b}{|A|}$, α a centering constant (in general $\alpha \in]0,1]$, conf. [Rieder et al. (2001)], p. 14). Again η has to satisfy the following conditions, where the first one is used to abbreviate the second one:

$$\begin{aligned} 0 &= \mathbb{E}\eta = \mathbb{E}(x^2-\alpha)\min\{1, \frac{c}{|x^2-\alpha|}\} && (4.36)\\ 1 &= A\mathbb{E}|x^2-1|\min\{|x^2-\alpha|, c\} && (4.37)\\ r^2c &= \mathbb{E}(|x^2-\alpha|-c)_+ && (4.38) \end{aligned}$$

$(* = v)$:

Equations (4.7) to (4.9) for the optimum robust IC specialize (with $\theta = 1$) to

$$\eta = c \vee A(x^2-1) \wedge (c+b) = A\cdot\{[g \vee x^2 \wedge (g+c')] - 1\} \tag{4.39}$$

for some $A \in \mathbb{R}$, $g = \frac{c}{A}$, $c' = \frac{b}{A}$ and

$$\begin{aligned} 0 &= \mathbb{E}(g-x^2)_+ - \mathbb{E}(x^2-g-c')_+ && (4.40)\\ 1 &= A\mathbb{E}x^2\{[g \vee x^2 \wedge (g+c')] - 1\} && (4.41)\\ r^2c' &= \mathbb{E}(g-x^2)_+ && (4.42) \end{aligned}$$

Approach by k-step estimators

We write $\psi_t(\,\cdot\,)$ for $\psi(\frac{\cdot}{t})$. In the scale model a monotone IC cannot be expected and for general F there will also be no symmetry for ψ, too. Of course, if one is willing to settle on the symmetric case with $F = \mathcal{N}(0,\sigma^2)$, for example, the IC could be cut into monotone pieces. The results belonging to these parts perhaps could be united later on by symmetry arguments. But proceeding in this sense requires a modification of (4.27) in

an appropriate way to fit the scale model. The characteristics of S_n have to be interpreted by the characteristics of ψ_t. Up to now this question is open.
Alternatively we catch up P. Ruckdeschel's suggestion to think of an approach local to a $\sqrt{n}$–consistent starting estimator θ_n^0 by use of the implicit function theorem and then define a k-step-estimator

$$\theta_n^{(k)} := \theta_n^{(k-1)} + \frac{1}{n}\sum_{i=1}^{n} \eta_{\theta_n^{k-1}}(X_i). \tag{4.43}$$

Proceeding this way we avoid complications concerning monotonicity or symmetry. But we do not pursue this further within this thesis.

We add two more remarks related to [Kohl (2005)] Remark 2.3.2 (b) on the connection between M- and k- or one-step-estimators, respectively.

Remark 4.7. *a) In practice one-step estimators have a clear advantage: Given some strict and $\sqrt{n}$ consistent starting estimator the one-step estimator is very fast to compute and additionally unique. Estimates derived from M equations, however, besides being more difficult to determine, need not be unique; confer [Reeds (1985)], for instance.*

b) In case of robust estimators with Hampel-type influence curves, the higher order asymptotics of [Ruckdeschel (2005b)] and [Ruckdeschel (2005d)] for the MSE show that in the case of normal location the M estimators and the one-step estimators are asymptotically equivalent up to second order. Without symmetry this is only true for the two-step estimator.

Chapter 5

A first simulation study

Before stating the theoretical results of the thesis in the next chapters, we summarize the results of a simulation study that lead us to a closer examination of the higher order expansions of the MSE. Although we prove our conjecture (3.23) originated by Remark 3.8, the result of the Main Theorem 6.13 is of asymptotic character, however. With respect to this the results of a preliminary simulation study are not only illustrative but provide insight to the fact that asymptotics kick in from sample size $n = 50$, already. This gives us evidence that the result of Theorem 6.13 already is valid for relatively small sample sizes.

5.1 Simulation design

Under R 2.4.1, we simulated `anzahl`=10000 runs of sample size $n = 50$ to $n = 100$ in the ideal location model $F = \mathcal{N}(\theta, 1)$ at $\theta = 0$.

Furthermore, we assume that for n large enough the finite ψ does not differ much from the asymptotically optimal ψ derived in (4.17) and (4.21), respectively:

Assumption 5.1. *In the context of an approximating simulation study we make the preliminary assumptions:*

(1) $F = \mathcal{N}(0, 1)$

(2) The IC ψ *is of asymptotically optimal form* (4.17) *and odd, i.e. for some* $A \in \mathbb{R}\backslash\{0\}$ *and* $b \in (0, \infty)$ *we assume* $c = -b/2$ *in* (4.7) *and have*

$$\psi(x) = -\frac{b}{2} \vee Ax \wedge \frac{b}{2} = A\left\{-\frac{g}{2} \vee x \wedge \frac{g}{2}\right\} \tag{5.1}$$

(3) The number of manipulated observations is binomial distributed: $K \overset{\text{i.i.d.}}{\sim} \text{Bin}(1, r/\sqrt{n})$.

Remark 5.2. *To get* $\mathbb{E}[\psi\Lambda_f] = 1$*, for* $F = \mathcal{N}(0, 1)$ *the Lagrange multiplier* A *is determined by* $A^{-1} = 2\Phi(g) - 1$.

Proof. In the Gaussian location model with $\theta = 0$ we gain for the scores function $\Lambda = x$, confer (4.12). Then by Assumption 5.1 (2) we get for the IC

$$\psi(x) = A \cdot \left\{ x \min\left(1, \frac{g}{|x|}\right) \right\}$$

Then with $\mathbb{E}[\psi\Lambda] = 1$ we have

$$\begin{aligned} A^{-1} &= \mathbb{E}x^2 \min\left(1, \frac{g}{|x|}\right) = 2\left[\int_0^g x^2\varphi(x)dx + g\int_g^\infty x\varphi(x)dx\right] \\ &= 2\left[-\int_0^g x\dot{\varphi}(x)dx - g\int_g^\infty \dot{\varphi}(x)dx\right] \\ &= 2\left[-x\varphi\big|_0^g + \int_0^g \varphi(x)dx - g\varphi(x)\big|_g^\infty\right] \\ &= 2\left[-g\varphi(g) + \Phi(g) - \frac{1}{2} + g\varphi(g)\right] \\ &= 2\Phi(g) - 1 \end{aligned}$$

□

Suppressing the conditioning w.r.t. $K \leq n/2$ in the context of thinned out neighborhoods, in the contiguous total variation situation, we have to deal with observations stemming from

$$Q_n = \bigotimes_{i=1}^n \left(dF + \frac{r}{\sqrt{n}}d\Delta_i\right)$$

In order to produce these observations we approximate them by a straight forward algorithm generating

> Q_n the measure resulting, when the K smallest observations under F^n
> $x_{(1)}, \ldots, x_{(K)}$ are transformed to $-x_{(1)}, \ldots, -x_{(K)}$ by changing sign.

Remark 5.3. *a) The mechanism of modification just sketched here is introduced and discussed in detail in section 8.3.*

b) As in Assumption 5.1 (2) the IC ψ is of form (5.1) we especially have that ψ is monotone. Otherwise we would have to do an ordering of the sample w.r.t. to $\psi(x)$.

c) The fact of ordering the sample creates a correlation of the sample, so we loose the assumption of independence. However, it is shown in Lemma 8.5 that we stay in the scenario generating a sample from $B_v(F, r/\sqrt{n})$. In Theorem 8.20 it is shown that under the Assumption 8.19 (p) the correlation vanishes for n large enough. Assumption 5.1 (2) implies Assumption 8.19 (p) even with $p = 1$.

d) Assumption 5.1 (3) is chosen for reasons of simplicity and motivated by the condition on the expectation of K in (3.27)*. A binomial distributed variable as chosen for the simulation study fulfills the condition $\mathbb{E}K = r\sqrt{n}$. But as Assumption 8.21 (VK), i.e. $\mathrm{Var}K = \frac{1}{2}r\sqrt{n}$, has to be chosen to gain the sufficiently high negligibility in Theorem 8.14, we repeat the introductory words of Assumption 5.1 just aiming at an approximating empirical result.*

e) In section 11.3 of [Kohl (2005)] a different algorithm is used to calculate a Finite-Sample Risk and Box-Cox plots[1] *comparable to ours. The approach consists not in a direct manipulation of the sample, but in a maximization of (deviation) probabilities followed by computation of the actual distribution of the data. With Notation 3.7 the probability of $\sum_{i=1}^n \chi_0(y_i) > 0$ and $\sum_{i=1}^n \chi_0(y_i) \geq 0$, respectively, under $Q_{-\tau_n} \in \mathcal{U}_*(-\tau_n)$ is maximal if*

$$Q_{-\tau_n}(\chi_0(y) = b) = Q_{-\tau_n}(y \geq b) = Q_0(y \geq b + \tau_n) = \max! \tag{5.2}$$

where $b \in (0, \infty)$ is some given clipping bound. For total variation neighborhoods ($* = v$) this leads to the c.d.f. (3.21) *and* (3.22)*, respectively, as already stated in Notation 3.7. The distribution of χ_0 under $Q'_{-\tau_n}$, for example, then calculates to*

$$\begin{aligned}
&Q'_{-\tau_n}(\chi_0(y) = -b) = (\Phi(-b + \tau_n) - \delta_n)_+ \\
&Q'_{-\tau_n}(-b < \chi_0(y) < t) = (\Phi(t + \tau_n) - \delta_n)_+ - (\Phi(-b + \tau_n) - \delta_n)_+ \quad t \in (-b, b) \\
&Q'_{-\tau_n}(\chi_0(y) = b) = 1 - (\Phi(b + \tau_n) - \delta_n)_+
\end{aligned}$$

So, analogous to our "piece-by-piece"-algorithm, in case of $Q'_{-\tau_n}$ mass δ_n is moved from the left tail to $[\tau_n + b, \infty)$.
For more details concerning this algorithm we refer to 11.3.2.1 and C.2 in [Kohl (2005)].

As estimator S_n we considered a three-step-estimator with the median as a starting estimate with IC ψ of form (5.1) and $g = 1.0$. For 10000 samples $X_1, ..., X_n$, abbreviated by (X_{i_j}) with $i = 1...n$ and $j = 1...10000$, we compute the empirical MSE by

$$\overline{\mathrm{empMSE}}_n = n \cdot \frac{1}{10000} \sum_{j=1}^{10000} S_n^2(X_{i_j}), \quad i = 1...n. \tag{5.3}$$

Furthermore, we compute the empirical asymptotic MSE according to

$$\overline{\mathrm{asyempMSE}}_n = n \cdot \frac{1}{10000} \sum_{j=1}^{10000} \left(\frac{1}{n} \sum_{i=1}^{n} \psi^2(X_{i_j}) \right) + r^2 \cdot \omega_v^2(\psi(X_{i_j})), \quad i = 1...n \tag{5.4}$$

with $\omega_v(\psi(X_{i_j})) = \sup \psi(X_{i_j}) - \inf \psi(X_{i_j})$ the total variation bias term (2.53) from Proposition 2.28 as used in risk (4.3), and consider $y = \overline{\mathrm{empMSE}}_n - \overline{\mathrm{asyempMSE}}_n$ for

[1] We show an excerpt of the results of the Box-Cox power transformation in [Kohl (2005)] in figure 5.7.

which we apply the Box-Cox power transformation provided by the MASS package of [Venables and Ripley (1999)]; i.e., we estimate λ by means of maximum likelihood such that $y^\lambda \approx 1/n$. That is, $\lambda \approx 1$ indicates $y = O(n^{-1})$. For further details we refer to the original paper by Box and Cox [Box and Cox (1964)]. For more details concerning the algorithm we refer to the appendix, subsection E.1.1 and E.1.2, respectively.

We anticipate the numerical result that the estimated values of λ are relatively close to 1, indeed, which may confirm our conjecture that we have a convergence of order n^{-1}; see Figures 5.1, 5.3 and 5.5. So in the next step we take a closer look and fit a linear model to the empirical MSE, i.e we establish

$$\overline{\text{empMSE}}_n = \beta_0 + \beta_1 \cdot 1/\sqrt{n} + \beta_2 \cdot 1/n. \tag{5.5}$$

By looking at the p-value for the corresponding t- and F-test we try to reduce formula (5.5) to a model less complex using just one regressor, i.e. we hope to see that a linear model with just the regressor of order n^{-1} shows the best fit to the data.

Remark 5.4. *We have to point out that the application of the t- and F-test as well as the AIC is based on heuristic assumptions like the postulation of (at least asymptotically) Gaussian variables. But as the whole character and the result of this chapter is a heuristic one, we do not bother about this too much.*

For the t- and F-test the following hypotheses and statistics hold (confer [Sachs and Hederich (2006)] with $p = 2$, for example):

$$\begin{aligned} H_0 : &\qquad \beta_i = 0 \\ H_A : &\qquad \beta_i \neq 0 \end{aligned}$$

$$\begin{aligned} \hat{F} &= \tfrac{(1/2)\cdot(SSY-RSS)}{(1/(n-3))\cdot RSS} \sim F_{n,(n-3)} \\ \hat{t}_i &= \tfrac{\hat{\beta}_i}{se(\hat{\beta}_i)} = \tfrac{((X'X)^{-1}X'y)_i}{\sqrt{(X'X)^{-1}_{ii}\hat{\sigma}^2)}} \sim t_{(n-3)} \end{aligned}$$

for $i \in \{1, 2\}$ with $\hat{\sigma}^2 = \frac{RSS}{n-3}$ and RSS the residual sum of squares.

For an interpretation of the p-value we offer the following tableau taken from [Sachs and Hederich (2006)], p. 324:

p-value	arguments against H_0
$> 15\%$	none
10% to $< 15\%$	hardly
5% to $< 10\%$	some
1% to $< 5\%$	many
$< 1\%$	lots of

We also make use of the stepAIC-procedure provided by the MASS package, which uses the Akaike Imformation Criterium (AIC) to indicate an appropriate structure of the linear model (confer [Sachs and Hederich (2006)] and [Venables and Ripley (1999)]). The AIC is defined as a measure for the "distance" between an unknown (true) mechanism, which may have generated the data and a model adapted to the data:

$$AIC = -2 \cdot \text{maximized log-likelihood} + 2 \cdot \#\text{parameters} \tag{5.6}$$

The second summand illustrates the "costs" for a too high account of parameters. So the optimal choice uses the fewest parameters. But whereas a low number of parameters raises the risk of an underfit, i.e. missing important effects or relations, a too high number of parameters leads to an overfit, i.e. pseudo effects or artefacts. Here the AIC criterium offers a balance between these two failure possibilities in the model setup.
Since the log-likelihood is defined only up to a constant depending on the data, this is also true for the AIC. For a linear model with n observations, p parameters and Gaussian errors the log-likelihood is

$$L(\beta, \sigma^2; y) = \text{const} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}|y - X\beta|^2$$

and by maximization over β we have

$$L(\hat{\beta}, \sigma^2; y) = \text{const} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\text{RSS}$$

with RSS$= \hat{\varepsilon}'\hat{\varepsilon}$ the residual sum of squares for the regression of $y - X\beta$. Thus if σ^2 is known, we can take

$$AIC = \frac{RSS}{\sigma^2} + 2p + const$$

but if σ^2 is unknown,

$$L(\hat{\beta}, \hat{\sigma}^2; y) = \text{const} - \frac{n}{2}\log\hat{\sigma}^2 - \frac{n}{2}, \qquad \hat{\sigma}^2 = \text{RSS}/n$$

and so

$$AIC = n\log RSS/n + 2p + const$$

The aim is to achieve a relatively small value for the AIC.

Within the stepAIC-procedure we can also use the attribute `test=F`. With the help of the specific F-statistic

$$\hat{F} = \frac{RSS_{(p-1)} - RSS_{(p)}}{RSS_{(p)}/(n - (p+1))},$$

p the number of influence parameters, the variable β_j with the smallest F-value should be eliminated, as it has no significant influence in the sense of $H_0 : \beta_j = 0$. A more detailed description of the step-procedure is given in [Hastie and Pregibon (1992)].

In order to contrast our result to the convex-contamination case we add a short look at the corresponding Box-Cox-plots and regression results. For every chosen radius we can clearly see the different structure of the convergence speed to be achieved: the peaks of the Box-Cox-plots are always shifted strongly to the left when convex contaminated data is used instead of total variation neighborhoods, see figures 5.2, 5.4 and 5.6.

5.2 Numerical evaluations

As the R output is very well structured and readable we prescind from a transformation to formatted text. The quotes on the R output contain all relevant and necessary information and levels of significance, for instance, are explained right away within the code.

5.2.1 r=0.1

Figure 5.1 shows the resulting Box-Cox-plot. Its maximum at about $\lambda = 1$ is consistent with the conjecture of a MSE convergence speed by the order of n^{-1}.

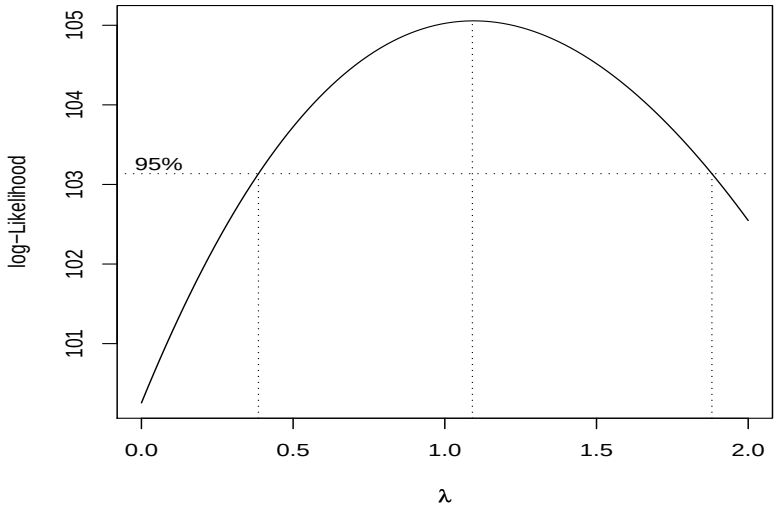

Figure 5.1: BoxCox-Plot for $r_v = 0.1, g = 1.0$ and $F = \mathcal{N}(0, 1)$.

Analysis of the linear model:[2]

```
> summary(lm(n3[,2]~I(1/sqrt(n3[,1]))+I(1/n3[,1])))

Call:
lm(formula = n3[, 2] ~ I(1/sqrt(n3[, 1])) + I(1/n3[, 1]))

Residuals:
       Min          1Q      Median          3Q         Max
-0.0519336 -0.0113140 -0.0007302  0.0105312  0.0376190

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.6506     0.2641   2.463   0.0174 *
I(1/sqrt(n3[, 1]))  11.8898     4.4551   2.669   0.0104 *
I(1/n3[, 1])       -45.0923    18.6304  -2.420   0.0193 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01744 on 48 degrees of freedom
Multiple R-Squared: 0.4277,     Adjusted R-squared: 0.4038
F-statistic: 17.93 on 2 and 48 DF,  p-value: 1.525e-06 }
```

Even so the Box-Cox plot in figure 5.1 indicates a single regressor of order n^{-1} there can no decision be made upon the p-values of the t-test in the summary of the full linear model above. In fact there is one item that does not fit the data: the (separately) computed asymptotic MSE for his model is about 1.33, which is far away from the estimated intercept even including the standard error. So we might look for a smaller model better suited to the data.

```
> stepAIC(lm(n3[,2]~I(1/sqrt(n3[,1]))+I(1/n3[,1])),test="F")$anova

Start:  AIC= -410.07
 n3[, 2] ~ I(1/sqrt(n3[, 1])) + I(1/n3[, 1])

                     Df Sum of Sq     RSS     AIC F Value   Pr(F)
<none>                             0.01 -410.07
- I(1/n3[, 1])        1  0.001782    0.02 -406.20    5.86 0.01934 *
- I(1/sqrt(n3[, 1]))  1  0.002167    0.02 -405.01    7.12 0.01035 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Stepwise Model Path Analysis of Deviance Table

Initial Model:
n3[, 2] ~ I(1/sqrt(n3[, 1])) + I(1/n3[, 1])
```

[2]name of the sample variable: `n3`

```
Final Model:
n3[, 2] ~ I(1/sqrt(n3[, 1])) + I(1/n3[, 1])


  Step Df Deviance Resid. Df Resid. Dev       AIC
1                         48 0.01460514 -410.0686
```

In this case there can no partial model be found based upon an analysis using an F-test and the AIC. But still the intercept is not the expected one. Hence we employ the minor linear model that was indicated by the Box-Cox result straight forward:

```
> summary(lm(n3[,2]~I(1/n3[,1])))

Call:
lm(formula = n3[, 2] ~ I(1/n3[, 1]))

Residuals:
       Min          1Q      Median          3Q         Max
-0.0499920 -0.0087438 -0.0002707  0.0085461  0.0386541

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.35475    0.01283 105.556  < 2e-16 ***
I(1/n3[, 1])  4.57626    0.90527   5.055  6.4e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0185 on 49 degrees of freedom
Multiple R-Squared: 0.3428,     Adjusted R-squared: 0.3294
F-statistic: 25.55 on 1 and 49 DF,  p-value: 6.401e-06
```

Now the intercept is a really good estimate for the asymptotic MSE, as 1.33 lies within the range of 1.355 ± 0.012. Even more the p-values are highly significant both for the intercept and the regressor coefficient.

For the sake of completeness we look at the other case, i.e. the regressor only belongs to order $n^{-1/2}$, too:

```
> summary(lm(n3[,2]~I(1/sqrt(n3[,1]))))

Call:
lm(formula = n3[, 2] ~ I(1/sqrt(n3[, 1])))

Residuals:
      Min         1Q     Median         3Q        Max
-0.050465 -0.008592 -0.000722  0.008692  0.038222
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.28722    0.02522  51.050  < 2e-16 ***
I(1/sqrt(n3[, 1]))  1.11814    0.21398   5.225 3.56e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01829 on 49 degrees of freedom
Multiple R-Squared: 0.3578,     Adjusted R-squared: 0.3447
F-statistic:  27.3 on 1 and 49 DF,  p-value: 3.556e-06
```

Again the results are highly significant, but the real value for the asymptotic MSE cannot be found within the calculated range of 1.287 ± 0.025.

Hence, for the case of $r = 0.1$ and $g = 1.0$ it seems evident that the MSE of an estimation settled upon a total variation neighborhood system skips the order $n^{-1/2}$ in its converging process and goes along with $O(n^{-1})$.

Comparison to convex contamination

According to (3.20) a total variation ball can be interpreted as being generated by two convex contamination balls. Consequently, this means that in the sense of a comparison of results in total variation and convex contamination neighborhoods we have to choose double the radius of the total variation setup for the convex contamination case. Thus in the present situation we have $r_c = 0.2$ instead of $r_v = 0.1$.

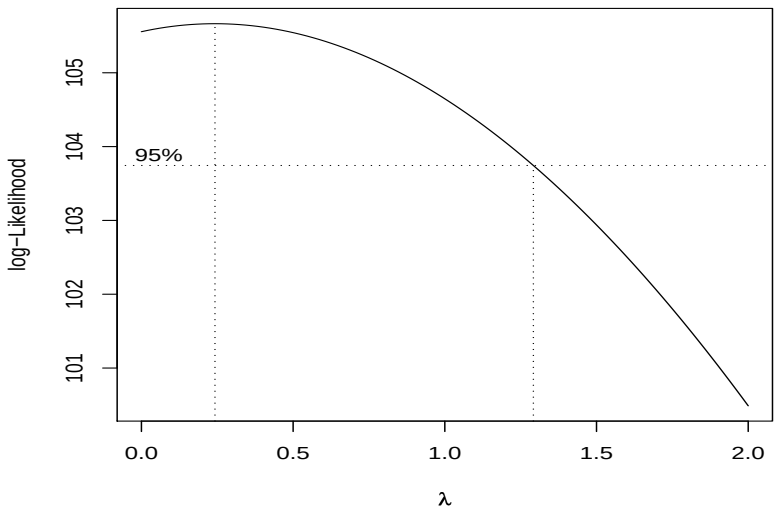

Figure 5.2: BoxCox-Plot for $r_c = 0.2$, $c = 1.0$ and $F = \mathcal{N}(0, 1)$.

The picture now totally differs from our result in the total variation case. Even though

there is no sharp maximum at $\lambda = 0.5$, it is obvious that we cannot expect a rate of convergence by the order n^{-1}. The already theoretically proven order $O(n^{-1/2})$ seems more likely.

Analysis of the linear model:[3]

An analysis of the underlying linear model shows that with respect to a lower AIC there can no model be found more appropriate than the full model with main term of order $n^{-1/2}$:

```
> stepAIC(lm(c1[,2]~I(1/sqrt(c1[,1]))+I(1/c1[,1])))$anova

Start:  AIC= -405.58
 c1[, 2] ~ I(1/sqrt(c1[, 1])) + I(1/c1[, 1])

                     Df Sum of Sq     RSS     AIC
<none>                              0.02 -405.58
- I(1/sqrt(c1[, 1]))  1 0.0006794    0.02 -405.45
- I(1/c1[, 1])        1 0.0007414    0.02 -405.27

Stepwise Model Path Analysis of Deviance Table

Initial Model:
c1[, 2] ~ I(1/sqrt(c1[, 1])) + I(1/c1[, 1])

Final Model:
c1[, 2] ~ I(1/sqrt(c1[, 1])) + I(1/c1[, 1])


  Step Df Deviance Resid. Df Resid. Dev       AIC
1                         48 0.01594798 -405.5827
```

5.2.2 r=0.25

The following figure looks very similar to the case of smaller radius $r = 0.1$ in figure 5.1. Again, a maximum at $\lambda = 1$ can be read of.

Analysis of the linear model:[4]

```
> summary(lm(n5[,2]~I(1/sqrt(n5[,1]))+I(1/n5[,1])))

Call:
lm(formula = n5[, 2] ~ I(1/sqrt(n5[, 1])) + I(1/n5[, 1]))

Residuals:
```

[3] name of the sample variable: `c1`
[4] name of the sample variable: `n5`

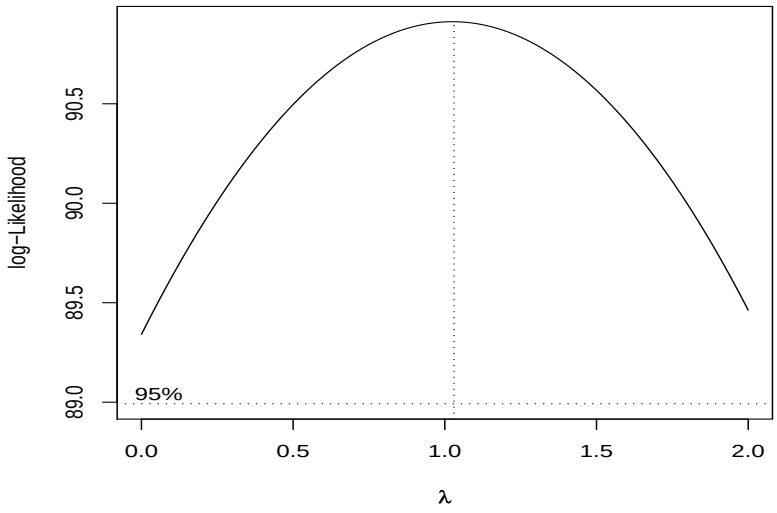

Figure 5.3: BoxCox-Plot for $r_v = 0.25$, $g = 1.0$ and $F = \mathcal{N}(0, 1)$.

```
      Min        1Q    Median        3Q       Max
-0.050076 -0.014657 -0.001082  0.014600  0.061132

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.7579     0.3679   4.779 1.71e-05 ***
I(1/sqrt(n5[, 1]))  0.4913     6.2052   0.079    0.937
I(1/n5[, 1])        5.1413    25.9490   0.198    0.844
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0243 on 48 degrees of freedom
Multiple R-Squared: 0.4328,     Adjusted R-squared: 0.4091
F-statistic: 18.31 on 2 and 48 DF,  p-value: 1.231e-06
```

The real asymptotic MSE now computes to 1.689, which lies within the suggested bounds for the intercept. However, both the calculated error terms and the p-values for the regressor coefficients are far too high to be trusted. Again we use the step-function to find a model more satisfying.

```
> stepAIC(lm(n5[,2]~I(1/sqrt(n5[,1]))+I(1/n5[,1])),test="F")\$anova

Start:  AIC= -376.27
 n5[, 2] ~ I(1/sqrt(n5[, 1])) + I(1/n5[, 1])

                     Df Sum of Sq     RSS     AIC F Value  Pr(F)
- I(1/sqrt(n5[, 1]))  1 3.700e-06    0.03 -378.27    0.01 0.9372
```

```
- I(1/n5[, 1])        1 2.317e-05    0.03 -378.23    0.04 0.8438
<none>                               0.03 -376.27

Step:  AIC= -378.27
 n5[, 2] ~ I(1/n5[, 1])

               Df Sum of Sq     RSS     AIC F Value     Pr(F)
<none>                        0.03 -378.27
- I(1/n5[, 1])  1      0.02    0.05 -351.36   37.37 1.564e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Stepwise Model Path Analysis of Deviance Table

Initial Model:
n5[, 2] ~ I(1/sqrt(n5[, 1])) + I(1/n5[, 1])

Final Model:
n5[, 2] ~ I(1/n5[, 1])


                  Step Df     Deviance Resid. Df Resid. Dev       AIC
1                                           48 0.02833364 -376.2721 2
- I(1/sqrt(n5[, 1]))  1 3.700147e-06        49 0.02833734 -378.2654
```

This time, with respect to the smaller AIC, the initial full model is reduced to a minor one. Indeed, the final model is the expected one, including only the regressor of order n^{-1}. We state the summary for this model:

```
> summary(lm(n5[,2]~I(1/n5[,1])))

Call:
lm(formula = n5[, 2] ~ I(1/n5[, 1]))

Residuals:
       Min         1Q     Median         3Q        Max
-0.0500586 -0.0146548 -0.0007811  0.0144541  0.0613530

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.78700    0.01668 107.118  < 2e-16 ***
I(1/n5[, 1])  7.19360    1.17668   6.113 1.56e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02405 on 49 degrees of freedom
Multiple R-Squared: 0.4327,     Adjusted R-squared: 0.4211
F-statistic: 37.37 on 1 and 49 DF,  p-value: 1.564e-07
```

Comparison to convex contamination

Again, with $r_c = 2r_v$, we get a similar figure for $r_c = 0.5$ as for $r_c = 0.2$:

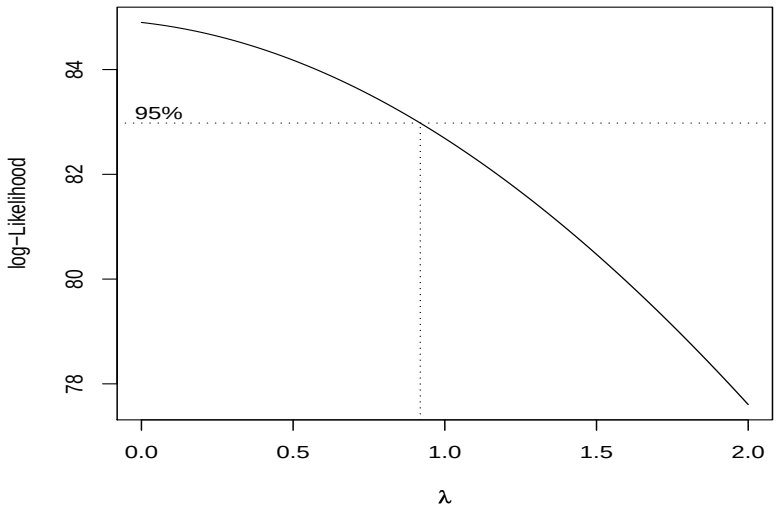

Figure 5.4: BoxCox-Plot for $r_c = 0.5$, $c = 1.0$ and $F = \mathcal{N}(0, 1)$.

Analysis of the linear model:[5]

The analysis of the linear model by the stepAIC-procedure gives:

```
stepAIC(lm(c5n4[,2]~I(1/sqrt(c5n4[,1]))+I(1/c5n4[,1])),test="F")$anova

Start:  AIC= -360.5
 c5n4[, 2] ~ I(1/sqrt(c5n4[, 1])) + I(1/c5n4[, 1])

                       Df Sum of Sq     RSS     AIC F Value  Pr(F)
- I(1/c5n4[, 1])        1 0.0005523    0.04 -361.78    0.69 0.4114
- I(1/sqrt(c5n4[, 1]))  1 0.0012420    0.04 -360.89    1.54 0.2200
<none>                                 0.04 -360.50

Step:  AIC= -361.78
 c5n4[, 2] ~ I(1/sqrt(c5n4[, 1]))

                       Df Sum of Sq     RSS     AIC F Value     Pr(F)
<none>                                 0.04 -361.78
- I(1/sqrt(c5n4[, 1]))  1      0.07    0.11 -313.42   82.54 4.406e-12 ***
---
```

[5] name of the sample variable: `c5n4`

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Stepwise Model Path Analysis of Deviance Table

Initial Model:
c5n4[, 2] ~ I(1/sqrt(c5n4[, 1])) + I(1/c5n4[, 1])

Final Model:
c5n4[, 2] ~ I(1/sqrt(c5n4[, 1]))


              Step Df     Deviance Resid. Df Resid. Dev       AIC
1                                         48 0.03860170 -360.5005 2
- I(1/c5n4[, 1])  1 0.0005523123        49 0.03915402 -361.7760
```

This time the stepAIC-procedure confirms the suggestion of the boxcox-plot. With the maximum closer to 0.5 than to 1.0 this leads to a reduced linear model containing only the $1/\sqrt{n}$ term, too. Again, this differs from the results computed for the total variation case above.

5.2.3 r=0.5

The Box-Cox-plot again leads to the conjecture of an exponent -1, exactly as in the cases for smaller radii:

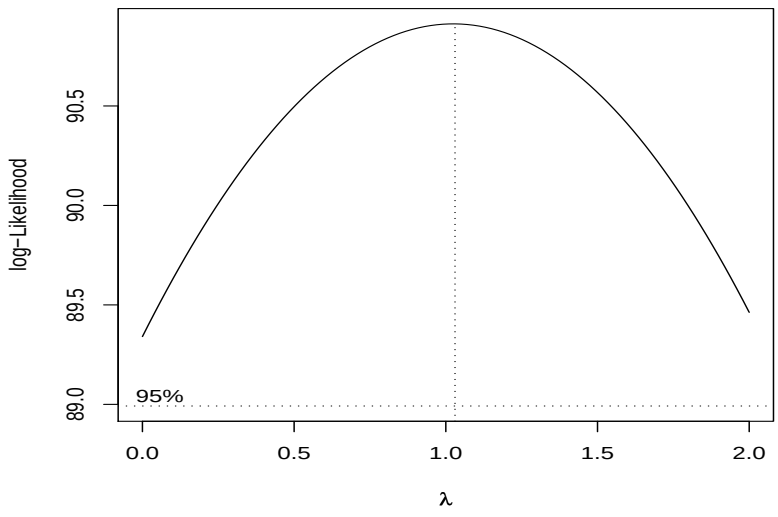

Figure 5.5: BoxCox-Plot for $r_v = 0.5$, $g = 1.0$ and $F = \mathcal{N}(0, 1)$.

Analysis of the linear model:[6]

[6]name of sample variable: `n7`

To avoid repetitions of all-too similar results, we skip an extensive application of step-procedures and just state the resulting parameters of the reduced linear model of order n^{-1}:

```
> summary(lm(n7[,2]~I(1/n7[,1])))

Call:
lm(formula = n7[, 2] ~ I(1/n7[, 1]))

Residuals:
       Min          1Q      Median          3Q         Max
-0.1100391 -0.0300190  0.0003800  0.0243646  0.0981380

