Dortmund, im Jahr 2013

Robuste Verfahren zur Periodendetektion in ungleichmässig beobachteten Lichtkurven

Dissertation
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften

Der Fakultät Statistik der Technischen Universität Dortmund vorgelegt von

Anita Monika Thieler

aus Husby bei Flensburg

Gutachter:	Prof. Dr. Roland Fried
	Prof. Dr. Christine Müller
Tag der mündlichen Prüfung:	30. Januar 2014

Ich möchte mich sehr herzlich bei meinem Betreuer Prof. Dr. Fried bedanken, der meiner Forschung viel Zeit und Aufmerksamkeit geschenkt hat und damit maßgeblich zu ihrem Gelingen beigetragen hat. Auch bin ich meiner Familie und meinem Freund dankbar, die in dieser teils sehr anstrengenden Zeit so geduldig mit mir waren und mich stets unterstützt und ermutigt haben.

Ich möchte mich auch bei Prof. Dr. Wolfgang Rhode und seiner Arbeitsgruppe für die anregende Zusammenarbeit bedanken. Ebenso gilt mein Dank meinen Lektoren Dr. Swaantje Casjens und Jonathan Rathjens, durch deren wertvolle Hinweise und konstruktive Kritik ich meine Arbeit entscheidend in ihrer Lesbarkeit verbessern konnte.

Finanziell wurde meine Arbeit durch das Graduiertenkolleg „Statistische Modellbildung" der Fakultät Statistik und den Sonderforschungsbereich SFB 876 „Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung" unterstützt, wofür ich mich bedanke. Ebenfalls dankbar bin ich dem IT & Medien Centrum der TU Dortmund für die Bereitsstellung von Computerressourcen in Form des Großrechners LiDo und für die allzeit schnelle und freundliche Unterstützung durch das Support-Team.

Mein Dank geht auch an alle anderen Diskussionspartner, Korrekturleser, Mitforscher, R-Experten, Rechercheure, Reparierer, Einhornbäcker, Forderer, Förderer, Geduldhaber, Kaffeeeinschenker, Techniker, Tippgeber, Verteidiger, Wissensteiler, Mitfreuer, Motivatoren, Kritiker, Macher, Panikreduzenten, Teamathmosphärenschaffer, und Systemadministratoren, die alle auf ihre Weise zum Entstehen dieser Arbeit beigetragen haben.

Inhaltsverzeichnis

1. Einleitung

Eine wichtige Aufgabe sowohl in der Astroteilchenphysik als auch in der Astrophysik ist die Suche nach Periodizität in ungleichmäßig beobachteten Zeitreihen, Lichtkurven genannt. Die Messwerte einer Lichtkurve entsprechen den gemessenen Photonenemissionen eines Objekts im Weltraum. Fluktuieren diese Emissionen periodisch und ist die Fluktuationsperiode bekannt, liefert dies wertvolle physikalische Erkenntnisse über das beobachtete Objekt. Diese Erkenntnisse beziehen sich beispielsweise auf die Umlaufzeit eines binären sternenähnlichen Objekts (vgl. Albert et al. 2006 sowie Albert et al. 2009), auf die Emissionsmechanismen in speziellen Aktiven Galaxien (vgl. Rieger und Mannheim 2000), oder auf die Form eines beobachteten Asteroiden (vgl. Warner 2006, Kapitel 11). Aufgrund unkontrollierbarer Faktoren wie etwa geeigneten Wetterbedingungen, die zur Durchführung einer Messung erfüllt sein müssen, liegen die Messzeiten nicht auf einem gleichabständigen Gitter, sondern sind ungleichmäßig verteilt. Bedingt durch die Rotationen der Himmelskörper können sie sogar eine periodische Verteilung aufweisen. Diese Periodizität der Messzeiten ist dabei unabhängig von der gesuchten Periodizität der Messwerte. Weiterhin steht bei Lichtkurvendaten üblicherweise für jeden Messwert ein Messfehler zur Verfügung. Dieser beschreibt, wie genau die Messung durchgeführt werden konnte.
Zur Ermittlung der Fluktuationsperiode ist das Fourier-Periodogramm als standardmäßiges Analysewerkzeug für gleichmäßig beobachtete Zeitreihen im Falle ungleichmäßig beobachteter Lichtkurven ungeeignet. Periodogramme für Lichtkurven werden daher häufig berechnet, indem periodische Funktionen verschiedener Periodenlängen (Testperioden) an die Lichtkurve angepasst werden. Am häufigsten wird eine Sinusfunktion mittels Kleinste-Quadrate-Regression angepasst. In solchen Periodogrammen können die Messfehler durch den Einsatz gewichteter Regression berücksichtigt werden.
In dieser Arbeit werden verschiedene Periodogrammmethoden verglichen, die alle auf der Anpassung einer periodischen Funktion basieren. Einige dieser Methoden werden dabei erstmals in dieser Arbeit und in einhergehenden Publikationen (vgl. Thieler 2011, Thieler 2012, Thieler et al. 2013 sowie Thieler, Fried und Rathjens 2013) vorgeschlagen. Im Vergleich werden sowohl auf gewichteter als auch auf ungewichteter Regression basierende Methoden betrachtet. Zur Anpassung werden sechs verschiedene periodische Funktionen verwendet. Sie decken die in der Astroteilchen - und Astrophysik beliebtesten und einige andere in der Literatur vorgeschlagene Funktionen ab. Da es in Lichtkurvendaten häufig zu Störungen kommt, etwa zu Zeitintervallen stark erhöhter Messwerte („Intervallstörungen"), werden neben der Kleinste-Quadrate-Regression verschiedene robuste Regressionstechniken untersucht. Einige dieser Techniken werden dabei erstmals im Rahmen dieser Arbeit für den Einsatz in der Periodogrammberechnung vorgeschlagen.

Eine Lichtkurve muss nicht notwendigerweise eine periodische Fluktuation enthalten. Es ist deshalb notwendig, sich vor Falschdetektionen zu schützen, die entstehen können, wenn in einem Periodogramm die zum höchsten Balken gehörige Periode stets als detektierte Fluktuationsperiode gilt. Zur Auswertung eines Periodogramms wird daher ein Detektionsmechanismus zur Erkennung auffällig hoher Periodogrammbalken benötigt, die auf eine Periodizität hindeuten. Die übliche Verteilungsannahme für die Periodogrammbalken ist dabei unter Nullhypothese eine Betaverteilung mit festen Parametern. Sie resultiert aus der Annahme unabhängig normalverteilter Messwerte und der Nutzung von Kleinste-Quadrate-Regression. Liegt ein Periodogrammbalken über einem zuvor spezifizierten Quantil der Verteilung, gilt er bzw. die zugehörige Periode als detektiert. Da in dieser Arbeit auch andere Regressionstechniken genutzt werden und sich die Annahme von ausschließlich normalverteiltem Rauschen als zu restriktiv herausstellt, wird hier statt der festen eine angepasste Betaverteilung genutzt. Damit ein hoher Periodogrammbalken detektiert wird und die Verteilungsschätzung nicht zu stark beeinflusst, erfolgt diese Anpassung robust mittels Cramér-von-Mises-Distanz-Minimierung. Zur Periodendetektion ist ein solches auf der Idee der Ausreißeridentifikation (vgl. Davies und Gather 1993) basierendes Vorgehen neu. Zwecks Reduktion der Abhängigkeiten unter den Periodogrammbalken werden zur Detektion auffällig hoher Periodogrammbalken drei Auswertungstypen verglichen: Bei zweien wird das Periodogramm vor Anpassung der Verteilung, bei dem dritten Auswertungstyp wird die Verteilung an das vollständige Periodogramm angepasst.

Die Suche nach einer Fluktuationsperiode in Zeitreihen mit ungleichmäßigem Messzeitmuster ist auch Forschungsgegenstand anderer Fachgebiete. Methoden zur Periodendetektion sind damit auch abseits der Lichtkurvenanalyse von Interesse, etwa in medizinischen Gebieten wie der Genetik (vgl. Ahdesmäki et al. 2007), der Chronobiologie (vgl. Ruf 1999) und der Hämatologie (vgl. Fortin und Mackey 1999) oder in geologischen Gebieten wie der Seismologie (vgl. Baisch und Bokelmann 1999) und der Paläoklimatologie (vgl. Petit et al. 1999).

Es ist keine Arbeit bekannt, in der systematisch Periodogramm- oder Detektionsmethoden verglichen werden, bei der periodische Funktionen mittels robuster Regression angepasst werden. In vielen Arbeiten werden mehrere hauptsächlich Kleinste-Quadrate-basierte Methoden zur Analyse echter Daten verwendet, ohne dass das wahre Modell und damit das wünschenswerte Analyseergebnis bekannt ist. Systematischere Vergleiche verschiedener Periodogrammmethoden, die auf der Anpassung einer periodischen Funktion basieren, werden von Reimann (1994) sowie Graham et al. (2013) unternommen. Dort werden jedoch keine auf robuster Regression basierende Methode betrachtet. In einer Studie von Oh et al. (2004) wird neben auf Kleinste-Quadrate-Regression basierenden Periodogrammmethoden auch eine auf robuster Filterung basierende Methode untersucht. Der Vergleich schließt jedoch keine Periodogrammmethode ein, bei der eine vorgegebene parametrische Funktion mittels robuster Regression angepasst wird. Bei den genannten Arbeiten wird die Präsenz einer periodischen Fluktuation vorausgesetzt, womit die Gefahr einer Falschdetektion in einer unperiodischen Lichtkurve nicht gegeben und die Untersuchung eines Detektionskriteriums nicht notwendig ist.

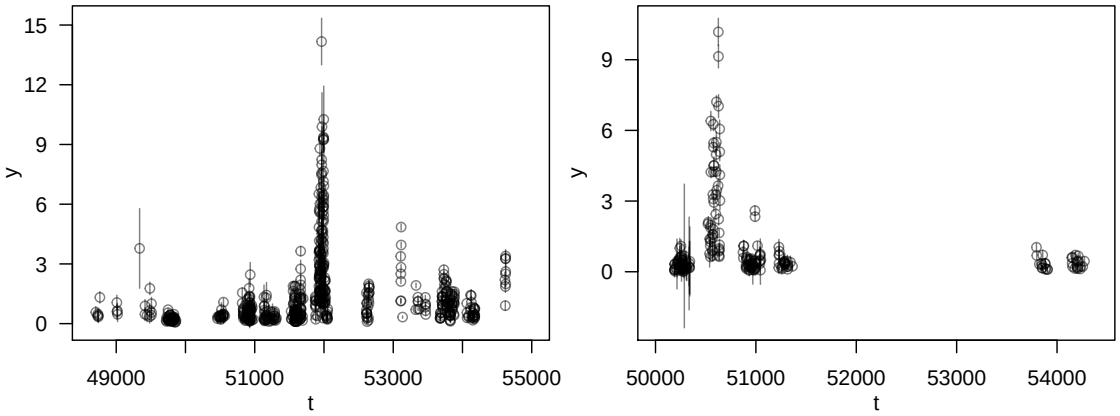
In dieser Arbeit erfolgt der Vergleich neuer und bekannter Periodogrammmethoden bzw. darauf aufbauender Detektionsmethoden in einer Simulationsstudie und in der Anwendung auf reale Daten. Dabei ist diese Arbeit wie folgt aufgebaut: In Kapitel 2 werden die Modellannahmen für die Lichtkurvendaten und das Prinzip der Periodogramme und der Detektion beschrieben. Dazu werden die spezifischen Eigenschaften von Lichtkurvendaten erklärt. Anschließend werden die zur Periodogrammberechnung verwendeten periodischen Funktionen und Regressionstechniken behandelt und bezüglich ihres bisherigen Einsatzes zur Periodogrammberechnung eingeordnet. Es folgt eine Motivierung und Erörterung des hier entwickelten auf Anpassung einer Betaverteilung basierenden Detektionskriteriums. Bestehende Probleme bei der Detektion durch Abhängigkeiten der Periodogrammbalken untereinander werden diskutiert und führen zur Entwicklung der bereits oben erwähnten Auswertungstypen mit Periodogrammausdünnung. Das Kapitel schließt mit einer Abgrenzung der hier betrachteten von anderen Periodogrammmethoden, welche nicht dem auf dem Prinzip der Anpassung periodischer Funktionen basieren und damit nicht Gegenstand dieser Arbeit sind. Kapitel 3 liefert Details zur Implementierung des R-Pakets `RobPer` (vgl. Thieler, Fried und Rathjens 2013), das zur Anwendung der Detektionsmethoden auf simulierte und echte Daten genutzt wird. `RobPer` enthält neben einer Funktion zur Periodogrammberechnung auch eine Funktion zur Anpassung einer Betaverteilung mittels Cramér-von-Mises-Distanz an ein Periodogramm und Funktionen zur Generierung der künstlichen Lichtkurven. Kapitel 4 beschreibt die Anwendung auf simulierte Daten. In einer Simulationsstudie werden hierzu die verschiedenen Detektionsmethoden, bestehend aus einer Periodogrammmethode und einem Auswertungstyp, verglichen. Dabei werden 252 Detektionsmethoden jeweils auf 20 verschiedene Lichtkurventypen angewendet. In Kapitel 5 erfolgt die Anwendung der besten Detektionsmethoden der Simulationsstudie und einiger Vergleichsmethoden auf reale Daten. Kapitel 6 liefert Ansätze für mögliche Modifikationen der verwendeten Periodogramme und der zu Grunde liegenden Modellannahmen. Der mögliche Einsatz anderer Regressionstechniken und die Anpassung anderer periodischer Funktionen wird in Hinblick auf Schwierigkeiten diskutiert, die bei der Periodendetektion in Kapitel 4 und 5 beobachtet werden. Die Präsenz einer speziellen Rauschkomponente in den Messwerten kann die Periodendetektion erschweren und macht einen weiteren Verarbeitungsschritt notwendig. Zum Umgang mit diesem sogenannten roten Rauschen werden hier erste Vorschläge für ein Rauschfilter gemacht, welches erfolgreich auf einige simulierte und ein reales Datenbeispiel angewendet werden kann. Als weitere Lockerung der Modellannahmen wird die zeitliche Veränderung der periodischen Fluktuation oder ihrer Fluktuationsperiode diskutiert. Die Veränderung der periodischen Fluktuation bei bekannter Fluktuationsperiode wird mit einem Problem aus dem Maschinenbau illustriert. Im Rahmen dieser Arbeit sind erste Ansätze zu dessen Lösung entstanden und publiziert worden (vgl. Fried, Raabe und Thieler 2012 sowie Raabe et al. 2012). Kapitel 7 schließt die Arbeit mit einer Zusammenfassung der erzielten Forschungsergebnisse und einem Ausblick.

2. Periodendetektion

In diesem Kapitel wird der zu analysierende Datentyp und das vorgeschlagene Analyseverfahren vorgestellt. Bei dem Datentyp handelt es sich um Lichtkurven. Lichtkurven sind Zeitreihen aus der Astroteilchenphysik mit speziellen Eigenschaften. Diese werden in Abschnitt 2.1 erläutert. Zur Detektion von Periodizität in Lichtkurven werden so genannte Periodogramme verwendet, deren Grundprinzip in Abschnitt 2.2 vorgestellt wird. Insbesondere wird auf Verfahren eingegangen, die auf Anpassung einer periodischen Funktion mittels linearer Regression basieren. Die in dieser Arbeit betrachteten Periodogrammmethoden arbeiten nach diesem Schema. Die Auffassung der Periodogrammmethoden als Regressionsergebnis ermöglicht die Modifizierung bereits bestehender Methoden durch Verwendung anderer Modelle (Abschnitt 2.3) und anderer, zum Beispiel robuster, Regressionstechniken (Abschnitt 2.4) sowie die Anwendung gewichteter Regression (Abschnitt 2.5). Hierdurch entstehen teils neue Methoden, die im Rahmen dieser Arbeit und daraus hervorgehenden Publikationen (Thieler 2011, Thieler 2012, Thieler et al. 2013 und Thieler, Fried und Rathjens 2013) erstmalig vorgestellt werden.
Es ergibt sich die Frage, wie eine Periode mit Hilfe eines Periodogramms entdeckt werden kann, ohne zufällige Effekte fälschlicherweise als Periodizität zu deuten. In Abschnitt 2.6 wird dazu ein bestehendes Verfahren vorgestellt, das in der Anwendung oft zu viele Perioden detektiert. Ferner wird ein neues Konzept vorgeschlagen, das auf Ausreißeridentifikation basiert und in Simulationen bessere Ergebnisse erzielt. Abschnitt 2.7 gibt eine Übersicht über andere Periodogrammmethoden, die nicht auf dem in dieser Arbeit verfolgten Ansatz (Anpassung periodischer Modelle mittels linearer Regression) basieren.

2.1. Lichtkurven

In diesem Abschnitt wird die angenommene Struktur der zu untersuchenden Daten vorgestellt. Bei den in dieser Arbeit zu analysierenden Daten handelt es sich um Lichtkurven, eine Datenstruktur, die typischerweise in der Astroteilchenphysik vorgefunden wird. Lichtkurven sind spezielle Zeitreihen $\mathfrak{Y} = (t_i, y_i, s_i)_{i=1,\ldots,n}$, bestehend aus Messzeiten, Messwerten und Messfehlern. Jeder Messwert y_i zu einer Messzeit t_i, $i = 1, \ldots, n$, enthält dabei die Emissionsmessung von Photonen einer bestimmten Energieklasse. Die Quelle dieser Photonen liegt im Weltraum und kann beispielsweise ein so genannter Blazar sein (vgl. Grupen 2000, S. 166–167). Zum Beispiel sind die in Abbildung 2.1 beobachteten Lichtkurven der Blazare Makarian 421 (Mrk 421) und Makarian 501 (Mrk 501) zu sehen (Daten aus Tluczykont et al. 2010 und dort enthaltenen Referenzen). Diese Lichtkurven enthalten die Emissionsmessungen hochenergetischer Photonen mit einer Energie von mindestens einem Terraelektronenvolt (Licht mit einer Wellenlänge von weniger als dem 10^{-13}-fachen der



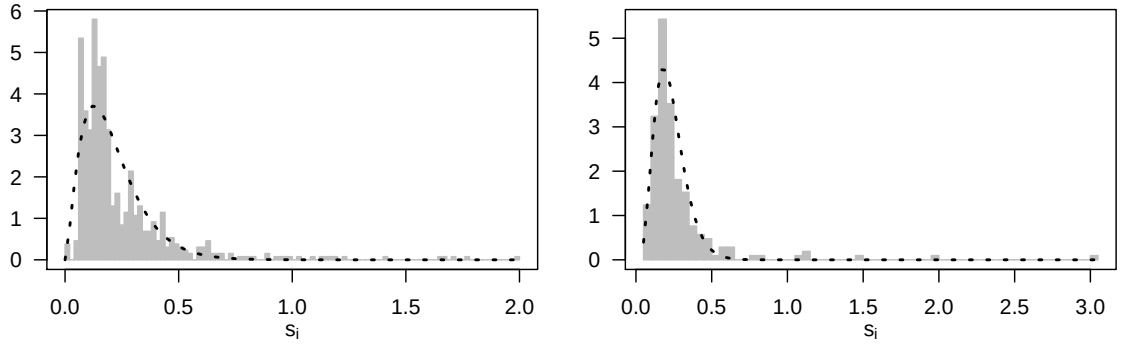
(a) (b)

Abbildung 2.1.: Lichtkurven aus Tluczykont et al. (2010). Quellen: (a) Makarian 421, (b) Makarian 501. Ordinate: Messwerte y_i der γ-Emissionen in Crab Units (relativ zu den Emissionen des Krebsnebels). Abszisse: Messzeiten t_i angegeben in Tagen seit dem 17. November 1858. Die Messfehler s_i sind als vertikale Balken nach oben und unten an jede Beobachtung eingezeichnet.

Wellenlänge sichtbaren Lichts, vgl. Grupen 2000, S. 86). Diese Photonen gehören in die Klasse der Gammateilchen (Photonen mit mehr als 100 Kiloelektronenvolt) und werden mit spezialisierten Teleskopen wie zum Beispiel dem MAGIC-Teleskop (Albert et al. 2008, Aleksić et al. 2012) gemessen.

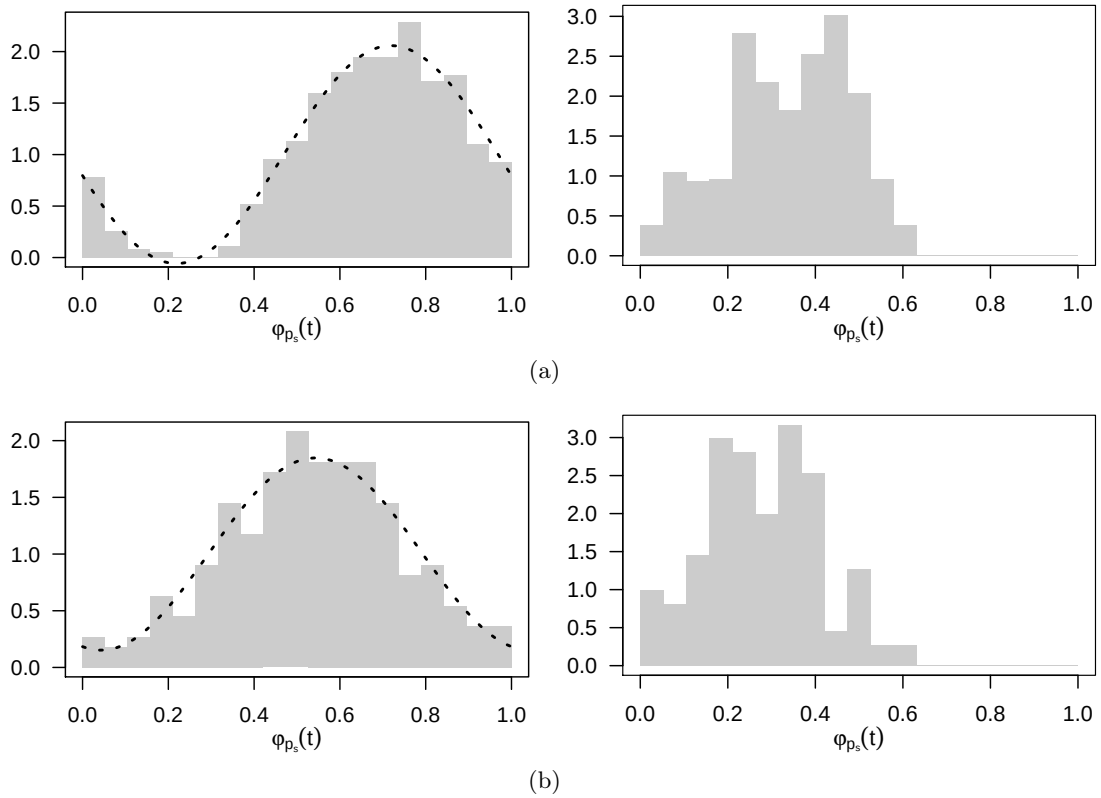
2.1.1. Ein Datenmodell

Die Messung hochenergetischer Gammaemissionen ist nur indirekt und mit Hilfe von Schätzungen möglich, so dass zu jedem realisierten Messwert y_i auch ein so genannter Messfehler s_i vorliegt, der die Genauigkeit der Messung beschreibt und als Standardabweichung interpretiert werden kann. Große Messfehler stehen also für eine ungenaue Beobachtung. In dieser Arbeit wird angenommen, dass y_i einen von den anderen Messwerten unabhängig normalverteilten additiven Fehlerterm $y_{i;w}$ beinhaltet (w wie „weißes Rauschen“), dessen Varianz durch s_i geschätzt wird. Abbildung 2.2 zeigt die Verteilung der s_i für die in Abbildung 2.1 gezeigten Lichtkurven. Für beide Lichtkurven ist sie rechtsschief.
Das Tripel (t_i, y_i, s_i) wird im Folgenden Beobachtung genannt, die Menge aller n Beobachtungen $(t_i, y_i, s_i)_{i=1,\ldots,n}$ einer Quelle ihre Lichtkurve. Hierbei ist zu beachten, dass die t_i im Allgemeinen nicht äquidistant liegen. Abgesehen von dem erzwungenen Gitter, das sich durch die begrenzte Anzahl Nachkommastellen bei digitaler Datenverarbeitung ergibt, sind sie im Allgemeinen auch nicht bis auf fehlende Messwerte äquidistant. Es gibt viele Gründe dafür, dass die Messzeiten nicht in gleichmäßigen Abständen liegen. Zum Einen sind dafür zufällige Faktoren verantwortlich: Da Messungen nur bei klarem Himmel möglich sind, müssen sie bei schlechten Wetterverhältnissen verschoben werden. Vor allem Cherenkov-Teleskope wie das MAGIC-Teleskop sind hier häufig betroffen (vgl. Abdo et al. 2011). Die Messung einer Quelle kann nicht erfolgen, während der Beobachtung einer anderen Quelle Vorrang gegeben wird. Und schließlich können technische Probleme im Teleskopbetrieb dazu



(a) (b)

Abbildung 2.2.: Histogramme der Messfehler s_i für die Lichtkurven aus Abbildung 2.1: (a) zu Makarian 421, (b) zu Makarian 501. Die eingezeichneten Gammadichten wurden durch Cramér-von-Mises-Distanz-Minimierung (vgl. Abschnitt 2.6) angepasst.



(a)

(b)

Abbildung 2.3.: Histogramme der umgeklappten Messzeiten. Lichtkurven (vgl. Abbildung 2.1) (a) zu Makarian 421, (b) zu Makarian 501. Messzeiten umgeklappt nach Mondzyklus (links, Samplingperiode p_s zwischen 27 und 28 Tagen) und Jahreszyklus (rechts, p_s = 365 Tage). Die Linie zeigt eine mittels Kleinste-Quadrate-Regression an die Histogrammbalken angepasste Sinusschwingung.

führen, dass Messungen ausfallen, verspätet stattfinden oder Beobachtungen nachträglich entfernt werden müssen.
Zusätzlich zu diesen Schwierigkeiten gibt es noch periodisch auftretende Hindernisse bei der Datennahme. So funktioniert die Messung nur, wenn die Quelle von der Erde aus, speziell dem Standort des Teleskops, sichtbar ist. Weiterhin sind Beobachtungen nur bei dunklem Himmel, also nachts und nicht bei Vollmond, möglich. Diese Aspekte beeinflussen die Datennahme periodisch in wiederkehrender Weise und hängen von der Konstellation der Himmelskörper ab. In Abbildung 2.3 wird die Verteilung der Messzeiten für die oben genannten Quellen im Mond- und Jahreszyklus gezeigt. Im Jahreszyklus wird deutlich sichtbar, dass die Quellen nicht während des ganzen Jahres beobachtbar sind. Im Mondzyklus ähnelt die Verteilung der Messzeiten einer Sinusfunktion. Die Vollmondphase, in der nicht beobachtet werden kann, ist deutlich erkennbar. Eine sinusförmige Verteilung der Messzeiten über den Mondzyklus ist wegen der Rotationsbewegung von Erde und Mond plausibel (vgl. Höfler 1913, S. 255 ff). Es wird also angenommen, dass die t_i Realisationen von mit periodischer Dichte verteilten Zufallsvariablen T_i sind. Die Periode der Messzeitverteilung sei im Folgenden mit p_s bezeichnet und wird Samplingperiode genannt.
Es wird noch eine weitere Periodizität angenommen, die von der Verteilung der Messzeiten unabhängig ist: Es wird erwartet, dass die Messwerte y_i eine periodische Komponente $y_{f;i} = f\left(\frac{t_i}{p_f}\right)$ beinhalten, die im Folgenden Signal oder periodische Fluktuation genannt wird. Dabei ist f eine stetige Funktion mit Periode 1, p_f wird Fluktuationsperiode genannt und ist die Periode des Signals. Die Funktion f kann konstant sein, da es auch möglich ist, dass die Messwerte keine periodische Komponente enthalten.
Mir diesen Annahmen und Überlegungen wird für die Zufallsvariablen $(T_i, Y_i, S_i)_{i=1,\dots,n}$, von denen die Lichtkurve $(t_i, y_i, s_i)_{i=1,\dots,n}$ eine Realisation ist, folgendes Modell aufgestellt:

$$T_i = T^\star_{(i)}, \qquad T^\star_1, \dots, T^\star_n \sim \mathcal{D}(p_s), \tag{2.1}$$

$$Y_i = Y_{f;i} + Y_{w;i}, \tag{2.2}$$

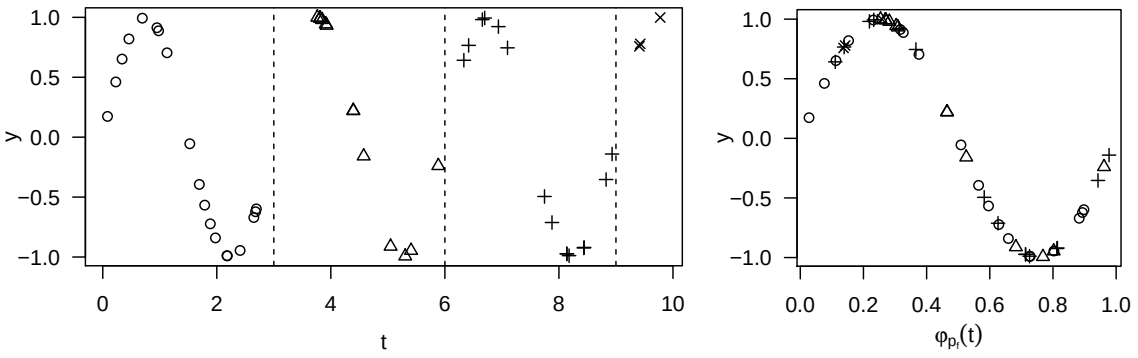
$$Y_{f;i} = f\left(\frac{T_i}{p_f}\right), \qquad f(\xi) = f(\xi + 1) \forall \xi \in \mathbb{R} \tag{2.3}$$

$$Y_{w;i} \sim \mathcal{N}(0, \sigma_i^2), \tag{2.4}$$

s_i gegebene Schätzung für σ_i unabhängig von $Y_1, \dots, Y_n$.

Hierbei ist $T^\star_{(i)}$ die i-te geordnete Zufallsvariable von $T^\star_1, \dots, T^\star_n$ und $\mathcal{D}(p_s)$ eine periodische Messzeitverteilung mit Periode p_s.
Dieses Modell ist bei Bedarf erweiterbar, zum Beispiel durch Ausreißer. Hierbei sei darauf hingewiesen, dass es in diesem Modell nicht zu großen Ausreißern in den Messzeiten, so genannten leverage points (vgl. Maronna, Martin und Yohai 2006, S. 115–124), kommen kann. Die Verteilungsannahme der Messzeiten würde durch deren Störung verletzt. Da für den realisierten Fluktuationswert (vgl. Formel (2.3))

$$y_{f;i} = f\left(\frac{t_i}{p_f}\right) = f\left(\frac{t_i + c \cdot p_f}{p_f}\right), c \in \mathbb{N}$$



(a) (b)

Abbildung 2.4.: Beispiel für das Phasendiagramm einer Zeitreihe. (a) Messwert y gegen die Messzeit t abgetragen, vertikale gestrichelte Linien deuten die Zyklen der Länge $p_f = 3$ an, unterschiedliche Symbole ($\circ$, $\triangle$, $+$, $\times$) stehen für unterschiedliche Zykluszugehörigkeiten. (b) Messwerte mit gleichem Symbol wie zuvor abgetragen gegen die Phase $\varphi_{p_f}(t)$. Anhand der vermischt auftretenden Symbole ist die Umordnung der Daten erkennbar.

gilt, kann eine Messzeit jedoch nur um maximal $\frac{p_f}{2}$ von einem zu $y_{f;i}$ passenden Wert abweichen. Abgesehen davon ist durch die Gestalt des Datennahmeprozesses nicht mit Ausreißern bei den Messzeiten zu rechnen. Anstatt f als periodische Funktion der mit der Fluktuationsperiode p_f normierten Messzeiten zu betrachten, kann sie auch auf dem Intervall $[0, 1[$ definiert werden und als Funktion der Phase $\varphi_{p_f}(t_i)$ mit

$$\varphi_p(t) = \frac{t - \left\lfloor \frac{t}{p} \right\rfloor p}{p} \in [0, 1] \tag{2.5}$$

aufgefasst werden (vgl. z.B. Dupuy und Hoffman 1985, Jurkevich 1971). Durch Anwendung von φ_{p_f} auf t_i geht die Information verloren, im wievielten Zyklus $z_p(t_i)$ mit

$$z_p(t) = \frac{t - \varphi_p(t)p}{p} + 1 = \left\lfloor \frac{t}{p} \right\rfloor + 1$$

des Signals die i-te Beobachtung gemessen wurde. Abbildung 2.4 zeigt zur Veranschaulichung eine realisierte Zeitreihe mit ungefährer Dauer $T = t_n - t_1 = 10$ und periodischer Fluktuation der Periode $p_f = 3$ (eine Sinusfunktion), einmal abgetragen gegen die Messzeit t, einmal gegen die Phase $\varphi_{p_f}(t)$. Letztere Darstellung wird auch Phasendiagramm (phase diagram, vgl. Dupuy und Hoffman 1985) genannt.
Wenn mit $\varphi_p(t_i)$ statt t_i gearbeitet wird, wird häufig davon gesprochen, dass die Lichtkurve nach Periode p „gefaltet" (folded) oder „umgeklappt" (collapsed, vgl. Dupuy und Hoffman 1985) wird. Man kann sich vorstellen, dass die Zeitreihe entlang der Abszisse in Abschnitte der Länge p_f unterteilt wird, und diese übereinandergelegt nach passender Normierung der horizontalen Achse das Phasendiagramm zeigen. Der plastisch passendere Ausdruck statt „falten" wäre eigentlich „aufrollen", vergleiche Abbildung 2.5, so dass der zyklische Grundgedanke nicht vergessen werden kann, durch den $\varphi = 0{,}9$ genauso weit von $\varphi = 0{,}05$ entfernt ist wie von $\varphi = 0{,}75$. Da dieser Begriff aber nicht üblich ist, sei er aber nur für

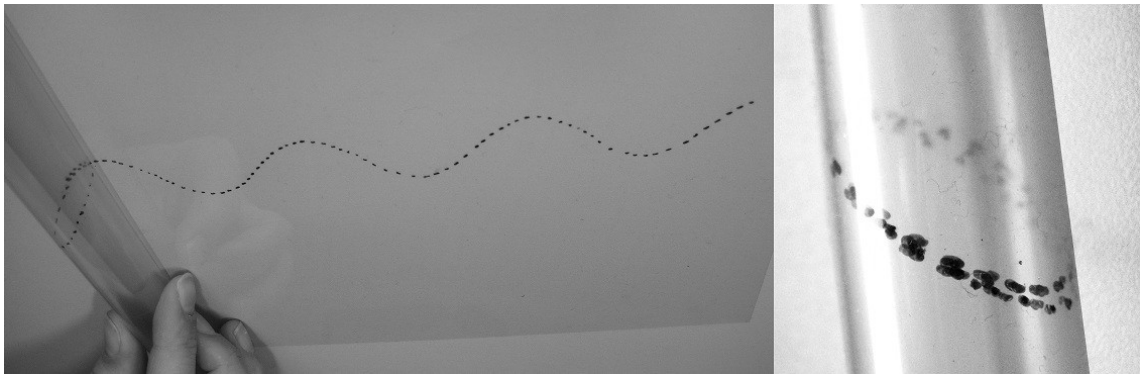


Abbildung 2.5.: Illustration, für das „Falten" beziehungsweise „Umklappen" einer Zeitreihe nach p_f: Die auf Folie notierte Zeitreihe wird buchstäblich aufgerollt, der entstehende Zylinder hat den Umfang p_f. Der Begriff „Aufrollen" würde den Vorgang eigentlich präziser umschreiben als „Falten" oder „Umklappen", ist aber nicht üblich.

intuitiveres Verständnis des Vorganges erwähnt, im Folgenden wird der Begriff "umklappen" verwendet. Wo es nicht zu Missverständnissen führen kann, wird $\varphi_p(t_i)$ mit φ_i abgekürzt. Es ist zu beachten, dass die Phase in einigen Publikationen anders normiert ist, häufig auf das Intervall $[0, p_f[$ (z.B. bei Stellingwerf 1978). Alle hier gemachten Aussagen gelten bei entsprechender Skalierung auch bei solchen Phasendefinitonen.

2.1.2. Weitere Notationen

Zur Erläuterung der auf linearer Regression basierenden Periodogramme werden in diesem Abschnitt weitere Notationen eingeführt. Sei $\mathfrak{Y} = (t_i, y_i, s_i)_{i=1,\dots,n}$ eine Lichtkurve und $g : \mathbb{R} \to \mathbb{R}$ eine periodische Funktion, die als lineares Modell formuliert werden kann. Weiter sei

$$\mathcal{X} : \mathbb{R} \times \mathbb{R}_{>0} \to \mathbb{R}^{1\times m} \tag{2.6}$$

die Funktion zur Bildung des Regressors, wobei $\mathcal{X}(t,p)$ zu gegebener Messzeit t und Testperiode p den Regressor angibt. Weiter sei

$$Y_i = X_i(p)\beta + Y_{w;i} \qquad \text{mit } Y_{w;i} \underset{\text{u.i.v.}}{\sim} \mathcal{N}(0,\sigma^2), \tag{2.7}$$

wobei $\sigma > 0$ unbekannt und $X_i(p) = \mathcal{X}(t_i, p)$, ein einfaches lineares Modell zur Anpassung von $g\left(\frac{t}{p}\right)$. Die Anpassung dieses Modelles wird im Folgenden ungewichtete Regression genannt. Die alternative Darstellung in Matrixschreibweise lautet

$$\begin{aligned} Y =& X(p)\beta(p) + Y_w \qquad \text{mit } Y_w \sim \mathcal{N}(0_n, \sigma^2\mathbb{I}_n) \\ \text{mit } & Y = (Y_1, Y_2, \dots, Y_n)^\mathsf{T}, \\ & X(p) = \left(X_1(p)^\mathsf{T}, X_2(p)^\mathsf{T}, \dots, X_n(p)^\mathsf{T}\right)^\mathsf{T}, \\ & \beta(p) \in \mathbb{R}^m, \\ & 0_n = (0,\dots,0)^\mathsf{T} \in \mathbb{R}^n, \end{aligned} \tag{2.8}$$

und $\mathbb{I}_n$ als $n \times n$-Einheitsmatrix.

Sei $\widehat{y}_i(p)$, $i = 1, \ldots, n$, die Anpassung an den Messwert y_i, dann ist $r_i(p) = y_i - \widehat{y}_i(p)$ das zugehörige Residuum und $r(p) = (r_1(p), \ldots, r_n(p))^\mathsf{T}$ der Residualvektor. Die Funktion $\zeta : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ sei die je nach Regressionstechnik zu minimierende Zielfunktion für den Residualvektor. Beispielsweise ist für die Kleinste-Quadrate (KQ)-Regression

$$\zeta_{KQ}(r) = \sum_{i=1}^{n} r_i^2 \tag{2.9}$$

und für die Kleinste-Beträge (L1)-Regression (vgl. Abschnitt 2.4) gilt

$$\zeta_{L1}(r) = \sum_{i=1}^{n} |r_i|. \tag{2.10}$$

Im Spezialfall der Anpassung eines Modells nur mit konstantem Term (also einer Lokationsschätzung) werden abweichend von den in den Gleichungen (2.7) und (2.8) eingeführten Bezeichnungen die folgenden verwendet: μ statt β und $\mathtt{i}$ statt X. Im Falle ungewichteter Regression gilt $\mathtt{i} = 1_n$. SY sei als der Wert der Zielfunktion bei Anpassung dieses Lokationsmodells definiert:

$$\mathrm{SY} = \zeta(y - \widehat{\mu}\mathtt{i}) \tag{2.11}$$

SE sei der Minimierungswert bei Anpassung des vollen Modells:

$$\mathrm{SE}(p) = \zeta(y - X(p)\widehat{\beta}(p)) \tag{2.12}$$

Im Falle ungewichteter Kleinste-Quadrate-Regression ist $\frac{1}{n-1}\,\mathrm{SY}$ die empirische Varianz der Beobachtungen, $\frac{1}{n-1}\,\mathrm{SE}(p)$ die der Residuen und $R^2(p) = 1 - \mathrm{SE}(p)/\mathrm{SY}$ das Bestimmtheitsmaß der Anpassung.

2.2. Periodogramme

Sei $\mathbb{L}$ die Menge möglicher beobachtbarer Lichtkurven und $\mathbb{R}^{>0}$ die Menge der streng positiven reellen Zahlen. Eine Funktion $\mathrm{Per} : \mathbb{L} \times \mathbb{R}_{>0} \to \mathbb{R}$, definiert für eine Testperiode $p \in \mathbb{R}_{>0}$ und eine Lichtkurve $\mathfrak{Y} = (t_i, y_i, s_i)_{i=1,\ldots,n}$ einen so genannten Periodogrammbalken. Eine Menge von Periodogrammbalken zu einer festen Lichtkurve wird ihr Periodogramm genannt. Sofern es nicht zu Doppeldeutigkeiten führt, wird auch die grafische Darstellung, in der $\mathrm{Per}(p)$ gegen p abgetragen wird, Periodogramm genannt.
Periodogrammbalken sind üblicherweise so konstruiert, dass ein Periodogrammbalken, dessen Periode nahe der Fluktuationsperiode liegt, einen relativ zu den anderen Balken extremen (je nach Methode besonders hohen oder niedrigen) Wert annehmen soll. Es stellt sich daher die Frage, ob wirklich das Periodogramm von Interesse ist, oder ob es sinnvoller wäre, zur Detektion einer Periodizität in den Messwerten die Funktion Per über p zu optimieren. Der

Nachteil des zweitgenannten Vorgehens ist, dass hierbei auch dann eine Periode entdeckt wird, wenn in den Messwerten gar keine periodische Fluktuation vorliegt. Bei der Auswertung eines Periodogramms mit q Balken kann mit Hilfe von Verteilungsannahmen und der Theorie des multiplen Testens ein Signifikanzbegriff (Abschnitt 2.6) verwendet werden, der auch die relative Höhe eines Balkens im Vergleich zu den anderen berücksichtigt, bei Optimierung ist dies so nicht möglich. Daher werden in dieser Arbeit Periodogramme mit einer definierten Menge an Testperioden $\{p_1, \ldots, p_q\}$ betrachtet. Details zur Wahl dieser Menge werden in Abschnitt 2.6.2 diskutiert.
Das bekannteste Periodogramm basiert auf der Fourieranalyse und ist für gleichabständige Zeitreihen definiert. Es wird zusammen mit den in dieser Arbeit benötigten Grundlagen der Fourieranalyse in Abschnitt 2.2.1 erläutert und eine der üblichen Erweiterungen auf ungleichmäßige Messzeiten vorgestellt. Anschließend wird das Prinzip der hier verwendeten Periodogramme in Abschnitt 2.2.2 erörtert.

2.2.1. Grundlagen der Fourieranalyse

Sei $y(t)$ eine Funktion, die ohne Beschränkung der Allgemeinheit auf dem Intervall $[0, T]$ definiert sei und sich dann in positive und negative Richtung periodisch fortsetze. In der Fourieranalyse wird genutzt, dass $y(t)$ durch eine Fourierreihe, eine Überlagerung von Sinus- und Kosinusschwingungen der Perioden $T, T/2, T/3, T/4$, usw., darstellbar ist:

$$y(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos\left(k \frac{2\pi}{T} t\right) + b_k \sin\left(k \frac{2\pi}{T} t\right) \right) \qquad t \in \mathbb{R}. \tag{2.13}$$

Die Funktionen $\mathbb{1}_{\mathbb{R}}(t)$, $\cos(1\frac{2\pi}{T}t)$, $\cos(2\frac{2\pi}{T}\omega t)$,..., $\sin(1\frac{2\pi}{T}t)$, $\sin(2\frac{2\pi}{T}t)$,..., sind paarweise orthogonal und es existiert eine eindeutige Darstellung von $y(t)$ gemäß Gleichung (2.13). Die Fourierkoeffizienten $a_0, a_1, \ldots, b_1, b_2, \ldots$ geben dabei die Amplitude der jeweiligen Schwingung in der Funktion an. Die Summe $a_k^2 + b_k^2$ beschreibt den Beitrag der Sinus- und Kosinusschwingung mit Periode T/k und lässt sich auch berechnen durch $|\mathfrak{F}(p)|^2$ mit

$$\mathfrak{F}(p) = \int_{-\infty}^{\infty} y(t) \exp\left(-\mathrm{i} \frac{2\pi}{p} t\right) dt \qquad \text{für } p \in \{T, T/2, T/3, T/4, \ldots\},$$

wobei $\mathrm{i} = \sqrt{-1}$ die imaginäre Einheit ist. $\mathfrak{F}(p)$ wird auch Fouriertransformierte genannt, $|\mathfrak{F}(p)|^2$ Spektrum oder Fourier-Periodogramm. In ihm ist ablesbar, mit welcher Amplitude die jeweilige phasenverschobene Sinusschwingung in y vertreten ist.
Eine diskretisierte Zeitreihe mit gleichabständigen Messzeiten $t_1 = 1, \ldots, t_n = n$, n gerade, lässt sich durch eine endliche Fouriersumme

$$y(t_j) = \frac{a_0}{2} + \sum_{k=1}^{T/2-1} \left(a_k \cos\left(k \frac{2\pi}{T} t_j\right) + b_k \sin\left(k \frac{2\pi}{T} t_j\right) \right) + a_{T/2} \cos\left(\frac{T}{2} \frac{2\pi}{T} t_j\right) \tag{2.14}$$
$$j = 1, \ldots, n, \quad a_0, \ldots, a_{n/2-1}, b_1, \ldots, b_{n/2} \in \mathbb{R},$$

ausdrücken, wobei $T = n$ die Dauer der Zeitreihe ist (vgl. Chatfield 2003, Kapitel 7, auch für ähnliche Formulierungen für n ungerade und anderen Abstand $\delta = t_i - t_{i-1}$). Die Funktion

kann damit über Sinus- und Kosinus-Schwingungen der Perioden $T, T/2, \ldots, T/(\frac{1}{2}T) = 2$ und eine Konstante modelliert werden. Die Vektoren

$$1_n, \left[\cos\left(1\frac{2\pi}{T}t_j\right)\right]_{j=1,\ldots,n}, \left[\cos\left(2\frac{2\pi}{T}t_j\right)\right]_{j=1,\ldots,n}, \ldots, \left[\cos\left(\frac{T}{2}\frac{2\pi}{T}t_j\right)\right]_{j=1,\ldots,n},$$
$$\left[\sin\left(1\frac{2\pi}{T}t_j\right)\right]_{j=1,\ldots,n}, \left[\sin\left(2\frac{2\pi}{T}t_j\right)\right]_{j=1,\ldots,n}, \ldots, \left[\sin\left(\left(\frac{T}{2}-1\right)\frac{2\pi}{T}t_j\right)\right]_{j=1,\ldots,n} \tag{2.15}$$

sind für gleichabständige Messzeiten weiterhin paarweise orthogonal. Die Inversen der verwendeten Perioden werden auch Fourierfrequenzen genannt, sie sind gleichabständig. Die Koeffizienten $a_0, \ldots, a_{n/2-1}, b_1, \ldots, b_{n/2}$ heißen (reellwertige) Fourierkoeffizienten. Die diskrete Fouriertransformierte ergibt sich zu

$$\mathfrak{F}_n(p) = \frac{1}{n}\sum_{j=1}^{n} y_j \exp\left(-\mathrm{i}\frac{2\pi}{p}t_j\right) \qquad \text{für } p \in \{T, T/2, T/3, T/4, \ldots, 2\} \tag{2.16}$$

und das zugehörige Periodogramm oder Spektrum zu

$$\mathrm{Per}_{\mathrm{Fourier}}(p) = c_n \left|\sum_{j=1}^{n} y_j \exp\left(-\mathrm{i}\frac{2\pi}{p}t_j\right)\right|^2 \text{ für } p \in \{T, T/2, T/3, T/4, \ldots, 2\}\,. \tag{2.17}$$

Eine übliche Wahl ist $c_n = \frac{1}{n}$ (Bloomfield 2000, S. 143, Deeming 1975). Das Fourier-Periodogramm von p entspricht der quadrierten geschätzten Amplitude (Schwarzenberg-Czerny 1998a) bzw. der erklärten Varianz $\mathrm{SY} - \mathrm{SE}(p)$ (Scargle 1982) bei Anpassung einer um Null zentrierten Sinusschwingung mit Periode p durch Kleinste-Quadrate-Regression an die zentrierte Zeitreihe.

Das Messzeitmuster hat einen Effekt darauf, welche Testperioden zusammen mit p_f einen erhöhten Periodogrammbalken haben. Bei gleichabständigen Messzeiten kommt es eher zu sogenanntem Aliasing, weil Sinusschwingungen unterschiedlicher Perioden auf gleichabständigen Messzeiten nicht immer unterscheidbar sind (vgl. Abbildung 2.6 und Deeming 1975 und Mudelsee et al. 2009). Aus diesem Grund empfehlen Dawson und Fabrycky (2010) sogar explizit die Verwendung ungleichmäßiger Beobachtungszeiten.

Bei Vorliegen unregelmäßiger Messzeiten kommt in der Praxis häufig das Deeming- oder Schuster-Periodogramm (Deeming 1975) zur Anwendung (beispielsweise bei Webb et al. 1988, Uttley, McHardy und Papadakis 2002, Vaughan et al. 2003). Dabei wird $\mathrm{Per}_{\mathrm{Fourier}}(p)$ einfach mit den ungleichmäßig vorliegenden Messzeiten berechnet. Dann sind jedoch die Vektoren in (2.15) nicht mehr notwendigerweise orthogonal (vgl. Schwarzenberg-Czerny 1998a), und $\mathrm{Per}_{\mathrm{Fourier}}(p)$ entspricht nicht mehr dem Quadrat der durch eine Kleinste-Quadrate-Regression angepassten Amplitude. Der so genannte Leakage-Effekt (vgl. Deeming 1975), bei dem Periodogrammbalken bedingt durch die fehlende Orthogonalität erhöht werden (vgl. Abbildung 2.7), wird verstärkt.

Für ungleichmäßige Messzeiten ist allgemein bekannt (vgl. z.B. Hall und Li 2006), dass diese Methode auch auf periodische Strukturen im Sampling reagiert. Zur leichteren Nachvollziehbarkeit dieses Sachverhaltes wird das folgende Beispiel konstruiert (vgl. Abbildung 2.8): Gegeben seien zwei Zeitreihen $\mathfrak{Y} = (t_i^y, y_i)_{i=1,\ldots,20}$ und $\mathfrak{Z} = (t_i^z, z_i)_{i=1,\ldots,28}$ mit

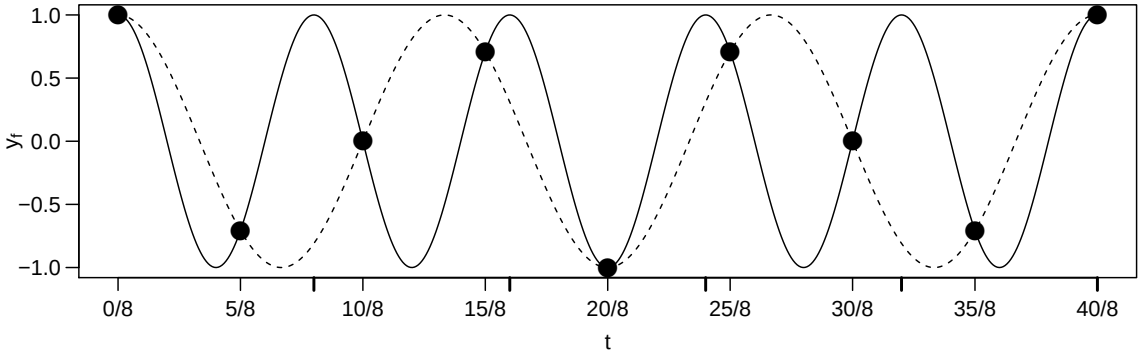


Abbildung 2.6.: Beispiel für samplingbedingtes Aliasing. Die Beobachtungen (schwarze Kreise) folgen sowohl einer Sinusschwingung mit Periode 1 (durchgezogene Linie), als auch einer der Periode $\frac{5}{3}$ (gestrichelte Linie). Ein ungleichmäßigeres Sampling erleichtert häufig die Unterscheidung.

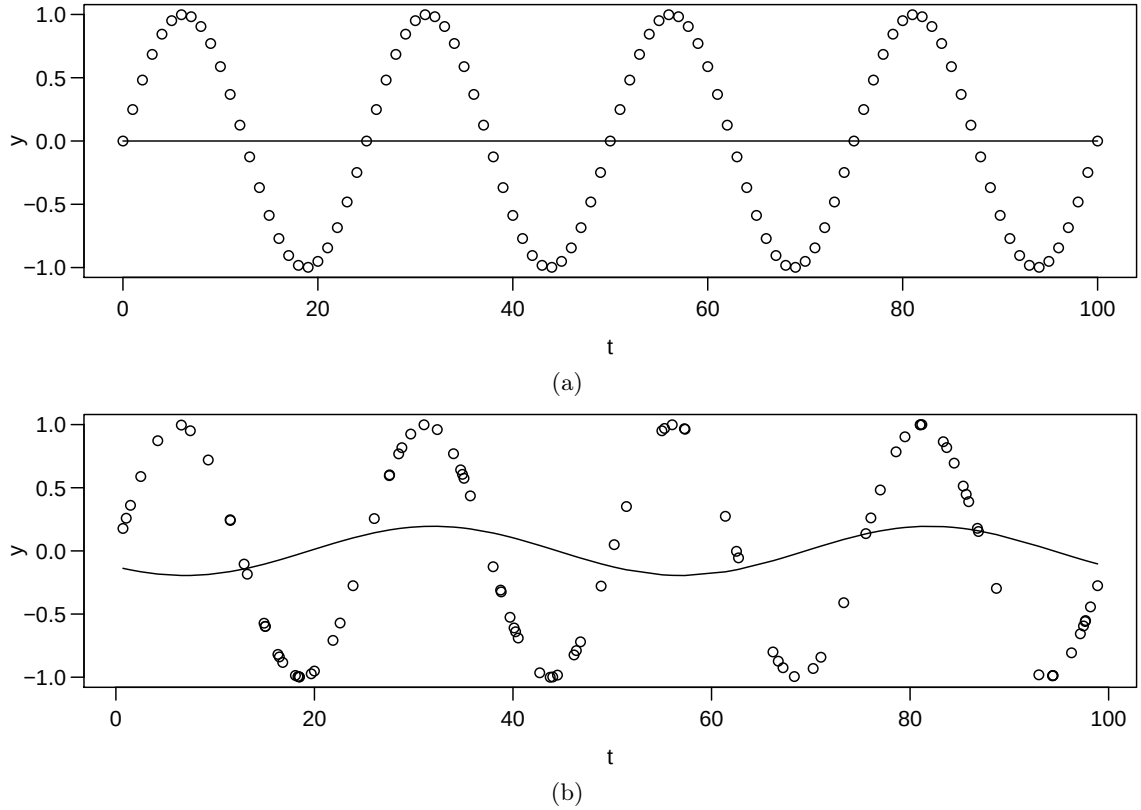
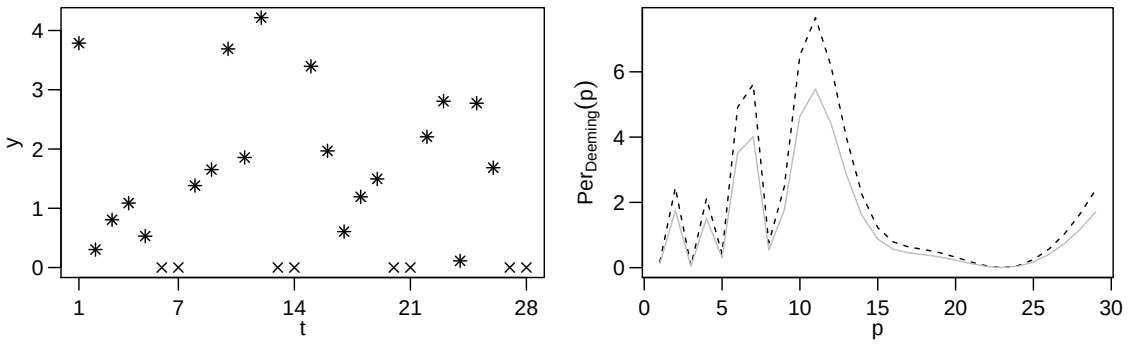


Abbildung 2.7.: Illustration der (Un-)Abhängigkeit zweier Testperioden. (a) Sinusfunktion der Periode $n/4$ (Punkte) mit KQ-angepasster Sinusfunktion der Periode $n/2$ (durchgezogene Linie), alle angepassten Koeffizienten sind null, (b) gleiches Szenario mit ungleichmäßigem Messzeitmuster.