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.20655    0.02791 114.875  < 2e-16 ***
I(1/n7[, 1]) 12.29438    1.96884   6.244 9.82e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04024 on 49 degrees of freedom
Multiple R-Squared: 0.4431,     Adjusted R-squared: 0.4318
F-statistic: 38.99 on 1 and 49 DF,  p-value: 9.818e-08
```

Comparison to convex contamination

As the results are consistent with the previous ones for convex contamination, we confine ourselves to presenting the associated Box-Cox-plot, which again differs clearly from figure 5.5 for the total variation case.

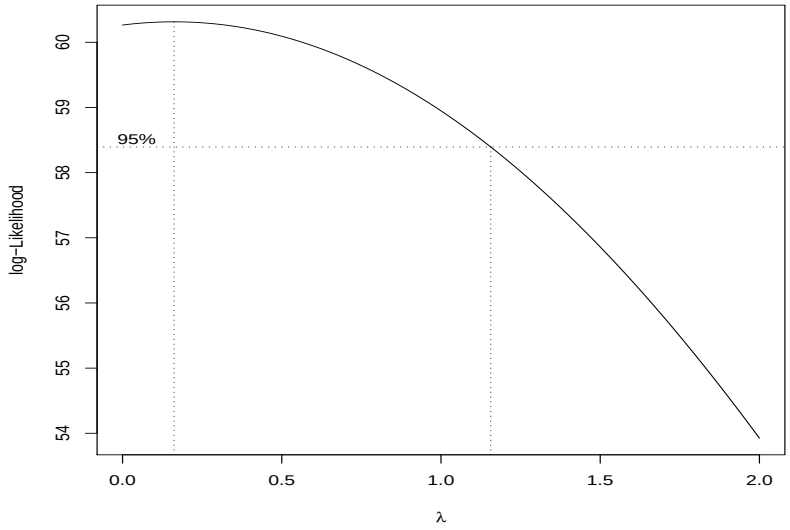

Figure 5.6: BoxCox-Plot for $r_c = 1.0$, $c = 1.0$ and $F = \mathcal{N}(0, 1)$.

5.2.4 Cross-check

Our results stay in line with the results of section 11.4 in [Kohl (2005)]. According to paragraph 11.4.2.2 (ibid), the plots obtained by the algorithm sketched in Remark 5.3 e) and a Box-Cox power transformation strongly confirm the conjecture for the finite-sample risk of the finite sample minimax estimator having speed of convergence of order $O(n^{-1})$. Similar to ours, the results in [Kohl (2005)] show to be almost independent from parameters like the radius δ_n corresponding to our r_v. The plots in figure 5.7 were computed by M. Kohl for sample sizes from $n = 2$ to $n = 250$. We reproduce figure 11.12 of [Kohl (2005)] with kindly permission of M. Kohl to illustrate the similarity to our figures 5.1, 5.3 and 5.5.

5.3 Summary

The results of the preceding section underline the results from M. Kohl and P. Ruckdeschel mentioned in subsection 3.2 earlier and affirm our conjecture (3.23). This encourages further investigations in the theoretical field, which are carried out in the next chapter, where we first take a look at the already proven results on MSE convergence in a convex contaminated setup. We then use the techniques suggested there as a guideline to investigate the total variation case.

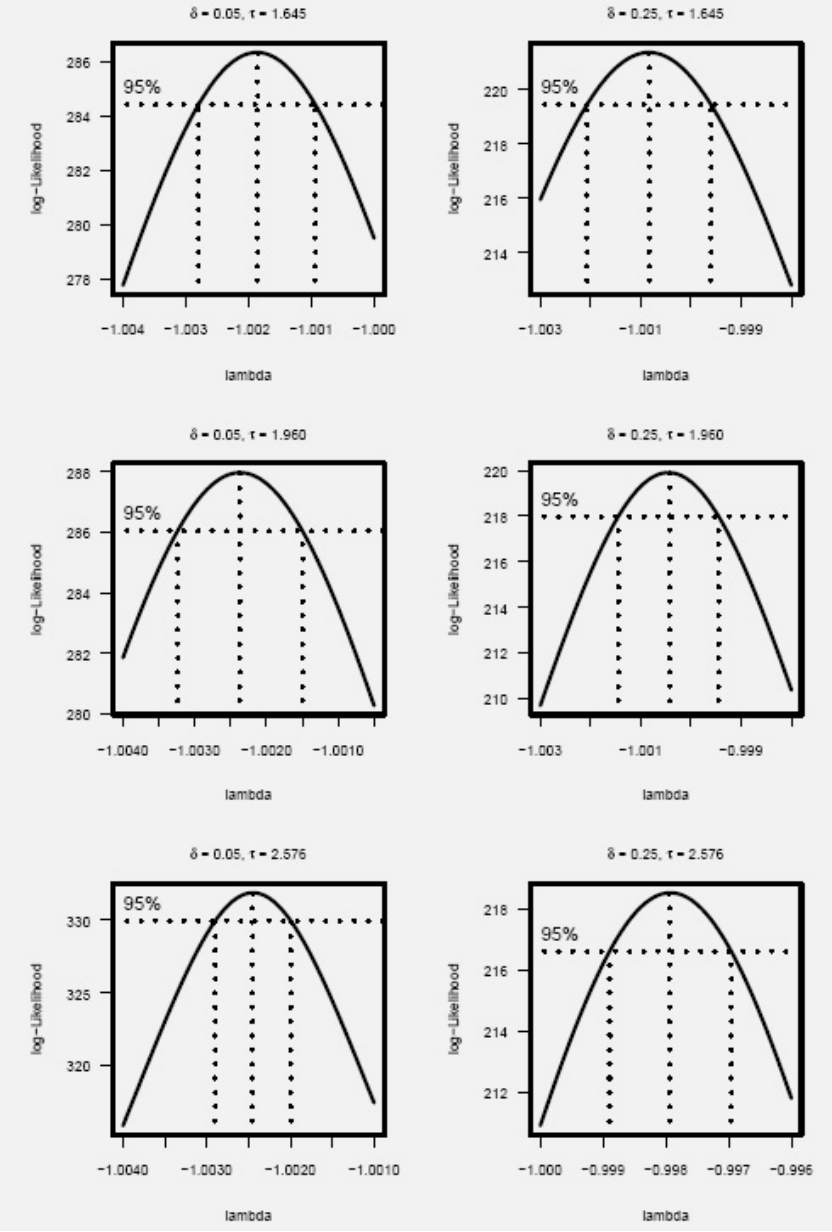

Figure 5.7: Results of the Box-Cox power transformation for the speed of convergence in the case of the finite-sample risk of the finite-sample minimax estimator and total variation neighborhoods; taken from [Kohl (2005)], p. 394.

Chapter 6

Higher Order Asymptotics for the MSE in the One–dimensional location model

In this chapter we focus on the question of a higher order expansion for the MSE of robust M-estimators of location on shrinking total variation neighborhoods. For reasons of comparison we first repeat the result for convex contamination.

6.1 Convex-Contamination neighborhoods

P. Ruckdeschel worked out the case of convex contamination neighborhoods for the one–dimensional location model in [Ruckdeschel (2005b)]. For reasons of comparison to the total variation case later on, we state his main result together with some necessary notations and assumptions.

We cite the notation used in subsection 3.1 of [Ruckdeschel (2005b)] together with Definition 3.1 and the assumptions made in subsection 3.2 (ibid.):

Notation 6.1. *To $\psi : \mathbb{R} \to \mathbb{R}$ monotone let $\psi_t(x) := \psi(x - t)$ and define the following functions*

$$\begin{aligned} L(t) &:= \mathbb{E}\psi(X - t), & V(t)^2 &:= \mathrm{Var}\psi(X - t), \\ \rho(t) &:= \mathbb{E}[(\psi(X - t) - L(t))^3]/V(t)^3, & \kappa(t) &:= \mathbb{E}[(\psi(X - t) - L(t))^4]/V(t)^4 - 3 \end{aligned}$$

Let $\check{y}_n$ and $\hat{y}_n$ sequences in $\mathbb{R}$ such that for some $\gamma > 1$

$$\psi(\check{y}_n) = \inf \psi + o\left(\frac{1}{n^\gamma}\right), \qquad \psi(\hat{y}_n) = \sup \psi + o\left(\frac{1}{n^\gamma}\right) \tag{6.1}$$

To state the main theorem, we need the following notation:

For $H \in \mathcal{M}_1(\mathbb{B}^n)$ and an ordered set of indices $I = (1 \le i_1 < \ldots < i_k \le n)$ denote H_I the marginal of H with respect to I.

Definition 6.2. *Consider three sequences c_n, d_n, and κ_n in $\mathbb{R}$, in $(0,\infty)$, and in $\{1,\dots,n\}$, respectively. We say that the sequence $(H^{(n)}) \subset \mathcal{M}_1(\mathbb{B}^n)$ is κ_n–*concentrated left [right] of c_n up to $o(d_n)$*, if for each sequence of ordered sets I_n of cardinality $i_n \le \kappa_n$*

$$1 - H_{I_n}^{(n)}\big((-\infty; c_n]^{i_n}\big) = o(d_n) \qquad \Big[\, 1 - H_{I_n}^{(n)}\big((c_n,\infty)^{i_n}\big) = o(d_n) \,\Big] \tag{6.2}$$

Assumption 6.3. *(bmi)* $\sup|\psi| = b < \infty$, *ψ monotone, $\psi \in \Psi_2$*

(D) For some $\delta \in (0,1]$, L, V, ρ, and κ from Notation 6.1 allow the expansions

$$\begin{aligned} L(t) &= l_1 t + \frac{1}{2} l_2\, t^2 + \frac{1}{6} l_3\, t^3 + O(t^{3+\delta}) \\ V(t) &= v_0(1 + \tilde{v}_1\, t + \frac{1}{2}\tilde{v}_2\, t^2) + O(t^{2+\delta}) \\ \rho(t) &= \rho_0 + \rho_1\, t + O(t^{1+\delta}) \\ \kappa(t) &= \kappa_0 + O(t^{\delta}) \end{aligned}$$

(Vb) $V(t) = O(|t|^{-(1+\delta)})$ for $|t| \to \infty$ and some $\delta \in (0,1]$

(C) Let f_t be the characteristic function of $\psi_t(X^{\mathrm{id}})$; then

$$\lim_{t_0 \to 0} \limsup_{s \to \infty} \sup_{|t| \le t_0} |f_t(s)| < 1$$

The main theorem appears as Theorem 3.5 in subsection 3.4 of [Ruckdeschel (2005b)]:

Theorem 6.4. *In the location model* (4.10) *with* (4.11) *assume (bmi) to (C). Then for sample size n,*

(a) the following expansion of the maximal MSE of an an M-estimator S_n to scores-function ψ holds

$$R_n(S_n, r, \varepsilon_0) = r^2 b^2 + {v_0}^2 + \frac{r}{\sqrt{n}}\, A_1 + \frac{1}{n}\, A_2 + o(n^{-1}) \tag{6.3}$$

with

$$A_1 = {v_0}^2 \Big(\pm (4\,\tilde{v}_1 + 3\,l_2)b + 1 \Big) + b^2 + [\,2\,b^2 \pm l_2\, b^3\,]\, r^2 \tag{6.4}$$

$$\begin{aligned} A_2 = {} & {v_0}^3 \Big((l_2 + 2\,\tilde{v}_1)\rho_0 + \frac{2}{3}\,\rho_1 \Big) + {v_0}^4 \,(3\,\tilde{v}_2 + \frac{15}{4}\,{l_2}^2 + l_3 + 9\,\tilde{v}_1^2 + 12\,\tilde{v}_1\, l_2\,) + \\ & + [\, {v_0}^2 \Big((3\,\tilde{v}_2 + 3\,\tilde{v}_1^2 + \frac{15}{2}\,{l_2}^2 + 2\,l_3 + 12\,\tilde{v}_1\, l_2\,)b^2 + 1 \pm (8\,\tilde{v}_1 + 6\,l_2\,)\, b \Big) + \\ & \pm 3\, l_2\, b^3 + 5\, b^2\,]\, r^2 + \Big((\frac{5}{4}\,{l_2}^2 + \frac{1}{3}\,l_3\,)b^4 \pm 3\, l_2\, b^3 + 3\, b^2 \Big)\, r^4 \end{aligned} \tag{6.5}$$

and we are in the $-$ [$+$]-case depending on whether (6.6) *or* (6.7) *below applies.*

(b) let $P_n^{\mathrm{di}} := \bigotimes_{i=1}^n P_{n,i}^{\mathrm{di}}$ *be contaminating measures for* (3.16)*. Then* Q_n *with* P_n^{di} *as contaminating measures achieves the maximal risk in* (6.18) *if for* $k_1 > 1$ *and* $k_2 > 6 \vee (\frac{3}{2} + \frac{3}{2\delta})$ *with* δ *from (Vb) and* $K_1(n) = \ulcorner k_1 r\sqrt{n}\urcorner$ *either*

$$(P_n^{\mathrm{di}}) \text{ is } K_1(n)\text{–concentrated left of } \check{y}_n - b\sqrt{k_2 \log(n)/n} \text{ up to } o(n^{-1}) \tag{6.6}$$

or

$$(P_n^{\mathrm{di}}) \text{ is } K_1(n)\text{–concentrated right of } \hat{y}_n + b\sqrt{k_2 \log(n)/n} \text{ up to } o(n^{-1}) \tag{6.7}$$

More precisely, if $\sup\psi < [>] - \inf\psi$*, the maximal MSE is achieved by contaminations according to* (6.6) *[*(6.7)*]. In case* $\sup\psi = -\inf\psi$*,* (6.6) *[*(6.7)*] applies if*

$$\tilde{v}_1 > [<] - \frac{l_2}{4}\Big(\frac{b^2}{v_0^2}(r^2+3)(1+\frac{r}{\sqrt{n}} - \frac{2r^2}{n}) + 3(1-\frac{b^2}{v_0^2})\Big) \tag{6.8}$$

If $\sup\psi = -\inf\psi$ *and there is "=" in* (6.8)*,* (6.6) *and* (6.7) *generate the same risk up to order* $o(n^{-1})$*.*

6.2 Total variation neighborhoods

We go on with some preparative Definitions, Notations and Lemmata in subsection 6.2.1 until we can state our Main Theorem 6.13. Therein we give the explicit expansion of form (1.1) for the total variation case. The key idea of the approach consists of transferring decomposition (3.19) to the expectation and variance terms $L_{\mathrm{re},i}(t) := \mathbb{E}_{\mathrm{re}}\psi(x_i - t)$ and $V_{\mathrm{re},i}^2(t) := \mathrm{Var}\psi(x_i - t)$ directly, before expanding them by Taylor series to get access to the coefficients defining the terms A_1 and A_2 (conf. Assumption 6.7 and 6.18, respectively). Thereby putting in more structure of / information about the basic total variation neighborhoods we get expressions more complex than in the proof for the convex contamination case in [Ruckdeschel (2005b)], confer Remark 6.16.

6.2.1 The Main Theorem

First, we give some definitions where in contrast to the functions defined in (6.1) we introduce the subscript i for every single observation. This is due to the fact that in the context of total variation neighborhoods we interpret sequences of shrinking balls $\mathcal{Q}_n^{(v)}(r)$ as in (3.19) by

$$Q_n = \bigotimes_{i=1}^n Q_{n,i} = \bigotimes_{i=1}^n (F + r_n\Delta_i) = P_n^{\mathrm{id}} + r_n P_{n,i}^{\mathrm{di}}.$$

Hence, we introduce:

Notation 6.5. *To* $\psi : \mathbb{R} \to \mathbb{R}$ *monotone let* $\psi_t(x) := \psi(x - t)$ *and define the following functions*

$$L_{\mathrm{re},i}(t) := \mathbb{E}\psi_t(X), \qquad V_{\mathrm{re},i}^2(t) := \mathrm{Var}\psi_t(X),$$
$$\rho_i(t) := \mathbb{E}[(\psi_t(X) - L_{\mathrm{re},i}(t))^3]/V_{\mathrm{re},i}(t)^3, \qquad \kappa_i(t) := \mathbb{E}[(\psi_t(X - L_{\mathrm{re},i}(t))^4]/V_{\mathrm{re},i}(t)^4 - 3,$$

where the subscript re, i *indicates the mean to be calculated under the real distribution* $Q_{n,i}$*, and with the centering* $\psi^0 := \psi - L_{re,i}$

$$\begin{aligned} L_{\mathrm{id}}(t) &:= \mathbb{E}_F \psi_t(X), & L_{\mathrm{c},i}(t) &:= \mathbb{E}_{\Delta_i} \psi_t(X), \\ V_{\mathrm{id}}(t) &:= \mathbb{E}_F (\psi_t^0)^2(X), & V_{\mathrm{c},i}(t) &:= \mathbb{E}_{\Delta_i} (\psi_t^0)^2(X), \end{aligned}$$

the subscript c, i *indicating the mean to be evaluated under the signed law* Δ_i*.*

Lemma 6.6. *With the functions defined in (6.5) we can write*

$$L_{\mathrm{re},i}(t) = L_{\mathrm{id}}(t) + \frac{r}{\sqrt{n}} L_{\mathrm{c},i}(t) \tag{6.9}$$

$$V^2_{\mathrm{re},i}(t) = V_{\mathrm{id}}(t) + \frac{r}{\sqrt{n}} V_{\mathrm{c},i}(t) \tag{6.10}$$

Proof.

$$\begin{aligned} L_{\mathrm{re},i}(t) &= \mathbb{E}\psi_t(x_i) = \int \psi_t dQ_{n,i} = \int \psi_t dF + \int \psi_t d(\frac{r}{\sqrt{n}} q_i F) \\ &= L_{\mathrm{id}}(t) + \frac{r}{\sqrt{n}} \int \psi_t d\Delta_i \\ &= L_{\mathrm{id}}(t) + \frac{r}{\sqrt{n}} L_{\mathrm{c},i}(t) \end{aligned}$$

and analogously for $V^2_{\mathrm{re},i}(t)$. □

To improve readability we drop the index i and assume the existence of Taylor expansions for $L_{\mathrm{id}}(t)$ and $L_{\mathrm{c}}(t)$ as well as for $V_{\mathrm{id}}(t)$ and $L_{\mathrm{c}}(t)$.

Assumption 6.7. *For some* $\delta \in (0, 1]$*,* L_{id}*,* V_{id}*,* L_{c}*,* L_{c} *and* $\rho(t)$ *from Notation 6.5 allow the expansions*

$$L_{\mathrm{id}}(t) = l_{id,0} + l_{id,1} t + \frac{1}{2} l_{id,2}\, t^2 + O(t^{2+\delta}) \tag{6.11}$$

$$L_{\mathrm{c}}(t) = l_{c,0} + l_{c,1} t + \frac{1}{2} l_{c,2}\, t^2 + O(t^{2+\delta}) \tag{6.12}$$

$$V_{\mathrm{id}}(t) = V_{id,0} + V_{id,1} t + O(t^{1+\delta}) \tag{6.13}$$

$$V_{\mathrm{c}}(t) = V_{c,0} + V_{c,1} t + O(t^{1+\delta}) \tag{6.14}$$

$$\rho(t) = \rho_{re,0} + O(t^{\delta}) \tag{6.15}$$

Remark 6.8. *a) For proving our conjecture* (3.23)*, we are satisfied with the second-order MSE, hence only assume expansion (6.15) of* $\rho(t)$*, neglecting the additional structural information of the total variation neighborhood, because according to the convex contamination case there don't appear any coefficients of* $\rho(t)$ *in the second-order MSE, confer (6.19) and (6.5). Furthermore, we don't pay attention to* $\kappa(t)$ *because of the same reason. P. Ruckdeschel pays attention to this effect in [Ruckdeschel (2005b)], Remark 3.6 c), and conjectures that this is probably due to the special loss function.*

b) We recall that $l_{c,0} = L_c(0)$ is the first-order bias term.

Further assumptions are (similar to the convex-contamination case)

Assumption 6.9. *(bmi)* $\sup|\psi| = \tilde{b} < \infty$, *$\psi$ monotone, $\psi \in \Psi_2$*

(Vb) $V_{\mathrm{id}}(t) = O(|t|^{-(1+\delta)})$ *for* $|t| \to \infty$ *and some* $\delta \in (0, 1]$

(C) Let f_t be the characteristic function of $\psi_t(X)$; then

$$\lim_{t_0 \to 0} \limsup_{s \to \infty} \sup_{|t| \le t_0} |f_t(s)| < 1 \tag{6.16}$$

Condition (C) is a local uniform Cramér condition, motivated by the fact that we apply Edgeworth expansion to functions of ψ_t. The Cramér condition ensures convergence and therefore is central to much of the theory. For more detailed information we refer to [Hall (1992)]. The Cramér condition is implied by

Lemma 6.10. *Assume $\mathcal{L}(\psi(X))$ has a nontrivial absolute continuous part and that ψ is continuous. Then (C) is fulfilled.*

Proof. confer [Ruckdeschel (2005b)], Lemma 3.1, Proof 7.2. The proof uses the Lebesgue Lemma. □

As already mentioned we are mainly interested in the vanishing of the $n^{-1/2}$-term A_1. So if one is content with an expansion of the MSE up to order $o(n^{-1/2})$, we may use the following weakened Cramér condition of nonlatticeness. This was first proved by Esseen in his path breaking work on convergence rates [Esseen (1945)]:

Assumption 6.11. *(C') "Uniformly" for t around $t = 0$, $\mathcal{L}\big(\psi_t(X)\big)$ is not a lattice distribution, that is, there exist $t_0 > 0$, $s_0 > 0$ such that for all $s_1 > s_0$*

$$\hat{f}_{s_0,t_0}(s_1) := \sup_{s_0 \le s \le s_1} \sup_{|t| \le t_0} |f_t(s)| < 1 \tag{6.17}$$

Remark 6.12. *a) P. Ruckdeschel mentions in [Ruckdeschel (2005b)] that although (C) implies (C'), contrary to (C), in (C') the case $\sup_{s_1} \hat{f}_{s_0,t_0}(s_1) = 1$ for all $s_0 > 0$ and all $t_0 > 0$ is allowed.*

b) Hall refers to Esseen in the bibliographical notes of chapter 2 in [Hall (1992)] and mentions that for non-lattice random variables additional terms of all orders must be added to take account of errors in approximating to discrete distributions by smooth distributions.

c) By condition (bmi) — as $\psi \in \Psi_2$ — it holds that $l_{id,0} = 0$ and $l_{id,1} = -1$.

Theorem 6.13 (**Main Theorem**). *In the location model* (4.10) *with* (4.11) *assume (bmi) to (C) or (C'), respectively, from Assumption 6.9. Then for sample size* n,

(a) The following expansion of the maximal MSE of an an M-estimator S_n *to IC* ψ *holds*

$$R_n(S_n, r) = r^2 b^2 + {v_{id,0}}^2 + \frac{r}{\sqrt{n}} A_1 + \frac{1}{n} A_2 + o(n^{-1}) \tag{6.18}$$

with

$$A_1 = (\pm l_{id,2} b^3 + 2 l_{c,1} b^2) r^3 + v_{id,0}^2 (2 l_{c,1} + 2 v_{c,0} \pm b(3 l_{id,2} + 4 v_{id,1})) \tag{6.19}$$

and A_2 *a polynomial in* r, b *and the Taylor coefficients from the momentum functions as in Assumption 6.7. We are in the* $-[+]$*-case depending on whether* (6.20) *or* (6.21) *below applies.*

(b) Let the summands of the decomposition $P_n^{\mathrm{di}} := \bigotimes_{i=1}^n \Delta_i = \bigotimes_{i=1}^n (\Delta_i^+ + \Delta_i^-) =: (P_n^{\mathrm{di}})^+ + (P_n^{\mathrm{di}})^-$ *be signed (modifying) measures for* (3.19). *Then* Q_n *with* P_n^{di} *as modifying measures achieves the maximal risk if for* $k_2 > 6 \vee (\frac{3}{2} + \frac{3}{2\delta})$ *with* δ *from (Vb) either*

$$\begin{gathered} (P_n^{\mathrm{di}})^- \text{is } r\sqrt{n}\text{-concentrated left of } \check{y}_n - b\sqrt{k_2 \log(n)/n} \text{ up to } o(n^{-1}) \\ \text{and} \\ (P_n^{\mathrm{di}})^+ \text{is } r\sqrt{n}\text{-concenrated right of } \hat{y}_n + b\sqrt{k_2 \log(n)/n} \text{ up to } o(n^{-1}) \end{gathered} \tag{6.20}$$

or

$$\begin{gathered} (P_n^{\mathrm{di}})^- \text{ is } r\sqrt{n}\text{-concentrated right of } \hat{y}_n + b\sqrt{k_2 \log(n)/n} \text{ up to } o(n^{-1}) \\ \text{and} \\ (P_n^{\mathrm{di}})^+ \text{is } r\sqrt{n}\text{-concenrated left of } \check{y}_n - b\sqrt{k_2 \log(n)/n} \text{ up to } o(n^{-1}) \end{gathered} \tag{6.21}$$

More precisely, if $\sup\psi < [>] - \inf\psi$, *the maximal MSE is achieved by modifications according to* (6.20) *[*(6.21)*]. In case* $\sup\psi = -\inf\psi$, (6.20) *[*(6.21)*] applies if*

$$v_{id,1} > [<] - \frac{l_{id,2}}{4}\Big(\frac{b^2 r^2}{v_{id,0}^2} + 3\Big) \tag{6.22}$$

If $\sup\psi = -\inf\psi$ *and there is "=" in* (6.22), (6.20) *and* (6.21) *generate the same risk up to order* $o(n^{-1})$.

Remark 6.14. *(a) As for the exact calculation of the A_2-term Assumption 6.7 has to be expanded to Assumption 6.18, we source the explicit A_2-term out to the appendix, where it can be read of in section C. However, we spell out the full A_2-term for the symmetric case in Corollary 6.19.*

(b) Similar to the convex contaminated case in [Ruckdeschel (2005b)] no ρ_0 [κ_0]-term shows up in the correction term A_1 [A_2], although being of corresponding order. We repeat P. Ruckdeschel's conjecture already mentioned in Remark 6.8 a) that this is probably due to the special loss function.

(c) Obviously, for $r = 0$, we get an approximation that is one order faster than under contamination, which again is similar to the convex contaminated case, confer Remark 3.6 (d) in [Ruckdeschel (2005b)].

(d) Apart from the infinitesimal neighborhood setup, i.e. for $r = 0$, we still get a better insight on the MSE as there are remaining terms in A_2 for $r = 0$. But as Assumption 6.18 is necessary for the full A_2-term to be calculable, we don't pay attention to this instance until Corollary 6.20.

6.2.2 Proof of the Main Theorem 6.13

The proof of Theorem 6.13 is prefixed by an outline packing the rather laborious character of the proof into 15 steps, before going on to the detailed deduction of the main result: after a partition of the real line, we show the negligibility of several cases (via the Chebyshev inequality and a Hoeffding bound, conf. Appendix B) and hence can confine ourselves to a shrinking compactum[1], wherein we apply an Edgeworth expansion[2] to the centered and standardized influence curve $\psi_{t,i}$. The massive use of the CAS `MAPLE`[3] enables us to compute several complex Taylor expansions of the integrand by keeping hold on the order of hundreds of terms. Additionally, the detection of a least favorable modification of the data with respect to the total variation bias term (conf. (6.64) and (6.65), respectively) leads us to the calculation of the final terms.

Outline

Following P. Ruckdeschel in the proof of Theorem 6.4 and motivated by the proof being rather laborious, we give an outline of the proof by listing the significant steps shortly one after the other:

(0) To get access to the number of modified observations we condition the MSE w.r.t. $K = k$.

(1) In order to apply the identity (4.27) we decompose the (conditioned) mean squared error like

$$n\,\mathrm{MSE}(S_n, Q_n|K = k) = \int_0^\infty P(S_n \geq \sqrt{t}|K = k)\,dt + \int_0^\infty P(S_n \leq -\sqrt{t}|K = k)\,dt$$

[1] conf. interval I in Figure 6.1.
[2] conf. Theorem A.5.
[3] The `MAPLE`-algorithm used is described in section E.2.

(2) Centering and standardization of ψ_t:

$$\frac{\psi_t - L_{\mathrm{re}}(t)}{V_{\mathrm{re}}(t)} =: \xi_{t,i} \quad \text{and} \quad s_n(t) := \frac{-\sqrt{n}L_{\mathrm{re}}(t)}{V_{\mathrm{re}}(t)}$$

(3) By exploitation of the monotonicity of ψ we get

$$P(S_n \leq t) = P\left(\frac{\sum \psi_t - nL_{\mathrm{re}}(t)}{\sqrt{n}V_{\mathrm{re}}(t)} < s_n(t)\right) + O(e^{-\gamma n})$$

(4) We do the partitioning

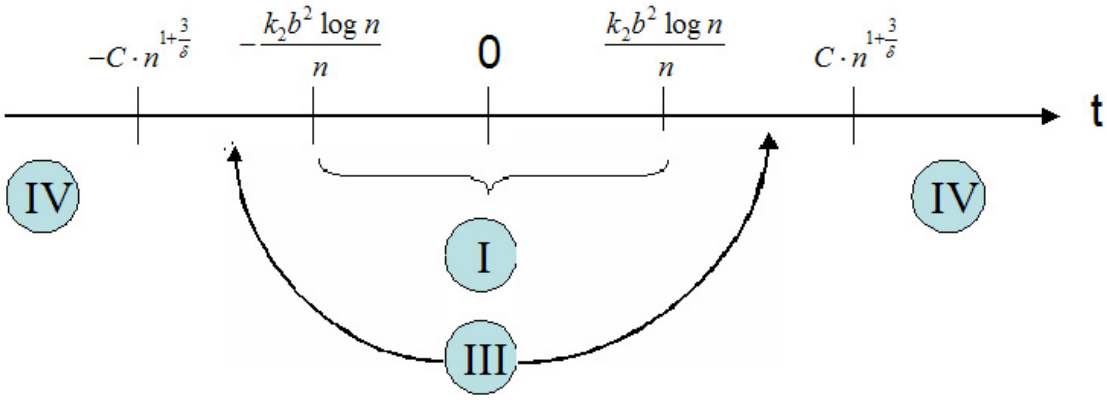

Figure 6.1: Partition of the real line by the values of the observations.

where it additionally holds for K that

	$K < r\sqrt{n}$	$r\sqrt{n} \leq K < n/2$	$K \geq n/2$
$\lvert t\rvert \leq k_2 b^2 \log(n)/n$	(I)	(II)	excluded
$k_2 b^2 \log(n)/n < \lvert t\rvert \leq C n^{1+3/\delta}$	(III)		
$\lvert t\rvert > C n^{1+3/\delta}$	(IV)		

(5) Case (IV) is of negligible order $o(n^{-1})$, because of the Chebyshev inequality and condition (Vb).

(6) Case (II) is of negligible order $o(e^{-rn^d})$, because of Hoeffding's Lemma.

(7) Case (III) is of negligible order $o(n^{-1})$, because of the Hoeffding bound, too.

(8) We apply an Edgeworth expansion to ξ_t by identification $t \rightsquigarrow \sqrt{t}$.

(9) Integration by parts gives an expression of form

$$n\,\mathrm{MSE}(S_n, Q_n) = \int_{-b\cdot\sqrt{k_2 \log n}}^{+b\cdot\sqrt{k_2 \log n}} s^2\left(\frac{u}{\sqrt{n}}\right) G_n'\left(s\left(\frac{u}{\sqrt{n}}\right)\right) du + o(n^{-1}) + R_n$$

where the first term is due to the compactum (I), the second illustrates the negligible cases (II) and (III), and R_n is a rest term.

(10) We show by Mills' ratio that $R_n = o(n^{-1})$, and therefore is negligible, too.

(11) We use the CAS `MAPLE` for Taylor expansions of the functions $s\left(\frac{u}{\sqrt{n}}\right)$, $s'\left(\frac{u}{\sqrt{n}}\right)$, $G'\left(s\left(\frac{u}{\sqrt{n}}\right)\right)$ and $G''\left(s\left(\frac{u}{\sqrt{n}}\right)\right)$. This yields an integrand of form

$$v_{id,0}\cdot\varphi\left(s\left(\frac{u}{\sqrt{n}}\right)\right)\cdot\left[1+\frac{1}{\sqrt{n}}P_1(u,t)+\frac{1}{n}P_2(u,t)\right]$$

with some elaborate polynomials P_1 and P_2.

(12) A Taylor expansion (by `MAPLE`) of $\varphi\left(s\left(\frac{u}{\sqrt{n}}\right)\right)$ results in

$$n\,\mathrm{MSE}(S_n, Q_n) = \int_I h_n(s)\varphi(s)\,\lambda(ds) + o(n^{-1})$$

with

$$h_n(s) := f(s)\cdot\left[1+\frac{1}{\sqrt{n}}\tilde{P}_1(u,t)+\frac{1}{n}\tilde{P}_2(u,t)\right]$$

for some function $f(s)$ and polynomials $\tilde{P}_1$ and $\tilde{P}_2$.

(13) To detect the least favorable deviation we use the fact that $h_n(s)$ is convex in $t = rL_c(0) \le rb$. So the maximum is to be found at the outer borders of the support. Thus the deviation looks something like figure 6.2.

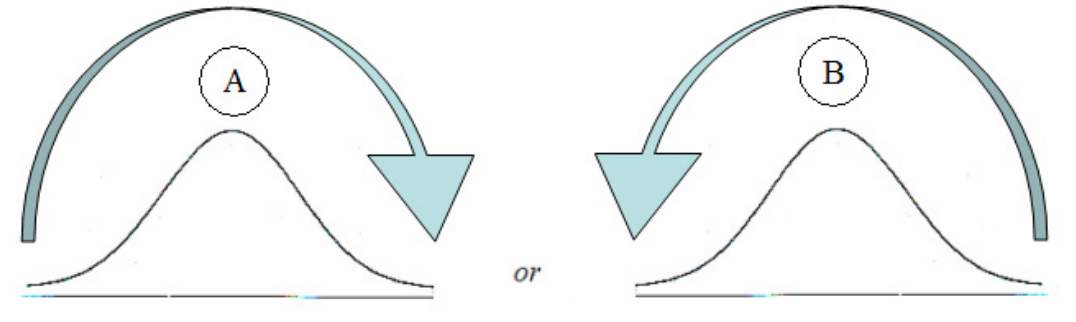

Figure 6.2: The least favorable deviation.

(14) Integrating out the integral by `MAPLE`.

(15) We give a decision criterium for the deviation cases (A) or (B) in figure 6.2.

Conversion

Ad (0) and (1):

In order to achieve the partitioning mentioned in item (2) we start with conditioning the mean squared error w.r.t. the number $K = k$ of substituted observations. After that we plug in $(X_i) \sim Q_n$ for some $Q_n \in \tilde{\mathcal{Q}}_n(r)$ into the defining relations for M-estimators of (4.26) and derive from (4.1) with $\theta = 0$

$$\begin{aligned} n\,\mathrm{MSE}(S_n, Q_n|K=k) &= \int_0^\infty P(S_n^2 \geq t|K=k)\,dt = \\ = \int_0^\infty P(S_n \geq \sqrt{t}|K=k)\,dt &+ \int_0^\infty P(S_n \leq -\sqrt{t}|K=k)\,dt \end{aligned} \tag{6.23}$$

Ad (2):

Next, we define

$$\tilde{t} := L_{\mathrm{c}}(0) = l_{c,0} \qquad s_n := s_n(t) = \frac{-\sqrt{n}\ L_{\mathrm{re}}(t)}{V_{\mathrm{re}}(t)} = \frac{-r\tilde{t} - \sqrt{n}\ \tilde{L}_{\mathrm{re}}(t)}{V_{\mathrm{re}}(t)} \tag{6.24}$$

with $\tilde{L}_{\mathrm{re}}(t) = L_{\mathrm{re}}(t) - \frac{r}{\sqrt{n}}\tilde{t}$, and

$$T_n(t) :=\ T_{\mathrm{re},n}(t) := \sum_i^n \psi_t(X_i) \tag{6.25}$$

Ad (3):

With these abbreviations we get by Lemma 4.5

$$\begin{aligned} P\{S_n \leq t\} &\overset{(4.27)}{=} P(T_n(t) < 0) + R_n^{(0)} \\ &= P\left(\frac{T_n(t) - nL_{\mathrm{re}}(t)}{\sqrt{n}V_{\mathrm{re}}(t)} < \frac{-nL_{\mathrm{re}}(t)}{\sqrt{n}V_{\mathrm{re}}(t)}\right) + R_n^{(0)} \\ &= P\left(\frac{T_n(t) - nL_{\mathrm{re}}(t)}{\sqrt{n}V_{\mathrm{re}}(t)} < \frac{-r\sqrt{n}\tilde{t} - \sqrt{n}\ \tilde{L}_{\mathrm{re}}(t)}{V_{\mathrm{re}}(t)}\right) + R_n^{(0)} \\ &= P\left(\frac{T_n(t) - nL_{\mathrm{re}}(t)}{\sqrt{n}V_{\mathrm{re}}(t)} < s_n(t)\right) + R_n^{(0)} \end{aligned}$$

where $R_n^{(0)} = O(\exp(-\gamma n))$, $\gamma > 0$ by Lemma 4.5, and $R_n^{(0)} \neq 0$ can only happen for mass points of $\mathcal{L}(T_n(t))$.

By Remark 6.12 c) we can simplify the expressions for $L_{\mathrm{re}}(t)$ and $V_{\mathrm{re}}(t)$ of Lemma 6.6:

$$L_{\mathrm{re}}(t) = \frac{r}{\sqrt{n}}l_{c,0} + (-1 + \frac{r}{\sqrt{n}}l_{c,1})t + (l_{id,2} + \frac{r}{\sqrt{n}}l_{c,2})\frac{t^2}{2} + O(t^3) \tag{6.26}$$

and

$$V_{\mathrm{re}}^2(t) = \left(V_{id,0} + \frac{r}{\sqrt{n}}V_{c,0}\right) + \left(V_{id,1} + \frac{r}{\sqrt{n}}V_{c,1}\right)t + O(t^2). \tag{6.27}$$

For our purpose we are interested in the square root of the last expression. As we do not want to loose the structure of (G) we use the Taylor expansion of the square root up to first order.

$$V_{\mathrm{re}}(t) = v_{id,0}\Big[(1 + \frac{r}{\sqrt{n}}v_{c,0}) + (v_{id,1} + \frac{r}{\sqrt{n}}v_{c,1})t\Big] + O(t^2). \tag{6.28}$$

with $v_{id,0} := \sqrt{V_{id,0}}$ and $v_{*,i} := \frac{V_{*,i}}{2V_{id,0}}$ otherwise.

Ad (4):

As we will show the negligibility of the integrand except for a shrinking compactum, we look at a partition according to the following tableau, where $C > 0$ is some constant and δ is the exponent from assumption (Vb):

	$K < k_1 r\sqrt{n}$	$k_1 r\sqrt{n} \le K < n/2$	$K \ge n/2$
$\lvert t\rvert \le k_2 b^2 \log(n)/n$	(I)	(II)	excluded
$k_2 b^2 \log(n)/n < \lvert t\rvert \le Cn^{1+3/\delta}$	(III)		
$\lvert t\rvert > Cn^{1+3/\delta}$	(IV)		

For the constants we anticipate their values:

constant	k_1	k_2
value	> 1	$> 6 \vee (\frac{3}{2} + \frac{3}{2\delta})$

Ad (5): Negligibility of case (IV)

As - after suitable substitution - the proof is similar to the proof of the negligibility in [Ruckdeschel (2005b)] we confine ourselves to the statement of the substitutions and add the detailed proof in the appendix (see subsection B.3). In subsection 8.4.4. of [Ruckdeschel (2005b)] we set

$$\tilde{t} := T_{\mathrm{c},n}(\sqrt{t}) = \sum \psi_t(X_i), \qquad X_i \sim \Delta_i \tag{6.29}$$

The application of the Chebyshev inequality delivers the desired result.

Ad (6): Negligibility of case (II)

For K binomial distributed, i.e. $K \sim \mathrm{Bin}(n, r/\sqrt{n}\,)$, this is an immediate consequence of Lemma 8.1 from [Ruckdeschel (2005b)], see subsection B.1 of the appendix. For K as defined in Assumption 8.21 (K) we get the same result from Theorem 8.22. Both Lemma and Theorem essentially are applications of Hoeffding's Lemma A.1.

Ad (7): Negligibility of case (III)

By a suitable substitution we can apply the proof from subsection 8.4.6 of [Ruckdeschel (2005b)]. In this sense we identify

$$\Delta := -\tilde{L}_{\mathrm{re}}(\sqrt{t}\,) - \frac{r\tilde{t}}{\sqrt{n}} \tag{6.30}$$

and get something of order $o(n^{-1})$. For more details we refer to subsection B.2 of the appendix.

Ad (8): Application of an Edgeworth expansion

We drop the dependency on the actual contamination $K = k$ and set $k \equiv r\sqrt{n}$, employing the structure of a total variation neighborhood. Hence, as then $K < k_1 r\sqrt{n}$, we stay in case (I).

The idea in the context of Theorem A.5 is to apply the CLT to $\frac{T_n(\sqrt{t}\,) - nL_{\mathrm{re}}(\sqrt{t}\,)}{\sqrt{n}\,V_{\mathrm{re}}(\sqrt{t}\,)}$ in order to achieve an Edgeworth expansion of the MSE.

On (I), by Lemma 4.5

$$P\big\{S_n \geq \sqrt{t}\big\} = P\left(\frac{T_n(\sqrt{t}\,) - nL_{\mathrm{re}}(\sqrt{t}\,)}{\sqrt{n}\,V_{\mathrm{re}}(\sqrt{t}\,)} > s_n(t)\right) + O(e^{-\gamma n}) \tag{6.31}$$

for some $\gamma > 0$, uniformly in t. We may apply Theorem A.5(b) to (6.31), identifying

$$\xi_{i,t} := \frac{1}{V_{\mathrm{re}}(t)}[\psi_t(X_i) - L_{\mathrm{re}}(t)], \qquad i = 1, \ldots, n \tag{6.32}$$

and setting $\Theta := \Theta_n = \{|t| \leq k_2 b^2 \log(n)/n\}$. This application is possible, as $|\psi| < \tilde{b}$, so $\sup_{t\in\Theta_n} \mathbb{E}|\tilde{\xi}_{i,t}|^5 < \infty$. By condition (C) of our assumptions, Cramér condition (A.13) of the theorem holds if n is large enough.

Remark 6.15. *a) The assumptions of Theorem A.5 seem to indicate that we reduce ourselves to a special case, i.e. the case of $\xi_{i,t}$ is a sequence of i.i.d. real-valued random variables; this means that the laws $Q_{n,i}$ and Δ_i stay fix for each index i. But as we are interested in the maximum MSE, choosing the supremal $Q_{n,i}$ for each X_i, it leads us to the decision upon a least favorable modification of the data, confer step (13). However, in section 6.5 we show that among the least favorable distributions there always is an i.i.d. one, so the simplific assumption of identically distributed variables actually is no limitation.*

b) In [Kohl (2005)], section 11.3.3, where M. Kohl raises the question and conjecture mentioned in subsection 3.2, he uses roughly the same ξ_i as we do in order to get an Edgeworth expansion via the CLT. He states that the ξ_i are i.i.d. a priori, but after all his setup is different from ours. Whereas we are interested in an approximation of the maxMSE, working in the framework of asymptotic theory, [Kohl (2005)] chapter

11.3 deals with the computation of the finite sample risk $Risk(S,*)$ that is defined by returning the maximum of the under- and overshooting probability of an M-estimator S. Therefore, and in contrast to our setup, the "least favorable" neighborhoods can be defined a priori and stay fix later on.

c) *In the framework of this proof and to improve readability we limit ourselves to the term A_1, hence only assume (C') and may apply Theorem A.5(a). For reasons of illustration (confer Remark 6.16) and with respect to the calculation of the A_2-term (confer Appendix C) we add terms of order n^{-1} at times, then always implicating assumption (C) instead of (C'); confer* (6.42) *and* (6.53)*, for instance.*

We apply Theorem A.5. With $G_{n,t}(s)$ from (A.9) we define

$$\tilde{G}_{n,t}(u) := G_{n,t}(s_{n,k}(u)), \qquad \tilde{G}_n(t) := \tilde{G}_{n,t}(t) \tag{6.33}$$

With these definitions we have for $|t| \le k_2 b^2 \log(n)/n$ uniformly in t:

$$\begin{aligned} & O(\exp(-\gamma n)) + P\{S_n \ge \sqrt{t}\} = \\ = \;& P\Big(\sum_{i=1}^{n} \xi_{i,\sqrt{t}} > s_n(\sqrt{t}\,)\Big) = 1 - \tilde{G}_n(\sqrt{t}\,) + O(n^{-3/2}) \end{aligned} \tag{6.34}$$

We set

$$l_n = \sqrt{k_2 \log(n)}, \qquad l_n^{(0)} = k_2 b^2 \log(n)/n \tag{6.35}$$

and obtain by the negligibility of (II), (III) and (IV):

$$\begin{aligned} & n\,\mathrm{MSE}(S_n, Q_n) = n \int_0^{l_n^{(0)}} 1 - \tilde{G}_n(\sqrt{t}\,) + \tilde{G}_n(-\sqrt{t}\,)\,dt + o(n^{-1}) = \\ = \;& \int_0^{bl_n} u\Big(1 - \tilde{G}_n(\frac{u}{\sqrt{n}}) + \tilde{G}_n(-\frac{u}{\sqrt{n}})\Big)\,du + o(n^{-1}) \end{aligned} \tag{6.36}$$

Ad (9):

$\tilde{G}_n$ is arbitrarily smooth. So integration by parts is available and gives

$$n\,\mathrm{MSE}(S_n, Q_n) = R_n + \int_{-bl_n}^{bl_n} \frac{u^2}{\sqrt{n}}\, G_n'(\frac{u}{\sqrt{n}})\,du + o(n^{-1}) \tag{6.37}$$

with

$$R_n := k_2\, \log(n)\, b^2 \left[1 - \tilde{G}_n(b\sqrt{\frac{k_2 \log(n)}{n}}) - \tilde{G}_n(-b\sqrt{\frac{k_2 \log(n)}{n}})\right] \tag{6.38}$$

Negligibility of the remainder term R_n

A closer investigation of $s_n(\pm b\sqrt{\frac{k_2 \log(n)}{n}})$ shows that

$$\begin{aligned} s_{n,k}(\pm b\sqrt{\frac{k_2 \log(n)}{n}}) &\overset{(G)}{=} \frac{O(\sqrt{n}) \pm b\sqrt{\frac{k_2 n^2 \log(n)}{n}} + O(\frac{n\,\log(n)}{n})}{\sqrt{n}\,(v_{id,0} + o(n^0))} = \\ &= \frac{\pm b\sqrt{k_2 \log(n)}}{v_{id,0}}(1 + o(n^0)) \end{aligned} \tag{6.39}$$

By condition (bmi) it holds that

$$v_{id,0}^2 = \mathbb{E}[\psi^2] \leq b^2, \tag{6.40}$$

hence $b/v_{id,0} > 1$. In particular, eventually in n,

$$|\tilde{s}_n(\pm b\sqrt{k_2 \log(n)}\,)| > \sqrt{6\log(n)} \tag{6.41}$$

The term $\tilde{G}_n(t) = G_{n,t}(s)$ reads according to definition (A.9) (up to order n^{-1})

$$G_{n,t}(s) := \quad \Phi(s) - \frac{\varphi(s)}{\sqrt{n}}\frac{\rho_t}{6}(s^2-1) \tag{6.42}$$

$$-\frac{\varphi(s)}{n}\Big[\frac{\kappa_t}{24}(s^3-3s) + \frac{\rho_t^2}{72}(s^5-10s^3+15s)\Big] \tag{6.43}$$

Ad (10):

We transfer from $G_{n,t}(s)$ to $1-G_{n,t}(s)$ as used in the rest term R_n in (6.38) and apply Mills' ratio[4] as defined in A.3. By Gordon's inequality (confer Lemma A.4) we get with $1-\Phi(s) = \Phi(-s)$ and $|\psi| \leq \tilde{b}$ by (bmi), $|\kappa| \leq \tilde{b}^4$, $|\rho| \leq \tilde{b}^3$:

$$\begin{aligned}
1 - G_n(s,t) &= \Phi(-s) - \frac{\varphi(s)}{\sqrt{n}}\frac{\rho_t}{6}(s^2-1) - \frac{\varphi(s)}{n}\Big[\frac{\kappa_t}{24}(s^3-3s) + \frac{\rho_t^2}{72}(s^5-10s^3+15s)\Big] \\
&\leq \varphi(s)\cdot\left[\frac{1}{s} - \frac{\rho_t}{6\sqrt{n}}(s^2-1) - \frac{1}{n}\Big[\frac{\kappa_t}{24}(s^3-3s) + \frac{\rho_t^2}{72}(s^5-10s^3+15s)\Big]\right] \\
&\leq k_3|s|^5\,\exp(-s^2/2)
\end{aligned}$$

with some $0 < k_3 < \infty$, independent of t and n (as we choose $n = 1$ in the last inequality). So

$$\max\big(1-G_{n,t}(s),\, G_{n,t}(-s)\big) \leq k_3|s|^5\,\exp(-s^2/2) \tag{6.44}$$

Thus for n sufficiently large

$$1 - \tilde{G}_n(b\sqrt{\frac{k_2\log(n)}{n}}\,) = \exp\left(-\frac{k_2 b^2\log(n)}{2v_{id,0}^2} + o(n^0)\right) = O\left(\frac{\log(n)^{5/2}}{n^{1+\delta}}\right) \tag{6.45}$$

for some $\delta > 0$. The same goes for $\tilde{G}_n(-2b\sqrt{\frac{\log(n)}{n}}\,)$, and therefore,

$$R_n = O(\log(n)^{7/2}/n^{1+\delta}) = o(n^{-1}) \tag{6.46}$$

So the rest term R_n is negligible up to suitable order and we have the expansion for the MSE:

$$n\,\mathrm{MSE}(S_n, Q_n\,) = \int_{-bl_n}^{bl_n} \frac{u^2}{\sqrt{n}}\; G_n'(\frac{u}{\sqrt{n}}\,)\,du + o(n^{-1}) \tag{6.47}$$

[4]Named after J. F. Mills and first mentioned in [Mills (1926)]

Ad (11): Extensive Taylor expansions via the CAS MAPLE

Following [Ruckdeschel (2005b)] we introduce the following notation to make more transparent, which terms are bounded to which degree.

$$t^\natural := r\tilde{t}, \qquad \tilde{s}_n(x) = s_n(\frac{x}{\sqrt{n}}) \tag{6.48}$$

A second cause for this transcription is the fact that it is easier for the CAS MAPLE to ignore irrelevant terms. Now, in our compactum (I)

$$u = O(\sqrt{\log(n)}), \qquad t^\natural = O(n^0).$$

The remainder terms of the Taylor expansions of assumption 6.7 aren't affected, too.
To ease readability, we drop the index of s_n and $\tilde{s}_n$, where it is clear from the context. Some more abbreviations for the derivatives of G pursue the same purpose:

$$\mathcal{G}_n(s,t) := G_{n,t}(s), \quad G^{(1)}_{n,t}(s) := [\frac{\partial}{\partial s}G_n](s,t), \quad G^{(2)}_{n,t}(s) := [\frac{\partial}{\partial t}G_n](s,t) \tag{6.49}$$

With $\tilde{s}'_n(x) = s'_n(\frac{x}{\sqrt{n}})/\sqrt{n}$ we spell out $\tilde{G}'_n(u)$ in (6.47) more explicitly and get by the chain rule

$$\begin{aligned} \tilde{G}'_n(\frac{u}{\sqrt{n}}) &= [G^{(1)}_{n,x}(s(x))s'(x) + G^{(2)}_{n,x}(s(x))]\Big|_{x=\frac{u}{\sqrt{n}}} = \\ &= G^{(1)}_{n,u/\sqrt{n}}(\tilde{s}(u))\,\tilde{s}'(u)\sqrt{n} + G^{(2)}_{n,u/\sqrt{n}}(\tilde{s}(u)) =: \tilde{g}_n(u)\sqrt{n} \end{aligned} \tag{6.50}$$

By the abbreviation

$$\tilde{g}_n(u) := G^{(1)}_{n,u/\sqrt{n}}(\tilde{s}(u))\,\tilde{s}'(u) + \frac{1}{\sqrt{n}}G^{(2)}_{n,u/\sqrt{n}}(\tilde{s}(u)) \tag{6.51}$$

we get as a next extension for the MSE

$$n\,\mathrm{MSE}(S_n, Q_n) = \int_{-bl_n}^{bl_n} u^2\tilde{g}_n(u)\,du + o(n^{-1}) \tag{6.52}$$

Up to now we did preparations in order to "feed" our MAPLE algorithms with appropriate terms. A summarizing documentation for these algorithms used in the sequel can be looked up in section E.2 of the appendix. We now expand the terms according to assumption 6.7. The MAPLE procedures asS gives

$$\begin{aligned} \tilde{s}(u) &= \frac{-t^\natural - \sqrt{n}\tilde{L}_{\mathrm{re}}(\frac{u}{\sqrt{n}})}{V_{\mathrm{re}}(\frac{u}{\sqrt{n}})} \\ &= \frac{1}{v_{id,0}}\Big[(u - t^\natural) - \frac{1}{\sqrt{n}}\big(-l_{c,1}ru - \frac{1}{2}l_{id,2}u^2 + (t^\natural - u)(v_{c,0}r + v_{id,1}u)\big)\Big] + o(n^{-(1/2+\delta)}) \end{aligned}$$

and asS1

$$\begin{aligned}\tilde{s}'(u) &= -\frac{\tilde{L}'_{\mathrm{re}}(\frac{u}{\sqrt{n}})}{V_{\mathrm{re}}(\frac{u}{\sqrt{n}})} + \frac{(t^\natural + \tilde{L}_{\mathrm{re}}(\frac{u}{\sqrt{n}}))V'_{\mathrm{re}}(\frac{u}{\sqrt{n}})}{V^2_{\mathrm{re}}(\frac{u}{\sqrt{n}})} \\ &= \frac{1}{v_0}\Big[1 + \frac{1}{\sqrt{n}}\Big(- l_{c,1}r - l_{id,2}u - (v_{c,0}r + v_{id,1}u) - ((-t^\natural + u)v_{id,1})\Big)\Big] + o(n^{-(1/2+\delta)})\end{aligned}$$

By application of the algorithm `asg` we get (up to order n^{-1})

$$\tilde{g}_n(u) = v_{id,0}\varphi(\tilde{s})[1 + \frac{1}{\sqrt{n}}P_1(u, t^\natural) + \frac{1}{n}P_2(u, t^\natural)] + o(n^{-(1+\delta)}) \tag{6.53}$$

for $P_1(u, t^\natural)$ and $P_2(u, t^\natural)$ polynomials in u, $t^\natural$, $v_{id,0}$, $v_{c,0}$, $v_{id,1}$, $l_{c,1}$, $l_{id,2}$ and ρ_0. As being rather longish - the polynomial P_2 consisting of not less than 98 summands, e.g. - and of no excessively informative character we abandon spelling out the explicit terms and refer for exact expressions to the `MAPLE` procedure `asg`.

Remark 6.16. *In contrast to the proof of the case ($* = c$) in [Ruckdeschel (2005b)] we get the polynomials P_1 and P_2 more complex; in the case of the polynomial P_2, which appears by calculation of the A_2-term assuming condition (C), we have 98 summands compared to 63 summands in the convex contaminated case, for example. The higher complexity is caused by putting in more structure of / information about the basic total variation neighborhoods than P. Ruckdeschel did when using convex contamination. Instead of staying in the ideal distributed setup and just expanding $L_{\mathrm{id}}(t)$ as P. Ruckdeschel did, we assume Taylor expansions of both the summands of $L_{\mathrm{re}}(t) = L_{\mathrm{id}}(t) + \frac{r}{\sqrt{n}}L_{\mathrm{c}}(t)$. As a result we obtain up to double the amount of coefficients for ($* = v$) than in the case ($* = c$).*

Ad (12):

To be able to calculate the integrals, the next candidate to be expanded is $\varphi(\tilde{s})$. This is done by the `MAPLE` procedure `dfac`. It expands $\varphi(\tilde{s})$ in a Taylor series about

$$s_1 = (u - t^\natural)/v_{id,0} \tag{6.54}$$

as

$$\varphi(\tilde{s}) = \varphi(s_1)[1 - s_1(\tilde{s} - s_1) + (s_1^2 - 1)(\tilde{s} - s_1)^2/2] + o(n^{-(1+\delta)}) \tag{6.55}$$

and hence

$$\tilde{g}_n(u) = v_{id,0}\varphi(s_1)g_n(s_1) + o(n^{-(1+\delta)}) \tag{6.56}$$

with

$$g_n(s_1) := 1 + \frac{1}{\sqrt{n}}\tilde{P}_1(s_1, t^\natural) + \frac{1}{n}\tilde{P}_2(s_1, t^\natural) \tag{6.57}$$

for $\tilde{P}_1(s_1, t^\natural)$ and $\tilde{P}_2(s_1, t^\natural)$ polynomials again to be looked up from a `MAPLE` procedure named `asgns`.

So we finally arrive at the MSE containing the density of the normal distribution as part of the integrand:

$$n\,\mathrm{MSE}(S_n, Q_n\,) = \int_{-bl_n/v_{id,0}}^{bl_n/v_{id,0}} h_n(s)\varphi(s)\,\lambda(ds) + o(n^{-1}) \tag{6.58}$$

for

$$h_n(s) = u_1(s)^2 g_n(s), \qquad u_1(s) = sv_{id,0} + t^\natural \tag{6.59}$$

Ad (13): Selection of the least favorable modification

Function $h_n(s)$ from (6.59) is a polynomial in s. So on (I), where $|s| = O(\log g(n))$, we may ignore terms of (pointwise-in-s) order $O(n^{-(1+\delta)})$. For an expansion of $h_n(s)$ we use the `MAPLE`-procedure `ashn`. It delivers a complicated expression of form

$$h_n(s) = (sv_{id,0} + t^\natural)^2 + \frac{1}{\sqrt{n}} Q_1 \tag{6.60}$$

where Q_1 is a polynomial in s, $t^\natural$, $v_{id,0}$, $v_{c,0}$, $v_{id,1}$, $l_{c,1}$, $l_{id,2}$, ρ_0 with $\deg(Q_1, s) = 5$ and $\deg(Q_1, t) = 4$; again, the exact expressions may be generated by our `MAPLE`-procedure `ashn`. Denoting the second partial derivative w.r.t. $t^\natural$ by an index t, t we get by calculation of the `MAPLE`-procedure `HND2s`

$$h_{n,t,t}(s) = 2 + \frac{1}{\sqrt{n}} Q_{1,t,t} \tag{6.61}$$

where $\deg(Q_{1,t,t}, s) = 3$. That is, on (I), uniformly in s, $h_{n,t,t}(s) = 2 + O(\log g(n)^3/\sqrt{n})$. Hence eventually in n, uniformly in s, h_n is strictly convex in $t^\natural$. Hence it takes its maximum on the boundary, that is for $|t^\natural|$ maximal. For fixed n we repeat that

$$t^\natural = r\tilde{t} = rL_{\mathrm{c}}(0) \tag{6.62}$$

with

$$L_{\mathrm{c}}(0) = \tilde{t} = \mathbb{E}_{\Delta_i}\psi_t\Big|_{t=0} = \mathbb{E}_F\psi_0 q_i = \mathbb{E}_F\psi_0 q_i^+ - \mathbb{E}_F\psi_0 q_i^- \le \sup_F \psi_0 - \inf_F \psi_0 = b \tag{6.63}$$

and $b \in]0, +\infty[$. The last inequality gets an equality in the symmetric case.
Hence we have

$$t^\natural = rb$$

as by (6.63) $\tilde{t}$ is bounded in absolute value by b. We now have to pay attention to the location case as we may disturb the original distribution, but as $\psi_t(x) = \psi(x-t)$ we have to shift the compactum, in which $\sqrt{t}$ is localized by the evaluation spot y_n. Hence this amounts to throwing essentially all the mass either from right of $\hat{y}_n + b\sqrt{k_2 \log g(n)/n}$ to the left of $\check{y}_n - b\sqrt{k_2 \log g(n)/n}$ or the other way round. We will deal with the decision which of the two alternatives is least favorable later on.

So we look at the decomposition of the signed measure $P_n^{\mathrm{di}} := \bigotimes_{i=1}^n \Delta_i = \bigotimes_{i=1}^n (\Delta_i^+ + \Delta_i^-) =: (P_n^{\mathrm{di}})^+ + (P_n^{\mathrm{di}})^-$ into positive and negative part. Then Q_n with P_n^{di} achieves the maximal risk if for $k_2 > 6 \vee (\frac{3}{2} + \frac{3}{2\delta})$ with δ from (Vb) either

$$\begin{gathered}(P_n^{\mathrm{di}})^-\text{is } r\sqrt{n}\text{-concentrated left of } \check{y}_n - b\sqrt{k_2 \log(n)/n} \text{ up to } o(n^{-1}) \\ \text{and} \\ (P_n^{\mathrm{di}})^+\text{is } r\sqrt{n}\text{-concentrated right of } \hat{y}_n + b\sqrt{k_2 \log(n)/n} \text{ up to } o(n^{-1})\end{gathered} \tag{6.64}$$

or

$$\begin{gathered}(P_n^{\mathrm{di}})^- \text{ is } r\sqrt{n}\text{-concentrated right of } \hat{y}_n + b\sqrt{k_2 \log(n)/n} \text{ up to } o(n^{-1}) \\ \text{and} \\ (P_n^{\mathrm{di}})^+\text{is } r\sqrt{n}\text{-concentrated left of } \check{y}_n - b\sqrt{k_2 \log(n)/n} \text{ up to } o(n^{-1})\end{gathered} \tag{6.65}$$

We remark that, as already mentioned, on (I), $|t^\natural|$ is bounded, so smallness of the probabilities in (6.64) resp. (6.65) entails that also the expectations of $(t^\natural)^j$, $j = 1, \ldots, 4$ arising in $h_n(s)$ are $O(n^{-1})$.

Let a distribution in $\tilde{\mathcal{Q}}_n$ which is modified according to (6.64) resp. (6.65) be denoted by Q_n^0. By the previous considerations, under Q_n^0, we may consider $|\tilde{t}|$ as being exactly b, and we will consider the cases $\tilde{t} = \pm b$ simultaneously.

Ad (14): Integration w.r.t. s

For the final integration of $\tilde{h}_n(s)$ w.r.t. s we use `ash` for the substitution $t^\natural = \pm rb$ in $\tilde{h}_n(s)$. Furthermore, as $bl_n/v_{id,0} > \sqrt{2 \log g(n)}$, by Lemma A.6, we may drop the integration limits and get

$$n \operatorname{MSE}(S_n, Q_n^0) = \int_{-\infty}^{\infty} \tilde{h}_n(s) \varphi(s) \, \lambda(ds) + O(n^{-1}) \tag{6.66}$$

For the integration, we also need the moments of the normal distribution, i.e. for $X \sim \mathcal{N}(0,1)$, $\mathbb{E}[X^j] = 0$, for $j = 1, 3, 5, 7$:

$$\mathbb{E}[X^2] = 1, \quad \mathbb{E}[X^4] = 3, \quad \mathbb{E}[X^6] = 15, \quad \mathbb{E}[X^8] = 115 \tag{6.67}$$

Next we apply the `MAPLE` procedures `intesout` and `asMSEK` and achieve the desired expansion for the MSE:

$$\begin{aligned} & n \operatorname{MSE}(S_n, Q_n^0) = \\ = \; & r^2 b^2 + v_{id,0}^2 + \\ & + \frac{r}{\sqrt{n}} \Big[(\pm l_{id,2} b^3 + 2 l_{c,1} b^2) r^2 + v_{id,0}^2 (2 l_{c,1} + 2 v_{c,0} \pm b(3 l_{id,2} + 4 v_{id,1}) \Big] + O(n^{-1}) \end{aligned}$$

Ad (15): Decision upon the alternative (6.64) **or** (6.65)

We join the declaration in 8.4.13 of [Ruckdeschel (2005b)] and denote Q_n^- a modified member in $\tilde{\mathcal{Q}}_n(r)$ according to (6.64) and correspondingly Q_n^+ according to (6.65). Now we have the differentiation of three cases, partly taking up the declaration in figure 6.2:

(A) $\sup\boldsymbol{\psi} < -\inf\boldsymbol{\psi}$: the maximal MSE is achieved by Q_n^-

(B) $\sup\boldsymbol{\psi} > -\inf\boldsymbol{\psi}$: the maximal MSE is achieved by Q_n^+

(AB) $\sup\boldsymbol{\psi} = -\inf\boldsymbol{\psi}$: the terms in A_1 are decisive:

$$\begin{aligned} & n(\mathbb{E}_{Q_n^+}[S_n^2] - \mathbb{E}_{Q_n^-}[S_n^2]) = \\ = \; & \frac{2rb}{\sqrt{n}}\Big(l_{id,2}b^2r^2 + v_{id,0}^2(3l_{id,2} + 4v_{id,1})\Big) + O(n^{-1}) \end{aligned} \tag{6.68}$$

Hence, Q_n^- $[Q_n^+]$ is least favorable up to $O(n^{-1})$, if

$$v_{id,1} \; > \; [<] - \frac{l_{id,2}}{4}\Big(\frac{b^2r^2}{v_{id,0}^2} + 3\Big) \tag{6.69}$$

If there is "=" in (6.69), no decision can be taken up to order $O(n^{-1})$.

6.3 The symmetric case for total variation

In order to prove our conjecture (3.23) we confine ourselves to the symmetric case. Actually, we are able to confirm the vanishing of the A_1-term in Corollary 6.19.

For F symmetric the asymptotically optimal IC ψ for one dimensional location is of form (4.17) and odd, i.e. for some $A \in \mathbb{R}\backslash\{0\}$ and $b \in (0,\infty)$ we assume $c = -b/2$ in (4.7) and have

$$\psi(x) = -\frac{b}{2} \vee Ax \wedge \frac{b}{2} = A\Big\{-\frac{g}{2} \vee x \wedge \frac{g}{2}\Big\} \tag{6.70}$$

Remark 6.17. *Considering only skew symmetric ICs is no restriction, as for any IC ψ, $\tilde{\psi} := -\psi(-.)$ is an IC, too, and so is the skew-symmetrized $\psi^{(s)} := \frac{1}{2}(\psi+\tilde{\psi})$. By convexity of the MSE (which is due to the quadratic loss function chosen in* (2.40)*), $\psi^{(s)}$ will be at least as good choice as ψ w.r.t. the MSE.*

For the generation of least favorable deviations from the ideal distribution, the modification mechanism, which is symbolized by our signed measure Δ_i takes mass from the outer LHS and throws it to the RHS, or the other way round. Furthermore, to maximize the total variation bias term $\omega_v(\psi) = \sup\psi - \inf\psi$ there must be sufficient mass in the areas, where ψ attains $\sup\psi$ or $\inf\psi$, respectively. Combining these two aspects with a look at (6.70) it is obvious that there is symmetry with respect to mass, i.e. Δ_i is odd, too.

As we expect the A_1-term to vanish for symmetry, we are interested in the third-order MSE, i.e. the explicit expression of the A_2-term. But therefore we need to make further assumptions:

Assumption 6.18. *For some $\delta \in (0,1]$, L_{id}, V_{id}, L_{c}, L_{c} and $\rho(t)$ from Notation 6.5 allow the expansions*

$$\begin{aligned}
L_{\mathrm{id}}(t) &= l_{id,0} + l_{id,1}t + \frac{1}{2}l_{id,2}\,t^2 + \frac{1}{6}l_{id,3}\,t^3 + O(t^{3+\delta}) && (6.71)\\
L_{\mathrm{c}}(t) &= l_{c,0} + l_{c,1}t + \frac{1}{2}l_{c,2}\,t^2 + \frac{1}{6}l_{c,3}\,t^3 + O(t^{3+\delta}) && (6.72)\\
V_{\mathrm{id}}(t) &= V_{id,0} + V_{id,1}t + \frac{1}{2}V_{id,2}t^2 + O(t^{2+\delta}) && (6.73)\\
V_{\mathrm{c}}(t) &= V_{c,0} + V_{c,1}t + \frac{1}{2}V_{c,2} + O(t^{2+\delta}) && (6.74)\\
\rho(t) &= \rho_0 + \rho_1 t + O(t^{1+\delta}) && (6.75)\\
\kappa(t) &= \kappa_0 + O(t^{\delta}) && (6.76)
\end{aligned}$$

Then we get indeed

Corollary 6.19. *Under symmetry of F it holds*

a) for the coefficients in (6.71), (6.72), (6.73) *or* (6.74), *respectively, that*

$$v_{c,0} = v_{id,1} = l_{c,1} = l_{id,2} = 0 \qquad (6.77)$$

and

$$v_{c,2} = l_{c,3} = 0, \qquad (6.78)$$

b) so by (6.77) *the term A_1 vanishes completely, and*

c) the maximum MSE reads

$$\begin{aligned}
n\,\mathrm{MSE}(S_n, Q_n^0) &= r^2b^2 + v_{id,0}^2 + \\
&\quad + \frac{1}{n}\Big[(\frac{1}{3}l_{id,3}b^4 \pm l_{c,2}b^3)r^4 + (\pm(4v_{c,1} + 3l_{c,2})v_{id,0}^2 b \\
&\quad + ((2l_{id,3} + 3v_{id,2})v_{id,0}^2)b^2)r^2 + (l_{id,3} + 3v_{id,2})v_{id,0}^4 \\
&\quad + \frac{2}{3}\rho_1 v_{id,0}^3\Big] + o(n^{-1}).
\end{aligned}$$

Proof. We recall that according to (G) $v_{*,i} := \frac{V_{*,i}}{2V_{id,0}}$ with $V_{id,0} = v_{id,0}^2 > 0$. With ψ odd, ψ^2 is even for $t = 0$. Hence the integrand of $V_{c,0}$, i.e. $(\psi^0)^2(0)d\Delta_i$, is odd. So

$$V_{c,0} = V_{c,i}(0) = \int (\psi^0)^2(t)\Big|_{t=0} d\Delta_i = 0. \qquad (6.79)$$

Hence $v_{c,0} = 0$ in (G). Furthermore

$$V_{id,1} = \int \frac{\partial}{\partial t}(\psi^0)^2(t)\Big|_{t=0} dF = 0, \tag{6.80}$$

therefore $v_{id,1} = 0$ as well. Looking at the coefficients of $L_{re}(t)$ we get

$$l_{c,1} = \int \frac{\partial}{\partial t}\psi(t)\Big|_{t=0} d\Delta_i = 0 \tag{6.81}$$

and

$$l_{id,2} = \int \frac{\partial^2}{\partial t^2}\psi(t)\Big|_{t=0} dF = 0. \tag{6.82}$$

Therefore the A_1-term vanishes completely under symmetry.

Under symmetry there are now too more coefficients equal to zero:

$$v_{c,2} = 0 \qquad \text{and} \qquad l_{c,3} = 0 \tag{6.83}$$

Hence

$$L_{\mathrm{re}}(t) = -\frac{r}{\sqrt{n}} l_{c,0} - t + \frac{r}{\sqrt{n}} l_{c,2}\frac{t^2}{2} + l_{id,3}\frac{t^3}{6} + O(t^{3+\delta}) \tag{6.84}$$

and

$$V_{\mathrm{re}}(t) = v_{id,0}\Big(1 + \frac{r}{\sqrt{n}} v_{c,1} t + v_{id,2}\frac{t^2}{2}\Big) + O(t^{2+\delta}). \tag{6.85}$$

Using the same methods as worked out in the non-symmetric second-order case in subsection 6.2.2, by the diverse vanishing coefficients we achieve a cut down version of the explicit A_2-term spelt out in section C of the appendix. So the MSE calculates to

$$\begin{aligned} n\,\mathrm{MSE}(S_n, Q_n^0) \;=\; & r^2 b^2 + v_{id,0}^2 + \\ & + \frac{1}{n}\Big[(\frac{1}{3} l_{id,3} b^4 \pm l_{c,2} b^3) r^4 + (\pm(4v_{c,1} + 3l_{c,2}) v_{id,0}^2 b \\ & + ((2l_{id,3} + 3v_{id,2}) v_{id,0}^2) b^2) r^2 + (l_{id,3} + 3v_{id,2}) v_{id,0}^4 \\ & + \frac{2}{3}\rho_1 v_{id,0}^3\Big] + o(n^{-1}). \end{aligned}$$

□

6.4 Cross-Checks

We deliver some cross-checks to confirm our result. First, we look at the terms of the higher order expansion in the symmetric convex-contaminated case and compare the remaining summands. Second, by the higher order expansion there are consequences for the ideal model, i.e. in the case $r = 0$, as we show in Corollary 6.20 and compare to the convex contaminated case, too. We see the expected consistency there.

6.4.1 The symmetric case for convex contamination

Let Q_n^0 be any distribution in $\mathcal{Q}_n$ attaining maximal risk in Theorem 6.4. Then according to Remark 3.7 c) of [Ruckdeschel (2005b)] under symmetry of F or more specifically if

$$l_2 = v_1 = \rho_0 = 0, \tag{6.86}$$

it holds in the convex contamination case that the result (6.3) of Theorem 6.4 becomes

$$n\,\mathrm{MSE}(S_n, Q_n^0) = (r^2b^2 + v_0^2)\Big(1 + \frac{r}{\sqrt{n}}\Big) + \frac{r}{\sqrt{n}}\big(b^2(1+r^2)\big) + O(n^{-1}) \tag{6.87}$$

Besides the fact of the term A_1 not vanishing we take a closer look at the remainder term $O(n^{-1})$ in order to compare the summands of the two different A_2-terms $A_{2,c}$ and $A_{2,v}$. Plugging in (6.86) in (6.5) we get

$$A_{2,c} = \frac{2}{3}\rho_1 {v_0}^3 + (3\tilde{v}_2 + l_3){v_0}^4 + (3\tilde{v}_2 b^2 + 2l_3 b^2 + 1){v_0}^2 r^2 + 5b^2r^2 + \Big(\frac{1}{3}l_3 b^4 + 3b^2\Big)r^4 \tag{6.88}$$

So we have the following table opposing the terms of $A_{2,c}$ and $A_{2,v}$:

$A_{2,c}$	$A_{2,v}$	identical
$\frac{2}{3}\rho_1 {v_0}^3$	$\frac{2}{3}\rho_1 v_{id,0}^3$	✓
$3\tilde{v}_2 {v_0}^4$	$3v_{id,2}v_{id,0}^4$	✓
$l_3 {v_0}^4$	$l_{id,3}v_{id,0}^4$	✓
$3\tilde{v}_2 b^2 {v_0}^2 r^2$	$3v_{id,2}v_{id,0}^2 b^2 r^2$	✓
$2l_3 b^2 {v_0}^2 r^2$	$2l_{id,3}v_{id,0}^2 b^2 r^2$	✓
$\frac{1}{3}l_3 b^4 r^4$	$\frac{1}{3}l_{id,3} b^4 r^4$	✓
$3b^2r^4$	$l_{c,2}b^3r^4$	-
${v_0}^2 r^2$	$4v_{c,1}v_{id,0}^2 br^2$	-
$5b^2r^2$	$3l_{c,2}v_{id,0}^2 br^2$	-

6.4.2 Consequences in the ideal model

If we abandon the infinitesimal approach introduced in section 2.3 by setting the radius $r = 0$, we get additional insights on the MSE, nevertheless, if we look at the term of order n^{-1}.

Corollary 6.20. *For $r = 0$ we get the MSE*

$$R_n(S_n, r) = {v_{id,0}}^2 + \frac{1}{n}\,A_2^0 + o(n^{-1})$$

with $A_{2,*}^0 \neq 0$:

($* = v$) By Assumption 6.18 and besides in the setup of Theorem 6.13 we have

$$\begin{aligned} A^0_{2,v} &= \Big[15(\frac{1}{2}l_{id,2}+v_{id,1})(-l_{id,2}-2v_{id,1})+\frac{15}{2}(-l_{id,2}-2v_{id,1})v_{id,1}+ \\ &+45(-\frac{1}{2}l_{id,2}-v_{id,1})^2+6v^2_{id,1}+3(l_{id,2}+v_{id,1})v_{id,1}+\frac{3}{2}l_{id,2}v_{id,1}+ \\ &+3v_{id,2}+l_{id,3}\Big]v^4_{id,0}+\Big(\frac{1}{2}\rho_0(\frac{3}{2}l_{id,2}+3v_{id,1})+\frac{2}{3}\rho_1+ \\ &+\rho_0(-l_{id,2}-2v_{id,1})+10(\frac{1}{2}l_{id,2}+v_{id,1})\rho_0+ \\ &+\frac{15}{2}\rho_0(-\frac{1}{2}l_{id,2}-v_{id,1})\Big)v^3_{id,0}-\frac{21}{4}l_{id,2}v^2_{id,0} \end{aligned}$$

($* = c$) In the setup of Theorem 6.4 we get

$$A^0_{2,c} = {v_0}^3\left((l_2+2\,\tilde{v}_1\,)\rho_0+\frac{2}{3}\,\rho_1\right)+{v_0}^4\,(3\,\tilde{v}_2+\frac{15}{4}\,{l_2}^2+l_3+9\,\tilde{v}_1^2+12\,\tilde{v}_1\,l_2\,)$$

Under symmetry of F the cases ($* = c$) and ($* = v$) coincide:

$$F \text{ symmetric} \quad \Rightarrow \quad A^0_{2,c} = \frac{2}{3}\rho_1 v_0^3+(3\,\tilde{v}_2+l_3){v_0}^4 \tag{6.89}$$

$$A^0_{2,v} = \frac{2}{3}\rho_1 v^3_{id,0}+(3\,v_{id,2}+l_{id,3}){v_{id,0}}^4 \tag{6.90}$$

Proof. Simply to be read of from the explicit A_2-term, confer appendix C. □

6.5 Negligibility of the non-i.i.d. case

We pick up the problem mentioned in remark 6.15 and show in the Lemmata 6.21 and 6.22 by an argument taken from [Feller (1971)] that the simplific assumption of identically distributed variables in the proof of Theorem 6.13 actually is no limitation.

In this sense we perform the single steps to obtain an Edgeworth expansion. The i.i.d. case can be looked up in [Field and Ronchetti (1990)], p.10, for example.

Lemma 6.21. *For $X_i \overset{\text{i.i.d.}}{\sim} F$ let the signed measure $\Delta_i \in \mathcal{M}_1(\mathbb{B})$ be defined as in (3.18) so that $Q_n = \otimes_{i=1}^n Q_{n,i}$ with $dQ_{n,i} = dF + r_n d\Delta_i$. Furthermore, let $V_{\text{re}}(t)$, $L_{\text{re}}(t)$ be functions as defined in Notation 6.5 and $\psi : \mathbb{R} \to \mathbb{R}$ an IC satisfying condition (bmi) from Assumption 6.9 with $\psi_t(x) := \psi(x-t)$. Then by assumptions (C) or (C'), respectively, for the non–identical but independently distributed variables*

$$\xi_{i,t} := \frac{1}{V_{\text{re}}(t)}[\psi_t(X_i)-L_{\text{re}}(t)], \qquad i=1,\ldots,n \tag{6.91}$$

with fourth moments, it holds for the error term $D_n(x_i)$ generated in the Edgeworth expansion compared to the case of identical distributed variables that

$$|D_n(x_i)| = o(1/\sqrt{n}), \qquad i=1,\ldots,n. \tag{6.92}$$

Proof. Denote by $\chi_{S_n}(t)$ the characteristic function of the (M-estimator) functional S_n. Then with $\xi_j := \xi_{j,t}$ independent

$$\begin{aligned}\chi_{S_n/\sqrt{n}}(t) &= \mathbb{E}e^{itS_n/\sqrt{n}} = \chi_{S_n}(t/\sqrt{n}) = \mathbb{E}e^{it/\sqrt{n}\sum \xi_j} = \\ &= \mathbb{E}\prod_j e^{it/\sqrt{n}\xi_j} = \prod_j \mathbb{E}e^{it/\sqrt{n}\xi_j} = \\ &= \prod_j \chi_{\xi_j}(t/\sqrt{n}). \end{aligned} \tag{6.93}$$

We now apply a Taylor-expansion to $\chi_{\xi_j}(t/\sqrt{n})$ at $t = 0$:

$$\begin{aligned}\prod_j \chi_{\xi_j}(t/\sqrt{n}) &= \prod_j \left[\chi_{\xi_j}(0) + \chi'_{\xi_j}(0)\frac{t}{\sqrt{n}} + \chi''_{\xi_j}(0)\frac{t^2}{2\sqrt{n}} + \chi^{(3)}_{\xi_j}(0)\frac{t^3}{6n^{3/2}} + \chi^{(4)}_{\xi_j}(0)\frac{t^4}{24n^2} + o\left(\frac{1}{n^2}\right)\right] \\ &= \prod_j \left[1 + (i\mathbb{E}\xi_j)\frac{t}{\sqrt{n}} - \mathbb{E}\xi_j^2\frac{t^2}{2n} - i\mathbb{E}\xi_j^3\frac{t^3}{6n^{3/2}} + \mathbb{E}\xi_j^4\frac{t^4}{24n^2} + o\left(\frac{1}{n^2}\right)\right] \\ &= \prod_j \left[1 - \frac{t^2}{2n} - \frac{it^3\mathbb{E}\xi_j^3}{6n^{3/2}} + \frac{t^4\mathbb{E}\xi_j^4}{24n^2} + o\left(\frac{1}{n^2}\right)\right] \end{aligned} \tag{6.94}$$

Next, we use $\prod_j[.] = \exp(\sum_j \ln[.])$. Hence

$$(6.94) = \exp\left\{\sum_j \ln\left[1 - \frac{t^2}{2n} - \frac{it^3\mathbb{E}\xi_j^3}{6n^{3/2}} + \frac{t^4\mathbb{E}\xi_j^4}{24n^2} + o\left(\frac{1}{n^2}\right)\right]\right\}. \tag{6.95}$$

The ln in (6.95) denotes the main branch of the complex logarithm defined in $\mathbb{C}\backslash\{z \in \mathbb{R} : z \leq 0\}$. As long as the increment of the logarithm does not take values on $\mathbb{R}_0^-$ we stay in the main branch[5]. Because of the Taylor expansion above at $t = 0$ we have $t \in \mathrm{B}_\varepsilon(0)$, $\varepsilon > 0$ small. With the increment of the ln according to (6.95) we get

$$1 - \frac{t^2}{2n} - \frac{it^3\mathbb{E}\xi_j^3}{6n^{3/2}} + \cdots = 1 - O(t^2) \approx 1 > 0 \tag{6.96}$$

Thus, with the increment only little varying we stay in the main branch.

We now apply a Taylor expansion once again to the logarithm as $\ln(1+x) = x - x^2/2 + o(x^2)$ at $x = 0$:

$$\begin{aligned}\chi_{S_n}(t/\sqrt{n}) &= \exp\left\{\sum_j \left[-\frac{t^2}{2n} - \frac{it^3\mathbb{E}\xi_j^3}{6n^{3/2}} + \frac{t^4\mathbb{E}\xi_j^4}{24n^2} + o\left(\frac{1}{n^2}\right)\right]\right\} \\ &= \exp\left(-\frac{t^2}{2}\right) \cdot R_{i,n}(t) \end{aligned} \tag{6.97}$$

[5]For details concerning the complex logarithm we refer to subsection 5.2 of [Jänich (1999)] and Section V, §1 of [Fischer and Lieb (1992)], respectively.

Next we check for independent variables whether we attain an error term small enough to achieve an Edgeworth expansion up to $o(1/\sqrt{n})$. We use an argument from [Feller (1971)], p. 546:

With

$$\mathbb{E}_{Q_{n,i}}(\xi_i) = 0, \qquad \mathbb{E}_{Q_{n,i}}(\xi_i^2) = 1, \qquad \mathbb{E}_{Q_{n,i}}(\xi_i^3) = \rho_i$$

put

$$s_n^2 = n, \qquad r_n = \rho_1 + \cdots + \rho_n$$

and denote by F_n the distribution of the sum $\xi_1 + ... + \xi_n$. Then for all x and n consider the one-term Edgeworth expansion with rest term $D_n(x)$

$$F_n(x) - \Phi(x) - \frac{r_n}{6s_n^2\sqrt{n}}\varphi(x) = D_n(x) \tag{6.98}$$

In the case of equal components it holds that $D_n(x) = o(1/\sqrt{n})$. Now D_n is the sum of various error terms which in the present situation need not be of comparable magnitude. But according to [Feller (1971)], if the ξ_i have fourth moments (which is satisfied in our case as $\sup_t \kappa_t < \infty$) and the ξ_i have no lattice distributions with the same span (satisfied by condition (C')) it holds that

$$|D_n(x)| = O(n^2 s_n^{-6}) + O(n s_n^{-4}). \tag{6.99}$$

Hence

$$\frac{n^2}{s_n^6} = \frac{n^2}{n^3} = o\left(\frac{1}{\sqrt{n}}\right) \tag{6.100}$$

and

$$\frac{n}{s_n^4} = \frac{n}{n^2} = o\left(\frac{1}{\sqrt{n}}\right). \tag{6.101}$$

Thus we get the desired exactness with $|D_n(x)| = o(1/\sqrt{n})$.

□

The next Lemma shows that, as already mentioned in Remark 6.15, the least favorable modification can indeed be achieved by interpretation of the ξ_i as i.i.d. variables.

Lemma 6.22. *Under the assumptions of lemma 6.21 a least favorable substitution as in 6.13 b) is achieved in the i.i.d. case, i.e $\Delta_i = \Delta$ so that $dQ_n = \otimes_{i=1}^n dF + r_n d\Delta = dF^n + r_n d\Delta^n$.*

Proof. The first factor in (6.97) is identically the same as in the i.i.d. case, whereas the second factor $R_{i,n}(t)$ comprises an index i. Applying a Taylor expansion to $R_{i,n}(t)$ at $t = 0$ (i.e. expanding $\exp(x)$), we get some polynomial in t with leading term 1. Finally the Fourier transformation delivers the Edgeworth expansion.
Important in the non-i.i.d. case now is that plugging in ξ_i considering the structure of the random variables stemming from a total variation neighborhood, we just will get some

differing additional summands of order $r/\sqrt{n}$ for each ξ_j in $\mathbb{E}\xi_j^3$ and $\mathbb{E}\xi_j^4$. But first, this only affects terms at least of order $1/n$, and second the additional summands do not affect the structure of the sum with the denominators increasing in $\sqrt{n}$-steps.
So in the non-i.i.d. case we will get the same structure of the Edgeworth expansion, but perhaps different terms from order $1/n$ on.
As the the first factor in (6.97) represents the first-order expansion term (and is nothing else than the characteristic function of the normal density $\varphi(x)$), it finally delivers the first-order MSE term, i.e. the sum of ideal variance and squared (total variation) bias multiplied by the radius.
But now we recall that in step (13) of subsection 6.2.2 the least favorable modification is chosen w.r.t. of the maximum first-order bias, represented by $L_{c,i}$ in (6.63) and bounded by b. The optimal first-order MSE reads

$$nMSE(S_n, Q_n) = v_{id,0}^2 + n\left(\frac{r}{\sqrt{n}} L_{c,i}(0)\right)^2 = v_{id,0}^2 + r^2 L_{c,i}^2(0) = v_{id,0}^2 + r^2 \left(\int \psi_t(X_i) d\Delta_i\right)^2 \Big|_{t=0}.$$

The question for the least favorable deviation now yields in maximizing the last summand, which is, for example, done right away by treating every X_i in **the same** "bad" way in order to achieve a maximum bias $L_{c,i} = b$ for every i – and this, for instance, happens in the i.i.d. case, dropping the index i at Δ_i. □

6.6 Illustration for $F = \mathcal{N}(0,1)$

In this subsection we illustrate the results for the total variation case in the symmetric setup $F = \mathcal{N}(0,1)$ and $\psi \in \Psi_2$ of form (6.70). W.l.o.g. we confine ourselves to the case (6.20). Then, according to Proposition 3.3 in [Ruckdeschel (2005b)], it holds for the total variation case as well as for the case of convex contamination:

Proposition 6.23. *For $F = \mathcal{N}(0,1)$ and ψ an IC of Hampel–form, assumptions (bmi) to (C) are in force; in particular the bound in (Vb) holds even exponentially.*

Proof. Section 7.3: "Proof to Proposition 3.3 and Remark 3.4" in [Ruckdeschel (2005b)]. □

Proposition 6.24. *For $F = \mathcal{N}(0,1)$, ψ to be an IC, $A = (2\Phi(g) - 1)^{-1}$. For $(* = v)$ we get, with $\Phi(x)$ the c.d.f. of $\mathcal{N}(0,1)$ and $\varphi(x)$ its density*

$$v_{c,0} = 0, \quad v_{id,1} = 0, \quad v_{c,2} = 0, \quad l_{c,1} = 0, \quad l_{id,2} = 0, \quad l_{c,3} = 0 \tag{6.102}$$

For $g \in (0, \infty)$, we get

$$
\begin{aligned}
l_{c,2} &= \frac{2g^3\varphi(g)}{2\Phi(g)-1} = g^2 \cdot l_{id,3} && (6.103)\\
l_{id,3} &= \frac{2g\varphi(g)}{(2\Phi(g)-1)} && (6.104)\\
v_{id,0}^2 &= 2b^2(1-\Phi(g)) + A(1-2b\varphi(g)) && (6.105)\\
v_{c,1} &= \frac{-g^3\varphi(g)A^2}{v_{id,0}^2} = \frac{-g^3\varphi(g)}{2g^2(1-\Phi(g)) + 2\Phi(g) - 1 - 2g\varphi(g)} && (6.106)\\
v_{id,2} &= \frac{6\Phi(g) - 4\Phi(g)^2 - 2 - 2g\varphi(g)}{2g^2(1-\Phi(g)) + 2\Phi(g) - 1 - 2g\varphi(g)} && (6.107)\\
\rho_1 &= \frac{3A^3\left(1 - 2\Phi(g) + 2g\varphi(g)\right)}{v_{id,0}^3} + 3v_{id,0}^{-1} && (6.108)
\end{aligned}
$$

where $b = g \cdot A$ and all the $\cdot_{id,\cdot}$-coefficients are the same as in the convex contamination case.

Proof. For F symmetric Corollary 6.19 is in force and by item b) (ibid.) we have

$$v_{c,0} = v_{id,1} = v_{c,2} = l_{c,1} = l_{id,2} = l_{c,3} = 0$$

For $l_{c,2}$ we get:

$$
\begin{aligned}
L_c(t) &= \mathbb{E}_\Delta \psi_t = \int A(-g \vee x - t \wedge g) d\Delta \\
&= A\left[\int_{-\infty}^{-g+t} (-g) d\Delta + \int_{-g+t}^{g+t} (x-t) d\Delta + \int_{g+t}^{\infty} g d\Delta\right] \\
&= A\left[\int_{-\infty}^{-g+t} (-g)^2 \varphi(x) dx + \int_{g+t}^{\infty} g^2 \varphi dx\right] \\
&= A[g^2(\Phi(t-g) - \Phi(t-\infty)) + g^2(\Phi(t+\infty) - \Phi(t+g))] \\
&= Ac^2[1 + \Phi(t-g) - \Phi(t+g)]
\end{aligned}
$$

$$
\begin{aligned}
l_{c,2}/A &= \left(\frac{\partial}{\partial t^2} L_c(t)\right)\Big|_{t=0} = \left(\frac{\partial}{\partial t}[0 + g^2(\varphi(t-g) - \varphi(t+g))]\right)\Big|_{t=0} \\