(a) (b)

Abbildung 2.8.: Illustration des Verhaltens des Deeming-Periodogramms bei periodischem Sampling. (a) Zwei Zeitreihen, identisch bis auf die Messzeiten 6,7,13,14,20,21,27,28, zu denen eine den Wert 0 annimmt (×), während die andere zu diesen Zeitpunkten nicht beobachtet wird (+, Überlagerung beider Reihen: ∗). Zeitreihe × ist periodisch in den Messwerten, + ist periodisch im Messzeitmuster. (b) Deeming-Periodogramm von + (gestrichelt) und × (grau), bis auf c_n gleich.

täglichen Messungen. Werktags zeigen beide Zeitreihen den gleichen Wert (unabhängig identisch normalverteiltes Rauschen um 1.5 mit Varianz 1). Am Wochenende (Messzeiten $6, 7, 13, 14, 20, 21, 27, 28$) liegen für $\mathfrak{Y}$ keine Messungen vor, für $\mathfrak{Z}$ stets der Messwert 0. Die Zeitreihe $\mathfrak{Y}$ weist ein periodisches Sampling der Periode 7 auf, aber keine Periodizität in den Messwerten selbst. Die Zeitreihe $\mathfrak{Z}$ hat ein gleichmäßiges unperiodisches Sampling, aber ein periodisches Signal der Länge 7. An Gleichung (2.17) ist leicht erkennbar, dass $\mathfrak{Y}$ und $\mathfrak{Z}$ bis auf die vom Stichprobenumfang abhängige Konstante c_n das gleiche Deeming-Periodogramm haben, obwohl in der einen eine Periodizität in den Messwerten vorliegt, in der anderen nur eine Periodizität im Sampling. Um einzuschätzen, welche Effekte den Messzeiten und welche den Messwerten zuzuordnen sind, empfiehlt Deeming (1975), auch das Spektralfenster

$$W(p) = c_w \left| \sum_{j=1}^{n} e^{\mathrm{i}\frac{2\pi}{T} t_j} \right|^2,$$

zu betrachten, das die Samplingeffekte widerspiegelt. Dies erschwert eine automatisierte Auswertung des Deeming-Periodogramms.

2.2.2. Prinzip der hier verwendeten Periodogramme

In dieser Arbeit geht es um Periodogramme, die das Resultat der Anpassung periodischer Funktionen sind. Eine Vielzahl der für ungleichmäßig beobachtete Zeitreihen vorgestellten Periodogrammmethoden basieren, teils ohne dass dies bei ihrer Einführung betont wurde, auf der separaten Anpassung mehrerer periodischer Funktionen $g\left(\frac{t}{p_1}\right), \ldots, g\left(\frac{t}{p_q}\right)$ an die Zeitreihen, wobei g selbst Periode eins hat. Die Menge $\{p_1, \ldots, p_q\}$ wird Testperiodenmenge genannt. Es wird nicht angenommen, dass g der wahren periodischen Fluktuation f aus (2.3) entspricht, da diese typischerweise unbekannt ist (vgl. Hall, Reimann und Rice 2000). Die den Periodogrammbalken bestimmende Funktion Per hängt von den angepassten Parametern

von g ab. Zum Beispiel passen Zechmeister und Kürster (2009) für das Generalized-Lomb-Scargle-Periodogram eine Sinusfunktion mittels Kleinste-Quadrate-Regression an die Lichtkurve an und verwenden

$$\mathrm{Per}_{GLS}(p) = R^2(p),$$

wobei $R^2(p)$ das Bestimmtheitsmaß bei Anpassung einer Sinusfunktion mit Testperiode p darstellt. Für verschiedene Testperioden p wird das Modell angepasst. Es wird erwartet, dass die Anpassung für die wahre Fluktuationsperiode p_f besonders gut, der Periodogrammbalken also auffällig höher ist als für die anderen Testperioden.
In dieser Arbeit soll dieses Periodogrammprinzip durch das Anwenden anderer linearer Modelle und Regressionstechniken sowie die Möglichkeit der gewichteten Regression erweitert werden. Die Grundstruktur der verwendeten Periodogrammmethoden ist in Abbildung 2.9 skizziert. Modifikationen bezüglich der angepassten Modelle und Regressionstechniken sowie der Einbezug der Messfehler durch gewichtete Regression wurden in der Vergangenheit bereits vorgenommen (vgl. Tabelle 2.1), bisher fehlte jedoch eine strukturierte Kombination der resultierenden Möglichkeiten und ein experimenteller Vergleich der entstehenden Methoden. Dies soll im Rahmen dieser Arbeit geschehen.

2.3. Verwendete Modelle für die periodische Fluktuation

Dieser Abschnitt behandelt die periodischen Funktionen, die in dieser Arbeit an die Lichtkurve angepasst werden, und die Abbildung $\mathcal{X}$ (vgl. Gleichung (2.6)), mit deren Hilfe die Anpassung durch ein lineares Modell erfolgen kann. Da die Periodogrammbalken unabhängig voneinander und stets nach dem gleichen Prinzip berechnet werden, wird hier häufig die Abhängigkeit von der Testperiode nicht mitnotiert, zum Beispiel bei R^2 statt $R^2(p)$.

2.3.1. Sinusfunktion

Die Sinusfunktion mit Periode 1, Amplitude A, Mesor β_1 und Phase ϕ

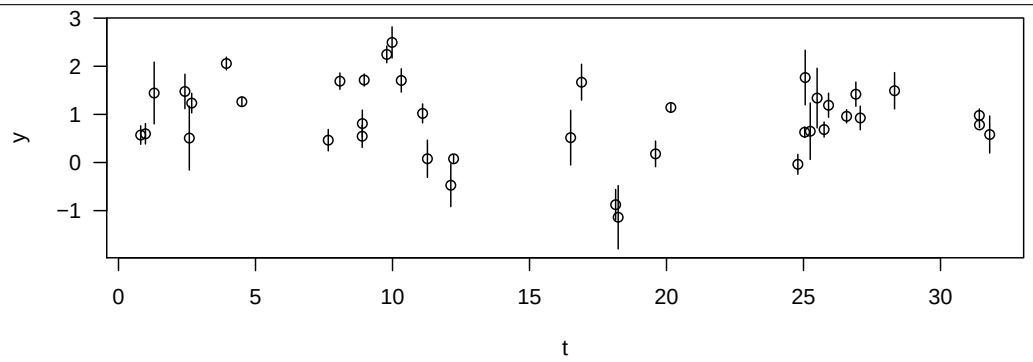
$$g(t) = \beta_1 + A\sin\left(2\pi t + \phi\right), \phi \in [0, 2\pi[\tag{2.18}$$

lässt sich mit einem Additionstheorem der Trigonometrie (vgl. Hackbusch, Schwarz und Zeidler 2003, S. 56)

$$\sin(u+v) = \sin(u)\cos(v) + \cos(u)\sin(v) \tag{2.19}$$

leicht linearisieren zu:

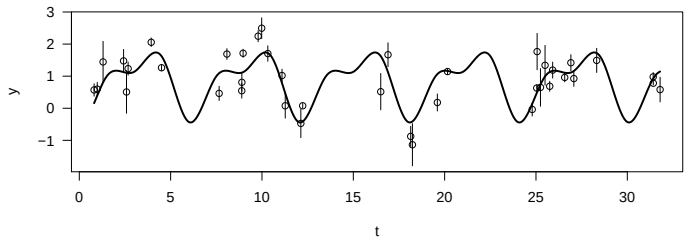
$$g(t) \underset{(2.19)}{=} \beta_1 + \beta_2\cos\left(2\pi t\right) + \beta_3\sin\left(2\pi t\right) \tag{2.20}$$
$$\text{mit } \beta_2 = A\sin(\phi), \beta_3 = A\cos(\phi).$$



Daten (Abschnitt 2.1): Lichtkurve $(t_i, y_i, s_i)_{i=1,\dots,n}$
Messfehler s_i sind als vertikale Balken eingezeichnet.

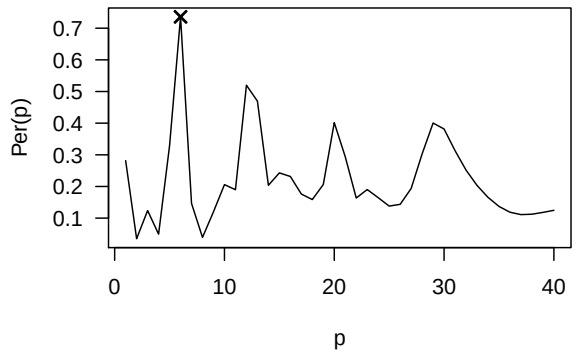
Für jede Testperiode $p \in \{p_1, \dots, p_q\}$

Funktion g mit Periode p an die Lichtkurve anpassen.



Wähle dazu:
- Periodische Funktion g (Abschnitt 2.3)
- Regressionstechnik (Abschnitt 2.4)
- Ob Messfehler s_i mittels gewichteter Regression einbezogen werden sollen (Abschnitt 2.5)

Anpassungsgüte ist Periodogrammbalken $\mathrm{Per}(p)$ (Abschnitt 2.6)



Entdecke auffällige Periode im Periodogramm $\mathrm{Per}(p_1), \dots, \mathrm{Per}(p_q)$ (Abschnitt 2.6)

Abbildung 2.9.: Schema zum Prinzip der in dieser Arbeit verwendeten Periodogramme und gleichzeitig zum Aufbau dieses Kapitels

Funktion – Regressionstechnik		Publikation (Periodogrammbezeichnung)
Einfachstufenfunktion	– KQ	Leahy et al. (1983) (Epoch-Folding)
	– KQ	Schwarzenberg-Czerny (1989) (Analysis-of-Variance)
Doppelstufenfunktion	– KQ	Stellingwerf (1978) (Phase-Dispersion-Minimization)
Sinus	– KQ	Scargle (1982) (Lomb-Scargle)
	– KQ	Zechmeister und Kürster (2009) (*Generalized-Lomb-Scargle*)
	– KQ	Cumming, Marcy und Butler (1999) (*Floating Mean*)
	– KQ	Ferraz-Mello (1981) (*Date-Compensated-Fourier-Transform**)
	– KQ	Reegen (2007) (*SigSpec**)
	– L1	Li (2009)*, Li (2010)*
	– LTS	Ahdesmäki et al. (2007)*
	– MCD	Ahdesmäki et al. (2007)*
	– MT	Ahdesmäki et al. (2007)*
	– MH	Zhang und Chan (2005)*
Fouriers. 2./3. Grades	– KQ	Hall, Reimann und Rice (2000)
	– KQ	Palmer (2009) (*Fast-Chi-Square*)
Splinefunktion	– KQ	Akerlof et al. (1994)
	– KQ	Hall, Reimann und Rice (2000)
	– KQ	Oh et al. (2004) (Generalized-Cross-Validation)
	– MH	Oh et al. (2004) (Robust-Cross-Validation)

Tabelle 2.1.: Übersicht bisher publizierter Periodogrammmethoden, die auf Anpassung einer periodischen Funktion an eine ungleichmäßig beobachtete Zeitreihe basieren. Regressionstechniken: Kleinste-Quadrate (KQ), Kleinste-Beträge (L1), Least-Trimmed-Squares (LTS), Minimum-Covariance-Determinant (MCD), M-Regression mit der Tukey-Funktion (MT), M-Regression mit der Huber-Funktion (MH). Kursiv gesetzte Methoden berücksichtigen Messfehler durch Nutzung gewichteter Regression. Die mit * gekennzeichneten Methoden verwenden zur Berechnung der Periodogrammbalken die angepassten Parameter (Amplitude, bei SigSpec auch Phasenverschiebung). Bei allen anderen Methoden hängen die Periodogrammbalken von SE und SY ab (vgl. Abschnitt 2.1.2).

Mit Hilfe des Regressors

$$\mathcal{X}(t,p) = \left(1, \cos\left(\frac{2\pi}{p}t\right), \sin\left(\frac{2\pi}{p}t\right)\right) \tag{2.21}$$

kann damit die Designmatrix zur Anpassung einer Sinusfunktion aufgestellt werden.
Das Anpassen einer Sinusfunktion wurde in verschiedenen Fachgebieten vorgeschlagen, zum Beispiel in der Chronobiologie (Mojón, Fernández und Hermida 1992, Mattes et al. 1991), in der Genetik (Ahdesmäki et al. 2007), in der Signalverarbeitung (Zhang und Chan 2005) und in der Astroteilchenphysik (Ferraz-Mello 1981, Scargle 1982, Cumming, Marcy und Butler 1999, Rojo-Álvarez et al. 2002, Reegen 2007, Zechmeister und Kürster 2009, Li 2012). Eine der am häufigsten[1] verwendeten Methoden ist das Lomb-Scargle-Periodogramm von Scargle (1982), bei dem die Sinusschwingung mittels Kleinste-Quadrate-Regression angepasst wird. Unter Verwendung der Definitionen (2.11) und (2.12) (vgl. Seite 11) ist es definiert als

$$\mathrm{Per}_{LS} = \mathrm{SY} - \mathrm{SE}\,.$$

Das Lomb-Scargle-Periodogramm wird auch in der Geophysik (Kirchner, Feng und Neal 2000), der Biologie (Ruf 1999), der Chronobiologie im Speziellen (Someren et al. 1999), der Medizin (Laguna, Moody und Mark 1998, Fortin und Mackey 1999, Skrøvseth und Godtliebsen 2011), der Neurophysiologie (Ruskin et al. 1999) und der Molekularbiologie (Glynn, Chen und Mushegian 2006) verwendet.
Das Lomb-Scargle-Periodogramm hat jedoch einen Nachteil, der von späteren Autoren kritisiert wurde: Bei dieser Methode wird bei der Anpassung der Sinusschwingung in Gleichung (2.18) auf die Konstante β_1 verzichtet und stattdessen mit der mittelwertbereinigten Zeitreihe gearbeitet. Bei gleichmäßig verteilten Messzeiten und der Annahme, dass eine ganzzahlige Anzahl Zyklen beobachtet wurde (also den Voraussetzungen der klassischen Fourieranalyse), ist der Kleinste-Quadrate-Schätzer für β_1 das arithmetische Mittel der Daten und Per_{LS} äquivalent zum Fourier-Periodogramm $\mathrm{Per}_{Fourier}$. Bei der in dieser Arbeit angenommenen Datenlage ist dies nicht der Fall. Hier wird daher – analog zu dem bereits erwähnten Generalized-Lomb-Scargle-Periodogram von Zechmeister und Kürster (2009) und Arbeiten von Ferraz-Mello (1981) und Cumming, Marcy und Butler (1999) – die Funktion mit Konstante angepasst, wie sie in Gleichung (2.18) eingeführt wurde.

[1] Unter www.scholar.google.com werden am 28.11.2013 insgesamt 3230 wissenschaftliche Arbeiten gefunden, die den dazugehörigen Artikel zitieren.

2.3.2. Stufenfunktion

Eine weitere periodische Funktion, die in der Astroteilchenphysik klassischerweise angepasst wird, ist die periodische Stufenfunktion mit m Stufen

$$g(t) = \begin{cases} \gamma_1 & \varphi_p(t) \in [0, k_1[\\ \gamma_2 & \varphi_p(t) \in [k_1, k_2[\\ \vdots & \\ \gamma_m & \varphi_p(t) \in [k_{m-1}, 1[\end{cases} \quad \text{mit } 0 < k_1 < k_2 < \ldots < k_{m-1} < 1. \tag{2.22}$$

Die Periodogrammmethoden, die auf ihr basieren, werden jedoch häufig nicht explizit als auf einem linearen Modell beruhende Methoden vorgestellt. Die Autoren verstehen sie als Streuungsminimierungsverfahren (z.B. das Phase-Dispersion-Minimization-Periodogramm von Stellingwerf 1978). Gesucht wird nach der Periode, die zu der kleinsten Streuung innerhalb der Klasse und der größten Streuung zwischen den Klassenmitteln führt, in welche die Messwerte gemäß ihrer Phase klassiert werden. Schwarzenberg-Czerny (1998b) erwähnt, dass beim Analysis-of-Variance-Periodogramm (Schwarzenberg-Czerny 1989) und beim Epoch-Folding-Periodogramm (Leahy et al. 1983) Klassen mit Klassengrenzen k_i zu bilden nichts anderes bedeutet als eine Stufenfunktion mit Sprungstellen k_i zu definieren, und Streuung minimieren einer Kleinste-Quadrate-Anpassung entspricht. Speziell gilt für das Epoch-Folding-Periodogramm

$$\mathrm{Per}_{EF} = (n-1)R^2, \tag{2.23}$$

für das Analysis-of-Variance-Periodogramm

$$\mathrm{Per}_{AoV} = \frac{n-m}{m-1}\left(\frac{\mathrm{SY}}{\mathrm{SE}} - 1\right) \tag{2.24}$$

und für das Jurkevich-Periodogramm (Jurkevich 1971)

$$\mathrm{Per}_{Jur} = \mathrm{SY} - \mathrm{SE}\,. \tag{2.25}$$

Das Phase-Dispersion-Minimization-Periodogramm ist für überlappende Klassen definiert, wodurch Stellingwerf (1978) glattere Periodogramme erreichen will. Die direkte Analogie zu einer Stufenfunktion ist damit nicht möglich. Wenn man jedoch davon ausgeht, dass sich die Klassen in zwei überschneidungsfreie Mengen aufteilen lassen, die jeweils eine Partition des Intervalls $[0, 1[$ bilden, lassen sich zwei Stufenfunktionen g_1 und g_2 anpassen. Das Phase-Dispersion-Minimization-Periodogramm kann dann umgerechnet werden zu

$$\mathrm{Per}_{PDM} = \frac{n-1}{n-m}\left(1 - \frac{R_1^2 + R_2^2}{2}\right), \tag{2.26}$$

wobei R_1^2 das Bestimmtheitsmaß der Anpassung von g_1 und R_2^2 das der Anpassung von g_2 ist.

Für das Lafler-Kinman-Periodogramm (Lafler und Kinman 1965), bei dem keine Klassengrenzen definiert sind, lässt sich ebenfalls ein regressionsbasierter Ansatz finden. Ist der Stichprobenumfang n geradzahlig, müssen auch zwei Stufenfunktionen g_1 und g_2 angepasst werden, es gilt:

$$\mathrm{Per}_{LK} = \frac{2}{\mathrm{SY}} \left(\mathrm{SE}_1 + \mathrm{SE}_2\right), \tag{2.27}$$

wobei SE_1 bzw. SE_2 entsprechend Gleichung (2.12) auf Seite 11 für die Anpassung von g_1 bzw. g_2 berechnet wird. Ist n ungerade, sind es sogar n Stufenfunktionen $g_1, \ldots, g_n$ und es gilt

$$\mathrm{Per}_{LK} = \frac{4}{(n-1)\,\mathrm{SY}} \sum_{j=1}^{n} \mathrm{SE}_j, \tag{2.28}$$

mit SE_j gemäß Gleichung (2.12) für die Anpassung von Funktion g_j, $j = 1, \ldots, n$. Dabei liegen die Sprungstellen der Funktionen stets zwischen aufeinanderfolgenden umgeklappten Messwerten, die Stufenfunktion ist also von den Messzeiten abhängig.
Die Gleichungen (2.23) bis (2.28) sind bei Schwarzenberg-Czerny (1998b) nicht hergeleitet und die Schwierigkeit im Fall von überlappenden Klassen, das Periodogramm weiterhin als Ergebnis einer Regression zu definieren, wird nicht thematisiert. Daher befindet sich in Anhang A dieser Arbeit die Herleitung der oben genannten Formeln.
Wird eine einzelne Stufenfunktion g angepasst, ist im Folgenden von einer Einfachstufenfunktion die Rede. Zur Berücksichtigung des Ansatzes von Stellingwerf (1978) beim Phase-Dispersion-Minimization-Periodogramm werden auch zwei Stufenfunktionen g_1 und g_2, deren Stufenmitten jeweils auf den Sprungstellen der anderen Funktion liegen, separat angepasst. Dieses Vorgehen wird im Folgenden das Anpassen einer Doppelstufenfunktion genannt. Sowohl bei der Einstufen- als auch bei der Doppelstufenfunktion werden alle Stufen gleich breit und unabhängig von der Lage der Messzeiten gewählt. Die parallele Anpassung von mehr als zwei Stufenfunktionen erfolgt in dieser Arbeit nicht.
In der Simulationsstudie in Kapitel 4 und in den Anwendungen in Kapitel 5 werden Stufenfunktionen mit zehn Stufen angepasst. Im Vergleich mit einem vierstufigen Modell traten in den Vorstudien zu dieser Arbeit keine großen Unterschiede bezüglich der Detektierbarkeit von Perioden auf. In Thieler et al. (2013) konnte allerdings beobachtet werden, dass eine Stufenfunktion mit zehn Stufen vor allem für die Periodenerkennung bei bestimmten periodischen gipfelförmigen Fluktuationen (hier f_{peak}, vgl. Gleichung (3.9) auf Seite 48) besser geeignet ist als andere Modelle, vermutlich, weil die Stufenfunktion mit mehr Stufen flexibler anpassbar ist. Der Nachteil ist eine längere Laufzeit. Für die Zukunft kann eine Optimierung über die Stufenzahl und Sprungstellen im Anpassungsverfahren angedacht werden (vgl. Abschnitt 6.1.2).

2.3.3. Fouriersumme

Deutlich seltener als Sinus- und Stufenfunktionen werden in der Literatur Periodogrammmethoden mit komplexeren periodischen Funktionen vorgeschlagen. Hall, Reimann und Rice (2000) schlagen vor, eine Fouriersumme $\frac{m-1}{2}$-ten Grades

$$g(t) = \beta_1 + \sum_{i=1}^{\frac{m-1}{2}} \beta_{i+1} \sin(i2\pi t) + \sum_{i=1}^{\frac{m-1}{2}} \beta_{i+\frac{m+1}{2}} \cos(i2\pi t), \qquad \frac{m-1}{2} \in \mathbb{N} \tag{2.29}$$

anzupassen. Für $m > 1$ und ungerade entsteht ein Periodogrammbalken hier also durch die Anpassung einer Sinusschwingung und einige ihrer Oberschwingungen, während er bei der klassischen Fourieranalyse nur auf der Anpassung einer Sinusschwingung basiert (vgl. Abschnitt 2.2.1). In dieser Arbeit werden Fouriersummen mit $m = 5$ (zweiten Grades, wie auch bei Hall, Reimann und Rice 2000) und $m = 7$ (dritten Grades, auch bei Hall, Reimann und Rice 2000 und Palmer 2009) betrachtet, außerdem ist die Sinusschwingung aus Gleichung (2.20) eine Fouriersumme ersten Grades.

2.3.4. Splinefunktion

Eine andere komplexere periodische Funktion ist die periodische Splinefunktion. Zur Verwendung in der Periodogrammberechnung wird sie von Akerlof et al. (1994) vorgeschlagen. Eine Splinefunktion vom Grad k mit m Knoten $\lambda_0 < \lambda_1 < \ldots < \lambda_{m+1}$ ist eine $(k-1)$-mal stetig differenzierbare Funktion, die auf jedem Intervall $[\lambda_i, \lambda_{i+1}]$, $i = 0, \ldots, m$, als Polynom vom Grad k ausdrückbar ist (vgl. Dierckx 1993, S. 3). Splinefunktionen können als Linearkombination von Basisfunktionen, welche nur von den gewählten Knoten abhängen, formuliert werden.
In dieser Arbeit werden periodische kubische ($k = 3$) Splinefunktionen mit $m = 4$ gleichabständigen Knoten pro Zyklus angepasst. Kubische Splines werden auch bei Akerlof et al. (1994), Hall, Reimann und Rice (2000) und Oh et al. (2004) verwendet. Die Knotenanzahl stellt einen Kompromiss zwischen Komplexität und Flexibilität des Modelles dar, kann aber auch anders gewählt werden. Akerlof et al. (1994) verwenden beispielsweise sechs Knoten und Hall, Reimann und Rice (2000) bis zu 13 Knoten. In beiden Fällen wird ein ausreißerfreies Szenario betrachtet. Bei Oh et al. (2004) wird die Knotenanzahl adaptiv gewählt. Abbildung 2.10 zeigt die vier benötigten Basisfunktionen M_0, M_1, M_2 und M_3 einer Splinefunktion g mit Periode 1. Ihre Konstruktion gemäß Dierckx (1993) wird in Anhang B erläutert. Da durch sie alle kubischen Splines (inklusive einer konstanten Funktion) mit den oben genannten Knotenpunkten ausgedrückt werden können, ist die regressorbildende Funktion

$$\mathcal{X}(t,p) = \Big(M_0\big(\varphi_p(t)\big), M_1\big(\varphi_p(t)\big), M_2\big(\varphi_p(t)\big), M_3\big(\varphi_p(t)\big)\Big)$$

und es wird keine eigene Einserspalte zur Lokationsmodellierung benötigt.

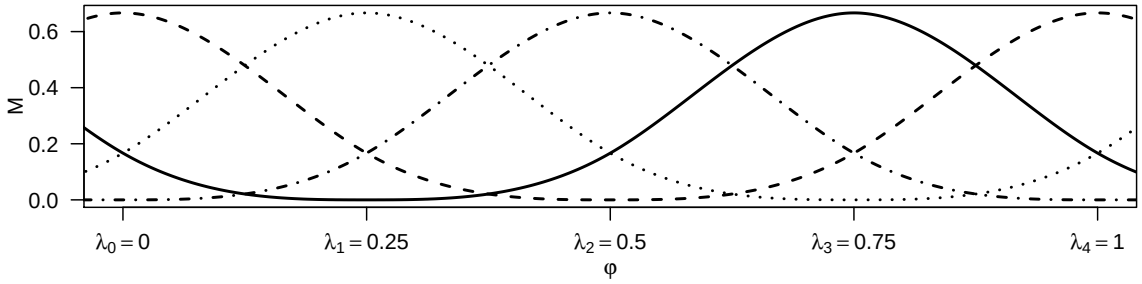


Abbildung 2.10.: Verwendete Basisfunktionen (M_0: gestrichpunktet, M_1: durchgezogen, M_2: gestrichelt, M_3: gepunktet) für eine kubische Splinefunktion mit vier gleichabständigen Knoten pro Zyklus und Periode eins. Zur Generierung dieser Grafik wurde das R-Paket `splines` (R Core Team 2013) verwendet.

Zusammenfassung und Anmerkung

Sechs periodische Funktionen werden im Rahmen dieser Arbeit an Lichtkurven angepasst: Die Einfachstufenfunktion, die Doppelstufenfunktion, die Sinusfunktion, die Fouriersummen zweiten und dritten Grades und die kubische Splinefunktion mit vier gleichabständigen Knoten pro Zyklus (im Folgenden einfach Splinefunktion).
Erwünscht ist, dass die Funktion für Fluktuationsperiode p_f gut angepasst werden kann, für die anderen Testperioden schlecht. Dies kann dann nicht der Fall sein, wenn die Lichtkurvendaten nicht nur durch p_f, sondern auch durch eine andere Testperiode modellierbar sind. Ein klassisches Beispiel ist der Aliasingeffekt (vgl. Abschnitt 2.2.1 und Abbildung 2.6), der bei ungleichmäßigen Messzeiten weniger stark ausfällt.
Zu einem ähnlichen Effekt, der jedoch nicht durch die Änderung des Messzeitenmusters gemindert werden kann, kommt es, wenn die angepasste Funktion mehr Parameter als nötig hat. In Abbildung 2.11 werden hierfür Beispiele gegeben. Durch Betrachtung des kompletten Periodogramms und des Phasendiagramms der interessierenden Testperiode kann man die kleinste passende Periode wählen.

2.4. Robuste Regressionstechniken

Gegeben sei das Modell

$$y = X\beta + e,$$

wobei $y \in \mathbb{R}^n$ der Vektor der Beobachtungen, $X \in \mathbb{R}^{n\times m}$ die Designmatrix, $\beta \in \mathbb{R}^m$ der anzupassende Parametervektor und $e \in \mathbb{R}^n$ der Fehlervektor ist. Bekanntermaßen ist der Kleinste-Quadrate (KQ)-Schätzer $\widehat{\beta}_{KQ}$ von β definiert als

$$\beta_{KQ} = \arg\min_{\beta} \sum_{i=1}^{n} (y - X\beta)_i^2 = \arg\min_{\beta} \zeta_{KQ}(y - X\beta). \tag{2.30}$$

Der KQ-Schätzer ist nicht robust: Ein einziger großer Ausreißer kann genügen, um die Schätzung beliebig stark zu stören.

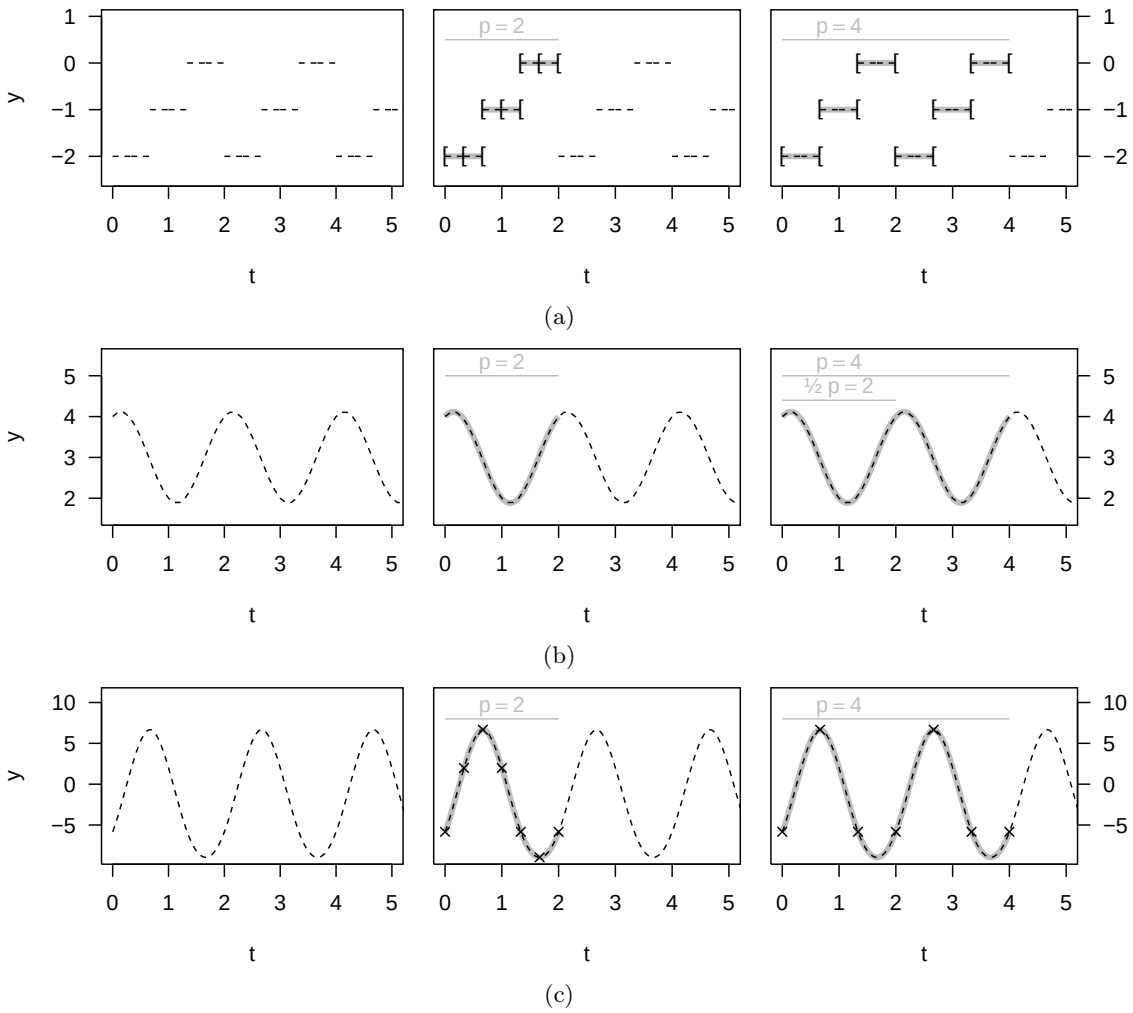


Abbildung 2.11.: Drei Beispiele für periodische Fluktuationen f (gestrichelt schwarz), die wegen Überparametrisierung der Funktion g (grau und fett) durch verschiedene Fluktuationsperioden modelliert werden können. Jeweils Links: Die Originalfluktuation f mit $p_f = 2$. Mitte: Beschreibung von f durch g mit Testperiode $p = 2$. Rechts: Beschreibung von f durch g mit Testperiode $p = 4$. Funktionen: (a) Stufenfunktionen, f mit drei, g mit sechs Stufen (Sprungstellen durch [angedeutet). (b) f Sinusfunktion, g Fouriersumme zweiten Grades. (c) Splinefunktionen, f mit vier, g mit sieben Knoten pro Zyklus (durch Kreuze angedeutet).

Die in Abschnitt 2.3 vorgestellten Modelle wurden bisher meist durch KQ-Regression angepasst, es wurden aber auch erste Vorschläge zur robusten Anpassung unterbreitet (vgl. Tabelle 2.1 auf Seite 18). Im Folgenden werden die robusten Regressionstechniken vorgestellt, die in dieser Arbeit Anwendung finden: M-Regression, Least-Trimmed-Squares (LTS)-Regression, τ-Regression und S-Regression. Ein bisheriger Einsatz der beiden letztgenannten Techniken in Periodogrammmethoden ist nicht bekannt.

2.4.1. Least-Trimmed-Squares (LTS) -Regression

Bei der Least-Trimmed-Squares-Regression von Rousseeuw und Yohai (1984) fließen die größten Residuen nicht in die Zielfunktion ein, es gilt

$$\beta_{LTS} = \arg\min_{\beta} \zeta_{LTS}(y - X\beta), \tag{2.31}$$
$$\text{mit } \zeta_{LTS}(r) = \sum_{i=1}^{h} r_{(i)}^2, \; r \in \mathbb{R}^n,$$

wobei der mit Klammern versehene Index (i) für den i-ten geordneten Wert steht und $h \in \mathbb{N} \cap [\frac{n}{2}, n]$ gewählt wird. Im Fall $h = n$ entspricht die LTS- der KQ-Regression. Im Fall $h < n$ werden die $n - h$ größten Residuen ignoriert, wodurch die Regressionstechnik robuster auf Ausreißer reagiert. Andererseits führt der nichtmonotone Einfluss der Residuen dazu, dass die Zielfunktion ζ_{LTS} lokale Minima aufweisen kann.
Wenn die Regressoren in allgemeiner Lage liegen, das heißt alle Matrizen $\widetilde{X}$ aus $\left\{\widetilde{X} \in \mathbb{R}^{m \times m} : \{\widetilde{X}_{1\cdot}, \ldots, \widetilde{X}_{m\cdot}\} \subset \{X_{1\cdot}, \ldots, X_{n\cdot}\}\right\}$ vollen Rang haben, so hat die Regressionstechnik für

$$h = \left\lfloor \frac{n}{2} \right\rfloor + \left\lfloor \frac{m+1}{2} \right\rfloor$$

einen asymptotischen Bruchpunkt von 0,5 (vgl. Rousseeuw und Yohai 1984). Liegen die Daten nicht in allgemeiner Lage, wie es zum Beispiel bei den Stufenfunktionen (Abschnitt 2.3.2) der Fall ist, kann der Bruchpunkt niedriger liegen. Der LTS-Schätzer ist, gemessen an seinem hohen Bruchpunkt, robust und außerdem regressions-, skalen- und affin äquivariant (Rousseeuw und Leroy 1987, S. 132). Er hat aber eine niedrige asymptotische Effizienz an der Normalverteilung von 0,07 (vgl. Maronna, Martin und Yohai 2006, S. 132), das heißt bei normalverteiltem Fehlerterm beträgt die Varianz des KQ-Schätzers nur 7% der Varianz der LTS-Schätzung. Ahdesmäki et al. (2007) verwenden LTS-Regression zur Periodogrammberechnung. Da sie mit multivariaten Zeitreihen arbeiten, verwenden die Autoren auch die Minimum-Covariance-Determinant- (MCD-)Regression von Rousseeuw (1985). Im in dieser Arbeit betrachteten univariaten Kontext reduziert sich die MCD-Regression auf die LTS-Regression (vgl. Rousseeuw 1985, S. 291).

2.4.2. M-Regression

Während bei der LTS-Regression ein fester Anteils großer Residuen das Zielkriterium nicht beeinflusst, wird der Einfluss auf das Zielkriterium bei der M-Regression durch Verwendung

anderer Abstandsmaße geschwächt. Ein M-Regressionsschätzer $\widehat{\beta}_M$ ist im Allgemeinen (vgl. Maronna, Martin und Yohai 2006, S. 98) gegeben durch

$$\begin{aligned} \widehat{\beta}_M &= \arg\min_{\beta} \zeta_M(r(\beta)) \\ \text{mit } r(\beta) &= \frac{y - X\beta}{\sigma}, \\ \text{und } \zeta_M(r) &= \sum_{i=1}^{n} \rho(r_i), \end{aligned} \tag{2.32}$$

wobei σ die Varianz des Fehlerterms ist und $\rho : \mathbb{R} \to \mathbb{R}_{\geq 0}$ eine Abstandsfunktion, die den Einfluss der einzelnen Residuen steuert. Die Funktion ρ ist um Null symmetrisch, steigt für positive Werte monoton, nimmt an der Stelle null den Wert null an und strebt, sofern sie beschränkt ist, für große Werte den Funktionswert eins an. Sie wird ρ-Funktion genannt (vgl. Maronna, Martin und Yohai 2006, S. 31). KQ-Regression mit $\rho_{KQ}(\nu) = \nu^2$ ist ein skaleninvarianter Spezialfall, für den σ nicht bekannt sein muss. Ebenso verhält es sich bei der ρ-Funktion für die L1-Regression $\rho_{L1}(\nu) = |\nu|$. Der resultierende L1-Schätzer $\widehat{\beta}_{L1}$ ist nicht unbedingt eindeutig (vgl. Maronna, Martin und Yohai 2006, S. 99). Ein einfaches Beispiel ist hier die Lokationsschätzung an eine gerade Anzahl von n Beobachtungen $y_1, \ldots, y_n$, bei der alle Werte im Intervall $[y_{(\frac{n}{2})}, y_{(\frac{n}{2}+1)}]$ die Zielfunktion in (2.32) minimieren. L1-Anpassung einer Sinusfunktion wird in Li (2009) und Li (2010) zur Periodogrammberechnung verwendet.
Für ρ-Funktionen mit streng monoton steigender erster Ableitung existieren eindeutige Lösungen der M-Schätzung (vgl. Maronna, Martin und Yohai 2006, S. 350). So auch bei der M-Schätzung mit Huberfunktion (vgl. Maronna, Martin und Yohai 2006, S. 26)

$$\rho_{MH}(\nu) = \begin{cases} \nu^2, & |\nu| \leq k \\ 2k|\nu| - k^2, & |\nu| > k \end{cases}, \quad k \in \mathbb{R}_{>0}, \tag{2.33}$$

im Folgenden M-Huber-Regression genannt. Über den Wert k lässt sich die Effizienz an der Normalverteilung steuern. Je größer k, umso effizienter ist die M-Huber-Regression, Ausreißer haben aber auch einen höheren Einfluss auf das Regressionsergebnis. In dieser Arbeit wird $k = 1.345$ gewählt, was einer Effizienz von 0,95 entspricht (vgl. Maronna, Martin und Yohai 2006, S. 27). M-Huber-Regression wird in Zhang und Chan (2005) zur Periodogrammberechnung durch Anpassung einer Sinusfunktion verwendet und bei Oh et al. (2004) durch Anpassung einer Splinefunktion. Ahdesmäki et al. (2007) nutzen die Tukey- (auch Biweight- oder Bisquare-) Funktion

$$\rho_{MT}(\nu) = \begin{cases} 1 - \left(1 - \left(\frac{\nu}{k}\right)^2\right)^3, & |\nu| \leq k \\ 1, & |\nu| > k \end{cases}, \quad k \in \mathbb{R}_{>0}. \tag{2.34}$$

Diese Regressionstechnik wird im Folgenden M-Tukey-Regression genannt. Ihre Effizienz an der Normalverteilung ist wieder über k kontrollierbar und wird für diese Arbeit als 0,95 gewählt, dies entspricht $k = 4.68$ (vgl. Maronna, Martin und Yohai 2006, S. 30). Die

Funktion ρ_{MT} ist für Werte betraglich größer k konstant, das bedeutet, jedes (skalierte) Residuum kann maximal den Einfluss 1 erreichen. Dies macht den Schätzer sehr robust, andererseits ist damit die Ableitung von ρ_{MT} nicht streng monoton steigend und β_{MT} ist nicht nur nicht eindeutig, die Zielfunktion in (2.32) kann auch lokale Optima haben. In der Periodogrammberechnung wird M-Tukey-Regression von Ahdesmäki et al. (2007) zur Anpassung eines Sinus verwendet.
In dieser Arbeit werden als M-Regressionstechniken L1-Regression, M-Huber-Regression, M-Tukey-Regression und zum Vergleich KQ-Regression zur Anpassung periodischer Modelle verwendet. Sowohl bei M-Huber- als auch bei M-Tukey-Regression besteht die Problematik, dass σ nicht bekannt ist. Sie wird üblicherweise gelöst, indem entweder β und σ alternierend optimiert werden, oder indem β zunächst durch eine andere robuste skaleninvariante Regressionstechnik angepasst und σ anhand der resultierenden Residuen geschätzt wird, bevor der M-Schätzer für β mittels der fest gehaltenen Schätzung für σ bestimmt wird (vgl. Maronna, Martin und Yohai 2006, Kap. 4.4). In dieser Arbeit wird die zweite Vorgehensweise verwendet, da in Vorversuchen kein starker Unterschied zwischen diesen beiden Lösungen festgestellt werden konnte und die zweite deutlich weniger Rechenschritte benötigt.

2.4.3. S-Regression

Der Wert der minimierten KQ-Zielfunktion ζ_{KQ} (vgl. Definition (2.30) auf Seite 23) kann auch als Schätzer für die n-fache Varianz der Residuen betrachtet werden. Analog minimiert der S-Schätzer von Rousseeuw und Yohai (1984)

$$\widehat{\beta}_S = \arg\min_{\beta} \zeta_S(y - X\beta),$$

wobei $\zeta_S(r)$ ein M-Schätzer der Varianz des Vektors r ist, also (vgl. Maronna, Martin und Yohai 2006, S. 35) so, dass

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i}{\zeta_S(r)}\right) = \delta.$$

Der Wert $\delta \in \mathbb{R}_{>0}$ wird als Erwartungswert von $\mathrm{E}(\rho(r))$ gewählt, üblicherweise unter der Annahme $r_1, \ldots, r_n \underset{\text{u.i.v.}}{\sim} \mathcal{N}(0,1)$. Die Funktion ρ ist dabei eine spezielle ρ-Funktion: Sie ist stetig und differenzierbar, steigt auf dem Intervall $[0, c[$ streng monoton und ist auf $[c, \infty[$ konstant. Die Tukey-Funktion aus Gleichung (2.34) erfüllt diese Anforderung. In dieser Arbeit wird wie in Rousseeuw und Yohai (1984) $c = 1.547$ gewählt, woraus $\delta = 0{,}5$ und ein asymptotischer Bruchpunkt von 0,5 folgt. Wenn die Daten in allgemeiner Lage sind, ist der asymptotische Bruchpunkt des S-Schätzers wie beim LTS-Schätzer 0,5, die maximale asymptotische Effizienz an der Normalverteilung von 0,287 (vgl. Rousseeuw und Leroy 1987, Tabelle 19) ist jedoch höher. S-Regression wurde bisher nicht zur Periodogrammberechnung eingesetzt.

2.4.4. τ-Regression

Wie bei der S-Regression wird bei der τ-Regression von Yohai und Zamar (1988) ein Skalenschätzer minimiert:

$$\begin{aligned} \widehat{\beta}_\tau &= \arg\min_\beta \zeta_\tau(y - X\beta) \\ \text{mit } \zeta_\tau(r) &= \widehat{\sigma}_M(r)^2 \frac{1}{n a_2} \sum_{i=1}^{n} \rho_2\left(\frac{r_i}{\widehat{\sigma}_M(r)}\right) \\ \text{und } \frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{r_i}{\widehat{\sigma}_M}\right) &= a_1. \end{aligned} \tag{2.35}$$

Dabei beruht ζ_τ wie die S-Regression auf einem M-Schätzer der Varianz, welcher hier jedoch nicht die zu minimierende Zielfunktion ist. Für die Wahl $a_1 = 0{,}5$, $a_2 = 0{,}128$, $\rho_1(\nu) = \rho(\nu, c = 1.214)$ und $\rho_2(\nu) = \rho(\nu, c = 3.270)$ mit

$$\rho(\nu, c) = \begin{cases} 1.38\left(\frac{\nu}{c}\right)^2 & \left|\frac{\nu}{c}\right| \le \frac{2}{3} \\ 0{,}55 - 2.69\left(\frac{\nu}{c}\right)^2 + 10{,}76\left(\frac{\nu}{c}\right)^4 - 11.66\left(\frac{\nu}{c}\right)^6 + 4.04\left(\frac{\nu}{c}\right)^8 & \frac{2}{3} < \left|\frac{\nu}{c}\right| \le 1 \\ 1 & \left|\frac{\nu}{c}\right| > 1 \end{cases}$$

hat der Schätzer eine asymptotische Effizienz von 0,95 und einen asymptotischen Bruchpunkt von 0,5 (vgl. Salibian-Barrera, Willems und Zamar 2008).

Zusammenfassung

In Abschnitt 2.4 werden die robusten Regressionstechniken vorgestellt, die in dieser Arbeit zur Periodogrammberechnung eingesetzt werden sollen. Diese sind: LTS-, L1-, M-Huber-, M-Tukey-, S- und τ-Regression. Die vier erstgenannten Techniken wurden in der Literatur schon zur Periodogrammberechnung eingesetzt (vgl.Tabelle 2.1). Eine bisherige Untersuchung zur Eignung von S- und τ-Regression zur Periodogrammberechnung ist nicht bekannt. Als Regressionstechniken, deren Minimierungskriterium, wie auch bei der LTS-Regression, zugleich ein robuster Varianzschätzer ist, sind sie für die Periodogrammberechnung jedoch interessant: Die entstehenden Periodogrammbalken können, wie auch bei der Kleinsten-Quadrate-Regression, als Anteil der über das Modell erklärten Variabilität interpretiert werden.

2.5. Gewichtete Regression als Umgang mit Messfehlern

Bei der Verwendung des klassischen linearen Modells aus Gleichung (2.7) (vgl. Seite 10) zur Anpassung periodischer Funktionen an Lichtkurven werden sowohl die in Gleichung (2.4) (vgl. Seite 8) angenommene Heteroskedastizität der Fehlervarianz als auch die gegebenen Informationen in Form von Messfehlern s_i außer Acht gelassen. Das heteroskedastische Modell lautet

$$Y_i = x_i^\mathsf{T}\beta + Y_{w;i} \qquad \text{mit } Y_{w;i} \underset{\text{u.v.}}{\sim} \mathcal{N}(0, \sigma_i^2) \tag{2.36}$$

mit $\sigma_1, \ldots, \sigma_n > 0$. Unter der Annahme, dass die exakten Fehlervarianzen bekannt sind, also $s_i^2 = \sigma_i^2$ für $i = 1, \ldots, n$, lässt sich Gleichung (2.36) leicht zu einem homoskedastischen Modell umformen:

$$\widetilde{Y}_i = \widetilde{x}_i^\mathsf{T}\beta + \widetilde{Y}_{w;i} \qquad \text{mit } \widetilde{Y}_{w;i} \underset{\text{u.i.v.}}{\sim} \mathcal{N}(0,1) \tag{2.37}$$
$$\text{und } \widetilde{Y}_i = \frac{Y_i}{s_i}, \qquad \widetilde{x}_i = \frac{1}{s_i} x_i, \qquad \widetilde{Y}_{w;i} = \frac{Y_{w;i}}{s_i}.$$

Die Berücksichtigung der Messfehler s_i kann also durch das Anpassen des Modells (2.37) geschehen, was im Folgenden gewichtete Regression heißt. Die Annahme exakt bekannter Fehlervarianzen ist hierbei sehr stark und vermutlich nicht praxisnah. Für eine erste Untersuchung zum Nutzen der Messfehlerberücksichtigung kann sie hier jedoch getroffen werden. Dieses Thema wird in Abschnitt 4.3.3 nochmals aufgegriffen, wo die Periodendetektion mit gewichteten Periodogrammen im Rahmen einer Simulationsstudie eingehender untersucht wird.

Die Messfehler s_i werden von einigen bereits verwendeten Periodogrammmethoden wie hier beschrieben berücksichtigt (vgl. Tabelle 2.1 auf Seite 18). Eine Ausnahme bildet Reimann (1994), der zur Gewichtung nicht die Messfehler $s_i, i = 1, \ldots, n$ verwendet, sondern $\check{s}_i$ mit $\check{s}_n = \sqrt{s_n + s_1}$ und $\check{s}_i = \sqrt{s_i + s_{i+1}}$ für $i = 1, \ldots, n-1$. Dies ergibt sich aus einem anderen Verständnis des verwendeten Lafler-Kinman-Periodogramms als Streuungsminimierungsverfahren statt regressionsbasierter Methode. Wie Tabelle 2.1 zu entnehmen ist, wird in allen Arbeiten zu gewichtet einsetzbaren Periodogrammen die KQ-Regression verwendet. Zur Anwendung gewichteter robuster Regression zur Berücksichtigung hetereoskedastischer Fehler in der Periodogrammberechnung ist mir keine Arbeit bekannt.

Der Einsatz gewichteter Regression kann in das bisher vorgestellte Verfahren unabhängig von der gewählten Regressionstechnik und dem gewählten Modell als Vorschritt eingebaut werden, indem das Modell nach Aufstellen der Designmatrix X entsprechend (2.37) modifiziert wird. Zur Anpassung des Lokationsmodells wird dann statt des Vektors $\mathtt{i} = 1_n$ (vgl. Abschnitt 2.1.2) der Vektor $\widetilde{\mathtt{i}} = (1/s_i)_{i=1,\ldots,n}$ verwendet.

2.6. Periodogrammbalken und Detektion auffälliger Balken

Unter Verwendung einer Regressionstechnik aus Abschnitt 2.4 kann nun für eine Testperiode p eine periodische Funktion $g(\cdot/p)$ aus Abschnitt 2.3 angepasst werden. Der Periodogrammbalken soll die Güte der Anpassung angeben, weil davon ausgegangen wird, dass die Anpassung einer periodischen Funktion mit der Fluktuationsperiode besser gelingt als die mit einer falschen Testperiode. In der Literatur (vgl. Tabelle 2.1, Seite 18) wurden Funktionen vorgeschlagen, die auf SY und SE (Gleichungen (2.11) und (2.12), vgl. Seite 11) basieren. In dieser Arbeit wird das Bestimmtheitsmaß

$$R^2 = 1 - \frac{\mathrm{SE}}{\mathrm{SY}}$$

verwendet, das außer für die Kleinste-Quadrate-Regression in dieser Form auch für die M-Regression (Maronna, Martin und Yohai 2006, S. 171) und für eine robuste Varianzschätzung minimierenden Regressionstechniken wie LTS-, S- und τ-Regression (Croux und Dehon 2003) definiert wurde.
Einige Periodogrammmethoden, bei denen eine Sinusfunktion angepasst wird, verwenden als Periodogrammbalken die quadrierte Amplitude (z.B. Ferraz-Mello 1981, Ahdesmäki et al. 2007). Dazu gehören alle in Tabelle 2.1 aufgeführten Methoden, bei denen eine Sinusfunktion mit robuster Regression angepasst wird. In dieser Arbeit wird diese Periodogrammdefinition nicht verwendet, da die Amplitude für andere als trigonometrische Funktionen nicht definiert ist.

2.6.1. Entwicklung eines Detektionskriteriums

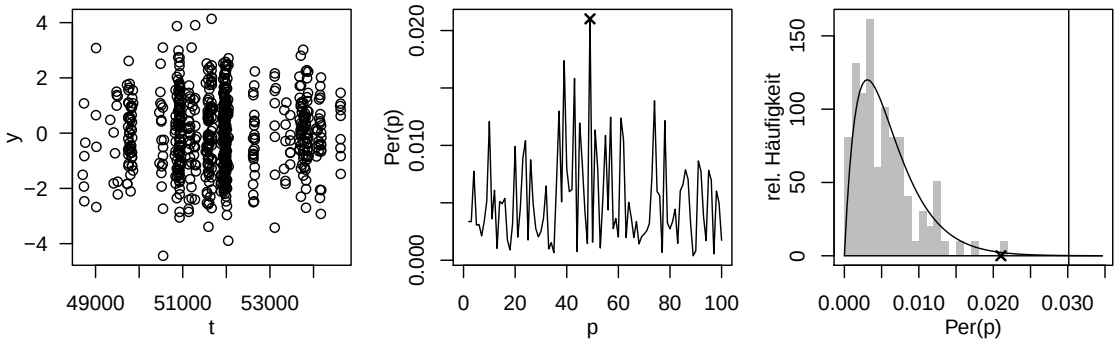
Von Interesse ist ein automatisiertes Verfahren, um mit Hilfe eines berechneten Periodogramms zu entscheiden, ob eine Testperiode auffällig gut angepasst werden konnte oder ob kein Periodogrammbalken auffällig hoch ist. Da nach Abschnitt 2.1 auch Lichtkurven möglich sind, die nicht periodisch fluktuieren, das heißt $f \equiv \text{const.}$, genügt es nicht, die Periode mit dem höchsten Periodogrammbalken als auffällige Periode zu definieren (vgl. auch Abb. 2.12). Es bietet sich daher an, die Verteilung der Periodogrammbalken zu berücksichtigen.
Wenn die n Messwerte unabhängig identisch normalverteilt sind, und ein Modell mit vollrangiger Designmatrix $X \in \mathbb{R}^{n\times m}$ mittels Kleinste-Quadrate-Regression angepasst wird, gilt für das Bestimmtheitsmaß (vgl. Schwarzenberg-Czerny 1998b)

$$R^2 \sim \mathcal{B}\left(\frac{m-1}{2}, \frac{n-m}{2}\right), \tag{2.38}$$

wobei $\mathcal{B}(\theta_1, \theta_2)$ für die Betaverteilung mit streng positiven Parametern θ_1 und θ_2 steht. Der Nachweis wird in Anhang C ausgeführt, da er in der Literatur nicht gefunden werden konnte.
Sei $b_\alpha(\theta_1, \theta_2)$ das α-Quantil der $\mathcal{B}(\theta_1, \theta_2)$-Verteilung. Dann gilt

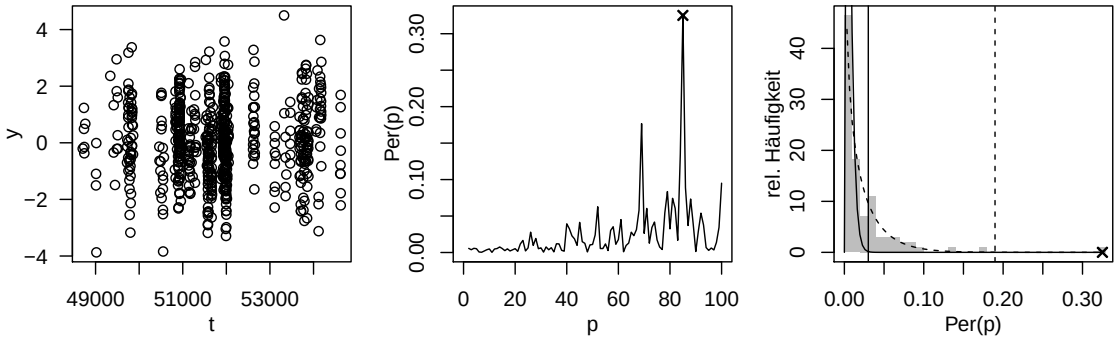
$$P\left(\text{Per}(p) > b_{1-\alpha}\left(\frac{m-1}{2}, \frac{n-m}{2}\right) \middle| f \equiv \text{const.}\right) = \alpha.$$

Ein Überschreiten des kritischen Wertes $b_{1-\alpha}\left(\frac{m-1}{2}, \frac{n-m}{2}\right)$ durch den Periodogrammbalken $\text{Per}(p)$ bedeutet, dass das angepasste periodische Modell signifikant mehr Varianz in der Lichtkurve als in beliebigem normalverteiltem Rauschen erklären kann. Dieses Vorgehen entspricht einem Test im Modell $Y_i = g(t_i/p) + Y_{w;i}$ mit der Nullhypothese „$g(t) = const.$" und der Alternativhypothese „$g(t) \neq const.$".