&= \left(g^2(\varphi'(t-g) - \varphi'(t+g))\right)\Big|_{t=0} \\
&= g^2(\varphi'(-g) - \varphi'(g)) = 2g^3\varphi(g)
\end{aligned}
$$

Hence $l_{c,2} = \frac{2g^3\varphi(g)}{2\Phi(g)-1} = g^2 l_{id,3}$. A similar calculation has to be done for $v_{c,1}$:

$$\begin{aligned} V_c(t) &= \int A^2((x-t)^2 \wedge g^2)d\Delta \\ &= A^2\left[\int_{-\infty+t}^{-g+t} g^2 d\Delta + \int_{-g+t}^{g+t}(x-t)^2 d\Delta + \int_{g+t}^{\infty+t} g^2 d\Delta\right] \\ &= A^2\left[\int_{-\infty+t}^{-g+t} -g^3\varphi(x)dx + \int_{g+t}^{\infty+t} g^3\varphi dx\right] \\ &= A^2[-g^3(\Phi(t-g)-\Phi(t-\infty)) + g^3(\Phi(t+\infty)-\Phi(t+g))] \\ &= Ac^3[1-\Phi(t-g)-\Phi(t+g)] \end{aligned}$$

$$\begin{aligned} V_{c,1}/A^2 &= \left(\frac{\partial}{\partial t}V_c(t)\right)\Big|_{t=0} = \big(g^3(-\varphi(t-g)-\varphi(t+g))]\big)\Big|_{t=0} \\ &= g^3(-\varphi(-g)-\varphi(g)) = -2g^3\varphi(g) \end{aligned}$$

Hence, with

$$V_{id}(0) = v_{id,0}^2 = 2A^2g^2(1-\Phi(g)) + A^2(1-2g\varphi(g))$$

we get

$$v_{c,1} = \frac{V_{c,1}}{2v_{id,0}^2} = \frac{-g^3\varphi(g)A}{2Ag^2(1-\Phi(g)) + 1 - 2Ag\varphi(g)}$$

The $\cdot_{id,\cdot}$-coefficients are the same as in the convex contamination case and can be read of from Remark 6.25.

For $\rho_1 =: \rho_{re,1}$ we recall the decomposition scheme $\cdot_{re,i} = \cdot_{id,i} + \frac{r}{\sqrt{n}} \cdot \cdot_{c,i}$. As ρ_1 only appears in the A_2-term, it belongs to order $1/n$ and so does the first summand in the decomposition $\rho_{id,1}$; but $\rho_{c,1}$ is no part of the A_2-term as it belongs to order $n^{-3/2}$. Hence, up to order $o(1/n)$ we can identify $\rho_1 = \rho_{id,1}$, which can be read of from Remark 6.25, too. □

We add the result for the convex contaminated case derived in [Ruckdeschel (2005b)] as Remark 3.6:

Remark 6.25. *For η_c to be an IC, $A_c = (2\Phi(c)-1)^{-1}$. As to the terms from (D) we get, with $\Phi(x)$ the c.d.f. of $\mathcal{N}(0,1)$ and $\varphi(x)$ its density*

$$l_2 = 0, \qquad \tilde{v}_1 = 0, \qquad \rho_0 = 0 \tag{6.109}$$

For $c \in (0, \infty)$, we get

$$
\begin{aligned}
l_3 &= 2c\varphi(c)/(2\Phi(c) - 1) && (6.110)\\
v_0^2 &= 2b^2(1 - \Phi(c)) + A_c(1 - 2b\varphi(c)) && (6.111)\\
\tilde{v}_2 &= \frac{6\Phi(c) - 4\Phi(c)^2 - 2 - 2c\varphi(c)}{2c^2(1 - \Phi(c)) + 2\Phi(c) - 1 - 2c\varphi(c)} && (6.112)\\
\rho_1 &= \frac{3A_c^3\,(1 - 2\Phi(c) + 2c\varphi(c))}{v_0^3} + 3v_0^{-1} && (6.113)\\
\kappa_0 &= \frac{2c^4\,(1 - \Phi(c)) - 2c(c^2 + 3)\varphi(c) + 3(2\Phi(c) - 1)}{[2c^2\,(1 - \Phi(c)) + 2\Phi(c) - 1 - 2c\varphi(c)]^2} - 3 && (6.114)
\end{aligned}
$$

For $c \downarrow 0$, $l_3 = 1$, $v_0^2 = \frac{\pi}{2}$, $\tilde{v}_2 = -\frac{2}{\pi}$, $\rho_1 = 2\sqrt{\frac{2}{\pi}}$, $\kappa_0 = -2$, and, formally, for $c \uparrow \infty$, $l_3 = 0$, $v_0 = 1$, $\tilde{v}_2 = 0$, $\rho_1 = 0$, $\kappa_0 = 0$.

Chapter 7

Numerical investigation of the Higher Order MSE

Using the derived coefficients for the symmetric setup we examine the behavior of the exact asymptotic MSE and compare the results of first, second and third order asymptotics. We confine ourselves to the symmetric case discussed in section 6.3. Thus Assumptions 6.18 and Corollary 6.19 are in force. Furthermore, we specialize on $F = \mathcal{N}(0,1)$, which is used throughout section 6.6 for illustration. Thus we use the terms calculated in Remark 6.25 for the convex contaminated and the terms calculated in Proposition 6.24 for the total variation case.

7.1 Convex Contamination

P. Ruckdeschel already examined the convex contamination case in [Ruckdeschel (2005b)] and concluded that "under symmetry and for large enough n, the maximal MSE on $\tilde{\mathcal{Q}}_n$ is always underestimated by first-order asymptotics". Indeed, this can be read of from the higher order terms in the following remark, appearing as Remark 3.6 (e) in [Ruckdeschel (2005b)]:

Remark 7.1. *Let Q_n^0 be any distribution in $\tilde{\mathcal{Q}}_n$ attaining maximal risk in Theorem 6.4. Under symmetry or more specifically if $l_2 = v_1 = \rho_0 = 0$, (6.18) becomes*

$$n\,\mathbb{E}_{Q_n^0}[S_n^2] = \left(r^2b^2 + {v_0}^2\right)\left(1 + \frac{r}{\sqrt{n}}\right) + \frac{r}{\sqrt{n}}\left(b^2(1+r^2)\right) + O(n^{-1}) \tag{7.1}$$

Thus under symmetry and for large enough n, **the maximal MSE on $\tilde{\mathcal{Q}}_n$ is always underestimated by first-order asymptotics!**

We show this fact in a little simulation study and use the SWEAVE package by F. Leisch, confer appendix E.3 for further details.

With $n = 100$ we set $c = 1.0$ and do our runs for the radii $r = 0.2$, $r = 0.5$ and $r = 1.0$.

Radius	first-order	second-order	third-order
$r = 0.2$	1.330740	1.392818	1.395806
$r = 0.5$	1.688780	1.879779	1.906810
$r = 1.0$	2.967496	3.605236	3.761374

7.2 Total Variation

7.2.1 Numerical results

With the same setup as in the convex contamination case we get for total variation

Radius	first-order	second-order	third-order
$r = 0.2$	1.330740	1.330740	1.329567
$r = 0.5$	1.688780	1.688780	1.687389
$r = 1.0$	2.967496	2.967496	2.975421

Just looking at the numerical results it seems that in contrast to the convex contamination case, the total variation setup leads to the conjecture that for small radii "under symmetry and for large enough n, the maximal MSE on $\tilde{\mathcal{Q}}_n$ is overestimated slightly by first-order (and second-order) asymptotics", but as the result for $r = 1.0$ indicates, the situation seems to chance to underestimation for the radius increasing.

7.2.2 Dependence on g and r

We take a look at the A_2-term in the second-order MSE:

$$\begin{aligned} A_2 \;=\; & \frac{1}{n}\Big[(\frac{1}{3}l_{id,3}b^4 \pm l_{c,2}b^3)r^4 + (\pm(4v_{c,1} + 3l_{c,2})v_{id,0}^2 b \\ & +((2l_{id,3} + 3v_{id,2} - 6)v_{id,0}^2 + 3)b^2)r^2 + (l_{id,3} + 3v_{id,2})v_{id,0}^4 \\ & +\frac{2}{3}\rho_1 v_{id,0}^3\Big] \end{aligned} \tag{7.2}$$

with the coefficients from 6.24 and $b = g \cdot A$. It holds

Coefficient	value for $g = 1.0$	sign for all $g > 0$
A	1.46	+
$l_{c,2}$	0.71	+
$l_{id,3}$	0.71	+
$v_{id,0}$	1.11	+
$v_{c,1}$	-0.47	-
$v_{id,2}$	-0.52	-
$\rho_{re,1}$	1.24	+

Remark 7.2. *We confine ourselves to the "+"-case of A_2, as according to the table above $b \cdot l_{c,2}|_{(6.20)}(b) = -b \cdot l_{c,2}|_{(6.20)}(-b) > 0$, for instance.*

With r small, $r < 1$ as in section 7.2, for example, the summands of order $O(r^4)$ and $O(r^2)$ in 7.3 are small. Calculating the remaining terms of order $O(r^0)$ for $g = 1.0$ we see that the main negative part stems from the summand $(2l_{id,3} + 3v_{id,2})v_{id,0}^2$, because the coefficient $v_{id,2}$ is negative.
But this does not have to hold for radii approaching 1. Hence, we suppose that the situation changes, when the sample modification by total variation gets to a specific amount, large enough to enable the terms of higher order in r to establish their influence on the MSE.

In this sense we evaluate $A_2(g, r)$ for $r \in [0; 1]$ and $c \in \{1.0;\ 1.6;\ 2.0\}$ and take a look at the plots in figure 7.1.

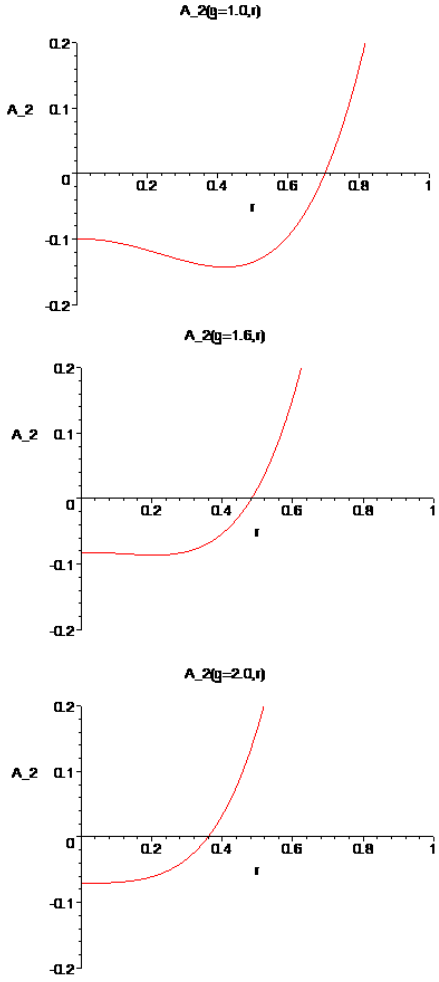

Figure 7.1: Numerical behavior of $A_2(r)$ for $g = 1.0$ (top), $g = 1.6$ (middle) and $g = 2.0$ (bottom).

First of all we can clearly see, that for small radii, we always get a negative contribution of the A_2-term to the MSE, indeed. Hence we get an overestimation of the MSE. But the situation changes by increasing (both) the radius r and the clipping height.There is always

a specific radius (depending on the clipping height), where the A_2-term gets positive. We then have to join the conclusion given in the convex-contaminated case by P. Ruckdeschel, i.e. the MSE suddenly is underestimated.
We might perhaps explain this circumstance with the fact that we can see a total variation neighborhood as the union of two convex contamination balls, as we have illustrated in (3.20). For r large enough "the convex contamination character of the total variation neighborhood system breaks through". But for small r it is not visible.
In other words, there is always a radius r'_g beyond which we cannot apply, or achieve, respectively, the least favorable deviation as we do in chapter 6, in order to get the asymptotic expansion of the MSE. The result is an MSE that turns more and more to infinity with each additional "bad member" of the sample.

Remark 7.3. *For $F = \mathcal{N}(0,1)$, n large enough but finite, we get numerically that*

a) for radii r_g small enough, like $r_g < 0.5$ with $g < 1.6$, for example, the maximal MSE on $\tilde{\mathcal{Q}}_n$ is overestimated by first-order and second-order asymptotics!

b) for radii r_g large enough, like $r_g > 0.5$ with $g > 1.6$, for example, the maximal MSE on $\tilde{\mathcal{Q}}_n$ is underestimated by first-order and second-order asymptotics!

Remark 7.4. *In a finite setup there is at least in the symmetric case $F = \mathcal{N}(0,1)$ a difference in the behavior of the asymptotic maxMSE on convex contaminated versus total variation neighborhoods.*

Chapter 8

Generation of least favorable deviations in total variation for finite sample

In Theorem 6.13 we declared a least favorable deviation $P_n^{di} = (P_n^{di})^+ + (P_n^{di})^-$ with $(P_n^{di})^+$ and $(P_n^{di})^-$ defined as in (6.20) or (6.21), respectively. However, in the finite scenario with original sample $(x_1, \ldots, x_n) \overset{\text{i.i.d.}}{\sim} P_n^{id}$ to be manipulated by the signed measure $\Delta_{(i)}^n$ as defined in (3.18) and Lemma 6.21, respectively, the least favorable deviation may not be possible. This means that we have to find and declare a suitable mechanism explaining the effect of Δ for every finite sample according to previous given conditions. For example, the amount of observations to be manipulated has to be determined and guaranteed (in probability) as well as the bound on the x_i for having maximum influence on the mean squared error according to the value of $\psi(x_i)$ with ψ a influence curve satisfying certain assumptions.
Furthermore, we settle on the symmetric case for $F = P^{id}$ symmetric on the Borel set $\mathbb{B}$ and show that for a certain kind of manipulation mechanism we gain the result of Corollary 6.19 even in the finite setup up to suitable order.
In this sense we first carry out a reordering of the sample by conditioning with respect to the arrangement. The influence curve ψ is seen as monotone and - in the symmetric case - as odd. Actually we confine ourselves to influence curves of Hampel-type form, at least attaining their maximum for $|x| > c_n$ with a general increasing sequence c_n initially. The amount k of manipulable observations is given by a random variable K with first moment $\mathbb{E}K = r\sqrt{n}$ chosen to satisfy the requirement of staying in a total variation ball $B_v(F, r/\sqrt{n})$. The second moment $\mathrm{Var}K = \frac{1}{2}r\sqrt{n}$ is a result of a deeper investigation of all terms in the expansion of the MSE given by a k-step approach. By ordering the sample the observations become (weakly) correlated, however, confer Proposition 8.16 and Theorem 8.17. But in in Theorem 8.20 we can show that under certain assumptions to choose the correlation vanishes.
Without application of further symmetry arguments we are confronted with the common law of the k- and $n-k+1$-quantile $X_{[k:n]}$ and $X_{[n-k+1:n]}$, which lead to order statistics. But the integrals to be evaluated in this setup show up to be very hard to handle, so we just give a short impression of these circumstances in subsection 8.5.1 and make use of

a symmetry argument loosely inspired by the reflection principle known from elementary stochastic. By consideration of several samples $\{x_1, \ldots, x_n\}_j \overset{\text{i.i.d.}}{\sim} F,\ j \in \mathbb{N}$, at once, we are able to neglect the difference between the lower and upper k-quantile.
It shows up that we only get access to the result of Corollary 6.19 in the finite context if we require the finite sample to attain the minimum and maximum of the given influence curve ψ with a certain probability already. Depending on this probability we derive a lower bound on the sample length n in Theorem 8.20, after having conjectured the existence of a such condition by preceding numerical investigations.
Finally, we give a restrictive condition on the distribution of K in assumption 8.21 (PK) that guarantees[1] the desired realization of $X_{[k:n]}$ beyond a now concrete bound c_n. The bound c_n is explicitly calculated for $F = \mathcal{N}(0,1)$ in Proposition 8.24 and at last suitable four-point distributions for K are given in section 8.9.3 that satisfy all the previous claimed conditions.

8.1 Division of the support

In order to generate least favorable deviations we assume that there exist intervals, where the influence curve ψ (almost) attains its minimum and maximum. As a preparation we begin with a partition of the real line:

Notation 8.1. *Let $c_n \geq 0$ be an increasing sequence. We denote*

$$\begin{aligned} I &:= \ ]-\infty,\ -c_n[, \\ II &:= \ [-c_n,\ c_n], \\ III &:= \ ]c_n,\ \infty[\end{aligned}$$

Furthermore we make the assumption that for n large enough ψ does not differ much from the asymptotic optimal influence curve described in (4.17).

Assumption 8.2. *In addition to Assumption 6.9 (bmi) we assume:*

(o) ψ is an odd function, i.e. $\psi(-x) = -\psi(x)$.

(F) ψ is of form

$$\psi(x_i) = \begin{cases} -b/2 + o(n^{-1}) & \text{for } x_i \in I \\ A\Lambda(x_i) = Ax_i & \text{for } x_i \in II \\ b/2 - o(n^{-1}) & \text{for } x_i \in III \end{cases}$$

(Z) $F(\psi < -\frac{b}{2}) = F(\psi > \frac{b}{2}) = 0$

[1] Theorem 8.22 shows the probability of exceeding the bound c_n negligible exponentially by assumption (PK).

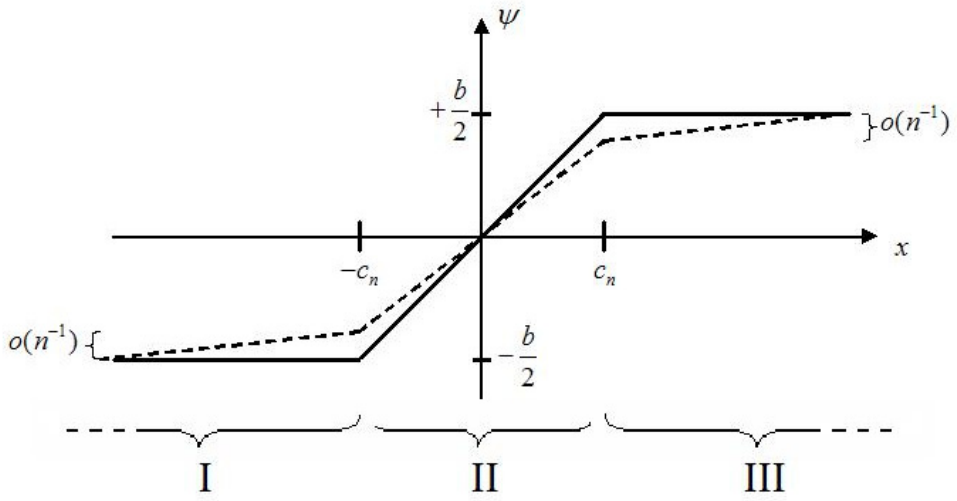

Figure 8.1: The considered IC with the divided support.

Remark 8.3. *a) The sequence c_n in notation 8.1 will show up to be of order $O(r/\sqrt{n})$.*

b) For $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} F$ and Q_n as defined in (3.19) we have the decomposition

$$dQ_n = dF - (dQ_n - dF)_- + (dQ_n - dF)_+$$

For the case $Q_n = Q_n^-$ with signed measures as in (6.20)*, for example, we can identify*

$$\begin{aligned} I &= \{dQ_n < dF\} \\ II &= \{dQ_n = dF\} \\ III &= \{dQ_n > dF\} \end{aligned}$$

c) Unless the Lagrangian multiplier A is calculated explicitly (conf. (8.23)*, for instance), we set $A := 1$. This improves readability by merely neglecting a multiplicative constant.*

8.2 Conditioning w.r.t. the arrangement of the sample

Now let $(X_i)_{i \le n} \sim F_n$ be a random vector and K a fixed number with $K \sim P(n, r_n)$, P an arbitrary measure initially. Without loss of generality the (worst case) signed measure Δ moves mass from I to III. To detect, which observations to choose for modification we now order the sample according to $\psi(x_i)$. Doing this, we loose the stochastic independence of the X_i.

Remark 8.4. *By assumption 6.9 (bmi) the influence function ψ is monotone. Thus ordering the sample according to $\psi(x_i)$ is equivalent to a simple ordering of the sample itself.*

But even by loosing independence we stay in the scenario generating a sample from $B_v(F, r/\sqrt{n})$ if we choose a distribution for K with $\mathbb{E}K = r\sqrt{n}$:

Lemma 8.5. *Let $(X_i)_{i\le n} \sim F_n$ and $K \sim P(n, r_n)$ with radius $r_n = \frac{r}{\sqrt{n}}$ and $\mathbb{E}K = r\sqrt{n}$. Then for the ordered sample $x_{(1)}, \ldots, x_{(n)}$ it still holds that $Q_n = B_v(F, r/\sqrt{n})$.*

Proof. The decision whether X_i will be modified or not, is invariant concerning permutation[2]. Hence we get by symmetry that

$$P(X_i \text{ modified},\ K = k) = \frac{k}{n} \tag{8.1}$$

and as K is stochastically independent from the rest

$$\mathbb{E}(X_i \text{ modified },\ K) = \mathbb{E}(X_i \text{ modified }) \cdot \mathbb{E}K = \frac{1}{n} \cdot \mathbb{E}K = \frac{r\sqrt{n}}{n} = \frac{r}{\sqrt{n}}$$

Additionally, the total variation distance does not change as Δ does not act on the interval II. So

$$d_v(F, Q) = \frac{1}{2} \int_{I \cup II \cup III} |dF - dQ| = \frac{1}{2} \left(\frac{r}{\sqrt{n}} + 0 + \frac{r}{\sqrt{n}} \right) = \frac{r}{\sqrt{n}}$$

□

8.3 The mechanism of modification

As already mentioned above, we order the sample $(X_i)_{i\le n} \sim F_n$ in order to detect which observations to choose for modification. Without loss of generality the (worst case) signed measure Δ moves mass from I to III. Therefore we define our explicit mechanism of total variation modification as to count of the k smallest observations and change their sign.

We now introduce new intervals $I^\natural \subset I$, $II^\natural$ and $III^\natural \subset III$, which are motivated by the finite setup and represent the actual partition of the real line by the realizations x_i of the X_i.

Notation 8.6. *Let k be the actual amount of substituted members in the sample $x_1, \ldots, x_n$. We denote*

$$\begin{aligned} I^\natural &:= [x_{(1)}, x_{(k)}], \\ III^\natural &:= [-x_{(k)}, \max(-x_{(1)}, x_{(n)})], \\ II^\natural &:= \mathbb{R} \setminus (I^\natural \cup III^\natural) \end{aligned}$$

[2]We refer in this context of so called "Exchangeability" to chapter 2.7 of [Krengel (1991)] or part I of [Aldous et al. (1985)], respectively.

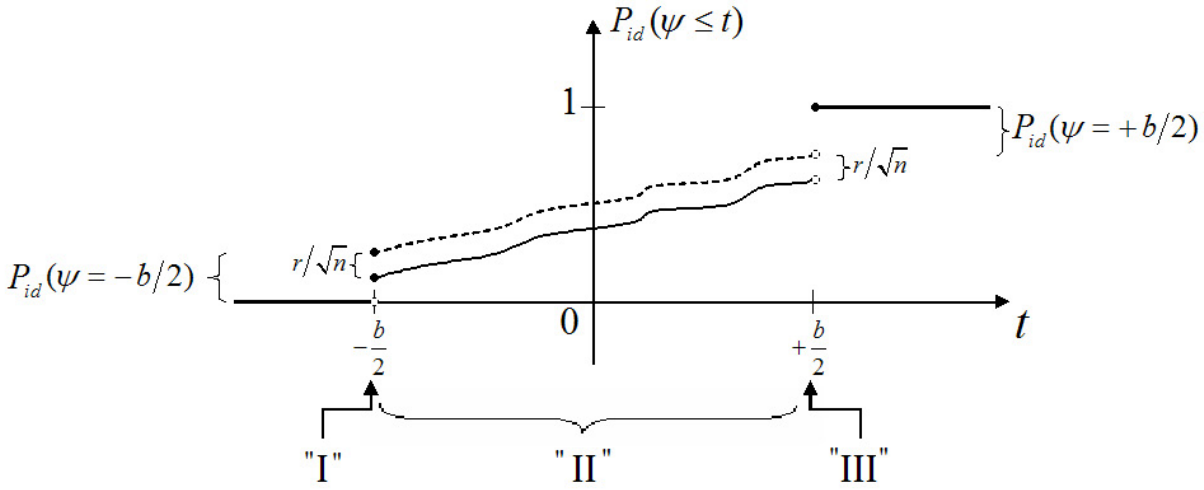

Figure 8.2: Illustration of the modified situation by total variation. The declarations "I", "II" and "III" denote the corresponding codomain of ψ for the intervals I,II and III. By substraction of $r/\sqrt{n}$ at "I" and "III" (or in I and III, respectively) the monotone graph in II is only shifted downwards for the density not changing; identify: $P_{id} = F$.

By changing sign the observations $x_i \le x_{(k)}$ fall into interval $III^\natural$. But there may already be observations bigger than $-x_{(k)}$. We count those members of the sample by the number K', which is identically distributed to K.

Remark 8.7. *a) $x_{(k)}$ is the k-quantile of the c.d.f. of the X_i, i.e. $x_{(k)} = X_{[k:n]}$*

b) Under symmetry of F it is obvious that $K \approx K'$. But only if the c.d.f. of X_i is exactly symmetric, then $k = k'$, i.e. $rk(-x_{(k)}) = n - 2k + 1$.

We confine ourselves to the case of no ties. Furthermore, there is $k \neq k'$, generally. Hence to the right of $-x_{(k)}$ there are $k - 1 + k'$ elements.

As we want to confirm Corollary 6.19 for finite n we have to show that the discrepancy between the stochastically independent and the dependent case is only of order $o(1/\sqrt{n})$, i.e. we have to compare the situations

(A) $Q_n = \bigotimes_{i=1}^n \left(dF + \frac{r}{\sqrt{n}} d\Delta\right)$

and

(B) Q_n the measure resulting, when the K smallest observations (concerning ψ) under F in I are substituted by observations in III by changing sign.

As an illustration for the difference of the two situations we get the different probabilities

$$P(X_i \text{ and } X_j \text{ modified}|K, (A)) = \left(\frac{K}{n}\right)^2,$$

and

$$P(X_i \text{ and } X_j \text{ modified}|K,(B)) = P(X_i \dots) \cdot P_{X_i}(X_j \dots) = \frac{K}{n} \cdot \frac{K-1}{n-1}.$$

8.4 Two-step approach

We recall that $nMSE(S_n) = n\mathbb{E}(S_n - \theta)^2$. It is appropriate for our purpose of investigation to use a Two-step approach, a generalized version of the concept of an one-step estimator, a procedure first used in [Bickel (1975)] in the context of Huber's M-estimators.

The one-step estimator $S_n^{(1)}$ to some starting estimator θ_0 is given by

$$S_n^{(1)} = \theta_0 + \frac{1}{n}\sum_{i=1}^{n} \psi_{\theta_0}(x_i)$$

and analogously, as already sketched in (4.43), a k-step estimator for $k \geq 1$ reads

$$S_n^{(k)} = S_n^{(k-1)} + \frac{1}{n}\sum_{i=1}^{n} \psi_{S_n^{(k-1)}}(x_i)$$

Notation 8.8. *We write*

$$\begin{aligned} \overline{\psi} &:= \frac{1}{n}\sum \psi, \\ d\theta_0 &:= \theta_0 - \theta, \\ \psi^0 &:= \psi - \mathbb{E}\psi = \psi - L_{\text{re}} \end{aligned}$$

with starting estimator θ_0 the arithmetic mean of the sample.

Lemma 8.9. *Let F be symmetric and ψ an IC according to Assumption 8.2. Then it holds for a two-step estimator $S_n^{(2)}$ that*

$$nMSE(S_n^{(2)}) = n\mathbb{E}(S_n^{(2)} - \theta)^2 = n\mathbb{E}A_0^2 + 2n\mathbb{E}A_0A_1 + o(n^{-1/2}), \tag{8.2}$$

with $A_0 := \overline{\psi}_\theta^0 + \frac{r}{\sqrt{n}} l_{c,0}$ and $A_1 := \overline{\dot{\psi}_\theta^0}(\overline{\psi}_\theta^0 + \frac{r}{\sqrt{n}} l_{c,0}) + \frac{r}{\sqrt{n}} l_{c,2} \frac{d\theta_1^2}{2}$.

Proof. We apply a Taylor expansion to ψ_{θ_0} and proceed to the first Newton-Raphson step. Accessorily we specialize on the symmetric case, i.e. $l_{re,1} = -1$ and $l_{re,2} = \frac{r}{\sqrt{n}} l_{c,2}$.

$$\begin{aligned} d\theta_1 &= S_n^{(1)} - \theta = d\theta_0 + \overline{\psi}_{\theta_0} \\ &= d\theta_0 + \overline{\psi}_\theta + \overline{\dot{\psi}_\theta} d\theta_0 + \overline{\ddot{\psi}_\theta} \frac{d\theta_0^2}{2} + \dots \\ &= d\theta_0 + (\overline{\psi}_\theta^0 + l_{re,0}) + (\overline{\dot{\psi}_\theta^0} + l_{re,1}) d\theta_0 + (\overline{\ddot{\psi}_\theta^0} + l_{re,2}) \frac{d\theta_0^2}{2} + \dots \\ &= d\theta_0 + \overline{\psi}_\theta^0 + \frac{r}{\sqrt{n}} l_{c,0} + \overline{\dot{\psi}_\theta^0} d\theta_0 - d\theta_0 + \frac{r}{\sqrt{n}} l_{c,2} \frac{d\theta_0^2}{2} + \dots \\ &= \overline{\psi}_\theta^0 + \frac{r}{\sqrt{n}} l_{c,0} + \overline{\dot{\psi}_\theta^0} d\theta_0 + \frac{r}{\sqrt{n}} l_{c,2} \frac{d\theta_0^2}{2} + \dots \end{aligned}$$

Hence

$$\begin{aligned} d\theta_2 &= \overline{\psi}^0_\theta + \frac{r}{\sqrt{n}} l_{c,0} + \overline{\dot{\psi}^0_\theta} d\theta_1 + \frac{r}{\sqrt{n}} l_{c,2} \frac{d\theta_1^2}{2} + \dots \\ &= \overline{\psi}^0_\theta + \frac{r}{\sqrt{n}} l_{c,0} + \overline{\dot{\psi}^0_\theta} (\overline{\psi}^0_\theta + \frac{r}{\sqrt{n}} l_{c,0}) + \frac{r}{\sqrt{n}} l_{c,2} \frac{d\theta_1^2}{2} + \dots \\ &= B_0 + B_1 + O(n^{-1/2}) \end{aligned}$$

with $B_0 := \overline{\psi}^0_\theta + \frac{r}{\sqrt{n}} l_{c,0}$ and $B_1 := \overline{\dot{\psi}^0_\theta} (\overline{\psi^0_\theta} + \frac{r}{\sqrt{n}} l_{c,0}) + \frac{r}{\sqrt{n}} l_{c,2} \frac{d\theta_1^2}{2}$. We note that $B_0 = O(n^{-1/2})$ and $B_1 = O(n^{-1})$. Then

$$\begin{aligned} (d\theta_2)^2 &= (B_0 + B_1 + O(n^{-1/2}))^2 \\ &= B_0^2 + 2B_0B_1 + B_1^2 + 2B_0 \cdot O(n^{-1/2}) + 2B_1 \cdot O(n^{-1/2}) + (O(n^{-1/2}))^2 \\ &= B_0^2 + 2B_0B_1 + O(n^{-2}) + O(n^{-1}) + O(n^{-3/2}) + O(n^{-1}) \\ &= B_0^2 + 2B_0B_1 + o(n^{-1/2}) \end{aligned}$$

Hence

$$n\mathbb{E}(d\theta_2)^2 = n\mathbb{E}B_0^2 + 2n\mathbb{E}B_0B_1 + o(n^{-1/2}).$$

□

Investigation of the term B_0^2 :

We now split the powers of ψ:

$$\begin{aligned} n\mathbb{E}B_0^2 &= \mathbb{E}(n\frac{1}{n^2} \sum_{i,j}^{n} \psi^0_\theta(x_i)\psi^0_\theta(x_j) + 2r\sqrt{n} l_{c,0} \overline{\psi}^0_\theta(x_i)) + r^2 l_{c,0}^2 \\ &= \mathbb{E}(\psi^0_\theta(x_1))^2 + (n-1)\mathbb{E}\psi^0_\theta(x_1)\psi^0_\theta(x_2) + 0 + r^2 l_{c,0}^2 \\ &= B_{0,1} + B_{0,2} \end{aligned}$$

with $B_{0,1} := \mathbb{E}(\psi^0_\theta(x_1))^2 + r^2 l_{c,0}^2$ and $B_{0,2} := (n-1)\mathbb{E}\psi^0_\theta(x_1)\psi^0_\theta(x_2)$. The term $B_{0,1}$ is the same in the independent and the dependent case and doesn't contribute to the error term. Hence we confine ourselves to term $B_{0,2}$.

We take a look at the following tableau:

III_2		**X**	
II_2	**X**		**X**
I_2		**X**	
	I_1	II_1	III_1

Figure 8.3: Grid of the two dimensional support with marked areas, confer Lemma 8.10.

Lemma 8.10. *In the areas marked with* **X** *in figure 8.3 it holds that* $B_{0,2} = 0$.

Proof. We use the concept of simple perturbations (conf. (2.35)) by plugging in

$$dQ_n(x_i) = dF(x_i) + \frac{r}{\sqrt{n}} d\Delta(x_i) = (1 + \frac{r}{\sqrt{n}} q(x_i)) dF(x_i), \tag{8.3}$$

with tangents $q \in \mathcal{G}_v(\theta)$ as defined in (2.33).
Hence

$$\begin{aligned} B_{0,2}/(n-1) &= \mathbb{E}_{Q_n} \psi_\theta^0(x_1)\psi_\theta^0(x_2) \\ &= \mathbb{E}_F \left\{ \psi_\theta^0(x_1)(1 + \frac{r}{\sqrt{n}} q(x_1))\psi_\theta^0(x_2)(1 + \frac{r}{\sqrt{n}} q(x_2)) \right\} \\ &= \mathbb{E}_F \psi_\theta^0(x_1)\psi_\theta^0(x_2) + \frac{r}{\sqrt{n}} \mathbb{E}_F \left\{ \psi_\theta^0(x_1) q(x_2) + \psi_\theta^0(x_2) q(x_1) \right\} + \\ &\quad + \frac{r^2}{n} \mathbb{E}_F \psi_\theta^0(x_1)\psi_\theta^0(x_2) q(x_1) q(x_2) \end{aligned}$$

In the areas marked with **X** all the summands in (8.4) vanish, because either one of the tangents $q(x_i) = 0$ or a symmetry argument applies. □

8.5 General approach via order statistics

We take a closer look at the integral to be evaluated:

$$
\begin{aligned}
& \mathbb{E}\psi_\theta^0(x_1)\psi_\theta^0(x_2) \\
= \;& \mathbb{E}_k[\mathbb{E}_{F(X_1,X_2)}\psi_\theta^0(x_1)\psi_\theta^0(x_2)|K=k] \\
= \;& \mathbb{E}_k[\mathbb{E}_{F(X_1,X_2)}\psi_\theta^0(x_1)\psi_\theta^0(x_2)(\mathbb{I}_{I_1^\natural\times I_2^\natural}+\cdots+\mathbb{I}_{III_1^\natural\times III_2^\natural})|K=k] \\
= \;& \mathbb{E}_k[\mathbb{E}_{F(X_1,X_2)}\psi_\theta^0(x_1)\psi_\theta^0(x_2)(\mathbb{I}_{I_1^\natural\times I_2^\natural}|K=k]+\cdots+ \\
& +\mathbb{E}_k[\mathbb{E}_{F(X_1,X_2)}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\mathbb{I}_{III_1^\natural\times III_2^\natural})|K=k] \\
= \;& \mathbb{E}_k[(\mathbb{E}_{\mathcal{L}(X_1,X_2|K=k,x_1\in I_1^\natural,x_2\in I_2^\natural)}\psi_\theta^0(x_1)\psi_\theta^0(x_2))\cdot P(x_1\in I_1^\natural,x_2\in I_2^\natural|K=k)|K=k]+\cdots+ \\
& +\mathbb{E}_k[(\mathbb{E}_{\mathcal{L}(X_1,X_2|K=k,x_1\in III_1^\natural,x_2\in III_2^\natural)}\psi_\theta^0(x_1)\psi_\theta^0(x_2))\cdot P(x_1\in III_1^\natural,x_2\in III_2^\natural|K=k)|K=k]
\end{aligned}
$$

The out spelt situation shows clearly that we are in need of the common law of X_1 and X_2. We first stay in the case of the unmodified situation, where the observations to be modified are already marked. As an example we spell out the case $x_1 \in I_1^\natural,\ x_2 \in I_2^\natural$:

$$\mathcal{L}(X_1,X_2|x_1\in I_1^\natural,x_2\in I_2^\natural)=\mathcal{L}(X_1,X_2|x_1\le x_{(k)}=X_{[k:n]},x_2\le x_{(k)}=X_{[k:n]})$$

After that we carry out the modification and are confronted with the reordered sample $Y_1,\ldots,Y_n$. The common law reads

$$\mathcal{L}(Y_1,Y_2)=\mathcal{L}\left(-X_1,-X_2|x_1\le x_{(k)}=X_{[k:n]},x_2\le x_{(k)}=X_{[k:n]}\right)$$

8.5.1 Showcase $I^\natural \times III^\natural$

Appendix D provides the techniques for dealing with the common law of two quantiles. We determine the parameters ν_1 and ν_2 in (D.1) just for the case $I^\natural \times III^\natural$:

$$\nu_1 = K \qquad \nu_2 = n-K'+1$$

So, by Lemma D.2 we have the marginal c.d.f.'s

$$
\begin{aligned}
P^Y(s) &:= \sum_{j=K}^{n}\binom{n}{j}F(s)^j\left(1-F(s)\right)^{(n-j)} \\
P^Z(t) &:= \sum_{j=n-K'+1}^{n}\binom{n}{j}F(t)^j\left(1-F(t)\right)^{(n-j)}
\end{aligned}
$$

and the marginal densities

$$
\begin{aligned}
p^Y(s) &:= n\binom{n-1}{K-1}F(s)^{(K-1)}\left(1-F(s)\right)^{n-K}f(s) \\
p^Z(t) &:= n\binom{n-1}{n-K'}F(t)^{n-K'}\left(1-F(t)\right)^{(K'-1)}f(t)
\end{aligned}
$$

As conditional c.d.f. and density, we obtain $P^Z(dt)$–a.e.

$$\begin{aligned} P^{Y|Z=t}(s) &= \mathbb{I}_{\{s>t\}} + \mathbb{I}_{\{s\le t\}} \sum_{j=K}^{n-K'} \binom{n-K'}{j} \big(\frac{F(s)}{F(t)}\big)^j \big(1-\frac{F(s)}{F(t)}\big)^{(n-K'-j)} \\ p^{Y|Z=t}(s) &= \mathbb{I}_{\{s\le t\}} (n-K') \binom{n-K'-1}{K-1} \big(\frac{F(s)}{F(t)}\big)^{(K-1)} \big(1-\frac{F(s)}{F(t)}\big)^{(n-K'-K)} f(s) \end{aligned}$$

Hence by symmetry, i.e. with $\psi_\theta^0(-x) = -\psi_\theta^0(x)$, we get

$$\begin{aligned} & \mathbb{E}_{\mathcal{L}(Y_1=-X_1, Y_2=X_2|K=k, x_1\in I_1^\natural, x_2\in III_2^\natural)} \psi_\theta^0(-x_1)\psi_\theta^0(x_2) = \\ &= \int -\psi_1(s)\psi_2(t) p^{X_1|X_2=t} p^{X_2}(t) ds dt = \\ &= n\binom{n-1}{n-K'} \int \psi_2(t) F(t)^{n-K'} (1-F(t))^{K'-1} f(t)\cdot S(t) dt \end{aligned}$$

with the inner integral

$$\begin{aligned} S(t) &= \int_{-\infty}^t (n-K')(-\psi_1(s)) \binom{n-K'-1}{K-1} \big(\frac{F(s)}{F(t)}\big)^{(K-1)} \big(1-\frac{F(s)}{F(t)}\big)^{(n-K'-K)} f(s) ds \\ &= (b-\eta) \int_{-\infty}^t (n-K') \binom{n-K'-1}{K-1} F_t(s)^{(K-1)} (1-F_t(s))^{(n-K'-K)} f(s) ds \\ &= (b-\eta) \int_{-\infty}^t g_{n,K,K'}(s) ds \end{aligned}$$

where we define $F_t(s) := F(s)/F(t)$ and the integrand

$$g_{n,K,K'}(s) := (n-K') \binom{n-K'-1}{K-1} F_t(s)^{(K-1)} (1-F_t(s))^{(n-K'-K)} f(s)$$

Applying the Stirling approximation (A.23) to the constants, we get

$$\binom{n-K'-1}{K-1} = \Big[\frac{n-K'-1}{K-1}\Big]^{K-1} \Big[\frac{n-K'-1}{n-K'+K}\Big]^{n-K'-K} \gamma_{n,K,K'}$$

with $K-1 \ge 1$ and

$$\gamma_{n,K,K'} := \sqrt{\frac{n-K'-1}{(n-K'-K)(K-1)2\pi}} \, (1+\rho_{m,K,K'}) \tag{8.4}$$

The product $F_t(s)^{K-1}(1-F_t(s))^{n-K'-K}$ suggests an asymptotic decay. So the idea is (similar to the approach in [Ruckdeschel (2005a)]) to expand the integrand of $S(t)$ at the

mode of $F_t(s)^{K-1}(1-F_t(s))^{n-K'-K}$. Therefore we look for the maximum of $F_t(s)^{K-1}(1-F_t(s))^{n-K'-K}$. By differentiation we get

$$\begin{aligned}
& F_t(s)^{(K-1)}(1-F_t(s))^{(n-K'-K)} = \max! \\
\Leftrightarrow\ & (K-1)\log F_t(s) + (n-K'-K)\log(1-F_t(s))\cdot(-1) = \max! \\
\Leftrightarrow\ & \frac{K-1}{F_t(s)} - \frac{n-K'-K}{1-F_t(s)} = 0 \\
\Leftrightarrow\ & K-1-KF_t(s)+F_t(s)-nF_t(s)+K'F_t(s)+KF_t(s) = 0 \\
\Leftrightarrow\ & F_t(s) = \frac{K-1}{n-K'-1}
\end{aligned}$$

Hence

$$F_t(s)^{(K-1)}(1-F_t(s))^{(n-K'-K)} \le \Big(\frac{K-1}{n-K'-1}\Big)^{K-1}\Big(\frac{n-K'-K}{n-K'-1}\Big)^{n-K'-K}$$

with equality iff $t = x_{n,K,K'}$ for

$$x_{n,K,K'} := F^{-1}(\frac{K-1}{n-K'-1})$$

So

$$g_{n,K,K'}(s) = (n-K')f(s)\left(\frac{n-K'-1}{K-1}F_t(s)\right)^{K-1}\left(\frac{n-K'-1}{n-K'-K}(1-F_t(s))\right)^{n-K'-K}\gamma_{n,K,K'}$$

For better readability we define the abbreviation

$$\Delta F_{n,K,K'} := F_t(s) - \frac{K-1}{n-K'-1} = F_t(s) - F(x_{n,K,K'}),$$

Now we see that

$$\begin{aligned}
\frac{n-K'-1}{K-1}F_t(s) &= 1 + \frac{n-K'-1}{K-1}\,\Delta F_{n,K,K'} \\
\frac{n-K'-1}{n-K'-K}(1-F_t(s)) &= 1 - \frac{n-K'-1}{n-K'-K}\,\Delta F_{n,K,K'}
\end{aligned}$$

and hence

$$g_{n,K,K'}(s) = (n-K')\gamma_{n,K,K'}f(s)\left[1+\frac{n-K'-1}{K-1}\,\Delta F_{n,K,K'}\right]^{K-1}\left[1-\frac{n-K'-1}{n-K'-K}\,\Delta F_{n,K,K'}\right]^{n-K'-K}$$

In [Ruckdeschel (2005a)] P. Ruckdeschel did an approach similar to the proof of our main theorem 6.13 in subsection 6.2.2 and fixed some constants $k_1 > 1$ and $k_2 > \sqrt{5/2}$ to split up the proof according to the following tableau that is taken from [Ruckdeschel (2005a)]:

	$K \le k_1 r\sqrt{n}$	$k_1 r\sqrt{n} < K \le n/2$	$K > n/2$
$\lvert s\rvert < k_2\sqrt{\log(n)/n}/f_0$	(I)	(III)	excluded
$\lvert s\rvert \ge k_2\sqrt{\log(n)/n}/f_0$	(II)		

The constant f_0 results from a assumed Taylor expansion of the Lebesgue density f of F about 0 as

$$f(x) = f_0 + f_1 x + \frac{1}{2} f_2 x^2 + O(x^{2+\delta}), \qquad f_0 > 0$$

for some $\delta > 0$.

For cases (II) and (III) it can be shown analogously to [Ruckdeschel (2005a)] that they attribute only terms of order $o(n^{-1})$ to $n\,\mathrm{MSE}(\mathrm{Med}_n)$ and hence can be neglected.

Case (I):

In order to take a look at the problems to face we restrict ourselves here to the case (I), i.e. $k \le k_1\sqrt{n}r$ and $|\frac{K-1}{n-K'-1} - F_t(s)| \le k_2\sqrt{\log(n)/n}$.

First of all we set

$$u := s - x_{n,K,K'}$$

To ease the access to a starting value, we reparametricise K by $K = n - K' + \kappa$ with some constant $\kappa \in \mathbb{Z}$. This gives

$$\frac{K-1}{n-K'-1} = 1 + \frac{\kappa}{n} + \frac{\kappa(K'+1)}{n^2} + o(n^{-3})$$

Thus, to get an approximation to $x_{n,K,K'}$, the $\frac{K-1}{n-K'-1}$–quantile of F, we use the approximation sketched in A.4 and perform the first step of a Newton–procedure to solve

$$G(x) = \frac{K-1}{n-K'-1} - 1 - f_0 x - f_1 x^2/2 - f_2 x^3/6 = 0$$

A starting value $x_0 = (\frac{\kappa}{n} + \frac{\kappa(K+1)}{n^2})/f_0$ is given by the equivalent to (A.15), as

$$\begin{aligned} G(x_0) &= \frac{K-1}{n-K'-1} - 1 - f_0 x_0 - f_1 x_0^2/2 - f_2 x_0^3/6 = \\ &= \frac{\kappa}{n} + \frac{\kappa(K+1)}{n^2} - f_0\left(\frac{\kappa}{f_0 n} + \frac{\kappa(K'+1)}{f_0 n^2}\right) + O(n^{-2}) \\ &= o(n^{-1}) \end{aligned}$$

and see that

$$x_{n,K,K'} = \frac{\kappa}{f_0 n} + \frac{\kappa K' + \kappa}{f_0 n^2} = x_0 \tag{8.5}$$

Obviously $|G(x_{n,K,K'})| = O(\kappa^2 n^{-2}) = |G(x_0)|$. This means that we cannot get better than our x_0 already is.

Additionally,

$$\begin{aligned} f(x_{n,K,K'}) &= f_0 + f_1 x_{n,K,K'} + f_2 x_{n,K,K'}^2 + o\left(\frac{1}{n}\right) = \\ &= f_0 + \frac{f_1\kappa}{f_0 n} + \frac{f_0 f_1 \kappa + f_0 f_1 \kappa K' + f_2 \kappa^2)}{f_0^2 n^2} + o\left(\frac{1}{n}\right) \end{aligned}$$

This implies that in (I), using $|\frac{K-1}{n-K'-1} - F_t(s)| \leq k_2\sqrt{\log(n)/n}$, applying a Taylor expansion to $F_t^{-1}(x)$ at $a = F_t(x_{n,K,K'}) = \frac{K-1}{n-K'-1}$ (i.e. $s = x_{n,K,K'}$) and plugging in that $(F_t^{-1})'(F_t(s)) = (f_t(s))^{-1}$,

$$\begin{aligned} u &= s - x_{n,K,K'} = \\ &= F_t^{-1}(F_t(s)) - F_t^{-1}(F_t(x_{n,K,K'})) = \\ &= (F_t^{-1}(F_t(x_{n,K,K'})) - x_{n,K,K'}) + (F_t^{-1})'(F_t(x_{n,K,K'})) \cdot (F_t(s) - F_t(x_{n,K,K'})) + O(\log(n)/n) \\ &= f_t(x_{n,K,K'})^{-1} \cdot (F_t(t) - F_t(x_{n,K,K'})) + o(\sqrt{\log(n)/n}) = O(\sqrt{\log(n)/n}). \end{aligned}$$

Again, abbreviating $\Delta F_{n,K,K'} := F_t(s) - F_t(x_{n,K,K'})$, and expanding this in a Taylor series around 0, we get

$$\begin{aligned} \Delta F_{n,K,K'} &= F(0) + [f_0 + f_1(u + x_{n,K,K'})/2 + f_2(u + x_{n,K,K'})/6](u + x_{n,K,K'}) - \\ &\quad -F(0) - [f_0 + f_1 x_{n,K,K'}/2 + f_2 x_{n,K,K'}/6]x_{n,K,K'} + o(n^{-3/2}) \\ &= f_0 u + f_1(u^2/2 + u x_{n,K,K'}) + \\ &\quad + f_2(u^3/6 + u x_{n,K,K'}(u + x_{n,K,K'})/2) + o(n^{-3/2}) \\ (\Delta F_{n,K,K'})^2 &= \cdots = (1 + 2\frac{f_1}{f_0}x_{n,K,K'} + (\frac{f_2}{f_0} + \frac{f_1^2}{f_0^2})x_{n,K,K'}^2)u^2 f_0^2 + \\ &\quad + (\frac{f_2}{f_0} + \frac{f_1^2}{f_0^2}x_{n,K,K'})u^3 f_0^2 + o(n^{-1}) \\ (\Delta F_{n,K,K'})^3 &= \cdots = (f_0^3 + 3f_0^2 f_1)u^3 + o(n^{-1}) \end{aligned}$$

and

$$f(t) = f_0 + f_1(u + x_{n,K,K'}) + f_2((u + x_{n,K,K'}))^2/2 + o(n^{-1})$$

The next step is to investigate the constant factors. We use Taylor approximations performed by `MAPLE` for the terms arising by applications of the Stirling formulas of section A.6.

$$(n - K')\sqrt{\frac{n - K' - 1}{(n - K' - K)(K - 1)}} = \frac{-\sqrt{-\kappa}12n}{12\kappa - 1}[1 - \frac{2K' + \kappa}{n} + \frac{9\kappa^2 - 10\kappa}{n^2} + o(n^{-3/2})]$$

$$\frac{(n - K' - 1)^2}{2(K - 1)} + \frac{(n - K' - 1)^2}{2(n - K' - K)} = \frac{n^2}{-2\kappa}[1 + \frac{-\kappa - 2 - 2K'}{n}] + o(n^{-1/2})$$

This might look a little bit odd, but as $\kappa = K + K' - n = O(n^{1/2}) - n < 0$ for $K + K' > 1$ we might rewrite the above equations with the positive $\tilde{\kappa} = -\kappa = n - (K + K')$:

$$(n-K')\sqrt{\frac{n-K'-1}{(n-K'-K)(K-1)}} = \frac{\sqrt{\tilde\kappa}12n}{12\tilde\kappa+1}[1-\frac{2K'-\tilde\kappa}{n}+\frac{9\tilde\kappa^2+10\tilde\kappa}{n^2}+o(n^{-3/2})]$$
$$\frac{(n-K'-1)^2}{2(K-1)}+\frac{(n-K'-1)^2}{2(n-K'-K)} = \frac{n^2}{2\tilde\kappa}[1+\frac{\tilde\kappa-2-2K'}{n}]+o(n^{-1/2})$$

With these terms we get

$$\begin{aligned}
& [1+\frac{n-K'-1}{K-1}\,\Delta F_{n,K,K'}]^{K-1}[1-\frac{n-K'-1}{n-K'-K}\,\Delta F_{n,K,K'}]^{n-K'-K} = \\
= \;& \exp\big\{-(\Delta F_{n,j,k})^2[\frac{(n-K'-1)^2}{2(K-1)}+\frac{(n-K'-1)^2}{2(n-K'-K)}]+O(log(n)/n)\big\} = \\
= \;& \exp\big\{-\frac{f_0^2}{2\tilde\kappa}n^2u^2+O(n)\big\}
\end{aligned}$$

We plug in (8.5) and set

$$\sigma_n^2 := \frac{f_0^2}{\tilde\kappa}n^2, \qquad y := u\sigma_n$$

Hence we get

$$\begin{aligned}
& [1+\frac{n-K'-1}{K-1}\,\Delta F_{n,K,K'}]^{K-1}[1-\frac{n-K'-1}{n-K'-K}\,\Delta F_{n,K,K'}]^{n-K'-K} \\
= \;& \exp(-y^2/2)\,h(y,K,K',n)+o(n^{-1})
\end{aligned}$$

with $h(y,K,K',n)$ some complex polynomial in y, K, K' and n after application of the Taylor expansion $\exp(x) = 1+x+x^2/2+o(x^2)$.

Finally, we define a further abbreviation: $\tilde x_{n,K,K'} := x_{n,K,K'}\sigma_n$. This gives with φ the density of $\mathcal{N}(0,1)$

$$\begin{aligned}
S(t) \;=\; & (b-\eta)\int_{-\infty}^{t} g_{n,K,K'}(s)\mathbb{I}_{\{|\frac{K-1}{n-K'-1}-F_t(s)|\le k_2\sqrt{\log(n)/n}\}}(t)\,ds = \int (c_{n,K,K'}+o(n^{-1}))\times \\
& \times f(s)\,\varphi(y(u(s)))\,h(y(u(s)),K,K',n)\mathbb{I}_{\{|s|\le k_2\sqrt{\frac{\log(n)}{n}}(1+o(n^0))/f_0\}}(s)\,ds
\end{aligned}$$

with a constant $c_{n,K,K'}$ derived from $\gamma_{n,K,K'}$ from (8.4):

$$c_{n,K,K'} := \frac{\sqrt{\tilde\kappa}12n}{12\tilde\kappa+1}[1-\frac{2K'-\tilde\kappa}{n}+\frac{9\tilde\kappa^2+10\tilde\kappa}{n^2}+o(n^{-3/2})]$$

Evaluation of the above integral will lead to the desired explicit expression for the inner integral $S(t)$. Then the explicit expression for the whole integral (8.4) may be calculated in the same way as it was shown for $S(t)$.

8.6 A symmetry argument inspired by the reflection principle

Without further assumptions and mechanisms of symmetrization we have to deal with the two different quantiles K and K' and their common distribution, leading towards an approach by order statistics as shown in the previous section exemplarily. By now we offer a way out of this dilemma using an argument loosely inspired by the reflection principle from the thematic background of stochastic processes and Brownian motion. The reflection principle itself is common knowledge in elementary stochastic and can be found in [Borodin and Saminen (1996)], p. 49, [Durrett (1999)], p. 247, or [Schmitz (1996)], p.276, for example.

According to Theorem 6.13 both situations (6.20) and (6.21) produce a least favorable situation in the asymptotic sense, i.e. with the notation in subsection 6.2.2

$$MSE^- := nMSE(S_n, Q_n^-) = nMSE(S_n, Q_n^+) =: MSE^+ \quad \text{for} \quad n \to \infty$$

In the finite setup we have

$$nMSE(S_n, Q_n^-) \neq nMSE(S_n, Q_n^+)$$

as the MSE depends on the direction of modification. But as we are in the symmetric context, we may conclude that the different terms in MSE^- and MSE^+ cancel out right away. Hence we aggregate both spaces by employing a randomization, i.e. throwing a coin in order to decide which case we are in and go on to the mean of both situations:

$$MSE = \overline{MSE} = \frac{1}{2}MSE^+ + \frac{1}{2}MSE^-$$

This means, by mutual compensation of the two possible cases of "mass transport" we may assume that the quantiles $X_{[K:n]}$ and $X_{[K':n]}$ coincide. Hence, for the sequel we assume

$$K = K'$$

8.6.1 A look at the convex-contaminated case

In the convex-contaminated case we have according to Theorem 6.4:

$$R_n(S_n, r, \varepsilon_0) \;=\; r^2b^2 + {v_0}^2 + \frac{r}{\sqrt{n}}\,A_1 + \frac{1}{n}\,A_2 + o(n^{-1})$$

with

$$A_1 = {v_0}^2\Big(\pm(4\,\tilde{v}_1 + 3\,l_2\,)b + 1\Big) + b^2 + [\,2\,b^2 \pm l_2\,b^3\,]\,r^2$$

and we are in the $-\,[+]$-case depending on whether (6.6) or (6.7) applies.

In the symmetric case we get $l_2 = 0$ and $\tilde{v}_1 = 0$, which gives

$$A_1 = {v_0}^2 + b^2 + 2\,b^2\,r^2 \tag{8.6}$$

Although in contrast to the total variation case the A_1-term does not vanish, again we see that there is no influence on the A_1-term as b only appears squared. So up to the terms of order $1/\sqrt{n}$ we have in analogy to the total variation case:

$$\begin{aligned} A_1^- = A_1^+ \quad &\Rightarrow \quad A_1 = \frac{1}{2}A_1^- + \frac{1}{2}A_1^+ \\ &\Rightarrow \quad MSE = \frac{1}{2}MSE^- + \frac{1}{2}MSE^+ \end{aligned}$$

8.7 Insufficient negligibility

8.7.1 Excluding the $II \times II$-case

Notation 8.11. *For a certain $k < n/2$, $k \in \mathbb{N}$ let*

$$I_n(y;x) := I_n(y;x_1,\ldots,x_n) := \mathbb{I}(x_{[k:n]} \le y \le x_{[(n-k+1):n]}) \tag{8.7}$$

Assumption 8.12. *We assume*

(U) $\psi(x_i)$ only weakly correlated for $x_i \in II$, i.e.

$$n\mathbb{E}\left\{\psi(X_1)\psi(X_2)\mathbb{E}[I_n(X_1,X)I_n(X_2,X)|X_1,X_2]\right\} = o(n^{-1/2})$$

(EK) $\mathbb{E}K = r\sqrt{n}$

(VK) $\mathrm{Var}K = \frac{1}{2}r\sqrt{n}$

Remark 8.13. *Assumption (U) is straightforward and preliminary. But both from literature and from numerical computations a correlation of the $\psi(x_i)$ in the interval II has to be assumed. In [Aldous et al. (1985)] p. 8, for example, it is shown that for a sequence (Z_i) of N square integrable and permutation invariant random variables there is a correlation $\rho(Z_i,Z_j)$, $i \neq j$, namely $\rho \ge \frac{-1}{N-1}$, which would indicate an order of at least $O(n^{-1})$.*

Theorem 8.14. *Let $(X_i)_{i\le n} \sim F$, F symmetric, and $K \sim P$ with $\mathbb{E}_P K = r\sqrt{n}$ and $\mathrm{Var}_P K = \frac{1}{2}r\sqrt{n}$. Then it holds with assumption (U) from 8.12 for the difference of the mean squared error in the compared situations (A) and (B) that*

$$n\left|\mathrm{MSE}(S_n^{(2)}|(B)) - \mathrm{MSE}(S_n^{(2)}|(A))\right| = o(n^{-1/2}). \tag{8.8}$$

Proof. We exclude the case $II \times II$. According to figure 8.3 the remaining interesting cases are summarized in the following table:

Area	Probability	Integrand Value	Quantity
I^2	$\frac{K}{n}\frac{K-1}{n-1}$	$\frac{b^2}{4}$	1
$I \times III$	$\frac{K}{n}\frac{K}{n-1}$	$\frac{b^2}{4}$	2
III^2	$\frac{K}{n}\frac{K-1}{n-1}$	$\frac{b^2}{4}$	1

Integrating out w.r.t. K gives

$$\begin{aligned} B_{0,2} &= (n-1)\frac{b^2}{4}\frac{\mathbb{E}(4K^2-2K)}{n(n-1)} \\ &= \frac{b^2}{4n}[4(\mathbb{E}K)^2 + 4\mathrm{Var}(K) - 2\mathbb{E}K] \\ &= \frac{b^2}{4n}[4r^2 n + 2r\sqrt{n} - 2r\sqrt{n}] \\ &= r^2 b^2 \end{aligned}$$

and this is exactly the same as in the independent case.

Investigation of the term $B_0 B_1$:

$$\begin{aligned} 2n\mathbb{E}B_0B_1 &= \mathbb{E}2n\left(\overline{\psi_\theta^0}(x_i) + l_{re,0}\right)\left[\dot{\overline{\psi_\theta^0}}(x_j)(\overline{\psi_\theta^0}(x_k) + l_{re,0}) + \frac{1}{2}l_{re,2}(\overline{\psi_\theta^0}(x_j) + l_{re,0})^2\right] \\ &= \mathbb{E}n\big(2\overline{\psi_\theta^0}(x_i)\overline{\psi_\theta^0}(x_k)\dot{\overline{\psi_\theta^0}}(x_j) + 4l_{re,0}\overline{\psi_\theta^0}(x_i)\dot{\overline{\psi_\theta^0}}(x_j) + 2l_{re,0}^2\dot{\overline{\psi_\theta^0}}(x_j) + \\ &\quad + l_{re,2}(\overline{\psi_\theta^0}(x_i)\overline{\psi_\theta^0}(x_j)\overline{\psi_\theta^0}(x_k) + 4l_{re,0}\overline{\psi_\theta^0}(x_i)\overline{\psi_\theta^0}(x_j) + 3l_{re,0}^2\overline{\psi_\theta^0}(x_i) + l_{re,0}^3\big) \\ &= T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7 \end{aligned}$$

with

$$T_1 := \frac{2}{n^2}\sum_{i,j,k}^{n}\mathbb{E}\psi_\theta^0(x_i)\psi_\theta^0(x_j)\dot{\psi}_\theta^0(x_k), \tag{8.9}$$

$$T_2 := \frac{4r}{n\sqrt{n}}l_{c,0}\sum_{i,j}^{n}\mathbb{E}\psi_\theta^0(x_i)\dot{\psi}_\theta^0(x_j), \tag{8.10}$$

$$T_3 := \frac{2r}{n}l_{c,0}^2\sum_{i}^{n}\mathbb{E}\dot{\psi}_\theta^0(x_i) = -\frac{2r}{n}l_{c,0}, \tag{8.11}$$

$$T_4 := \frac{r}{n^2\sqrt{n}}l_{c,2}\sum_{i,j,k}^{n}\mathbb{E}\psi_\theta^0(x_i)\psi_\theta^0(x_j)\psi_\theta^0(x_k), \tag{8.12}$$

$$T_5 := \frac{4r^2}{n^2}l_{c,0}l_{c,2}\sum_{i,j}^{n}\mathbb{E}\psi_\theta^0(x_i)\psi_\theta^0(x_j) \tag{8.13}$$

$$T_6 := \frac{3r^3}{n\sqrt{n}}l_{c,0}^2 l_{c,2}\sum_{i}^{n}\mathbb{E}\psi_\theta^0(x_i), \tag{8.14}$$

$$T_7 := \frac{r^4}{n}l_{c,0}^3 l_{c,2}. \tag{8.15}$$

As ψ is an IC according to Assumption 8.2 (F), $\dot{\psi}$ is $o(n^{-1})$ in the intervals I and III. Hence $T_2 = o(n^{-1})$ under Δ. T_3 and T_7 are constants. Obviously T_6 is zero and T_5 leads

to the quadratic case already treated above. So the only remaining terms are the triple T_1 and T_4. Explicit multiplication again gives terms of order 0, 1, 2 and 3. The terms of smaller order than three have already been discussed in the B_0 paragraph. Only the new triples have to be examined:

$$\begin{aligned} T_1^* &:= \frac{2}{n^2} n(n-1)(n-2)\mathbb{E}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\dot{\psi}_\theta^0(x_3) \qquad (8.16)\\ &= \frac{2(n-1)(n-2)}{n}\mathbb{E}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\dot{\psi}_\theta^0(x_3) \\ T_4^* &:= \frac{r}{n^2\sqrt{n}} n(n-1)(n-2) l_{c,2}\mathbb{E}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3) \qquad (8.17)\\ &= \frac{r(n-1)(n-2)l_{c,2}}{n\sqrt{n}}\mathbb{E}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3) \end{aligned}$$

The term T_4^*:

Again, we look at the notation of $\mathbb{E}_{Q_n}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3)$ by simple perturbations:

$$\begin{aligned} & \mathbb{E}_{Q_n}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3) \\ = \; & \mathbb{E}_{F_n}\psi_\theta^0(x_1)(1+\frac{r}{\sqrt{n}}q_1)\psi_\theta^0(x_2)(1+\frac{r}{\sqrt{n}}q_2)\psi_\theta^0(x_3)(1+\frac{r}{\sqrt{n}}q_3) \\ = \; & \mathbb{E}_{F_n} q\cdot\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3) \end{aligned}$$

with the "tangent term"

$$q := 1 + \frac{r}{\sqrt{n}}(q_1+q_2+q_3) + \frac{r^2}{n}(q_1q_2+q_2q_3+q_3q_1) + \frac{r^3}{n\sqrt{n}}q_1q_2q_3 \qquad (8.18)$$

As in the two-dimensional case concerning the term B_0^2 in Lemma 8.10, for a possible difference in the $n$maxMSE we are only interested in the scenario, where no factor is evaluated in the interval II, because otherwise we have the following cases:

(a) **One** factor is in II, i.e. without loss of generality: $\psi_1^0 \in II_1$. Then first, it holds for the tangents that by evaluation q_1 delivers 0 and q_2, q_3 deliver $\pm\frac{b}{2}+o(n^{-1})$. Second, the integrand $\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3)$ is odd, hence $\mathbb{E}_F\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3)=0$. Thus

$$\mathbb{E}_{Q_n}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3) = 0$$

(b) **Two** factors are in II, i.e. without loss of generality: $\psi_1^0 \in II_1$ and $\psi_2^0 \in II_2$. Then we have

$$|\mathbb{E}_{Q_n}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3)| \le \sup|\psi_3^0|\cdot|\mathbb{E}_{Q_n}\psi_\theta^0(x_1)\psi_\theta^0(x_2)|$$

By assumption 8.12 (U) the $\psi(x_i)$ are only weakly correlated for $x_i \in II$, i.e. $\mathbb{E}_{Q_n}\psi_\theta^0(x_1)\psi_\theta^0(x_2) = o(n^{-1/2})$, hence

$$|\mathbb{E}_{Q_n}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3)| = o(n^{-1/2})$$

(c) **All three** factors are in II. Then it holds for the tangents that by evaluation all three q_1, q_2 and q_3 deliver 0. This indicates for the tangent term to be $q = 1$, hence we have the ideal distributed situation $r = 0$. But as the integrand $\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3)$ is odd, we have

$$\mathbb{E}_{Q_n}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3) = \mathbb{E}_F\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3) = 0$$

The residual cases represent the corners of cube $[I, II, III]^3$:

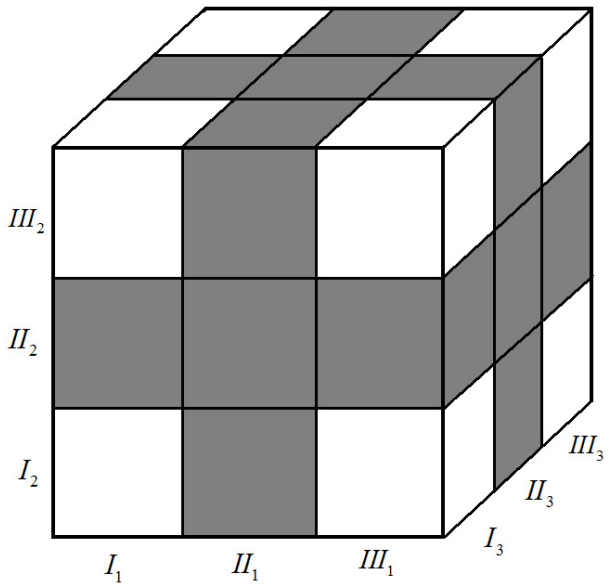

Figure 8.4: The cube $[I, II, III]^3$ with the vanishing cases darkened for the term T_4^*.

Area	Probability	Integrand Value	Quantity
$I^2 \times III$	$\frac{K^2(K-1)}{n(n-1)(n-2)}$	$\frac{b^3}{8}$	3
$III^2 \times I$	$\frac{K^2(K-1)}{n(n-1)(n-2)}$	$\frac{b^3}{8}$	3
III^3	$\frac{K(K-1)(K-2)}{n(n-1)(n-2)}$	$\frac{b^3}{8}$	1
I^3	$\frac{K(K-1)(K-2)}{n(n-1)(n-2)}$	$\frac{b^3}{8}$	1

Hence

$$\begin{aligned} & \mathbb{E}_{Q_n}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3) \\ = \; & \frac{b^3}{8}\frac{1}{n(n-1)(n-2)}\mathbb{E}8K^3 - 12K^2 + 4K \end{aligned}$$

We define $\mu := \mathbb{E}K$, $v := \mathrm{Var}K = \mathbb{E}K^2 - (\mathbb{E}K)^2$, $K_1 := K - \mu$ and $\rho := \mathbb{E}K_1^3 = \mathbb{E}(K-\mu)^3$.

Then

$$\begin{aligned}\mathbb{E}K^3 &= \mathbb{E}(K_1+\mu)\\ &= \rho + 3\mathbb{E}(K-\mu)^2\mu + 3\mu^2\mathbb{E}(K-\mu) + \mu^3\\ &= \rho + 3(\mathbb{E}K^2 - 2\mu^2\mathbb{E}K + \mu^2) + 0 + \mu^3\\ &= \rho + 3\mu v + \mu^3\end{aligned}$$

and

$$\mathbb{E}K^2 = v + \mu^2. \tag{8.19}$$

With this abbreviations we get

$$\begin{aligned}\mathbb{E}8K^3 - 12K^2 + 4K &= 8(\rho + 3\mu v + \mu^3) - 12(v+\mu^2) + 4\mu)\\ &= 8\rho + 24\mu v + 8\mu^3 - 12v - 12\mu^2 + 4\mu\end{aligned}$$

We apply that $\mu = r\sqrt{n}$ and $v = \frac{1}{2}r\sqrt{n}$:

$$\begin{aligned}\mathbb{E}8K^3 - 12K^2 + 4K &= 8\rho + 12r^2n + 8r^3n^{3/2} - 6r\sqrt{n} - 12r^2n + 4r\sqrt{n}\\ &= 8\rho + 8r^3n^{3/2} + 4r\sqrt{n}\end{aligned}$$

This leads us to the final calculation of the order of the T_4^*-term in the maximum mean squared error:

$$\begin{aligned}T_4^* &= \frac{r(n-1)(n-2)l_{c,2}}{n\sqrt{n}}\mathbb{E}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\psi_\theta^0(x_3)\\ &= \frac{r(n-1)(n-2)l_{c,2}}{n\sqrt{n}} \cdot \frac{b^3}{8}\frac{1}{n(n-1)(n-2)} \cdot (8\rho + 8r^3n^{3/2} + 4r\sqrt{n})\\ &= b^3 \cdot O(n) \cdot O(n^{-1/2}) \cdot O(n^{-3}) \cdot (\rho + O(n^{3/2}) + O(n^{1/2}))\\ &= b^3 \cdot O(n^{-1})\\ &= o(n^{-1/2}),\end{aligned}$$

where we used that $\rho = \mathbb{E}(K-\mu)^3 = O(n^{3/2})$.

The term T_1^*:

As $\dot{\psi}_\theta^0$ appears as a factor in T_1^*, we have to pay attention to the interval II, too. Additionally we remember that $\dot{\psi}_\theta^0$ is the centered derivative of the influence curve and reads $\dot{\psi}_\theta + 1$ explicitly. As we want to use the fact that ψ_θ^0 is zero in the intervals I and III, we decompose T_1^* into two summands and analyze each one separately:

$$\begin{aligned}T_1^* &= \frac{2(n-1)(n-2)}{n}\mathbb{E}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\dot{\psi}_\theta^0(x_3)\\ &= \frac{2(n-1)(n-2)}{n}[\alpha+\beta]\end{aligned} \tag{8.20}$$

with

$$\alpha = \mathbb{E}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\dot{\psi}_\theta(x_3), \tag{8.21}$$
$$\beta = \mathbb{E}\psi_\theta^0(x_1)\psi_\theta^0(x_2). \tag{8.22}$$

First we stay with **the term α:**

Differently to the two-dimensional case concerning the term B_0^2 in Lemma 8.10, on the one hand we have to deal with three dimensions and on the other hand regard that the third factor now has to be evaluated only in the interval II. Then we get the following cases, which represent the corners of the middle slice in the cube $[I, II, III]^3$:

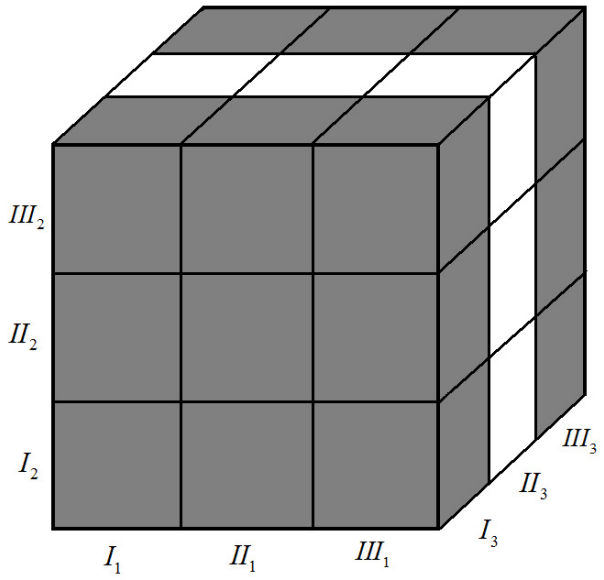

Figure 8.5: The cube $[I, II, III]^3$ with the vanishing cases darkened for the term α in T_1^*.

The residual cases again are negligibly small or equal to zero according to a similar argumentation as we did for the term T_4^* before. As the statements are quite the same we just go on to the non-negligible situation.

Area	Probability	Integrand value	Quantity
$I^2 \times II$	$\frac{K(K-1)(n-2K)}{n(n-1)(n-2)}$	$A\frac{b^2}{4}$	1
$III^2 \times II$	$\frac{K(K-1)(n-2K)}{n(n-1)(n-2)}$	$A\frac{b^2}{4}$	1
$I \times III \times II$	$\frac{K^2(n-2K)}{n(n-1)(n-2)}$	$A\frac{b^2}{4}$	2

As ψ_θ is an influence curve we also have that

$$\mathbb{E}\dot{\psi}_\theta^0(x_3) = \mathbb{E}\dot{\psi}_\theta(x_3) + 1 = 0$$

and this gives according to the probability of $\frac{n-2K}{n}$ in interval II

$$A = -\frac{n}{n-2K}, \tag{8.23}$$

with $\mathbb{E}\dot{\psi} = A$ the Lagrangian multiplier.

Hence

$$\begin{aligned}
\alpha &= \mathbb{E}_{Q_n}\psi_\theta^0(x_1)\psi_\theta^0(x_2)\dot{\psi}_\theta(x_3) \\
&= \mathbb{E}_{F_n}\psi_\theta^0(x_1)(1+\frac{r}{\sqrt{n}}q_1)\psi_\theta^0(x_2)(1+\frac{r}{\sqrt{n}}q_2)\psi_\theta(x_3)(1+\frac{r}{\sqrt{n}}q_3) \\
&= A\frac{b^2}{4}\frac{1}{n(n-1)(n-2)}\mathbb{E}2K^2(n-2K)+2K(K-1)(n-2K) \\
&\overset{(8.23)}{=} \frac{b^2}{4}\frac{1}{n(n-1)(n-2)}\mathbb{E}2nK-4nK^2.
\end{aligned}$$

Now we examine the **term** β:

As the third factor is only represented by 1, there are all three intervals possible areas for the realizations of X_3. Then we get the following cases, which represent the horizontal rows behind the corners of the front panel $[I, II, III]^2$ in the cube $[I, II, III]^3$:

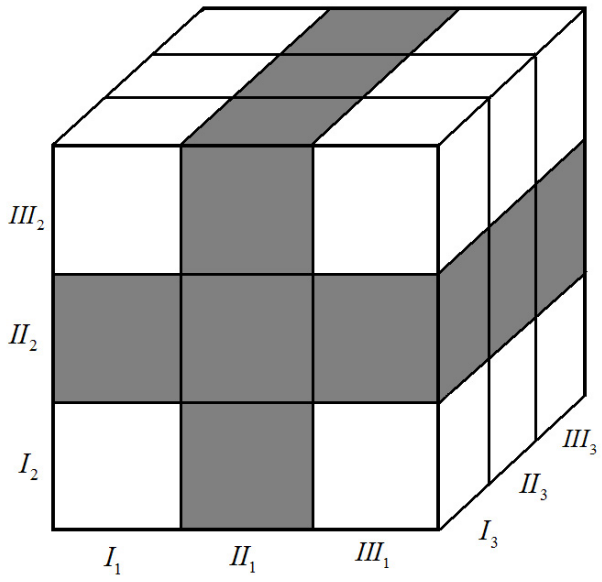

Figure 8.6: The cube $[I, II, III]^3$ with the vanishing cases darkened for the term β in T_1^*.

The residual cases again are negligibly small or equal to zero according to a similar argumentation as we did for the term T_4^* before. We just go on to the non-negligible situation.

Area	Probability	Integrand value	Quantity
I^3	$\frac{K(K-1)(K-2)}{n(n-1)(n-2)}$	$\frac{b^2}{4}$	1
$I^2 \times II$	$\frac{K(K-1)(n-2K)}{n(n-1)(n-2)}$	$\frac{b^2}{4}$	1
$I^2 \times III$	$\frac{K^2(K-1)}{n(n-1)(n-2)}$	$\frac{b^2}{4}$	1
$I \times III \times I$	$\frac{K^2(K-1)}{n(n-1)(n-2)}$	$\frac{b^2}{4}$	2
$I \times III \times II$	$\frac{K^2(n-2K)}{n(n-1)(n-2)}$	$\frac{b^2}{4}$	2
$I \times III \times III$	$\frac{K^2(K-1)}{n(n-1)(n-2)}$	$\frac{b^2}{4}$	2
$III^2 \times I$	$\frac{K^2(K-1)}{n(n-1)(n-2)}$	$\frac{b^2}{4}$	1
$III^2 \times II$	$\frac{K^2(n-2K)}{n(n-1)(n-2)}$	$\frac{b^2}{4}$	1
III^3	$\frac{K^2(K-1)}{n(n-1)(n-2)}$	$\frac{b^2}{4}$	1

Hence

$$\begin{aligned}
\beta &= \mathbb{E}_{Q_n}\psi_\theta^0(x_1)\psi_\theta^0(x_2) \\
&= \mathbb{E}_{F_n}\psi_\theta^0(x_1)(1+\frac{r}{\sqrt{n}}q_1)\psi_\theta^0(x_2)(1+\frac{r}{\sqrt{n}}q_2) \\
&= \frac{b^2}{4}\frac{1}{n(n-1)(n-2)}\mathbb{E}2(K^3-K^2+nK^2-2K^3+K^3-K^2)+ \\
&\quad +2(K^3-3K^2+2K+nK^2-2K^3-nK+2K^2+K^3-K^2) \\
&= \frac{b^2}{4}\frac{1}{n(n-1)(n-2)}\mathbb{E}4nK^2-8K^2-2nK+4K.
\end{aligned}$$

Now we **sum up** α **and** β and calculate T_1^*:

$$\begin{aligned}
T_1^* &= \frac{2(n-1)(n-2)}{n}[\alpha+\beta] \\
&= \frac{b^2}{4}\frac{2(n-1)(n-2)}{n}\frac{1}{n(n-1)(n-2)}\mathbb{E}2nK-4nK^2+4nK^2-8K^2-2nK+4K \\
&= \frac{b^2}{4}\frac{2}{n^2}\mathbb{E}4K-8K^2
\end{aligned}$$

We apply that $\mathbb{E}K = r\sqrt{n}$ and $\mathbb{E}K^2 = \frac{1}{2}r\sqrt{n}+r^2n$ and end up with the order of the T_1^*-term in the maximum mean squared error:

$$\begin{aligned}
T_1^* &= \frac{2b^2}{n^2}(r\sqrt{n}-r\sqrt{n}-2r^2n) \\
&= -\frac{b^2r^2}{n} \\
&= o(n^{-1/2})
\end{aligned}$$

Hence,

$$n\mathbb{E}\left\{\psi(X_1)\psi(X_2)\mathbb{E}[I_n(X_1,X)I_n(X_2,X)|X_1,X_2]\right\} = o(n^{-1/2})$$

□

8.7.2 The case $II \times II$

Up to now the result of Theorem 8.14 is based upon Assumption 8.12 (U) proposing the $\psi(x_i)$ to be only weakly correlated for $x_i \in II$. In this subsection we take a closer look and check this assumption for validity. Proposition 8.16 delivers a first result, namely the fact that after all we cannot hope for $E = 0$ under all the assumptions made so far.

Notation 8.15. *Let*

$$\begin{aligned} m &:= m(x_1,x_2) = \min(x_1,x_2) \quad \textit{and} \quad M := M(x_1,x_2) = \max(x_1,x_2) \\ E &:= \mathbb{E}\left[\psi(X_1)\psi(X_2)\mathbb{E}[I_n(X_1,X)I_n(X_2,X)|X_1,X_2]\right] \\ g(x_1,x_2) &:= \mathbb{E}[I_n(X_1,X)I_n(X_2,X)|X_1=x_1,X_2=x_2] \end{aligned}$$

Proposition 8.16. *With Notation 8.15 it holds that $E > 0$.*

Proof. The definition of E or g, respectively, shows that again we are in need of the common law of X_1 and X_2. By Notation 8.11

$$\begin{aligned} & I_n(x_1,x)\cdot I_n(x_2,x) = \mathbb{I}(x_{[k:n]} \le x_1, x_2 \le x_{[(n-k+1):n]}) = \\ = \;& \mathbb{I}(\#\{x_i\colon x_i \le m \mid i=3,\dots,n\} \ge k)\cdot \mathbb{I}(\#\{x_i\colon x_i \ge M \mid i=3,\dots,n\} \ge k) \end{aligned}$$

and X_i, $i = 3,\dots,n$ stochastically independent from (m, M).

Hence for m, M fix (modulo zero-sets)

$$\begin{aligned} & I_n(x_1,x)I_n(x_2,x) = \mathbb{I}(x_{[k:(n-2)]} \le m)\mathbb{I}(x_{[(n-k+1):(n-2)]} \ge M) = \\ = \;& \mathbb{I}(x_{[k:(n-2)]} \le m) - \mathbb{I}(x_{[k:(n-2)]} \le m)\mathbb{I}(x_{[(n-k+1):(n-2)]} \le M) \end{aligned}$$

With this we get

$$g(x_1,x_2) := H_{n-2}(m,M) = \mathbb{E}[\mathbb{I}(x_{[k:(n-2)]} \le m) - \mathbb{I}(x_{[k:(n-2)]} \le m)\mathbb{I}(x_{[(n-k+1):(n-2)]} \le M)]$$

For $m \le M$ we have (after partial derivation according to s) and $\nu_1 = k$, $\nu_2 = n-k+1$

$$H_{n-2}(m,M) = \int_{-\infty}^{m}\int_{M}^{\infty} h_{n-2}(t,s)\,ds\,dt$$

for

$$h_{n-2}(t,s) = n(n-1)\binom{n-2}{k-1,k-1} f(s)f(t)F(t)^{k-1}(1-F(s))^{k-1}(F(s)-F(t))^{n-2k}$$

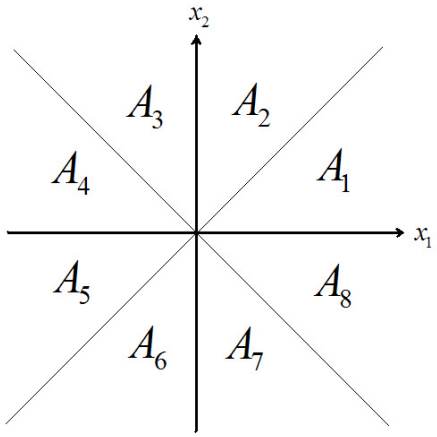

Figure 8.7: The "Compass Card - Partition" of the x_1, x_2-plane.

Because of the symmetry of F it holds that

$$H_{n-2}(m, M) = H_{n-2}(-M, -m) \tag{8.24}$$

We now look at the following "Compass Card - Partition" of the x_1, x_2-plane in Figure 8.7 and define

$$\begin{aligned}
A_1(x_1, x_2) &:= \{(x_1, x_2) \in \mathbb{R}^2 \colon\ 0 \leq x_2 \leq x_1\} \\
A_2(x_1, x_2) &:= \{(x_1, x_2) \in \mathbb{R}^2 \colon\ 0 \leq x_1 \leq x_2\} \\
A_3(x_1, x_2) &:= \{(x_1, x_2) \in \mathbb{R}^2 \colon\ 0 \leq -x_1 \leq x_2\} \\
A_4(x_1, x_2) &:= \{(x_1, x_2) \in \mathbb{R}^2 \colon\ 0 \leq x_2 \leq -x_1\} \\
A_5(x_1, x_2) &:= \{(x_1, x_2) \in \mathbb{R}^2 \colon\ 0 \leq -x_2 \leq -x_1\} \\
A_6(x_1, x_2) &:= \{(x_1, x_2) \in \mathbb{R}^2 \colon\ 0 \leq -x_1 \leq -x_2\} \\
A_7(x_1, x_2) &:= \{(x_1, x_2) \in \mathbb{R}^2 \colon\ 0 \leq x_1 \leq -x_2\} \\
A_8(x_1, x_2) &:= \{(x_1, x_2) \in \mathbb{R}^2 \colon\ 0 \leq -x_2 \leq x_1\}
\end{aligned}$$

Then it holds that

$$E = E_1 + E_2 + E_3 + E_4 + E_5 + E_6 + E_7 + E_8$$

for

$$E_i = \mathbb{E}\left[\psi(x_1)\psi(x_2)H_{n-2}(m(x_1, x_2), M(x_1, x_2))\mathbb{I}((x_1, x_2) \in A_i)\right]$$

Because of symmetry for the exchange of x_1 and x_2, it even holds that

$$E = 2(E_2 + E_3 + E_6 + E_7)$$

as

$$H_{n-2}(m(x_1,x_2),M(x_1,x_2)) = H_{n-2}(m(x_2,x_1),M(x_2,x_1))$$

and

$$\begin{aligned}
\{\mathbb{I}((x_1,x_2)\in A_1\} &= \{\mathbb{I}((x_2,x_1)\in A_2\}\\
\{\mathbb{I}((x_1,x_2)\in A_6\} &= \{\mathbb{I}((x_2,x_1)\in A_5\}\\
\{\mathbb{I}((x_1,x_2)\in A_7\} &= \{\mathbb{I}((x_2,x_1)\in A_4\}\\
\{\mathbb{I}((x_1,x_2)\in A_8\} &= \{\mathbb{I}((x_2,x_1)\in A_3\}
\end{aligned}$$

According to the definition of A_2 and A_3 we have $-x_2 \le x_1 \le x_2$ and $x_2 > 0$. Hence $m(\pm x_1,x_2) = \pm x_1$, $M(\pm x_1,x_2) = x_2$ and

$$\begin{aligned}
E_2+E_3 &= \int_0^\infty\int_{-x_2}^{x_2} \psi(x_1)\psi(x_2)H(\min(x_1,x_2),\max(x_1,x_2))f(x_1)f(x_2)\,dx_1\,dx_2 =\\
&= \int_0^\infty\int_0^{x_2} \psi(x_1)H(\min(x_1,x_2),\max(x_1,x_2))f(x_1) +\\
&\quad +\psi(-x_1)H(\min(-x_1,x_2),\max(-x_1,x_2))f(-x_1)\,dx_1\,\psi(x_2)f(x_2)\,dx_2 =\\
&\overset{\psi \text{ odd}}{=} \int_0^\infty\int_0^{x_2} \psi(x_1)f(x_1)[H(x_1,x_2)-H(-x_1,x_2)]\,dx_1\,\psi(x_2)f(x_2)\,dx_2 \quad (8.25)
\end{aligned}$$

Analogously holds by changing sign of x_2 in a first and of x_1 in a second step

$$\begin{aligned}
E_6+E_7 &= \int_{-\infty}^0\int_{x_2}^{-x_2} \psi(x_1)\psi(x_2)H(\min(x_1,x_2),\max(x_1,x_2))f(x_1)f(x_2)\,dx_1\,dx_2 =\\
&= -\int_0^\infty\int_{-x_2}^{x_2} \psi(x_1)\psi(x_2)H(\min(x_1,-x_2),\max(x_1,-x_2))f(x_1)f(x_2)\,dx_1\,dx_2 =\\
&= \int_0^\infty\int_{-x_2}^{x_2} \psi(x_1)\psi(x_2)H(\min(-x_1,-x_2),\max(-x_1,-x_2))f(x_1)f(x_2)\,dx_1\,dx_2 =\\
&\overset{(8.24)}{=} \int_0^\infty\int_{-x_2}^{x_2} \psi(x_1)\psi(x_2)H(-\max(x_1,x_2),-\min(x_1,x_2))f(x_1)f(x_2)\,dx_1\,dx_2 =\\
&= E_2+E_3
\end{aligned}$$

thus

$$E = 4(E_2+E_3)$$

Unfortunately, by (8.25) it holds that the integrand of E_2+E_3 is strictly positive as with $h_{n-2} > 0$ we have

$$H(x_1,x_2)-H(-x_1,x_2) = \int_{-\infty}^{x_1\ge 0} S_{n-2}(t)\,dt - \int_{-\infty}^{-x_1<0} S_{n-2}(t)\,dt > 0$$

with $S_{n-2}(t) = \int_{x_2}^\infty h_{n-2}(t,s)\,ds > 0$. This gives

$$E > 0$$

□

But after all, we can show that E is negligible, even though not of desired order:

Theorem 8.17. *It holds*

a) for ψ (only) fulfilling assumption 6.9 (bmi) it holds that

$$\mathbb{E}\left[\psi(X_1)\psi(X_2)\mathbb{E}[I_n(X_1,X)I_n(X_2,X)|X_1,X_2]\right] \leq 4\sup|\psi|^2\frac{(\log n)^2}{n^{3/2}}\left(1+o(n^0)\right) \tag{8.26}$$

b) for ψ as defined in assumption 8.2, i.e. $\psi(t) = sup|\psi| = b/2$ for $t \geq c_n$, the bound above on the expectation is tight and (8.26) reads

$$\mathbb{E}\left[\psi(X_1)\psi(X_2)\mathbb{E}[I_n(X_1,X)I_n(X_2,X)|X_1,X_2]\right] = b^2\frac{r(\log n)^2}{n^{3/2}}\left(1+o(n^0)\right)$$

Proof. We recall from the proof of Proposition 8.16 that

$$E_2 + E_3 = \int_0^\infty \int_0^{x_2} \int_{-x_1}^{x_1} \int_{x_2}^\infty \psi(x_1)f(x_1)\psi(x_2)f(x_2)\, h_{n-2}(t,s)\, ds\, dt\, dx_1\, dx_2$$

First, we use the general transformation formula with the substitutions $\tilde{x}_i = F(x_i)$, $\tilde{t} = F(t)$, $\tilde{s} = 1 - F(s)$ and achieve with $dt = d\tilde{t}/f(t)$, $d\tilde{x}_i = f(x_i)dx_i$

$$E_2 + E_3 = \int_{1/2}^1 \int_{1/2}^{\tilde{x}_2} \int_{1-\tilde{x}_1}^{\tilde{x}_1} \int_0^{1-\tilde{x}_2} \psi(F^{-1}(\tilde{x}_1))\psi(F^{-1}(\tilde{x}_1))h_{n-2}(\tilde{t},\tilde{s})\, d\tilde{s}\, d\tilde{t}\, d\tilde{x}_1\, d\tilde{x}_2$$

with

$$h_{n-2}(\tilde{t},\tilde{s}) = n(n-1)\binom{n-2}{k-1,k-1}\tilde{t}^{k-1}\tilde{s}^{k-1}(1-\tilde{s}-\tilde{t})^{n-2k} \tag{8.27}$$

We now set $k = r\sqrt{n}$, $\sigma_n = \log n/n^{1/4}$ and get by the Stirling formula from section A.6

$$\binom{n-2}{k-1,k-1} = (\frac{r}{\sqrt{n}})^{2-2r\sqrt{n}}(1-2\frac{r}{\sqrt{n}})^{n-2r\sqrt{n}}\frac{1}{(2\pi)r\sqrt{n}}(1+n^0)$$

Then by Lemma D.3 it holds that

$$h(\tilde{t},\tilde{s}) \leq \frac{n^{3/2}}{(2\pi)r}(1+n^0) \tag{8.28}$$

We recall that

$$h(\tilde{t},\tilde{s}) = n(n-1)\left(\frac{r}{\sqrt{n}}\right)^{2-2r\sqrt{n}}\left(1-2\frac{r}{\sqrt{n}}\right)^{n-2r\sqrt{n}}\frac{1}{2\pi r\sqrt{n}}\cdot\tilde{t}^{k-1}\tilde{s}^{k-1}(1-\tilde{s}-\tilde{t})^{n-2k}(1+n^0)$$

Thus let $\tilde{s} = \frac{r}{\sqrt{n}}(1+\sigma_n u)$, $\tilde{t} = \frac{r}{\sqrt{n}}(1+\sigma_n v)$ for $\sigma_n = \log n/n^{1/4}$. Then it holds that

$$\begin{aligned}
& \tilde{t}^{k-1}\tilde{s}^{k-1}(1-\tilde{s}-\tilde{t})^{n-2k} \\
= & \left(\frac{r}{\sqrt{n}}(1+\sigma_n v)\right)^{r\sqrt{n}-1}\left(\frac{r}{\sqrt{n}}(1+\sigma_n u)\right)^{r\sqrt{n}-1}\left(1-\frac{2r}{\sqrt{n}}-\frac{r}{\sqrt{n}}\sigma_n(u+v)\right)^{n-2r\sqrt{n}} \\
= & \left(\frac{r}{\sqrt{n}}\right)^{-(2-2r\sqrt{n})}[(1+\sigma_n v)(1+\sigma_n u)]^{r\sqrt{n}-1}\left(1-\frac{r\sigma_n(u+v)}{\sqrt{n}(1-\frac{2r}{\sqrt{n}})}\right)^{n-2r\sqrt{n}}
\end{aligned}$$

Hence

$$h(\tilde{t},\tilde{s}) = \frac{n^{3/2}}{2\pi r}\left[(1+\sigma_n u)(1+\sigma_n v)\right]^{\sqrt{n}r-1}\left[1-\frac{r}{\sqrt{n}}\frac{\sigma_n(v+u)}{1-2\frac{r}{\sqrt{n}}}\right]^{n-2r\sqrt{n}}(1+n^0) \tag{8.29}$$

We now use that $\exp\log x = x$ and apply a Taylor expansion to the logarithm. Then it holds by Lemma D.4 that

$$\begin{aligned}
h(\tilde{t},\tilde{s}) &= \frac{n^{3/2}}{2\pi r}\exp\left\{-r\sqrt{n}\sigma_n^2(u^2+v^2)/2+o(n^0)\right\}(1+n^0) = \\
&= \frac{n^{3/2}}{2\pi r}\exp\left\{-r(\log n)^2(u^2+v^2)/2+o(n^0)\right\}(1+n^0)
\end{aligned}$$

So for our inner integral only the domain $|u|,|v| < C$, for some $C > 0$, is interesting, otherwise this expression tends to zero exponentially.

Going back to the last statement of $E_2 + E_3$ we see that the integration domain is limited as

$$\begin{aligned}
\frac{1}{2} &\leq \tilde{x}_1 \leq \tilde{x}_2 && (8.30)\\