(a) (b) (c)

Abbildung 2.12.: Periodendetektion für eine unperiodische Lichtkurve: (a) Lichtkurve unabhängig identisch normalverteilter Messwerte ohne periodische Fluktuation, (b) Periodogramm (M-Tukey-Anpassung einer Splinefunktion), der höchste Balken ist mit × gekennzeichnet, (c) Histogramm der Periodogrammbalken mit Dichte der $\mathcal{B}\left(\frac{m-1}{2} = 1{,}5,\ \frac{n-m}{2} = 325{,}5\right)$-Verteilung (durchgezogene Linie), Wert des höchsten Periodogrammbalken auf der Abszisse mit × gekennzeichnet. Der maximale Periodogrammbalken liegt nicht über dem $\sqrt[q]{1-0{,}05}$-Quantil der Verteilung (durchgezogene vertikale Linie), wird also nicht detektiert.



(a) (b) (c)

Abbildung 2.13.: Periodendetektion für eine periodische Lichtkurve: (a) Lichtkurve mit Messwerten mit Fluktuationskomponente ($p_f = 85$) und unabhängig identisch normalverteiltem Rauschen, (b) Periodogramm (M-Tukey-Anpassung einer Splinefunktion), der höchste Balken ist mit × gekennzeichnet, (c) Histogramm der Periodogrammbalken, mit Dichte der vordefinierten $\mathcal{B}\left(\frac{m-1}{2} = 1{,}5,\ \frac{n-m}{2} = 325{,}5\right)$-Verteilung (durchgezogene Linie) und einer angepassten $\mathcal{B}(0{,}78,\ 30{,}72)$-Verteilung (gestrichelt), Wert des höchsten Periodogrammbalken auf der Abszisse mit × gekennzeichnet. Vertikale Linien markieren das $\sqrt[q]{1-0{,}05}$-Quantil der jeweiligen Verteilung. Die angepasste Verteilung beschreibt die Periodogrammbalkenverteilung (außer für $\mathrm{Per}(p_f)$) besser als die vordefinierte. $\mathrm{Per}(p_f)$ liegt über dem Quantil, wird also detektiert.

Bei Vorliegen eines Periodogramms mit mehreren Balken (der übliche Fall) kann unter der Annahme, dass diese wie in der klassischen Fourieranalyse unabhängig identisch verteilt sind, die Verteilung des Maximums genutzt werden:

$$P\left(\max_{i=1,\ldots,q}\left(\mathrm{Per}(p_i)\right) > b_{\sqrt[q]{1-\alpha}}\left(\frac{m-1}{2}, \frac{n-m}{2}\right)\middle| f \equiv \text{const.}\right) = \alpha. \tag{2.39}$$

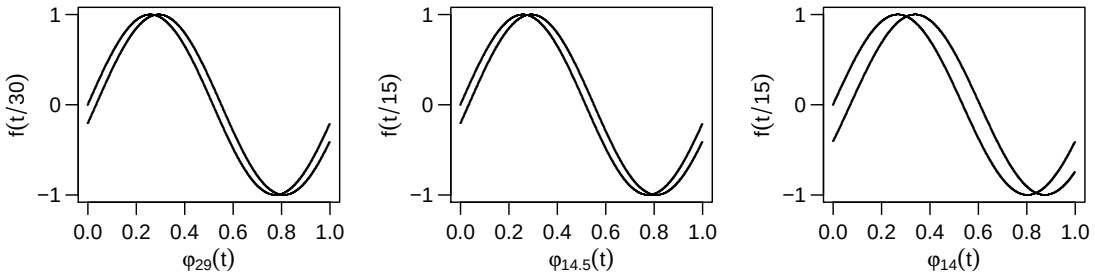
Ein solches Vorgehen (auch mit anderen Verteilungen, wenn der Periodogrammbalken anders definiert wurde) ist in der Astroteilchenphysik bereits üblich (beispielsweise bei Cumming 2004 oder Sturrock und Scargle 2010). Bei der Anwendung dieses Verfahrens im vorliegenden Fall gibt es einige Schwierigkeiten, die zu diskutieren sind:

KQ-Regression ist in dieser Arbeit nicht die einzige verwendete Regressionstechnik. Zusätzlich werden auch robuste Regressionstechniken angewandt. Simulationen (vgl. Anhang D) rechtfertigen, dass auch hier eine Betaverteilung (mit freien Parametern) geeignet ist.

Null- und Alternativhypothese sind falsch formuliert. Die Nullhypothese sollte für Testperiode p auch umfassen, dass eine periodische Fluktuation mit Periode $p_f \neq p$ vorliegt. Im Modell $Y_i = g(t_i/p) + Y_{w;i}$ ist also die Nullhypothese "$g(t) = const. \vee p \neq p_f$". In Abbildung 2.13 ist zu sehen, dass für $p_f \neq p$ die $\mathcal{B}\left(\frac{m-1}{2}, \frac{n-m}{2}\right)$-Verteilung weniger zur Verteilung der Periodogrammbalken passt, rein heuristisch scheint aber eine andere Betaverteilung denkbar zu sein. Die Parameter der Betaverteilung werden daher in dieser Arbeit angepasst. Diese Anpassung an die Periodogrammbalken erfolgt robust, da erwartet wird, dass für mindestens ein $p \approx p_f$ im Vergleich zu den anderen Testperioden eine bessere Anpassung erfolgt und der entsprechende Periodogrammbalken damit ein Ausreißer nach oben ist. Zur robusten Anpassung einer Gammaverteilung haben Clarke, McKinnon und Riley (2012) gute Erfahrungen mit der Minimierung der Cramér-von-Mises-Distanz gemacht. Die Cramér-von-Mises-Distanz ist für beliebige univariate Verteilungen definiert als

$$\begin{aligned}\Delta_{CvM}(\theta) &= \int_0^\infty \left(F_n(u) - F_\theta(u)\right)^2 dF_\theta(u) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(F_\theta(u_{(i)}) - \frac{i-0{,}5}{n}\right)^2 + \frac{1}{12n^2},\end{aligned} \tag{2.40}$$

wobei θ der Parametervektor der Verteilung ist, F_θ ihre Verteilungsfunktion, $u_{(1)}$, $u_{(2)}, \ldots, u_{(n)}$ die geordnete Stichprobe und F_n die empirische Verteilungsfunktion zur Stichprobe. In Vorstudien zu der vorliegenden Arbeit funktionierte diese Technik auch für Betaverteilungen. Die robuste Anpassung einer univariaten Verteilung und die Auswahl der Beobachtungen, die über einem gewissen Quantil liegen, wird zur Ausreißeridentifikation in Davies und Gather (1993) vorgeschlagen. In dieser Arbeit und einhergehenden Publikationen wird erstmalig Ausreißeridentifikation zur Bestimmung auffälliger Perioden in Periodogrammen eingesetzt.



(a) (b) (c)

Abbildung 2.14.: Illustration zur relativen statt absoluten Betrachtung von Periodenabständen: Phasendiagramme von Sinusschwingungen mit Periode p_f, im Phasendiagramm nach Periode p. (a) $p_f = 30$, $p = 29$, (b) p_f=15, $p = 14{,}5$,(c) p_f=15, $p = 14$. In (a) und (b) ist das Verhältnis p_f/p gleich, in (a) und (c) ist die Differenz $p_f - p$ gleich.

Im Falle gewichteter Regression mit nicht exakten Messfehlern s_i ist der Fehlerterm nicht mehr normalverteilt (vgl. Abschnitt 2.5). Wie sich dies auf die Leistung des Verfahrens ausübt, wird in Kapitel 4.3.3 untersucht.

Die Annahme der unabhängig identisch verteilten Periodogrammbalken ist nicht gerechtfertigt. Speziell kann dies dazu führen, dass viele von der wahren Periode p_f abhängige Periodogrammbalken ebenfalls erhöhte Werte zeigen und trotz robustem Vorgehen eine Verteilung mit höherem Erwartungswert angepasst wird, so dass $\mathrm{Per}(p_f)$ nicht über dem gewünschten $\mathcal{B}$-Quantil liegt. Der Test wird also konservativer. Dieses Problem und mögliche Lösungen werden eingehender in Abschnitt 2.6.2 diskutiert.

2.6.2. Testperioden und Umgang mit Abhängigkeiten

Wie bereits in Abschnitt 2.2.1 erläutert wurde, sind die in der klassischen Fourieranalyse verwendeten $\frac{n}{2}$ Testperioden in Frequenz gleichabständig und die mit ihnen gebildeten Regressoren sind bei Anpassung von Sinusschwingungen auf gleichabständigen Messzeiten paarweise orthogonal. Diese Testperioden werden auch für andere Periodogrammmethoden verwendet, z.B. von Israel und Stella (1996) für unregelmäßig beobachtete Lichtkurven.
Der Vorteil dieser Wahl ist bei gleichabständigen Zeitreihen, dass eine Zeitreihe auf diesen Messzeiten in Sinus-und Kosinuskomponenten zerlegbar ist. Dieser Unabhängigkeitsvorteil geht verloren, wenn das Messzeitmuster ungleichmäßig ist oder mit einer anderen periodischen Fluktuation gearbeitet wird. Hinzu kommt, dass diese Testperioden häufig nur unzureichend die Perioden abdecken, für die man sich interessiert (vgl. Islam 2011). Für die letztgenannte Problematik gibt es zwei übliche Auswege: Einige Anwender wählen Testperioden, jedoch weiterhin so, dass ihre Frequenzen gleichabständig sind (vgl. Ismailov und Adygezalzade 2012, Sturrock und Scargle 2010). Andere Anwender wählen die Testperioden selbst gleichabständig, etwa Benlloch et al. (2001), Kudryavtseva und Pyatunina (2006) und Rödig et al. (2009). Bei letzterem Vorgehen, dass auch hier verwendet wird, sollte die Distanz zweier Testperioden dennoch als relativ betrachtet werden. Ob der Abstand eins zwischen zwei Perioden groß ist, hängt von ihrer eigenen Größenordnung ab (vgl. auch Abbildung 2.14).

Wie im vorigen Abschnitt erwähnt, ist es für die Ausreißeridentifikation problematisch, wenn die Periodogrammbalken nicht unabhängig voneinander sind, speziell wenn der zur Fluktuationsperiode gehörende hohe Periodogrammbalken durch viele andere hohe Periodogrammbalken nicht mehr als Ausreißer erkannt wird. Hierbei gibt es vier spezielle Abhängigkeiten:

1. Die periodische Fluktuation y_f wird auf dem vorliegenden Messzeitmuster nicht eindeutig durch eine anzupassende Funktion g beschrieben (vgl. Abbildung 2.6, Seite 14).

2. Ein Teil der Streuung der periodischen Fluktuation y_f kann messzeitmusterbedingt durch verschiedene Testperioden erklärt werden (vgl. Abbildung 2.7, Seite 14).

3. Die periodische Fluktuation y_f wird durch die anzupassende Funktion g wegen deren Überparametrisierung nicht eindeutig beschrieben (vgl. Abbildung 2.11, Seite 24).

4. Speziell ist eine periodische Fluktuation der Periode p_f auch leidlich durch eine passende Funktion g mit Periode nahe p_f beschreibbar (vgl. Abbildung 2.14).

Alle vier Fälle führen dazu, dass mit dem Auftreten einer Periodizität in der Lichtkurve nicht nur der Periodogrammbalken der Fluktuationsperiode, sondern auch Balken anderer Testperioden erhöht sind. In den ersten drei hier genannten Fällen müssen diese anderen Testperioden nicht in der Nähe der Fluktuationsperiode p_f liegen. Der letztgenannte Fall führt dazu, dass der Periodogrammbalken zur Fluktuationsperiode auch von erhöhten Periodogrammbalken umgeben ist. Der Vorteil dieser Gipfelbildung um die Fluktuationsperiode p_f ist, dass bei Fehlen der Fluktuationsperiode in der Menge der Testperioden noch eine Testperiode detektiert werden kann, die nahe p_f liegt. Unabhängig davon, ob man die Testperioden oder die Testfrequenzen gleichabständig wählt, wird es ohne weiteres Vorwissen nicht möglich sein, sicherzustellen, dass die wahre Fluktuationsperiode p_f unter den Testperioden ist.
Die genannten positiven Abhängigkeiten lassen befürchten, dass der zur Fluktuationsperiode gehörige Periodogrammbalken nicht mehr als Ausreißer hervorsticht. Daher werden drei verschiedene Vorgehen zur Findung der passenden Betaverteilung verfolgt. Sie werden Auswertungstypen genannt und im Rahmen einer Simulationsstudie in Kapitel 4 verglichen. Es wird unterschieden in:

Typ 1: Ignorieren vorliegender Abhängigkeiten: Bei diesem Vorgehen wird an das komplette Periodogramm wie beschrieben eine Betaverteilung angepasst. Eine erhöhte Anzahl positiver Periodogrammbalken könnte durch die Robustheit des Verfahrens verträglich sein.

Typ 2: Entfernen von Periodogrammbalken nahe dem höchsten Balken: Bei diesem Vorgehen wird die Betaverteilung an ein reduziertes Periodogramm angepasst, bei dem die Periodogrammbalken um den höchsten fehlen. So soll verhindert werden, dass Abhängigkeiten unter benachbarten Perioden die Anpassung beeinflussen. Die Anzahl q beteiligter Periodogrammbalken wird dementsprechend in (2.39)

reduziert. Bei Zechmeister und Kürster (2009) wird für das Generalized-Lomb-Scargle-Periodogramm (Anpassung einer Sinusfunktion mit Kleinste-Quadrate) angegeben, die Gipfelbreite bei Testfrequenzen betrage ungefähr $1/T$, wobei T die Dauer der Lichtkurve ist. Die daraus resultierende Gipfelbreite von ungefähr p_i^2/T bei Testperiode p_i konnte in Vorstudien zu der vorliegenden Arbeit für sinusförmige Fluktuationen f simulativ bestätigt werden. Für andere periodische Fluktuationen wurde die Gipfelbreite teils unter-, aber nie systematisch überschritten. Ein Entfernen aller Perioden in einer p_i^2/T-Umgebung scheint daher den Gipfel sicher zu entfernen. Detailliertere Ausführungen hierzu befinden sich in Anhang E.

Typ 3: Betrachtung einer Auswahl potentiell unabhängiger Perioden: Zechmeister und Kürster (2009) schlagen vor, in (2.39) statt der Anzahl q der berechneten Periodogrammbalken die verminderte Anzahl $\widetilde{q}$ der lokal maximalen Periodogrammbalken zu verwenden, da diese angeblich unabhängig seien. Diesen Ansatz fortführend wird bei diesem Vorgehen nur an die lokal maximalen Periodogrammbalken eine Betaverteilung angepasst. Einerseits wird ein breiter Gipfel im Periodogramm die Betaverteilung damit nicht nach oben verzerren, andererseits fehlen im reduzierten Periodogramm die niedrigen Werte.

Eine weitere Gefahr besteht darin, eine falsche Periode, die weit entfernt von p_f liegt und deren Periodogrammbalken stark von p_f abhängt, als auffällig zu erkennen, wenn p_f selbst keine Testperiode ist. Dies tritt nur bei starker Abhängigkeit der Perioden auf, zum Beispiel bei Vielfachen von p_f, und kann eventuell erkannt werden, wenn die nach der detektierten Periode $p^\star$ umgeklappte Lichtkurve $(\varphi_{p^\star}(t_i), y_i)$ betrachtet wird.

2.6.3. Das vorgeschlagene Verfahren

In dieser Arbeit wird damit das folgende Verfahren vorgeschlagen, um in Lichtkurven periodische Signale zu detektieren:

1. Wähle Testperioden $p_1, \ldots, p_q$ (Abschnitt 2.6.2), ein Modell für die periodische Funktion g, die angepasst werden soll (Abschnitt 2.3), eine Regressionstechnik (Abschnitt 2.4) und ob die s_i mittels gewichteter Regression einbezogen werden sollen (Abschnitt 2.5).

2. Passe für jede Testperiode p_i, $i = 1, \ldots, q$, die Funktion g mittels der gewählten Regressionstechnik, gegebenenfalls gewichtetet, an. Berechne den Periodogrammbalken $\mathrm{Per}(p_i)$ als das Bestimmtheitsmaß R^2.

3. Passe an die Menge der Periodogrammbalken $\mathrm{Per}(p_1), \ldots, \mathrm{Per}(p_q)$ (oder eine wie in Abschnitt 2.6.2 beschriebene reduzierte Menge) eine Betaverteilung an und bestimme α-Ausreißer gemäß Gleichung (2.39) (Abschnitt 2.6.1). Dies sind die detektierten Perioden.

Dieses Vorgehen kann nur als heuristisch bezeichnet werden, dennoch sei darauf hingewiesen, dass zumindest Hoffnung besteht, nur mit Wahrscheinlichkeit α im Falle von $f \equiv \text{const.}$ eine (falsche) Periode zu detektieren. Dies wird in Abschnitt 4.3 näher untersucht. Um zu

betonen, dass das hier vorgeschagene Verfahren nicht vollständig theoretisch untermauert ist, ist im Folgenden stets von „detektierten", nicht aber von „signifikanten" Perioden und Periodogrammbalken die Rede.

2.7. Andere Ansätze in der Literatur

Obwohl der Ansatz der Anpassung periodischer Funktionen speziell durch das oft verwendete Lomb-Scargle-Periodogramm weit verbreitet ist, gibt es auch andere Ansätze zur Periodenfindung. Viele basieren auf der klassischen Fourieranalyse (Abschnitt 2.7.1) und/oder der Autokovarianz (Abschnitt 2.7.2) oder setzen Glättungsfilter ein (Abschnitt 2.7.3). Weitere bekannte Ansätze sind in Abschnitt 2.7.4) beschrieben.

2.7.1. Fourierbasierte Ansätze

Das Deeming-Periodogramm als Erweiterung des Fourier-Periodogramms auf ungleichmäßige Zeitreihen wurde bereits in Abschnitt 2.2.1 vorgestellt. Ein weiterer Ansatz ist das SparSpec-Periodogramm von Bourguignon, Carfantan und Idier (2007). Dabei wird das Modell (2.14) von Seite 12 mittels LASSO-Regression (vgl. Bühlmann und van der Geer 2011, S.9) angepasst, die zu minimierende Zielfunktion entspricht der der normalen Kleinste-Quadrate-Regression, erweitert um einen Bestrafungsterm, der umso kleiner wird, je weniger der Fourierkoeffizienten ungleich 0 sind. Die Periodogrammbalken entsprechen wie im Fourier-Periodogramm der quadrierten Amplitude der jeweiligen Schwingung. Von Nachteil ist neben der Unrobustheit dieses Verfahrens, dass der Anwender bei den Testperioden auf die Fourierfrequenzen festgelegt ist (vgl. auch Abschnitt 2.6.2).

2.7.2. Auf der Autokovarianz basierte Ansätze

Einige Methoden für gleichmäßig beobachtete Zeitreihen nutzen aus, dass für eine solche, um Null zentrierte Zeitreihe gilt (vgl. Bloomfield 2000, S. 145):

$$\mathrm{Per}_{\mathrm{Fourier}} = \sum_{k=-(n-1)}^{n-1} \widehat{V}(k) \exp\left(-\mathrm{i}\frac{2\pi}{p}k\right) \text{ für } p \in \{T, T/2, T/3, T/4, \ldots\},$$

wobei $\mathrm{Per}_{\mathrm{Fourier}}$ wie in Formel (2.17) (Seite 13) mit $c_n = \frac{1}{n}$ definiert ist und

$$\widehat{V}(k) = \frac{1}{n} \sum_{i:i,i+k\in\{1,\ldots,n\}} y_i y_{i+k}$$

die Autokovarianz zum Zeitabstand k ist.
Die auf dieser Tatsache basierenden Periodogramme (Blackman-Tukey-Korrelogramm, vgl. Pearson et al. 2003, Methoden nach Pearson et al. 2003, nach Huijse et al. 2011 und nach Liu et al. 2011) unterscheiden sich nur in ihrem Schätzer für die Autokovarianz. Dabei eignet sich nur die Methode nach Liu et al. 2011 für ungleichmäßige Messzeiten. Eyer und Genton (1999) betrachten die Autokovarianzfunktion selbst und verwenden den robusten Q_n-Schätzer von Rousseeuw und Croux (1993).

2.7.3. Auf Glättungsfiltern basierte Ansätze

Während bei den fourierbasierten Ansätzen, zu denen im weiteren Sinne auch die korrelationsbasierten Ansätze gehören, ein Sinusmodell die Grundlage ist, setzen die im Folgenden vorgestellten Filterverfahren und die Methoden in Abschnitt 2.7.4 kein bestimmtes Modell für die periodische Fluktuation voraus. Sie nehmen nur an, dass eine kleine Änderung in der Phase auch eine kleine Änderung in der periodischen Fluktuation bedeutet.
Ansätze, bei denen ein Glättungsfilter eingesetzt wird, sind bei McDonald (1986) und Hall, Reimann und Rice (2000) zu finden. Hier wird die Lichtkurve wie in Abschnitt 2.1 erläutert bezüglich einer Testperiode umgeklappt, geglättet und die Abweichungen (Residuen) von den ungeglätteten Daten betrachtet. Da die umgeklappte Lichtkurve keine Ränder hat, sondern zyklisch definiert ist (vergleiche auch Abbildung 2.5, Seite 10), ist beim Filtern keine Randbehandlung nötig. Wurde die Lichtkurve bezüglich der Fluktuationsperiode umgeklappt, ist eine besonders niedrige Streuung der Residuen zu erwarten, das Periodogramm ist daher eine Funktion dieser Streuung.
Die Ansätze unterscheiden sich im verwendeten Filter, wobei alle mit ungleichmäßigen Messzeiten umgehen können, jedoch nur bei McDonald (1986) durch Anwendung einer Ausreißererkennungsregel auf Robustheit geachtet wird.

2.7.4. Weitere Ansätze

Drei weitere bekannte Methoden sind die String-Length von Dworetsky (1983), das Renson-Periodogramm nach Renson (1978) und die Structure-Function von Simonetti, Cordes und Heeschen (1985). Bei der String-Length und dem Renson-Periodogramm werden benachbarte Beobachtungen im Phasendiagramm mit Strecken verbunden und können damit als auf nichtparametrischer Regression beruhende Methoden angesehen werden. Die String-Length misst die Länge des Streckenzuges, beim Renson-Periodogramm werden die quadrierten Anstiege der Teilstrecken summiert. Beide Periodogramme zeigen für die Fluktuationsperiode einen niedrigen Wert an. Die Structure-Function selbst ist nur für gleichabständige Lichtkurven definiert. Durch Erweiterungen für ungleichmäßige Messzeiten (wie etwa von Paltani et al. 1997 oder Liu et al. 2011) ähnelt sie dem Renson-Periodogramm, nutzt aber nicht die komplette Information des Phasendiagramms. Da große Ausreißer zu einer starken Verlängerung der Strecken und steileren An- und Abstiegen im Streckenzug führen, können alle drei Methoden als nicht robust angesehen werden.

Zusammenfassung

In diesem Kapitel wird der zu untersuchende Datentyp, Lichtkurven, vorgestellt sowie das Prinzip, nach dem in dieser Arbeit Periodogramme berechnet und Fluktuationsperioden detektiert werden. Zur Berechnung der Periodogramme werden periodische Funktionen mit unterschiedlichen Perioden (Testperioden) an die Lichtkurve angepasst und das Bestimmtheitsmaß als Periodogrammbalken verwendet. Die zur Anpassung verwendeten Funktionen und Regressionstechniken wurden ebenfalls in diesem Kapitel vorgestellt. Der Schwerpunkt

liegt hierbei auf robusten Regressionstechniken. Die Möglichkeit der gewichteten Regression wurde diskutiert.
Zur Detektion einer Periode wird im Periodogramm nach auffällig hohen Werten gesucht. Dazu wird eine Betaverteilung mittels Cramér-von-Mises-Distanz-Minimierung robust an die Periodogrammbalken angepasst. Es werden Perioden identifiziert, deren Periodogrammbalken über einem hohen Quantil der Verteilung liegt. Mögliche Abhängigkeiten im Periodogramm und auf Ausdünnung des Periodogramms beruhende Techniken zum Umgang damit werden diskutiert.
Abschließend werden andere Ansätze zur Periodogrammberechnung beschrieben, die in der Vergangenheit vorgeschlagen wurden und nicht auf der Anpassung einer periodischen Funktion durch eine Regressionstechnik basieren.

3. Das R-Paket `RobPer`

Die Implementierung der in Kapitel 2 vorgestellten Periodogrammmethoden erfolgt in der statistischen Programmiersprache R (R Core Team 2013) und wurde im Paket `RobPer` (vgl. Thieler, Fried und Rathjens 2013) veröffentlicht und ist online erhältlich[2]. Das Paket enthält neben Programmen zur Bestimmung der Periodogrammbalken (Abschnitt 3.1) auch Funktionen zur Generierung künstlicher Lichtkurven (Abschnitt 3.2). Eine weitere Funktion ermöglicht die Anpassung einer Betaverteilung an die berechneten Periodogrammbalken durch Cramér-von-Mises-Distanz-Minimierung zur Detektion auffälliger Perioden wie in Abschnitt 2.6 beschrieben. Der hierzu verwendete Code ist eine modifizierte Version einer von Brenton R. Clarke zur Verfügung gestellten R-Funktion zur Minimierung der Cramér-von-Mises-Distanz bei Gamma-Verteilungen. Alle in diesem Kapitel vorgestellten R-Funktionen sind, sofern es nicht explizit anders erwähnt ist, im Rahmen dieser Arbeit entstanden.

3.1. Implementierung der Periodogrammmethoden

Mit der R-Funktion `RobPer` wird die Anpassung der periodischen Funktionen an die Lichtkurve durchgeführt. Ausgegeben wird das Periodogramm. Eingabeparameter sind neben der Lichtkurve (Abschnitt 2.1), den gewählten Testperioden, einer periodischen Funktion (Abschnitt 2.3), einer Regressionstechnik (Abschnitt 2.4) und einer dichotomen Variable zur Entscheidung, ob gewichtete Regression durchgeführt werden soll (Abschnitt 2.5), weitere Parameter zur Kontrolle des Anpassungsvorgangs. Eine Übersicht aller Eingabeparameter, eine Auflistung der verwendeten Funktionen und R-Pakete sowie Schaubilder zum Ablauf des Programmes befinden sich in Anhang F. In diesem Abschnitt werden die bei der Implementierung aufgetretenen Probleme und die gewählten Lösungen diskutiert. Es geht dabei stets um die Frage, ob eine und welche bereits vorhandene R-Funktion, gegebenenfalls modifiziert, zur Durchführung der Regression verwendet werden kann, und wenn nein, worauf bei der Neuimplementierung einer Regressionstechnik geachtet werden muss.
Es sind fünf andere R-Pakete bekannt, die Funktionen zur Periodogrammberechnung von Zeitreihen mit ungleichmäßigen Messzeiten beinhalten. Vier davon berechnen das Lomb-Scargle-Periodogramm, welches der KQ-Anpassung einer Sinusfunktion entspricht. Dies sind die R-Pakete `lomb` (Ruf 1999, Funktion `lsp`), `nuspectral` (Mathias et al. 2004, Funktion `lombcoeff`), `cts` (Wang 2013, Funktion `spec.ls`) und `nlts` (Bjornstad 2013, Funktion `spec.lomb`). Weiterhin gibt es das Paket `GeneCycle` (Ahdesmäki, Fokianos und Strimmer 2012, Funktion `robust.spectrum`). Hier sind die Periodogrammbalken die quadrierten Schätzer der Amplitude einer mittels M-Tukey-Regression angepassten Sinusfunktion. Keine

[2] http://cran.r-project.org/web/packages/RobPer

der Periodogrammfunktionen erlaubt die Berücksichtigung von Messfehlern. Die Testperioden entsprechen für die Funktion `spec.ls` und `robust.sprectrum` den Fourierfrequenzen. Für die Funktion `lsp` sind die Testperioden so wählbar, dass entweder sie selbst oder ihre Frequenzen gleichabständig sind. Bei den Funktionen `nuspectral` und `spec.lomb` ist die Testperiodenmenge frei wählbar.

3.1.1. M-Tukey- und M-Huber-Regression

Zur Berechnung eines auf M-Regression basierenden Periodogrammbalkens $\mathrm{Per}(p) = 1 - \frac{\mathrm{SE}(p)}{\mathrm{SY}}$ müssen die in Abschnitt 2.1.2 eingeführten Werte

$$\mathrm{SY} = \min_{\mu} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathtt{i}_i\mu}{\widehat{\sigma}_\mu}\right) \tag{3.1}$$

$$\text{und}\,\mathrm{SE}(p) = \min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i(p)^\mathsf{T}\beta}{\widehat{\sigma}_\beta(p)}\right) \tag{3.2}$$

berechnet werden, die den minimalen Wert der jeweiligen Zielfunktion bei Anpassung einer Lokationsschätzung (SY, Gleichung (2.11)), bzw. des vollen Modells mit einer periodischer Funktion der Periode p (SE(p), Gleichung (2.12)) bezeichnen. Dabei ist ρ eine Abstandsfunktion wie in Abschnitt 2.4.2. Der Vektor `i` besteht im Falle ungewichteter Regression aus Einsen (vgl. Abschnitt 2.1.2), im Falle gewichteter Regression aus den inversen Messfehlern (vgl. Abschnitt 2.5).
Der Wert $\widehat{\sigma}_\beta(p)$ wird durch eine Initialschätzung gewonnen, bei der die periodische Funktion mit einer robusten skalenäquivarianten Regressionstechnik angepasst und ein robuster Skalenschätzer der Residuen $\widehat{\sigma}_\beta(p)$ berechnet wird (vgl. Abschnitt 2.4.2). Zur Berechnung des Bestimmtheitsmaßes empfehlen Maronna, Martin und Yohai (2006, S. 171), dem Skalenschätzer $\widehat{\sigma}_\mu$ des Lokationsmodells ebenfalls diesen Wert $\widehat{\sigma}_\beta(p)$ zuzuordnen. Ein Nachteil ist, dass SY damit über $\widehat{\sigma}_\beta(p)$ von der jeweiligen Testperiode p abhängt und nicht global berechnet werden kann. Nur dieses Vorgehen stellt jedoch sicher, dass das volle Modell $y = X(p)\beta(p) + Y_w, Y_{w;1}, \ldots, Y_{w;n} \underset{u.i.v}{\sim} \mathcal{N}(0, \sigma_\beta(p))$, eine Verallgemeinerung des Lokationsmodells $y = \mathtt{i}\mu + Y_w, Y_{w;1}, \ldots, Y_{w;n} \underset{u.i.v}{\sim} \mathcal{N}(0, \sigma_\mu)$ darstellt, womit $\mathrm{SE}(p) \leq \mathrm{SY}$ und damit $\mathrm{Per}(p) \geq 0$ sichergestellt ist.
Außerdem wird im Falle gewichteter Regression (also Ersetzung von y_i, x_i und $\mathtt{i}_i$ durch die mit s_i normierten Werte $\widetilde{y}_i$, $\widetilde{x}_i$, und $\widetilde{\mathtt{i}}_i$, vgl. Abschnitt 2.5) davon ausgegangen, dass im Falle exakter Messfehler $s_i = \sigma_i$ gilt $\sigma_\beta = 1$. Es könnte vernünftig sein, $\widehat{\sigma}_\beta$ nicht zu schätzen, sondern bereits im Vorhinein auf eins zu fixieren.
Es wird also eine Implementierung der Regressionstechnik benötigt, bei der die Skalenschätzung vorgegeben werden kann und nicht im Laufe der Regression iterativ verbessert wird. Diese Anforderung können die mir bekannten Funktionen zur M-Tukey- bzw. M-Huber-Regression in R nicht erfüllen. Dies sind die Funktionen `rlm` (R-Paket `MASS`, Venables und Ripley 2002), `lmrob..M..fit` (R-Paket `robustbase`, Rousseeuw et al. 2012), `iwlsm` (R-Paket `RSiena`, Ripley, Snijders und Preciado 2012) und die Funktionen `robustregBS` und `robustRegH` (beide R-Paket `robustreg`, Johnson 2011). Die M-Regression wird daher für `RobPer` neu implementiert. Dazu wird ein Iteratively-Reweighted-Least-Squares (IRWLS)-

Ansatz verwendet (vgl. Maronna, Martin und Yohai 2006, S.104–105), der auch den soeben genannten R-Funktionen zu Grunde liegt. Ausgehend von einer Initialschätzung $\widehat{\beta}^{(0)}$ wird dabei iterativ eine gewichtete Kleinste-Quadrate-Regression an das Modell $y = X\beta + Y_w$ durchgeführt, deren Gewichte $\sqrt{W\left(\frac{y_i - x_i^\top \beta(j)}{\widehat{\sigma}}\right)}$ für die Schätzung von $\widehat{\beta}(j+1), j \in \mathbb{N}_0$, von der vorigen Anpassung abhängen. Hierbei ist

$$W(\nu) = \begin{cases} \frac{1}{\nu}\frac{\partial\rho(\nu)}{\partial\nu} & \nu \neq 0 \\ \left.\frac{\partial\rho(\nu)}{\partial^2\nu}\right|_{\nu=0} & \nu = 0 \end{cases}, \tag{3.3}$$

wobei ∂ für den Ableitungsoperator steht. Die konkreten Gewichtungsfunktionen W_{MT} und W_{MH} für M-Tukey- und M-Huber-Regression sind in Anhang F gegeben.
Diese Prozedur wird fortgesetzt, bis sich die Residuen der Anpassung von Regressionsschritt zu Regressionsschritt um weniger als einen gewählten Schwellwert unterscheiden. In Versuchen hat sich gezeigt, dass ein Schwellwert von 10^{-3} zu guten Ergebnissen führt.
Da im Falle gewichteter Regression für den additiven Fehler eine Varianz von ungefähr eins erwartet wird (vgl. Abschnitt 2.5), wurde auch die Option implementiert, den Wert $\widehat{\sigma} = \widehat{\sigma}_\beta(p)$ auf eins fest einzustellen anstatt ihn zu schätzen. Soweit nicht anders erwähnt, wird diese Einstellung in den Analysen in Kapitel 5 für gewichtete M-Regression verwendet, da durch diese in Vorversuchen leicht bessere Ergebnisse erzielt werden konnten.
Für den hier verwendeten Algorithmus zur M-Schätzung werden also drei Initialschätzungen benötigt: Eine Skalenschätzung $\widehat{\sigma}_\beta$ (sofern nicht auf eins voreingestellt) und Schätzer $\widehat{\beta}^{(0)}$ und $\widehat{\mu}^{(0)}$ für den Parameter(-vektor) im vollen und im Lokationsmodell. Für den Schätzer $\widehat{\mu}^{(0)}$ wird wie bei Maronna, Martin und Yohai (2006, S. 105) vorgeschlagen der (bei Berücksichtigung der s_i gewichtete) Median verwendet. Zur initialen Anpassung des vollen Modells wird LTS-Regression (vgl. Abschnitt 3.1.4) verwendet. Bei der Anpassung von q verschiedenen periodischen Funktionen (je eine pro Testperiode $p_1, \ldots, p_q$) ist zu erwarten, dass einige Funktionen nur schlecht zu den vorliegenden Daten passen. Die LTS-Regression hat einen höheren Bruchpunkt als die Kleinste-Beträge-Regression und ist damit besser für solche Situationen und damit zur initialen Schätzung geeignet.
Für M-Huber-Regression konvergiert IRWLS gegen die eindeutige Lösung des Minimierungsproblems aus Gleichung (3.1) bzw. (3.2) (vgl. Maronna, Martin und Yohai 2006, S. 328–329), für M-Tukey-Regression gegen ein lokales Optimum. Dies kann bei M-Tukey-Regression dazu führen, dass für das Lokationsmodell eine im Sinne der Zielfunktion bessere Anpassung erzielt wird als für das volle Modell, obwohl letzteres das Lokationsmodell enthält. Die Konsequenz ist $\mathrm{SY} < \mathrm{SE}(p)$ und damit ein negatives Bestimmtheitsmaß bzw. ein negativer Periodogrammbalken. Um zu niedrige oder gar negative Periodogrammbalken zu verhindern, wird die Initialschätzung $\widehat{\beta}^{(0)}$ des vollen Modells mittels der Funktion `genoud` (R-Paket `rgenoud` von Mebane Jr. und Sekhon (2011), vgl. auch Abschnitt 3.1.5 dieser Arbeit) optimiert. Es konnte im Rahmen dieser Arbeit beobachtet werden, dass das Ergebnis von `genoud` durch anschließende IRWLS-Anwendung noch leicht verbessert werden kann. Daher wird auch bei M-Tukey-Regression IRWLS angewendet, mit der durch `genoud` optimierten Schätzung als Initialschätzung.

3.1.2. S-Regression

Zur S-Regression sind mir die R-Pakete `robeth` (Marazzi 2011, Funktion `hysest`), `FRB` (Roelant, Van Aelst und Willems 2011, Funktion `Sest_multireg`), `robustbase` (Rousseeuw et al. 2012, Funktion `lmrob.S`) und `robust` (Wang et al. 2012, Funktion `lmRob(...,estim=Initial)`) sowie die in Salibian-Barrera und Yohai (2006) publizierte R-Funktion `fast.s` bekannt. Letztere wurde in leicht modifizierter Form `FastS` zur Durchführung der S-Regression verwendet. Die Funktion basiert auf dem Fast-S-Algorithmus, der im Folgenden zusammen mit den wichtigsten Modifikationen in `FastS` erläutert wird. Eine Auflistung aller Veränderungen der Ursprungsfunktion ist in Anhang F zu finden.

Der Fast-S-Algorithmus basiert auf der Idee, aus einer großen Menge von Kandidaten für den Regressionsparameter β den mit dem kleinsten Zielfunktionswert $\zeta_S(y - X\widehat{\beta})$ (vgl. Abschnitt 2.4.3) zu wählen. Ein Großteil dieser Kandidaten entsteht durch Anpassung des Modells an eine Teilstichprobe der Größe m in allgemeiner Lage, wobei m die Dimension des Modells ist. Diese Kandidaten werden anschließend iterativ lokal bezüglich $\zeta_S(y - X\widehat{\beta})$ optimiert, allerdings nicht bis zum Erreichen des lokalen Minimums, sondern nur wenige Iterationen. Von diesen "leicht verbesserten" Kandidaten werden dann nur die besten komplett lokal optimiert und schließlich der Kandidat mit dem kleinsten Zielfunktionswert als Anpassungsergebnis gewählt. Dieses Vorgehen, nur einige statt aller Kandidaten bis zum lokalen Minimum zu optimieren, verkürzt die Rechenzeit der Funktion.

Zur Gewinnung eines Kandidaten wird im Originalalgorithmus eine Teilstichprobe der Größe m gezogen und geprüft, ob die zugehörige $m \times m$-Designmatrix vollen Rang hat. In diesem Fall kann mit dieser Teilstichprobe ein Kandidat ermittelt werden, anderenfalls wird eine neue Teilstichprobe aus allen vorliegenden Datenpunkten gezogen und auf allgemeine Lage geprüft. In `fast.s` wird dieser Schritt durchgeführt, bis eine Stichprobe in allgemeiner Lage gefunden werden konnte. In der modifizierten Version `FastS` gibt die Funktion nach 100 erfolglosen Versuchen `NA` aus. In `RobPer` wird dann der entsprechende Periodogrammbalken auf `NA` gesetzt und eine Warnmeldung ausgegeben.

Ist die anzupassende Funktion eine Stufenfunktion, enthält eine minimale Stichprobe in allgemeiner Lage genau eine Beobachtung je Stufe. Um die Auffindung passender Stichproben sicherzustellen und zu beschleunigen, gibt es in der modifizierten Funktion daher einen Mechanismus, der Stufenfunktionen erkennt und eine entsprechende Stichprobe zieht.

Zusätzlich zu den so gewonnenen Kandidaten für den Regressionsparameter β erlaubt `FastS` auch noch die Eingabe eines zusätzlichen Kandidaten, der wie die anderen behandelt wird. In `RobPer` ist $\widehat{\beta}_\mu$ dieser zusätzliche Kandidat, wobei

$$\widehat{\beta}_\mu = \begin{cases} (\hat{\mu}, \ldots, \hat{\mu})^\mathsf{T} \in \mathbb{R}^m & g \text{ Stufen- oder Splinefunktion} \\ (\hat{\mu}, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^m & g \text{ Fouriersumme vom Grad } k, k = 1, 2, 3 \end{cases} \tag{3.4}$$

die Darstellung des Lokationsschätzers im vollen Modell ist. Wenn β_μ als mögliche Lösung für das volle Modell mitberücksichtigt wird, ist sichergestellt, dass die gefundene Lösung für

das volle Modell mindestens so gut ist wie die des Lokationsmodells und damit $\mathrm{SE}(p) \leq \mathrm{SY}$ gilt.
Der Hauptgrund, `fast.s` (modifiziert) in `RobPer` zur S-Regression einzusetzen, ist die Möglichkeit, in einem lesbaren und gut dokumentierten R-Code einen Zusatzkandidaten $\widehat{\beta}_\mu$ und einen Mechanismus zur leichteren Kandidatenfindung bei Stufenmodellen einzubauen. Eine Anpassung der anderen oben genannten Funktionen an die Anforderungen des hier vorliegenden Kontextes scheint ohne fundierte Kenntnisse in den Programmiersprachen C oder Fortran nicht möglich.
In `FastS` kann über einen Eingabeparameter eingestellt werden, dass der eingegeben Designmatrix eine zusätzliche Interceptspalte hinzugefügt werden soll. Da diese Funktionalität in `RobPer` nicht benötigt wird, wird sie dort ausgeschaltet. Die Anzahl der Kandidaten, die Iterationenanzahl, die bei allen Kandidaten zur Verbesserung durchgeführt wird, die Anzahl der Kandidaten, die anschließend komplett optimiert werden, die Variablen c und δ (vgl. Abschnitt 2.4.3) sowie ein Startwert für die Zufallsgeneratoren zur Reproduzierbarkeit der Regressionsergebnisse können in `RobPer` durch den Nutzer verändert werden. Standardmäßig sind die meisten Werte auf die in `fast.s` verwendeten und in Salibian-Barrera und Yohai (2006) empfohlenen Werte eingestellt. Eine Ausnahme bildet der c-Wert, der gemäß Rousseeuw und Yohai (1984) auf 1.547 gesetzt wurde, um einen asymptotischen Bruchpunkt von 0,5 zu erhalten, und die Anzahl der Kandidaten. Bei Variation dieses Parameters stellte sich in Vorversuchen heraus, dass im Falle ungewichteter Regression eine Kandidatenzahl von 50 stets ähnlich gute Ergebnisse brachte wie eine Kandidatenzahl von 200, nur in geringerer Zeit. Im Falle gewichteter Regression konnte die Detektionsfähigkeit der Methoden durch Verwendung von 200 Kandidaten noch gesteigert werden. Außer bei Anpassung der Einfachstufenfunktion brachte eine Erhöhung auf 600 Kandidaten keine zusätzlichen Vorteile. Im Falle von gewichteter Anpassung der Einfachstufenfunktion konnten mit 600 Kandidaten noch bessere Ergebnisse erzielt werden, eine weitere Erhöhung der Kandidatenanzahl auf 1000 erbrachte keine Verbesserung der Ergebnisse.

3.1.3. τ-Regression

Zur recheneffizienten Durchführung der τ-Regression führen Salibian-Barrera, Willems und Zamar (2008) den Fast-τ-Algorithmus ein und liefern eine R-Funktion `FastTau`, die diesen verwendet. Diese Funktion wird leicht modifiziert in `RobPer` verwendet.
Der zugrunde liegende Algorithmus folgt dem gleichen Prinzip wie der Fast-S-Algorithmus, nämlich geringe lokale Optimierung vieler Parameterkandidaten und komplette lokale Optimierung der dabei entstehenden besten Kandidaten (vgl. Abschnitt 3.1.2). Dabei werden die Kandidaten auch hier aus minimalen Teilstichproben in allgemeiner Lage gewonnen und wie bei der S-Regression wurde die Funktion auch hier so modifiziert, dass in dem hier beschriebenen Zusammenhang effizienter minimale Teilstichproben zur Gewinnung von Parameterkandidaten gefunden werden können.
In der Originalfunktion enthält eine weitere Teilstichprobe die $\left\lfloor \frac{n}{2} \right\rfloor$ Beobachtungen, die am wenigsten vom Median der gesamten Stichprobe abweichen. In der in `RobPer` verwendeten modifizierten Version des Algorithmus' wird der so gewonnene Kandidat bei Anpassung des

vollen Modells durch $\widehat{\beta}_\mu$ aus Gleichung (3.4) ersetzt, damit $\mathrm{SE}(p) \leq \mathrm{SY}$ gilt. Eine Aufstellung aller Veränderungen des ursprünglich veröffentlichen R-Codes sind im Anhang F zu finden.
In Vorstudien wurde der Einfluss der Kandidatenanzahl N auf das Detektionsergebnis der Periodogrammmethode untersucht. Es ergab sich, dass eine Anzahl $N = 100$ gegenüber $N = 300$ zu großer Rechenzeitersparnis führt, aber kaum zu nennenswert schlechteren Detektionsergebnissen. Weiterhin bietet `FastTau` die Option, die Zielfunktion nicht exakt zu berechnen, sondern lediglich zu approximieren. Da diese Option in Vorversuchen zu keiner Rechenzeitersparnis oder besseren Ergebnissen geführt hat, eine exakte Berechnung der Zielfunktion aber besser zu rechtfertigen ist, wird die Option zur Approximation im Folgenden nicht angewendet. Alle anderen Voreinstellungen werden übernommen wie von Salibian-Barrera, Willems und Zamar (2008) empfohlen.

3.1.4. LTS-Regression

Damit die LTS-Anpassung des vollen Modells die LTS-Anpassung des Lokationsmodells einschließt, soll für ζ_{LTS} stets das gleiche Trimming $h(m)$ (vergleiche Gleichung (2.31), Seite 25) verwendet werden, wobei m die Anzahl der Regressoren des Modells ist. Bei Modellen mit Stufenfunktionen kann es im Zusammenhang mit periodischen Lücken im Messzeitmuster leicht dazu kommen, dass für eine Stufe keine Beobachtungen vorliegen. In diesem Fall wird der entsprechende Regressor aus dem Modell entfernt. Daher sind m und damit ζ_{LTS} im Stufenmodell von der Testperiode p abhängig und SY kann nicht für alle p gemeinsam bestimmt werden.
Zur Durchführung von LTS-Regression sind mir die R-Pakete `robeth` (Marazzi 2011, Funktion `hyltse`), `MASS` (Venables und Ripley 2002, Funktion `ltsreg`) und `robustbase` (Rousseeuw et al. 2012, Funktion `ltsReg`) bekannt. Da der Trimmingparameter h bei `hyltse` unveränderbar auf $\left\lfloor \frac{n}{2} \right\rfloor + 1$ eingestellt ist, eignet sich diese Funktion nicht zur Verwendung in `RobPer`. Beim Vergleich der Funktionen `ltsreg` und `ltsReg` konnte in Vorversuchen festgestellt werden, dass `ltsreg` schneller (etwas mehr als doppelt so schnell) zu einem Ergebnis kommt, `ltsReg` die Zielfunktion aber besser minimiert. Daher wird in `RobPer` die LTS-Regression mit der R-Funktion `ltsReg` durchgeführt. Dies geschieht sowohl wenn die LTS-Schätzung als Startschätzung für eine M-Regression verwendet wird (vgl. Abschnitt 3.1.1), als auch wenn die LTS-Regression die Hauptregression ist.
In Vorversuchen konnte auch festgestellt werden, dass sich der im vollen Modell durch `ltsReg` erreichte Wert der Zielfunktion manchmal noch leicht verringern lässt. Daher besteht in `RobPer` die Möglichkeit, der LTS-Anpassung eine Optimierung mit der Funktion `genoud` (Paket `rgenoud`, Abschnitt 3.1.5) nachzuschalten, sofern LTS die Hauptregression ist. In Vorversuchen hat sich gezeigt, dass dieses Vorgehen hauptsächlich bei Anpassung eines Stufenmodells Vorteile bringt, daher wird `genoud` in den Simulationen und Analysen in Kapitel 5 bei diesem Modell eingesetzt.
`ltsReg` zeigt zudem die Eigenschaft, in seltenen Fällen wegen eines zufälligen Schrittes in der Anpassung abzubrechen, der nicht durch `set.seed` (R-Paket `base`, vgl. R Core Team 2013), beeinflusst werden kann. Ein wiederholter Aufruf der Funktion führt dann meist zum Erfolg. `RobPer` versucht bis zu dreimal, die LTS-Regression durchzuführen. Nach dem

dritten misslungenen Versuch wird der Periodogrammbalken entweder auf `NA` gesetzt oder eine Kleinste-Beträge-Schätzung berechnet. Letzteres geschieht, wenn der Schätzung noch eine Optimierung mit `genoud` oder eine Anpassung mittels M-Regression folgt.

3.1.5. Optimierung mit `genoud`

Die bereits genannte Funktion `genoud` ist Teil des R-Paketes `rgenoud` von Mebane Jr. und Sekhon (2011) und kombiniert evolutionäre Algorithmen und ableitungsbasierte Verfahren zur Lösung von Optimierungsproblemen. In einem evolutionären Algorithmus wird aus einer Menge von Lösungskandidaten (eine "Generation") eine neue Menge von Lösungskandidaten gewonnen. In dieser neuen Generation sollen sich Kandidaten befinden, die die Zielfunktion besser optimieren als die Kandidaten der letzten Generation. Zur Generierung der neuen Generation werden die Kandidaten der letzten Generation nach gewissen Regeln zufällig zu neuen Kandidaten kombiniert. Zusätzlich zu diesem bei evolutionären Algorithmen üblichen Vorgehen wird bei `genoud` der jeweils beste Kandidat einer Generation lokal unter Verwendung einer ableitungsbasierten Methode optimiert. Für weitere Details vergleiche Mebane Jr. und Sekhon (2011).
Die Funktion `genoud` kommt in `RobPer` zum Einsatz, um ζ_{MT} und ζ_{LTS} zu optimieren. In beiden Fällen ist ohne Verwendung einer zusätzlichen Optimierung das Problem aufgetreten, dass das volle Modell schlechter an die Lichtkurve angepasst wird als das Lokationsmodell. Die Folge ist ein negatives Bestimmtheitsmaß. Da das Lokationsmodell aber ein Teilmodell des vollen Modells ist, muss stets $\mathrm{SE}(p) \leq \mathrm{SY}$ gelten. Das Problem der negativen Bestimmtheitsmaße konnte gelöst werden, indem die Funktion `genoud` neben einer Initialschätzung $\widehat{\beta}^{(0)}$ für β auch eine Lokationsschätzung $\widehat{\beta}_\mu$ (vgl. Gleichung (3.4)) in die erste Generation aufnimmt. Zusätzlich wuchsen in Vorversuchen durch zusätzliche Anwendung von `genoud` nicht nur bis dahin negative Periodogrammbalken, sondern auch andere, für deren Testperioden das volle Modell nun offensichtlich besser angepasst werden konnte (vgl. Abbildung 3.1).
Die Funktion `genoud` hat über 45 Parameter. In `RobPer` wird abhängig von der verwendeten Regressionstechnik ζ_{MT} bzw. ζ_{LTS} als zu minimierende Funktion gewählt. Als voreingestellte Lösungskandidaten der ersten Generation werden neben den von der Funktion zufällig selbst generierten eine Initialschätzung $\widehat{\beta}^{(0)}$ und der Schätzer für das Lokationsmodell $\widehat{\beta}_\mu$ verwendet. Parameter, die die Genauigkeit der Lösung beeinflussen (Größe einer Generation, Höhe der Konvergenzschranke und Anzahl der Generationen, die maximal oder ohne weitere Optimierung der Zielfunktion generiert werden bevor die Optimierung abgebrochen wird), werden der Funktion `RobPer` übergeben.
In Vorstudien hat sich gezeigt, dass eine Generationengröße (`pop.size`) von 100 in einigen Fällen Vorteile gegenüber einer von 50 hat. Als Konvergenzschranke (`tol`) hat sich 10^{-3} bewährt. Als maximale Generationenanzahl, die ohne zusätzliche Optimierung der Zielfunktion generiert werden (`wait.generations`), genügen bei LTS-Regression fünf Generationen, bei M-Tukey-Regression ließ sich eine Verbesserung der Detektionen mit zehn Generationen erzielen. Als maximale Generationenzahl (`max.generations`) hat sich 50 als ausreichend groß erwiesen. Die übrigen Parameter von `genoud` werden in ihrer Voreinstellung belassen.

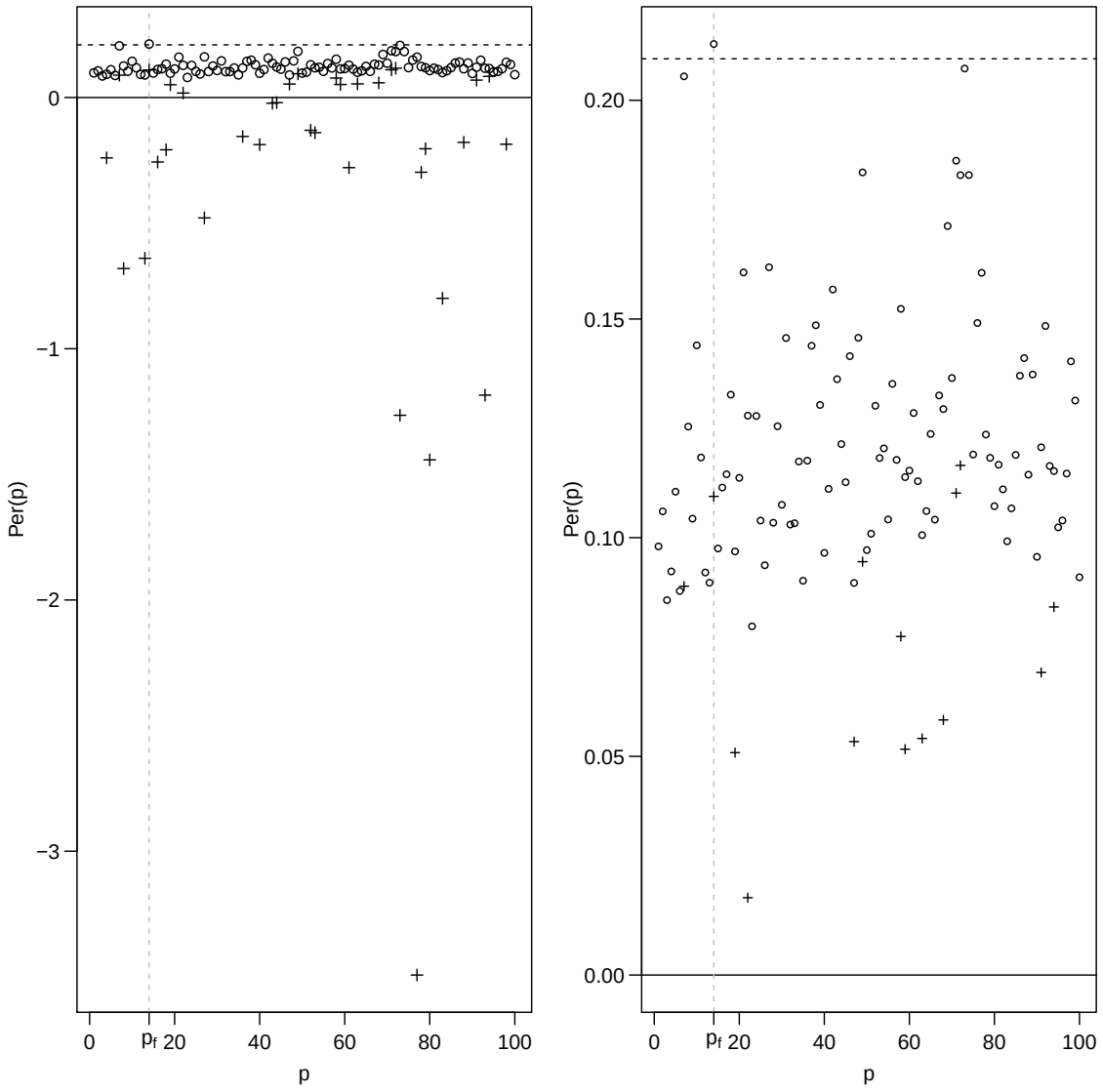


Abbildung 3.1.: Illustration des Einsatzes der Funktion `genoud` zur Vermeidung negativer Periodogrammbalken: (a) Periodogramme (LTS-Regression, Stufenmodell) einer Lichtkurve mit Fluktuationsperiode $p_f = 14$, ohne (+) und mit (∘) zusätzlicher Optimierung. Fehlende +-Balken kommen vor, wenn die R-Funktion bei dreimaligem Ausführen kein Ergebnis erzielt. Viele +-Periodogrammbalken sind negativ, obwohl sie als Bestimmtheitsmaße im Intervall [0,1] liegen müssten. (b) Ausschnittvergrößerung des positiven Wertebereichs aus (a).

3.2. Lichtkurvengenerator

Der hier entwickelte Lichtkurvengenerator `tsgen` eignet sich zur künstlichen Erstellung von Lichtkurvendaten zu Simulationszwecken. Er ruft nacheinander die im Folgenden beschriebenen Unterfunktionen auf, um zufällige Messzeiten t_i zu ziehen (`sampler`), Werte der periodischen Fluktuation $y_{f;i}$ zu berechnen (`signalgen`), eine Rauschkomponente zu addieren (`lc_noise`) und die resultierende Lichtkurve wahlweise zu stören (`disturber`). Die Rauschkomponente kann sich aus von den Messfehlern s_i abhängigen und unabhängigen Summanden zusammensetzen. Dies stellt, wie auch das Stören der fertigen Lichtkurve mit `disturber`, eine Erweiterung des Datenmodells aus Abschnitt 2.1 dar. Die Ausgabe ist eine Lichtkurve $(t_i, y_i, s_i)_{i=1,\dots,n}$. Im Folgenden wird die Funktionsweise der Unterfunktionen erläutert. Dabei wird auf die Verwendung der im R-Programm genutzten Bezeichnungen der einzustellenden Parameter verzichtet. Diese sind im Anhang G aufgelistet.

3.2.1. Messzeitgenerierung

Mit Hilfe der Funktion `sampler` werden die Messzeiten $t_1, \dots, t_n$ aus einer periodischen Verteilung $\mathcal{D}(p_s)$ (vgl. Formel (2.1), Seite 8) gezogen. Eingabeparameter sind hierbei die Samplingperiode p_s, die Anzahl der Samplingzyklen n_s, der Stichprobenumfang n und die Verteilung $\mathcal{D}(p_s)$ der ungeordneten Messzeiten. Hier sind vier Verteilungen möglich: Bei äquidistantem Sampling sind die Messzeiten gleichabständig mit $t_i = i\frac{p_s n_s}{n}$ und $t_i - t_{i-1} = \frac{p_s n_s}{n}$. Bei einer Rechteckverteilung stammen die ungeordneten Messzeiten aus einer Rechteckverteilung aus dem Intervall $[0, n_s p_s]$, das Messzeitmuster ist ungleichmäßig, aber nicht periodisch.
Zur Generierung periodischer, ungleichmäßig abständiger Messzeitmuster stehen die Sinus- und eine unsymmetrische Dreiecksverteilung zur Verfügung. Bei der Sinusverteilung werden die Phasen $\varphi_i^\star = \varphi_{p_s}(t_i^\star)$ der ungeordneten Messzeiten $t_i^\star$ unabhängig identisch gemäß der Dichte

$$d_{sin}(x) = \sin(2\pi x) + 1 \tag{3.5}$$

generiert, bei der Dreiecksverteilung gemäß der Dichte

$$d_{trian}(x) = \begin{cases} 3x & 0 \le x \le \frac{2}{3} \\ 6 - 6x & \frac{2}{3} < x \le 1 \end{cases}. \tag{3.6}$$

Bei beiden Einstellungen stammen die Zyklen $z_i^\star = z_{p_s}(t_i^\star)$ der ungeordneten Messzeiten $t_i^\star$ aus einer Gleichverteilung auf der Menge $\{1, \dots, n_s\}$ und die ungeordneten Messzeiten lassen sich durch

$$t_i^\star = \varphi_i^\star + (z_i^\star - 1)p_s$$

berechnen. Diese getrennte Randomisierung von Phase und Zyklus zur Generierung ungleichmäßiger Beobachtungen mit periodischer Verteilung erfolgt nach dem Vorbild von

Hall und Yin (2003). Die Sinusverteilung passt zur monatlichen Messzeitpunktverteilung der gegebenen Lichtkurven im betrachteten Anwendungsfall (vgl. Abbildung 2.3, Seite 7). Die Einstellung `trian` wird als alternatives periodisches Sampling betrachtet. Zur Ziehung einer Zufallszahl mit einer der oben genannten Dichten wird die entsprechende inverse Verteilungsfunktion F^{-1} auf eine rechteckverteilte Zufallszahl $u \sim \mathcal{U}_{[0,1]}$ angewendet, wobei

$$F^{-1}_{trian}(u) = \begin{cases} \sqrt{\frac{2u}{3}} & 0 \le u \le \frac{2}{3} \\ 1 - \sqrt{\frac{1-u}{3}} & \frac{2}{3} < u \le 1 \end{cases}$$

ist. Zur Approximation von F^{-1}_{sin} wird die Funktion `BBsolve` aus dem R-Paket `BB` (Varadhan und Gilbert 2009) verwendet.
Die Ausgabe von `sampler` ist ein geordneter Messzeitenvektor mit Werten im Intervall $[0, n_s p_s]$.

3.2.2. Generierung der periodischen Fluktuation

Zur Generierung des Signals

$$y_{f;i} = f\left(\frac{t_i}{p_f}\right)$$

(vgl. Gleichung 2.3, Seite 8) wird die Funktion `signalgen` verwendet. Eingabeparameter sind hierbei Messzeiten $t_1, \ldots, t_n$, zum Beispiel durch die Funktion `sampler` gewonnen, eine Fluktuationsperiode p_f und die Funktion f. Für f kann hierbei eine konstant auf den Wert 0 abbildende Funktion, eine Sinusfunktion f_{sin}, eine Dreiecksfunktion f_{trian} oder eine Peakfunktion f_{peak} gewählt werden. Dabei ist

$$f_{sin}(t) = \sin\left(\frac{2\pi t}{p_f}\right), \tag{3.7}$$

$$f_{trian}(t) = \begin{cases} 3\varphi_1(t) & 0 \le \varphi_1(t) \le \frac{2}{3} \\ 6 - 6\varphi_1(t) & \frac{2}{3} < \varphi_1(t) \le 1 \end{cases}, \tag{3.8}$$

$$f_{peak:p_f}(t) = \begin{cases} 9\exp\left(-3p_f^2\left(\varphi_1(t) - \frac{2}{3}\right)^2\right) & 0 \le \varphi_1(t) \le \frac{2}{3} \\ 9\exp\left(-12p_f^2\left(\varphi_1(t) - \frac{2}{3}\right)^2\right) & \frac{2}{3} < \varphi_1(t) \le 1 \end{cases}, \tag{3.9}$$

wobei $\varphi_1(t)$ die nach Periode 1 umgeklappte Messzeit ist, also $\varphi_1(t) = t - \lfloor t \rfloor$ (vgl. Gleichung 2.5, Seite 9). Die Funktion f_{sin} wird häufig verwendet (Schwarzenberg-Czerny 1989), zum Beispiel auch bei klimatologischen Daten (Mann und Lees 1996). Die Funktion f_{trian} wurde zunächst als beliebige zusätzliche asymmetrische periodische Funktion zu Versuchszwecken in den Generator aufgenommen. Es zeigt sich aber, dass ähnlich geformte Funktionen durchaus realistisch sind (vgl. Lichtkurve von CoRoT ID 0105288363 bei Chadid et al. 2011). Die Funktion $f_{peak:p_f}$ wird in dieser Arbeit zur Modellierung eines asymmetrischen hohen Ausschlags konstruiert. Sie ist stetig und die Breite des Ausschlags ist $\frac{1}{p_f}$ und für $y_{f;i}$ damit konstant eine Zeiteinheit (vgl. Abb. 3.2). Beide Nachweise sind im Anhang H zu finden.

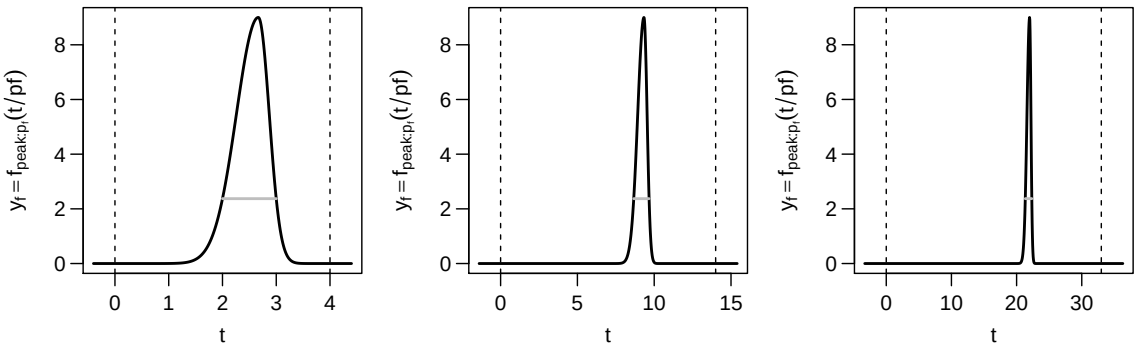


Abbildung 3.2.: Funktion $f_{peak:p_f}(t/p_f)$ in Abhängigkeit von der Messzeit t. Fluktuationsperioden: (a) p_f=4, (b) $p_f = 14$, (c) $p_f = 33$. Die Breite des Ausschlags auf Höhe $9\exp(-4/3)$ (graue Linie) beträgt stets eine Zeiteinheit. Die gestrichelten Linien markieren einen Zyklus.

3.2.3. Hinzufügen der Rauschkomponente und der Messfehler

Mit Hilfe der Funktion `lc_noise` können zu einer bestehenden periodischen Fluktuation y_f additives Rauschen und die dazugehörigen Messfehler generiert werden. Eingabeparameter sind unter anderem Messzeiten $t_1, \ldots, t_n$, zum Beispiel durch die Funktion `sampler` gewonnen, und Werte $y_{f;1}, \ldots, y_{f;n}$ der periodischen Fluktuation, zum Beispiel aus der Ausgabe von `signalgen`. Inspiriert von den beobachteten Verteilungen bei den vorliegenden realen Lichtkurven (vgl. Abbildung 2.2, Seite 7) werden die Standardabweichungen σ_i aus einer $\Gamma(3, 10)$-Verteilung gezogen. Die Funktion gibt als Messfehler s_i keine Schätzung von σ_i, sondern σ_i selbst aus. Dies stellt eine Idealisierung dar.
Bisher wurde angenommen, dass (vgl. Gleichungen (2.2) bis (2.4), Seite 8) gilt:

$$Y_i = Y_{f;i} + Y_{w;i},$$
$$Y_{w;i} \sim \mathcal{N}(0, \sigma_i^2),$$
s_i gegebene Schätzung für σ_i unabhängig von $Y_1, \ldots, Y_n$.

Die hier vorgestellte Funktion `lc_noise` arbeitet mit einem erweiterten Modell, das eine weitere Rauschkomponente y_r zulässt:

$$Y_i = Y_{f;i} + Y_{w;i} + Y_{r;i}.$$

Dabei ist $y_{r;i}$ Realisation einer Rauschkomponente, die nicht von s_i abhängt. Ihr Anteil am gesamten Rauschen wird durch das Verhältnis

$$\psi = \mathrm{var}(y_r)/(\mathrm{var}(y_w) + \mathrm{var}(y_r)) \tag{3.10}$$

gesteuert, welches Eingabeparameter der Funktion ist. Hierbei stehen var() für die empirische Varianz eines Vektors und y_w und y_r für den simulierten Vektor der jeweiligen Rauschkomponente. Letztere kann zum Beispiel unabhängig normalverteiltes Rauschen sein, was durch einen weiteren Eingabeparameter erreicht wird. In diesem Fall unterscheidet

sich das Modell nur von dem bisher verwendeten, wenn bei der Periodogrammberechnung gewichtete Regression verwendet wird, da die s_i nicht mehr die Varianz aller Rauschkomponenten berücksichtigen und ihre Informativität somit herabgesetzt ist. Wenn y_r kein unabhängiges normalverteiltes Rauschen ist, unterscheidet sich das neue Modell auch bei ungewichteter Regression vom bisherigen. Denkbar wäre beispielsweise so genanntes Rotes Rauschen (vgl. Abschnitt 6.2).
Mit einem weiterem Eingabeparameter kann das Verhältnis

$$\eta_1 = \mathrm{var}(y_f)/\mathrm{var}(y_w + y_r) \tag{3.11}$$

kontrolliert werden. Hiermit soll die Relation von Signal und Rauschen kontrolliert werden. Der Nachteil dieses Wertes ist, dass er für beobachtete Lichtkurven nicht bekannt ist und damit nicht gemäß bekannten Datenszenarien realistisch gewählt werden kann. In Vorarbeiten wurde auch mit anderen Parametern experimentiert, so zum Beispiel mit (vgl. Thieler et al. 2013)

$$\eta_2 = \mathrm{var}(y)/\ \overline{s^2}, \tag{3.12}$$

wobei $\overline{s^2}$ für das arithmetische Mittel der quadrierten Messfehler steht. Wegen Unsicherheiten bezüglich der Güte der Schätzungen s_i wird jedoch auch die Zuverlässigkeit der in der Realität messbaren Werte für η_2 in Zweifel gezogen. Damit scheint η_2 dem Parameter η_1, der leichter zu interpretieren ist, nicht überlegen zu sein.

3.2.4. Störung der Lichtkurve

Mit Hilfe der Funktion `disturber` können Störungen in eine eingegebene Lichtkurve eingebaut werden. Hierbei können entweder zufällig gewählte Messfehler s_i oder zeitlich aufeinander folgende Messwerte y_i durch atypische Werte ersetzt werden oder beides zugleich. Eine Störung der Messzeiten t_i ist nicht vorgesehen.
Zusätzlich zur Lichtkurve benötigt die Funktion die Eingabe, wie hoch der Anteil der Messfehler s_i sein soll, die durch Ausreißer ersetzt werden. Bei gewichteter Regression werden die Beobachtungen mit kleinem Messfehler s_i als besonders zuverlässig eingestuft, die anzupassende Funktion wird durch den Punkt (t_i, y_i) oder nahe an ihn heran gezwungen. Daher ist zu erwarten, dass bei Messfehlern Ausreißer nach unten stärker stören als solche nach oben. Deshalb werden gestörte Messfehler durch den niedrigeren Wert $\dot{s} = \frac{1}{2}\min(s_1, \ldots, s_n)$ ersetzt. Für einen Anteil $\alpha \in]0, 1]$ zu störender Messwerte werden $\langle n\alpha \rangle$ zufällig gezogene Messfehler durch $\dot{s}$ ersetzt, wobei $\langle\rangle$ für kaufmännisches Runden steht.
Ein weiterer Eingabeparameter legt fest, ob eine Intervallstörung durchgeführt werden soll. Bei einer Intervallstörung werden die Messwerte y_i der Lichtkurve auf einem Intervall der Länge $3p_s$ (Samplingperiode p_s ebenfalls Eingabeparameter) durch Werte ausgetauscht, die einer eingipfeligen Funktion mit vergleichsweise hohem Maximum folgen. Für eine Intervallstörung wird ein Startpunkt $t_{start} = t_{i^\star}$, $i^\star$ gemäß $\mathcal{U}_{\{i \in \{1,\ldots,n\}: t_i < t_n - 3p_s\}}$ gezogen, und

die Messwerte von drei Samplingzyklen ab t_{start} zu einem erhöhten Ausschlag mit Maximum $6\tilde{y}_{0,9}$ verändert:

$$y_i^{neu} = 6\ \tilde{y}_{0,9}\ \frac{d_{\mathcal{N}(t_{start}+1.5p_s,p_s^2)}(t_i)}{d_{\mathcal{N}(0,p_s^2)}(0)} \quad \forall\ i : t_i \in [t_{start}, t_{start} + 3p_s],$$

wobei $\tilde{y}_{0.9}$ das 0,9-Quantil der Messwerte $y_1, \ldots, y_n$ ist und $d_{\mathcal{N}(\theta_1,\theta_2)}$ die Dichte der $\mathcal{N}(\theta_1, \theta_2)$-Verteilung. Die Intervallstörung ist durch die Beobachtung realer Quellen motiviert (vgl. beispielsweise Messungen zu Mrk 421 in Abbildung 2.1 um 52 000 MJD) und es ist bekannt, dass die Gründe für solche Störungen nicht mit einer periodischen Funktion modellierbar sind (Andrew et al. 1969).
Liegt der Anteil der zu störenden Messfehler bei 0 und wird keine Intervallstörung gewählt, gibt die Funktion `disturber` die eingegebene Lichtkurve unverändert wieder aus.

Zusammenfassung

Kapitel 3 behandelt die Implementierung der zuvor vorgestellten Periodogrammmethoden und eines Lichtkurvengenerators im R-Paket `RobPer`. Mehrere Regressionstechniken werden dazu selbst implementiert, da die vorhandenen R-Funktionen nicht alle Anforderungen erfüllen. So wird für die gewichtete Regression eine M-Regression implementiert, deren Varianzschätzung auf eins fixiert werden kann. Ein zusätzlicher Optimierer wird verwendet, um das M-Tukey- und die LTS- Regressionsergebnis zu optimieren. Dieser Optimierer und die Algorithmen für die S- und die τ-Regression arbeiten mit Mengen von Parameterkandidaten. Es wird sichergestellt, dass der Lokationsschätzer auch einer der Kandidaten ist. So kann gewährleistet werden, dass die Anpassung des vollen Modells zu einer mindestens so guten Anpassung führt wie das Lokationsmodell, womit ein nicht negatives Bestimmtheitsmaß garantiert ist.
Mit Hilfe des Lichtkurvengenerators können Messzeiten, Messwerte und Messfehler generiert werden. Die Messzeiten folgen einer periodischen Verteilung. Die Messwerte können eine periodische Fluktuation enthalten. Die Messfehler entsprechen exakt den Standardabweichungen der in den Messwerten enthaltenen Rauschenkomponente. Es kann eine zweite Rauschkomponente hinzugeschaltet werden, um die Informativität der Messfehler zu verringern. Es ist möglich, die Lichtkurvendaten durch das Einfügen von Intervallstörungen oder Ersetzen von Messfehlern durch Ausreißer zu stören.

4. Simulationsstudie

In diesem Kapitel werden die in Kapitel 2 entwickelten Periodogramm- und Detektionsmethoden in der in Kapitel 3 vorgestellten Implementierung auf künstliche Lichtkurvendaten angewendet, um die Eigenschaften der Methoden zu untersuchen. Hierbei geht es vor allem um die Frage, welche Methoden gut zur Detektion periodischer Fluktuationen geeignet sind und zugleich in Abwesenheit einer periodischen Fluktuation das Niveau einhalten. Wie in den vorigen Kapiteln erläutert, werden dabei die folgenden Grundeinstellungen verwendet:

- Die periodische Stufenfunktion hat zehn Stufen pro Zyklus (vgl. Abschnitt 2.3.2).
- Die für M-Regression bzw. genoud-Optimierung benötigte Konvergenzschranke ist `tol` = 10^{-3} (vgl. Abschnitte 3.1.1 und 3.1.5).
- Die Einstellungen `pop.size`=100 und `max.generations`=5 werden für die Optimierung mit `genoud` verwendet, für LTS-Regression ist dabei `wait.generations`=5, für M-Tukey-Regression `wait.generations`=10 (vgl. Abschnitt 3.1.5).
- Im Falle der LTS-Regression wird das Regressionsergebnis mit `genoud` nachträglich verbessert (vgl. Abschnitt 3.1.4).
- Im Fast-S-Algorithmus wird im ungewichteten Fall mit 50 Kandidaten gearbeitet, bei gewichteter Anpassung der einfachen Stufenfunktion mit 600 Kandidaten und in allen anderen gewichteten Fällen mit 200 Kandidaten.
- Im Fast-τ-Algorithmus wird mit 100 Kandidaten gearbeitet und das Zielkriterium exakt berechnet (vgl. Abschnitt 3.1.3).

Damit weichen die Einstellungen teilweise von den Voreinstellungen von `RobPer` (vgl. Tabelle F.1 im Anhang) ab. Die Voreinstellungen wurden als Kompromiss zwischen Detektionsgüte und Rechenzeit gewählt, während in den hier vorgestellten Simulationsstudien längere Rechenzeiten zugunsten besserer Detektionsgüte in Kauf genommen wurden.

4.1. Studienaufbau

In der Simulationsstudie werden die Periodogrammmethoden auf simulierte Daten angewendet. Das Vorgehen ist hierbei immer identisch: In jedem Durchlauf i, $i = 1, \ldots, 1000$ mit vorgegebenen Einstellungen erfolgt

1. Generierung einer Lichtkurve mit den vorgegebenen Einstellungen.

2. Berechnung eines Periodogramms mit den vorgegebenen Einstellungen für die Testperioden $p_1 = 1, p_2 = 2, \ldots, p_{100} = 100$.

3. Bei Vorliegen einer periodischen Fluktuation der Periode p_f: Definition von $p^\star$ als die Testperiode, die auf der logarithmischen Skala am nächsten an p_f liegt. Die Perioden werden logarithmisch betrachtet, um einen relativen Abstand zu erhalten (vgl. auch Abbildung 2.14, Seite 33).

4. Bei Einstellungen ohne periodische Fluktuation: Definition von $p^\star$ als die Testperiode mit dem höchsten Periodogrammbalken.

5. Anpassung einer Betaverteilung $\mathcal{B}(\widehat{\theta}_1, \widehat{\theta}_2)$ an das Periodogramm $\mathrm{Per}(p_1), \ldots, \mathrm{Per}(p_{100})$ mittels Cramér-von-Mises-Distanz-Minimierung.

6. Für Testperiode $p^\star$: Bestimmung des niedrigsten Niveaus α_i, zu dem der Periodogrammbalken $\mathrm{Per}(p^\star)$ detektiert wird.

Für $\alpha \in \{\alpha_1, \ldots, \alpha_{1000}\}$ wird der Anteil $H(\alpha)$ der wie oben generierten Periodogramme bestimmt, bei dem $\mathrm{Per}(p^\star)$ zum Niveau α detektiert wird. Für $\alpha \in [0,1] \backslash \{\alpha_1, \ldots, \alpha_{1000}\}$ wird dieser Anteil linear interpoliert. Die entstehende Funktion $H : [0,1] \to [0,1]$ (interpolierter) Detektionsanteile wird Detektionskurve genannt. Beispiele für Detektionskurven werden im weiteren Verlauf dieses Kapitels diskutiert und sind in den Abbildungen 4.3 (Seite 57) und 4.4 (Seite 61) zu finden.
Im Folgenden werden die verwendeten Lichtkurven und Periodogrammmethoden eingehender erläutert. Anschließend werden die zur Beurteilung der Detektionskurven verwendeten Gütemaße vorgestellt, die bei der Auswertung der Simulationsstudie Anwendung finden.

4.1.1. Lichtkurven

Bei den Simulationsstudien werden Lichtkurven verwendet, deren Messzeiten, sofern nicht anders erwähnt, gemäß einer periodischen Verteilung mit Sinus- oder Dreiecksdichte („Sinussampling" oder „Dreiecksampling", vgl. Definitionen (3.5) und (3.6) in Abschnitt 3.2.1) mit Samplingperiode $p_s = 25$ generiert werden. Diese Zahl ist willkürlich gewählt, hat aber die gleiche Größenordnung wie beispielsweise der Mondzyklus. Es werden $n_c = 30$ Zyklen beobachtet und insgesamt 750 Beobachtungen generiert. Dies entspricht einer durchschnittlichen Stichprobengröße von 25 pro Zyklus. Nach Einschätzung der Kooperationspartner aus der Astroteilchenphysik (vgl. Thieler et al. 2013) ist die Gewinnung von 25 Messungen in einem beobachtbaren Samplingzyklus ein realistisches Ziel. Die Messwerte y_i enthalten eine normalverteilte weiße Rauschkomponente, die Messfehler s_i entsprechen genau den zugehörigen Standardabweichungen und sind ausreißerfrei. Die Messwerte enthalten entweder keine periodische Fluktuation (zur Überprüfung der Einhaltung des Niveaus) oder eine Sinus- oder Peakfluktuation (vgl. Definitionen (3.7) und (3.9) in Abschnitt 3.2.2) mit einer Fluktuationsperiode von $p_f = 14$ oder $p_f = 33$. Es werden sowohl störungsfreie Lichtkurven als auch solche mit Intervallstörung (vgl. Abschnitt 3.2.4) betrachtet. Die Kombination all dieser Möglichkeiten führt zu 20 verschiedenen Einstellungen zur Generie-

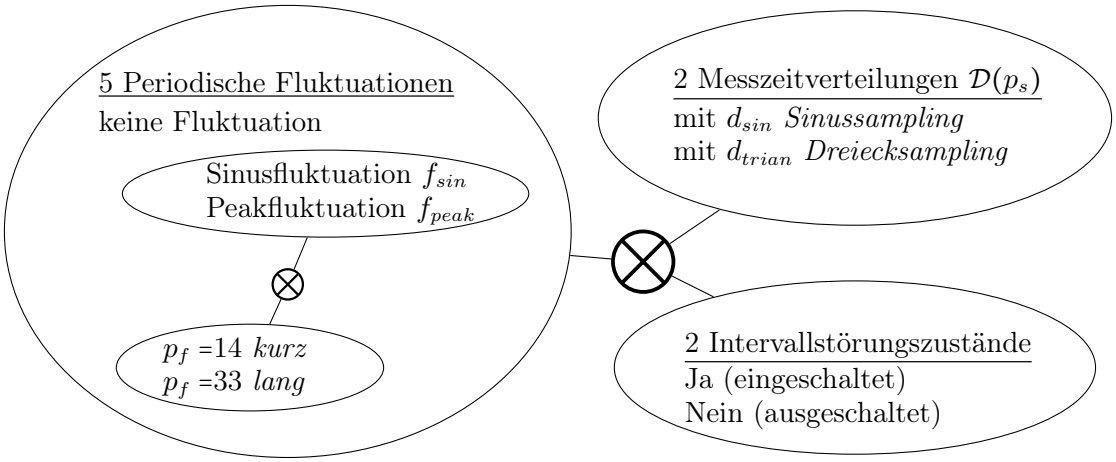


Abbildung 4.1.: Die Kombination mehrerer Parameter führt zu insgesamt 20 unterschiedlichen Lichtkurventypen bzw. Datenszenarien. Kursiv gedruckte Worte stehen für alternative Bezeichnungen.

rung der Lichtkurven in Schritt 1. Abbildung 4.1 zeigt eine Übersicht der verschiedenen Kombinationsmöglichkeiten.
Bei den verwendeten periodischen Fluktuationen wurde auf die Dreiecksfunktion (3.8) zwecks Eingrenzung der Szenarienanzahl verzichtet. Dies ist auch dadurch gerechtfertigt, dass die von Thieler et al. (2013) untersuchten Szenarien mit periodischer Dreiecksfluktuation neben den anderen untersuchten Szenarien mit Sinus- oder Peakfluktuation zu wenig zusätzlichen Erkenntnissen für die verglichenen Periodogrammmethoden führten. Verglichen wurden dabei die auch hier verwendeten Periodogramme basierend auf KQ-, L1- und M-Regression.

4.1.2. Detektionsmethoden

Für die Periodogrammberechnung in Schritt 2 werden 84 verschiedene Methoden verwendet: Es kann zwischen sechs Modellen gewählt werden, die mit sieben verschiedenen Regressionstechniken gewichtet oder ungewichtet angepasst werden können.
Gemäß Abschnitt 2.6.2 werden die Detektionskurven nicht nur für das volle Periodogramm berechnet (Auswertungstyp 1), sondern zusätzlich auch für reduzierte Periodogramme. Dazu werden im Durchlauf zwischen Schritt 3 und Schritt 4 Balken aus dem Periodogramm entfernt. Im einen Fall werden diejenigen entfernt, die nahe des höchsten Balkens liegen (Typ 2), im anderen Fall wird das Periodogramm auf die Balken reduziert, die im ursprünglichen Periodogramm $\mathrm{Per}(p_1), \ldots, \mathrm{Per}(p_{100})$ lokal maximal sind (Typ 3).
Durch Ansiedelung der Reduktion zwischen Schritt 3 und Schritt 4 findet insbesondere die Festlegung von $p^\star$ in Abwesenheit einer periodischen Fluktuation auf dem reduzierten Periodogramm statt. Bei Vorliegen einer periodischen Fluktuation wird $p^\star$ dagegen aus der vollen Testperiodenmenge $p_1 = 1, p_2 = 2, \ldots, p_{100} = 100$ gewählt. Sollte im letzteren Fall $p^\star$ bei der Reduktion aus der Testperiodenmenge entfernt werden, kann hier keine Detektion der Periodizität erfolgen. Während die Detektionskurve für Typ-1-Periodogramme stets den Punkt $H(1) = 1$ enthält, zu einem Signifikanzniveau von $\alpha = 1$ also jede Periode

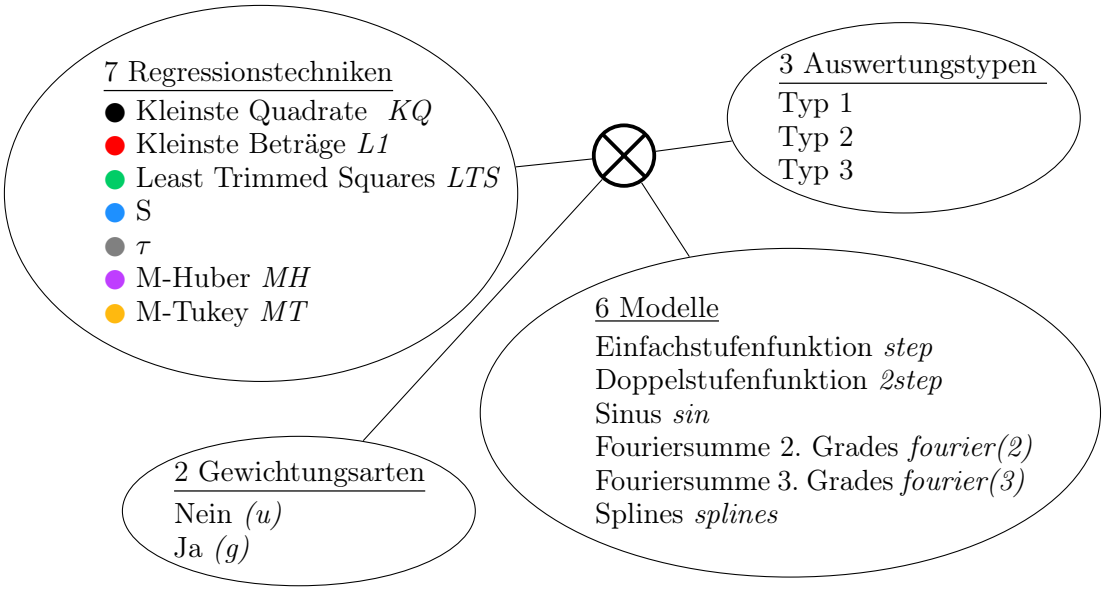


Abbildung 4.2.: Vier verschiedene Parameter werden zu insgesamt 252 unterschiedlichen Detektionsmethoden kombiniert. Alternativ verwendete Abkürzungen sind kursiv gesetzt. Die farbigen Markierungen zeigen den in den farbigen Abbildungen dieses Kapitels angewendeten Farbschlüssel für die Regressionstechniken.

detektiert wird, kann der Wert $H(1)$ für Typ 2 und Typ 3 in Anwesenheit einer periodischen Fluktuation damit geringer ausfallen. Ein solcher Fall ist in Abbildung 4.3(a) zu sehen, wo die Detektionskurve der KQ-Regression (schwarz) nur maximal 0,924 erreicht.
Durch Kombination der durch (gewichtete) Regressionstechnik und anzupassende Funktion definierten 84 Periodogrammmethoden mit den drei verschiedenen Auswertungstypen entstehen 252 Detektionsmethoden. Abbildung 4.2 zeigt eine Übersicht der verschiedenen Kombinationsmöglichkeiten.

4.1.3. Gütemaße

Abbildung 4.3 zeigt Detektionskurven ausgewählter Periodogrammmethoden für verschiedene Datenszenarien. In den Diagrammen ist erkennbar, dass die Kurven sich mitunter schneiden. Damit ist keine allgemeine Aussage mehr möglich, welche Kurve zur besser detektierenden oder besser das Niveau einhaltenden Methode gehört. Im Folgenden werden daher Gütemaße eingeführt, die durch Integration der Detektionskurven die Güte der jeweiligen Detektion beurteilen sollen.
In Abbildung 4.3(a) wird die Detektionskurve für eine Lichtkurve dargestellt, die eine periodische Fluktuation mit Fluktuationsperiode p_f enthält. In diesem Beispiel ist p_f Element der Testperiodenmenge, es gilt also $p^\star = p_f$ und $H(\alpha)$ beschreibt, mit welcher relativen Häufigkeit p_f zum Niveau α detektiert wird. In Anwesenheit einer periodischen Fluktuation sollte die Detektionskurve einer erfolgreichen Methode hohe Werte aufweisen. KQ-Regression (schwarz) erreicht in den gezeigten Beispielkurven für kleine α höhere Detektionsanteile als LTS-Regression (grün). Bei $\alpha \approx 0{,}06$ liegt aber der Schnittpunkt beider Detektionskurven und für $\alpha > 0{,}06$ erreicht LTS-Regression mehr Detektionen. Ebenfalls

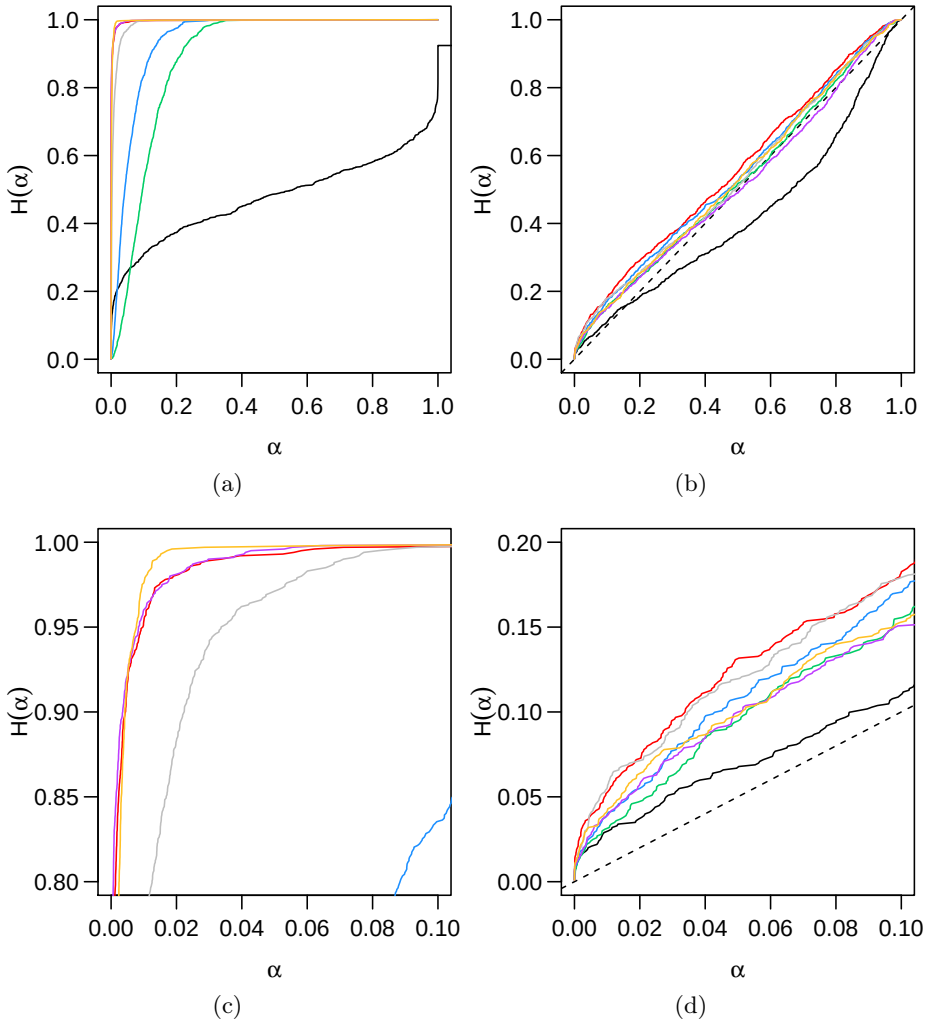


Abbildung 4.3.: Beispiele für Detektionskurven $H(\alpha)$. Detektionsmethoden: Gewichtete Anpassung einer Fouriersumme dritten Grades, Typ-3-Auswertung, Farbkodierung der Regressionstechniken gemäß Tabelle 4.1. Datenszenario: Sinussampling und Intervallstörung, (a) mit periodischer Sinusfluktuation der Länge $p_f = 14$ (Ausschnittsvergrößerung: (c)), (b) ohne periodische Fluktuation (Ausschnittsvergrößerung: (d)). Die gestrichelte Linie in (b) und (d) markieren die Winkelhalbierende.

Regression	KQ	L1	LTS	S	τ	M-Huber	M-Tukey
Gütemaß	●	●	●	●	●	●	●
MDα in (a)	0,245	0,981	0,223	0,506	0,917	0,984	0,983
MDN in (a)	0,211	0,869	0,076	0,213	0,652	0,919	0,855
MDα in (b)	0,065	0,117	0,090	0,101	0,114	0,093	0,097
MAAW in (b)	0,015	0,067	0,040	0,051	0,064	0,043	0,047
M(A)GAW in (b)	2,527	10,583	6,209	7,944	10,121	6,801	7,473

Tabelle 4.1.: Gütemaße für die Detektionskurven in Abbildung 4.3. MAGAW und MGAW sind identisch (alle Kurven in Abbildung 4.3(b) liegen über der Winkelhalbierenden).

schwierig sind L1-Regression (rot), M-Tukey-Regression (gelb) und M-Huber-Regression (violett) gemäß ihrer Detektionsfähigkeit in diesem Szenario in eine Reihenfolge zu bringen.
Als Vergleichswert wird hier die mittlere Detektionsrate nach α

$$\mathrm{MD}\alpha(H) = \frac{1}{\alpha_{\max}} \int_0^{\alpha_{\max}} H(\alpha) d\alpha$$

vorgeschlagen. Sie liegt für alle Detektionskurven zwischen null (keine Detektion für $\alpha \leq \alpha_{\max}$) und eins (perfekte Detektion für beliebiges α) und kann als durchschnittliche Detektionsrate auf dem Intervall $[0, \alpha_{\max}]$ interpretiert werden. Die obere Integralgrenze $\alpha_{\max} = 0{,}1$ ist so gewählt, dass in der Praxis übliche Signifikanzniveaus abgedeckt sind. Statt der mittleren Detektionsrate nach α kann auch die mittlere Detektionsrate nach dem tatsächlich erreichten Niveau

$$\mathrm{MDN(H)} = \frac{1}{\alpha_{\max}} \int_0^{\alpha_{\max}} H(H_0(\alpha)) dH_0(\alpha)$$

betrachtet werden. Dabei beschreibt H_0 die Detektionskurve für Szenarien, die sich von den in H verwendeten Szenarien nur durch das Fehlen einer periodischen Fluktuation unterscheiden. Die MDN berücksichtigt neben der Fähigkeit einer Methode, vorhandene Perioden zu detektieren, auch ihre Fähigkeit zur Niveauerhaltung. Der Wert α hat dabei die Rolle eines Tuning-Parameters, der Sensitivität und Spezifität gegenläufig beeinflusst. Wie die MDα liegt die MDN zwischen null und eins, wobei höhere Werte für eine erfolgreichere Methode sprechen. Je exakter eine Methode das Niveau einhält, umso stärker ähneln sich MDα und MDN.
Neben Gütemaßen für Szenarien mit periodischer Fluktuation werden auch Kennzahlen für fluktuationsfreie Szenarien benötigt. Die Detektionskurven für Szenarien ohne periodische Fluktuation messen den Anteil an Simulationsdurchläufen, in dem eine beliebige Periode detektiert wurde. Beispiele sind in den Abbildungen 4.3(b) und 4.4 gegeben. Die Detektionskurve einer Methode, die das Niveau exakt einhält, sollte ungefähr auf der Winkelhalbierenden liegen. Methoden, deren Detektionskurven unter der Winkelhalbierenden liegen, sind konservativ: Sie halten das Niveau zwar ein, schöpfen es aber nicht aus.
Zur Beurteilung der Einhaltung des Niveaus werden zwei Gütemaße im fluktuationsfreien Szenario betrachtet: Die bereits eingeführte MDα und die mittlere absolute Abweichung von der Winkelhalbierenden

$$\mathrm{MAAW} = \frac{1}{\alpha_{\max}} \int_0^{\alpha_{\max}} |H(\alpha) - \alpha| d\alpha.$$

Für Methoden, die das Niveau einhalten, liegt die MDα bei $\alpha_{\max}/2$, für Methoden, die das Niveau nicht ausschöpfen, niedriger. Die MAAW liegt zwischen null und $1 - \alpha_{\max}/2$ und für exakt das Niveau einhaltende Methoden nahe null. Für Methoden, die das Niveau nicht ausschöpfen, steigt sie ebenso an wie für Methoden, die das Niveau nicht einhalten.
Neben MAAW und MDα werden noch zwei weitere, gewichtete, Kennzahlen betrachtet. Sie beruhen auf dem Gedanken, dass die Detektionskurve einer das Niveau einhaltenden Methode ungefähr auf der Winkelhalbierenden liegen sollte, allerdings mit einer variierenden,

von α abhängigen Streuung um die Winkelhalbierende zu rechnen ist. Es werden daher auch die mittlere gewichtete Abweichung von der Winkelhalbierenden

$$\text{MGAW} = \frac{1}{\alpha_{\max} - \alpha_{\min}} \int_{\alpha_{\min}}^{\alpha_{\max}} \frac{H(\alpha) - \alpha}{\sqrt{\frac{\alpha(1-\alpha)}{1000}}}\, d\alpha$$

und die mittlere absolute gewichtete Abweichung von der Winkelhalbierenden

$$\text{MAGAW} = \frac{1}{\alpha_{\max} - \alpha_{\min}} \int_{\alpha_{\min}}^{\alpha_{\max}} \left| \frac{H(\alpha) - \alpha}{\sqrt{\frac{\alpha(1-\alpha)}{1000}}} \right| d\alpha$$

betrachtet. Hier ist die Abweichung von der Winkelhalbierenden mit der jeweiligen Standardabweichung einer binomialverteilten Zufallsvariable mit 1000 Versuchen und Erfolgswahrscheinlichkeit α gewichtet. Für eine exakt das Niveau einhaltende Methode und festes α kann $H(\alpha)$ als Realisation einer solchen Zufallsvariable angesehen werden. Die sich ergebenden Werte für die MGAW (MAGAW) können als Anzahl der Standardabweichungen interpretiert werden, mit der $H(\alpha)$ auf dem Intervall $[\alpha_{\min}, \alpha_{\max}]$ durchschnittlich vom Sollwert α (absolut) abweicht. Hierbei ist $\alpha_{\min} = 0{,}01$ größer null gewählt, da für $\alpha = 0$ der Nenner des Integrands null und der Integrand somit nicht definiert ist. Die obere Integralgrenze ist wie zuvor $\alpha_{\max} = 0{,}1$ gewählt.
Die Gütemaße zu den in Abbildung 4.3 gezeigten Detektionskurven sind in Tabelle 4.1 notiert. Da im fluktuationsfreien Szenario alle hier gewählten Detektionskurven im Bereich $[0, \alpha_{\max} = 0{,}1]$ über der Winkelhalbierenden liegen (vgl. Abbildung 4.3(d)), sind die Werte MGAW und MAGAW identisch und MDα und MAAW unterscheiden sich nur um $\alpha_{\max}/2 = 0{,}05$. L1- (rot), M-Huber- (violett) und M-Tukey-Regression (gelb) haben im Beispiel in Abbildung 4.3(a) sehr ähnliche MDα-Werte knapp über 0,98, die vergleichsweise niedrigen Detektionsraten der M-Tukey-Regression für kleine α können durch den anschließenden rapiden Anstieg der Detektionskurve ausgeglichen werden. Im fluktuationsfreien Szenario in (b) liegt die Detektionskurve der M-Huber-Regression jedoch meist unter der der L1- und der M-Tukey-Regression (näher an der Winkelhalbierenden). Dies macht sich in niedrigeren MDα-, MAAW-, MGAW- und MAGAW-Werten und in einem höheren MDN-Wert für das Szenario (a) mit periodischer Fluktuation bemerkbar. Ebenfalls ähnlich in MDα und unterschiedlich in MDN für das Beispielszenario mit periodischer Fluktuation sind KQ- (schwarz) und LTS-Regression (grün).
In den Simulationen zeigt sich, dass die Ergebnisse von MDα und MAAW stets ähnlich zu interpretieren sind wie die ihrer gewichteten Pendants MGAW und MAGAW. Die Rangfolge der Detektionsmethoden bezüglich MDα bzw. MAAW weicht nur geringfügig von der bezüglich MGAW bzw. MAGAW ab. Im Weiteren werden daher nur MDα und MAAW betrachtet, die nicht vom zusätzlichen Tuningparameter $\alpha_{\min}$ abhängen und als intuitiver interpretierbar eingestuft werden.
Im Folgenden werden die Ergebnisse der Simulation besprochen.

4.2. Einhaltung des Signifikanzniveaus

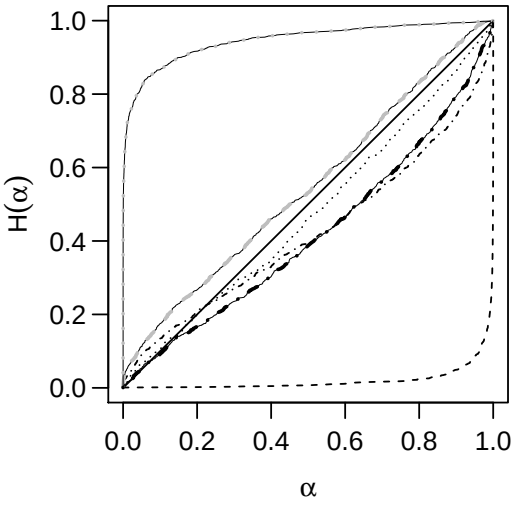
In diesem Abschnitt werden die Szenarien ohne periodische Fluktuation und sowohl mit als auch ohne Intervallstörung betrachtet (vgl. Abbildung 4.1). Zur Analyse werden die in Abschnitt 4.1.3 eingeführten Gütemaße MDα und MAAW verwendet.

4.2.1. Festlegung von Schwellwerten

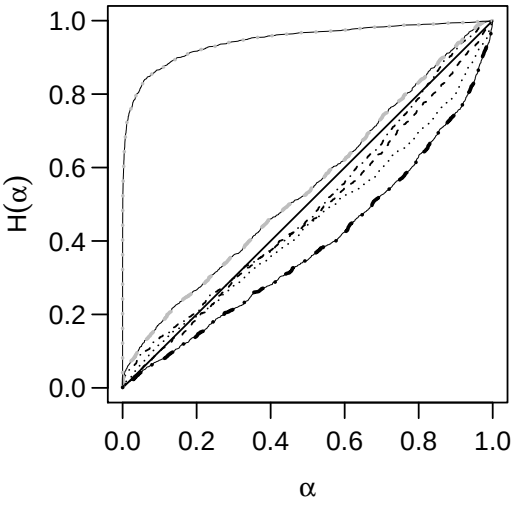
Um zu entscheiden, welcher Wert für das jeweilige Gütemaß akzeptabel ist, werden je Gütemaß die Quartile, Minimum und Maximum aller Detektionskurven berechnet und für das jeweilige Resultat ein Repräsentant gesucht. Die Detektionskurven der Repräsentanten sind in Abbildung 4.4 dargestellt. Sie sind nicht repräsentativ für die jeweilige Methode, Auswertungstyp oder Datenlage, sondern sollen als Repräsentanten für Kurven mit ähnlicher MDα bzw. MAAW verstanden werden. Anhand dieser Grafiken und der Betrachtung weiterer Repräsentanten wurde entschieden, für Methoden mit MDα-Wert kleiner 0,0605 bzw. MAAW-Wert kleiner 0,0105 anzunehmen, dass sie das Niveau zufriedenstellend einhalten bzw. auch ausschöpfen. Tabelle 4.2 zeigt, für welche der vier Szenarien ohne periodische Fluktuation die Methoden dies vermögen. Nach diesen Schwellwerten hält in Abbildung 4.4(c) die gepunktete Detektionskurve mit MDα 0,060 das Niveau ein, die gestrichpunktete Kurve mit MDα 0,082 jedoch nicht. In Abbildung 4.4(d) hält die fett schwarz gestrichpunktete Kurve mit MAAW 0,010 das Niveau ein und schöpft es aus, die gepunktete Linie mit MAAW 0,013 nicht.
Es wird anhand dieser Tabelle deutlich, dass die Auswertungstypen 2 und 3 meist Nachteile bezüglich der Einhaltung des Signifikanzniveaus bringen. In fast allen Fällen wird mit ihnen in weniger Situationen oder nach weniger Kriterien das Niveau eingehalten als mit Auswertungstyp 1. Bei L1- und S-Regression ist die Verwendung von Typ 1 immer von Vorteil. Um das Niveau einzuhalten ist damit für alle Periodogrammmethoden die Empfehlung, Typ 1 zu verwenden.
Bei eingehender Betrachtung der Tabelle fällt auf, dass LTS-Regression wie auch L1-Regression häufig das Niveau nicht einhalten kann, bei der Analyse von Lichtkurven ohne periodische Fluktuation wird also oft eine Periode detektiert. Durch Anpassung einer Fouriersumme dritten Grades und Verwendung von Auswertungstyp 1 kann mit allen Techniken außer der LTS-Regression in allen Situationen das Niveau eingehalten werden, häufig sogar exakt.

4.2.2. LTS-Regression

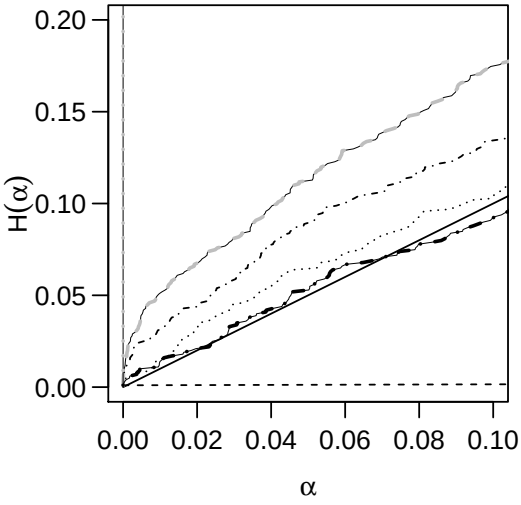
Das Unvermögen vor allem der LTS-Regression zur Niveauerhaltung wird exemplarisch genauer untersucht. Hierzu wird aus einem Datenszenario mit Sinussampling und ohne Intervallstörung ein spezieller Durchlauf gewählt: In diesem Durchlauf wird mit ungewichteter LTS-Anpassung einer Splinefunktion und anschließender Auswertung des Periodogramms nach Typ 1 zum Niveau $\alpha = 0{,}01$ eine Periode detektiert. Bei Verwendung einer beliebigen anderen ungewichteten Regressionstechnik erfolgt aber zu keinem Niveau $\alpha < 0{,}1$ eine Detektion. Abbildung 4.5(a) zeigt die verschiedenen Periodogramme. Es fällt auf, dass



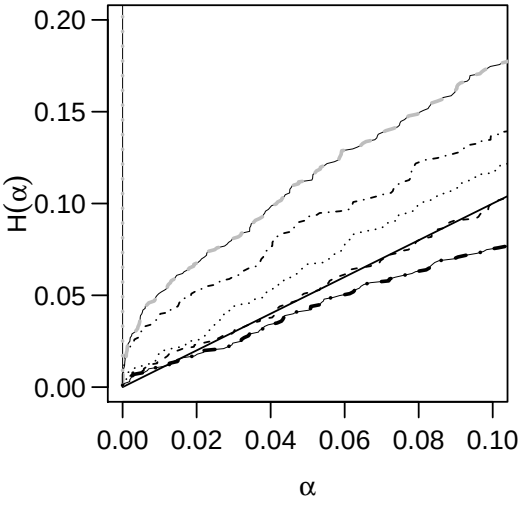
(a)



(b)



(c)



(d)

MDα-Werte:	Wskt. δ	δ-Quantil
- - - - -	0	0,000
—•—	0,116	0,050
·····	0,25	0,060
-·-·-	0,5	0,082
———	0,75	0,108
———	1	0,799

(e)

MAAW-Werte:	Wskt. δ	δ-Quantil
- - - - -	0	0,002
—•—	0,193	0,010
·····	0,25	0,013
-·-·-	0,5	0,034
———	0,75	0,058
———	1	0,749

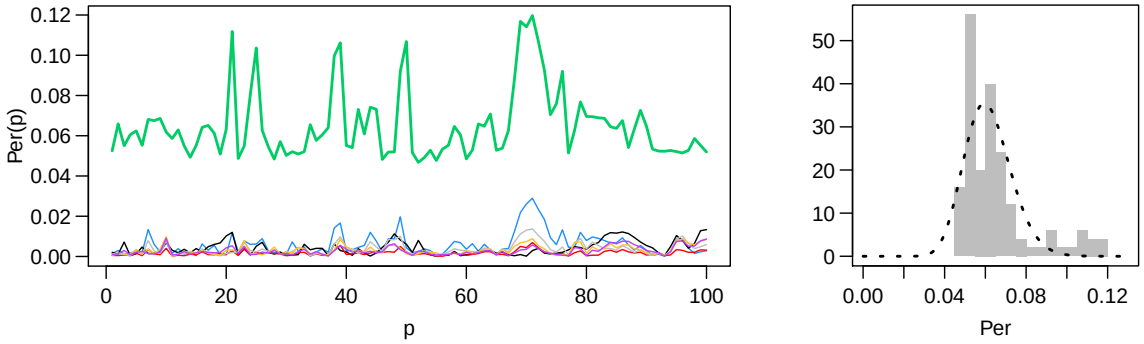
(f)

Abbildung 4.4.: Quantile der beobachteten Werte für Gütemaße. Gütemaß: (a) MDα,(b) MAAW. Darstellung: (oben) Detektionskurven, (mittig) Ausschnittsvergrößerungen von oben, (unten) Angabe der entsprechenden Quantile. Die durchgezogene Linie zeigt die Winkelhalbierende. Tabelle I.1 im Anhang gibt an, zu welchen Szenarien und Detektionsmethoden die hier gezeigten Detektionskurven gehören.

	step	(g)	2step	(g)	sin	(g)	fourier(2)	(g)	fourier(3)	(g)	splines	(g)
L2 (1)	1234	12..	1.34	.234	..34	.2.4	1234	12.4	1234	1234	1234	12.4
L2 (2)	..3.	.2..	..3.	....	..3.	....	..34	....	..34	1.34	..34	...4
L2 (3)	..34	....	..34	....	....	....	..34	....	..34	....	..34	....
L1 (1)	12..	12..	....	.2..	....	....	12..	1...	1234	12.4	12..	12..
L1 (2)	....	.2..	....	....	....	....	....	....	....	1...	....	....
L1 (3)	....	....	....	....	....	....	....	....	....	....	....	....
LTS (1)	12..	1234	....	1.3.	....	....	....	1234	....	1234	....	....
LTS (2)	....	..3.	....	....	....	....	....	.2..	....	1234	....	....
LTS (3)	....	....	....	....	....	....	....	....	....	....	....	....
S (1)	123.	1234	1.3.	123.	....	1.34	1.3.	1234	1234	1234	.23.	1234
S (2)	12..	1.34	....	....	....	....	....	..3.	....	....	....	..34
S (3)	....	....	....	....	....	....	....	....	....	....	....	....
τ (1)	1234	1234	1.3.	1234	....	1...	1.34	.234	1234	1234	1234	1234
τ (2)	....	1234	....	....	....	....	....	....	....	....	....	..3.
τ (3)	....	....	....	....	....	....	....	....	....	....	....	....
MH (1)	1234	12..	..34	12..	....	.2..	.2..	.2..	1234	12.4	....	12..
MH (2)	....	12..	....	....	....	....	....	....	..3.	.2..	....	....
MH (3)	..3.	....	..3.	....	....	....	....	....	..3.	....	....	....
MT (1)	.2..	1234	....	12.4	..34	..34	.234	.234	1234	1234	..34	1234
MT (2)	....	12.4	....	....	....	....	....	....	....	.2.4	....	....
MT (3)	....	....	....	....	....	....	....	....	....	....	....	....