\frac{1}{2} &\leq \tilde{x}_2 \leq 1 && (8.31)\\
1-\tilde{x}_1 &\leq \tilde{t} \leq \tilde{x}_1 && (8.32)\\
0 &\leq \tilde{s} \leq 1-\tilde{x}_2 && (8.33)
\end{aligned}$$

The first and second inequality give

$$\frac{1}{2} \leq \tilde{x}_1 \leq \tilde{x}_2 \leq 1, \tag{8.34}$$

the third inequality transforms to

$$1-\tilde{x}_1 \leq \max(\tilde{t}, 1-\tilde{t}) \leq \tilde{x}_1 \tag{8.35}$$

and the fourth inequality gives

$$\tilde{x}_2 \le 1 - \tilde{s} \le 1. \tag{8.36}$$

Combining (8.34) to (8.36) we have

$$1/2 \le \hat{t} := \max\{\tilde{t}, 1-\tilde{t}\} \le \tilde{x}_1 \le \tilde{x}_2 \le 1 - \tilde{s} \le 1 \tag{8.37}$$

Exchanging the order of integration (first: $\tilde{x}_1,\ \tilde{x}_2$), we get

$$\begin{aligned}
|E_2 + E_3| \quad &\le \quad \sup|\psi|^2 \int_{1/2}^1 \int_{1/2}^{\tilde{x}_2} \int_{1-\tilde{x}_1}^{\tilde{x}_1} \int_0^{1-\tilde{x}_2} n(n-1) \times \\
&\quad \times \binom{n-2}{k-1,k-1} \tilde{t}^{k-1}\tilde{s}^{k-1}(1-\tilde{s}-\tilde{t})^{n-2k}\, d\tilde{s}\, d\tilde{t}\, d\tilde{x}_1\, d\tilde{x}_2 = \\
&\overset{(8.37)}{=} \quad \sup|\psi|^2 \int_0^1 \int_{\hat{t}}^1 \int_{\hat{t}}^{1-\tilde{s}} \int_{\hat{t}}^{\tilde{x}_2} n(n-1) \times \\
&\quad \times \binom{n-2}{k-1,k-1} \tilde{t}^{k-1}\tilde{s}^{k-1}(1-\tilde{s}-\tilde{t})^{n-2k}\, d\tilde{x}_1\, d\tilde{x}_2\, d\tilde{s}\, d\tilde{t} = \\
&= \quad \sup|\psi|^2 \int_0^1 \int_{\hat{t}}^1 \int_0^{1-\tilde{s}-\hat{t}} n(n-1) \times \\
&\quad \times \binom{n-2}{k-1,k-1} \tilde{t}^{k-1}\tilde{s}^{k-1}(1-\tilde{s}-\tilde{t})^{n-2k}\, \tilde{x}_2\, d\tilde{x}_2\, d\tilde{s}\, d\tilde{t} = \\
&= \quad \sup|\psi|^2 \int_0^1 \int_{\hat{t}}^1 n(n-1) \times \\
&\quad \times \binom{n-2}{k-1,k-1} \frac{(\hat{t}-1+\tilde{s})^2}{2}\, \tilde{t}^{k-1}\tilde{s}^{k-1}(1-\tilde{s}-\tilde{t})^{n-2k}\, d\tilde{s}\, d\tilde{t}
\end{aligned}$$

We substitute $\tilde{s} = \frac{r}{\sqrt{n}}(1+\sigma_n u)$, $\tilde{t} = \frac{r}{\sqrt{n}}(1+\sigma_n v)$, which gives

$$d\tilde{s}\, d\tilde{t} = \frac{r^2(\log n)^2}{n\sqrt{n}}$$

As then $\hat{t}$ is in $1-\tilde{t}$ n-eventually we get with

$$\hat{t} - 1 + \tilde{s} = 1 - \tilde{t} - 1 + \tilde{s} = \tilde{s} - \tilde{t} = \frac{r}{\sqrt{n}}(\sigma_n(u-v))$$

that

$$\begin{aligned}
|E_2 + E_3| \quad &\le \quad \sup|\psi|^2 \frac{n^{3/2}}{(2\pi)r} \int_{-C}^{C} \int_{-C}^{C} \exp\left\{-\frac{r}{2}(\log n)^2(u^2+v^2)\right\} \times \\
&\quad \times \frac{1}{2}\left(\frac{r}{\sqrt{n}}(\sigma_n(u-v))\right)^2 \frac{\log(n)^2 r^2}{n\sqrt{n}}\, du\, dv\, (1+o(n^0)) = \\
&= \quad \sup|\psi|^2 \frac{r}{(2\pi)} \int_{-C\log n}^{C\log n} \int_{-C\log n}^{C\log n} \exp\left\{-\frac{r}{2}(u^2+v^2)\right\} \frac{(r(u-v))^2}{2} \times \\
&\quad \times \frac{\log(n)^2}{n^{3/2}}\, du\, dv\, (1+o(n^0))
\end{aligned}$$

For n sufficiently large we drop the integration limits and look at the integrand as the density of a random variable $X := U - V$ with $U, V \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/\sqrt{r})$ and stochastically independent. With $\mathbb{E}(U - V) = 0$ the variance calculates to

$$\mathrm{Var}(U - V) = \mathbb{E}(U - V)^2 = \mathrm{Var}U + \mathrm{Var}V = \frac{1}{r} + \frac{1}{r} = \frac{2}{r}. \tag{8.38}$$

So

$$\begin{aligned} |E_2 + E_3| &\leq \sup|\psi|^2 \frac{(\log n)^2}{n^{3/2}} \frac{r^2}{2} \frac{r}{2\pi} \int\int (u-v)^2 e^{-\frac{r}{2}(u^2+v^2)}\, du\, dv\, (1 + o(n^0)) \\ &= \sup|\psi|^2 \frac{(\log n)^2}{n^{3/2}} \frac{r^2}{2} \mathbb{E}_{\mathcal{N}_2(0,1/\sqrt{r})}(U-V)^2\, (1 + o(n^0)) \\ &\overset{(8.38)}{=} \sup|\psi|^2 \frac{r(\log n)^2}{n^{3/2}}\, (1 + o(n^0)) \end{aligned}$$

This bound is quite tight as long as $\psi(t) = \sup|\psi|$ for t large:

$$E = 4(E_2 + E_3) = \sup|\psi|^2 \frac{4r(\log n)^2}{n^{3/2}}\, (1 + o(n^0))$$

□

Corollary 8.18. *By Assumption 8.2 the case $II \times II$ is not negligible up to suitable order.*

Proof. For the probability we get in the case $II \times II$:

$$P_{II\times II} = \frac{n - 2K}{n} \frac{n - 2K - 1}{n - 1} \tag{8.39}$$

Additionally we recall that

$$B_{0,2} = (n-1) \cdot \mathbb{E}\psi_\theta^0(x_1)\psi_\theta^0(x_2) \tag{8.40}$$

and for the integrand value we have

$$\mathbb{E}\psi_\theta^0(x_1)\psi_\theta^0(x_2) = b^2 \frac{r(\log n)^2}{n^{3/2}}\, (1 + o(n^0)) \tag{8.41}$$

Hence,

$$\begin{aligned} B_{0,2}|_{II\times II} &= (n-1)\frac{n - 2K}{n} \frac{n - 2K - 1}{n - 1} b^2 \frac{r(\log n)^2}{n^{3/2}}\, (1 + o(n^0)) = \\ &= O\left(\frac{(\log n)^2}{\sqrt{n}}\right) \end{aligned}$$

and that is too big to be $o(n^{-1/2})$. □

8.8 Sufficient negligibility

In this section it shows up that we only get access to the result of Corollary 6.19 in the finite context if we require (in Assumption 8.19 (p)) the finite sample and the IC, respectively, to attain the minimum and maximum of the given influence curve ψ with a certain probability already. Depending on this probability we derive a lower bound on the sample length n (in Theorem 8.20), after having conjectured the existence of such a condition by preceding numerical investigations.

8.8.1 Preliminary simulation study

The same approximating assumptions as in Assumption 5.1 are in force. We compute the empirical and the empirical asymptotic MSE according to the definitions (5.3) and (5.4), respectively. Under R 2.6.0, we simulated $anzahl = 1000$ runs of sample size $n = 1$ to $n = 500$ in the ideal location model $F = \mathcal{N}(\theta, 1)$ at $\theta = 0$. Then we calculate the exact A_2-term as determined in Corollary 6.19 with the explicit coefficients as in Proposition 6.24 and compare it to the results from $\overline{\text{empMSE}}_n - \overline{\text{asyempMSE}}_n$. The results are shown in figure 8.8.

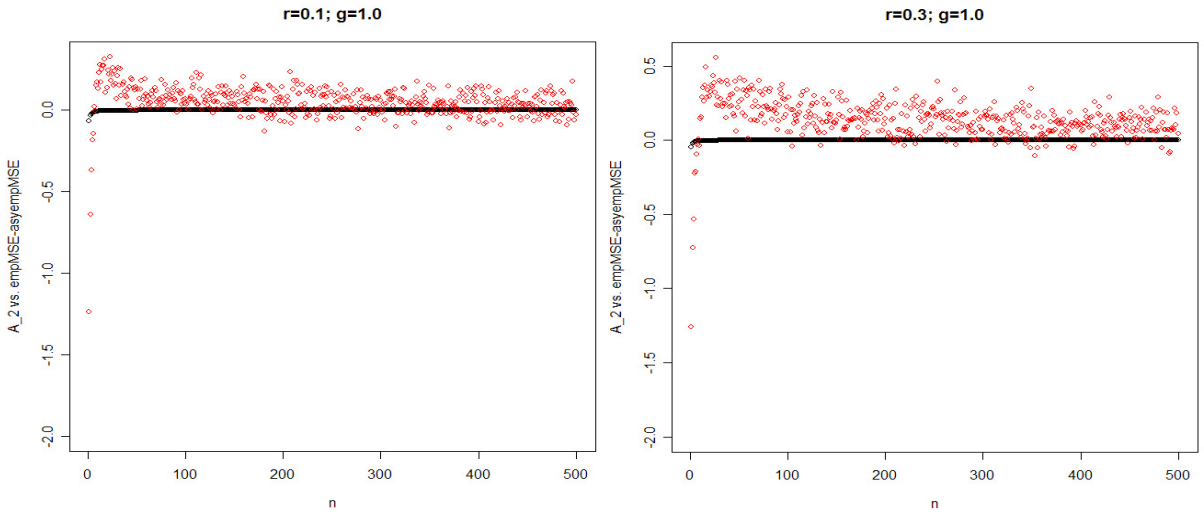

Figure 8.8: Comparison of the exact A_2-term to empirical calculations for $F = \mathcal{N}(0, 1)$, $g = 1.0$ and radii $r \in \{0.1,\ 0.3\}$

Of course, as we lack of the terms A_i, $i \geq 3$, and $A_2 < 0$ for small radii (confer chapter 7), we cannot rely on a solid interpretation of the gap between the exact and empirical values, because for relatively small n the influence of potentially positive terms of higher order may close the gap in some way. But after all we cannot deny a significant influence of the observation number n and the contamination radius r on the quality of approximation. Hence, we conjecture that there has to be a condition in n depending on r for the $II \times II$-term to vanish and therefore improve the quality of the simulation. Actually, Theorem 8.20 gives such a condition.

8.8.2 Stronger assumptions on the finite sample

We make stronger assumptions, enforcing the IC to attain its maximum and minimum, respectively, on the intervals I and III:

Assumption 8.19. *We specialize on*

(p) $F(\psi(x_i) = \pm b/2) =: p > 0$ *for all* $x_i \notin II$, *i.e.* $|x| \geq c_n$

with c_n *an increasing series and* $F(c_n) = O(n^{-1/2})$.

This leads to the desired result of the correlation vanishing in II x II:

Theorem 8.20. *For some* $\delta > 0$ *and* $n > ((r+\delta)/p)^2$ *it holds with notation 8.15 and assumption 8.19 (p) that* $E = 0$.

Proof. According to the definition of E_2 and E_3 (with $m = x_1,\ M = x_2$) in the proof of Theorem 8.17 we have

$$\begin{aligned} E_2 + E_3 &= \int_0^\infty \int_{-x_2}^{x_2} \psi(x_1)\psi(x_2) H(\min(x_1, x_2), \max(x_1, x_2)) f(x_1) f(x_2)\, dx_1\, dx_2 = \\ &= \int_0^\infty \int_{-x_2}^{x_2} \int_{-\infty}^{x_1} \int_{x_2}^{\infty} \psi(x_1) f(x_1) \psi(x_2) f(x_2)\, h_{n-2}(t, s)\, ds\, dt\, dx_1\, dx_2 \end{aligned}$$

Again, we use the general transformation formula with the substitutions $\tilde{x}_i = F(x_i)$, $\tilde{t} = F(t)$, $\tilde{s} = 1 - F(s)$ and achieve

$$E_2 + E_3 = \int_{1/2}^1 \int_{1-\tilde{x}_2}^{\tilde{x}_2} \int_0^{\tilde{x}_1} \int_{\tilde{x}_2}^1 \psi(F^{-1}(\tilde{x}_1)) \psi(F^{-1}(\tilde{x}_1))\, h_{n-2}(\tilde{t}, \tilde{s})\, d\tilde{s}\, d\tilde{t}\, d\tilde{x}_1\, d\tilde{x}_2$$

Again we see that the integration domain is limited as

$$\begin{aligned} 1/2 &\leq \tilde{x}_2 \leq 1 \\ 1 - \tilde{x}_2 &\leq \tilde{x}_1 \leq \tilde{x}_2 \\ 0 &\leq \tilde{s} \leq \tilde{x}_1 \\ \tilde{x}_2 &\leq \tilde{t} \leq 1 \end{aligned}$$

The second and third inequality can be matched to

$$\max(1 - \tilde{x}_2, \tilde{s}) \leq \tilde{x}_1 \tag{8.42}$$

Case I: $\tilde{s} \leq 1 - \tilde{x}_2$:

By Assumption 8.2 (o) ψ is odd, i.e.

$$\psi(-x) = \pm\psi(x) := \begin{cases} +\psi(x), & \text{for } x < 0 \\ -\psi(x), & \text{for } x > 0 \end{cases} \tag{8.43}$$

Let $\tilde{\psi} := \psi \circ F^{-1}$, then

$$\begin{aligned}\tilde{\psi}(1-\tilde{x}_2) &= \psi(F^{-1}(1-\tilde{x}_2)) = \psi(F^{-1}(1-F(x_2)) \\ &= \psi(F^{-1}(F(-x_2))) = \psi(-x_2) = \pm\psi(x_2) \\ &= \pm\psi(F^{-1}(\tilde{x}_2)) = \pm\tilde{\psi}(\tilde{x}_2) \end{aligned} \tag{8.44}$$

By (8.42) it holds that $\max(1-\tilde{x}_2, \tilde{s}) \leq \tilde{x}_1$. With $\tilde{s} \leq 1-\tilde{x}_2$ this gives

$$1-\tilde{x}_2 \leq \tilde{x}_1$$

and by the abbreviation for the inner integral

$$S(\tilde{x}_1,\ \tilde{x}_2) := \int_0^{\tilde{x}_1} \int_{\tilde{x}_2}^1 h_{n-2}(\tilde{t},\tilde{s})\, d\tilde{s}\, d\tilde{t} \tag{8.45}$$

we get that

$$\begin{aligned} E_2 + E_3 &= \int_{1/2}^1 \int_{1-\tilde{x}_2}^{\tilde{x}_2} \tilde{\psi}(\tilde{x}_1)\tilde{\psi}(\tilde{x}_1) S(\tilde{x}_1,\ \tilde{x}_2)\, d\tilde{x}_1\, d\tilde{x}_2 \\ &\overset{(8.44)}{=} 0 \end{aligned}$$

as $\tilde{\psi}(\tilde{x}_1)$ is odd with respect to the limits of integration.

Case II: $\tilde{s} > 1-\tilde{x}_2$:

We carry out the transformation

$$X \rightsquigarrow \psi(X) \tag{8.46}$$

By assumption 8.19 (p) we are now able to choose intervals I, II and III according to $Y_{[K:n]}$ with $Y := \psi(X)$, instead of $X_{[K:n]}$. So we do the ordering by $\psi(x)$, but as ψ is monotone by assumption 6.9 (*bmi*), the previous results and expressions still hold, especially $H_{n-2}(m, M)$. In order to improve readability we keep the notation of the expressions derived in Case I and stay with the variable x, keeping (8.46) in mind.

We recall that in the proof of Theorem 8.17 we had for $m < M$

$$H_{n-2}(m, M) = \int_{-\infty}^{m} \int_{M}^{\infty} h_{n-2}(t,s)\, ds\, dt$$

with

$$h_{n-2}(t,s) = n(n-1)\binom{n-2}{k-1,k-1} f(s)f(t)F(t)^{k-1}(1-F(s))^{k-1}(F(s)-F(t))^{(n-2k)}$$

Now it holds by Assumption 8.2 (Z) for $x_i \in I, III$ that

$$F(\psi < -b/2) = F(\psi > b/2) = 0$$

Thus

$$\begin{aligned} H_{n-2}(m, M) &= \int_{-\infty}^{m} \int_{M}^{\infty} h_{n-2}(t, s)\, ds\, dt \\ &= \int_{-b/2}^{m} \int_{M}^{b/2} h_{n-2}(t, s)\, ds\, dt \end{aligned}$$

Furthermore, we have by Assumption 8.19 (p) that

$$F(m = -b/2) = p \tag{8.47}$$
$$F(M = +b/2) = p \tag{8.48}$$

After the substitutions $\tilde{x}_i = F(x_i)$, $\tilde{t} = F(t)$, $\tilde{s} = 1 - F(s)$ we see that the integration domain is limited by

$$\begin{aligned} 1/2 &\leq \tilde{x}_2 \leq 1 \\ 1 - \tilde{x}_2 &\leq \tilde{x}_1 \leq \tilde{x}_2 \\ p &\leq \tilde{s} \leq \tilde{x}_1 \\ \tilde{x}_2 &\leq \tilde{t} \leq 1 - p \end{aligned}$$

The second and third inequality again can be matched to

$$\max(1 - \tilde{x}_2, p) \leq \tilde{s} \tag{8.49}$$

But as $\tilde{s} = \frac{r}{\sqrt{n}}(1 + \sigma_n u)$ for $\sigma_n = \log n / n^{1/4}$ and some u, bounded by $|u| < C$, $C > 0$, it holds that

$$\begin{aligned} \tilde{s} &= \frac{r}{\sqrt{n}} \cdot \left(1 \pm \text{const} \frac{\log n}{n^{1/4}}\right) \\ &= \frac{r + \eta_n}{\sqrt{n}} \end{aligned}$$

for some $\eta_n > 0$.

So for n sufficiently large $\tilde{s} < \max(1 - \tilde{x}_2, p)$. If $\max(1 - \tilde{x}_2, p) = 1 - \tilde{x}_2$, then this is a contradiction to the assumption $\tilde{s} > 1 - \tilde{x}_2$ and if $\max(1 - \tilde{x}_2, p) = p$ we get a contradiction to (8.49). So we see that for

$$n > \left(\frac{r + \eta_n}{p}\right)^2$$

case $II \times II$ does not exist and the integral vanishes. □

8.9 The distribution of K

In the next Theorem we see that if we assume K to be locked in some small interval (by assumption 8.21 (PK)), then almost surely, i.e. up to exponential negligibility, we have a sufficient number of observations in interval I.

8.9.1 A restrictive condition

Assumption 8.21. *In addition to assumptions 8.12 (EK), (VK) and 8.19 (p) we specialize on*

(PK) $P(K \in [r\sqrt{n}(1-\eta), r\sqrt{n}(1+\eta)] = 1 - O(e^{-n^\delta})$ *for some* $\delta, \eta > 0$.

This leads to

Theorem 8.22. *With assumptions 8.19 it holds that*

$$P(X_{[K:n]} > -c_n) = O(e^{-n^\delta}) \quad \textit{for} \quad c_n = -F^{-1}\left(\frac{\rho}{\sqrt{n}}\right)$$

for some $\rho > r + \eta$ *and some* $\delta > 0$.

Proof. Let

$$J := [(r-\eta)\sqrt{n},\ (r+\eta)\sqrt{n}] \tag{8.50}$$

for some $\eta > 0$.

We assume that there exists a distribution for K such that for some $\delta > 0$

$$P(K \in J) = 1 - O(e^{-n^\delta}). \tag{8.51}$$

Let

$$A := \{X_{[K:n]} > -c_n\}$$

By (A.1) and (A.2) we get by setting $\varepsilon := \varepsilon_n = \frac{\eta\rho}{\sqrt{n}}$

$$P(|\mathrm{Bin}(n, r/\sqrt{n}\,) - \rho\sqrt{n}| > n\varepsilon_n) = O(e^{-n\varepsilon_n^2}) = O(e^{-n^\delta}) \tag{8.52}$$

for some $\delta > 0$.

By application of the identity in equation (D.4) we get that

$$\begin{aligned} A &= \{X_{[K:n]} > -c_n\} \\ &= \{X_{[K:n]} \le -c_n\}^c \\ &= \{\sum I_{(X_i \le -c_n)} \ge K\}^c \\ &= \{\sum I_{(X_i \le -c_n)} < K\} \end{aligned}$$

and for the distribution

$$P(A) = P(\sum Y_i < K) = P(\overline{Y} < K/n) \tag{8.53}$$

with $Y_i := \mathbb{I}_{\{X_i \le -c_n\}}$ and $Y_i \overset{\text{i.i.d.}}{\sim} \text{Bin}(1, F(-c_n)) = \text{Bin}(1, \rho/\sqrt{n})$. Then

$$\begin{aligned}
P(A) &= P(X_{[K:n]} > -c_n) \\
&= P(\text{Bin}(n, \Phi(-c_n)) < K\,) \\
&\overset{(8.52)}{=} P(\text{Bin}(n, \Phi(-c_n)) < K \wedge (r+\eta)\sqrt{n} > K > (r-\eta)\sqrt{n}\,) + O(e^{-n^\delta}) \\
&\le P(\text{Bin}(n, \rho/\sqrt{n}) - \rho\sqrt{n} < (r+\eta)\sqrt{n} - \rho\sqrt{n}\,) \\
&= P\left(\frac{\text{Bin}(n, \rho/\sqrt{n})}{n} - \frac{\rho}{\sqrt{n}} < \frac{(r+\eta-\rho}{\sqrt{n}}\right) \\
&\overset{(A.2)}{=} O(e^{-(r+\eta-\rho)^2\sqrt{n}}) = O(e^{-n^\delta}).
\end{aligned}$$

for some $\rho > r + \eta$ and some $\delta > 0$.

□

8.9.2 Explicit upper bound c_n for $F = \mathcal{N}(0,1)$

For $F = \mathcal{N}(0,1)$ we are able to calculate the bound c_n explicitly. It turns out in Proposition 8.24 that $c_n = O(\sqrt{\log n})$.

Assumption 8.23. *We specialize on*

(N) $F = \mathcal{N}(0,1)$

Proposition 8.24. *For $F = \mathcal{N}(0,1)$ it holds that $c_n = \sqrt{\log n}(1 + o(n^0))$.*

Proof. We define some $\rho(r) = O(r)$ as a slightly disturbed radius r so that

$$\Phi^{-1}\left(\frac{\rho}{\sqrt{n}}\right) = -c_n. \tag{8.54}$$

It holds for some $s_1,\ s_2 \in [0,1]$, $-s_1 \le -s_2$ that

$$-s_1 \le \Phi^{-1}\left(\frac{\rho}{\sqrt{n}}\right) \le -s_2 \tag{8.55}$$

and therefore

$$\Phi(-s_1) \le \frac{\rho}{\sqrt{n}} \le \Phi(-s_2) \tag{8.56}$$

Applying Mills' ratio as defined in A.3 by using Gordon's inequality (Lemma A.4) we get with $1 - \Phi(s_i) = \Phi(-s_i)$

$$\frac{\varphi(s_i)s_i}{s_i^2 + 1} \le \Phi(-s_i) \le \frac{\varphi(s_i)}{s_i} \tag{8.57}$$

Plugging (8.57) into (8.56) we have

$$\frac{\varphi(s_1)s_1}{s_1^2+1} \leq \frac{\rho}{\sqrt{n}} \leq \frac{\varphi(s_2)}{s_2} \tag{8.58}$$

As we are interested in a displacement of $X_{[K:n]}$ to the right side of the exact K/n-quantile $\Phi^{-1}(K/n)$ in the direction of the critical area of the influence function ψ, we confine ourselves to the upper bound $-s_2$.
The inversion of the mapping $x \mapsto \frac{\varphi(x)}{x} = C \cdot \frac{e^{-x^2/2}}{x} =: y$ gives

$$\begin{aligned} \frac{e^{-x^2/2}}{x} &= \frac{y}{C} \\ \Leftrightarrow -\frac{x^2}{2} - \log x &= \log y - \log C \end{aligned} \tag{8.59}$$

For a first approximation we neglect the term $\log x$ as $x^2/2 \gg \log x$ for x large enough. This gives, just considering the positive branch of the square root (as x is the norm of a negative bound in (8.55)), renaming $x = c_n^0$ (as being a first approximation) and plugging in $y = \rho/\sqrt{n}$:

$$\begin{aligned} c_n^0 &= \sqrt{\log C^2/y^2} = \sqrt{\log(C^2/\rho^2) + \log n} \\ &= \sqrt{\log n}\sqrt{1 + \frac{\log C^2 - \log \rho^2}{\log n}} \end{aligned} \tag{8.60}$$

Plugging this into the upper Mills' bound we get

$$\begin{aligned} \Phi(-c_n^0) &= \frac{\varphi(c_n^0)}{c_n^0} = \frac{C \exp -1/2(\log(C^2/\rho^2) + \log n)}{\sqrt{\log(C^2/\rho^2) + \log n}} \\ &= \frac{C \exp(\log(\rho/C) + \log n^{-1/2})}{\sqrt{\log(C^2/\rho^2) + \log n}} \\ &= \frac{\rho n^{-1/2})}{\sqrt{\log n}\sqrt{\log(C^2/\rho^2)/\log n + 1}} \\ &= \frac{\rho}{\sqrt{n \log n}}(1 - O((\log n)^{-1})) \\ &= \frac{\rho}{\sqrt{n \log n}}(1 - o(n^0)) \end{aligned}$$

But as this is not the desired order as in (8.54). So now we want to take a closer look and set for a second approximation

$$c_n = c_n^0 \cdot (1 - \delta_n)$$

with some $\delta_n > 0$. Plugging this into (8.59) we get

$$(c_n^0)^2(1-2\delta_n+\delta_n^2)+2\log c_n^0+2\log(1-\delta_n)=\log C^2/\rho^2+\log n$$

Applying a Taylor expansion to $\log(1-\delta_n)$ we get

$$(c_n^0)^2(1-2\delta_n+\delta_n^2)+2\log c_n^0+2(-\delta_n-\delta_n^2/2)=\log C^2/\rho^2+\log n$$

For δ_n small ($\delta_n \to 0$) we neglect δ_n^2 and get

$$(c_n^0)^2(1-2\delta_n)+2\log c_n^0-2\delta_n=\log C^2/\rho^2+\log n$$

Hence

$$\delta_n=-\frac{\log C^2-\log\rho^2+\log n-2\log c_n^0-(c_n^0)^2}{2(c_n^0)^2+1}$$

Plugging in (8.60) we achieve

$$\delta_n=\frac{\log(\log C^2-\log\rho^2+\log n)}{2(\log C^2-\log\rho^2+\log n)+1}=\frac{\log(c_n^0)}{2(c_n^0)^2+1}$$

Using that $c_n^0=O(\sqrt{\log n}(1+o(n^0)))$ we have

$$\delta_n=O\left(\frac{\sqrt{\log n}}{1+\log n}\right)=o(n^0)$$

Summarizing the results, up to now we get

$$\begin{aligned} c_n &= \sqrt{\log n}\cdot s_n \\ s_n &:= (1-\delta_n)\sqrt{1+\frac{\log C^2-\log\rho^2}{\log n}} \\ \delta_n &= \frac{\log(\log C^2-\log\rho^2+\log n)}{2(\log C^2-\log\rho^2+\log n)+1} \end{aligned}$$

with

$$s_n=(1-o(n^0))(1+o(n^0))=1-o(n^0)$$

We now control again with the Mills' bounds:

$$\begin{aligned}
\frac{\varphi(c_n)}{c_n} &= \frac{\varphi(\rho\sqrt{\log n}s_n)}{\rho\sqrt{\log n}s_n} \\
&= \frac{C\exp(-(\log n\cdot s_n^2)/2)}{\sqrt{\log n}s_n} \\
&= \frac{C\exp(-\frac{1}{2}\log n\cdot(1-\delta_n)^2\cdot(1-(\log C^{-2}+2\log\rho)/\log n))}{\exp(\log(\sqrt{\log n})\cdot(1-\delta_n)\cdot\sqrt{1-(\log C^{-2}+2\log\rho)/\log n})} \\
&= C\exp\left\{(\log\frac{\rho}{\sqrt{n}C})(1-\delta_n)^2-\log(\sqrt{\log\frac{nC^2}{\rho^2}}(1-\delta_n))\right\} \\
&= C\exp\left\{(\log\frac{\rho}{\sqrt{n}C})(1-2\delta_n+\delta_n^2)-\frac{1}{2}\log(\log\frac{nC^2}{\rho^2}+\delta_n)\right\} \\
&= C\exp\left\{\log\frac{\rho}{\sqrt{n}C}(1+\delta_n^2)\right\}\cdot R(\rho,n,C)
\end{aligned}$$

with

$$\begin{aligned}
R(\rho,n,C) &= \exp\left\{\delta_n\cdot\log\left(\frac{\rho^2}{nC^2}\right)^{-1}-\frac{1}{2}\log\log\frac{nC^2}{\rho^2}+\delta_n\right\} \\
&= \exp\left\{\frac{\log\frac{\rho^2}{nC^2}\cdot\log\log\frac{nC^2}{\rho^2}}{2(1+\log\frac{nC^2}{\rho^2})}-\frac{1}{2}\log\log\frac{nC^2}{\rho^2}+\frac{\log\log\frac{nC^2}{\rho^2}}{2(1+\log\frac{nC^2}{\rho^2})}\right\} \\
&= \exp\left\{\frac{\log\frac{\rho^2}{nC^2}\cdot\log\log\frac{nC^2}{\rho^2}-\log\log\frac{nC^2}{\rho^2}-\log\frac{nC^2}{\rho^2}\cdot\log\log\frac{nC^2}{\rho^2}+\log\log\frac{nC^2}{\rho^2}}{2(1+\log\frac{nC^2}{\rho^2})}\right\} \\
&= \exp(0)=1
\end{aligned}$$

It remains by a final Taylor expansion of the exponential function

$$\begin{aligned}
\frac{\varphi(c_n)}{c_n} &= C\exp\left(\log\frac{\rho}{\sqrt{n}C}(1+\delta_n^2)\right) \\
&= \frac{\rho}{\sqrt{n}}\cdot\exp\left(\log\frac{\rho}{\sqrt{n}C}\delta_n^2\right) \\
&= \frac{\rho}{\sqrt{n}}\cdot\left(1-O\left(\frac{(\log\log n)^2}{8\log n}\right)\right) \\
&= \frac{\rho}{\sqrt{n}}\cdot(1-o(n^0))
\end{aligned}$$

An analogous calculation shows

$$\begin{aligned}
\frac{c_n\varphi(c_n)}{1+c_n^2} &= \frac{\sqrt{\log n}s_n\varphi(\sqrt{\log n}s_n)}{1+(\sqrt{\log n}s_n)^2} \\
&= \frac{\rho}{\sqrt{n}}\cdot(1+o(n^0))
\end{aligned}$$

Hence

$$\Phi(-c_n) = \frac{\rho}{\sqrt{n}}$$

for

$$c_n = \sqrt{\log n} \cdot s_n = \sqrt{\log n}(1 + o(n^0))$$

□

8.9.3 Concrete distributions of K

Summarizing the results of the previous sections we have to find a distribution for the number K of observations to be manipulated such that

(EK) $\mathbb{E}[K] = r\sqrt{n}$

(VK) $\mathrm{Var}[K] = \frac{1}{2}r\sqrt{n}$

(PK) for every $\eta > 0$ it holds that $P(|K - r\sqrt{n}| > \eta\sqrt{(n)}) = O(e^{-n^\delta})$ for some $\delta > 0$.

A counterexample - Two-point distribution

In order to fulfill (VK) we have

$$\mathrm{Var}X = \sigma^2 = \frac{1}{2}r\sqrt{n} \quad \Rightarrow \quad \sigma = \sqrt{\frac{1}{2}r\sqrt{n}} \tag{8.61}$$

and therefore, with respect to (EK), we look at a most simple two-point distribution with support

$$\begin{aligned} a_1 &= r\sqrt{n} + \sqrt{\frac{1}{2}r\sqrt{n}} + \gamma_n \\ a_2 &= r\sqrt{n} - \sqrt{\frac{1}{2}r\sqrt{n}} + \gamma_{n'} \end{aligned}$$

and $0 \leq \gamma_n,\ \gamma_{n'} < 1$ such that $a_1,\ a_2 \in \mathbb{N}$. We abbreviate $P(K = a_i) = p_i$, $i = 1, 2$.

Then, condition (PK) is fulfilled, because

$$|K - r\sqrt{n}| \leq 2\sigma = 2\sqrt{\frac{1}{2}r\sqrt{n}} = \sqrt{2r\sqrt{n}} = O(n^{-1/4}) < o(\sqrt{n})$$

Proposition 8.25. *A two-point distribution with support*

$$\begin{aligned} a_1 &= r\sqrt{n} + \sqrt{\frac{1}{2}r\sqrt{n}} + \gamma_n \\ a_2 &= r\sqrt{n} - \sqrt{\frac{1}{2}r\sqrt{n}} + \gamma'_n \end{aligned}$$

and $0 \leq \gamma_n,\ \gamma_n' < 1$ *such that* $a_1,\ a_2 \in \mathbb{N}$ ***does not exist*** *for* $P(K = a_i) = p_i,\ i = 1, 2$ *probabilities and*

$$n > \left(\frac{\gamma_n \gamma_n'}{\gamma_n' - \gamma_n}\right)^4 \cdot \frac{4}{r^2}$$

Proof. The conditions (EK), (VK) and the sum of all p_i deliver three equations

$$\begin{aligned} \mathbb{E}K &= r\sqrt{n} \\ \mathrm{Var}K &= \frac{1}{2} r\sqrt{n} \\ \sum_{i=1}^{4} p_i &= 1 \end{aligned}$$

which read explicitly

$$\begin{aligned} &(i) \quad a_1 p_1 + a_2 p_2 = r\sqrt{n} \\ &(ii) \quad (a_1 - r\sqrt{n})^2 p_1 + (a_2 - r\sqrt{n})^2 p_2 = \frac{1}{2} r\sqrt{n} \\ &(iii) \quad p_1 + p_2 = 1 \end{aligned}$$

With $p_2 = 1 - p_1$ from (iii) we get

$$\begin{aligned} &(i') \quad -2p_1\sqrt{\frac{1}{2} r\sqrt{n}} + \sqrt{\frac{1}{2} r\sqrt{n}} + +\gamma_n p_1 - \gamma_n' p_1 + \gamma_n' = 0 \\ &(ii') \quad -2p_1(\gamma_n + \gamma_n')\sqrt{\frac{1}{2} r\sqrt{n}} + 2\gamma_n'\sqrt{\frac{1}{2} r\sqrt{n}} + \gamma_n^2 p_1 - {\gamma_n'}^2 p_1 + {\gamma_n'}^2 = 0 \end{aligned}$$

We calculate $(ii') - (\gamma_n + \gamma_n') \cdot (i')$ and get

$$\sqrt{\frac{1}{2} r\sqrt{n}} \cdot (\gamma_n' - \gamma_n) = \gamma_n \gamma_n' \tag{8.62}$$

But as the LHS is unbounded and the RHS bounded by 1, there exists a $n_0(r) \in \mathbb{N}$ such that for $n > n_0(r)$ equation (8.62) doesn't hold:

$$n_0(r) = \left(\frac{\gamma_n \gamma_n'}{\gamma_n' - \gamma_n}\right)^4 \cdot \frac{4}{r^2}$$

□

Four-point distribution

We look at a four-point distribution with support

$$\begin{aligned} a_1 &= r\sqrt{n} + \sqrt{\frac{1}{2}r\sqrt{n}} + \gamma_n \\ a_2 &= r\sqrt{n} + \sqrt{\frac{1}{2}r\sqrt{n}} + \gamma_n - 1 \\ a_3 &= r\sqrt{n} - \sqrt{\frac{1}{2}r\sqrt{n}} + \gamma_{n'} \\ a_4 &= r\sqrt{n} - \sqrt{\frac{1}{2}r\sqrt{n}} + \gamma_{n'} - 1 \end{aligned}$$

and $0 \leq \gamma_n,\ \gamma_{n'} < 1$ such that $a_1, \ldots, a_4 \in \mathbb{N}$. We abbreviate $P(K = a_i) = p_i$, $i = 1, \ldots, 4$.

Then, condition (PK) is fulfilled, because surely

$$|K - r\sqrt{n}| < o(\sqrt{n})$$

The conditions (EK), (VK) and the sum of all p_i deliver three equations for four variables:

$$\begin{aligned} \mathbb{E}K &= r\sqrt{n} \\ \mathrm{Var}K &= \frac{1}{2}r\sqrt{n} \\ \sum_{i=1}^{4} p_i &= 1 \end{aligned}$$

which read explicitly

$$\begin{aligned} &a_1p_1 + a_2p_2 + a_3p_3 + a_4p_4 = r\sqrt{n} \\ &(a_1 - r\sqrt{n})^2p_1 + (a_2 - r\sqrt{n})^2p_2 + (a_3 - r\sqrt{n})^2p_3 + (a_4 - r\sqrt{n})^2p_4 = \frac{1}{2}r\sqrt{n} \\ &p_1 + p_2 + p_3 + p_4 = 1 \end{aligned}$$

As we want the equation system to be solved for all p_i probabilities, we make the assumption

$$p_i = \frac{1}{4} + q_i \quad \text{and} \quad -\frac{1}{4} \leq q_i \leq \frac{3}{4} \qquad \text{for } i = 1 \ldots 4$$

Solving the equations system with `MAPLE` gives

$$\begin{aligned} p_1 &= \frac{1}{4} - \frac{1}{2}(\gamma_n - \gamma_n^{'}) - q_4 + O(n^{-1/4}) \\ p_2 &= \frac{1}{4} + \frac{1}{2}(\gamma_n - \gamma_n^{'}) + q_4 + O(n^{-1/4}) \\ p_3 &= \frac{1}{4} - q_4 + O(n^{-1/4}) \\ p_4 &= \frac{1}{4} + q_4 \end{aligned}$$

and q_4 free to choose with $-\frac{1}{4} \leq q_4 \leq \frac{3}{4}$.

Four-point distribution - examples

We give some examples of the four-point distribution for radius $r = 0.5$ and different sample lengths n, calculated by a CAS.

As for $a_i \in \mathbb{N},\ i = 1 \dots 4$ we can choose $\gamma_n = \gamma'_n = 0$. Then, for example,

	a_1	a_2	a_3	a_4	p_1	p_2	p_3	p_4
$n = 16$	3	2	1	0	0.45	0.15	0.35	0.05
$n = 256$	10	9	6	5	0.25	5/12	1/12	0.25
$n = 4096$	36	35	28	27	0.15	0.45	0.05	0.35

Chapter 9

Outlook

Although this thesis gives an extensive treatise of the Higher Order Asymptotics for the MSE of Robust M-Estimators of Location on Shrinking Total Variation Neighborhoods, in a more general perspective the topic is a highly specific one. The are several extensions to think of:

- Obviously, the location model is only the simplest model. The approach based on M-estimators in section 4.3, especially, only works for this kind of model as the proof of the Main Theorem 6.13 uses the monotone character concerning the influence curves of those estimators, compare (4.27) or figure 4.1, respectively. Regarding the scale model, this approach won't work as the IC lacks global monotonicity. Therefore, in subsection 4.3.2 we already suggest a suitable and promising possibility in form of a k-step-approach according to (4.43). Of course, this alternative is not limited to the scale model but seems applicable for several estimating problems. The location model treated here may be handled in a more general way by the k-step-approach, too.

- The preliminary simulation study in chapter 5 was done by algorithms assuming the fact $K \overset{\text{i.i.d.}}{\sim} \text{Bin}(1, r/\sqrt{n})$. Of course, this worked out as an approximation. But one might think of an extensive simulation study using the theoretically derived conditions (EK), (VK) and (PK) in subsection 8.9.3, i.e. $K \overset{\text{i.i.d.}}{\sim} P$ with $\mathbb{E}_P K = r\sqrt{n}$, $\text{Var} K = \frac{1}{2} r\sqrt{n}$ and $\forall \eta > 0 : \; P(|K - r\sqrt{n}| > \eta\sqrt{(n)}) = O(\exp(-n^{\delta}))$. An example in form of a four-point distribution was derived in subsection 8.9.3 as a most simple one, and subsequently we computed a representative for radius $r = 0.5$ and sample lengths $n \in \{16,\ 256,\ 4096\}$. As an extension one might think of an algorithm generally adapting to the current situation and automatically switching to an appropriate distribution - maybe 4-point, maybe not.

- The explanation of the dependence of the behavior of the A_2-term on r and g and in this context the question of under- or overestimation of the MSE by first-order vs. second- or third-order asymptotics is only a heuristic one. Further investigations of the theoretical background might be interesting in order to bring more light to this situation.

- Up to now Theorem 8.20 only holds for optimal ICs required by assumption 8.19 (p). Probably it is possible to expand the class of ICs by weakening assumption (p). The question to answer is whether the constant p may be substituted by a sequence $p_n \to 0$ for $n \to \infty$.
- For the result of this thesis being a rather theoretical one, the potential application presented in the preface concerning a robustification of operational risk estimation in the financial sector by asymmetric ICs offers an opportunity for total variation neighborhoods and the derived distribution for K to show its value in practice.

Appendix A

Tools

A.1 Two Hoeffding Bounds

Lemma A.1. *Let $\xi_i \overset{\text{i.i.d.}}{\sim} P$, $i = 1, \ldots, n$ be real–valued random variables, $|\xi_i| \leq M$ Then for $\varepsilon > 0$*

$$P(\frac{1}{n}\sum_i \xi_i - \mathbb{E}[\xi_1] \geq \varepsilon) \leq \exp(-\frac{n\varepsilon^2}{2M^2}) \tag{A.1}$$

$$P(\frac{1}{n}\sum_i \xi_i - \mathbb{E}[\xi_1] \leq -\varepsilon) \leq \exp(-\frac{n\varepsilon^2}{2M^2}) \tag{A.2}$$

Proof. [Hoeffding (1963)], Thm. 2. □

Lemma A.2. *Let $\xi_i \overset{\text{i.i.d.}}{\sim} P$, $i = 1, \ldots, n$ be real–valued random variables, $|\xi_i| \leq 1$ Then for $\mu = \mathbb{E}[\xi_1]$ and $0 < \varepsilon < 1 - \mu$*

$$P(\frac{1}{n}\sum_i \xi_i - \mu \geq \varepsilon) \leq \left\{ \left(\frac{\mu}{\mu+\varepsilon}\right)^{\mu+\varepsilon} \left(\frac{1-\mu}{1-\mu-\varepsilon}\right)^{1-\mu-\varepsilon} \right\}^n \tag{A.3}$$

Proof. [Hoeffding (1963)], Thm. 1, inequality (2.1). □

A.2 Mills' ratio

Definition A.3. *The function R defined by*

$$R(x) = \frac{1-\Phi(x)}{\varphi(x)} = \frac{\int_x^{+\infty} e^{-t^2/2} dt}{e^{-x^2/2}} = e^{x^2/2} \int_x^{+\infty} e^{-t^2/2} dt \tag{A.4}$$

is called MILLS' ratio.

Lemma A.4. *For all $x > 0$ it holds that*

$$\frac{x}{x^2+1} \leq R(x) \leq \frac{1}{x}. \tag{A.5}$$

Proof. See [Gordon (1941)]. □

A.3 A uniform Edgeworth expansion

In the following theorem, a generalized form of Thm. 1 in [Ibragimov (1967)] and Thm. 3.3.1 in [Ibragimov and Linnik (1971)], proved in [Ruckdeschel (2005b)], to the situation where the law of ξ_i depends through an additional parameter t:

Theorem A.5. *For some set $\Theta \subset \mathbb{R}$ and fixed $t \in \Theta$ let $\xi_{i,t}$, $i = 1, 2, \ldots$ be a sequence of i.i.d. real-valued random variables with distribution F_t and with*

$$\mathbb{E}\xi_{i,t} = 0, \quad \mathbb{E}\xi_{i,t}^2 = 1, \quad \mathbb{E}\xi_{i,t}^3 = \rho_t, \quad \mathbb{E}\xi_{i,t}^4 - 3 = \kappa_t \tag{A.6}$$

Let $\Phi(s)$ and $\varphi(s)$ be the c.d.f. and p.d.f. of $\mathcal{N}(0,1)$ and

$$F_n(s,t) := P(\sum_{i=1}^{n} \xi_{i,t} < s\sqrt{n}) \tag{A.7}$$

$$H_n(s,t) := \Phi(s) - \frac{\varphi(s)}{\sqrt{n}} \frac{\rho_t}{6}(s^2 - 1) \tag{A.8}$$

$$G_n(s,t) := H_n(s,t) - \frac{\varphi(s)}{n}\Big[\frac{\kappa_t}{24}(s^3 - 3s) + \frac{\rho_t^2}{72}(s^5 - 10s^3 + 15s)\Big] \tag{A.9}$$

Let f_t be the characteristic function of F_t.

(a) If $\sup_t \kappa_t < \infty$ and if there is some $u_0 > 0$ such that for all u_1 the "no-lattice"-condition (C)'

$$\hat{f}_{u_0}(u_1) := \sup_{u_0 < u < u_1} \sup_t |f_t(u)| < 1 \tag{A.10}$$

is fulfilled, then

$$\sup_{s \in \mathbb{R}} \sup_t |F_n(s,t) - H_n(s,t)| = o(n^{-1/2}) \tag{A.11}$$

(b) If

$$\sup_t \mathbb{E}|\xi_{i,t}|^5 < \infty \tag{A.12}$$

and the uniform Cramér–condition (C)

$$\limsup_{u \to \infty} \sup_t |f_t(u)| < 1 \tag{A.13}$$

is fulfilled, then

$$\sup_{s \in \mathbb{R}} \sup_t |F_n(s,t) - G_n(s,t)| = O(n^{-3/2}) \tag{A.14}$$

A.4 A refined implicit function theorem

This strategy for a refined version of the implicit function theorem is taken from subsection 5.3.2 of [Ruckdeschel (2005a)]:

Let $F : \mathbb{R}^2 \to \mathbb{R}$ be continuously differentiable in $z_a = (x_a, y_a)$ and let $F(z_a) = 0$ and $F_x(z_a) \neq 0$. Then to a given y_b "near" y_a we look for a x_b such that $z_b = (x_b, y_b)$ is a

zero of F - at least up to a suitable order. The implicit function theorem leads to the approximation

$$x_0 = x_a - \frac{F_y(z_a)}{F_x(z_a)}(y_b - y_a) \tag{A.15}$$

To refine this approximation we will now consider the function $x \mapsto F(x, y_b)$ and apply the first step of a Newton procedure to get as approximation for x_b

$$x_1 = x_0 - \frac{F(x_0, y_b)}{F_x(x_0, y_b)} \tag{A.16}$$

This approximation is then proved to achieve the desired exactness in each case by controlling $F(x_1, y_b)$.

A.5 Decay of the standard normal

We note the following Lemma for $\mathcal{N}(0,1)$ variables, appearing as Lemma 5.7 in [Ruckdeschel (2005a)]

Lemma A.6. *Let $X \sim \mathcal{N}(0,1)$. Then for $k = 0, 1, 2, \ldots, 8$ and any sequence $(c_n)_n \subset \mathbb{R}$ with $\liminf_n c_n > \sqrt{2}$,*

$$\mathbb{E}[X^k \mathbb{I}_{\{X \geq c_n \sqrt{\log(n)}\}}] = o(n^{-1}) \tag{A.17}$$

Proof. Lemma 5.7 in [Ruckdeschel (2005a)] □

A.6 Stirling Approximations

These approximations for the factorials and the binomial coefficients derived from the Stirling formula can be found e.g. in [Abramowitz and Stegun (1984)], 6.1.37:

$$n! = (\frac{n}{e})^n \sqrt{2\pi n}\,(1 + \frac{1}{12n} + o(n^{-1})), \tag{A.18}$$

$$\binom{2n}{n} = 4^n/\sqrt{\pi n}(1 - \frac{1}{8n} + o(n^{-1})), \tag{A.19}$$

From this we easily derive

$$\binom{2n-k}{n-k} = (\frac{2n-k}{\max(n-k,1)})^{n-k}(\frac{2n-k}{n})^n \sqrt{\frac{n-k/2}{n(n-k)\pi}}(1+\rho_{n,k}),$$
$$\text{for } -1/2 - 1/(48n) \leq \rho_{n,k} \leq \frac{1}{12n}, \tag{A.20}$$

$$= (\frac{2n-k}{n-k})^{n-k}(\frac{2n-k}{n})^n \sqrt{\frac{n-k/2}{n(n-k)\pi}}(1 - \frac{1}{8n} + o(\frac{1}{n})),$$
$$\text{for } k = O(\sqrt{n}), \tag{A.21}$$

$$\binom{2n-k}{n-j} = (\frac{2n-k}{\max(n-j,1)})^{n-j}(\frac{2n-k}{n+j-k})^{n+j-k}\sqrt{\frac{2n-k}{(n+j-k)(n-j)2\pi}}(1+\rho_{n,j,k}),$$

$$\text{for } -1/2-1/(48n) \le \rho_{n,j,k} \le \frac{1}{12n}, \tag{A.22}$$

$$= (\frac{2n-k}{n-j})^{n-j}(\frac{2n-k}{n+j-k})^{n+j-k}\sqrt{\frac{2n-k}{(n+j-k)(n-j)2\pi}}(1-\frac{1}{8n}+o(\frac{1}{n})),$$

$$\text{for } j,k = O(\sqrt{n}), \tag{A.23}$$

Appendix B

Negligibility of cases (II) to (IV)

B.1 Case (II) for K binomial distributed

Lemma B.1. *Let $k_1(n) = 1 + d_n$ and assume that for some $\delta \in (0, 1/4)$,*

$$d_n n^{1/4-\delta} \to \infty, \qquad d_n n^{-1/4+\delta} \to 0 \qquad \text{for } n \to \infty$$

Let

$$\mathcal{K}_n := k_1(n) \log k_1(n) + 1 - k_1(n)$$

Then if $\liminf_n d_n > 0$ there is some $c > 0$ such that

$$P(\mathrm{Bin}(n, r/\sqrt{n}\,) > k_1(n) r\sqrt{n}\,) = o(e^{-cr\sqrt{n}}) \tag{B.1}$$

and, if $d_n = o(n^0)$, for any $0 < \delta_0 \le 2\delta$, it holds that

$$P(\mathrm{Bin}(n, r/\sqrt{n}\,) > k_1(n) r\sqrt{n}\,) = o(e^{-rn^{\delta_0}}) \tag{B.2}$$

Proof. Lemma 8.1 in [Ruckdeschel (2005b)] □

We add Remark 8.1 and Corollary 8.2 from [Ruckdeschel (2005b)]:

Remark B.2. *Even if d_n is increasing at a faster rate than $n^{1/4}$, assertion* (B.1) *remains true, as long as $\liminf_n d_n > 0$ —but this is not needed here.*

As in (II), $|t| < Cn^{1+3/\delta}$, the integrand of $n\,\mathrm{MSE}(S_n, Q_n \mid K = k\,)$ is bounded by some polynomial in n, and hence by Lemma B.1 the contribution of (II) is indeed $o(n^{-1})$. Another consequence of the exponential decay of (B.1)/(B.2) is

Corollary B.3. *Let $K \sim \mathrm{Bin}(n, r/\sqrt{n}\,)$. Then, in the setup of Lemma B.1, for any $j \in \mathbb{N}$,*

$$\mathbb{E}[K^j \mathbb{I}_{\{X \ge k_1(n) r\sqrt{n}\}}] = o(e^{-rn^d}) \tag{B.3}$$

for any $0 < d < \sqrt{n}$ if $\liminf_n d_n > 0$ and any $0 < d \le \delta_0$ if $\lim_n d_n = 0$.

Proof. Corollary 5.4 in [Ruckdeschel (2005a)] □

B.2 Case (III)

We apply Hoeffding's bound Lemma A.1:

$$\begin{aligned} P\{S_n > \sqrt{t}\} &\leq P(T_n(\sqrt{t}\,) \geq -\tilde{t}\,) \\ &= P(T_n(\sqrt{t}) - nL_{\mathrm{re}}(t)\,) \geq -r\sqrt{n}\tilde{t} - n\tilde{L}_{\mathrm{re}}(t)\,) \leq \exp(-2n\Delta^2/b^2) \end{aligned}$$

for $\Delta := -\tilde{L}_{\mathrm{re}}(\sqrt{t}\,) - \frac{r\tilde{t}}{\sqrt{n}}$. As ψ is isotone, $\tilde{L}_{\mathrm{re}}$ is antitone, hence in case (II),

$$\tilde{L}_{\mathrm{re}}(\sqrt{t}\,) \leq \tilde{L}_{\mathrm{re}}(b\sqrt{k_2\log(n)/n}\,) = -b\sqrt{k_2\log(n)/n} + o(\sqrt{\log(n)/n}\,) \tag{B.4}$$

Thus

$$\Delta \geq -\tilde{L}_{\mathrm{re}}(\sqrt{t}\,) - \frac{rb}{\sqrt{n}} \overset{(B.4)}{>} \frac{b}{\sqrt{n}}[\sqrt{k_2\log(n)} + o(\sqrt{\log(n)}\,)]$$

and

$$\exp(-2\frac{n\,\Delta^2}{b^2}) < n^{-2k_2}(1 + o(n^0))$$

This latter is $o(n^{-3-3/\delta})$ and thus integrating n MSE out along (II) we get something of order $o(n^{-1})$.

B.3 Case (IV)

Without loss, assume that $b = \hat{b}$. By monotonicity and boundedness in assumption (bmi), to given $0 < \eta < -\check{b}$ there is a $t_0 > 0$ such that for $t > t_0$,

$$\check{b} < L_{\mathrm{id}}(t) \leq \check{b} + \eta$$

Let $t_1 > t_0$, $\delta > 0$ and $C' > 0$ so that for $t > t_1$, by (Vb), $|V_{\mathrm{id}}(t)| \leq C' t^{-1-\delta}$. Then we apply the Chebyshev inequality to obtain for $t > t_1^2$

$$\begin{aligned} & P\{S_n > \sqrt{t}\} \leq P(T_{\mathrm{re},n}(\sqrt{t}) \geq 0) = P(T_{\mathrm{id},n}(\sqrt{t}) + T_{\mathrm{c},n}(\sqrt{t}) \geq 0) \\ = \;& P(T_{\mathrm{id},n}(\sqrt{t}) \geq -T_{\mathrm{c},n}(\sqrt{t})) = P(T_{\mathrm{id},n}(\sqrt{t}\,) \geq -\tilde{t}) \\ = \;& P\Big(T_{\mathrm{id},n}(\sqrt{t}\,) - (n-k)L_{\mathrm{id}}(\sqrt{t}\,) \geq -\tilde{t} - (n-k)L_{\mathrm{id}}(\sqrt{t}\,)\Big) \leq \\ \overset{\text{Cheb.}}{\leq} \;& \frac{(n-k)V_{\mathrm{id}}^2(\sqrt{t}\,)}{(\tilde{t} + (n-k)L_{\mathrm{id}}(\sqrt{t}\,))^2} \overset{\text{(Vb)}}{\leq} \frac{(n-k)C't^{-(1+\delta)}}{(\tilde{t} + (n-k)L_{\mathrm{id}}(\sqrt{t}\,))^2} \leq \frac{nC't^{-(1+\delta)}}{(\tilde{t} + (n-k)\check{b} + \eta)^2} \leq \\ \overset{\tilde{t}\leq k\hat{b}}{\leq} \;& \frac{nC't^{-(1+\delta)}}{[k\hat{b} + (n-k)\check{b} + \eta]^2} \leq \frac{nC't^{-(1+\delta)}}{[k(\hat{b} - \check{b}) + n\check{b} + \eta]^2} \overset{k\leq n/2}{\leq} \frac{nC't^{-(1+\delta)}}{(\check{b} + \eta)^2} \end{aligned} \tag{B.5}$$

and correspondingly (with $b = -\check{b}$) for $P\{S_n \leq -\sqrt{t}\}$; but according to the definition of case (IV)

$$\frac{C'n^2}{(b+\eta)^2}\int_{Cn^{1+3/\delta}}^{\infty} t^{-(1+\delta)}\,dt = \frac{C'C^{-\delta}n^{-1-\delta}}{\delta(\check{b}+\eta)^2} = o(n^{-1}) \tag{B.6}$$

Appendix C

The explicit A_2-term

$$
\begin{aligned}
A_2 \;=\; & \Big[3l_{c,1}^2b^2 \pm \big(-3l_{c,1}(-v_{id,1}-l_{id,2}) + 2l_{c,1}l_{id,2} + l_{c,2} - 2l_{c,1}v_{id,1} + l_{c,1}(-l_{id,2}-2v_{id,1}) - \\
& -(l_{id,2}+v_{id,1})l_{c,1} - 2l_{c,1}(-\frac{1}{2}l_{id,2}-v_{id,1})\big)b^3 + (\frac{1}{3}l_{id,3} + \frac{1}{4}l_{id,2}^2 - l_{id,2}(-v_{id,1}-l_{id,2}) + \\
& +(-v_{id,1}-l_{id,2})^2 - l_{id,2}(-\frac{1}{2}l_{id,2}-v_{id,1}) - l_{id,2}v_{id,1} + (l_{id,2}+v_{id,1})v_{id,1} + \\
& +\frac{1}{2}l_{id,2}(-l_{id,2}-2v_{id,1}) + (l_{id,2}+v_{id,1})(-v_{id,1}-l_{id,2}) + \frac{1}{2}(-2l_{id,2}-2v_{id,1})v_{id,1} - \\
& -\frac{1}{4}l_{id,2}/v_{id,0}^2)b^4\Big]r^4 + \Big[l_{id,3} + 15(\frac{1}{2}l_{id,2}+v_{id,1})(-l_{id,2}-2v_{id,1}) + \frac{15}{2}(-l_{id,2}-2v_{id,1})v_{id,1} + \\
& +3v_{id,2} + 45(-\frac{1}{2}l_{id,2}-v_{id,1})^2 + 6v_{id,1}^2 + 3(l_{id,2}+v_{id,1})v_{id,1} + \frac{3}{2}l_{id,2}v_{id,1}\Big]v_{id,0}^4 + \Big[(4l_{c,1}v_{c,0} + \\
& +v_{c,0}^2 + 3l_{c,1}^2)v_{id,0}^2 + \big((2l_{id,3} + 3v_{id,2} + 6(-v_{id,1}-l_{id,2})^2 - 6l_{id,2}(-\frac{1}{2}l_{id,2}-v_{id,1}) + \\
& +30(-v_{id,1}-l_{id,2})(-\frac{1}{2}l_{id,2}-v_{id,1}) + 6(-\frac{1}{2}l_{id,2}-v_{id,1})^2 + \frac{3}{2}l_{id,2}v_{id,1} + \\
& +6(l_{id,2}+v_{id,1})v_{id,1} + 6v_{id,1}^2 + \frac{3}{2}l_{id,2}(-l_{id,2}-2v_{id,1}) + 3(l_{id,2}+v_{id,1})(-v_{id,1}-l_{id,2}) + \\
& +\frac{9}{2}(-2l_{id,2}-2v_{id,1})v_{id,1} + 3(\frac{1}{2}l_{id,2}+v_{id,1})(-l_{id,2}-2v_{id,1}) + \frac{9}{2}(-l_{id,2}-2v_{id,1})v_{id,1} + \\
& +6(l_{id,2}+v_{id,1})(-\frac{1}{2}l_{id,2}-v_{id,1}) + 6(l_{id,2}+v_{id,1})(-l_{id,2}-2v_{id,1}) + 6(\frac{1}{2}l_{id,2} + \\
& +v_{id,1})(-v_{id,1}-l_{id,2}))v_{id,0}^2 + ((\frac{1}{2}l_{id,2}+v_{id,1})\rho_0 + \frac{3}{2}\rho_0(-\frac{1}{2}l_{id,2}-v_{id,1}) + \\
& +2(l_{id,2}+v_{id,1})\rho_0 + 3\rho_0(-v_{id,1}-l_{id,2}) + \frac{1}{3}\rho_0(3l_{id,2}+3v_{id,1}) + \\
& +\frac{1}{6}\rho_0(\frac{3}{2}l_{id,2}+3v_{id,1}))v_{id,0} - 6l_{id,2}\big)b^2 \pm \big((6l_{id,2}+4v_{id,1})v_{c,0} + 3l_{c,2} + \\
& +9l_{c,1}(-l_{id,2}-2v_{id,1}) + 4v_{c,1} + 4b^2 - 9l_{c,1}(-v_{id,1}-l_{id,2}) - 36l_{c,1}(-\frac{1}{2}l_{id,2}-v_{id,1}) - \\
& -6(\frac{1}{2}l_{id,2}+v_{id,1})l_{c,1} - 3(l_{id,2}+v_{id,1})l_{c,1} - 6l_{c,1}v_{id,1}\big)v_{id,0}^2b\Big]r^2 - \frac{21}{4}l_{id,2}v_{id,0}^2 \\
& +\big(\frac{1}{2}\rho_0(\frac{3}{2}l_{id,2}+3v_{id,1}) + \frac{2}{3}\rho_1 + \rho_0(-l_{id,2}-2v_{id,1}) + 10(\frac{1}{2}l_{id,2}+v_{id,1})\rho_0 + \frac{15}{2}\rho_0(-\frac{1}{2}l_{id,2}-v_{id,1})\big)v_{id,0}^3
\end{aligned}
$$

Appendix D

The common law of two quantiles

For all situations of the belonging intervals of the X_i we start with the common law of

$$(Y, Z) := (X_{[\nu_1:n]}, X_{[\nu_2:n]}) \tag{D.1}$$

for $1 \leq \nu_1 < \nu_2 \leq n,\ X_i \overset{\text{i.i.d.}}{\sim} F$, $i = 1, \ldots, n$ and $F(dx) = f(x)\,dx$. We write

$$P^{Y,Z}(s,t) := \mathbb{P}(Y \leq s,\ Z \leq t)$$

D.1 Distributions and densities

To determine distributions and densities we use the following lemma that already appears in [Ruckdeschel (2005a)]

Lemma D.1. *Let $X_i \overset{\text{i.i.d.}}{\sim} P$ real-valued random variables. Then*

$$P(X_{[k:n]} \leq t) = \sum_{l=k} \binom{n}{l} P(t)^l (1 - P(t))^{n-l} \tag{D.2}$$

If $dP = p\,d\lambda$, then $X_{[k:n]}$ has density

$$g(t) = np(t) \binom{n-1}{k-1} P(t)^{k-1} (1 - P(t))^{n-k} \tag{D.3}$$

Proof. The proof is standard, but as we will need some terms later on, we pass through the main steps here: For fixed $t \in \mathbb{R}$ we introduce $Y_i := I_{\{X_i \leq t\}}$. Then the following events are identical

$$\{X_{[k:n]} \leq t\} = \{\#i : \{X_i \leq t\} \geq k\} = \{\sum_{i=1}^{n} Y_i \geq k\} \tag{D.4}$$

The fact that $Y_i \overset{\text{i.i.d.}}{\sim} \text{Bin}(1, P(t))$ entails (D.2). (D.3) follows by simple differentiating.

□

Lemma D.2. *The marginal distribution functions and densities of Y, Z are*

$$\begin{aligned}
P^Y(s) &:= \sum_{j=\nu_1}^{n} \binom{n}{j} F(s)^j \big(1-F(s)\big)^{(n-j)} \\
P^Z(t) &:= \sum_{j=\nu_2}^{n} \binom{n}{j} F(t)^j \big(1-F(t)\big)^{(n-j)} \\
p^Y(s) &:= n\binom{n-1}{\nu_1-1} F(s)^{(\nu_1-1)} \big(1-F(s)\big)^{(n-\nu_1)} f(s) \\
p^Z(t) &:= n\binom{n-1}{\nu_2-1} F(t)^{(\nu_2-1)} \big(1-F(t)\big)^{(n-\nu_2)} f(t)
\end{aligned}$$

and, $P^Z(dt)$–a.e., the factorized regular conditional distribution function of Y given $Z=t$ reads

$$P^{Y|Z=t}(s) = \mathbb{I}_{\{s>t\}} + \mathbb{I}_{\{s\leq t\}} \sum_{j=\nu_1}^{\nu_2-1} \binom{\nu_2-1}{j} \big(\frac{F(s)}{F(t)}\big)^j \big(1-\frac{F(s)}{F(t)}\big)^{(\nu_2-j-1)}$$

so the conditional density is $P^Z(dt)$–a.e.

$$p^{Y|Z=t}(s) = \mathbb{I}_{\{s\leq t\}} (\nu_2-1)\binom{\nu_2-2}{\nu_1-1} \big(\frac{F(s)}{F(t)}\big)^{(\nu_1-1)} \big(1-\frac{F(s)}{F(t)}\big)^{(\nu_2-1-\nu_1)} f(s)$$

Proof. It's appropriate to employ a partitioning into several cases. In this sense we look at the last two realizations $Y=y$ and $Z=z$ and their relationship to each other. This technique is inspired by the proceeding in [Siddiqui (1970)] and [Ruckdeschel (2005a)].

1 $\mathbf{s > t}$:

$$\mathbb{P}(Y\leq s,\ Z\leq t | s>t) = \mathbb{I}_{\{s>t\}} \cdot \mathbb{P}(Y\leq s | Z\leq t)\cdot \mathbb{P}(Z\leq t) = \mathbb{P}(Z\leq t),$$

as $\mathbb{P}(Y\leq s|Z\leq t)=1$, because $\nu_1<\nu_2$ and therefore $Y\leq t<s$.

2 $\mathbf{s \leq t}$:

a) $\mathbf{Z > s}$ (i.e. $z>y$):

In this case we have to subtract the observations smaller or equal than s:

$$\mathbb{P}(Y\leq s,\ s<Z\leq t | s\leq t) = \mathbb{I}_{\{s\leq t\}} \sum_{j=\nu_1}^{\nu_2-1} \left[\mathbb{P}\left(\sum_{i\leq n} \mathbb{I}_{\{X_i\leq s\}} = j,\ \sum_{i\leq n} \mathbb{I}_{\{s<X_i\leq t\}} \geq \nu_2 - j\right)\right]$$

b) $\mathbf{Z \leq s}$ (i.e. $z\leq y$):

$$\mathbb{P}(Y\leq s,\ Z\leq s\leq t) = \mathbb{I}_{\{s\leq t\}} \cdot \mathbb{P}(Z\leq s)$$

Combining these three cases we get

$$\begin{aligned}
P^{Y,Z}(s,t) &:= \mathbb{I}_{\{s>t\}}\mathbb{P}(Z \le t) + \mathbb{I}_{\{s\le t\}}\mathbb{P}(Y \le s,\ Z \le t) = \\
&= \mathbb{I}_{\{s>t\}}\mathbb{P}(\sum_{i\le n}\mathbb{I}_{\{X_i\le t\}} \ge n) + \\
&\qquad + \mathbb{I}_{\{s\le t\}}\mathbb{P}(\sum_{i\le n}\mathbb{I}_{\{X_i\le s\}} \ge m,\ \sum_{i\le n}\mathbb{I}_{\{X_i\le t\}} \ge n) = \\
&= \mathbb{I}_{\{s>t\}}\mathbb{P}(\sum_{i\le n}\mathbb{I}_{\{X_i\le t\}} \ge \nu_2) + \mathbb{I}_{\{s\le t\}}\mathbb{P}(\sum_{i\le n}\mathbb{I}_{\{X_i\le s\}} \ge \nu_2) + \\
&\qquad + \mathbb{I}_{\{s\le t\}}\sum_{j=\nu_1}^{\nu_2-1}\mathbb{P}(\sum_{i\le n}\mathbb{I}_{\{X_i\le s\}} = j,\ \sum_{i\le n}\mathbb{I}_{\{s<X_i\le t\}} \ge \nu_2 - j)
\end{aligned}$$

In the last summand, for each X_i three cases are possible: $X_i \le s$, $s < X_i \le t$, $X_i > t$, so this summand may be treated as a trinomial variable and, splitting the cases where $\sum_{i\le n}\mathbb{I}_{\{s<X_i\le t\}} \ge \nu_2 - j$, we get

$$\begin{aligned}
P^{Y,Z}(s,t) &:= \mathbb{I}_{\{s>t\}}\sum_{j=\nu_2}^{n}\binom{n}{j}F(t)^j\big(1-F(t)\big)^{(n-j)} + \\
&\qquad + \mathbb{I}_{\{s\le t\}}\Big\{\sum_{j=\nu_2}^{n}\binom{n}{j}F(s)^j\big(1-F(s)\big)^{(n-j)} + \\
&\qquad + \sum_{j=\nu_1}^{\nu_2-1}\sum_{l=\nu_2-j}^{n-j}\binom{n}{j,\,l}F(s)^j\big(F(t)-F(s)\big)^l\big(1-F(t)\big)^{(n-l-j)}\Big\}
\end{aligned}$$

Also, we note that the marginal distribution functions and densities of Y, Z are

$$\begin{aligned}
P^Y(s) &:= \sum_{j=\nu_1}^{n}\binom{n}{j}F(s)^j\big(1-F(s)\big)^{(n-j)} \\
P^Z(t) &:= \sum_{j=\nu_2}^{n}\binom{n}{j}F(t)^j\big(1-F(t)\big)^{(n-j)} \\
p^Y(s) &:= n\binom{n-1}{\nu_1-1}F(s)^{(\nu_1-1)}\big(1-F(s)\big)^{(n-\nu_1)}f(s) \\
p^Z(t) &:= n\binom{n-1}{\nu_2-1}F(t)^{(\nu_2-1)}\big(1-F(t)\big)^{(n-\nu_2)}f(t)
\end{aligned}$$

In order to obtain conditional densities we determine

$$\begin{aligned}\frac{\partial}{\partial t}P^{Y,Z}(s,t) &= \mathbb{I}_{\{s>t\}}p^Z(t)+\mathbb{I}_{\{s\le t\}}\sum_{j=\nu_1}^{\nu_2-1}\binom{n}{j}F(s)^j\,(n-j)\binom{n-j-1}{n-\nu_2}\times\\ &\quad\times\big(F(t)-F(s)\big)^{(\nu_2-j-1)}\big(1-F(t)\big)^{(n-\nu_2)}f(t)=\\ &= n\Big\{\mathbb{I}_{\{s>t\}}\binom{n-1}{\nu_2-1}F(t)^{(\nu_2-1)}\big(1-F(t)\big)^{(n-\nu_2)}+\\ &\quad+\mathbb{I}_{\{s\le t\}}\sum_{j=\nu_1}^{\nu_2-1}\binom{n-1}{n-\nu_2,\,j}F(s)^j\big(F(t)-F(s)\big)^{\nu_2-j-1}\times\\ &\quad\times\big(1-F(t)\big)^{(n-\nu_2)}\Big\}f(t)\end{aligned}$$

So, $P^Z(dt)$–a.e., the factorized regular conditional distribution function of Y given $Z=t$ reads

$$P^{Y|Z=t}(s) = \mathbb{I}_{\{s>t\}}+\mathbb{I}_{\{s\le t\}}\sum_{j=\nu_1}^{\nu_2-1}\binom{\nu_2-1}{j}\big(\frac{F(s)}{F(t)}\big)^j\big(1-\frac{F(s)}{F(t)}\big)^{(\nu_2-j-1)}$$

and hence the conditional density is $P^Z(dt)$–a.e.

$$p^{Y|Z=t}(s) = \mathbb{I}_{\{s\le t\}}\,(\nu_2-1)\binom{\nu_2-2}{\nu_1-1}\big(\frac{F(s)}{F(t)}\big)^{(\nu_1-1)}\big(1-\frac{F(s)}{F(t)}\big)^{(\nu_2-1-\nu_1)}f(s)$$

□

D.2 Further Lemmata

By setting $\nu_1=k,\ \nu_2=n-k+1$, in subsection 8.7.2 we derive the density

$$h_{n-2}(t,s)=n(n-1)\binom{n-2}{k-1,k-1}f(s)f(t)F(t)^{k-1}(1-F(s))^{k-1}(F(s)-F(t))^{n-2k} \quad \text{(D.5)}$$

For the substitutions $\tilde{x}_i=F(x_i)$, $\tilde{t}=F(t)$, $\tilde{s}=1-F(s)$ (D.5) becomes

$$h_{n-2}(\tilde{t},\tilde{s})=n(n-1)\binom{n-2}{k-1,k-1}\tilde{t}^{k-1}\tilde{s}^{k-1}(1-\tilde{s}-\tilde{t})^{n-2k}$$

For this expression the following Lemma holds:

Lemma D.3. *Let*

$$h_{n-2}(\tilde{t},\tilde{s})=n(n-1)\binom{n-2}{k-1,k-1}\tilde{t}^{k-1}\tilde{s}^{k-1}(1-\tilde{s}-\tilde{t})^{n-2k} \quad \text{(D.6)}$$

Then it holds that the expression $l(\tilde{t},\tilde{s})=\tilde{t}^{k-1}\tilde{s}^{k-1}(1-\tilde{s}-\tilde{t})^{n-2k}$ gets maximal for

$$t=s=\frac{r\sqrt{n}-1}{n-2}=:\theta$$

and

$$h(\tilde{t},\tilde{s}) \leq h(\theta,\theta) \leq \frac{n^{3/2}}{(2\pi)r}(1+n^0) \tag{D.7}$$

Proof.

$$\begin{aligned}
& \tilde{t}^{k-1}\tilde{s}^{k-1}(1-\tilde{s}-\tilde{t})^{(n-2k)} = \max! \\
\Leftrightarrow\ & (k-1)\log\tilde{t} + (k-1)\log\tilde{s} + (n-2k)\log(1-\tilde{s}-\tilde{t}) = \max! \\
\Leftrightarrow\ & \frac{k-1}{\tilde{t}} + 0 + \frac{n-2k}{1-\tilde{s}-\tilde{t}}\cdot(-1) = 0 \\
& \wedge\quad \frac{k-1}{\tilde{s}} + 0 + \frac{n-2k}{1-\tilde{s}-\tilde{t}}\cdot(-1) = 0 \\
\Leftrightarrow\ & \frac{k-1}{\tilde{t}} = \frac{n-2k}{1-\tilde{s}-\tilde{t}} \\
& \wedge\quad \frac{k-1}{\tilde{s}} = \frac{n-2k}{1-\tilde{s}-\tilde{t}} \\
\Leftrightarrow\ & (k-1)(1-2\tilde{t}) = (n-2k)\tilde{t} \quad \wedge \quad \tilde{s} = \tilde{t} \\
\Leftrightarrow\ & \tilde{t} = \tilde{s} = \frac{k-1}{n-2} = \frac{r\sqrt{n}-1}{n-2}
\end{aligned}$$

But

$$l(\frac{r}{\sqrt{n}},\frac{r}{\sqrt{n}}) = l(\theta,\theta)(3-o(n^0)) \tag{D.8}$$

We derive this result by the following calculation:

$$b := \frac{r}{\sqrt{n}}\cdot\theta^{-1} = \frac{r\sqrt{n}-2\frac{r}{\sqrt{n}}}{r\sqrt{n}-1} = \frac{1-\frac{2}{n}}{1-\frac{1}{r\sqrt{n}}}$$

so

$$\begin{aligned}
l(\frac{r}{\sqrt{n}},\frac{r}{\sqrt{n}}) &= (\theta b)^{-2+2r\sqrt{n}}\cdot(1-2\theta b)^{n-2r\sqrt{n}} \\
&= \theta^{-2+2r\sqrt{n}}\cdot b^{-2+2r\sqrt{n}}\cdot(1-2\theta b)^{n-2r\sqrt{n}}\cdot\left(\frac{1-2\theta b}{1-2\theta}\right)^{n-2r\sqrt{n}} \\
&= l(\theta,\theta)\cdot\left(\frac{1-\frac{2}{n}}{1-\frac{1}{r\sqrt{n}}}\right)^{-2+2r\sqrt{n}}\cdot\left(\frac{1-2\frac{r}{\sqrt{n}}}{1-\frac{2r\sqrt{n}-1}{n-2}}\right)^{n-2r\sqrt{n}} \\
&= l(\theta,\theta)\cdot\left(1-\frac{2}{n}\right)^{-2+2r\sqrt{n}}\cdot\left(1-\frac{1}{r\sqrt{n}}\right)^{-(-2+2r\sqrt{n})}\cdot\left(\frac{1-\frac{2r}{\sqrt{n}}-\frac{2}{n}+\frac{4r}{n\sqrt{n}}}{1-\frac{2r}{\sqrt{n}}-\frac{1}{n}}\right)^{n-2r\sqrt{n}} \\
&= l(\theta,\theta)\cdot l_1(r,n)\cdot l_2(r,n)\cdot l_3(r,n)
\end{aligned}$$

Applying Taylor expansions to the terms $l_i(r,n),\ i=1\ldots3$ shows

$$\begin{aligned}
l_1(r,n) &= (1-\frac{2}{n})^{-2+2r\sqrt{n}} = 1+(-2+2r\sqrt{n})\cdot\frac{2}{n}+O(n^{-1}) = 1+o(n^0)\\
l_2(r,n) &= \left(1-\frac{1}{r\sqrt{n}}\right)^{-(-2+2r\sqrt{n})} = 1+(-2+2r\sqrt{n})\frac{1}{r\sqrt{n}}+O(n^{-1}) = 3+o(n^0)\\
l_3(r,n) &= \left(1-\frac{\frac{1}{n}+\frac{4r}{n\sqrt{n}}}{1-\frac{2r}{\sqrt{n}}-\frac{1}{n}}\right)^{n-2r\sqrt{n}}\\
&= (1-O(n^{-1})(1+O(n^{-1/2})))^{n-2r\sqrt{n}} = (1-O(n^{-1}))^{n-2r\sqrt{n}}\\
&= 1-(n-2r\sqrt{n})O(n^{-1}) = 1-O(n^0)
\end{aligned}$$

so

$$l_1(r,n)\cdot l_2(r,n)\cdot l_3(r,n) = 3-o(n^0) \tag{D.9}$$

Now

$$\begin{aligned}
l\left(\frac{r}{\sqrt{n}},\frac{r}{\sqrt{n}}\right) &= \left(\frac{r}{\sqrt{n}}\right)^{-1+r\sqrt{n}}\left(\frac{r}{\sqrt{n}}\right)^{-1+r\sqrt{n}}\left(1-2\frac{r}{\sqrt{n}}\right)^{-(n-2r\sqrt{n})}\\
&= \left(\frac{r}{\sqrt{n}}\right)^{-2+2r\sqrt{n}}\left(1-2\frac{r}{\sqrt{n}}\right)^{-(n-2r\sqrt{n})}
\end{aligned}$$

hence

$$\begin{aligned}
h(\tilde{t},\tilde{s}) &\leq h(\theta,\theta)\\
&= n(n-1)\left(\frac{r}{\sqrt{n}}\right)^{2-2r\sqrt{n}}\left(1-2\frac{r}{\sqrt{n}}\right)^{n-2r\sqrt{n}}\frac{1}{2\pi r\sqrt{n}}(1+n^0)\cdot l(\theta,\theta)\\
&\overset{(D.8)}{=} n(n-1)\left(\sqrt{r}\sqrt{n}\right)^{2-2r\sqrt{n}}\left(1-2\frac{r}{\sqrt{n}}\right)^{n-2r\sqrt{n}}\frac{1}{2\pi r\sqrt{n}}(1+n^0)\cdot\frac{l(\frac{r}{\sqrt{n}},\frac{r}{\sqrt{n}})}{3-o(n^0}\\
&= n(n-1)\left(\frac{r}{\sqrt{n}}\right)^{2-2r\sqrt{n}}\left(1-2\frac{r}{\sqrt{n}}\right)^{n-2r\sqrt{n}}\frac{1}{2\pi r\sqrt{n}}\times\\
&\quad\times\left(\frac{r}{\sqrt{n}}\right)^{-2+2r\sqrt{n}}\left(1-2\frac{r}{\sqrt{n}}\right)^{-(n-2r\sqrt{n})}(1+n^0)\\
&= n(n-1)\frac{1}{2\pi r\sqrt{n}}(1+n^0)\\
&\leq \frac{n^{3/2}}{(2\pi)r}(1+n^0)
\end{aligned}$$

□

The next Lemma is rather technical.

Lemma D.4. *let* $\tilde{s} = \frac{r}{\sqrt{n}}(1+\sigma_n u)$, $\tilde{t} = \frac{r}{\sqrt{n}}(1+\sigma_n v)$ *for* $\sigma_n = \log n/n^{1/4}$. *Then it holds that*

$$[(1+\sigma_n u)(1+\sigma_n v)]^{\sqrt{n}r-1} = \exp\left\{-r\sqrt{n}\frac{\sigma_n^2}{2}(u^2+v^2) + r\sqrt{n}\sigma_n(u+v) + o(n^0)\right\} \quad \text{(D.10)}$$

and

$$\left[1 - \frac{r}{\sqrt{n}}\frac{\sigma_n(v+u)}{1-2\frac{r}{\sqrt{n}}}\right]^{n-2r\sqrt{n}} = \exp\left\{-r\sqrt{n}\sigma_n(u+v) + o(n^0)\right\} \quad \text{(D.11)}$$

Proof.

$$\begin{aligned}
& [(1+\sigma_n u)(1+\sigma_n v)]^{\sqrt{n}r-1} \\
= \quad & \exp\left\{\log\left[1+\sigma_n u)(1+\sigma_n v)\right]^{\sqrt{n}r-1}\right\} \\
= \quad & \exp\left\{(\sqrt{n}r-1)\left[\log(1+\sigma_n u) + \log(1+\sigma_n v)\right]\right\} \\
= \quad & \exp\left\{(\sqrt{n}r-1)\left[\sigma_n u - \frac{\sigma_n^2 u^2}{2} + \sigma_n v + \frac{\sigma_n^2 v^2}{2})\right]\right\} \\
= \quad & \exp\left\{(\sqrt{n}r-1)\left[\sigma_n(u+v) - \frac{\sigma_n^2}{2}(u^2+v^2)\right]\right\} \\
\overset{\sigma_n = o(n^0)}{=} \quad & \exp\left\{-r\sqrt{n}\frac{\sigma_n^2}{2}(u^2+v^2) + r\sqrt{n}\sigma_n(u+v) + o(n^0)\right\}
\end{aligned}$$

and

$$\begin{aligned}
& \left[1 - \frac{r}{\sqrt{n}}\frac{\sigma_n(v+u)}{1-2\frac{r}{\sqrt{n}}}\right]^{n-2r\sqrt{n}} \\
= \quad & \exp\left\{\log\left[1 - \frac{r}{\sqrt{n}}\frac{\sigma_n(v+u)}{1-2\frac{r}{\sqrt{n}}}\right]^{n-2r\sqrt{n}}\right\} \\
= \quad & \exp\left\{(n-2r\sqrt{n})\log\left[1 - \frac{r}{\sqrt{n}}\frac{\sigma_n(v+u)}{1-2\frac{r}{\sqrt{n}}}\right]\right\} \\
= \quad & \exp\left\{(n-2r\sqrt{n})\left[-\left(\frac{r}{\sqrt{n}}\frac{\sigma_n(u+v)}{1-2\frac{r}{\sqrt{n}}} + \frac{r^2\sigma_n^2(u+v)^2}{n(1-2\frac{r}{\sqrt{n}})^2}\right)\right]\right\} \\
= \quad & \exp\left\{(n-2r\sqrt{n})\left[-\frac{r}{\sqrt{n}}\frac{\sigma_n(u+v)}{1-2\frac{r}{\sqrt{n}}} - o(n^0)\right]\right\} \\
= \quad & \exp\left\{-r\sqrt{n}\sigma_n(u+v) + o(n^0)\right\}
\end{aligned}$$

□

Appendix E

Description of the algorithms and software used

E.1 R

E.1.1 In chapter 5 - Computation by a `loop` structure

Function `simulation(anzahl,n,b,r)`:

(0) Initialization `i=1, sammel=0, sammel2=0`.

(1) Loop for `i in 1:anzahl`:

- Generation of the sample $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ by `rnorm(n)`.
- Function `Anorm`: calculating the Lagrangian multiplier A in the Gaussian location model, determined by $A^{-1} = 2\Phi(g) - 1$, confer Remark 5.2.
- Function `psi`: calculating the symmetric IC of Hampel-type form by Assumption 5.1 (2): $\psi(x) = -\frac{b}{2} \vee Ax \wedge \frac{b}{2}$.
- Ordering of the sample: $x_{(1)}, \ldots, x_{(n)}$ according to approach (B) in section 8.3.
- Generation of an approximate number of modified observations $K \overset{\text{i.i.d.}}{\sim} \mathrm{Bin}(1, r/\sqrt{n})$ according to 5.1 (3).
- Test for $K < 0.5n$ by a `while` loop because of restriction (3.28).
- Changing sign for the K smallest observation.
- Calculation of the three-step estimator $\theta_n^{(3)}$ with the median as starting estimator by a three times iteration of (4.43): $\theta_n^{(k)} := \theta_n^{(k-1)} + \frac{1}{n}\sum_{i=1}^n \psi_{\theta_n^{k-1}}(X_i)$.
- Calculation of the empirical MSE `empMSE`, i.e. $\mathrm{empMSE} = n(\theta_n^{(3)})^2$.
- Function `omega`: calculating the total variation bias term from (2.53) for $p = 1$: $\omega_v(\psi) = \sup\psi - \inf\psi$.
- Calculation of `asMSE_0`, i.e. the empirical asymptotic MSE $\tilde{R}(S_n, r) = r^2(\sup\psi - \inf\psi)^2 + \mathbb{E}_{\mathrm{id}}|\psi|^2$, confer (4.3).

  - Collection of the results by `sammel[i]<-empMSE` and `sammel2[i]<-asMSE_0`.
  - `i<-i+1`

(2) Calculation and output of the averaged MSEs `emMSE<-mean(sammel)` and `asMSE<-sammel2` as a two dimensional vector.

E.1.2 In chapter 5 - Computation by matrix operation

Function `simulation2(anzahl,n,b,r)`:

(0) Function `Anorm`: calculating the Lagrangian multiplier A in the Gaussian location model, determined by $A^{-1} = 2\Phi(g) - 1$, confer Remark 5.2.

(1) Function `psi`: calculating the IC of symmetric Hampel-type form by Assumption 5.1 (2): $\psi(x) = -\frac{b}{2} \vee Ax \wedge \frac{b}{2}$.

(2) Generation of `anzahl` samples $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ by `rnorm(anzahl*n)`.

(3) Initialization of a matrix `a1` filled with `anzahl` samples $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ line by line.

(4) Ordering of each row according to approach (B) in section 8.3.

(5) Generation of an approximate number of modified observations $Ka \overset{\text{i.i.d.}}{\sim} \mathrm{Bin}(\texttt{anzahl}, r/\sqrt{n})$ according to 5.1 (3) for each row.

(6) Definition of `Kb` by $Kb = n - Ka$ for the number of unmodified observations.

(7) Generation of a matrix `K1mat` with the same dimensions as the sample matrix $a1$ containing the number -1 on all places of observations to be modified and 1 at the rest.

(8) Test for $K < 0.5n$ in each row by summation over each row because of restriction (3.28).

(9) Item-wise multiplication of the matrices $a1$ and $K1mat$ resulting in changing sign for the K smallest observations in each row.

(10) Exclusion of lines not fulfilling the test in (8).

(11) Calculation of the three step estimator $\theta_n^{(3)}$ with the median as starting estimator by a three times iteration of (4.43): $\theta_n^{(k)} := \theta_n^{(k-1)} + \frac{1}{n}\sum_{i=1}^n \psi_{\theta_n^{k-1}}(X_i)$ for each row.

(12) Calculation of the averaged empirical MSE `emMSE`, i.e. $\overline{\mathrm{empMSE}}$ with $\mathrm{empMSE} = n(\theta_n^{(3)})^2$.

(13) Calculating of the total variation bias term from (2.53) for $p = 1$: $\omega_v(\psi) = \sup\psi - \inf\psi$ for each row.

(14) Calculation of the averaged `asMSE`, i.e. the mean of empirical asymptotic MSE $\tilde{R}(S_n, r) = r^2(\sup\psi - \inf\psi)^2 + \mathbb{E}_{\mathrm{id}}|\psi|^2$, confer (4.3).

(15) Output of the averaged MSEs `emMSE` and `asMSE` as a two dimensional vector.

Function `szenario2(start,ende,anzahl2,b,r)`:

(0) Initialization `ergebnis` as a matrix with $\sharp rows =$`ende-start+1` and $\sharp columns = 3$; `i=start`.

(1) Loop for `i in start:ende`:

- Call function `simulation(anzahl2,i,b,r)` or `simulation2(anzahl2,i,b,r)`, respectively.
- Collection of the results in matrix `ergebnis`.
- `i<-i+1`

(2) Output of matrix `ergebnis` with column one containing `[start, ende]`, column two `emMSE` and column three `emMSE - asMSE`.

E.1.3 In chapter 7 - maximal asymptotic MSE up to second order

(0) Function `asMSE0`: calculation of the asymptotically optimal first order MSE $R(S_n, r) = r^2b^2 + v_{id,0}^2$ from (4.3)

(1) Function `asMSE1`: calculation of the A_1-term $A_1 = (\pm l_{id,2}b^3 + 2l_{c,1}b^2)r^2 + v_{id,0}^2(2l_{c,1} + 2v_{c,0} \pm b(3l_{id,2} + 4v_{id,1}))$ from (6.19) and returning the asymptotically optimal second order MSE by `asMSE(r,b,vid0)+A1*r/sqrt(n)`

(2) Function `asMSE2symm`: calculation of the A_2-term for the symmetric case, confer Corollary 6.19:

$$\begin{aligned} A_2 \;=\; & (\frac{1}{3}l_{id,3}b^4 \pm l_{c,2}b^3)r^4 + (\pm(4v_{c,1} + 3l_{c,2})v_{id,0}^2 b \\ & +((2l_{id,3} + 3v_{id,2})v_{id,0}^2)b^2)r^2 + (l_{id,3} + 3v_{id,2})v_{id,0}^4 \\ & +\frac{2}{3}\rho_1 v_{id,0}^3 \end{aligned}$$

and returning the asymptotically optimal third order MSE by `asMSE1(n,r,b,vid0,0,0,0,0,1)+A2/n`.

(3) Function `Vnorm`: calculating $V_{id}(0) = v_{id,0}^2$ in the Gaussian location model as in (6.105), i.e. $V_{id}(0) = v_{id,0}^2 = 2A^2g^2(1 - \Phi(g)) + A^2(1 - 2g\varphi(g))$.

(4) Function `Anorm`: calculating the Lagrangian multiplier A in the Gaussian location model, determined by $A^{-1} = 2\Phi(g) - 1$, confer Remark 5.2.

(5) Function `lid3norm`: calculating $l_{id,3}$ in the Gaussian location model by $l_{id,3} = 2g\varphi(g)/(2\Phi(g) - 1)$, confer (6.104).

(6) Function `vid2norm`: calculating $v_{id,2}$ in the Gaussian location model by $v_{id,2} = \frac{6\Phi(g)-4\Phi(g)^2-2-2g\varphi(g)}{2g^2(1-\Phi(g))+2\Phi(g)-1-2g\varphi(g)}$, confer (6.107).

(7) Function `r1norm`: calculating ρ_1 in the Gaussian location model by $\rho_1 = \frac{3A^3(1-2\Phi(g)+2g\varphi(g))}{v_{id,0}^3} + 3v_{id,0}^{-1}$, confer (6.108).

(8) Function `lc2norm`: calculating $l_{c,2}$ in the Gaussian location model by $l_{c,2} = \frac{2g^3\varphi(g)}{2\Phi(g)-1} = g^2 \cdot l_{id,3}$, confer (6.103).

(9) Function `vc1norm`: calculating $v_{c,1}$ in the Gaussian location model by $v_{c,1} = \frac{-g^3\varphi(g)A^2}{v_{id,0}^2}$, confer (6.106).

(10) Functions `asMSE0.norm, asMSE1.norm, asMSE2.norm`: calculation of the asymptotically optimal first, second and - in the symmetric case - third order MSE in the Gaussian location model by application of all previously defined functions.

E.2 MAPLE

In chapter 6 - Higher Order Algorithms

E.2.1 In chapter 6 - Higher Order Algorithms

The following procedures were originally developed by P. Ruckdeschel in [Ruckdeschel (2005b)] for the examination of higher order asymptotics for the MSE on convex contamination neighborhoods. They have been modified in order to be applied to the total variation case. But as the structure of the procedures hardly changed, we stick by the procedure names given by P. Ruckdeschel.

(0) Procedure `asS`: calculation of asymptotic expansion of $s_n(t) = \frac{-r\tilde{t}-\sqrt{n}\ \tilde{L}_{\mathrm{re}}(t)}{V_{\mathrm{re}}(t)}$ from (6.24).

(1) Procedure `asS1`: calculation of asymptotic expansion of $s_n'(t)$.

(2) Procedure `asg`: calculation of asymptotic expansion of $\tilde{g}_n(u) = v_{id,0}\varphi(\tilde{s})[1+\frac{1}{\sqrt{n}}P_1(u,t^\natural)+\frac{1}{n}P_2(u,t^\natural)] + o(n^{-(1+\delta)})$ from (6.53) with the terms from Theorem A.5.

(3) Procedure `dfac`: calculation of asymptotic expansion of $\varphi(\tilde{s})$ about $s_1 = (u - t^\natural)/v_{id,0}$ from (6.54) as $\varphi(\tilde{s}) = \varphi(s_1)[1 - s_1(\tilde{s} - s_1) + (s_1^2 - 1)(\tilde{s} - s_1)^2/2] + o(n^{-(1+\delta)})$, confer (6.55).

(4) Procedure `asgns`: calculation of asymptotic expansion of $g_n(s_1) := 1 + \frac{1}{\sqrt{n}}\tilde{P}_1(s_1,t^\natural) + \frac{1}{n}\tilde{P}_2(s_1,t^\natural)$ from (6.57).

(5) Procedure `ashn`: calculation of asymptotic expansion of $h_n(s) = (sv_{id,0} + t^\natural)^2 + \frac{1}{\sqrt{n}}Q_1(+\frac{1}{n}Q_2)$ from (6.60).

(6) Procedure `HN`: sort by t in the asymptotic expansion of h_n.

(7) Procedure `HND2s`: calculation of the second partial derivative w.r.t. $t^\natural$ of h_n to achieve $h_{n,t,t}(s) = 2 + \frac{1}{\sqrt{n}}Q_{1,t,t}(+\frac{1}{n}Q_{2,t,t})$ as in (6.61).

(8) Procedure `ash`: substitution $t^\natural = \pm rb$ in h_n.

(9) Procedure `intesout`: integrating out s in h_n by usage of moments of $\mathcal{N}(0,1)$ as stated in (6.67).

(10) Procedure `asMSEK`: calculation of the final asymptotic expansion of the MSE in (6.68)

E.2.2 Translation Table

As a guidance to read the `MAPLE` code and its results, especially, we offer a translation table.

Plain Text	`MAPLE`
r	`r`
$\tilde{t}$	`t`
t	`u`
$\sqrt{n}$	`W`
$l_{\mathrm{id},2}$	`l3`
$l_{\mathrm{id},3}$	`l5`
$l_{\mathrm{c},1}$	`l2`
$l_{\mathrm{c},2}$	`l4`
$l_{\mathrm{c},3}$	`l6`
$v_{\mathrm{id},0}$	`v0`
$v_{\mathrm{id},1}$	`v2`
$v_{\mathrm{id},2}$	`v4`
$v_{\mathrm{c},0}$	`v1`
$v_{\mathrm{c},1}$	`v3`
$v_{\mathrm{c},2}$	`v5`
ρ_0	`r0`
ρ_1	`r1`

E.3 SWEAVE