Tabelle 4.2.: Szenarien, bei denen eine Detektionsmethode das Niveau nach den im Text genannten Schwellwerten einhält (gemäß MDα) bzw. auch ausschöpft (gemäß MAAW, unterstrichen). Die vier Szenarien sind Sinussampling (1) und Dreiecksampling (2) jeweils ohne Intervallstörung und Sinussampling (3) und Dreiecksampling (4) jeweils mit Intervallstörung. Die Zeilen geben die verwendete Regressionstechnik (Auswertungstyp in Klammern), die Spalten das angepasste Modell (jeweils erst ungewichtet, dann gewichtet) an. Die verwendeten Abkürzungen sind in Abbildung 4.2 eingeführt.

die LTS-Periodogrammbalken sehr viel höhere Werte annehmen als die der anderen Regressionstechniken. Alle LTS-Periodogrammbalken erreichen mit einer minimalen Höhe von 0,047 mindestens 39 Prozent des maximal beobachteten LTS-Periodogrammbalkens, während der minimale Periodogrammbalken bei anderen Regressionstechniken nur ein oder zwei Prozent des maximalen Ausschlags hoch ist. Gemessen am arithmetischen Mittel oder Median streuen die Balken des LTS-Periodogramms um ungefähr die Hälfte des maximalen Ausschlags, die der anderen Periodogramme um ein Drittel der jeweils maximalen Höhe oder weniger. Die an die Periodogrammbalken des LTS-Periodogramms angepasste Betaverteilung (vgl. Abbildung 4.5(b)) scheint nahezu symmetrisch, während das Histogramm der Periodogrammbalken rechtsschief wirkt. Die Annahme der Betaverteilung scheint hier also nicht gerechtfertigt zu sein. Eine schlechte Optimierung im LTS-Algorithmus scheint nicht der Grund für dieses Phänomen zu sein: Für verschiedene Testperioden, deren Periodogrammbalken für LTS-Regression relativ niedrig liegen, während sie bei anderen Regressionstechniken zumindest lokal hervorstechen, wird der angepasste Parametervektor $\widetilde{\beta}_{LTS}$ mit den Anpassungen der anderen Regressionstechniken verglichen. Keiner der anderen geschätzten Parametervektoren erreicht ein kleineres LTS-Kriterium. Dasselbe kann auch für andere stichprobenartig ausgewählte Simulationsdurchläufe des gleichen Szenarios und für Lichtkurven mit Dreiecksampling beobachtet werden.



(a) (b)

Abbildung 4.5.: (a) Periodogramme, durch Anpassung einer Splinefunktion mit verschiedenen Regressionstechniken (Farbkodierung vgl. Abbildung 4.2) für die gleiche Lichtkurve (Sinussampling, keine periodische Fluktuation, ohne Intervallstörung) gewonnen. (b) Histogramm des LTS-Periodogramms (grün in (a)) mit angepasster Betaverteilung.

4.2.3. Sinusfunktion

Bei Betrachtung der Tabelle 4.2 fällt weiterhin auf, dass bei Anpassung einer Sinusfunktion das Niveau besonders schlecht eingehalten wird. Das ist interessant, da die Sinusanpassung auch von Standardmethoden der Astroteilchenphysik, wie dem Lomb-Scargle-Periodogramm oder der Date-Compensated-Fourier-Transform, verwendet wird (vgl. Abschnitt 2.3.1).
Die Sinusanpassung wird zunächst für KQ-Regression auf Lichtkurven ohne periodische Fluktuation zum Niveau 0,05 eingehender analysiert. Speziell wird dabei der Frage nachgegangen, ob durch Nutzung der vordefinierten $\mathcal{B}\left(\frac{m-1}{2}, \frac{n-m}{2}\right)$-Verteilung (vgl. Gleichung 2.38, Abschnitt 2.6.1) oder in Szenarien ohne periodische Messzeitverteilung bessere Ergebnisse erzielt werden.
Tatsächlich hält die Detektionsmethode bei Verwendung der vordefinierten Betaverteilung das Niveau besser ein, der Fehler erster Art liegt je nach Datenszenario zwischen 0,033 und 0,041. Dies gilt allerdings nur für Datenszenarien ohne Intervallstörungen: Bei intervallgestörten Daten erhöht sich der Fehler erster Art auf 0,989 bis 1.000, während er bei Nutzung der CvM-angepassten Betaverteilung zwischen 0,015 und 0,108 liegt (ohne Intervallstörung zwischen 0,062 und 0,078). Bei Anwendung der Detektionsmethoden auf Lichtkurven mit rechteckverteilten Messzeiten, also ohne periodisches Sampling, liegen die Fehler erster Art ungefähr in der gleichen Größenordnung wie bei periodischem Sampling. Dabei wird in Szenarien ohne Intervallstörung bei allen betrachteten Messzeitverteilungen eine Testperiode tendenziell um so häufiger detektiert, je kleiner sie ist. Bei Szenarien mit Intervallstörung werden vor allem die sehr großen Testperioden detektiert. Bei Vorliegen einer periodischen Messzeitverteilung wird in letzterem Fall auch zusätzlich die Samplingperiode gehäuft detektiert.
Analoge Untersuchungen der Sinusanpassung wurden auch für L1-, Huber-M und τ-Regression durchgeführt. Hier führt die Nutzung der vordefinierten Betaverteilung zu einer starken Unterschreitung des Niveaus, der maximale Fehler erster Art bei vorgegebenem Niveau $\alpha = 0{,}05$ liegt für L1-Regression bei 0,004, für τ-Regression bei 0,01. Bei der M-Huber-Regression liegt der Fehler erster Art je nach Datenszenario zwischen 0 und

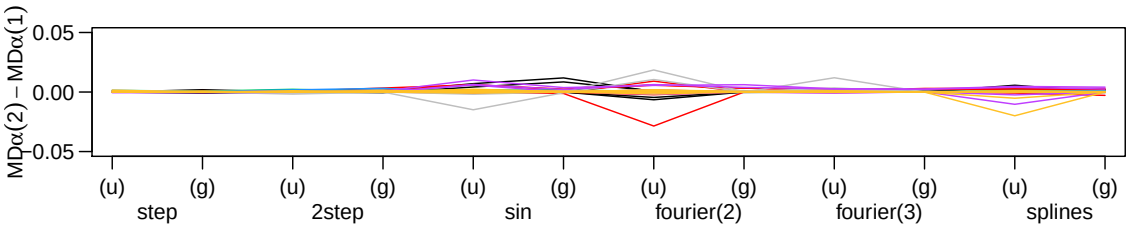
0,061. In den Szenarien mit Intervallstörung ändert sich der Fehler erster Art für L1- und τ-Regression nicht (Änderungen in der dritten Nachkommastelle). Für die gewichtete M-Huber-Regression sinkt der Fehler erster Art um eine Zehnerpotenz, bei ungewichteter Anpassung ist kein systematischer An- oder Abstieg erkennbar. Das Muster der detektierten Perioden ähnelt dem für KQ-Regression beschriebenen, allerdings detektieren die Methoden bei Vorliegen einer Intervallstörung nicht vorwiegend hohe Perioden.
Die Probleme der Sinusmethoden bei der Niveaueinhaltung entstehen also scheinbar nicht durch die periodische Messzeitverteilung, wenn die Samplingperiode bei Vorliegen einer Intervallstörung auch häufig Detektionsergebnis ist. Für KQ-Regression ließe sich durch Nutzung der vordefinierten Betaverteilung in intervallgestörten Szenarien besser das Niveau einhalten, bei Vorliegen einer Intervallstörung funktioniert dies nicht mehr. Für die eingehender untersuchten robusten Regressionstechniken führt die Verwendung der vordefinierten Betaverteilung zu einer sehr konservativen Detektion.

4.3. Detektionsvermögen

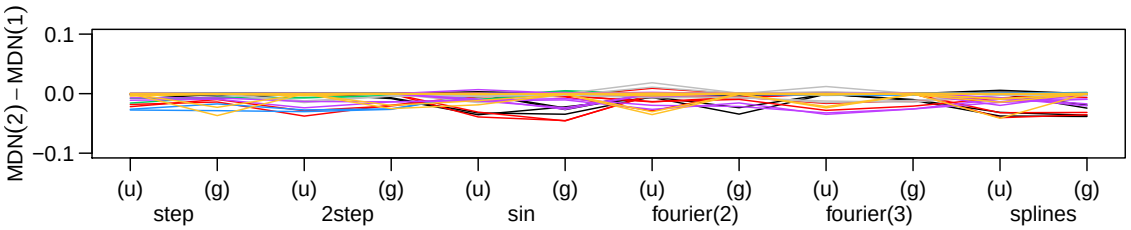
In diesem Abschnitt werden die Szenarien mit periodischer Fluktuation und ohne Intervallstörungen analysiert. Sofern andere Szenarien zum Vergleich hinzugezogen werden, wird explizit darauf hingewiesen. Die Analyse erfolgt mit Hilfe der MDα und der MDN (vgl. Abschnitt 4.1.3).

4.3.1. Auswertungstypen

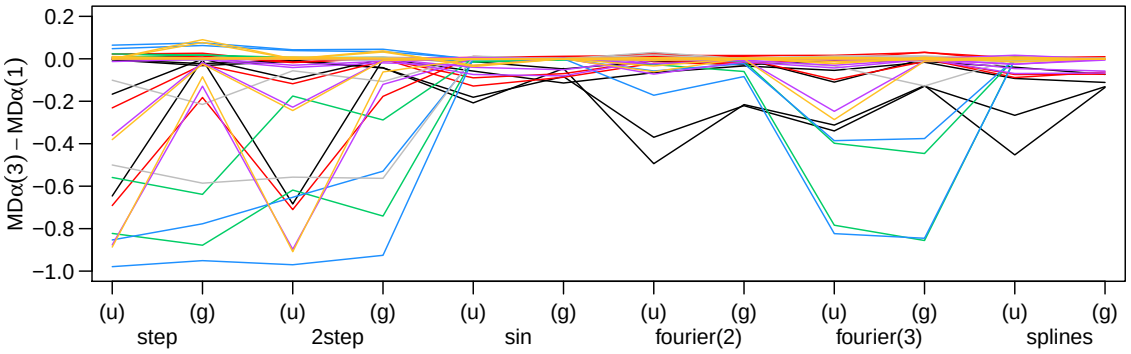
Zunächst wird die Auswirkung der verschiedenen Auswertungstypen auf das Detektionsvermögen einer Detektionsmethode betrachtet. Abbildung 4.6(a) zeigt die Differenzen der MDα-Werte zwischen Typ-2- und Typ-1-Auswertung. Die Differenzen liegen für alle Periodogrammmethoden nahe null und erreichen einen Wert von maximal 0,018 (zugunsten Typ 2) und minimal -0,028 (zugunsten Typ 1). Die entsprechenden Differenzen der MDN-Werte (Abbildung 4.6(b)) liegen zwischen -0,046 und 0,018 und meist im negativen Bereich, Auswertungstyp 1 erreicht also höhere MDN-Werte. Dies ist mit der in Abschnitt 4.2 gemachten Beobachtung zu erklären, dass mit Auswertungstyp 2 meist das Niveau nicht eingehalten wird.
In den Abbildungen 4.6(c) und 4.6(d) sind die entsprechenden Diagramme für die Differenzen zwischen Typ-3- und Typ-1-Auswertung zu sehen. Auswertungen von Typ 3 und Typ 1 unterscheiden sich stärker als solche von Typ 2 und Typ 1, die Differenzen liegen hier zwischen -0,98 (zugunsten Typ 1) und 0,09 (zugunsten Typ 3). Die MDN-Differenzen liegen zwischen -0,998 und 0,03, Typ 1 erreicht häufig viel bessere Werte als Typ 3. Zusammengefasst spricht alles dafür, mit Auswertungstyp 1 weiterzuarbeiten. Die hier beschriebenen Analysen wurden auch für Szenarien mit Intervallstörung durchgeführt und führten zu vergleichbaren Ergebnissen, die die Entscheidung für eine Typ-1-Auswertung weiter stützen. In den folgenden Analysen werden daher nur Detektionsmethoden mit Auswertungstyp 1 berücksichtigt.



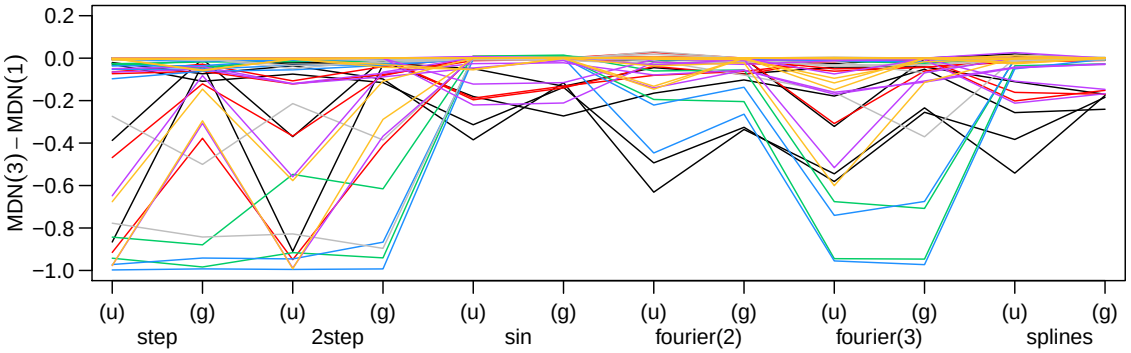
(a)



(b)



(c)



(d)

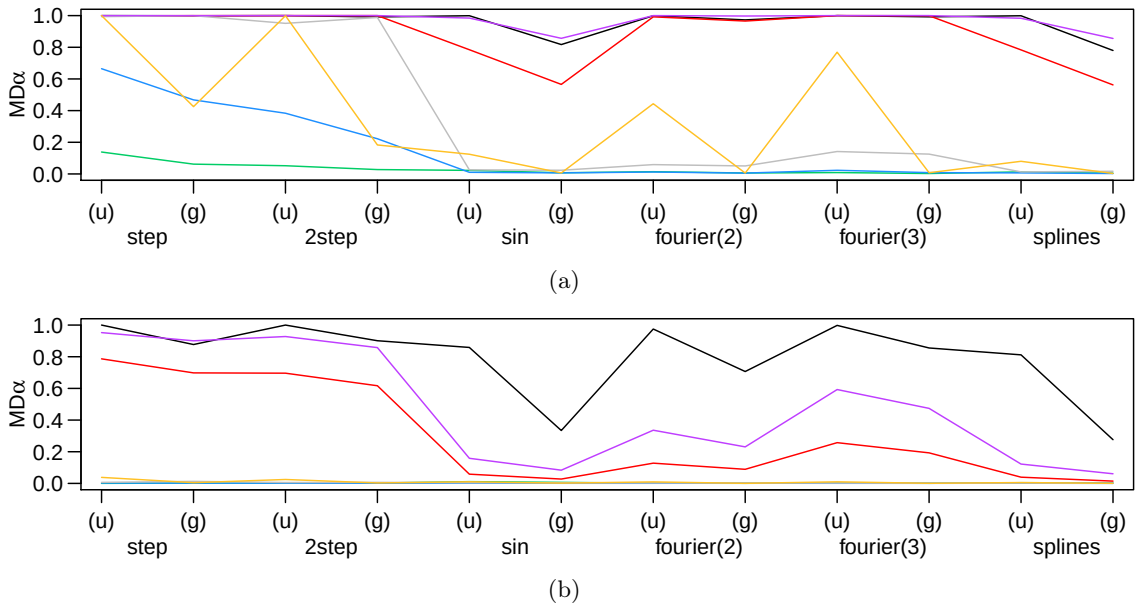
Abbildung 4.6.: Kennzahlendifferenzen unterschiedlicher Auswertungstypen für verschiedene Detektionsmethoden (Modell, ungewichtete (u) /gewichtete (g) Regression: Abszisse. Regressionstechnik: Farbe, vgl. Abbildung 4.2). Betrachtete Differenz: (a) MDα: Typ 2-Typ 1, (b) MDN: Typ 2-Typ 1, (c) MDα: Typ 3-Typ 1, (d) MDN: Typ 3-Typ 1. Die Verbindungslinien stellen keinen Verlauf dar, sondern dienen nur der Orientierung. Jede Linie steht für ein Datenszenario mit periodischer Fluktuation ohne Intervallstörung mit fester Regressionstechnik. Die unterschiedliche Skalierung der y-Achsen ist zu beachten.

4.3.2. Regressionstechniken

Im Falle einer Sinusfluktuation ergeben alle Detektionsmethoden mit Auswertungstyp 1 sehr gute MDα-Werte von mindestens 0,970 und mehr. Da eventuelle Unterschiede, etwa bezüglich der Regressionstechnik, somit nur gering sein können und nicht als relevant eingestuft werden, wird im Folgenden das Verhalten der Detektionsmethoden in Szenarien mit Peakfluktuation eingehender betrachtet.
In Abbildung 4.7 werden die MDα-Werte für verschiedene Datenszenarien mit periodischer Peakfluktuation gezeigt. LTS-Regression (grün) ergibt sehr kleine Werte für alle angepassten Modelle in allen gezeigten Szenarien. Mit dem Sinus-, dem Fourier- oder dem Splinemodell erreichen auch τ- (grau), S- (blau) und gewichtete M-Tukey-Regression (gelb, (g)) MDα-Werte nahe null. Im Falle einer periodischen Fluktuation der Periode $p_f = 33$ (Abbildung 4.7(b)) sinken zusätzlich die MDα-Werte der anderen robusten Regressionstechniken (L1 (rot), M-Huber (violett)) unter die MDα-Werte der KQ-Regression (schwarz). In diesem Fall werden also mit KQ-Regression die besten mittleren Detektionsraten erzielt.
Wegen des auffällig schlechten Abschneidens von LTS- und S-Regression werden diese beiden Techniken in Einzelbeispielen genauer betrachtet. Bei der Betrachtung verschiedener Beispiele mit Peakfluktuation mit $p_f = 14$ und Anpassung einer Stufenfunktion wird festgestellt, dass das S-Periodogramm bei der Fluktuationsperiode lokal maximal ist, sich jedoch nicht stark von den anderen Balken abhebt und häufig nicht global maximal ist. Eine fehlgeschlagene Optimierung scheint nicht der Grund zu sein, zumindest kann für die betrachteten Beispiele das S-Zielkriterium durch Verwendung eines mit einer anderen Regressionstechnik gewonnenen Schätzers nicht weiter verringert werden. Es ist möglich, dass diese robuste Regressionstechnik sensibler als die anderen Regressionstechniken auf Modellabweichungen reagiert, wenn also die anzupassende periodische Funktion g zu schlecht zur periodischen Fluktuation f passt. Während alle sieben periodischen Modelle die Sinusfluktuation noch leidlich nachzeichnen können, kommt es bei der Peakfunktion bei jeder Modellanpassung zu größeren Residuen. Hinzu kommt, dass bei der Peakfunktion nur wenige Messwerte (die im Peak) maßgeblich Informationen über die Fluktuationsperiode beinhalten und diese Messwerte bei robuster Anpassung als Ausreißers ignoriert oder heruntergewertet werden können. Es wird vermutet, dass dies auch das Problem beim LTS-Periodogramm ist, bei dem der zur Fluktuationsperiode gehörende Periodogrammbalken teilweise nicht einmal lokal maximal ist. Analog zur S-Regression durchgeführte Betrachtungen lassen auch hier kein Optimierungsproblem vermuten. Zusammen mit der Tatsache, dass LTS-Regression häufig das Niveau nicht einhält, scheint diese Regressionstechnik für die in dieser Arbeit vorgeschlagene Detektionsmethodik eher ungeeignet. Die S-Regression sollte nur angewendet werden, wenn zuverlässige Annahmen über die Gestalt einer potentiellen periodischen Fluktuation vorliegen.

4.3.3. Gewichtete Regression

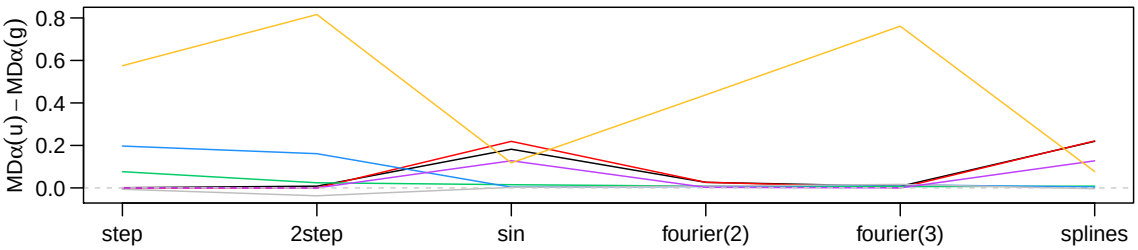
Bei Betrachtung der in Abbildung 4.7 gezeigten MDα-Werte fällt auf, dass die Verwendung gewichteter Regression keinen Vorteil gegenüber ungewichteter Regression zu haben scheint. Speziell in Abbildung 4.7(a) wird deutlich, dass M-Tukey-Regression (gelb) unter Einbezug



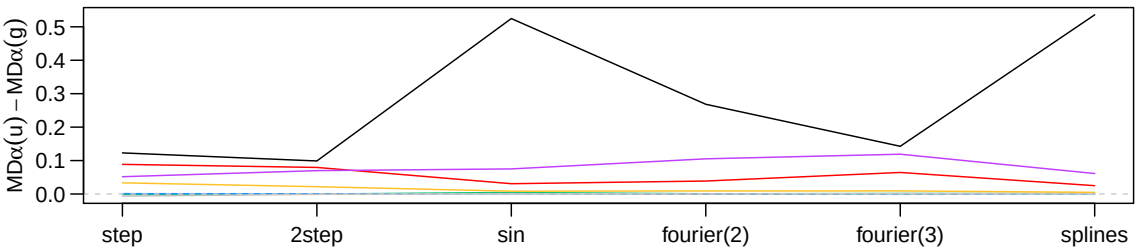
(a)

(b)

Abbildung 4.7.: MDα-Werte für Peakfluktuation und Sinussampling (Ergebnisse für Dreiecksampling sehr ähnlich). Fluktuationsperiode: (a) $p_f = 14$, (b) $p_f = 33$. Für Farbkodierung der Regressionstechniken und Abkürzungen vgl. Abbildung 4.2. Die Verbindungslinien stellen keinen Verlauf dar, sondern dienen nur der Orientierung.



(a)



(b)

Abbildung 4.8.: Differenzen (jeweils ungewichtet minus gewichtet) der MDα-Werte aus Abbildung 4.7. Fluktuationsperiode: (a) $p_f = 14$, (b) $p_f = 33$. Die Verbindungslinien stellen keinen Verlauf dar, sondern dienen nur der Orientierung.

der Gewichte stark an Detektionsfähigkeit einbüßt. In Abbildung 4.8 werden daher die Differenzen der MDα-Werte aus Abbildung 4.7 zwischen gewichteter und ungewichteter Regression dargestellt. Die Differenzen liegen größtenteils über null und damit zugunsten der ungewichteten Regression. Lediglich bei der τ-Anpassung (grau) von Einfach- und Doppelstufenfunktionen sind die Differenzen meist negativ, wenn betraglich auch klein. In diesem Fall werden also durch den Einsatz von Gewichten geringfügig höhere MDα-Werte erreicht. Dies gilt für fast alle Datenszenarien ohne Intervallstörung, die einzige Ausnahme bildet die Anpassung der Doppelstufenfunktion bei langer Peakfluktuation und Dreiecksampling.
Um die Praxistauglichkeit der gewichteten τ-Anpassung einer Stufenfunktion zu überprüfen, wird diese Periodogrammethode auch auf Lichtkurven angewendet, bei denen die Qualität der Messfehler geringer ausfällt als bei den anderen hier getätigten Simulationen: Die Messfehler s_i erklären hier nur einen Anteil ψ der Rauschkomponente (vgl. Definition (3.10), Abschnitt 3.2.3), werden zu fünf Prozent durch niedrigere Ausreißer ersetzt (vgl. Abschnitt 3.2.4) oder beides zugleich. Im Szenario mit der niedrigsten Messfehlerqualität wird nur rund ein Viertel der Rauschvarianz durch die (ausreißerfreien) s_i beschrieben und es sind Ausreißer vorhanden. In allen Szenarien können die Detektionsmethoden nach wie vor das Niveau einhalten, es wird jedoch nicht mehr ausgeschöpft. Periodische Sinusfluktuationen können unvermindert gut entdeckt werden (Änderungen der MDα bzw. MDN-Werte in vierter bzw. dritter Nachkommastelle). Bei Vorliegen einer Peakfluktuation mit Fluktuationsperiode $p_f = 14$ verschlechtert sich die Detektion allerdings massiv. Die MDα liegt dann nur noch bei ungefähr 0,4, wenn die s_i sich nur auf die Hälfte der Rauschvarianz beziehen, bei 0,22. Bei Vorliegen von fünf Prozent Ausreißern in den Messfehlern sinken die MDα-Werte unter 0,16, selbst wenn die ungestörten s_i die komplette Rauschvarianz erklären. Bei zusätzlich vorliegender Intervallstörung ändern sich die MDα-Werte und die damit verbundene Interpretation kaum. Die MDN-Werte der Szenarios mit verminderter Messfehlerqualität sind aufgrund des nicht ausgeschöpften Niveaus etwas höher, das sich ergebende Bild ist jedoch das gleiche. Die gewichtete τ-Regression erreicht bei Vorliegen perfekter Messfehler geringfügig bessere Detektionsergebnisse als die ungewichtete Variante. Bei Vorliegen unperfekter Daten, wie in der Praxis zu erwarten, sind die Detektionsergebnisse für ungewichtete Regression jedoch eindeutig besser. Auf Grundlage dieser Analysen ist damit im Allgemeinen vom Einsatz gewichteter Regression abzuraten, auch bei Vorlage von Messfehlern, die mit den wahren Fehlerstandardabweichungen identisch sind. Eine in Abschnitt 2.6.1 angeregte Diskussion, inwiefern sich nichtexakte Messfehler auf die Verteilung der Periodogrammbalken ausüben, ist damit hinfällig.

4.3.4. Angepasste Modelle

Bei Abbildung 4.7 fällt auch auf, dass das Sinusmodell bei allen Regressionstechniken tendenziell schlechtere MDα-Werte erreicht als die anderen angepassten Modelle. Die nicht detailliert besprochenen Analyseergebnisse der intervallgestörten Daten sehen ähnlich aus. Das kann mit der Punktsymmetrie der Sinusfunktion oder ihrer geringen Parameteranzahl erklärbar sein. Beide Eigenschaften führen womöglich dazu, dass die Peakfunktion mit dem

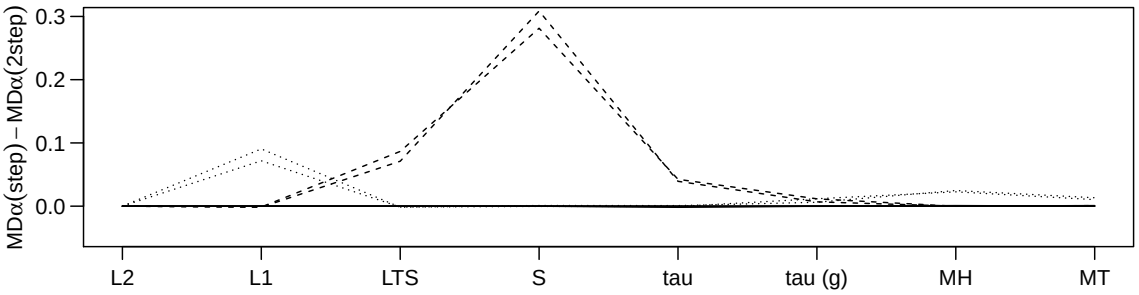


Abbildung 4.9.: MDα-Differenzen: Stufenmodell minus Doppelstufenmodell. Regressionstechnik: Abszisse. Datenszenarien: ohne Intervallstörung, mit Sinus- oder Dreiecksampling (je eine Linie), mit Sinusfluktuation (durchgezogen), Peakfluktuation mit $p_f = 14$ (gestrichelt) oder Peakfluktuation mit $p_f = 33$ (gepunktet). Die Verbindungslinien stellen keinen Verlauf dar, sondern dienen nur der Orientierung.

Sinus schlechter nachgeahmt werden kann als mit den anderen betrachteten Modellen. Es sei an dieser Stelle darauf hingewiesen, dass mit den Sinusperiodogrammen auch seltener als mit den anderen Methoden das Niveau eingehalten werden konnte (vgl. Abschnitt 4.2). Damit scheint die Anpassung von Sinusfunktionen zur Periodendetektion nach der in dieser Arbeit verwendeten Methode nicht empfehlenswert.
Die Anpassung einer Doppelstufenfunktion erweist sich nie als sinnvoller als die einer Einfachstufenfunktion. Abbildung 4.9 zeigt die Differenzen der MDα-Werte von Einfach- und Doppelstufenfunktion. Die Differenzen sind stets positiv, die MDα-Werte für die Einfachstufenfunktion sind also größer. Die gleiche Analyse werden auch für Situationen mit Intervallstörungen durchgeführt. Dabei hat die Doppelstufenfunktion für KQ-Regression geringfügig bessere Werte, jedoch in Szenarien, in denen die MDα-Werte für Einzel- und Doppelstufenfunktion beide sehr gering ausfallen. Die Anpassung einer Einfachstufenfunktion ist damit der Anpassung einer Doppelstufenfunktion vorzuziehen. Detektionsmethoden, die auf der Anpassung der letzteren basieren, werden im Folgenden nicht weiter betrachtet.

4.4. Robustheitsbetrachtung

In diesem Abschnitt werden Szenarien mit Intervallstörungen analysiert. Wegen der in Abschnitt 4.3 gemachten Beobachtungen werden im Folgenden nur Detektionsmethoden mit Auswertungstyp 1 betrachtet. Gewichtete und LTS-Regression werden nicht weiter verfolgt. Weiterhin werden nur noch folgende Modelle eingehender analysiert:

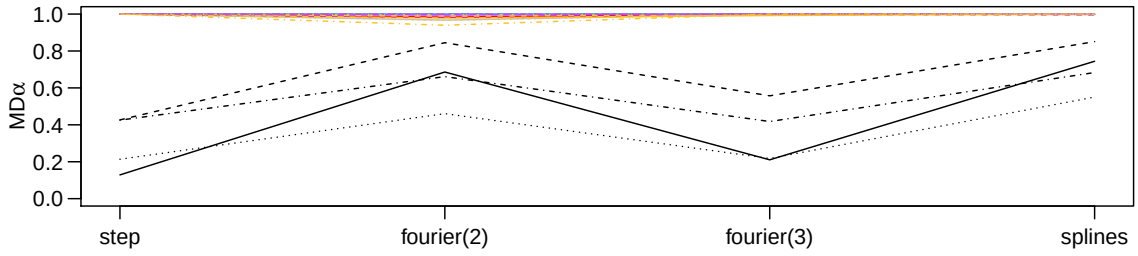
- Einfachstufenfunktionen für alle Regressionstechniken außer der LTS-Regression
- Fouriersummen zweiten und dritten Grades sowie Splines für KQ-, L1- und M-Regression

Es sei darauf hingewiesen, dass auch alle anderen sich aus den Schemata in Abbildung 4.1 und 4.2 ergebenden Kombinationen für intervallgestörte Szenarien analysiert wurden. Die hier nicht besprochenen Ergebnisse schienen jedoch keine weiteren interessanten Aspekte aufzuweisen.

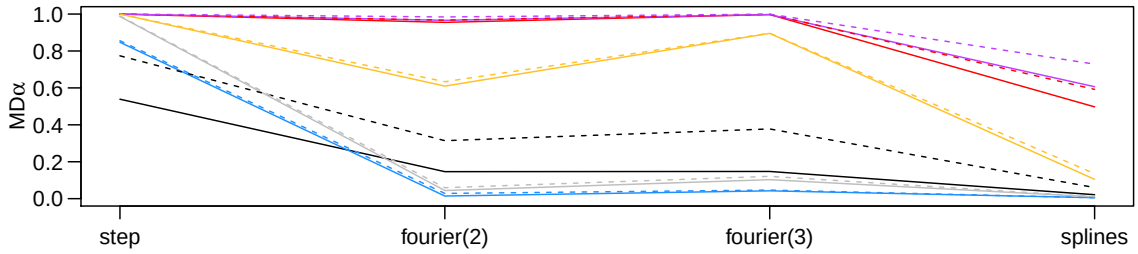
Abbildung 4.10(a) zeigt die MDα-Werte für Szenarien mit periodischer Sinusfluktuation und Intervallstörung. Hier sind unverkennbar Vorteile der robusten Regressionstechniken gegenüber der KQ-Regression (schwarz) erkennbar. Während die MDα für alle robusten Techniken Werte sehr nahe eins erzielt, liegen die der KQ-Regression deutlich niedriger. Die besten Werte für die KQ-Anpassung werden in allen Szenarien (gekennzeichnet durch unterschiedliche Linientypen) bei der Splineanpassung erreicht (ähnliche Werte auch für die hier nicht gezeigte Sinusanpassung).
Auch im Falle einer kurzen Peakfluktuation (vgl. Abbildung 4.10(b)) verliert die KQ-Regression an Detektionsfähigkeit. Auch Methoden, die ein Splinemodell anpassen, werden hier schwächer. Interessanterweise verbessert sich die MDα der S-Regression (blau). Dies kann geringfügig auch für die hier nicht gezeigte LTS-Regression festgestellt werden. Es scheint, als ermögliche das Vorliegen echt gestörter Beobachtungen den Regressionstechniken, mehr Information aus den nicht-gestörten, zum Peak gehörenden Beobachtungen zu ziehen. Dennoch liegen die entsprechenden MDα-Werte unter denen der anderen robusten Techniken.
Im Falle einer Peakfluktuation mit Fluktuationsperiode $p_f = 33$ (vgl. Abbildung 4.10(c)) sinken für KQ-Regression (schwarz), L1-Regression (rot) und M-Huber-Regression (violett) die MDα-Werte für alle Modelle. Wie in der Grafik zu erkennen ist, liegen die MDα-Werte der beiden robusten Regressionstechniken für Stufenfunktionen über den entsprechenden Werten der KQ-Regression. Gemessen an den MDN-Werten (hier nicht gezeigt), erzielt die KQ-Regression jedoch häufig die leicht besseren Detektionsergebnisse. Die anderen Regressionstechniken, S-, τ- und M-Tukey-Regression, eignen sich noch weniger zur Detektion einer periodischen Peakfluktuation der Länge $p_f = 33$. Die MDα (und auch die nicht gezeigte MDN) liegt hier stets nahe null.

4.5. Rechenzeiten

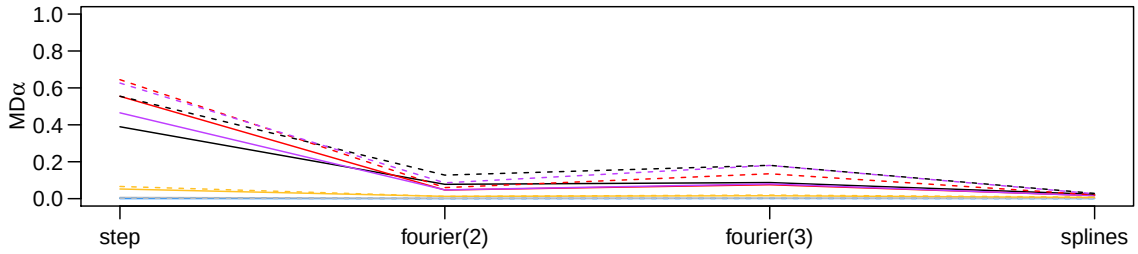
Die Simulation wurde auf dem Rechencluster LiDong (vgl. Becker 2010) der TU Dortmund durchgeführt. Gerechnet wurde auf 3 GHz-Prozessoren der Marke Intel Xeon E5450. Für alle Simulationsdurchläufe wurden die zur Berechnung des Periodogramms benötigten Rechenzeiten festgehalten. Tabelle 4.3 gibt einen groben Überblick über die benötigten Rechenzeiten.
Es ist deutlich erkennbar, dass die Anpassung einer Einfach- oder einer Doppelstufenfunktion jeweils am längsten benötigt. Dies ist verständlich, da die Stufenfunktion mit zehn Stufen von allen angepassten Funktionen die meisten Parameter hat. Da die Doppelstufenfunktion die Anpassung von zwei Stufenfunktionen bedeutet (vgl. Abschnitt 2.3.2), ist nachvollziehbar, dass die Berechnung eines Periodogramms mit dieser Funktion im Mittel bzw. Median ungefähr doppelt so lange dauert wie die Berechnung eines Periodogramms, dass auf der Einfachstufenfunktion basiert. Die einzige Ausnahme von dieser Regel ist für die S-Regression beobachtbar. Möglicherweise führt die effiziente Unterstichprobenfindung bei der Anpassung von Stufenfunktionen (vgl. Abschnitt 3.1.2) dazu, dass ein Großteil der Rechenzeit nicht zur Anpassung einzelner Funktionen benötigt wird, sondern zur Schätzung der Varianz SY in der Lichtkurve.



(a)



(b)



(c)

Abbildung 4.10.: MDα-Werte von Detektionsmethoden mit verschiedenen Modellen (Abszisse) und ungewichteten Regressionstechniken (Farbkodierung, vgl. Abbildung 4.2) auf verschiedenen intervallgestörten Szenarien. Periodische Fluktuation: (a) Sinus, $p_f = 14$ (durchgezogen: Sinussampling, gestrichelt: Dreiecksampling) oder $p_f = 33$ (gepunktet: Sinussampling, gestrichpunktet: Dreiecksampling), (b) Peak, $p_f = 14$ (durchgezogen: Sinussampling, gestrichelt: Dreiecksampling). (c) Peak, $p_f = 33$ (durchgezogen: Sinussampling, gestrichelt: Dreiecksampling). Die Verbindungslinien stellen keinen Verlauf dar, sondern dienen nur der Orientierung.

Neben den Zeitunterschieden durch Anpassung verschiedener Modelle treten jedoch die von der Regressionstechnik abhängigen Unterschiede deutlich stärker hervor. Zur Berechnung eines Periodogramms benötien die KQ- und L1-basierten Methoden nie mehr als drei Sekunden, bei Nutzung von M-Huber-Regression kann es fast drei Minuten dauern. Mit S-Regression dauert es im Mittel je nach Modell zwischen zwei und sechs Minuten, mit M-Tukey zwischen zwei und neun Minuten, mit LTS-Regression zwischen drei und zwölf Minuten, für τ-Regression zwischen fünf und 13 Minuten. Bei diesen Techniken wurden jedoch enorm höhere maximale Berechnungszeiten beobachtet. Die maximal beobachtete Zeit für die LTS-Anpassung einer Doppelstufenfunktion von 18 680 Sekunden (entspricht 5 Stunden, 11 Minuten und 20 Sekunden) kann anhand der ähnlichen Werte für Median und arithmetisches Mittel als Ausreißer identifiziert werden und ist vermutlich durch einen fehlerhaften Prozessor entstanden. Es kam in zwei von 40 000 Durchgängen mit LTS-Anpassung einer Doppelstufenfunktion zu einer Rechenzeit von über fünf Stunden, die nächstkürzere Dauer beträgt 21 Minuten und 16 Sekunden. Länger brauchen nur die M-Tukey- und die τ-Anpassung einer Doppelstufenfunktion mit maximal 2186 und 1680 Sekunden (36,433 und 28 Minuten).
Detailliertere Darstellungen der Rechenheiten sind in Anhang I.2 zu finden. Dort ist erkennbar, dass die Periodogrammberechnung vor allem unter Nutzung von M-Tukey- oder S-Regression bei intervallgestörten Daten tendenziell länger dauert. Möglicherweise benötigen die beiden entsprechenden Algorithmen in Anwesenheit von Intervallstörungen länger, bis das Regressionsergebnis konvergiert.
Die Einstellungen für `RobPer`, wie eingangs in diesem Kapitel bemerkt, wurden für die Simulationsstudien ohne Rücksicht auf die Rechenzeit gewählt. Doch auch bei der Nutzung anderer Einstellungen scheint es empfehlenswert, die Implementierung, die für diese Arbeit vor allem übersichtlich gehalten wurde, effizienter zu gestalten. Für die Anpassung von Stufenfunktionen ist dies für M-Regression leicht umzusetzen, indem nicht eine Stufenfunktion mit m Stufen, sondern stattdessen pro Stufe ein Lokationsschätzer berechnet wird. Für komplexere Regressionstechniken wie LTS-, S- und τ-Regression ist es leider nicht möglich, jede Stufe einzeln anzupassen. Hier könnte eine Auslagerung der Implementierung in C-Code nützlich sein.

Zusammenfassung und Fazit

Der Einsatz robuster Regressionstechniken bei der Periodogrammberechnung ist im Falle einer Intervallstörung dem KQ-Periodogramm vorzuziehen. Die Periodogrammberechnungen dauern allerdings deutlich länger, vor allem für anzupassende Funktionen mit vielen Parametern wie den Einfach- und Doppelstufenfunktionen. Zur Detektion hat sich der Auswertungstyp 1 bewährt, speziell wenn die gewählte Detektionsmethode das Niveau einhalten soll. Bei diesem Auswertungstyp wird das Periodogramm vor Anpassung einer Betaverteilung nicht ausgedünnt. Von der Nutzung der Messfehler s_i mittels gewichteter Regression ist abzuraten. Sie führen in den Simulationen nur bei τ-Anpassung einer Stufenfunktion zu einer geringfügigen Verbesserung und das nur, wenn sie ausreißerfrei und perfekt die Standardabweichung des in den Messwerten befindlichen Rauschens wiedergeben.

Als geeignete periodische Funktion erweisen sich vor allem die Fouriersumme dritten Grades, die zu sehr guten Niveaueigenschaften führt, und die Stufenfunktion, die in den anspuchsvollsten Szenarien die besten Ergebnisse erzielt. Die Doppelstufenfunktion erreicht gegenüber der Einfachstufenfunktion keine Vorteile.
Die Wahl der robusten Regressionstechnik ist umso entscheidender, je weniger Beobachtungen der Lichtkurve maßgeblich von der periodischen Fluktuation beeinflusst werden. Bei einer Peakfunktion der Länge $p_f = 14$ können nur L1-, τ-, M-Huber- und M-Tukey-Regression mit geeigneten Modellen beste Detektionen erreichen. Bei einer Peakfunktion der Länge $p_f = 33$ sind die auf robuster Regression basierenden Methoden zu robust, um die Fluktuationsperiode zu detektieren. Die besten Ergebnisse, die allerdings mit denen der KQ-Regression vergleichbar sind, werden mit M-Huber- oder L1-Anpassung einer Einfachstufenfunktion erreicht. Dies ist insofern nachvollziehbar, als dies von den verwendeten Regressionstechniken genau jene sind, bei denen der Einfluss jeder Beobachtung monoton ins Zielkriterium einfließt (vgl. Abschnitt 2.4). Die zum Peak gehörenden Messwerte können so nicht vollständig ignoriert werden.
Da mit der M-Huber-Anpassung einer Stufenfunktion in den Simulationen auch stets das Niveau eingehalten wird, ist diese Methode zu empfehlen. Sieht man von der periodischen Peakfluktuation der Länge $p_f = 33$ ab, zeigen auch die τ-Stufenanpassung und die L1- oder M-Huber-Anpassung einer Fouriersumme dritten Grades gute Eigenschaften sowohl in der Periodendetektion als auch in der Einhaltung des Signifikanzniveaus.

	step	2step	sin	fourier(2)	fourier(3)	splines
L2	0.552	1.136	0.351	0.399	0.452	0.473
	0.690	1.381	0.402	0.473	0.548	0.556
	0.690	1.385	0.391	0.455	0.558	0.539
	1.188	2.415	0.815	0.856	1.081	1.029
L1	0.694	1.392	0.433	0.497	0.572	0.567
	0.848	1.700	0.509	0.598	0.706	0.669
	0.840	1.699	0.494	0.585	0.712	0.669
	1.488	2.968	0.947	1.065	1.241	1.216
LTS	303.769	609.877	167.025	217.324	262.052	194.015
	347.519	695.564	188.805	242.777	298.391	217.601
	347.457	694.487	188.726	242.606	298.539	217.489
	660.957	18680.812	345.269	436.665	534.355	392.529
S	96.557	194.724	68.357	75.590	81.426	72.746
	273.371	327.621	119.262	129.122	138.196	124.653
	302.797	373.215	131.688	143.953	154.382	137.904
	756.836	763.465	268.499	293.955	303.664	282.080
τ	218.934	437.948	178.402	195.050	218.262	194.289
	381.217	758.964	340.375	337.357	340.745	335.380
	374.495	745.513	326.705	316.211	321.165	318.261
	881.459	1680.973	885.868	901.330	800.745	820.959
MH	7.159	14.983	3.997	4.741	5.325	5.344
	12.821	25.671	7.836	9.249	10.974	9.227
	11.185	22.372	6.636	7.684	8.949	8.087
	88.245	151.075	33.337	35.894	49.169	32.350
MT	127.644	256.502	63.029	79.407	96.794	77.667
	256.700	512.622	131.841	179.607	224.948	120.265
	172.593	345.445	88.244	115.640	142.616	101.685
	995.090	2186.316	641.397	735.681	906.810	639.523

Tabelle 4.3.: Rechenzeit in Sekunden je Periodogramm für eine Lichtkurve mit 750 Beobachtungen und 100 Testperioden nach Regressionstechnik (Zeilen) und anzupassendes Modell (Spalten). Die vertikal aufeinander folgende Zahlen geben Minimum, arithmetisches Mittel, Median und Maximum an. Nach den in Abbildung 4.1 gezeigten Parametern sowie Gewichtung/Nichtgewichtung wird nicht unterschieden. Differenziertere Aufstellungen zur Rechenzeit sind in Anhang I.2 zu finden.

5. Anwendungsbeispiele

In diesem Kapitel werden die entwickelten Periodogramm- und Detektionsmethoden auf reale Lichtkurvendaten angewendet. Dabei werden nur noch die Methoden verwendet, die in der Simulationsstudie in Kapitel 4 erfolgreich waren. Für jeden hier vorgestellten Datensatz werden die folgenden neun Periodogramme berechnet:

- Anpassung einer Stufenfunktion mittels KQ-, τ- und M-Huber-Regression,
- Anpassung einer Sinusfunktion mittels KQ-, L1- und M-Huber-Regression,
- Anpassung einer Fouriersumme dritten Grades mittels KQ-, L1- und M-Huber-Regression.

Die Regressionstechnik wird jeweils ungewichtet angewendet. Die Periodogramme werden mit dem R-Paket `RobPer` (vgl. Kapitel 3) mit den in Kapitel 4 verwendeten Grundeinstellungen berechnet (vgl. Seite 53). Als Testperiodenmenge wird die Menge $\{1, 2, \ldots, \widetilde{T} - 1, \widetilde{T}\}$ betrachtet, wobei $\widetilde{T}$ eine natürliche Zahl ist und ungefähr einem Zehntel der Gesamtdauer der Lichtkurve entspricht. Die Empfehlung, Perioden zu untersuchen, für die mindestens zehn Zyklen beobachtet wurden, stammt von Halpern, Leighly und Marshall (2003). Die gewählte Auflösung, also der Abstand 1 zwischen den Testperioden, ist willkürlich gewählt und von der im Datensatz gewählten Zeiteinheit abhängig. Nachdem andere Auswertungstypen in der Simulationsstudie nicht überzeugen konnten (vgl. Kapitel 4), wird eine Betaverteilung zur Bestimmung auffälliger Perioden an alle Periodogrammbalken angepasst (Auswertungstyp 1). Es gelten solche Perioden als detektiert, für die der entsprechende Periodogrammbalken über dem $\sqrt[q]{0.95}$-Quantil der angepassten Betaverteilung liegt, wobei q die Anzahl der Testperioden beschreibt.
Insgesamt wurden im Rahmen dieser Arbeit mehr reale Datensätze untersucht als in diesem Kapitel vorgestellt werden. Bei vielen der hier nicht besprochenen Analysen werden ähnliche Effekte wie in den gezeigten beobachtet. Bei anderen Analysen kommen alle Methoden zu vergleichbaren Detektionsergebnissen. Bei wieder anderen unterscheiden sich die Ergebnisse sehr stark, aufgrund fehlender Informationen über das zugrunde liegende wahre Datenmodell kann aber nicht entschieden werden, welche Methode bessere Ergebnisse erzielt.
Um den Umfang der vorliegenden Arbeit angemessen zu gestalten, werden hier nur exemplarisch einige der Analysen besprochen. In Abschnitt 5.1 werden die Ergebnisse der Makarian-Lichtkurven diskutiert, die als Motivation für diese Arbeit dienen (vgl. Abschnitt 2.1). Abschnitt 5.2 beschäftigt sich mit Lichtkurven mit ausgewiesenen Ausreißern, anhand derer sich die Vorteile von robuster Regression bei der Peridodogrammberechnung veranschaulichen lassen. In Abschnitt 5.3 werden Lichtkurven mit geringem Rauschanteil behandelt, deren periodische Fluktuationen aufgrund eines kompexeren Verlaufs dennoch

schwer zu detektieren sind. Anhand der Lichtkurven in Abschnitt 5.4 lassen sich vor allem die regressionsabhängige Reaktion der Periodogrammmethoden auf Intervallstörungen und die erhöhte Abhängigkeit der Periodogrammbalken bei einer zu hohen Parameterzahl der anzupassenden Funktion beobachten. In Abschnitt 5.5 werden die Periodogrammmethoden auf paläoklimatische Daten angewendet und unterschiedliches Verhalten der verschiedenen robusten Regressionstechniken festgestellt.
In diesem Kapitel muss der Signifikanzbegriff noch umsichtiger behandelt werden als bisher: Es wird eine Vielzahl von Periodogrammen berechnet und auf auffällige Perioden untersucht. Viele dieser Periodogramme sind untereinander abhängig, da sie sich auf die gleichen Daten beziehen. Zudem ist für die vorliegenden Daten nicht bekannt, inwieweit sie die Bedingungen erfüllen, unter denen die Methoden in Kapitel 4 auf ihre Fähigkeit zur Niveaueinhaltung überprüft wurden. Von einer signifikanten Detektion darf in diesem Kapitel somit nicht die Rede sein. Es wird stattdessen immer von einer detektierten Periode gesprochen.
Alle hier verwendeten Daten stammen aus dem Internet. Alle astrophysikalischen Lichtkurven (Abschnitte 5.1 bis 5.4) weisen auch Messfehler s_i auf, die aufgrund der Ergebnisse der Simulationsstudie in Kapitel 4 nicht berücksichtigt werden. Die Objekte, von denen Lichtkurven beobachtet werden können, werden Quellen genannt. Alle Abbildungen eines Abschnitts befinden sich jeweils am Ende desselben.

5.1. Photonenemissionen: Makarian-Blazare

In diesem Abschnitt werden die bereits in Abschnitt 2.1 vorgestellten Lichtkurven zu den Quellen Mrk 421 und Mrk 501[3] analysiert. Die Lichtkurve zu Mrk 421 umfasst 655 Beobachtungen in einem Zeitintervall der Länge 5 920 Tage. Die Lichtkurve zu Mrk 501 besteht aus 210 Beobachtungen und dauert 4 085 Tage. Die Zeiteinheit der Messzeiten sind Tage, angegeben als Modified Julian Date (MJD), Tage seit dem 17. November 1858. Gemessen wird die Gammaemission in Crab Units (relativ zu den Emissionen des Krebsnebels).
Bei der Lichtkurve Mrk 421 (Abbildung 2.1(a) auf Seite 6) wird keine Periode detektiert. Abbildung 5.1 zeigt eine Auswahl der berechneten Periodogramme. Auffällig ist, dass bei Anpassung einer Sinusfunktion unter den ersten 100 Testperioden die Periode 31 lokal hervorsticht (vgl. Abbildung 5.1(a) für KQ-Regression, L1- und M-Huber-Regression ähnlich), bei Verwendung eines anderen Modells (Abbildungen 5.1(c) und (e) für KQ- und τ-Anpassung einer Stufenfunktion) oder bei Betrachtung aller Testperioden ist der Balken allerdings nicht mehr auffällig.
Die auf KQ-Regression basierenden Periodogramme weisen einen in der Testperiode monoton steigenden Trend auf. Für robuste Periodogramme wachsen die Ausschläge mit der Testperiode. Dies kann ein Hinweis darauf sein, dass der Anteil der intervallgestörten Daten zu hoch für die Methoden ist. Zum Vergleich werden Periodogramme der Lichtkurve betrachtet, bei der die Beobachtungen des Zeitintervalls [51 872, 52 053] entfernt wurden, da diese augenscheinlich Teil einer Intervallstörung sein könnten. Ein Entfernen dieser Messungen

[3] Heruntergeladen am 30. Mai 2011 von:
http://nuastro-zeuthen.desy.de/magic_experiment/projects/light_curve_archive/index_eng.html

(ca. ein Viertel aller Beobachtungen) aus der Lichtkurve reduziert den Effekt der monoton steigenden Periodogrammwerte und des erhöhten Periodogrammbalkens bei 31, eine Periode wird aber weiterhin [51872, 52053] nicht detektiert (vgl. auch Abbildungen 5.1(b), (d) und (f)).
Bei der Lichtkurve Mrk 501 (vgl. Abbildung 2.1(b) auf Seite 6) wird ebenfalls keine Periode detektiert. Abbildung 5.2 zeigt eine Auswahl der für diese Reihe berechneten Periodogramme. Auffällig sind die erhöhten Periodogrammbalken zwischen 200 und 300. Dieser Effekt tritt außer beim τ-Stufenperiodogramm (Abbildung 5.2(d)) bei allen, auch den nicht abgebildeten Periodogrammen, auf. Auf der Anpassung einer Sinusfunktion basierende Periodogramme (Abbildungen 5.2(a), (c) und (e)) weisen zudem erhöhte Balken im Bereich um Testperiode 30 auf, wobei der Balken bei 30 selbst lokal minimal ist.
Die Betrachtung einiger Phasendiagramme (vgl. Beispiele in Abbildung 5.3) zeigt, dass die umgeklappten Messzeiten für Perioden zwischen 200 und 300 gleichmäßig über den Zyklus verteilt sind und die erhöhten Werte zugleich dicht beieinander liegen. Eine wahre Periode scheint hierfür nicht der Grund zu sein, denn die hohen Messwerte stammen alle aus dem gleichen Zyklus und nah in der Phase liegende Messwerte anderer Zyklen sind niedriger. Bei Entfernen der relativ hohen Messwerte aus dem Zeitraum [50 523, 50 643] verschwindet sowohl die Erhebung für Testperioden zwischen 200 und 300 als auch die Erhöhung um Testperiode 30 (vgl. Beispiele in Abbildung 5.4). Dabei fällt auf, dass das τ-Periodogramm (vgl. Abbildungen 5.2(d) und 5.4(d)) sich in Gestalt und Skalierung am wenigsten verändert. Die beiden lokalen Maxima bei 117 und 176 bleiben bestehen. Sie tauchen nach Entfernung der Intervallstörung nun teilweise auch in anderen Periodogrammen auf, insbesondere auch in den hier nicht gezeigten Periodogrammen mit Anpassung einer Fouriersumme dritten Grades, werden jedoch nie detektiert. Die Angleichung der anderen Periodogramme an das τ-Periodogramm bei Auslassung vermeintlicher Ausreißer spricht für die hohe Robustheit desselben.

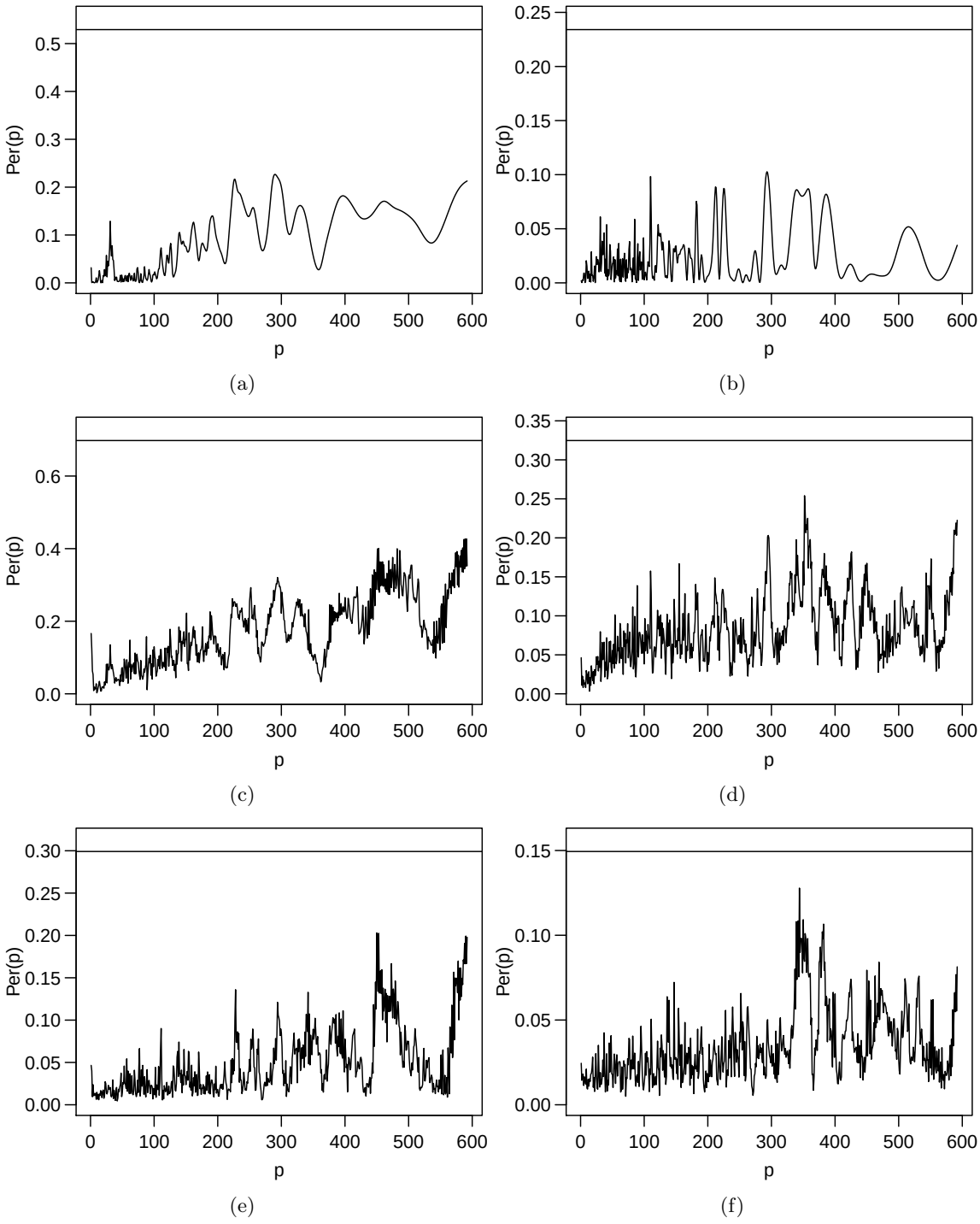


Abbildung 5.1.: Periodogramme der Lichtkurve von Mrk 421 (vgl. Abbildung 2.1(a) auf Seite 6). Angepasste Funktion: (a)–(b) Sinusfunktion, (c)–(f) Einfachstufenfunktion. Regressionstechnik: (a)–(d) KQ, (e)–(f) τ. Lichtkurve: (links) unverändert, (rechts) ohne Beobachtungen aus dem Intervall [51 872, 52 053]. Die horizontale Linie markiert den jeweiligen Schwellwert für Detektionen.

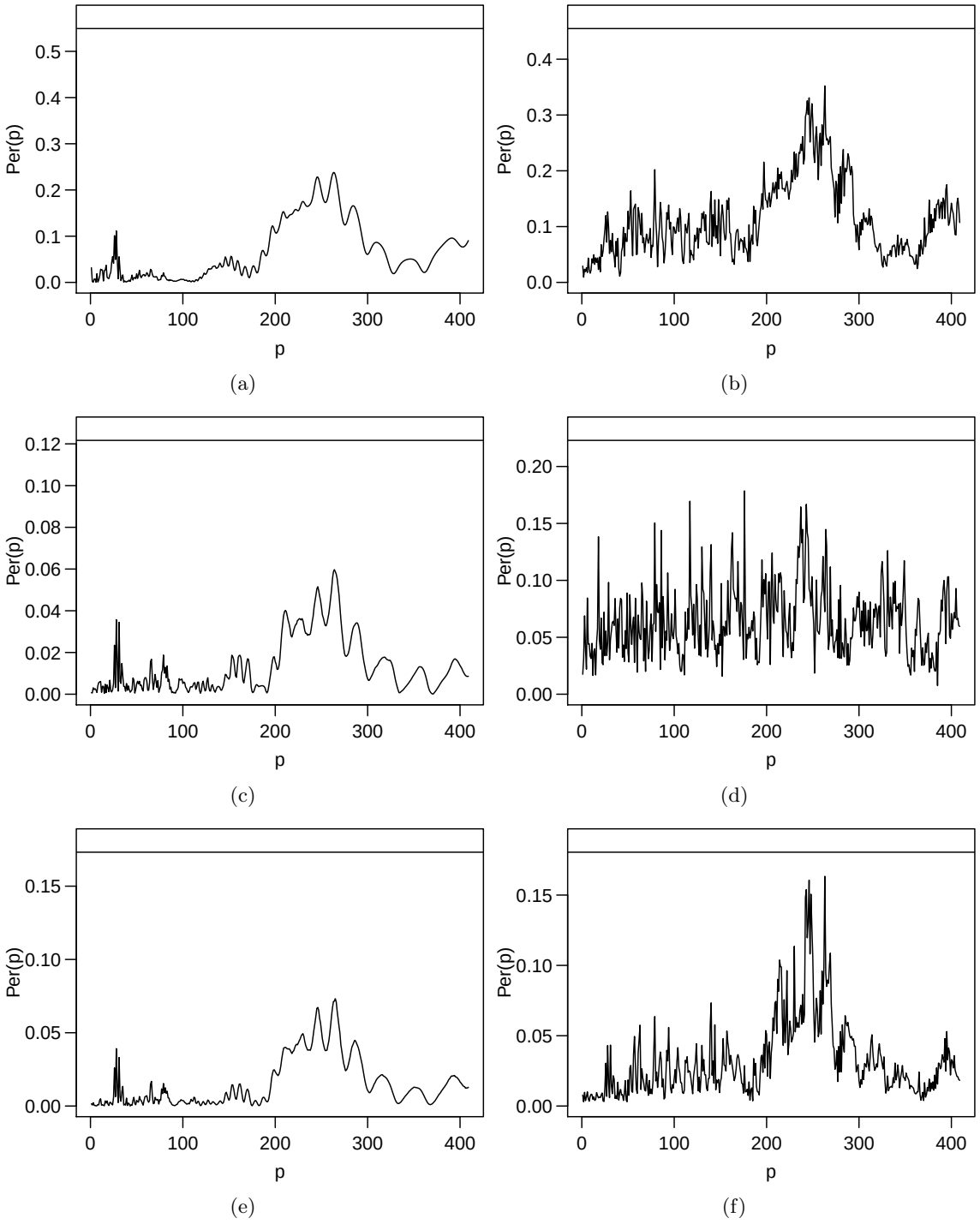
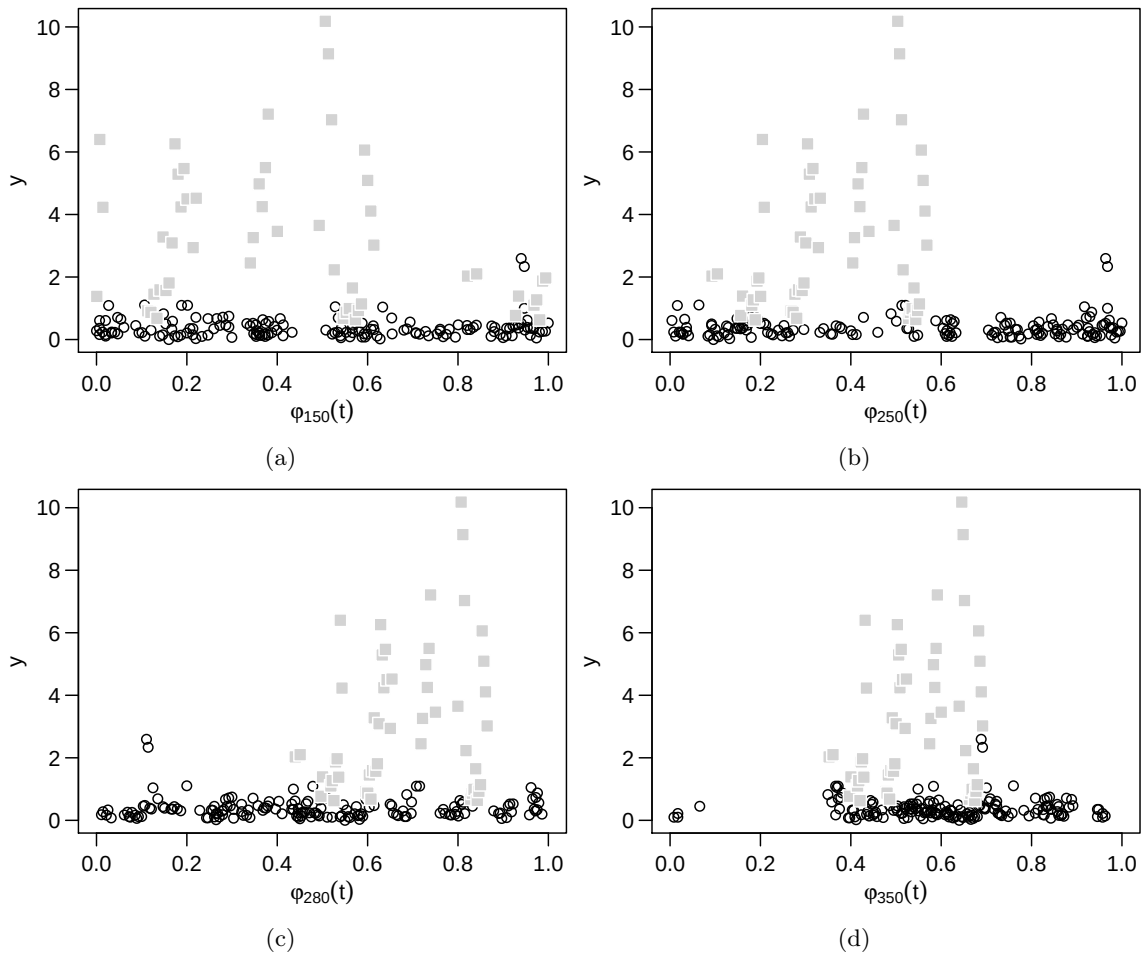


Abbildung 5.2.: Periodogramme der Lichtkurve von Mrk 501 (vgl. Abbildung 2.1(b) auf Seite 6). Angepasste Funktion: (links) Sinusfunktion, (rechts) Einfachstufenfunktion. Regressionstechnik: (oben) KQ, (c) L1, (d) τ, (unten) M-Huber. Die horizontale Linie markiert den jeweiligen Schwellwert für Detektionen.



(a) (b) (c) (d)

Abbildung 5.3.: Phasendiagramme der Lichtkurve von Mrk 501. Perioden: (a) 150, (b) 250, (c) 280 und (d) 350. Zu einer vermeintlichen Intervallstörung gehörende Beobachtungen sind grau hervorgehoben.

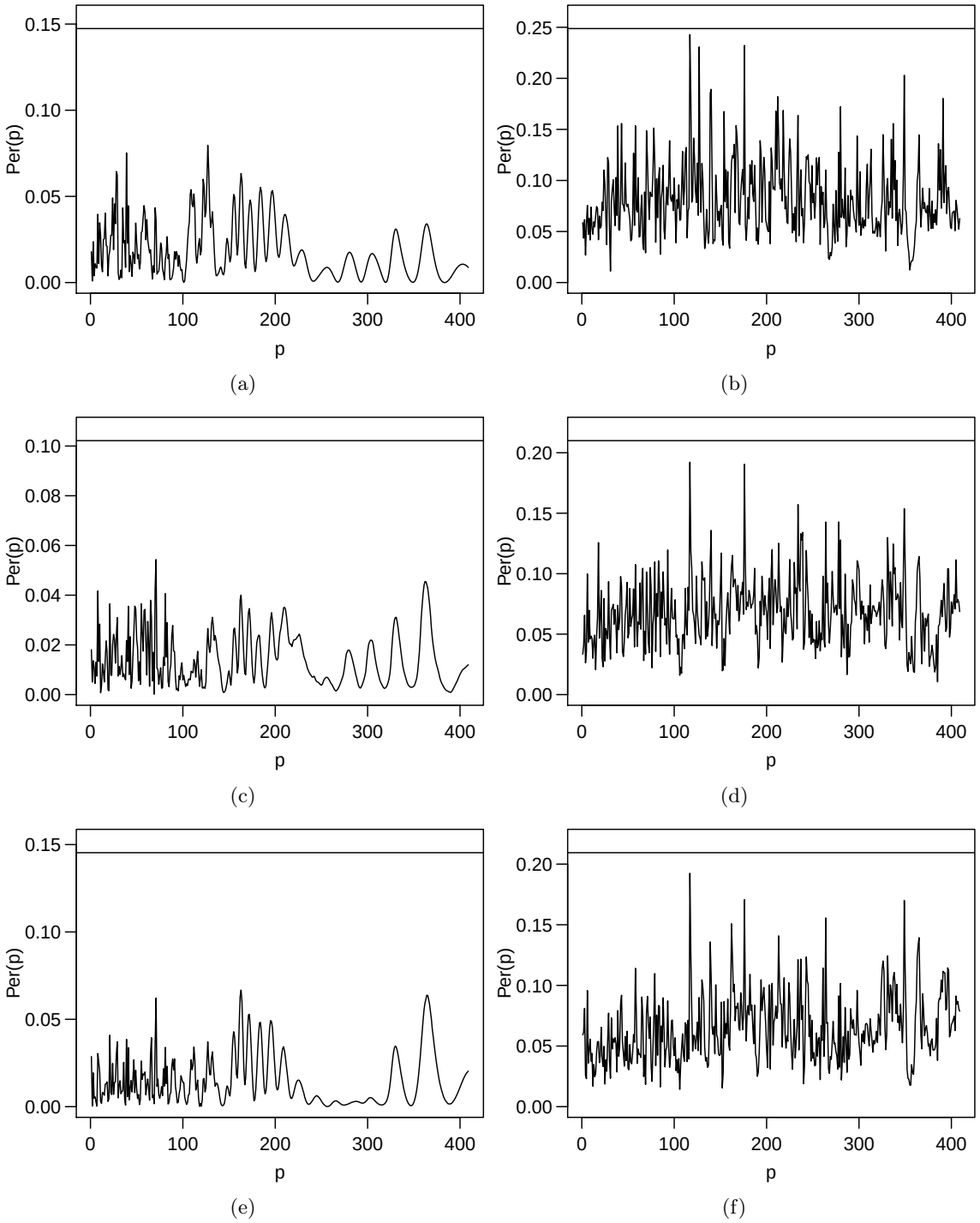


Abbildung 5.4.: Periodogramme der Lichtkurve von Mrk 501 ohne die Beobachtungen im Intervall [50 523, 50 643]. Angepasste Funktion: (links) Sinusfunktion, (rechts) Einfachstufenfunktion. Regressionstechnik: (oben) KQ, (c) L1, (d) τ, (unten) M-Huber. Die horizontale Linie markiert den jeweiligen Schwellwert für Detektionen.

5.2. Sichtbares Licht: Der All Sky Automated Survey

In diesem Abschnitt werden Lichtkurvendaten vom All Sky Automated Survey (ASAS, vgl. Pojmanski 1997) betrachtet[4]. Ziel dieses Projekts ist es, die Helligkeit von von der Erde aus sichtbaren Sterne, Kometen und Asteroiden mit auf der Erde verteilten Teleskopen zu messen. Die Koordinaten (RA , DEC) dieser Objekte sind dabei in Rektaszension (RA) und Deklination (DEC), den in der Astronomie üblichen Polarkoordinaten, angegeben. Beide Koordinaten werden als Zeitangabe angegeben, RA auf der Skala h:m:s, DEC auf der Skala d:m, wobei d für Tag steht, h für Stunde, m für Minute und s für Sekunde. Für diese Arbeit werden Objekte mit Rektaszension zwischen 12:00:00 und 13:00:00 betrachtet. Die Messzeiten werden in heliozentrischen Tagen angegeben (HJD), der entsprechende Kalender (Heliocentric Julian Date) und ist nicht linear in einen auf der Erde gültigen Kalender übertragbar. Gemessen wird die Helligkeit in dem in der Astronomie typischen Helligkeitsmaß Magnitude. Mehr Details zum Messverfahren sind bei Pojmanski und Maciejewski (2004) zu finden.

Zu jeder veröffentlichten Lichtkurve wird eine detektierte Periode (im Folgenden Referenzperiode) angegeben. Es kann davon ausgegangen werden, dass die Referenzperiode durch eine Analyse der Daten gewonnen wurde. Es ist jedoch unklar, wie diese geschätzt wurde und inwiefern sie signifikant ist. Eine Bestätigung der Referenzperiode mit den hier verwendeten Methoden darf daher nicht als Bestätigung der astroteilchenphysikalischen Theorie angesehen werden.

Über die abgesuchte Testperiodenmenge zur Detektion der Referenzperiode ist nichts bekannt. In den im Rahmen dieser Arbeit durchgeführten Analysen der ASAS-Lichtkurven werden häufig Perioden detektiert, die leicht von der jeweiligen Referenzperiode abweichen. Diese Abweichung kann mit unterschiedlichen Testperiodenmengen erklärbar sein. Ist sie es nicht, ist ohne weitere Hintergrundinformationen keine Aussage darüber machbar, ob die Ziel- oder die hier detektierte Periode eher der Wahrheit entspricht.

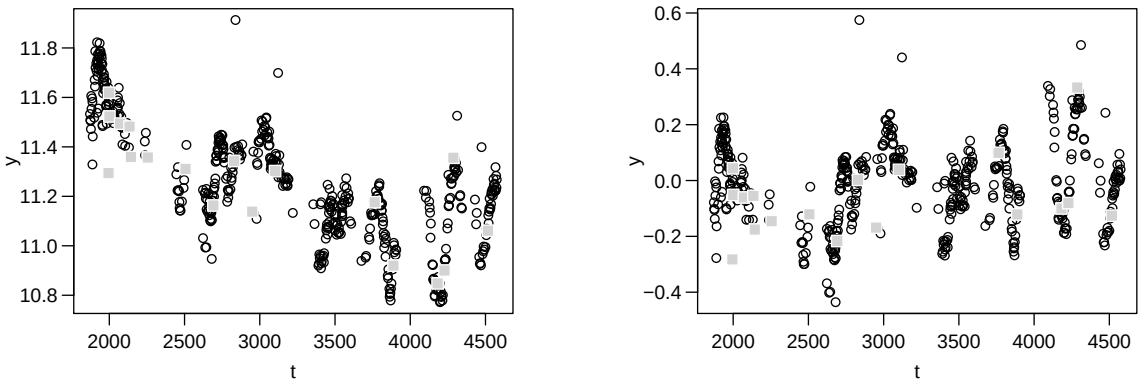
Einige der Messpunkte der jeweiligen ASAS-Lichtkurve werden als unbrauchbar eingestuft (vgl. Graham et al. 2013). Diese Messpunkte werden im Folgenden als bekannte Ausreißer behandelt. Der Grund für eine solche Einstufung ist nicht bekannt und vermutlich auf Hardware-Ebene zu suchen. Die Einstufung als unbrauchbar ist sind im Datensatz in einer zusätzlichen Indikatorvariable dokumentiert. Dies schafft die Möglichkeit, die Analyse der unbereinigten Lichtkurve mit der der ausreißerbereinigten Lichtkurve zu vergleichen. Bei Betrachtung der Phasendiagramme zu Perioden, die in der bereinigten Lichtkurve gefunden werden, zeigt sich jedoch, dass auch einige der als unbrauchbar eingestuften Beobachtungen dem jeweiligen periodischen Verlauf folgen und damit zur Detektion der Periode hätten beitragen können. Neben dem Wunsch, auch robust bezüglich unbekannter Ausreißer zu analysieren, motiviert dies zusätzlich, robuste Methoden auf der unbereinigten anstatt KQ-Methoden auf der bereinigten Lichtkurve zu verwenden.

Als erste Lichtkurve aus dem ASAS-Projekt wird die zum Stern mit den Koordinaten (12:00:09 / 67:52,8) betrachtet. Sie umfasst 525 Beobachtungen auf einem Intervall von 2702 heliozentrischen Tagen, wobei 34 Beobachtungen als Ausreißer eingestuft sind. Bei

[4]Heruntergeladen am 30. September 2013 von: http://www.astrouw.edu.pl/asas/?page=download

Betrachtung der Lichtkurve (Abbildung 5.5(a)) fällt auf, dass ein Trend vorzuliegen scheint, der die Periodendetektion erschweren kann. Deshalb wird eine einfache Trendbereinigung durchgeführt, bei der jeweils eine Gerade an die Beobachtungen vor Messzeit 4000 und nach Messzeit 4000 angepasst und von den Daten abgezogen wird. Die so trendbereinigte Zeitreihe wird in Abbildung 5.5(b) gezeigt.
Abbildung 5.6 zeigt Periodogramme für die nicht trendbereinigte Lichtkurve, sowohl in ihrem ursprünglichen Zustand als auch unter Auslassung der ausgewiesenen Ausreißer. Angepasst wurde jeweils eine Stufenfunktion unter Verwendung verschiedener Regressionstechniken. Das gezeigte KQ-Periodogramm für die ausreißerbereinigte Lichtkurve (Abbildung 5.6(b)) und die auf robuster Regression basierenden Periodogramme für die Lichtkurve vor und nach der Ausreißerbereinigung (Abbildungen 5.6(c) bis 5.6(f)) sehen einander ähnlich. Die als Ausreißer identifizierten Beobachtungen üben bei den auf robuster Regression basierenden Methoden also keinen starken Einfluss auf die Gestalt des Periodogramms aus. Dagegen hat das für die nichtbereinigte Lichtkurve berechnete KQ-Periodogramm (vgl. Abbildung 5.6(a)) eine völlig andere Gestalt. Zur Periodogrammberechnung sollte hier also entweder robuste Regression eingesetzt werden oder die Ausreißer müssen im Vorhinein identifiziert und entfernt werden (vgl. Abbildung 5.6(b)).
In den ausreißerbereinigten Daten sind erhöhte Periodogrammbalken bei ca. 160 und 260 auffällig, die aber nur für M-Huber-Regression (Abbildung 5.6(f)) zur Detektion führen. Ohne Ausreißerbereinigung wird durch M-Huber-Regression nur die Periode bei 260 detektiert (vgl. Abbildung 5.6(e)). Die τ-Regression detektiert nur in der nicht bereinigten Lichtkurve eine Periode von 260 (vgl. Abbildung 5.6(c)). Bei M-Huber-Anpassung einer Fouriersumme dritten Grades (hier nicht gezeigt) können beide Perioden sowohl mit als auch ohne Ausreißerbereinigung detektiert werden.
Bei Betrachtung der Skalierungen der Periodogramme in den Abbildungen 5.6(b)–(f) fällt auf, dass diese sich trotz eines ähnlichen Periodogrammverlaufs stark unterscheiden. Dies unterstreicht nochmal den Sinn einer adaptiven Schwellwertberechnung. Für τ-Regression ändert sich die Skalierung des Periodogramms bei Entfernung der Ausreißer am wenigsten. Dies konnte auch bei hier nicht gezeigten Periodogrammen beobachtet werden. Ein Erklärungsansatz hierfür ist, dass das für die τ-Regression genutzte Bestimmtheitsmaß mit robusten Varianzschätzern berechnet wird. Das Zielkriterium der robusten M-Regression dagegen führt zu einer robusten Parameterschätzung, ist selbst aber nicht robust.
Bei Anwendung der Periodogrammmethoden auf die trendbereinigte Lichtkurve ändert sich an der Gestalt der Periodogramme nur wenig (vgl. Abbildung 5.7). Allerdings wird keine Periode mehr detektiert. Eine möglicher Erklärungsansatz hierfür ist, dass in der vorverarbeiteten Lichtkurve ein höherer Anteil der Varianz in den Daten erklärbar ist, wodurch viele Periodogrammbalken steigen. An diese wird eine Betaverteilung mit höherem Erwartungswert angepasst, was zu höheren Quantilen für eine Detektion führt.
Abbildung 5.8 zeigt Phasendiagramme für die Referenzperiode 157.8947 und für die Periode 263, sowohl von der ausreißerbereinigten als auch von der ausreißer- und trendbereinigten Lichtkurve. Die Perioden scheinen gut zu passen. Eine Periode von 263 ist im vorliegenden Datensatz nicht dokumentiert. In der Abbildung kann nachvollzogen werden, dass einige der Beobachtungen trotz Kennzeichnung als Ausreißer gut zum Kurvenverlauf passen. Wie

eingangs erörtert, ist dies scheinbar für viele gekennzeichnete Ausreißer in ASAS-Lichtkurven der Fall.
Die zweite aus dem ASAS-Projekt betrachtete Lichtkurve wurde vom Stern mit den Koordinaten (12:35:48 / 73:21,8) aufgenommen. Sie umfasst 1319 Beobachtungen inklusive 168 Ausreißern und dauert 2705 heliozentrische Tage. Abbildung 5.9(a) zeigt die Lichtkurve. Alle verwendeten Methoden detektieren hier in den ausreißerbereinigten Daten eine Periode von ca. 156 (ungefähr zwei Tage kleiner als die Referenzperiode). Ohne Ausreißerbereinigung gelingt dies allen auf robuster Regression, aber keiner der auf Kleinste-Quadrate-Regression beruhenden Methoden. Abbildung 5.10 zeigt beispielhaft die mit verschiedenen Regressionstechniken gewonnenen Stufenperiodogramme für die Lichtkurve vor und nach Ausreißerbereinigung. Ein Problem mit der Periodendetektion bei Anpassung einer Funktion mit zu vielen Parametern zeigt sich bei der Lichtkurve zum Stern mit den Koordinaten (12:31:20 / 70:20,2) (Abbildung 5.11(a)). Diese Lichtkurve besteht aus 958 Beobachtungen inklusive 80 Ausreißern und dauert 2703 heliozentrische Tage. Abbildung 5.12 zeigt verschiedene Periodogramme der ausreißerbereinigten Lichtkurve, unter Anpassung einer Fouriersumme dritten Grades oder einer Sinusfunktion. In den auf KQ- und L1-Regression basierenden Sinusperiodogrammen (vgl. Abbildungen (a) und (c)) wird eine Periode bei ca. 100 detektiert (Referenzperiode: 102.2604). Bei Analyse der ausreißerbehafteten Lichtkurve (hier nicht gezeigt) wird diese Periode auch mit M-Huber-Regression detektiert, ist jedoch im KQ-Periodogramm nicht mehr auffällig hoch. Das Phasendiagramm in Abbildung 5.11(b) bestätigt einen sinusartigen Verlauf mit Periode 100. In den Fourier-Periodogrammen (vgl. Abbildungen 5.12(b), (d) und (f)) sind diese Effekte auch zu sehen. Dadurch, dass mit Testperiode $2p_f$ auch eine Sinusschwingung der Periode p_f angepasst werden kann (vgl. Abschnitt 2.3.4 und dort Abbildung 2.11), ist hier zusätzlich der Balken bei 200 erhöht. Die hohe Anzahl an Parametern führt zudem zu einer leidlichen Anpassung auch bei vielen falschen Perioden (overfitting). Diese Vielzahl an erhöhten Balken führt dazu, dass keiner als auffällig hoch detektiert wird. Der gleiche Effekt wird bei den berechneten Stufenperiodogrammen (hier nicht gezeigt) sichtbar.



(a) (b)

Abbildung 5.5.: Lichtkurve des Sterns mit Koordinaten (12:00:09 / 67:52,8). Messzeiten in HJD. Messwerte in Magnituden. Neben 18 hervorgehobenen Ausreißern (grau) befinden sich 16 weitere Ausreißer außerhalb des gezeigten y-Achsenabschnitts. Lichtkurve: (a) unverändert, nicht gezeigte Ausreißer bei 30 und 100, (b) trendbereinigt, nicht gezeigte Ausreißer zwischen 18 und 19 bzw. 88 und 89.

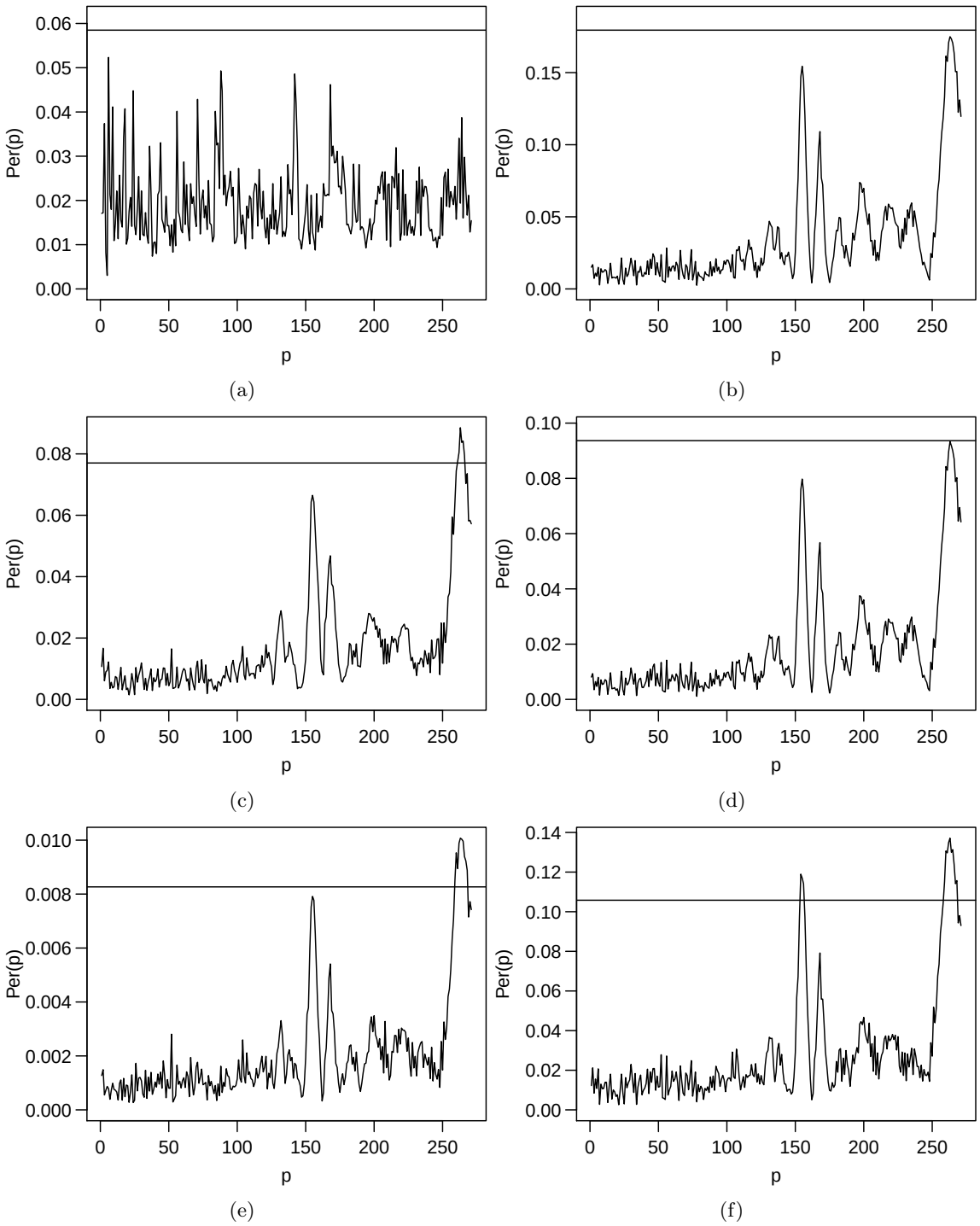


Abbildung 5.6.: Periodogramme der Lichtkurve des Stern mit Koordinaten (12:00:09 / 67:52,8). Angepasste Funktion: Einfachstufenfunktion. Lichtkurve: (links) unverändert, (rechts) ausreißerbereinigt. Regressionstechnik: (oben) KQ, (mittig) τ, (unten) M-Huber. Die horizontale Linie markiert den jeweiligen Schwellwert für Detektionen.

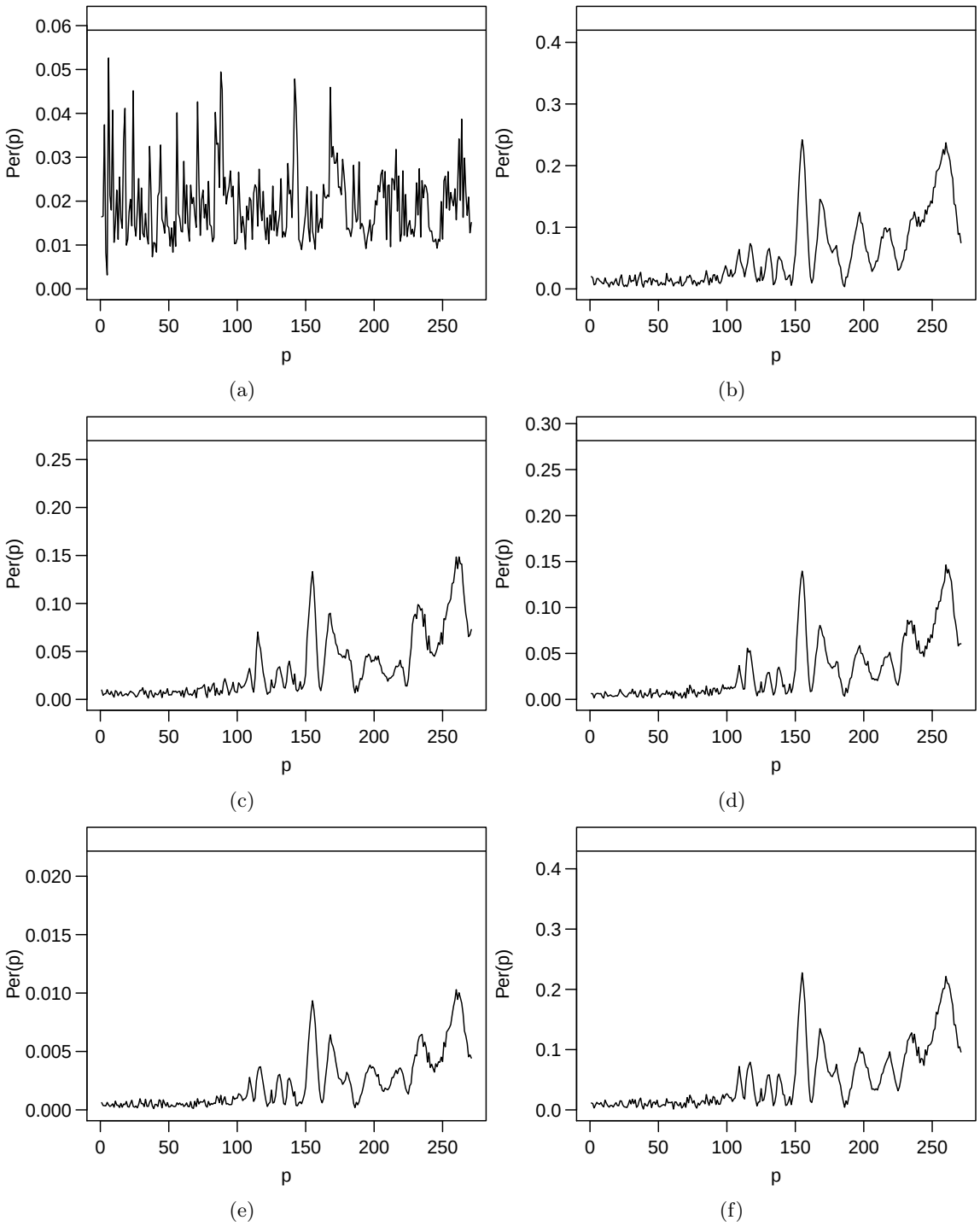
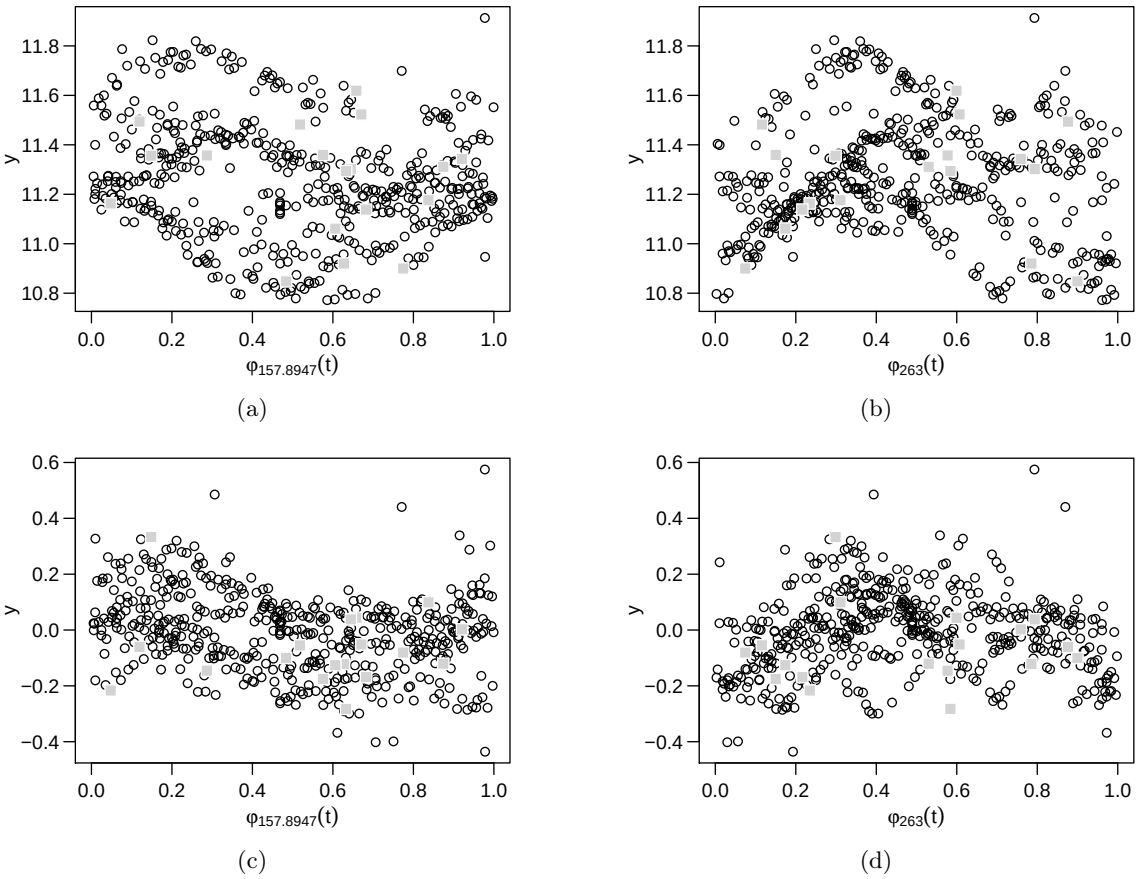
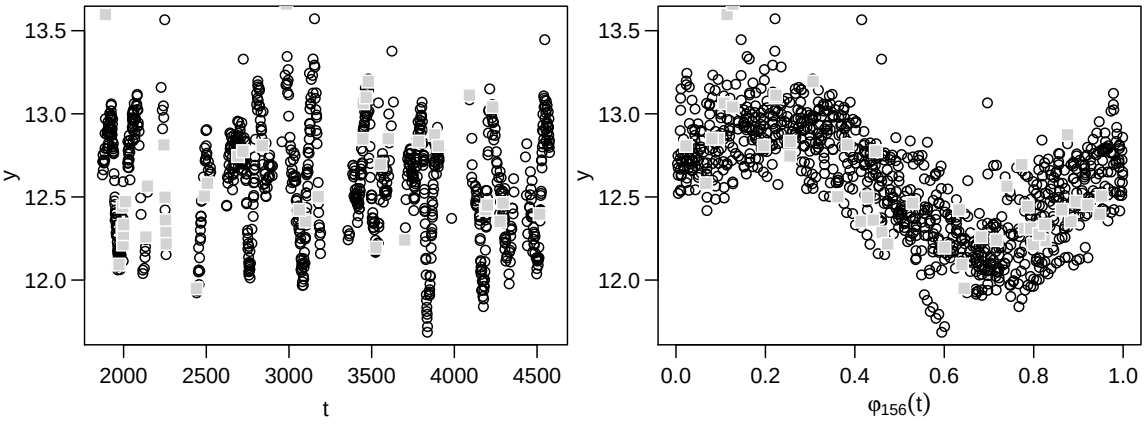


Abbildung 5.7.: Periodogramme der trendbereinigten Lichtkurve des Sterns mit Koordinaten (12:00:09 / 67:52,8). Angepasste Funktion: Einfachstufenfunktion. Lichtkurve: (links) unverändert, (rechts) ausreißerbereinigt. Regressionstechnik: (oben) KQ, (mittig) τ, (unten) M-Huber. Die horizontale Linie markiert den jeweiligen Schwellwert für Detektionen.



(a) (b) (c) (d)

Abbildung 5.8.: Phasendiagramme der Lichtkurven des Sterns mit Koordinaten (12:00:09 / 67:52,8). Lichtkurve: (oben) unverändert, (unten) trendbereinigt. Testperiode: (links) Referenzperiode 157,8947, (rechts) 263. Graue Hervorhebung wie in Abbildung 5.5.



(a) (b)

Abbildung 5.9.: Beobachtungen des Sterns mit Koordinaten (12:35:48 / 73:21,8). Messzeiten in HJD. Messwerte in Magnituden. Neben 44 hervorgehobenen Ausreißern (grau) befinden sich 31 zusätzliche Ausreißer außerhalb des gezeigten y-Achsenabschnitts (zwischen 13,6 und 100). Darstellung: (a) Lichtkurve, (b) Phasendiagramm zur Periode 156.

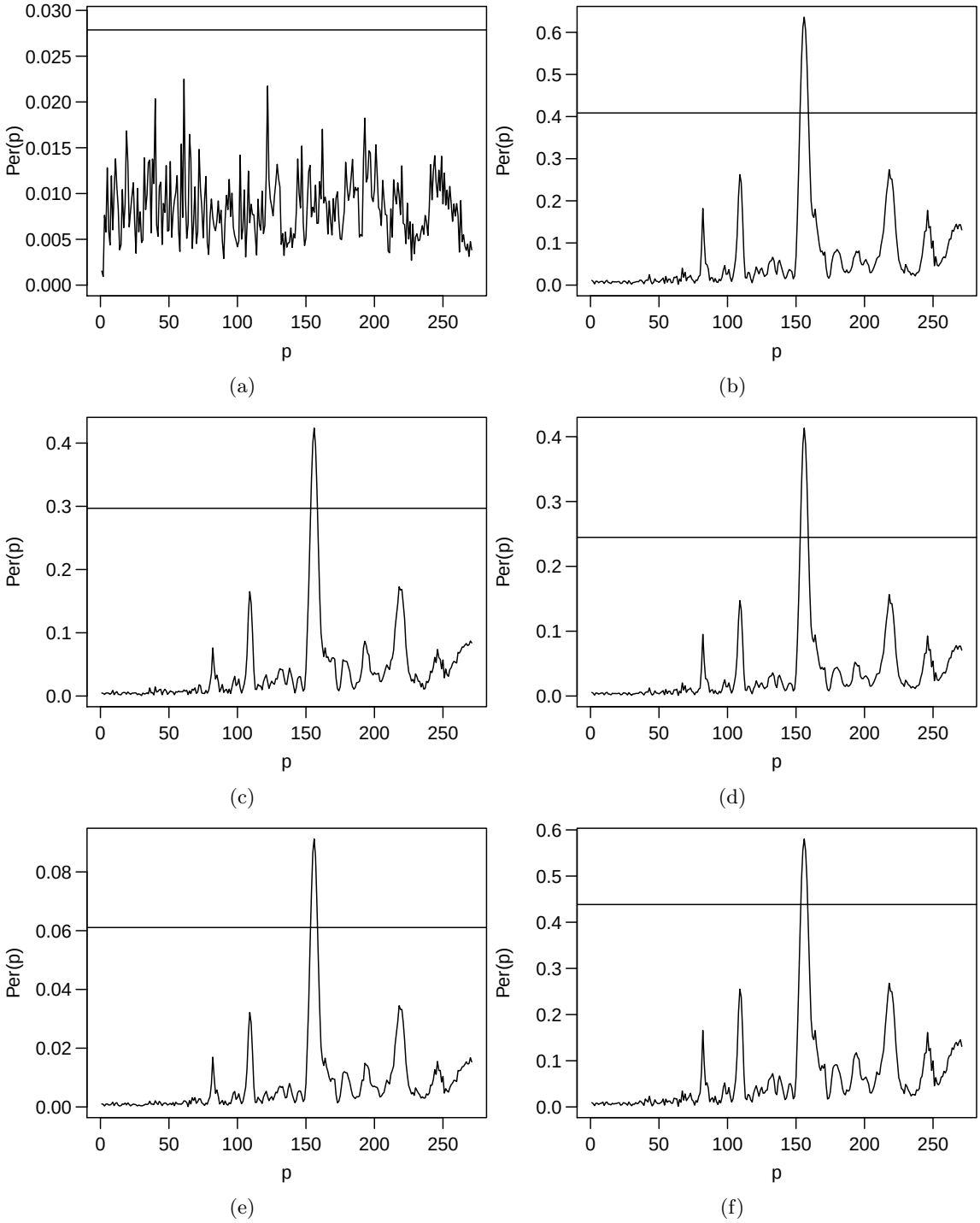
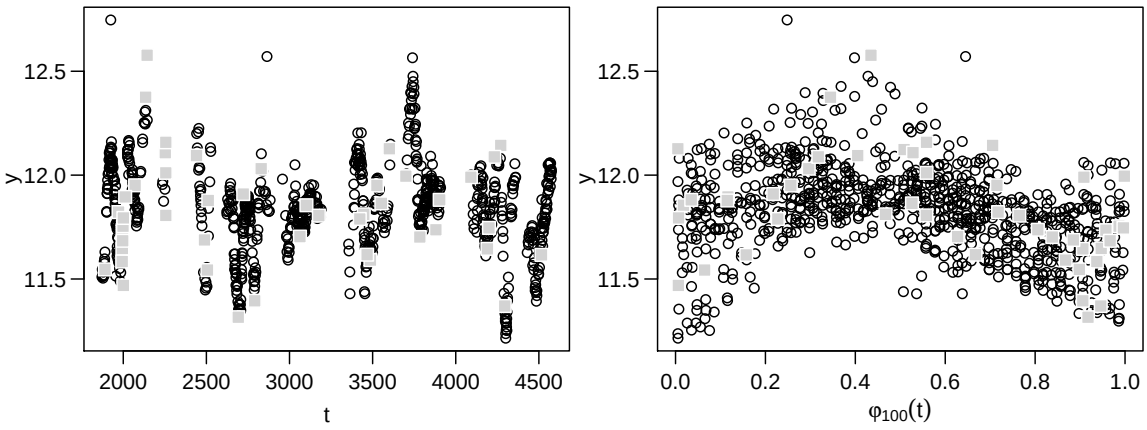


Abbildung 5.10.: Periodogramme der Lichtkurve des Sterns mit Koordinaten (12:35:48 / 73:21,8). Angepasste Funktion: Einfachstufenfunktion. Lichtkurve: (links) unverändert, (rechts) ausreißerbereinigt. Regressionstechnik: (oben) KQ, (mittig) τ, (unten) M-Huber. Die horizontale Linie markiert den jeweiligen Schwellwert für Detektionen.



(a) (b)

Abbildung 5.11.: Beobachtungen des Sterns mit Koordinaten (12:31:20 / 70:20,2). Messzeiten in HJD. Messwerte in Magnituden. Neben 47 hervorgehobenen Ausreißern (grau) befinden sich 33 weitere Ausreißer außerhalb des gezeigten y-Achsenabschnitts (bei 30 und 100). Darstellung: (a) Lichtkurve, (b) Phasendiagramm zur Periode 100.

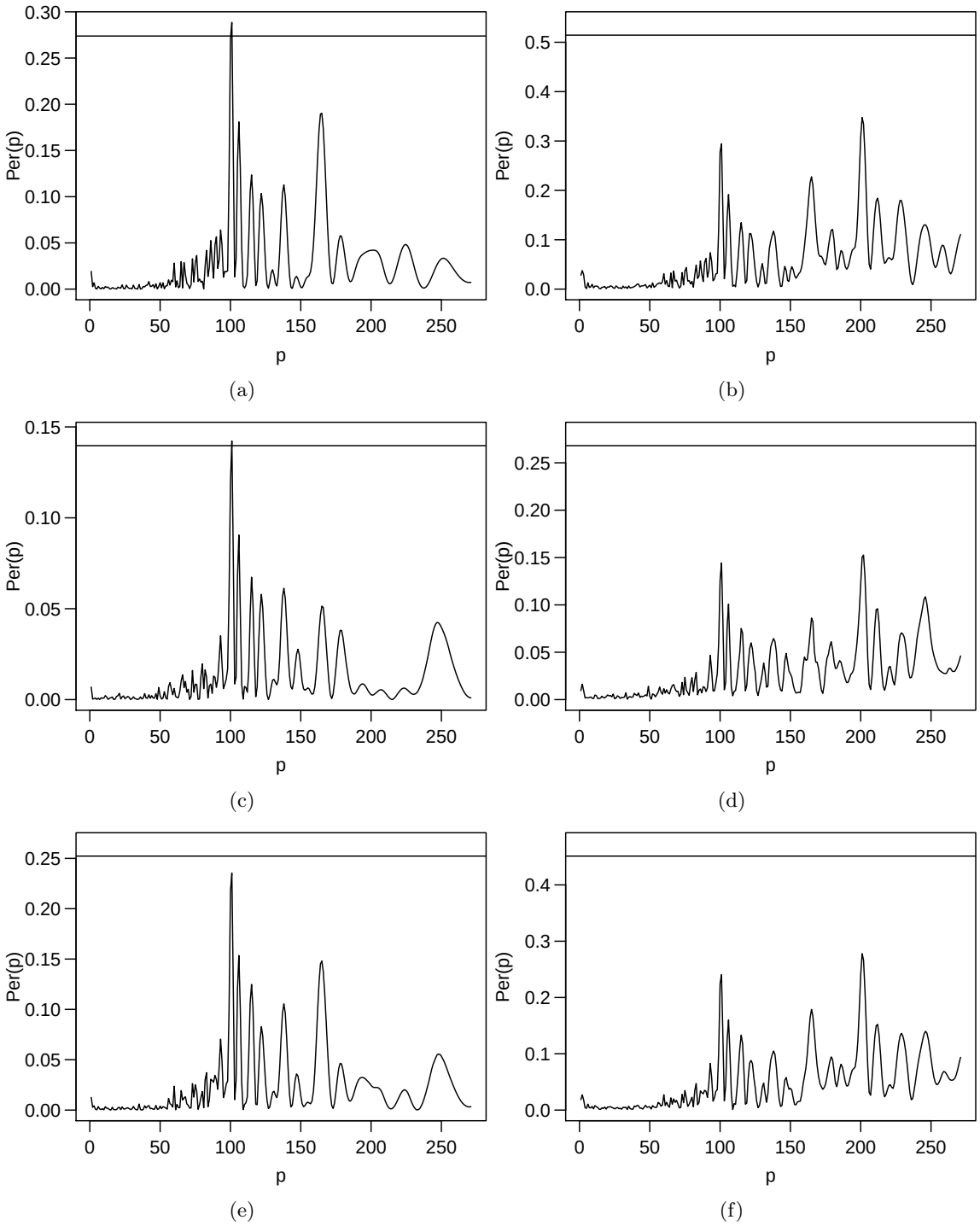


Abbildung 5.12.: Periodogramme der ausreißerbereinigten Lichtkurve des Sterns mit Koordinaten (12:31:20 / 70:20,2). Angepasste Funktion: (links) Sinusfunktion, (rechts) Fouriersumme dritten Grades. Regressionstechnik: (oben) KQ, (mittig) L1, (unten) M-Huber. Die horizontale Linie markiert den jeweiligen Schwellwert für Detektionen.

5.3. Sichtbares Licht: Catalina Survey Data Release

Die in diesem Abschnitt besprochenen Analysen beziehen sich auf Lichtkurven aus dem von Drake et al. (2009) veröffentlichten ersten Catalina Survey Data Release[5] (CSDR), die mit dem CSS Schmidt Teleskop aufgezeichnet wurden. Auch bei diesen Lichtkurven werden Referenzperioden angegeben. Sie liegen meist weit unter einem Tag und damit außerhalb der Testperiodenmenge, die in dieser Arbeit bisher verwendet wurde. Für die Daten des CSDR werden daher Periodogramme sowohl auf der üblichen ganzzahligen Testperiodenmenge als auch auf der Menge $\{0.1p^\star, 0.2p^\star, \ldots, 9.9p^\star, 10p^\star\}$ berechnet, wobei $p^\star$ die Referenzperiode ist.

In den meisten Analysen kann $p^\star$ bestätigt werden. Detektionen auf der ganzzahligen Testperiodenmenge tauchen nicht systematisch auf und können visuell im Phasendiagramm nicht bestätigt werden. Die betrachteten Daten scheinen keine starken Ausreißer aufzuweisen.

In einigen Lichtkurven kann eine besondere periodische Fluktuation mit interessantem Effekt auf die Periodogramme beobachtet werden. Als Beispiel für diesen Effekt wird im Folgenden die Lichtkurve zum Stern mit Catalina-ID 1001005030535721 (Abbildung 5.13) analysiert. Sie besteht aus 234 Beobachtungen und hat eine Dauer von 2736 Tagen.

Abbildung 5.14 zeigt Phasendiagramme für diese Lichtkurve zur Referenzperiode $p^\star = 0.67508$ und zur halben Referenzperiode $p^\star/2$. Die (auf eine Nachkommstelle gerundete) halbe Referenzperiode $p^\star/2$ ist jene Testperiode, welche von allen Methoden außer dem τ-Stufenperiodogramm detektiert wird (vgl. Abbildung 5.15 für Periodogramme Anpassung einer Einfachstufenfunktion oder einer Fouriersumme dritten Grades). In Abbildung 5.14(a) ist eine Fluktuation mit zwei lokalen Maxima (im Folgenden „periodisch bimodal") zu erkennen, während diese Maxima in Abbildung 5.14(b) übereinanderliegen (im Folgenden „periodisch unimodal"). Eine Form wie in Abbildung 5.14(a) weist auf einen bewegungsveränderlichen Stern hin („Eclipsing Binary", vgl. Warner 2006, Kapitel 2). Die dazugehörige Periode $p^\star$ wird nur von den eine Fouriersumme dritten Grades anpassenden Methoden erkannt, wobei der Periodogrammbalken bei $p^\star/2$ erheblich höher als bei $p^\star$ ist (vgl. Abbildungen 5.15(b), (d) und (f)). Mit Stufenperiodogrammen (vgl. Abbildungen 5.15(a), (c) und (e)) wird $p^\star$ nicht detektiert, bei den Sinusperiodogrammen (ohne Abbildung) ist der entsprechende Periodogrammbalken nicht einmal lokal maximal. Vermutlich lassen sich beide periodische Funktionen nicht präzise genug an die periodisch bimodale Form anpassen: die Stufenfunktion nicht wegen der starr vorgegebenen Sprungstellen, die Sinusfunktion nicht aufgrund ihrer zyklischen Unimodalität.

Auf der ganzzahligen Testperiodenmenge detektieren fast alle robusten Methoden ein bis zwei Perioden, die jedoch je nach Methode unterschiedlich sind. Eine Betrachtung der zugehörigen Phasendiagramme legt den Verdacht nahe, dass die robusten Regressionstechniken in dieser scheinbar ausreißerfreien Lichtkurve durch Vernachlässigung einiger hoher Messwerte die falsche Perioden erfolgreich anpassen können.