SWEAVE provides a flexible framework for mixing text and R code for an automatic document generation. A single source file contains both documentation text and R code, which are then woven into a final document containing the documentation text together with the R code and the output of the code by running the R code through the R engine. This allows to regenerate a report if the input data change. The R code of the complete analysis is embedded into a LaTeXdocument using the noweb syntax, confer [Ramsey (1998)]. For more details on SWEAVE we refer to [Leisch (2002a)], [Leisch (2002b)] and [Leisch (2003)].

Appendix F

Further classical results of asymptotic statistics

We assume a parametric family

$$\mathcal{P} = \{P_\theta \,|\, \theta \in \Theta\} \subset \mathcal{M}_1(\mathcal{A})$$

of probability measures on some sample space $(\Omega, \mathcal{A})$, whose parameter space Θ is an open subset of some finite dimensional $\mathbb{R}^k$. For a more detailed introduction to this topics we also refer to Chapter 2 of [Bickel et al. (1998)] and Chapters 6 - 9 of [van der Vaart (1998)], respectively.
We consider a parameter sequence (θ_n) about θ of the form

$$\theta_n = \theta + \frac{t_n}{\sqrt{n}} \qquad\qquad t_n \to t \in \mathbb{R}^k \tag{F.1}$$

Corresponding to this parametric alternatives (θ_n) two sequences of product measures are defined on the n-fold product measurable space $(\Omega^n, \mathcal{A}^n)$

$$P_\theta^n = \bigotimes_{i=1}^n P_\theta \qquad\qquad P_{\theta_n}^n = \bigotimes_{i=1}^n P_{\theta_n} \tag{F.2}$$

Theorem F.1. *If $\mathcal{P}$ is L_2 differentiable at θ, its L_2 derivative Λ_θ is uniquely determined in $L_2^k(P_\theta)$. Moreover,*

$$\mathbb{E}_\theta \Lambda_\theta = 0$$

and the alternatives given by (F.1) and (F.2) have the log likelihood expansion

$$\log \frac{dP_{\theta_n}^n}{dP_\theta^n} = \frac{t^\tau}{\sqrt{n}} \sum_{i=1}^n \Lambda_\theta(y_i) - \frac{1}{2} t^\tau \mathcal{I}_\theta t + o_{P_\theta^n}(n^0)$$

where

$$\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \Lambda_\theta(y_i) \right) \left(P_\theta^n\right) \overrightarrow{w} \mathcal{N}(0, \mathcal{I}_\theta)$$

Proof. Special case of Theorem 2.3.7 in [Rieder (1994)] □

The next result of asymptotic statistics that is known as *Le Cam's third lemma*.

Theorem F.2. *Let $P_n, Q_n \in \mathcal{M}_1(\mathcal{A}_n)$ be two sequences of probabilities with log likelihoods $L_n \in \log \frac{dQ_n}{dP_n}$, and S_n a sequence of statistics on $(\Omega_n, \mathcal{A}_n)$ taking values in some finite-dimensional $(\bar{\mathbb{R}}^p, \bar{\mathbb{B}}^p)$ such that for $a, c \in \mathbb{R}^p$, $\sigma \in [0, \infty)$ and $C \in \mathbb{R}^{p \times p}$,*

$$\begin{pmatrix} S_n \\ L_n \end{pmatrix} (P_n) \stackrel{\rightarrow}{w} \mathcal{N}\left(\begin{pmatrix} a \\ -\sigma^2/2 \end{pmatrix}, \begin{pmatrix} C & c \\ c^\tau & \sigma^2 \end{pmatrix} \right)$$

then

$$\begin{pmatrix} S_n \\ L_n \end{pmatrix} (Q_n) \stackrel{\rightarrow}{w} \mathcal{N}\left(\begin{pmatrix} a + c \\ \sigma^2/2 \end{pmatrix}, \begin{pmatrix} C & c \\ c^\tau & \sigma^2 \end{pmatrix} \right)$$

Proof. [Rieder (1994)], Corollary 2.2.6. □

The following definition corresponds to Definition 2.2.9 of [Rieder (1994)].

Definition F.3. *A sequence $(\mathcal{Q}_n)$ of statistical models on sample spaces $(\Omega_n, \mathcal{A}_n)$,*

$$\mathcal{Q}_n = \{Q_{n,t} \,|\, t \in \Theta_n\} \subset \mathcal{M}_1(\mathcal{A}_n)$$

with the same finite-dimensional parameter space $\Theta_n = \mathbb{R}^k$ (or at least $\Theta_n \uparrow \mathbb{R}^k$) is called asymptotically normal, *if there exists a sequence of random variables $Z_n \colon (\Omega_n, \mathcal{A}_n) \to (\mathbb{R}^k, \mathbb{B}^k)$ that are asymptotically normal,*

$$Z_n(Q_{n,0}) \stackrel{\rightarrow}{w} \mathcal{N}(0, C)$$

with positive definite covariance $C \in \mathbb{R}^{k \times k}$, and such that for all $t \in \mathbb{R}^k$ the log likelihoods $L_{n,t} \in \log \frac{dQ_{n,t}}{dQ_{n,0}}$ have the approximation

$$L_{n,t} = t^\tau Z_n - \frac{1}{2} t^\tau C t + o_{Q_{n,0}}(n^0)$$

The sequence $Z = (Z_n)$ is called the asymptotically sufficient statistic *and C the* asymptotic covariance *of the* asymptotically normal models *$(\mathcal{Q}_n)$.*

We present the convolution and the asymptotic minimax theorems in the parametric case; confer Theorems 3.2.3, 3.3.8 of [Rieder (1994)]. These two mathematical results of asymptotic statistics are mainly due to Le Cam and Hájek.

Assume a sequence of statistical models $(\mathcal{Q}_n)$ on sample spaces $(\Omega_n, \mathcal{A}_n)$,

$$\mathcal{Q}_n = \{Q_{n,t} \,|\, t \in \Theta_n\} \subset \mathcal{M}_1(\mathcal{A}_n)$$

with the same finite-dimensional parameter space $\Theta_n = \mathbb{R}^k$ (or $\Theta_n \uparrow \mathbb{R}^k$). The parameter of interest is Dt for some $p \times k$-matrix D of full rank $p \leq k$. Moreover we consider asymptotic estimators

$$S = (S_n) \qquad S_n \colon (\Omega_n, \mathcal{A}_n) \to (\mathbb{R}^p, \mathbb{B}^p)$$

We start with Definition 3.2.2 of [Rieder (1994)].

Definition F.4. *An asymptotic estimator S is called* regular *for the parameter transform D, with limit law $M \in \mathcal{M}_1(\mathbb{B}^p)$, if for all $t \in \mathbb{R}^k$,*

$$(S_n - Dt)(Q_{n,t}) \stackrel{\rightarrow}{w} M$$

that is, $S_n(Q_{n,t}) \stackrel{\rightarrow}{w} M * \mathbb{I}_{Dt}$ as $n \to \infty$, for every $t \in \mathbb{R}^k$.

We now may state the convolution theorem.

Theorem F.5. *Assume models $(\mathcal{Q}_n)$ that are asymptotically normal with asymptotic covariance $C \succ 0$ and asymptotically sufficient statistic $Z = (Z_n)$. Let $D \in \mathbb{R}^{p\times k}$ be a matrix of rank $p \leq k$. Let the asymptotic estimator S be regular for D with limit law M. Then there exists a probability $M_0 \in \mathcal{M}_1(\mathbb{B}^p)$ such that*

$$M = M_0 * \mathcal{N}(0, \Gamma) \qquad \Gamma = DC^{-1}D^\tau$$

and

$$(S_n - DC^{-1}Z_n)(Q_{n,0}) \stackrel{\rightarrow}{w} M_0$$

An asymptotic estimator S^* is regular for D and achieves limit law $M^* = \mathcal{N}(0, \Gamma)$ iff

$$S_n^* = DC^{-1}Z_n + o_{Q_{n,0}}(n^0)$$

Proof. Three variants of the proof are given in [Rieder (1994)], Theorem 3.2.3. □

For the specification of the asymptotic minimax theorem we need the definition of the set L of loss functions; confer pp. 78, 81 of [Rieder (1994)].

Definition F.6. *Let* L *be the set of all Borel measurable functions $\ell\colon \bar{\mathbb{R}}^p \to [0,\infty]$ that are*
(a) symmetric subconvex *on $\mathbb{R}^p$; that is, for all $z \in \mathbb{R}^p$ and all $c \in [0,\infty]$,*

$$\ell(z) = \ell(-z) \qquad \{z \in \mathbb{R}^p \,|\, \ell(z) \leq c\} \text{ is convex}$$

(b) upper semicontinuous at infinity*; that is, for every sequence $z_n \in \mathbb{R}^p$ with $z_n \to z \in \bar{\mathbb{R}}^p \setminus \mathbb{R}^p$,*

$$\limsup_{n\to\infty} \ell(z_n) \leq \ell(z)$$

This functions $\ell \in$ L will be called loss functions*. If there is an increasing function $v\colon [0,\infty] \to [0,\infty]$ and a symmetric positive definite matrix $A \in \mathbb{R}^{p\times p}$, then a loss function of type,*

$$\ell(z) = \begin{cases} v(z^\tau A z) & \text{if } |z| < \infty \\ v(\infty) & \text{if } |z| = \infty \end{cases}$$

will be called monotone quadratic.

For part (a) of the asymptotic minimax theorem we assume Θ_n open. Moreover asymptotic estimators with extended values can be allowed; i.e.,

$$S = (S_n) \qquad S_n\colon (\Omega_n, \mathcal{A}_n) \to (\bar{\mathbb{R}}^p, \bar{\mathbb{B}}^p)$$

Theorem F.7. *Assume models $(\mathcal{Q}_n)$ that are asymptotically normal with asymptotic covariance $C \succ 0$. Let $D \in \mathbb{R}^{p\times k}$ be a matrix of rank $p \le k$. Put*

$$\rho_0 = \int \ell \, d\mathcal{N}(0,\Gamma) \qquad \Gamma = DC^{-1}D^\tau$$

for any Borel measurable function $\ell\colon \mathbb{R}^p \to [0,\infty]$.
(a) *Then, if $\ell \in \mathrm{L}$ and ℓ is lower semicontinuous on $\bar{\mathbb{R}}^p$,*

$$\lim_{b\to\infty}\lim_{c\to\infty}\liminf_{n\to\infty}\inf_{S}\sup_{|t|\le c}\int b\wedge\ell(S_n - Dt)\,dQ_{n,t} \ge \rho_0$$

(b) *Suppose $\ell\colon \mathbb{R}^p \to [0,\infty]$ is continuous a.e. λ^ploss function!continuous a.e. λ^p and the asymptotic estimator S^* is asymptotically normal for every $c \in (0,\infty)$, uniformly in $|t| \le c$,*

$$(S_n^* - Dt)(Q_{n,t}) \xrightarrow{w} \mathcal{N}(0,\Gamma)$$

Then for all $c \in (0,\infty)$

$$\lim_{b\to\infty}\lim_{n\to\infty}\sup_{|t|\le c}\int b\wedge\ell(S_n^* - Dt)\,dQ_{n,t} = \rho_0$$

and necessarily

$$S_n^* = DC^{-1}Z_n + o_{Q_{n,0}}(n^0)$$

Proof. [Rieder (1994)], Theorem 3.3.8 and Remark 3.3.9 (d). □

The following lemma corresponds to Lemma 4.2.18 of [Rieder (1994)]. It is a consequence of Theorem F.1 together with Slutzky's lemma, the Cramér-Wold device and Le Cam's third lemma (Theorem F.2).

Lemma F.8. *Let the ALE S have the asymptotic expansion (2.20) involving some function $\psi_\theta \in L_2^k(P_\theta)$, $\mathbb{E}_\theta\psi_\theta = 0$. Then*

$$\mathbb{E}_\theta\psi_\theta\Lambda_\theta^\tau = \mathbb{I}_k$$
holds iff
$$\sqrt{n}\,(S_n - \theta)(P_{\theta+t_n/\sqrt{n}}) \xrightarrow{w} \mathcal{N}(t, \mathrm{Cov}_\theta(\psi_\theta))$$

for all convergent sequences $t_n \to t$ in $\mathbb{R}^k$.

In the parametric setup, and restricted to the class of ALEs, the convolution theorem (Theorem F.5) and the local asymptotic minimax theorem (Theorem F.7) coincide with the Cramér-Rao bound. Specializing these two theorems to the parametric context we get

$$\mathcal{Q}_n = \left\{P^n_{\theta+t/\sqrt{n}} \,|\, t \in \mathbb{R}^k\right\} \subset \mathcal{M}_1(\mathcal{A})$$

Then we have with Proposition 4.2.19 of [Rieder (1994)].

Proposition F.9. (a) *Let an asymptotic estimator R be regular for t with limit law $M \in \mathcal{M}_1(\mathbb{B}^k)$. Then there is a probability $M_0 \in \mathcal{M}_1(\mathbb{B}^k)$ such that*

$$M = M_0 * \mathcal{N}(0, \mathcal{I}_\theta^{-1})$$

A regular estimator R^* achieves the limit law $M^* = \mathcal{N}(0, \mathcal{I}_\theta^{-1})$ iff R^* is the standardization of an estimator S^* that is asymptotically linear at P_θ with IC $\psi_{h,\theta}$.
(b) *Let the loss function $\ell \in \mathrm{L}$ be lower semicontinuous on $\bar{\mathbb{R}}^k$. Then*

$$\lim_{b\to\infty} \lim_{c\to\infty} \liminf_{n\to\infty} \inf_R \sup_{|t|\le c} \int b \wedge \ell(R_n - t)\, dP^n_{\theta+t/\sqrt{n}} \ge \rho_0$$

where

$$\rho_0 = \int \ell\, d\mathcal{N}(0, \mathcal{I}_\theta^{-1})$$

If the function $\ell\colon \mathbb{R}^k \to [0,\infty]$ is continuous a.e. λ^k, and the estimator S^* is asymptotically linear at P_θ with IC $\psi_{h,\theta}$, then

$$\lim_{b\to\infty} \lim_{c\to\infty} \lim_{n\to\infty} \sup_{|t|\le c} \int b \wedge \ell(\sqrt{n}\,(S_n^* - \theta) - t)\, dP^n_{\theta+t/\sqrt{n}} = \rho_0$$

For regular ALEs in the sense of Definition F.4 we obtain the Cramér-Rao bound.

Proposition F.10. *Consider an estimator $S = (S_n)$ that is asymptotically linear at P_θ with IC $\rho_\theta \in \Psi_2(\theta)$.*
(a) *Then its standardization R is regular with normal limit law*

$$\mathcal{N}(0, Cov_\theta(\rho_\theta)) = \mathcal{N}(0, Cov_\theta(\rho_\theta) - \mathcal{I}_\theta^{-1}) * \mathcal{N}(0, \mathcal{I}_\theta^{-1})$$

(b) *Assume a loss function $\ell \in \mathrm{L}$ that is continuous a.e. λ^kloss function!continuous a.e. λ^k. Then*

$$\lim_{b\to\infty} \lim_{c\to\infty} \lim_{n\to\infty} \sup_{|t|\le c} \int b \wedge \ell(\sqrt{n}\,(S_n - \theta) - t)\, dP_{\theta+t/\sqrt{n}}$$
$$= \int \ell\, d\mathcal{N}(0, Cov_\theta(\rho_\theta)) \ge \int \ell\, d\mathcal{N}(0, \mathcal{I}_\theta^{-1})$$

The lower bound is achieved by $\rho_\theta = \psi_{h,\theta}$. If ℓ is monotone quadratic and not constant a.e. λ^k, the lower bound can be achieved only by $\rho_\theta = \psi_{h,\theta}$.

Proof. [Rieder (1994)], Proposition 4.2.20. □

Appendix G

Errata

- Page ii, formula (2): $\eta_c = A\Lambda_f \min\left\{1, \frac{c}{|\Lambda_f|}\right\}$
- Page ii, line 18: ~~n a und~~
- Page iii, line 33: ~~$\in \mathcal{M}_1(\mathbb{B})$~~
- Page vi, line 5: ~~Voraussetzung~~ Ursache
- Page xx, line 5: $[-1, 1] \subset \mathbb{R}$
- Page xx, line 24: ~~$\mathcal{X}$~~ χ
- Page xx, line 37: ~~$\Delta \in \mathcal{M}_1(\mathbb{B})$~~ with total mass 0
- Page xxi, line 26: $|O(r_n)/r_n|(P_n)$
- Page xxii, line 32: C_c^∞
- Page 1, line 6: at rate $1/\sqrt{n}$
- Page 2, formula (1.2): $\eta_c = A\Lambda_f \min\left\{1, \frac{c}{|\Lambda_f|}\right\}$
- Page 2, line 6: ~~s a and~~
- Page 3, line 3: ~~$\in \mathcal{M}_1(\mathbb{B})$~~
- Page 9, line 3: ~~distribution~~ neighborhood
- Page 9, line 4: ... of an ...
- Page 9, line 7: $r \in [0, \sqrt{n}]$
- Page 9, second formula: $a := \inf\left\{t > 0 | \mathrm{MSE}_{Q_n(r,t)}(\bar{X}_n) > \mathrm{MSE}_{Q_n(r,t)}(S_n^c)\right\}$
- Page 9, line 28: ... deviations can have ...
- Page 9, footnote 1: ... [Ruckdeschel (2006)], ...

- Page 13, line 0: Let
$$\mathcal{P} = \{P_\theta \,|\, \theta \in \Theta\} \subset \mathcal{M}_1(\mathcal{A})$$
a family of probability measures on some sample space $(\Omega, \mathcal{A})$, with an open parameter set $\Theta \subset \mathbb{R}^k$ of finite dimension k. Fix $\theta \in \Theta$.
- Page 13, in formula (2.12): $\mathbb{E}_\theta \psi_\theta \Lambda_\theta^\tau = \mathbb{I}_k$
- Page 13, in formula (2.13): $\mathbb{E}_\theta \psi_\theta \Lambda_\theta^\tau = D$
- Page 16, line 20: ... Appendix F.
- Page 16, line 32: ... [Ruckdeschel (2006)], ...
- Page 18, in formula (2.37): $\mathcal{N}_k(r\mathbb{E}_\theta \psi_\theta q,\, \mathrm{Cov}_\theta(\psi_\theta))$
- Page 18, line 29: ~~for all convergent sequences $t_n \longrightarrow t$ in $\mathbb{R}^k$.~~
- Page 19, line 29: ... mean squared error ...
- Page 25, line 23: ...of A. It is ...
- Page 27, Remark 3.4: a) ~~Simply~~ Roughly ...
- Page 37, line 4 to 5: ... the size of the bias. But ..., ~~the evaluated bias terms coincide and so does~~ the MSE coincide.
- Page 41, in formula (4.43): $\eta_{\theta_n}^{(k-1)}$
- Page 42, in Ass. 5.1 (3): $K \sim \mathrm{Bin}(n, r/\sqrt{n})$
- Page 42, in Rem. 5.2: $A^{-1} = 2\Phi(g/2) - 1$
- Page 43, in Proof: ~~g~~ $\frac{g}{2}$
- Page 45, line 7: ~~numerical~~
- Page 61, line 2: ~~(6.18)~~ (6.3)
- Page 64, line 0: By Remark (6.12) c) we can simplify the expressions for $L_{\mathrm{re}}(t)$ and $V_{\mathrm{re}}(t)$ of Lemma 6.6:
$$L_{\mathrm{re}}(t) = \frac{r}{\sqrt{n}} l_{c,0} + (-1 + \frac{r}{\sqrt{n}} l_{c,1})t + (l_{id,2} + \frac{r}{\sqrt{n}} l_{c,2})\frac{t^2}{2} + O(t^3)$$
and
$$V_{\mathrm{re}}^2(t) = \left(V_{id,0} + \frac{r}{\sqrt{n}} V_{c,0}\right) + \left(V_{id,1} + \frac{r}{\sqrt{n}} V_{c,1}\right) t + O(t^2).$$

For our purpose we are interested in the square root of the last expression. As we do not want to loose the structure of (G) we use the Taylor expansion of the square root up to first order.

$$V_{\mathrm{re}}(t) = v_{id,0}\Big[(1 + \frac{r}{\sqrt{n}}v_{c,0}) + (v_{id,1} + \frac{r}{\sqrt{n}}v_{c,1})t\Big] + O(t^2).$$

with $v_{id,0} := \sqrt{V_{id,0}}$ and $v_{*,i} := \frac{V_{*,i}}{2V_{id,0}}$ otherwise.

- Page 65, in last formula: ~~$\mathbb{P}$~~ Q_n
- Page 66, in (3): ~~$\mathbb{P}$~~ Q_n
- Page 68, in formula (6.23): ~~$\mathbb{P}$~~ Q_n
- Page 68/69: ~~By Remark (6.12) c) we can simplify ... and $v_{*,i} := \frac{V_{*,i}}{2V_{id,0}}$ otherwise.~~
- Page 70, in formula (6.31): ~~$\mathbb{P}$~~ Q_n
- Page 71, in formula (6.34): ~~$\mathbb{P}$~~ Q_n
- Page 75, in Ad (13): ~~Q_{T}~~ B_1, ~~$Q_{1,t,t}$~~ $B_{1,t,t}$
- Page 76, in formulas (6.64) and (6.65): ... concentrated ...
- Page 78, in Proof: $(\psi^0)^2(x-t)$, $\psi(x-t)$
- Page 80, in Cor. 6.20: $A^0_{2,*}$
- Page 81, in Lem. 6.21: ~~$\in \mathcal{M}_1(\mathbb{B})$~~
- Page 82, in formula (6.94): ...+$\chi''_{\xi_j}(0)\frac{t^2}{2n}$+...
- Page 89, line 13: ... ~~second~~ third-order ...
- Page 89, in Rem. 7.2: $b \cdot l_{c,2}|_{(6.20)}(b) = -b \cdot l_{c,2}|_{(6.21)}(-b) > 0$
- Page 90, line 10: ~~e~~ g
- Page 90, label to Fig. 7.1: ~~Numerical~~
- Page 92, line 25: ~~to choose~~
- Page 93, footnote 1: ... c_n to be negligible ...
- Page 95, Proof of Lem. 8.5: ~~$P(X_i \text{ modified, } K = k) = \frac{k}{n}$ ... ... $d_v(F,Q) = \frac{1}{2}\int_{I\cup II\cup III}|dF - dQ| = \frac{1}{2}\left(\frac{r}{\sqrt{n}} + 0 + \frac{r}{\sqrt{n}}\right) = \frac{r}{\sqrt{n}}$~~

$$P(\{X_i \text{ modified}\}|K = k) = \frac{k}{n}$$

and so

$$\mathbb{E}I(\{X_i \text{ modified }\}) = \mathbb{E}\mathbb{E}[I(\{X_i \text{ modified }\})|K] = \frac{r}{\sqrt{n}}.$$

Hence

$$d_v(F,Q) = \frac{1}{2}\int I(\{X_i \text{ modified }\})|dF - dQ| \leq \frac{r}{\sqrt{n}}$$

with $=$ when in order to attain Q the total F-mass is transported from I to III or the other way round.

- Page 97, in Lem 8.9: ~~A_0~~ B_0, ~~A_1~~ B_1
- Page 103, line 1: ...from an assumed ...
- Page 107, in Rem 8.13: ~~numerical~~
- Page 107, in Theorem 8.14: $\mathbb{E}K$, Var$K$
- Page 108, line 2: $\mathbb{E}B_0^2 = ...$
- Page 111, line 2: $\mathbb{E}K^3 = \mathbb{E}(K_1 + \mu)^3$
- Page 127, line 8: $= P\left(\frac{\text{Bin}(n,\rho/\sqrt{n})}{n} - \frac{\rho}{\sqrt{n}} < \frac{r+\eta-\rho}{\sqrt{n}}\right)$
- Page 128, line 18: $= \frac{\rho n^{-1/2}}{\sqrt{\log n}\sqrt{\log(C^2/\rho^2)/\log n + 1}}$
- Page 131, in (PK): $P(|K - r\sqrt{n}| > \eta\sqrt{n}) = ...$
- Page 131, line 23: $... = \sqrt{2r\sqrt{n}} = O(n^{1/4}) < o(\sqrt{n})$
- Page 135, line 19: ... $K \sim \text{Bin}(n, r/\sqrt{n})$.
- Page 135, line 21: ... $K \sim P$ with $\mathbb{E}K = r\sqrt{n}$ ...
- Page 142, line 5: ...in case (III),
- Page 151, in (1): ... $K \sim \text{Bin}(n, r/\sqrt{n})$ ...
- Page 152, in (5): ... $Ka \sim \text{Bin}(\texttt{anzahl}, r/\sqrt{n})$ ...
- Page 155, in (5): ~~Q_1~~ B_1, ~~Q_2~~ B_2
- Page 155, in (7): ~~$Q_{1,t,t}$~~ $B_{1,t,t}$, ~~$Q_{2,t,t}$~~ $B_{2,t,t}$
- Page 157, line 1: ~~that~~
- Page 159, line 7: ~~loss function!continuous a.e. λ^p~~
- Page 160, line 18: ~~loss function!continuous a.e. λ^k~~
- Page 165: [Ruckdeschel (2006)] ...

Bibliography

[Abramowitz and Stegun (1984)] Abramowitz M. and Stegun I.A. (Eds.) 1984: *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* National Bureau of Standards, Washington, D.C. Reprint of the 1972 edition. 5.4.1

[Aldous et al. (1985)] Aldous D.J., Ibragimov I.A., Jacod J. 1985: *École d'Été de Probabilités de Saint-Flour XIII - 1983.* Lecture Notes in Mathematics. Springer.

[Artzner et. al. (1998)] Artzner P., Delbaen F., Eber J.-M. and Heath D. 1999: *Coherent Measures of Risk. Mathematical Finance*, Vol. **9**, No. **3**, p. 203-228.

[Barndorff-Nielsen and Cox (1994)] Barndorff-Nielsen O. and Cox D. 1994: *Inference and asymptotics.*, Vol. 52 of *Monographs on Statistics and Applied Probability*. Chapman and Hall.

[BCBS (2004)] Basel Committee on Banking Supervision 2004: *International convergence of capital measurement and capital standards, a revised framework.*

[Beekmann and Stemper (2006)] Beekmann F. and Stemper P. 2006: *Ein Modell zur Quantifizierung operationeller Risiken in Banken* In: *OR News*, **26**.

[Bickel (1975)] Bickel P.J. 1975: *One-step Huber estimates in the linear model.* In: *J. Am. Statist. Assoc.*, **70**, 428–434.

[Bickel (1981)] —— 1981: Quelques aspects de la statistique robuste. In: *Ecole d'ete de probabilites de Saint-Flour IX-1979. Lect. Notes Math.*, **876**, 2–72.

[Bickel (1984)] —— 1984: *Robust regression based on infinitesimal neoghborhoods.* Ann. Stat., 12: 1349-1368.

[Bickel and Doksum (2001)] Bickel P.J. and Doksum K.A. 2001: *Mathematical statistics: basic ideas and selected topics.* Prentice Hall. 2nd edition.

[Bickel et al. (1998)] Bickel P.J., Klaassen C.A., Ritov Y. and Wellner J.A. 1998: *Efficient and adaptive estimation for semiparametric models..* Springer.

[Borodin and Saminen (1996)] Borodin A. N. and Saminen P. 1996: *Handbook of Brownian Motion - Facts and Formulae* Probability and its applications. Birkhäuser Basel.

[Box and Cox (1964)] Box, G. E. P. and Cox, D. R. 1964: *An analysis of transformations (with discussion). Journal of the Royal Statistical Society B*, **26**, 211–252.

[Brandl (2003)] Brandl M. 2003: *Das Poincarésche Zentrumproblem - Nicht degenerierter und degenerierter Fall* Diploma Thesis and Zulassungsarbeit. Also available in `http://www.old.uni-bayreuth.de/departments/math/org/mathe6/publ/da/brandl/brandl.pdf`

[Cont et. al. (2007)] Cont R., Deguest R. and Scandolo G. 2007: *Robustness and sensitivity analysis of risk measurement procedures Financial Engineering Report*, **2007-06**: 1–34.

[Delbaen (2002)] Delbaen F. 2002: *Coherent measures of risk on general probability spaces.* In: *Advances in Finance and Stochastics.* Essay in Honour of Dieter Sondermann. Springer, p. 1-37.

[Donoho and Huber (1983)] Donoho D.L. and Huber P.J. 1983: The notion of breakdown point. In: *A Festschrift for Erich L. Lehmann*, (P.J. Bickel, K. Doksum and J.L. Hodges, Jr., eds.), p. 157–184. Wadsworth, Belmont, CA.

[Dudley (1989)] Dudley R.M. 1989: *Real analysis and probability.* Wadsworth & Brooks/Cole, Pacific Grove, Calif.

[Durrett (1999)] Durrett R. 1999: *Essentials of Stochastic Processes* Springer texts in statistics. Springer.

[Esseen (1945)] Esseen C.-G. 1945: *Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law* In: Acta Math., **77**: 1–125.

[Feller (1971)] Feller W. 1971: *An Introduction to Probability Theory and Its Applications, Vol. II* John Wiley & Sons, Inc. New York.

[Fernandes et. al. (2007)] Fernandes J. L. B., Hasman A. and Peña J. I. (2007): *Risk premium: insights over the threshold Applied Financial Economics*, **18:1**, p. 41–59.

[Fernholz (1983)] Fernholz, L. T. 1983: *Von Mises Calculus for Statistical Functionals* Lecture Notes in Statistics 19. Springer, New York.

[Field and Ronchetti (1990)] Field C. and Ronchetti E. 1990: *Small sample asymptotics*, Vol. 13 of *IMS Lecture Notes - Monograph Series.* Institute of Mathematical Statistics., Hayward, CA.

[Filippova (1961)] Filippova, A. A. 1961: *Mises' theorem on the statistical behavior of functionals of empirical distribution functions and its statistical applications Theory Prob. Appl.* **7**: 24-57.

[Fischer and Lieb (1992)] Fischer W. and Lieb I. 1992: *Funktionentheorie* Vieweg, Braunschweig/Wiesbaden. 6. edition.

[Fraiman et al. (2001)] Fraiman R., Yohai V.J. and Zamar R.H. 2001: Optimal robust M-estimates of location. *Ann. Stat.*, **29**(1): 194–223.

[Gordon (1941)] Gordon R. D. 1941: Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Ann. Math. Stat.*, **12**: 364–366.

[Hall (1992)] Hall P. 1992: *The bootstrap and Edgeworth expansion.* Springer Series in Statistics. Springer-Verlag.

[Hampel (1968)] Hampel F.R. (1968): *Contributions to the theory of robust estimation.* Dissertation, University of California, Berkely, CA.

[Hampel (1974)] —— 1974: The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.*, **69**: 383–393.

[Hampel et al. (1986)] Hampel F.R., Ronchetti E.M., Rousseeuw P.J. and Stahel W.A. 1986: *Robust statistics. The approach based on influence functions.* Wiley Series in Probability and Mathematical Statistics. Wiley.

[Harville (1997)] Harville D. A. 1997: *Matrix algebra from a statistician's perspective.* New York: Springer.

[Hastie and Pregibon (1992)] Hastie, T. J. and Pregibon, D. 1992: *Generalized linear models.* , Chapter 6 of *Statistical Models* In: *S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.*

[Hoeffding (1956)] Hoeffding W. 1956. *"On the Distribution of the Number of Successes in Independent Trials"* In: Annals of Mathematical Statistics, **27**, 713–721.

[Hoeffding (1963)] —— 1963: Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.*, **58**: 13–30.

[Hoeffding and Wolfowitz (1958)] Hoeffding W. and Wolfowitz J. 1958: *Distinguishability of sets of distributions.* Ann. Math. Stat. **29**, 700-718.

[Huber (1964)] Huber P.J. 1964: Robust estimation of a location parameter. *Ann. Math. Stat.*, **35**: 73–101.

[Huber (1968)] —— 1968: Robust confidence limits. *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, **10**: 269–278.

[Huber (1981)] —— 1981: *Robust statistics*. Wiley Series in Probability and Mathematical Statistics. Wiley.

[Huber (1997)] —— 1997: *Robust statistical procedures.*, Vol. 68 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2. edition.

[Huber-Carol (1970)] Huber-Carol, C. 1970: Étude asymptotique de tests robustes Ph.D. thesis, ETH Zürich.

[Huber-Carol (1986)] —— 1986: Théorie de la robustnesse. (Theory of robustness). In *Probability and statistics, Lect. Winter Sch., Santiago de Chile*, Nr. 1215 in Lect. Noes Math., pp. 1-128.

[Ibragimov (1967)] Ibragimov I. 1967: The Chebyshev-Cramér asymptotic expansions. *Theor. Probab. Appl.*, **12**: 454–469.

[Ibragimov and Linnik (1971)] Ibragimov I. and Linnik Y. 1971: *Independent and stationary sequences of random variables*. Wolters-Noordhoff Series of Monographs and Textbooks on Pure and Applied Mathematics. Wolters-Noordhoff Publishing Company, Groningen. edited by J.F.C. Kingman.

[Jänich (1999)] Jänich K. 1999: *Funktionentheorie: eine Einfürung.* Springer, Berlin, Heidelberg. 5. edition.

[Kohl (2005)] Kohl M. 2005: *Numerical contributions to the asymptotic theory of robustness*. Dissertation, Universität Bayreuth, Bayreuth.

[Kohl et al. (2004)] Kohl M., Ruckdeschel P. and Stabla T. 2004: General Purpose Convolution Algorithm for Distributions in S4-Classes by means of FFT. Available in `http://www.uni-bayreuth.de/departments/math/org/mathe7/RUCKDESCHEL/pubs/comp.pdf`.

[Krengel (1991)] Krengel U. 1991: *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. 3. erw. Auflage. Vieweg.

[Le Cam (1986)] Le Cam L. 1986: *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer.

[Leisch (2002a)] F. Leisch. 2002: *Sweave: Dynamic generation of statistical reports using literate data analysis.* In W. Härdle and B. Rönz, editors, *Compstat 2002 - Proceedings in Computational Statistics*, p. 575-580. Physica Verlag, Heidelberg, 2002. `http://www.ci.tuwien.ac. at/leisch/Sweave`.

[Leisch (2002b)] —— 2002: *Sweave, part I: Mixing R and LATEX.* R News, **2(3)**, p. 28-31. `http://CRAN.R-project.org/doc/Rnews/`.

[Leisch (2003)] —— 2003: *Sweave, part II: Package Vignettes.* R News, **3(2)**, p. 21-24. `http://CRAN.R-project.org/doc/Rnews/`.

[Mills (1926)] Mills J. P. 1926 *Table of the ratio: area to bounding ordinate, for any portion of normal curve/.* In: Biometrika, 18, p. 395-400.

[Parthasarathy (1967)] Parthasarathy K.R. 1967: *Probability measures on metric spaces.* Probability and Mathematical Statistics. A Series of Monographs and Textbooks. Academic Press.

[Pfanzagl (1979)] Pfanzagl J. 1979: First order efficiency implies second order efficiency. In: *Contributions to statistics, Jaroslav Hajek Mem. Vol.*, p. 167-196.

[Ramsey (1998)] N. Ramsey 1998: *Noweb man page.* University of Virginia, USA. `http://www.cs.virginia.edu/nr/noweb`. version 2.9a.

[Reeds (1976)] Reeds, J. A. 1976: *On the definition of von Mises functionals.* Research Report S 44. Departement of Statistics, Harvard University, Cambridge, Mass.

[Reeds (1985)] —— 1985: *Asymptotic number of roots of Cauchy distribution likelihood equations.* In: *Ann. Stat.*, **13**: 775–784.

[Rieder (1980)] Rieder H. 1980: Estimates derived from robust tests. *Ann. Stat.*, **8**: 106–115.

[Rieder (1989)] —— 1989: *A finite-sample miminax regression estimator. Statistics*, **20(2)**: 211-221.

[Rieder (1994)] —— 1994: *Robust asymptotic statistics*. Springer Series in Statistics. Springer.

[Rieder et al. (2001)] Rieder H., Kohl M. and Ruckdeschel P. 2001: *The Costs of not Knowing the Radius.* Appeared as discussion paper Nr. 81. SFB 373 (Quantification and Simulation of Economic Processes), Humboldt University, Berlin; also available in
`http://www.uni-bayreuth.de/departments/math/org/mathe7/RIEDER/pubs/RR.pdf`.

[Ruckdeschel (2004)] Ruckdeschel P. 2006: *A Motivation for $1/\sqrt{n}$–Shrinking–Neighborhoods.* Metrika **63**(3), pp. 295–307. Also available in
`http://www.uni-bayreuth.de/departments/math/org/mathe7/RUCKDESCHEL/pubs/whysqn.pdf`.

[Ruckdeschel (2005a)] —— 2005a: *Higher order asymptotics for the MSE of the median on shrinking neighborhoods.* Unpublished manuscript. Also available in
`http://www.uni-bayreuth.de/departments/math/org/mathe7/RUCKDESCHEL/pubs/medmse.pdf`.

[Ruckdeschel (2005b)] —— 2005b: *Higher order asymptotics for the MSE of M-estimators on shrinking neighborhoods.* Unpublished manuscript.

[Ruckdeschel (2005c)] —— 2005c: *Optimally one-sided bounded influence curves.* In: *Mathematical Methods of Statistics*. **14**: 105–131.

[Ruckdeschel (2005d)] —— 2005d: *Higher order asymptotics for the MSE of one-step estimators on shrinking neighborhoods.* Unpublished manuscript.

[Ruckdeschel and Rieder (2004)] Ruckdeschel P. and Rieder, H. 2004: *Optimal Influence Curves for General Loss Functions.* Statistics and Decisions. **22**: 201–223.

[Sachs and Hederich (2006)] Sachs L. and Hederich J. 2006: *Angewandte Statistik - Methodensammlung mit R.* Zwölfte, vollständig neu bearbeitete Auflage. Springer

[Serfling (1980)] Serfling, R. J. 1980: *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

[Schmitz (1996)] Schmitz N. 1996: *Vorlesungen über Wahrscheinlichkeitstheorie* Teubner-Studienbücher: Mathematik. Stuttgart.

[Siddiqui (1970)] Siddiqui M. M. 1970: *Order statistics of a sample and of an extended sample.* In: Puri M. L. (Ed.): *Nonparametric Techniques in Statistical Inference.* Cambridge Univ. Pr.

[Staudte and Sheather (1990)] Staudte R.G. and Sheather S.J. 1990: *Robust Estimation and Testing.* Wiley, New York.

[Stigler (1977)] Stigler, S.M. 1977: *An Attack on Gauss*, Published by Legendre in 1820. In *Historia Mathematica*, **4**, pp. 31-35.

[Strassen (1965)] Strassen V. 1965: *The existence of probability measures with given marginals. Ann. Math. Statist.*, **36**, pp. 423–439.

[Tukey (1960)] Tukey J. 1960: *A survey of sampling from contaminated distribution.* In: Contrib. Probab. Stat., Essays in Honor of H. Hotelling, 448-485 .

[von Mises (1937)] von Mises, R. 1937: *Sur les fonctions statistiques.* In: Conf. de la Réunion Internationale des Math., Gauthier-Villars, Paris, pp. 1-8.

[von Mises (1947)] —— 1947: *On the asymptotic distribution of differentiable statistical functions* In: *Ann. Math. Statist.* **18**: 309-348.

[van der Vaart (1998)] van der Vaart A. W. 1998: *Asymptotic statistics.*, Vol. 3 of *Cambridge Series on Statistical and Probabilistic Mathematics*. Cambridge Univ. Press., Cambridge.

[Venables and Ripley (1999)] Venables W.N. and Ripley B.D. 1999: *Modern Applied Statistics with S-PLUS.*, Third Edition, Springer Series in Statistics and Computing. Springer.

[Wikipedia (2008)] Deimos 28, Tolstoy the (little black) Cat et. al. 2005–2008: *Robust statistics.*
`http://en.wikipedia.org/wiki/Robust_statistics`.

Author Index

Subject Index

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit

Higher Order Asymptotics for the MSE of Robust M-Estimators of Location on Shrinking Total Variation Neighborhoods

selbständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Augsburg, den 29. Mai 2008 Matthias Brandl.