Eine periodisch bimodale Fluktuation taucht auch in anderen Lichtkurven des CSDR auf und die Ergebnisse sind meist ähnlich wie hier beschrieben. Wenn die lokalen Maxima eine stark unterschiedliche Höhe aufweisen, ist bei Anpassung einer Fouriersumme dritten Grades der

[5] Heruntergeladen am 2. Oktober 2013 von: http://nesssi.cacr.caltech.edu/DataRelease/ExamplesCSS.html

Periodogrammbalken bei $p^\star$ höher als der bei $p^\star/2$. Für Sinus- und Stufenfunktionen gilt dies nicht. Auf der ganzzahligen Testperiodenmenge kommt es ab und zu zu Falschdetektionen.

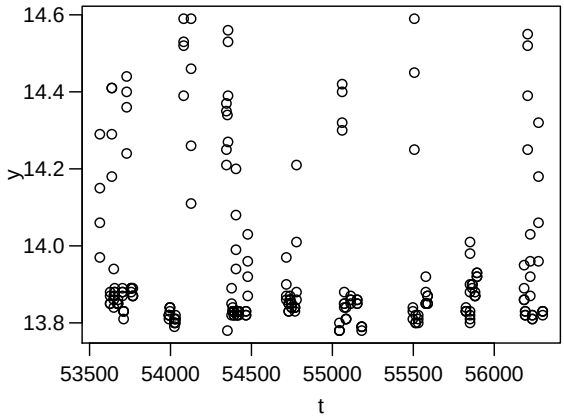
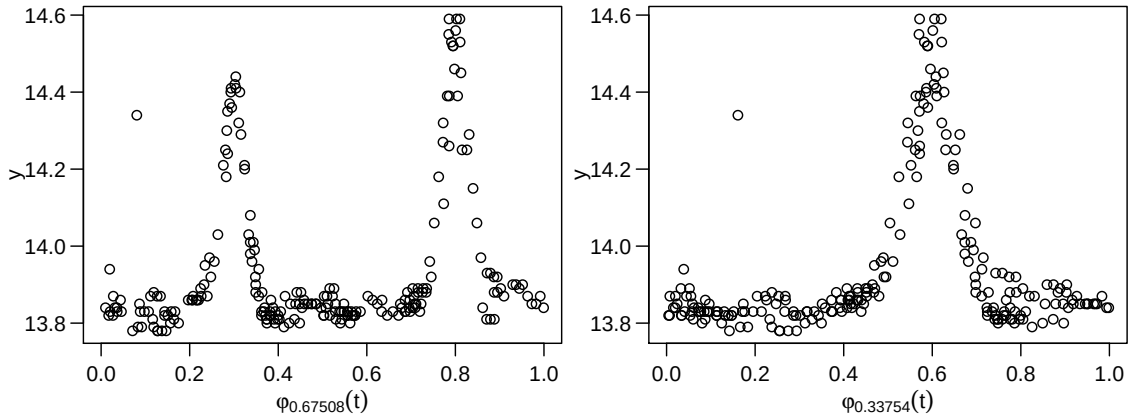


Abbildung 5.13.: Lichtkurve des Sterns mit Catalina-ID 1001005030535721. Messzeiten in MJD. Messwerte in Magnituden.



(a) (b)

Abbildung 5.14.: Phasendiagramme der Lichtkurve des Sterns mit Catalina-ID 1001005030535721. Periode: (a) Referenzperiode $p^\star = 0.67508$, (b) $p^\star/2$.

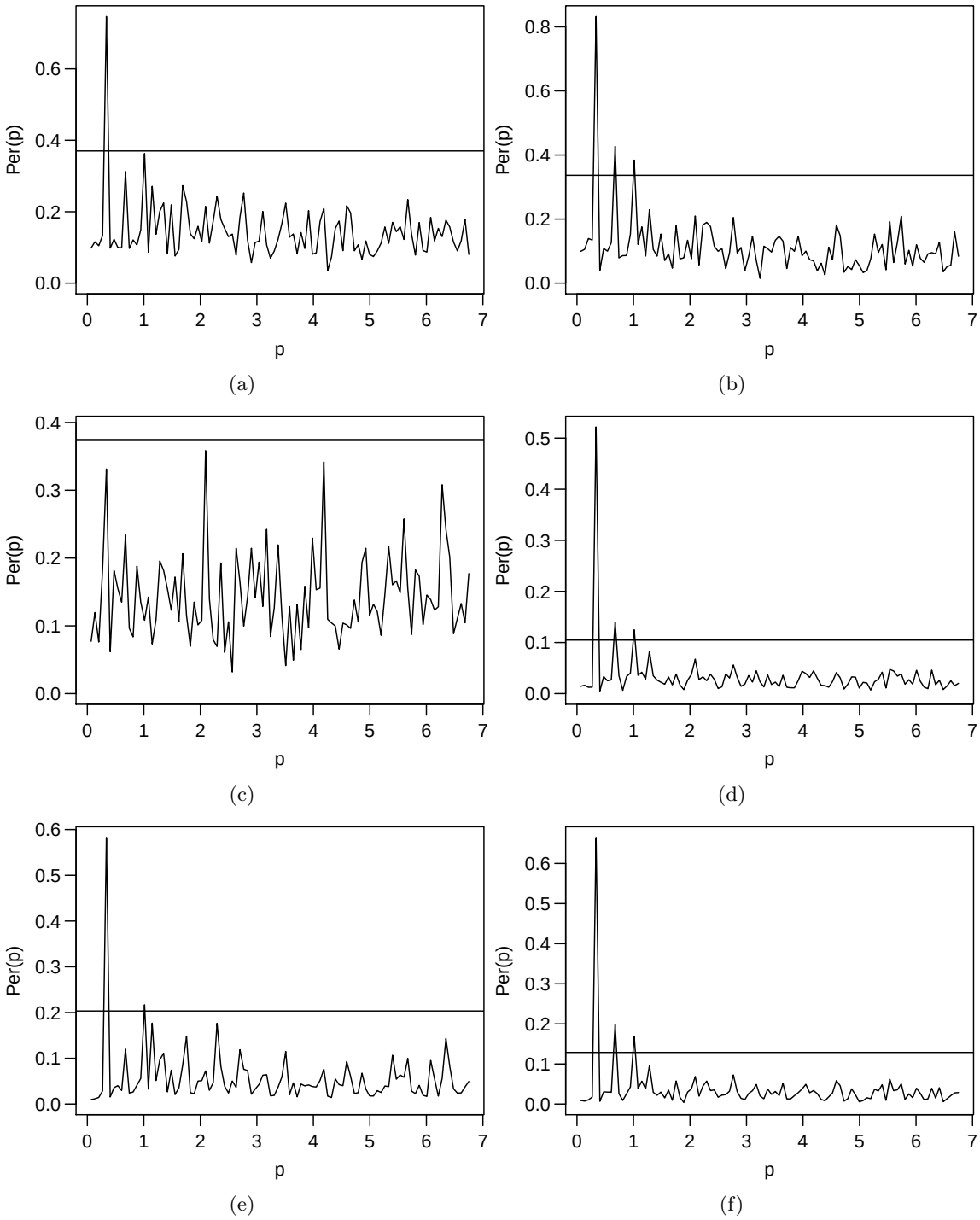


Abbildung 5.15.: Periodogramme der Lichtkurve des Sterns mit Catalina-ID 1001005030535721. Angepasste Funktion: (links) Einfachstufenfunktion, (rechts) Fouriersumme dritten Grades. Regressionstechnik: (oben) KQ, (c) τ, (d) L1, (unten) M-Huber. Die horizontale Linie markiert den jeweiligen Schwellwert für Detektionen.

5.4. Photonenemissionen: Das Burst and Transient Source Experiment

Die folgenden Lichtkurven stammen aus dem Burst and Transient Source Experiment[6] (vgl. Fishman et al. 1989). Während der Laufzeit des Experiments erfasste ein auf einem Satelliten befindlicher Detektor Photonenemissionen. Bei diesen Lichtkurven fällt kein periodisches Messzeitmuster auf. Dies ist damit zu erklären, dass ein auf einem Satellit befindlicher Detektor unabhängig von einem Tag-Nacht-Rhythmus und der Ausrichtung der Erde aufnimmt. Die Messzeiten sind wieder in MJD angegeben. Gemessen wird die mittlere Anzahl beobachteter Photonen mit einer Energie zwischen 20 und 70 Kiloelektronenvolt pro Sekunde pro Quadratzentimeter. In den Daten können Intervallstörungen, stark erhöhte Messwerte in einzelnen Intervallen der Messreihe, beobachtet werden. Diese erhöhten Messwerte entstehen nicht durch eine Fehlfunktion des Detektors, sondern durch für die Photonenquelle spezifische Mechanismen. Sie entstehen jedoch unabhängig von der interessierenden periodischen Schwankung und sind daher nicht bei ihrer Detektion hilfreich. Im Folgenden werden für den Umgang der Detektionsmethoden mit diesen Intervallstörungen einige Beispiele gegeben.

Zunächst wird die Lichtkurve von Gammaemissionen der Galaxie NGC 4151 (Abbildung 5.16) betrachtet. Diese Lichtkurve beinhaltet 778 Messungen, verteilt auf 3317 Tage. Abbildung 5.17 zeigt die entsprechenden Periodogramme. Das τ-Stufenperiodogramm detektiert eine Periode der Länge 52 (vgl. Abbildung 5.17(b)). Die Balken bei den Vielfachen der Periode sind ebenfalls erhöht. Dieses Phänomen tritt häufig in Anwesenheit periodischer Fluktuationen auf (vgl. Abschnitt 2.3.4 und dort Abbildung 2.11, sowie Periodogramme zum Stern bei den Koordinaten (12:31:20 / 70:20,2) in Abbildung 5.12) und kann ein Hinweis auf eine Periodizität sein. In den anderen Stufenperiodogrammen (vgl. Abbildungen 5.17(a) und (c)) oder Periodogrammen, bei denen eine Fouriersumme dritten Grades angepasst wird (vgl. Abbildungen 5.17(d), (e) und (f)), stechen diese Balken ebenfalls hervor, werden aber nur für KQ-Regression detektiert. Die Anpassung einer Sinusfunktion (vgl. Abbildungen 5.17(g), (h) und (i)) ergibt erhöhte (für keine Regressionstechnik detektierten) Balken bei 26 und 52. Da nur die Anpassung einer unimodalen Funktion zu einem höheren Periodogrammbalken bei Periode 26 als bei $2 \cdot 26 = 52$ führt, liegt die Vermutung einer periodoisch bimodalen Fluktuation, wie in Abschnitt 5.3, nahe. Abbildung 5.18 zeigt Phasendiagramme zu diesen beiden Perioden. Die Daten streuen sehr stark, doch eine periodisch uni- oder bimodale Fluktuation könnte möglich sein. Aufgrund der hohen Streuung sollten zukünftig weitere Lichtkurvendaten der gleichen Quelle untersucht werden, um die hier gefundenen Hinweise auf eine Periodizität zu verifizieren. Die Feststellung einer solchen Periodizität in der Vergangenheit ist der Autorin nicht bekannt. Parsons et al. (1998) analysieren eine kürzere Lichtkurve von Gammaemissionen, im Lomb-Scargle-Periodogramm sind die entsprechenden Periodogrammbalken lokal, aber nicht global maximal. Lyutyi und Oknyanskii (1987) beobachten in älteren Messungen der gleichen Galaxie eine Periode von ungefähr $5 \cdot 26 = 130$ Tagen. Sie beobachten allerdings sichtbares Licht und die Messzeiten liegen in größeren Abständen (560 Beobachtungen in 76 Jahren).

[6] Heruntergeladen am 20. Juni 2011 von: http://lheawww.gsfc.nasa.gov/users/craigm/batse-lc/

Eine weitere Lichtkurve ist in Abbildung 5.19(a) dargestellt. Diese Lichtkurve umfasst 679 Beobachtungen der Quelle Gro J1719–24, gemessen in einer Zeitspanne von 3316 Tagen. Auffällig hohe Messwerte fallen in zwei Intervallen auf (grau hervorgehoben). Die Periodogramme, die durch Anpassung einer Sinusfunktion entstanden, werden in den Abbildungen 5.20(a), (c) und (e) gezeigt. Während die robusten Periodogramme eine eindeutige Periode bei 51 Tagen indizieren, ist das KQ-Periodogramm dort nur lokal maximal und wächst ansonsten für steigende Testperioden. Dieses Phänomen, dass durch Vorliegen atypischer hoher Intervallstörungen ein KQ-basiertes Periodogramm mit der Testperiode wächst, wurde bereits von Thieler, Fried und Rathjens (2013) an einem halbsynthetischen Beispiel diskutiert. Auch die anderen KQ-Periodogramme (für Anpassung einer Stufenfunktion vgl. Abbildung 5.20(b)) sind auf diesem Datensatz derart unbrauchbar. Die auf Anpassung einer Fouriersumme dritten Grades basierenden robusten Periodogramme (ohne Abbildung) haben zusätzlich erhöhte Periodogrammbalken bei dem Doppelten und Dreifachen der Periode 51, die Stufenperiodogramme (vgl. Abbildung 5.20(d) und (f)) auch bei weiteren Vielfachen. Dies führt zur Detektion eines oder mehrerer Vielfachen der Periode oder im Falle von M-Huber-Anpassung einer Stufenfunktion zur Nichtdetektion.
Abbildung 5.19(b) zeigt das Phasendiagramm der besprochenen Lichtkurve zu Periode 51 unter Auslassung des y-Achsenbereiches, in dem nur Beobachtungen aus den vermeintlichen Intervallstörungen liegen. Eine sinusähnliche Form ist hier gut erkennbar, wenn sie auch stark verrauscht ist. Von einem früheren Bericht über eine derartige Periodizität ist nichts bekannt.
Ein anderes Verhalten der Detektionsmethoden wird durch die Lichtkurve zu Quelle GRS 0834–430 (Abbildung 5.21(a), 630 Beobachtungen, Dauer 3315 Tage) ausgelöst. Hier wird einzig mit dem KQ-Sinus-Periodogramm eine Periode von 114 detektiert (vgl. Abbildung 5.22 für auf Anpassung einer Sinus- oder einer Einfachstufenfunktion basierende Periodogramme). Die Betrachtung des zugehörigen Phasendiagramms (Abbildung 5.21(b)) zeigt, dass 114 eine Periode ist, bei der die hohen Messwerte des ersten Beobachtungsintervalls (grau dargestellt) dicht beieinanderliegen. Die auf unrobuster Anpassung einer periodisch unimodalen Funktion beruhende Detektionsmethode bevorzugt diese Anhäufung so stark, dass die entsprechende Periode detektiert wird. Eine genauere Betrachtung des Phasendiagramms der ersten Beobachtungen zeigt keinen eindeutigen Verlauf im Zyklus. Es scheint also nicht der Fall zu sein, dass mit KQ-Regression ein periodisches Phänomen gefunden und mit anderen Regressionstechniken übersehen wurde.

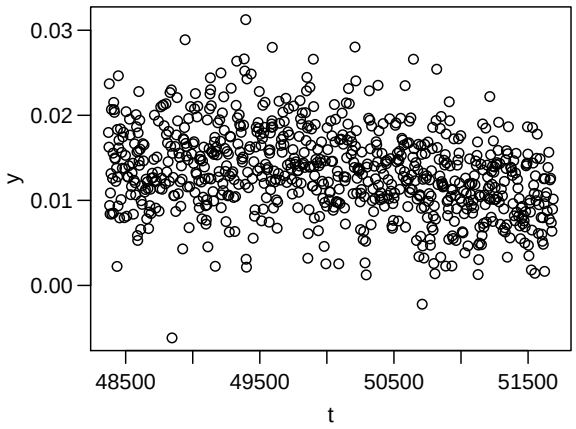


Abbildung 5.16.: Lichtkurve der Quelle NGC 4151. Messzeiten in MJD. Messwerte in mittlerer normierter Photonenanzahl.

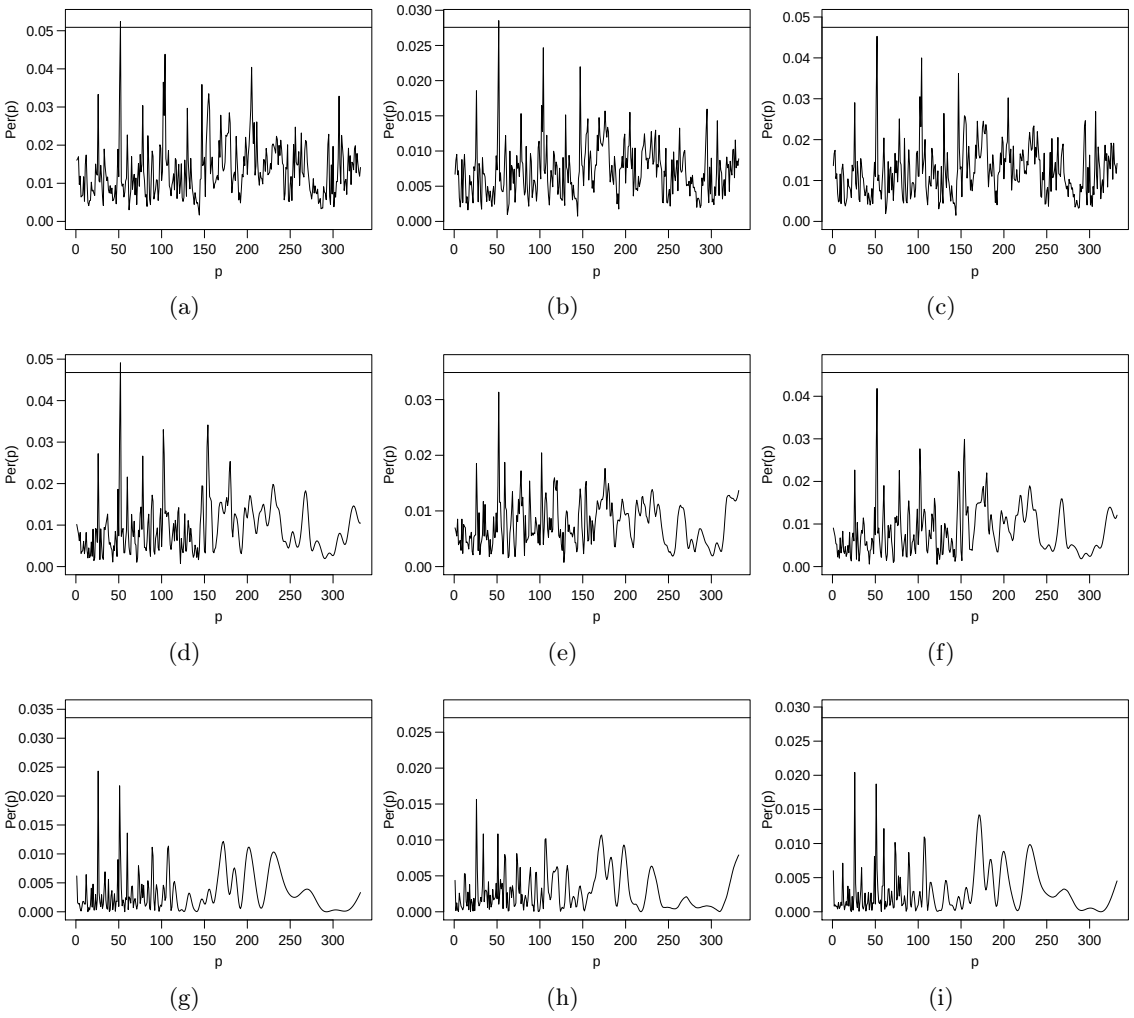
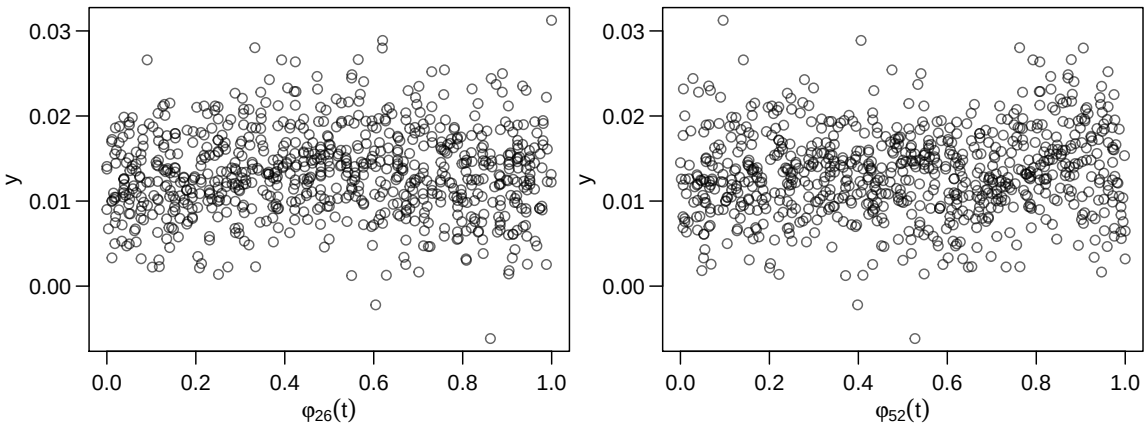
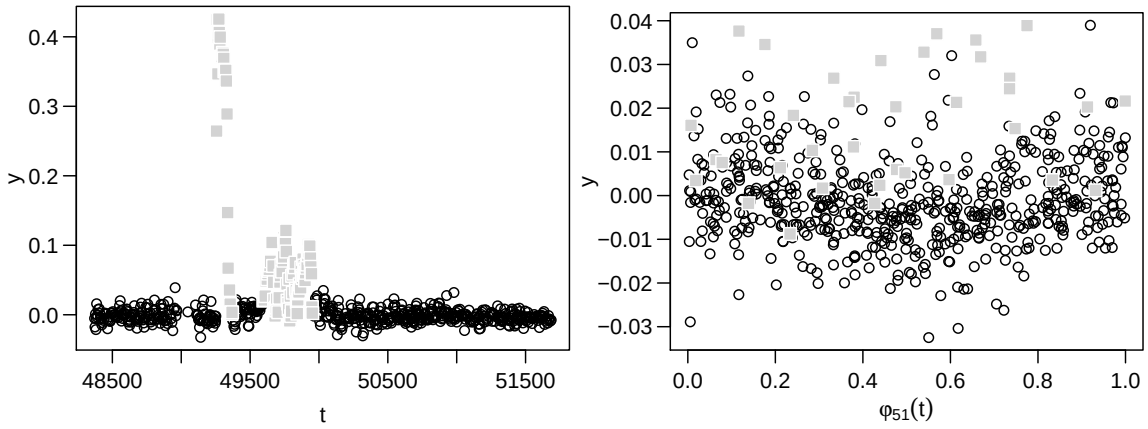


Abbildung 5.17.: Periodogramme der Lichtkurve von NGC 4151. Angepasste Funktion: (oben) Stufenfunktion, (vertikal mittig) Fouriersumme dritten Grades, (unten) Sinusfunktion. Regressionstechnik: (links) KQ , (b) τ , (e)/(h) L1, (rechts) M-Huber. Die horizontale Linie markiert den jeweiligen Schwellwert für Detektionen.



(a) (b)

Abbildung 5.18.: Phasendiagramme der Lichtkurve von NGC 4151. Perioden: (a) 26, (b) 52.



(a) (b)

Abbildung 5.19.: Beobachtungen der Quelle Gro J1719–24. Messzeiten in MJD. Messwerte in mittlerer normierter Photonenanzahl. Darstellung: (a) Lichtkurve (b) Phasendiagramm nach Periode 51, mit reduziertem y-Achsenausschnitt (vgl. Text). Zu einer vermeintlichen Intervallstörung gehörende Beobachtungen (in beiden Diagrammen dieselben) sind grau hervorgehoben.

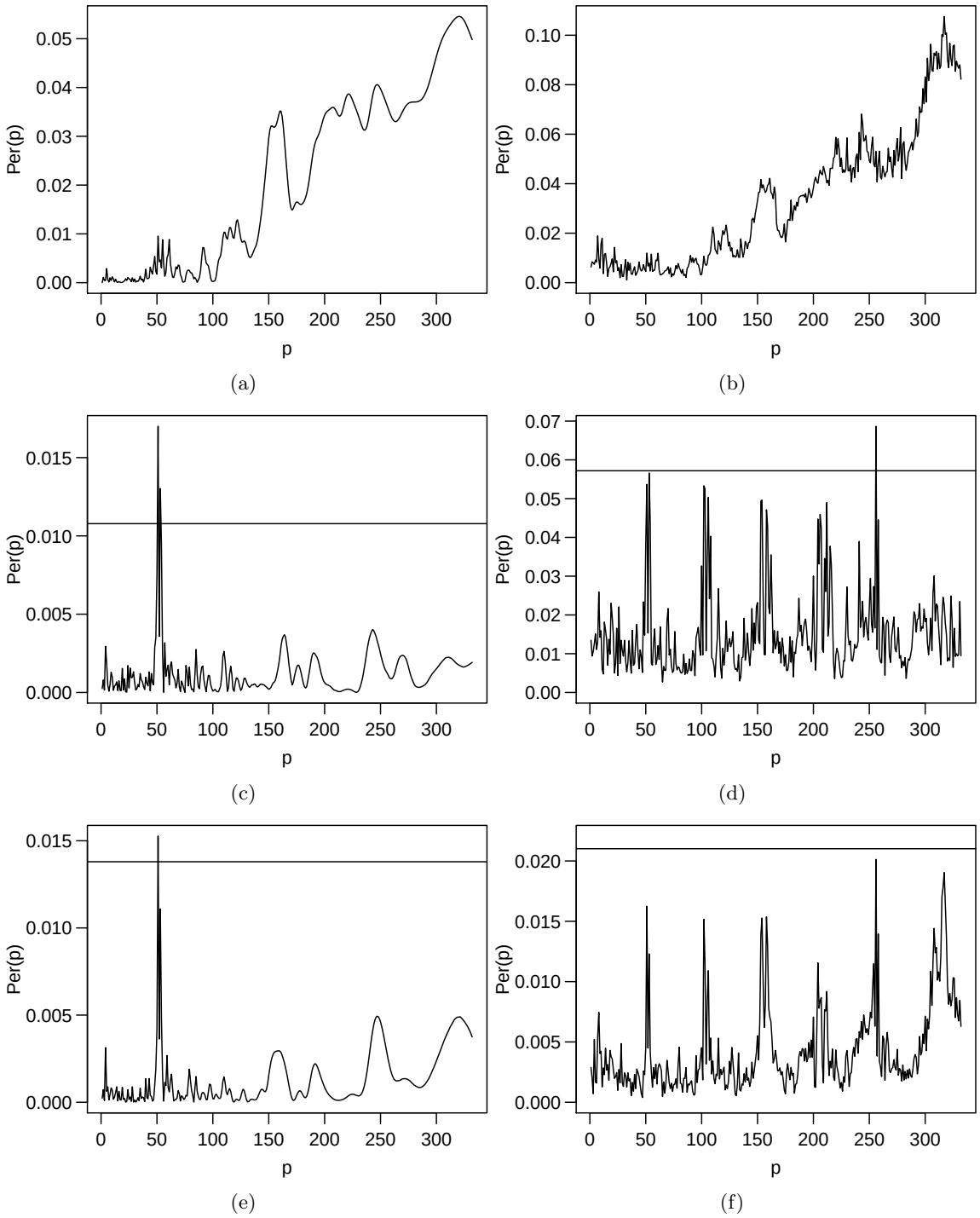
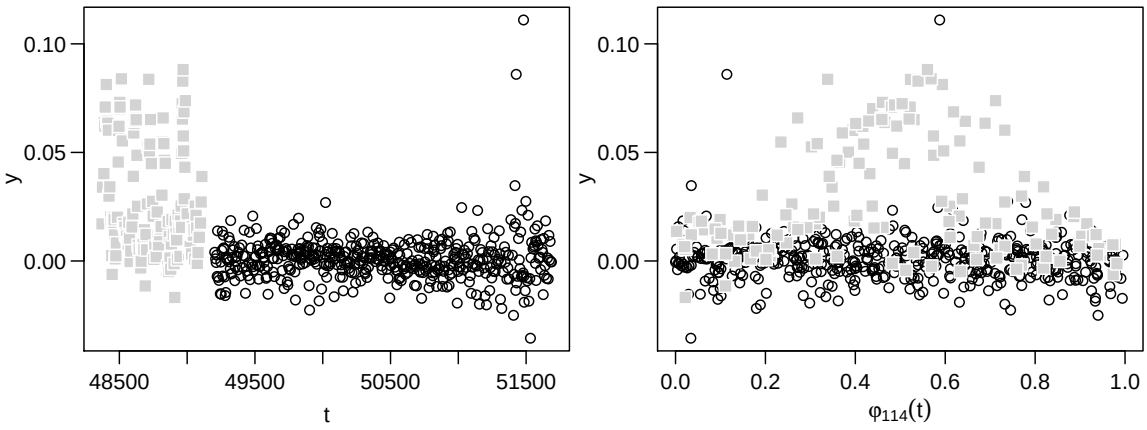


Abbildung 5.20.: Periodogramme der Lichtkurve von Gro J1719–24. Angepasste Funktion: (links) Sinusfunktion, (rechts) Einfachstufenfunktion. Regressionstechnik: (oben) KQ, (c) L1, (d) τ, (unten) M-Huber. Schwellwert zur Detektion: (a) 0.3, (b) 0.29, (c)–(f) eingezeichnet (horizontale Linie).



(a) (b)

Abbildung 5.21.: Beobachtungen zur Quelle GRS 0834–430. Messzeiten in MJD. Messwerte in mittlerer normierter Photonenanzahl. Darstellung: (a) Lichtkurve, (b) Phasendiagramm nach Periode 114. Zu einer vermeintlichen Intervallstörung gehörende Beobachtungen (in beiden Diagrammen dieselben) sind grau hervorgehoben.

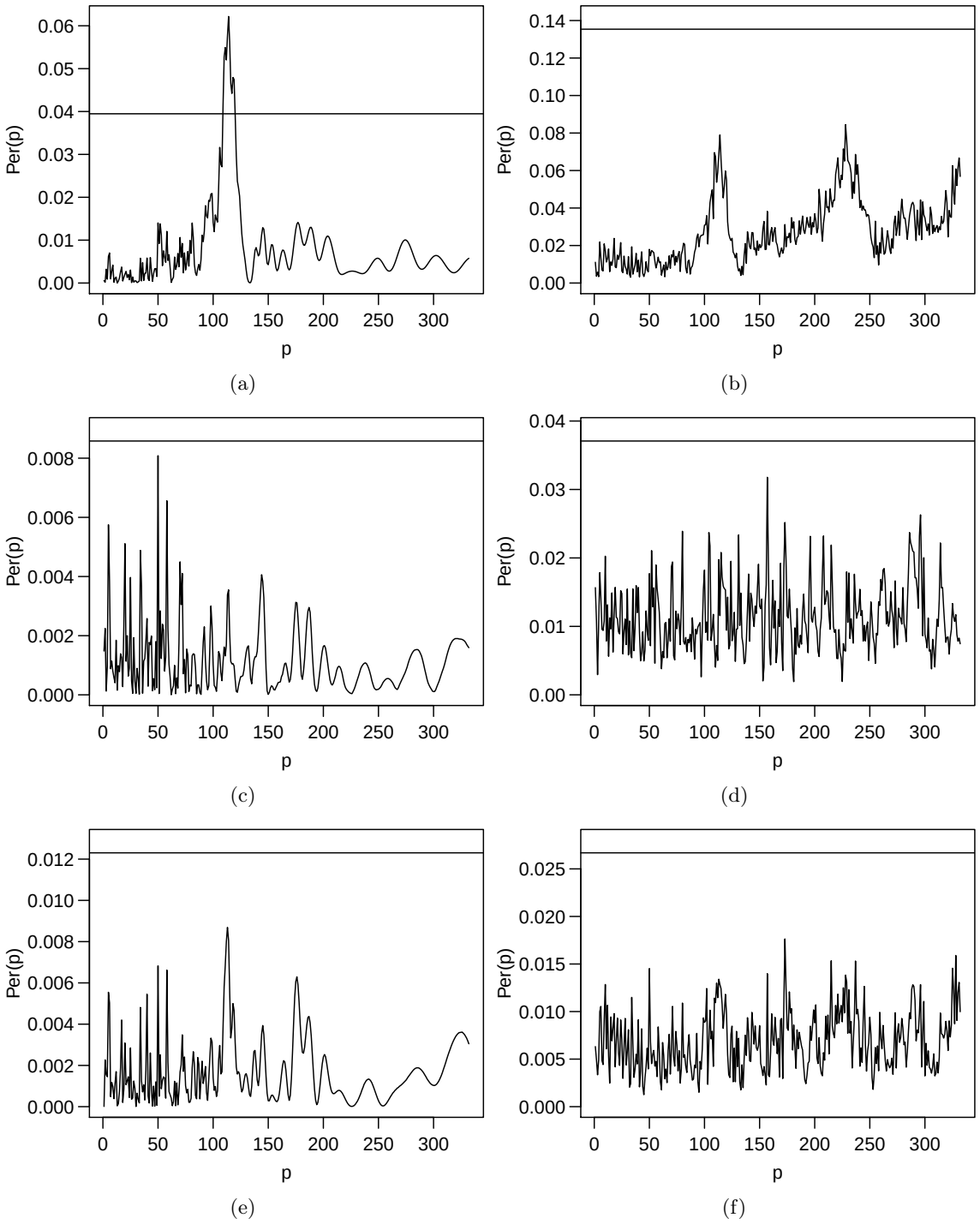


Abbildung 5.22.: Periodogramme der Lichtkurve von GRS 0834–430. Angepasste Funktion: (links) Sinusfunktion, (rechts) Einfachstufenfunktion. Regressionstechnik: (oben) KQ, (c) L1, (d) τ, (unten) M-Huber.

5.5. Stoffmengen: Eisbohrungen in der Antarktis

Als abschließendes Anwendungsbeispiel werden die in dieser Arbeit entwickelten und untersuchten Periodogrammmethoden auf Daten angewendet, die nicht aus der Astroteilchenphysik, sondern aus der Paläoklimatologie stammen. Die von Petit et al. (1999) veröffentlichten Daten stammen aus Eisbohrungen in der östlichen Antarktis[7]. Über die Bohrtiefe kann das Alter des Eises berechnet werden, wodurch die gemessenen Anteile bestimmter Stoffe als Zeitreihe darstellbar sind. Durch die Zuordnung der Bohrtiefe zu einem Alter entsteht ein ungleichmäßiges, vermutlich unperiodisches Messzeitmuster. Gemessen wird der Staubanteil im Eis in ppm (parts per million, 10^{-6}).
Die Betrachtung der Messwerte y_i in Abhängigkeit vom Eisalter t_i bedeutet, dass die Messpunkte t_i negativ mit der Zeit korrelieren: Eine Messung mit kleinem Wert t_i wurde in jungem Eis gefunden, der gemessene Staub hat sich also spät im Eis abgesetzt, während Messungen mit einem hohen Wert t_i zu altem Eis gehören, in denen sich früh Staub abgesetzt hat. Obwohl die t_i nicht positiv von der Zeit abhängen, werden sie im Folgenden wie gewohnt Messzeiten genannt. Da die in den Periodogrammmethoden angepassten periodischen Funktionen alle symmetrisch in der Zeit sind, sind die Periodogrammmethoden invariant bezüglich der Orientierung der Messzeiten.
Abbildung 5.23 zeigt den im Eis gemessenen Staubanteil. Es liegen hierzu 522 Messungen für ein Zeitintervall von 417 252 Jahren vor. Alle berechneten Periodogramme sind in Abbildung 5.24 zu finden. Mit den Stufenperiodogrammen wird stets eine Periode detektiert, mit τ-Regression bei ca. 30 000 Jahren, mit KQ- und Huber-M-Regression bei ca. 40 000 Jahren. Auch das auf L1-Anpassung einer Fouriersumme dritten Grades basierende Periodogramm detektiert eine Periode bei ca. 40 000 Jahren. Die anderen auf Anpassung einer Sinusfunktion oder Fouriersumme basierenden Periodogramme sind bei 40 000 Jahren deutlich erhöht, detektieren aber nichts.
Obwohl die Periodogramme relativ ähnlich aussehen, kommt es nicht bei allen zur Detektion. Dies hängt wahrscheinlich von der genauen Breite und Höhe der Gipfel im Periodogramm ab. Außerdem wurden für diese Periodogramme nach Standardvorgehen in diesem Kapitel q =41 726 Testperioden überprüft, das $\sqrt[q]{0.95}$-Quantil ist damit das 0.9999988-Quantil der Verteilung. Zum Vergleich wird die Betaverteilung auch an die $q = 417$ Periodogrammbalken einer reduzierten Testperiodenmenge $100, 200, \ldots, 41700$ angepasst. Die an das volle Periodogramm angepasste Betaverteilung erreicht leicht höhere Schwellwerte als die an das reduzierte Periodogramm angepasste, für $q = 0.95$ ist das $\sqrt[q]{0.95}$-Quantil maximal 4% höher. Die niedrigere Testperiodenanzahl q führt jedoch dazu, dass im reduzierten Periodogramm das niedrigere 0.9998770-Quantil den Schwellwert bestimmt. In jedem Periodogramm außer dem mit KQ-Anpassung einer Sinusschwingung überragt der höchste Periodogrammbalken diesen Schwellwert.
In den Abbildungen 5.25(a) und (b) sind Phasendiagramme für Perioden von ca. 30 000 und 40 000 zu sehen. Abbildung 5.25(c) zeigt zum Vergleich das Phasendiagramm zu einer willkürlich gewählten Periode von 300, welche von keiner Methode detektiert wird. Im

[7] Heruntergeladen am 12. Februar 2010 von:
http://www.nature.com/nature/journal/v399/n6735/extref/nature399429-s1.pdf

Phasendiagramm zu einer Periode von ca. 40 000 fällt auf, dass hohe Messwerte stark gruppiert sind. Dies konnte auch für andere Perioden nahe 40 000 festgestellt werden. Bei dem Phasendiagramm nach der vom τ-Periodogramm favorisierten Periode von ca. 30 000 sind die hohen Messwerte relativ stark gestreut. Die niedrigen Werte scheinen dagegen einem Verlauf zu folgen, gut sichtbar bei Phase $\varphi(t) = 0.8$. Es ist möglich, dass beide Periodizitäten gewisse Phänomene in der Reihe erklären, die eine etwa das hohe Staubaufkommen (mit einer Periode von ca. 40 000 Jahren), die andere (mit einer Periode von ca. 30 000 Jahren) das Verhalten der Reihe in der restlichen Zeit.
Bei Betrachtung der Zeitreihe in Abbildung 5.23 fällt noch eine andere mögliche Periode von ca. 106 000 Jahren (ca. vier Zyklen) auf. Ein Phasendiagramm hierzu ist in Abbildung 5.25(d) gegeben. Diese Periode wird vom Periodogramm nicht abgedeckt, da hier nur Perioden berücksichtigt werden, die mindestens in zehn Zyklen beobachtet werden konnten.

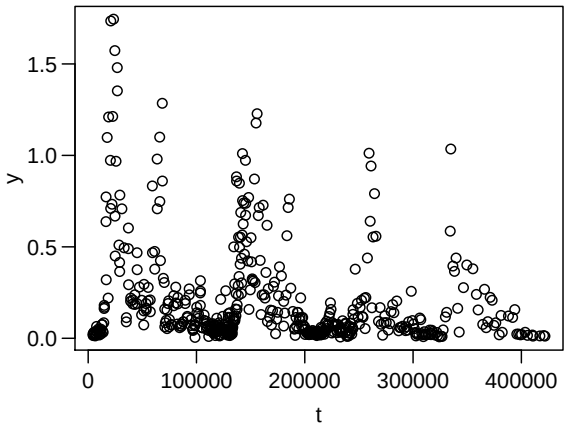


Abbildung 5.23.: Staubanteil im Eis der Antarktis-Bohrungen. Die Zeiteinheit ist Eisalter in Jahren. Messwerte in ppm.

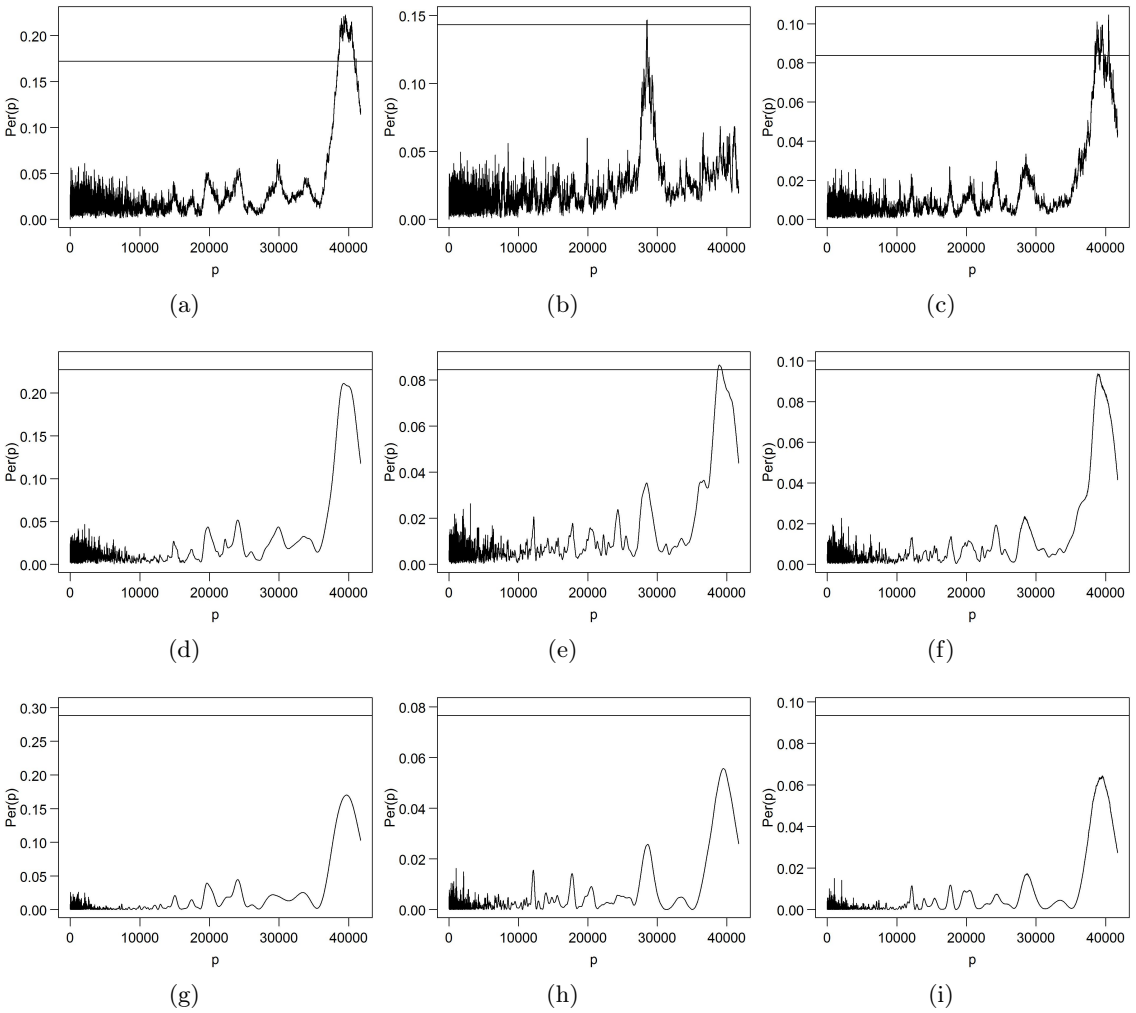
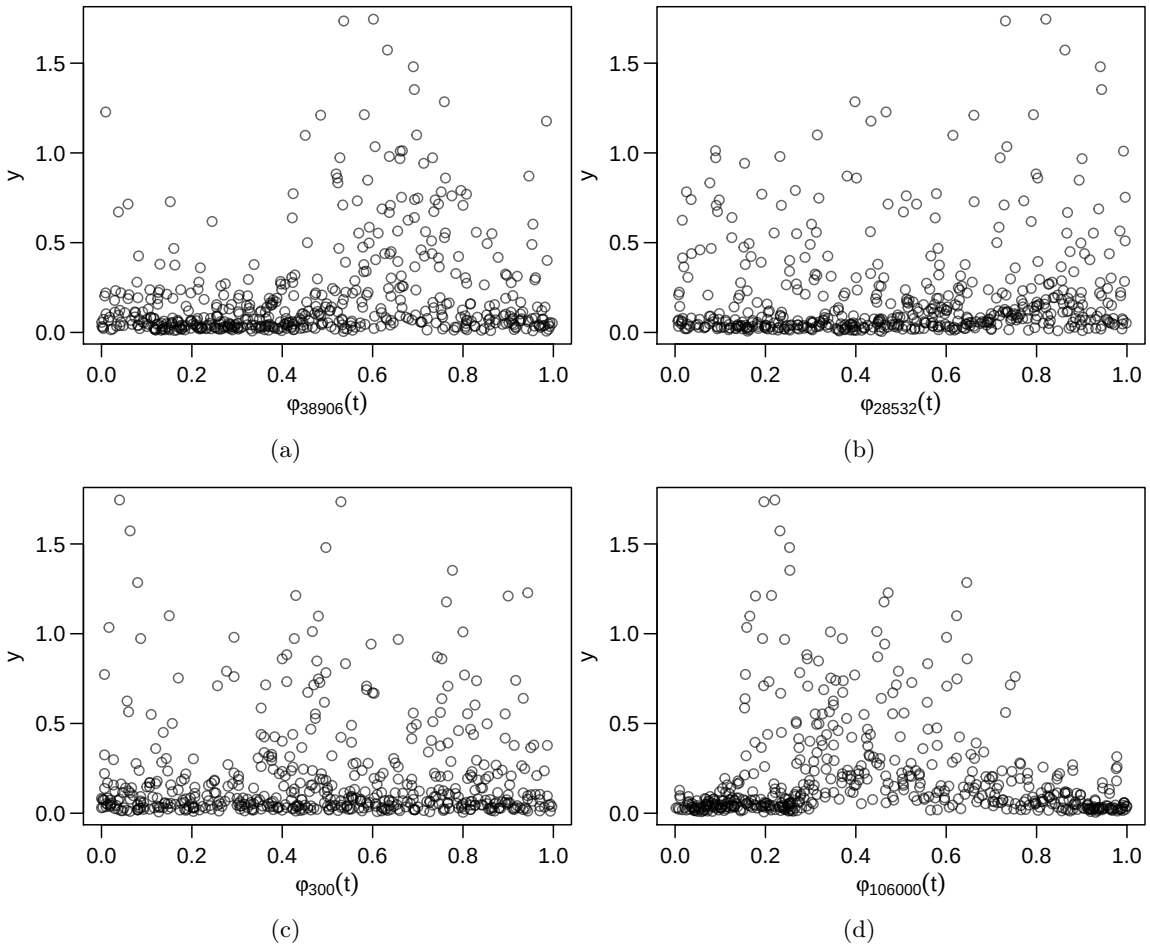


Abbildung 5.24.: Periodogramme der Staubanteil-Zeitreihe. Angepasste Funktion: (oben) Einfachstufenfunktion , (vertikal mittig) Fouriersumme dritten Grades, (unten) Sinusfunktion . Regressionstechnik: (links) KQ , ((b)) τ, (e)/(h) L1, (rechts) M-Huber .



(a) (b) (c) (d)

Abbildung 5.25.: Phasendiagramme der Staubanteil-Zeitreihe. Perioden: (a) 38 906, (b) 28 532, (c) 300, (d)106 000.

Zusammenfassung

Robuste Regression kann von großem Vorteil sein, wenn in den Daten Störungen auftauchen, etwa in den ASAS-Daten durch das Vorliegen vieler Ausreißer (vgl. Abschnitt 5.2) oder bei den Lichtkurven zu Gro J1719-24 und GRS 0834-430 durch Intervallstörungen (vgl. Abschnitt 5.4). Speziell im τ-Stufen-Periodogramm kommt es durch Ausreißerbereinigung der Lichtkurve häufig zu nur kleinen Änderungen im Periodogramm, während bei anderen robusten Periodogrammen nur die Relation der Balken untereinander ähnlich bleibt, sich aber die Skalierung ändert.
Bei der Anwendung robuster Regression ist es jedoch besonders wichtig, dass die potenziell vorliegende periodische Fluktuation auch mit der gewählten periodischen Funktion gut anzupassen ist, weil sonst Messwerte mit hohem Informationsgehalt als Ausreißer betrachtet werden könnnen. Dies ist zum Beispiel beobachtbar, wenn periodisch bimodale Lichtkurven aus dem CSDR (Abschnitt 5.3) auf der falschen Testperiodenmenge untersucht werden. Speziell bei multimodalen periodischen Fluktuationen wie bei einem Veränderlichen Stern sind die unimodale Sinusfluktuation und die Stufenfunktion mit ihren festen Sprungstellen nicht flexibel genug. Hier ist Expertenwissen um mögliche vorliegende Fluktuationsformen von immensem Vorteil und kann auch gewinnbringend bei der Wahl der Testperiodenmenge eingesetzt werden. Eine beliebig große Testperiodenmenge zu nutzen bedeutet lange Rechenzeiten und kann den von der Anzahl der Testperioden abhängigen Schwellwert für Detektionen in die Höhe treiben, wie beispielsweise bei den Staubanteildaten in Abschnitt 5.5, so dass es zu keiner Detektion kommt.
Von den in Abschnitt 2.6.2 beschriebenen Abhängigkeiten im Periodogramm können vor allem zwei beobachtet werden: die Abhängigkeit mehrerer Testperioden durch Überparametrisierung der angepassten Funktion und die Abhängigkeit dicht beieinander liegender Perioden. Die Überparametrisierung führt dazu, dass die periodische Fluktuation durch verschiedene Testperioden gleichermaßen gut angepasst werden kann. Sie wird beispielsweise für die Lichtkurve zum Stern mit den Koordinaten (12:31:20 / 70:20,2) (vgl. Periodogramme in Abbildung 5.12) deutlich und kann auch gut für die Quellen NGC 4151 (vgl. Abbildung 5.17) und Gro J1719–24 (vgl. Abbildung 5.20) beobachtet werden.
Mit den Detektionsmethoden konnten in den hier analysierten Lichtkurven teilweise Perioden detektiert werden, die bisher nicht dokumentiert zu sein scheinen. So wurden Hinweise auf eine Periode von 263 heliozentrischen Tagen in den ASAS-Daten zum Stern mit den Koordinaten (12:00:09 / 67:52,8) gefunden, auf eine Periode von 26 oder 52 Tagen bei den Gammaemissionen zur Galaxie NGC 4151 und auf eine Periode von 51 Tagen in den Emissionen der Quelle Gro J1719–24. Diese Perioden dürfen nicht als signifikant detektiert betrachtet werden, da in diesem Kapitel eine Vielzahl abhängiger Tests durchgeführt wurde und nicht jede Detektionsmethode die jeweilige Periode detektiert. Die Analysen weiterer Lichtkurven der jeweiligen Quelle könnte die hypothetische Periodizität verifizieren.

6. Erweiterung des Konzepts

Durch Erweiterung des in dieser Arbeit verfolgten Periodogrammprinzips und Lockerung der zu Grunde liegenden Modellannahmen ergeben sich einige interessante Forschungsfelder für die Zukunft. Dieses Kapitel bietet erste Ansätze und Vorarbeiten für ihre zukünftige Erschließung.
Die Stärke des hier vorgeschlagenen Periodogrammprinzips besteht in seiner beliebigen Erweiterbarkeit durch zusätzliche Regressionstechniken und periodische Funktionen. Vorschläge für weitere Techniken und Funktionen werden in Abschnitt 6.1 unterbreitet.
Darauf folgend werden Erweiterungen der zu Grunde liegenden Modellannahmen und damit einhergehende notwendige Modifikationen der Detektionsmethoden diskutiert: Abschnitt 6.2 behandelt eine weitere mögliche Rauschkomponente, das sogenannte rote Rauschen. In Abschnitt 6.3 werden periodische Fluktuationen betrachtet, die sich bezüglich ihrer Gestalt und Fluktuationsperiode verändern. Dabei werden auch die Arbeiten von Fried, Raabe und Thieler (2012) und Raabe et al. (2012) berücksichtigt, die im Rahmen einer Fragestellung aus dem Maschinenbau ebenfalls die Analyse periodischer Zeitreihen behandeln.
Die in diesem Kapitel gemachten Vorschläge sind erste, nicht im Detail erprobte Ideen zur Erweiterung der in dieser Arbeit betrachteten Detektionsmethoden. Sie können als erste Ansätze für zukünftige Forschungsvorhaben dienen.

6.1. Zusätzliche Periodogrammmethoden

Die Periodogrammberechnung basiert in dieser Arbeit auf Anpassung einer periodischen Funktion g mittels einer Regressionstechnik. In dieser Arbeit wurden sechs periodische Funktionen g (vgl. Abschnitt 2.3) mit sieben Regressionstechniken (vgl. Abschnitt 2.4) kombiniert. Mögliche Vorteile und Herausforderungen beim Einsatz weiterer Regressionstechniken werden in Abschnitt 6.1.1 besprochen. Einige Aspekte der Gestalt von g in Hinblick auf adaptive bis hin zu Filterverfahren werden in Abschnitt 6.1.2 diskutiert.

6.1.1. Einsatz anderer Regressionstechniken

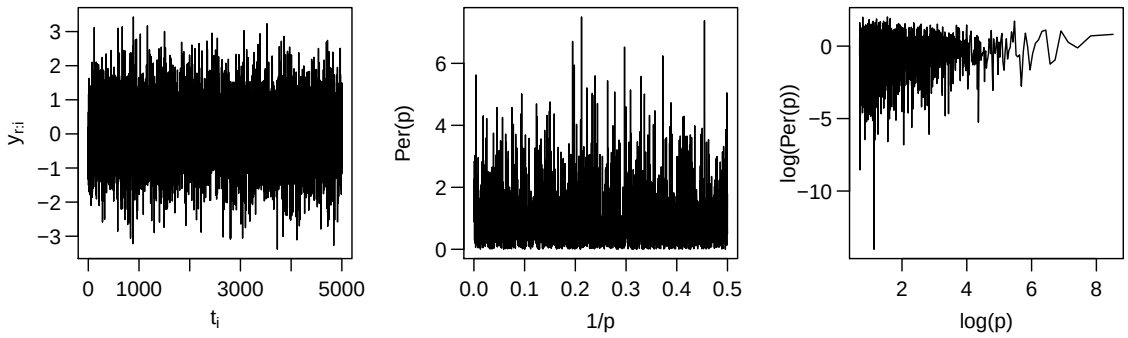
Neben den in dieser Arbeit genutzten können auch andere Regressionstechniken für die Periodogrammberechnung verwendet werden. Ein erster Ansatz wäre bei der M-, der S- und der τ- Regression andere ρ-Funktionen (Abschnitt 2.4.2) einzusetzen. Beispielsweise wurde die S-Regression (vgl. Abschnitt 2.4.3) bisher mit der Tukey-Funktion verwendet. Die Tukey-Funktion ist beschränkt, sodass alle als Ausreißer gewerteten Beobachtungen einen gleich starken Einfluss auf das Minimierungskriterium haben. Das macht die darauf basierenden Regressionstechniken einerseits sehr robust. Andererseits könnte dies der Grund

für die unzureichende Detektionsfähigkeit der auf S-Regression basierenden Methoden sein, die in der Simulationsstudie für Situationen beobachtet werden konnte, in denen sich die periodische Fluktuation f und die anzupassende Funktion g in ihrer Gestalt stark unterscheiden (vgl. Abschnitt 4.3.2). In diesem Fall könnte eine unbeschränkte ρ-Funktion von Vorteil sein (Modifikationen der anzupassenden Funktion werden in Abschnitt 6.1.2 behandelt).
Weiterhin gibt es auch die Möglichkeit, auf anderen Prinzipien beruhende Regressionstechniken zu verwenden. So untersucht Rathjens (2012) unter anderem die Least-Weighted-Regression von Víšek (2000) und die Repeated-Median-Regression von Siegel (1982) auf ihre Fähigkeiten zur robusten Anpassung einer Sinusfunktion in unregelmäßig beobachteten Zeitreihen. Anhand der dort gewonnenen Erkenntnisse könnte sich vor allem die Repeated-Median-Regression gut zur Periodendetektion nach dem hier verfolgten Prinzip eignen.
Die Schätzung des Parametervektors wird bei der Repeated-Median-Regression für jede Dimension separat durchgeführt. Anhand der anzupassenden periodischen Funktion g müsste zuvor eine geeignete Reihenfolge oder die simultane, unabhängige Durchführung der Schätzungen festgelegt werden. Eine sinnvolle Definition für einen Periodogrammbalken muss hier noch gefunden werden, da die Regressionstechnik nicht über die Minimierung einer Zielfunktion definiert ist und SY und SE (vgl. Definitionen (2.11) und (2.12), Seite 11) daher nicht analog zu den anderen Regressionstechniken definiert werden können. Bei einer anderen robusten Regressionstechnik, der Deepest Regression von Rousseeuw und Hubert (1999), existiert zwar eine Zielfunktion (die Regression Depth), sie ist allerdings nicht für das Lokationsmodell definiert. Auch hier müssten Überlegungen zur Definition des zugehörigen Periodogramms angestellt werden.
Weiterhin kann analog zur S- und τ-Regression zu jedem Varianzschätzer eine Regressionstechnik definiert werden, die diesen Schätzer minimiert. Das Bestimmtheitsmaß und damit der Periodogrammbalken können dann wie bisher definiert werden. Ein Beispiel wäre die durch Minimierung des Q_n-Varianzschätzers (Rousseeuw und Croux 1993) der Residuen definierte LQD-Regression (vgl. Croux, Rousseeuw und Hössjer 1994).

6.1.2. Periodische Fluktuationen komplexer Gestalt

In der Simulationsstudie (Kapitel 4) und der Anwendung auf reale Lichtkurven (Kapitel 5) wird deutlich, dass periodische Funktionen g mit wenig Parametern $\beta_1, \ldots, \beta_m$ in manchen Fällen zu unflexibel sind, um die periodische Fluktuation f gut anzupassen. Dies zeigt sich vor allem bei der Sinus- und der Splinefunktion zur Anpassung einer Peakfluktuation (vgl. Abbildung 4.7 auf Seite 67) oder bei der Sinusfunktion zur Anpassung einer periodisch bimodalen Fluktuation (vgl. Abschnitt 5.3). Bei jeder der vorgestellten Funktionen g ist es möglich, die Parameterzahl zu erhöhen, um eine flexiblere Anpassung zu erlauben: Es kann eine Stufenfunktion mit vielen Stufen pro Zyklus, eine Fouriersumme hohen Grades oder eine periodische Splinefunktion mit vielen Knoten pro Zyklus angepasst werden. Die Stufenfunktion mit zehn Stufen ist zum Beispiel in den Simulationsstudien sehr erfolgreich bei der Detektion einer Peakfluktuation (Abbildung 4.7 auf Seite 67).

Allein die Erhöhung der Parameteranzahl führt jedoch nicht immer zu besseren Detektionsergebnissen. So kann für Stufenfunktionen mit zehn Stufen und Fouriersummen dritten Grades eine erhöhte Abhängigkeit unter den Testperioden festgestellt werden. Es kann bisweilen passieren, dass eine periodische Fluktuation von der angepassten Funktion dann nicht nur gut mit der Fluktuationsperiode p_f, sondern auch mit Vielfachen $2p_f$ und $3p_f$ beschreibbar ist. Beispiele hierfür sind in Kapitel 5 zu finden (vgl. Periodogramme in Abbildungen 5.12(b)/(d)/(f), 5.17(a)-(c) und 5.20(d)/(f)). Außerdem stellt die Verwendung vieler Parameter nicht sicher, dass die Funktion gut angepasst wird: So kann die zehnstufige Stufenfunktion in Abschnitt 5.3 nicht die für einen veränderlichen Stern typische periodisch bimodale, sondern nur eine periodisch unimodale Fluktuation detektieren. Zehn Stufen sollten zwar für die Anpassung einer solchen Form genügen, die in gleichmäßigen Abständen gewählten Sprungstellen befinden sich aber in diesem konkreten Beispiel an ungünstigen Stellen. Analog genügen zwar vier Knotenpunkte, um eine Peakfluktuation $f_{peak;p_f}$ mit einer periodischen Splinefunktion annähernd zu beschreiben. Werden die Knotenpunkte jedoch, wie in dieser Arbeit, gleichabständig gewählt, ist das nicht der Fall. Zur Anpassung einer periodisch bimodalen Fluktuation werden außerdem mehr Knotenpunkte benötigt.
Eine Möglichkeit, diesen Problemen zu begegnen, ist, die Anzahl und Position der Sprungstellen und Knotenpunkte für die Stufen- und die Splinefunktion adaptiv zu wählen. Oh et al. (2004) passen dazu ein Modell mit hoher Knotenanzahl an und ergänzen das Minimierungskriterium um einen Bestrafungsterm. Dieser führt dazu, dass die Anzahl der Parameterkoeffizienten ungleich null möglichst niedrig gehalten wird, sodass nur wenige Knotenpunkte tatsächlich mit in die Modellierung einfließen. Ein Ansatz zur adaptiven Sprungstellenwahl bietet die Dissertation von Morell (2012). Dort werden unter anderem robuste Filter mit adaptiver Fensterbreite betrachtet, die aus der gefilterten Zeitreihe ein bis auf Sprünge konstantes Signal extrahieren.
Entsprechend dem Ansatz nach Morell (2012) ist es auch möglich, sich komplett von der Idee einer parametrischen periodischen Funktion g zu entfernen und als Anpassung $\widehat{g}$ stattdessen das Filterergebnis eines auf das Phasendiagramm angewendeten Filters zu verwenden. Der bisherige Einsatz von Filterverfahren zur Periodendetektion wurde bereits in Abschnitt 2.7.3 besprochen. Die Nutzung von Zeitreihenfiltern eröffnet mit vielen untersuchbaren Verfahren ein weites Forschungsfeld. Einige Verfahren müssten jedoch für ungleichmäßige Zeitreihen erweitert werden. Auf eine Erweiterung zur Berücksichtigung von Messfehlern kann nach den Simulationsergebnissen der vorliegenden Arbeit verzichtet werden (vgl. Kapitel 4).
Die adaptive Wahl der Sprungstellen bzw. Knotenpunkte bedeutet, dass viele Modelle zu falschen Testperioden erfolgreicher als bisher angepasst werden können, es also zu mehr hohen Periodogrammbalken kommen kann. Ein ähnlicher Effekt kann bei der Nutzung von Filtern erwartet werden. Das Anpassen einer Betaverteilung und die anschließende Ausreißerdetektion führt dann eventuell nicht mehr zu einer erfolgreichen Periodendetektion. In den Arbeiten von McDonald (1986), Hall, Reimann und Rice (2000) zur Periodensuche mittels Filterverfahren sowie Oh et al. (2004) wird dies nicht thematisiert, da das Vorhandensein einer Periodizität vorausgesetzt und das Periodogramm nur zur Gittersuche nach der optimalen Periode verwendet wird. Ist jedoch, wie in dieser Arbeit angenommen, die



(a) (b) (c)

Abbildung 6.1.: Beispiel für weißes Rauschen. (a) Unabhängig identisch normalverteilte Messwerte mit Varianz 1, (b) Fourier-Periodogramm von (a), (c) Diagramm aus (b) mit logarithmierten Achsen, Abszisse Periode statt Frequenz.

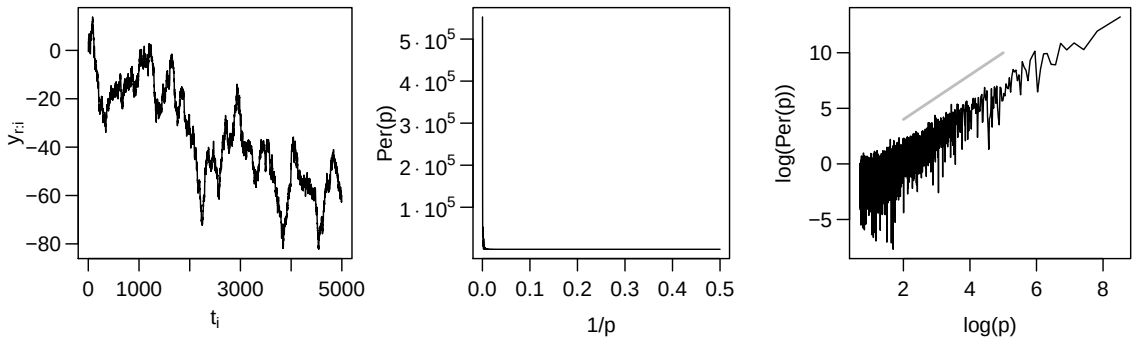
Abwesenheit jeglicher periodischer Fluktuation möglich, muss das Ausmaß an Adaptivität bzw. die weitere Modifikation des Detektionsverfahrens umsichtig umgesetzt werden.

6.2. Rotes Rauschen

Neben der Einführung neuer Periodogrammmethoden kann die Anwendbarkeit des entwickelten Detektionsprinzips auch durch Modifikationen im angenommenen Datenmodell (vgl. Abschnitt 2.1) erweitert werden.
Neben unkorreliertem weißen Rauschen wurde in der Vergangenheit bei Lichtkurven auch ein spezielles korreliertes Rauschen beobachtet, welches häufig rotes Rauschen, Potenzgesetzrauschen (Power Law Noise) oder $1/f$-Rauschen genannt wird. In Abschnitt 6.2.1 wird diese Rauschart vorgestellt. Die Definition von rotem Rauschen geschieht direkt über das Periodogramm der Fourieranalyse und wird in diesem Abschnitt daher nur für gleichmäßig abständige Zeitreihen und die Fourierfrequenzen bzw. die dazugehörigen Perioden betrachtet (vgl. Abschnitt 2.2.1). In Abschnitt 6.2.2 werden die durch rotes Rauschen entstehenden Probleme bei der Periodendetektion diskutiert und es wird ein kurzer Überblick darüber gegeben, wie mit diesen Problemen aktuell in der Praxis umgegangen wird. Erste Vorschläge zu einer anderen Strategie werden in Abschnitt 6.2.3 unterbreitet. Dort werden auch Maßnahmen zur Übertragung dieser Strategie auf ungleichmäßig abständige Messzeiten und die in dieser Arbeit verwendeten Periodogramme diskutiert.

6.2.1. Rotes Rauschen im Allgemeinen

Zeitreihen von unabhängig identisch verteilten Beobachtungen werden weißes Rauschen genannt. Die Balken eines Fourier-Periodogramms sind in diesem Fall unabhängig identisch verteilt (vgl. Busch 2004, S. 4, 9). Abbildung 6.1 zeigt beispielhaft eine aus weißem Rauschen bestehende Zeitreihe und ihr Fourier-Periodogramm.
In der Praxis treten neben weißem Rauschen gelegentlich korrelierte Rauschtypen auf, deren Periodogrammbalken testperiodenabhängig verteilt sind. Ein häufig auftretender



(a) (b) (c)

Abbildung 6.2.: Brown'sche Bewegung als Beispiel für rotes Rauschen mit $\alpha = -2$. (a) Messwerte, generiert gemäß Gleichung (6.2), (b) Fourier-Periodogramm von (a), (c) Diagramm aus (b) mit logarithmierten Achsen, Abszisse Periode statt Frequenz. Zum Vergleich ist eine Strecke mit Steigung $-\alpha = 2$ eingezeichnet (grau).

Rauschtyp ist dabei das rote Rauschen. Hierbei erklären niedrigfrequente Schwingungen (mit langer Periode) einen höheren Anteil der Varianz in den Messungen $y_{r;1}, \ldots, y_{r;n}$ als hochfrequente (vgl. Israel und Stella 1996). Diese Umschreibung findet zum Beispiel in der Paläoklimatologie Anwendung. So modellieren Schulz und Mudelsee (2002) die rote Rauschkomponente einer Zeitreihe als AR(1)-Prozess.
In der Signalübertragung und auch in der Astroteilchenphysik wird rotes Rauschen häufig als Prozess definiert, der sich modellieren lässt durch

$$\begin{aligned} \log\left(\mathrm{Per}_{Fourier}(p_i)\right) &= \mu - \alpha \log\left(p_i\right) + \epsilon_i, \quad i = 1, \ldots, q \\ &= \mu + \alpha \log\left(f_i\right) + \epsilon_i, \end{aligned} \tag{6.1}$$

wobei p_i eine Testperiode bzw. $f_i = p_i^{-1}$ eine zu den Fourierfrequenzen gehörende Testfrequenz ist und ϵ_i ein Fehlerterm. Abkürzend wird häufig $\mathrm{Per}_{Fourier} \sim f^{\alpha}$ geschrieben (vgl. Benlloch García 2003, S. 40). Man sagt auch: Das Rauschen folgt einem Potenzgesetz. Der exponentielle Anstieg α ist dabei charakteristisch für den Prozess. Weißes Rauschen, als ein spezielles rotes Rauschen, folgt einem Potenzgesetz mit $\alpha = 0$ (vgl. Abbildung 6.1). Ein anderer bekannter Vertreter des roten Rauschens ist die Brown'sche Bewegung mit

$$Y_{r;t} = \sum_{i=1}^{t} e_i, \quad e_1, e_2, \ldots \sim \mathcal{N}(0, \sigma^2), \quad \sigma > 0, \quad t = 1, 2, \ldots \ . \tag{6.2}$$

Es ist allgemein bekannt (Press 1978) und in Abbildung 6.2 nachvollziehbar, dass der Prozess einem Potenzgesetz mit $\alpha = -2$ folgt.
Die Abbildungen 6.3(a)–(c) zeigen drei verschiedene Realisationen roten Rauschens, die einem Potenzgesetz mit $\alpha = -1{,}5$ folgen. Sie wurden gemäß der entsprechenden Algorithmen von Timmer und König (1995), Kasdin (1995) und Milotti (2006) generiert. Die zugehörigen Periodogramme sind in den Abbildungen 6.3(d)–(f) zu sehen.

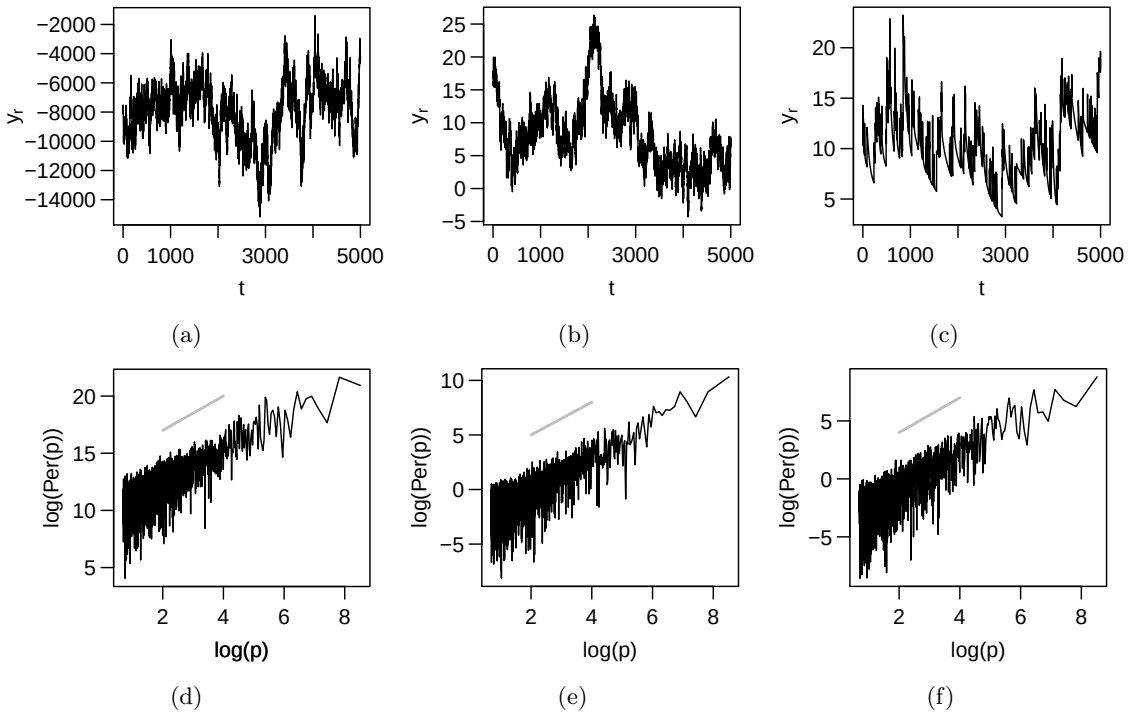


Abbildung 6.3.: Verschiedenartig generiertes Rotes Rauschen mit $\alpha = -1,5$. Generierung nach: (links) Timmer und König (1995), (mittig) Kasdin (1995), (rechts) Milotti (2006). Darstellung: (oben) Zeitreihe, (unten) Periodogramm mit logarithmierten Achsen. Zum Vergleich ist eine Strecke mit Steigung $-\alpha = 1,5$ eingezeichnet.

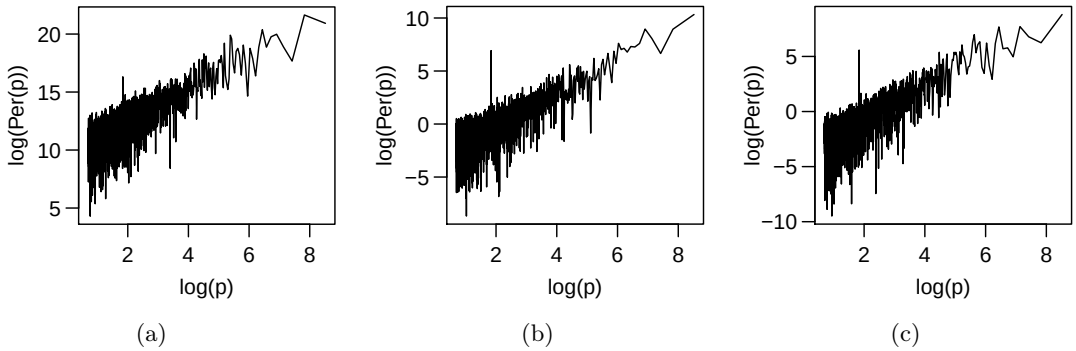


Abbildung 6.4.: Periodogramme mit logarithmierten Achsen drei verschiedener, mit einer Sinusschwingung der Periode 2π überlagerter Rauschtypen. Rauschgenerierung: (a) Timmer und König (1995), (b) Kasdin (1995), (c) Milotti (2006).

6.2.2. Die durch rotes Rauschen entstehende Problematik

Im Periodogramm einer Lichtkurve mit einer roten Rauschkomponente muss der zu einer periodischen Fluktuation gehörige Balken nicht der höchste im Periodogramm sein. Abbildung 6.4 zeigt die Periodogramme von Sinusfluktuationen der Periode $p_f = 2\pi$, die mit den Rauschzeitreihen aus Abbildung 6.3 überlagert wurden. Der Periodogrammbalken um $p_f = 2\pi$ ist lokal, aber nicht global maximal. Da die besprochenen Periodogrammmethoden darauf abzielen, einen Balken im Periodogramm auszumachen, der sich von den anderen Balken abhebt, stellt rotes Rauschen ein Problem für die Periodendetektion dar.
Die Umgehung dieses Problems wurde in der Praxis bereits behandelt. Israel und Stella (1996) schätzen mit Hilfe eines Mittelwertfilters ohne Einbezug der jeweils zentralen Beobachtung das erwartete Periodogramm $\widehat{\mathrm{Per}}(p_i), i = 1, \ldots, q$, und betrachten dann ein standardisiertes Periodogramm $\mathrm{Per}(p_i)/\widehat{\mathrm{Per}}(p_i)$. Wenn der zugrunde liegende Prozess ein AR-Prozess ist, besitzt dieses normierte Periodogramm laut den Autoren approximativ die gleiche Verteilung wie das Periodogramm von weißem Rauschen. Dieser Ansatz wird genauer bei Zelo (2013) untersucht. Benlloch García (2003), Kong et al. (2002), Halpern, Leighly und Marshall (2003) und Do et al. (2009) schätzen (mit unterschiedlichen Verfahren) den Parameter α und die Varianz des Rauschens, um dann durch die Simulation von rotem Rauschen die entsprechende Verteilung zu simulieren und für Signifikanzaussagen bezüglich der einzelnen Periodogrammbalken zu nutzen. Die Verfahren hängen damit stark von der simulierten Verteilung ab. Standardmäßig wird hierfür der Algorithmus von Timmer und König (1995) verwendet. Es ist jedoch unklar, inwiefern dieser allgemein genug ist, um zur Simulation der Verteilung eines beliebigen roten Rauschens zu dienen. Bei beiden Ansätzen wird die Struktur des roten Rauschens in der Periodogrammdarstellung geschätzt.

6.2.3. Ein Ansatz für ein Rauschfilter

Im hier skizzierten Ansatz wird vor der Periodogrammberechnung ein Filter auf die Zeitreihe angewendet, welches die rote Rauschkomponente entfernt. Hierzu wird eine Generierungsmethode von rotem Rauschen genauer untersucht und ein Vorschlag unterbreitet, wie mit ihrer Hilfe ein Filter entstehen kann. Das vorgeschlagene Filter wird auf Zeitreihen mit periodischer Fluktuation und roter Rauschkomponente angewendet. Erste Erfolge können, auch bei verschiedenartig simuliertem roten Rauschen oder echten Daten, festgestellt werden. Der Abschnitt schließt mit einem Ausblick.
Ein Ansatz zur Generierung von rotem Rauschen wird bei Kasdin (1995) vorgeschlagen. Dort wird die Rauschkomponente Y_r als AR(k)-Prozess mit $k = \infty$ generiert durch

$$Y_{r;t} = e_t + \sum_{i=1}^{k} (-a_i) Y_{r;t-i}, \qquad t \in \mathbb{Z} \tag{6.3}$$

$$\text{mit } a_i = \frac{\Gamma(i - \alpha/2)}{\Gamma(i+1)\Gamma(-\alpha/2)} \qquad \text{und } e_t \sim \mathcal{N}(0,1) \ \forall i \in \mathbb{Z},$$

wobei Γ die Gammafunktion bezeichnet. Es gilt $a_0 = 1$. Abbildung 6.3(b) zeigt die Approximation eines solchen Prozesses für $k = 6000$ statt $k = \infty$. Für die Approximation $k < \infty$ sind die Beobachtungen $Y_{r;k+1}, \ldots, Y_n$ gemäß Modell (6.3) auch in Matrixschreibweise

$$\begin{pmatrix} Y_{r;k+1} \\ Y_{r;k+2} \\ \vdots \\ Y_{r;n} \end{pmatrix} = \begin{pmatrix} Y_{r;1} & \ldots & Y_{r;k} \\ Y_{r;2} & \ldots & Y_{r;k+1} \\ \vdots & & \vdots \\ Y_{r;n-k} & \ldots & Y_{r;n-1} \end{pmatrix} \begin{pmatrix} a_k \\ \vdots \\ a_1 \end{pmatrix} + \begin{pmatrix} e_{k+1} \\ e_{k+2} \\ \vdots \\ e_n \end{pmatrix} \tag{6.4}$$

darstellbar.

Bei Vorliegen von Beobachtungen $y_{r;1}, \ldots, y_n$ ist die Idee, die Zufallsvariablen $Y_{r;1}, \ldots, Y_{r;n-1}$ auf der rechten Seite von Gleichung (6.4) durch die jeweiligen Beobachtungen zu ersetzen, den Parametervektor $(a_1, \ldots, a_k)^\mathsf{T}$ mit geeignetem k (vgl. unten) mittels Kleinste-Quadrate-Regression zu schätzen und die geschätzten Residuen $\widehat{e}_{k+1}, \ldots, \widehat{e}_n$ als gefilterte Zeitreihe zu betrachten. Auf dieser Zeitreihe kann dann Periodendetektion betrieben werden. Der Parameter α wird in dieser Prozedur nicht geschätzt.

Die Schätzung der durch das rote Rauschen in den Daten vorhandenen Struktur geschieht anhand der Zeitreihe und nicht anhand ihres Periodogramms. Damit ist die Prozedur unabhängig von der angewendeten Periodogrammmethode und für alle in dieser Arbeit verwendeten Regression-Modell-Kombinationen nutzbar. Bei Vorliegen einer Intervallstörung in den Daten (vgl. Abschnitt 3.2.4) müsste für das Filter vermutlich wie für die Regressionstechnik eine robuste Variante gewählt werden.

Abbildung 6.6 zeigt die Periodogramme der Rauschzeitreihen aus Abbildung 6.3 bei Filterung mit $k = 1, 4, 10$. Im Vergleich zu den Periodogrammen der ungefilterten Zeitreihen (vgl. Abbildung 6.3) zeigt sich, dass eine Abflachung des Periodogramms schon bei $k = 1$ stattfindet. Vor allem bei den gemäß Timmer und König (1995) und Milotti (2006) generierten Rauschzeitreihen wird dieser Effekt durch Erhöhen von k verstärkt. Andererseits sind in Abbildung 6.5 die Phasendiagramme drei künstlicher periodischer Fluktuationen vor und nach Anwendung eines Filters mit $k = 1, 2, 3$ zu sehen. Es ist deutlich zu erkennen, dass eine Filterung mit $k = 2$ die Sinusschwingung bereits vollständig entfernt. Dies wird theoretisch durch das Additionstheorem in Gleichung (2.19) (vgl. Seite 16) erklärt. Spätestens bei $k = p_f$ (sofern die Messzeitabstände und die Fluktuationsperiode ganzzahlig sind) wird jede periodische Fluktuation durch Anpassung der Koeffizienten $a_1 = a_2 = a_{p_f - 1} = 0, a_{p_f} = 1$ bei der Filterung vollständig entfernt.

Als erster Ansatz für das Entfernen roten Rauschens scheint eine Filterung mit $k = 1$ schon hilfreich. Dies entspricht der Anpassung eines AR(1)-Modells, dieses Filter könnte also auch in der Paläoklimatologie Anwendung finden. Hier soll allerdings der Einsatz bei rotem Rauschen untersucht werden, wie es in Gleichung (6.1) definiert ist. Die Periodogramme in Abbildung 6.7 wurden von den gleichen Zeitreihen wie die in Abbildung 6.4 berechnet, jedoch nach Filterung mit $k = 1$. Es ist deutlich erkennbar, dass bei allen die Fluktuationsperiode nach der Filterung die höchste ist.

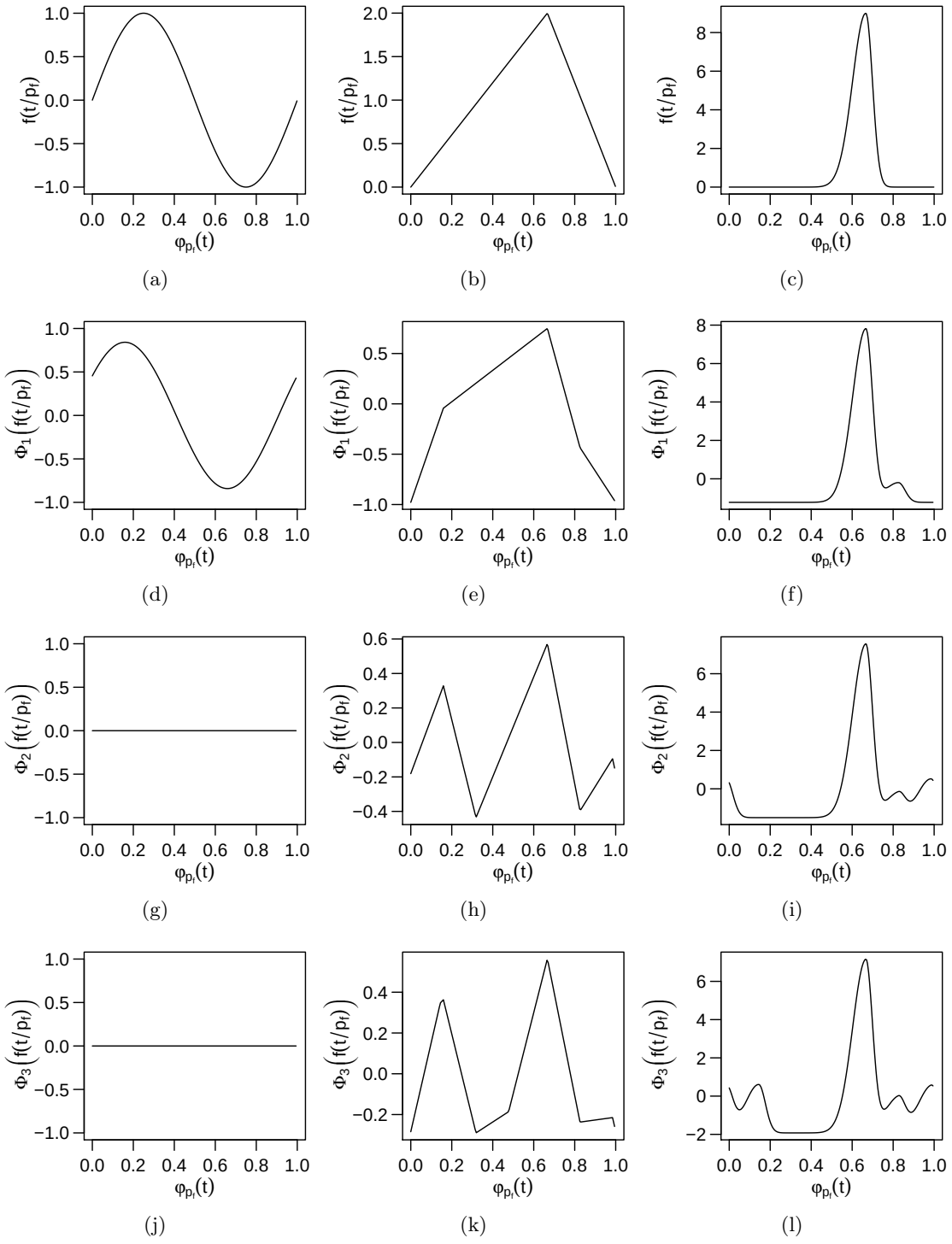


Abbildung 6.5.: Phasendiagramme drei verschiedener periodischer Fluktuationen mit Periode $p_f = 2\pi$ gefiltert mit verschiedenen k. Fluktuation: (links) Sinus, (horizontal mittig) Dreieck, (rechts) Peak. Filterung: (erste Zeile) ungefiltert, (zweite Zeile) $k = 1$, (dritte Zeile) $k = 2$, (vierte Zeile) $k = 3$.

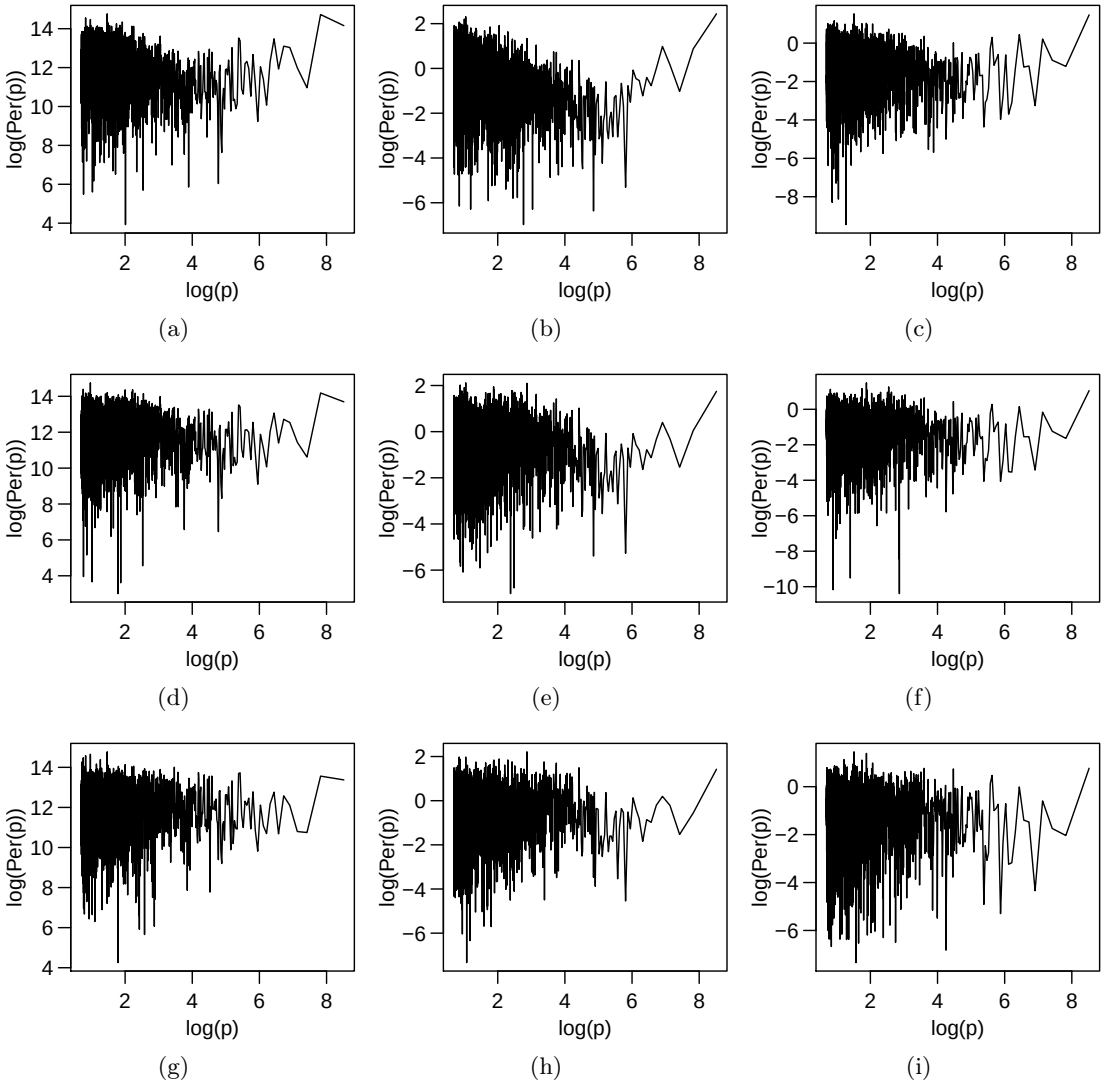
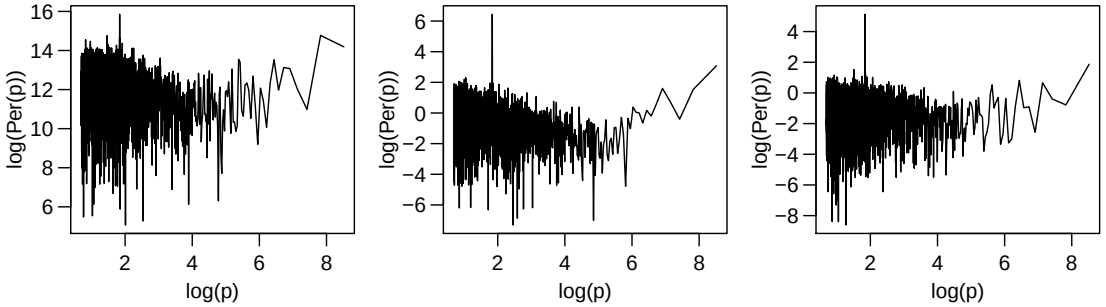


Abbildung 6.6.: Periodogramme mit logarithmierten Achsen verschiedener zuvor gefilterter Rauschzeitreihen mit $\alpha = 1,5$. Generierung nach: (links) Timmer und König (1995), (horizontal mittig) Kasdin (1995), (rechts) Milotti (2006). Filterung mit verschiedenen k: (oben) $k = 1$, (vertikal mittig) $k = 4$, (unten) $k = 10$. Für Periodogramme der ungefilterten Zeitreihen vgl. Abbildung 6.3.



(a) (b) (c)

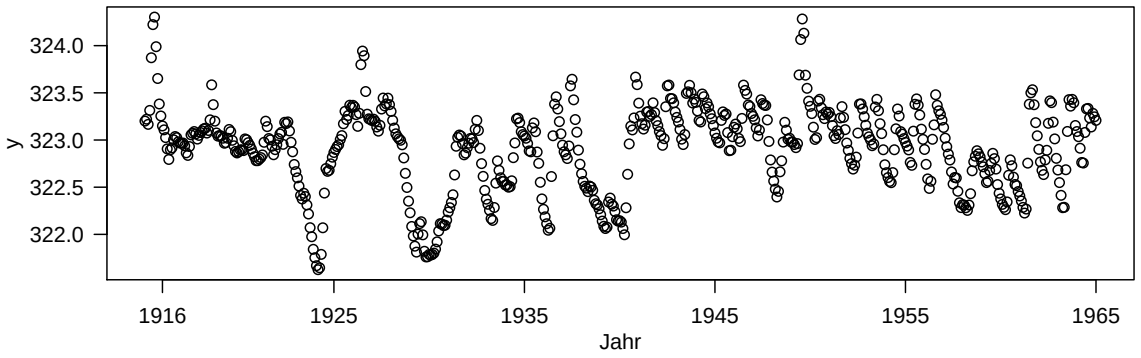
Abbildung 6.7.: Periodogramme mit logarithmierten Achsen drei verschiedener gefilterter mit Sinusschwingung überlagerter Rauschtypen. Filterung mit $k = 1$. Rauschgenerierung: (a) Timmer und König (1995), (b) Kasdin (1995), (c) Milotti (2006). Für Periodogramme der ungefilterten Zeitreihe vgl. Abbildung 6.4.

In Abbildung 6.8 wird die Anwendung auf ein echtes Datenbeispiel demonstriert: Die Zeitreihe[8] (Abbildung 6.8) stammt aus Water Survey of Canada (1992, S.51) und zeigt die monatliche Durchschnittstiefe des nordamerikanischen Sees Lake of the Woods in Metern. Im Periodogramm in Abbildung 6.8(b) ist ein doppelexponentialer Anstieg (linear durch die logarithmierten Achsen) zu erkennen. Die Periode zwölf (logarithmiert: 2.48, markiert durch ein Kreuz) sticht lokal hervor, ist aber im Gesamtperiodogramm nicht auffällig hoch. Abbildung 6.8(c) zeigt das mit $k = 1$ gefilterte Periodogramm (geschätzt wurde ein Koeffizient von $\widehat{a}_1 = 0.948$). Der Anstieg im Periodogramm ist verringert, der Periodogrammbalken bei Periode zwölf ist nun eindeutig maximal. Eine Periode von zwölf Monaten entspricht einem Jahresrhythmus und ist plausibel. Abbildung 6.8(d) zeigt die Messwerte gegen die Monate abgetragen. Auch hier ist eine periodische Fluktuation der Periode zwölf gut sichtbar.
Diese Ergebnisse motivieren eine zukünftige, detaillierte Untersuchung dieses Filterverfahrens einschließlich eines Vergleichs mit den in Abschnitt 6.2.2 erwähnten Vorschlägen zum Umgang mit rotem Rauschen. Dabei ist vor allem die Erweiterbarkeit auf ungleichmäßig beobachtete Zeitreihen und der Nutzen des Filters in Kombination mit den in dieser Arbeit betrachteten Periodogrammen ein interessantes Forschungsvorhaben.
Bei einer verfeinerten Methode könnte das Modell in Gleichung (6.4) durch das vom exponentiellen Anstieg α abhängige nichtlineare Modell (6.3) ersetzt werden. Hierbei würde unabhängig von k stets nur ein Parameter, nämlich α, angepasst. Dies sollte auch Filterungen mit $k > 1$ ermöglichen, ohne die periodische Fluktuation in den Messwerten zu entfernen. Gleichzeitig ließe sich das Modell umformulieren zu

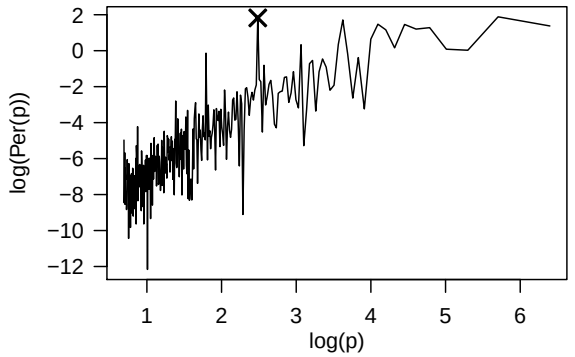
$$Y_{r;i} = e_i + \sum_{j=1}^{i-1} -a_\alpha(t_i - t_j) Y_{r;t_j}, \qquad i \in \{1, \dots, n\} \tag{6.5}$$

$$\text{mit } a_\alpha(\delta) = \begin{cases} \frac{\Gamma(\delta-\alpha/2)}{\Gamma(\delta+1)\Gamma(-\alpha/2)} & \delta \le k \\ 0 & \delta > k \end{cases} \qquad \text{und } e_i \sim \mathcal{N}(0,1), \quad i \in \{1, \dots, n\}.$$

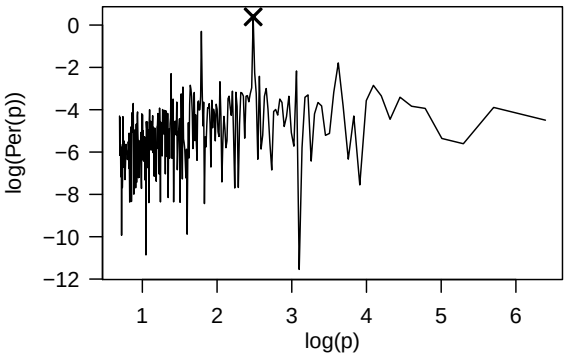
[8] Heruntergeladen am 27. August 2013 von:
http://datamarket.com/data/set/22vn/monthly-mean-water-levels-in-meters-lake-of-the-wood-at-warroad-1916-1965



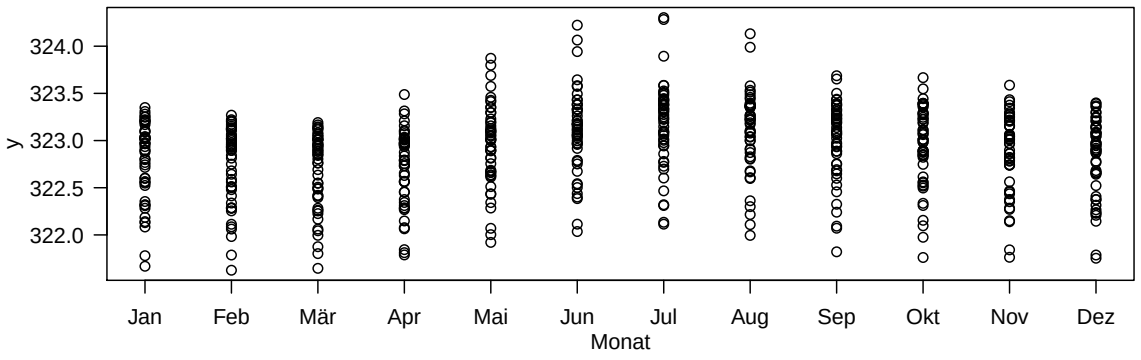
(a)



(b)



(c)



(d)

Abbildung 6.8.: Monatliche durchschnittliche Tiefe des Lake of the Woods, nordamerikanischer Kontinent, in Metern. Darstellung: (a) Zeitreihe, Abszisse beschriftet mit Jahreszahlen gemäß gregorianischem Kalender, (b) Periodogramm mit logarithmierten Achsen, (c) Periodogramm mit logarithmierten Achsen der mit $k = 1$ gefilterten Zeitreihe, (d) Phasendiagramm nach einer Periode von 12 Monaten. Die Kreuze in (b) und (c) markieren den Periodogrammbalken zu einer Testperiode von ungefähr zwölf.

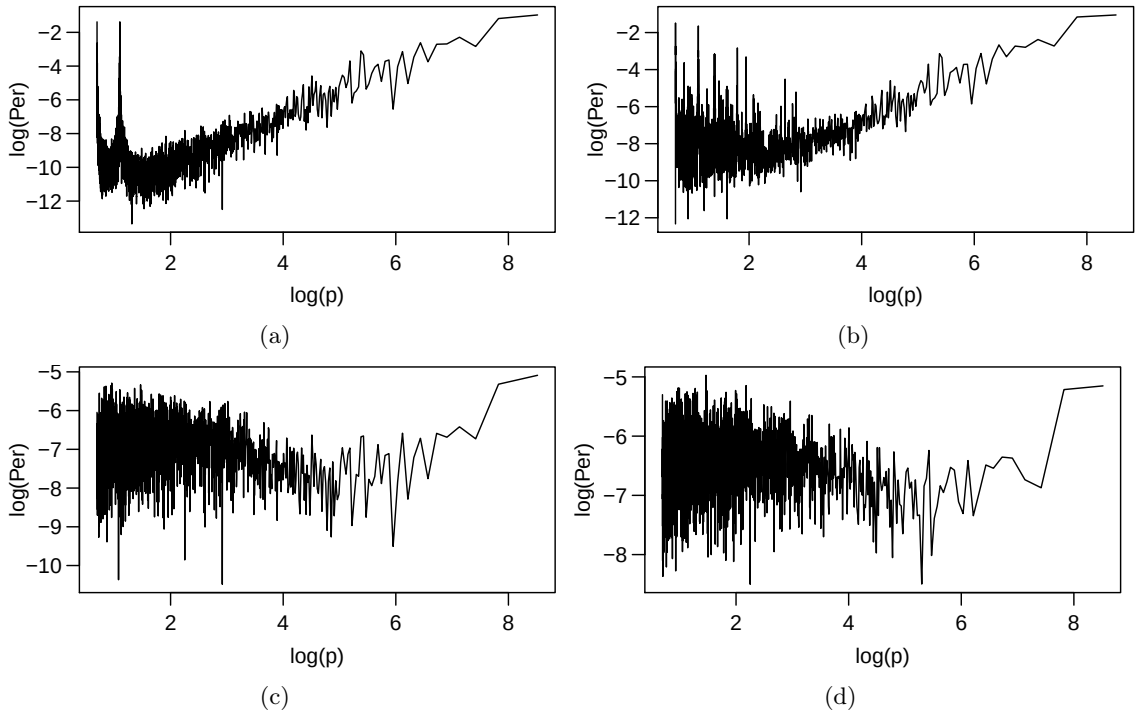


Abbildung 6.9.: Log-Log-Periodogramme der Rauschzeitreihe aus Abbildung 6.3(a), entstanden durch KQ-Regression. Angepasstes Modell: (links) Fouriersumme dritten Grades, (rechts) Einfachstufenfunktion. Filterung: (oben) keine, (unten) mit $k = 1$. Entsprechende Fourier-Periodogramme sind in den Abbildungen 6.3(d) und 6.6(a) dargestellt.

Für Messzeiten mit gleichen Abständen sind die Modelle (6.3) und (6.5) identisch, für ungleichmäßig beobachtete Messzeiten bietet jedoch nur Modell (6.5) einen Forschungsansatz.

Die Übertragbarkeit der Ergebnisse auf ein ungleichmäßiges Messzeitmuster ist von entscheidender Bedeutung für einen möglichen Einsatz in der Lichtkurvenanalyse. Zur Generierung von rotem Rauschen auf ungleichmäßigen Beobachtungszeiten wird üblicherweise eine aus rotem Rauschen bestehende, gleichmäßig beobachtete Zeitreihe mit möglichst hoher Auflösung generiert. Davon werden dann nur die für das ungleichmäßige Messzeitmuster benötigten Beobachtungen verwendet (vgl. Uttley, McHardy und Papadakis 2002). Eine entsprechende Version des Algorithmus' von Timmer und König (1995) zur Generierung von rotem Rauschen ist auch im R-Paket `RobPer` in der Funktion `lc_noise` implementiert (vgl. Abschnitt 3.2.3 und Tabelle G.3 im Anhang) und bei Thieler et al. (2013) zur Anwendung gekommen. Inwiefern die typischen Eigenschaften des jeweiligen Rauschprozesses bei diesem Vorgehen erhalten bleiben, muss noch untersucht werden. Zechmeister und Kürster (2009) weisen bei periodischem Messzeitmuster auf eine erhöhte Detektionshäufigkeit für die Samplingperiode p_s hin. Ähnliche Beobachtungen konnten auch auf den Simulationsdaten von Thieler et al. (2013) gemacht werden. Mögliche Erklärungen und Lösungen für dieses Phänomen wurden bisher noch nicht erarbeitet.

Weiterhin ist die Reaktion der Periodogramme dieser Arbeit auf rotes Rauschen ein interessanter Forschungsgegenstand. Abbildungen 6.9(a) und (b) zeigen durch KQ-Anpassung einer Fouriersumme dritten Grades oder einer Einfachstufenfunktion entstandene Periodo-

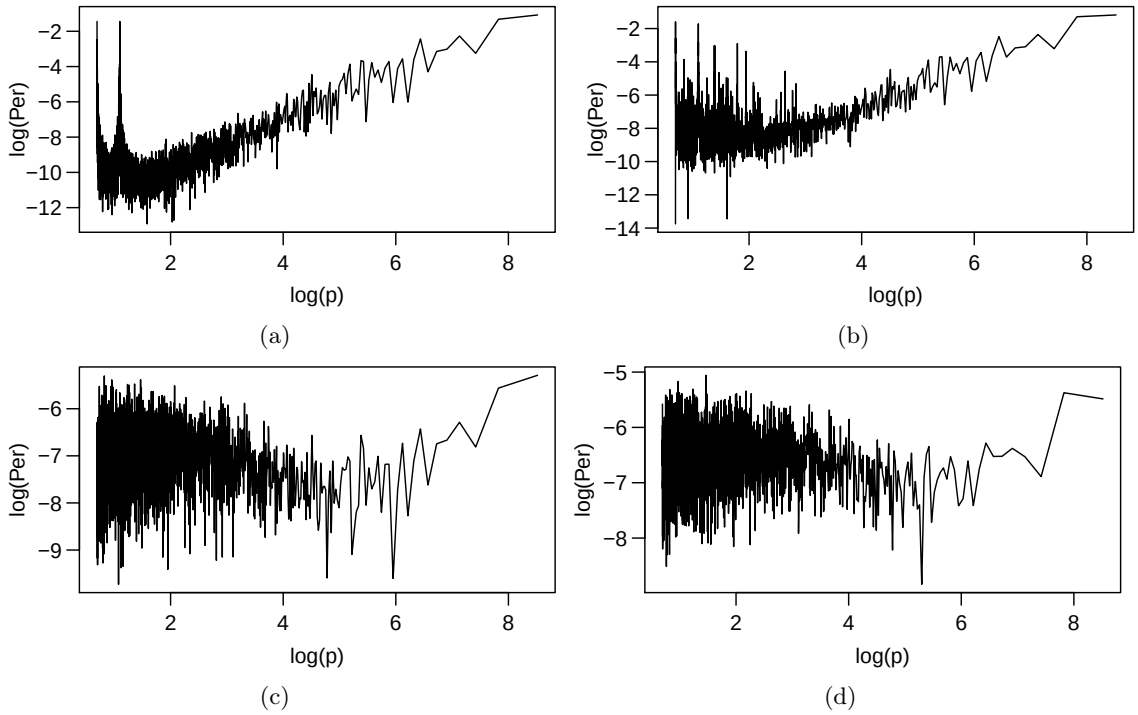


Abbildung 6.10.: Log-Log-Periodogramme der Rauschzeitreihe aus Abbildung 6.3(a), entstanden durch M-Huber-Regression. Angepasstes Modell: (links) Fouriersumme dritten Grades, (rechts) Einfachstufenfunktion. (Filterung: (oben) keine, (unten) mit $k = 1$. Vgl. Abbildung 6.9 für analog mit KQ-Regression berechnete Periodogramme.

gramme der Rauschzeitreihe aus Abbildung 6.3(a). Hier sind im Vergleich zum Periodogramm in Abbildung 6.3(d) (welches als Fourier-Periodogramm mit der Anpassung einer Sinusfunktion arbeitet, vgl. Abschnitt 2.2.1) zusätzlich erhöhte Periodogrammbalken zu erkennen. Diese kommen vermutlich durch die höhere Flexibilität der angepassten Funktionen und die damit entstehende zusätzliche Abhängigkeit zwischen den Periodogrammbalken zustande. Eine Filterung mit $k = 1$ (vgl. Abbildungen 6.9(c) und (d)) vermindert neben dem Anstieg der Periodogrammbalken auch diese zusätzlich erhöhten Balken. Die durch robuste Regressionstechniken entstandenen Periodogramme sehen ihrem jeweiligen KQ-Pendant ähnlich (vgl. Abbildung 6.10 für M-Huber-Regression). Detailliertere Vergleiche der Methoden auch in Situationen mit Intervallstörungen, Ausreißern und anderen periodischen Fluktuationen stellen ein weiteres Forschungspotential dar.

6.3. Zeitliche Strukturänderungen

Neben der Einführung einer zusätzlichen Rauschkomponente sind auch andere Modifikationen des Datenmodells aus Abschnitt 2.1 denkbar. Die Robustheit der entwickelten Methoden bezüglich kurzzeitiger Veränderungen der Rahmenbedingungen, wie etwa Ausreißer und Intervallstörungen, wurden in der Simulationsstudie (vgl. Kapitel 4) und exemplarisch in der Anwendung auf reale Daten (vgl. Kapitel 5) untersucht. Eine andere Herausforderung stellen

langfristige Veränderungen an die Rahmenbedingungen dar, zum Beispiel Veränderungen der Fluktuationsperiode oder der Gestalt der periodischen Fluktuation.
In diesem Abschnitt werden zwei Beispiele für langfristige Veränderungen behandelt. Im Zusammenhang mit Lichtkurven kann es zu einer Abhängigkeit der Periodenlänge von der Zeit kommen. Ziel ist hier vor allem, in Abhängigkeit von der Zeit die Fluktuationsperiode zu detektieren (vgl. Abschnitt 6.3.1). In Betonbohrexperimenten im Maschinenbau ändert die periodische Fluktuation sprungartig in jeweils einzelnen Segmenten des Zyklus ihren Verlauf (vgl. Abschnitt 6.3.2). Die Fluktuationsperiode ist dabei bekannt und ändert sich nicht, von Interesse ist die Modellierung der periodischen Fluktuation.

6.3.1. Veränderung der Fluktuationsperiode

Im Zusammenhang mit Lichtkurven kann die Länge der periodischen Fluktuation abhängig von der Zeit sein. In diesem Fall wird nicht ein Periodogramm der gesamten Lichtkurve, sondern fensterweise jeweils ein Periodogramm berechnet. Diese Technik wird Dynamic Power Spectrum genannt und für Zeitreihen mit gleichabständigen Beobachtungszeiten in der Astroteilchenphysik schon lange unter Verwendung des Fourier-Periodogramms genutzt (vgl. Oda et al. 1976). Für Lichtkurven mit ungleichmäßigen Beobachtungszeiten wurde statt des Fourier-Periodogramms auch schon das Lomb-Scargle-Periodogramm eingesetzt, etwa bei Wilms et al. (2001), Benlloch García (2003) oder Clarkson et al. (2003). Die Verwendung eines auf Anpassung einer Sinusfluktuation mittels L1-Regression basierenden Periodogramms wird im Bereich der Signalanalyse von gleichmäßig beobachteten Zeitreihen von Djurović, Katkovnik und Stanković (2001) vorgeschlagen. Eine Anwendung dieses Vorschlags im Bereich der Astroteilchenphysik ist nicht bekannt.
Auch der Einsatz der in dieser Arbeit betrachteten Periodogramme im Dynamic Power Spectrum ist möglich. Problematisch ist dabei der erhöhte Rechenaufwand. Eventuell kann dieser verringert werden, indem zunächst Periodogramme für einige über die Gesamtdauer der Lichtkurve verteilte Fenster berechnet werden. Für diese Periodogramme könnte zunächst die jeweilige Fluktuationsperiode geschätzt und die Gestalt der dazugehörigen periodischen Fluktuation betrachtet werden. Dieses Vorgehen ermöglicht dann die Wahl einer anzupassenden Funktion g mit möglichst niedrigdimensionalem Parametervektor $\beta \in \mathbb{R}^m$ für die komplette Analyse der Lichtkurve mittels gleitendem Periodogramm.
Die Verwendung von Detektionskriterien im Dynamic Power Spectrum ist bisher unüblich. Technisch gesehen kann der Ansatz dieser Arbeit entsprechend angepasst werden. Aufgrund seiner Adaptivität könnten hier weitere Untersuchungen lohnend sein, zum Beispiel um die Veränderung in der Fluktuationsperiode leichter zu verfolgen.
In Hinblick auf zeitveränderliche Strukturen stellt auch die Wavelet-Analyse (vgl. Bergh, Ekstedt und Lindberg 2007) einen interessanten Ansatz dar. Ähnlich wie bei der Fouriertransformation wird dabei eine Funktion in orthogonale Basisfunktionen zerlegt, die sich aber nicht wie bei der Fourieranalyse periodisch fortsetzen, sondern zeitlich verortet sind. Dies ermöglicht es, anhand der angepassten Koeffizienten auch zeitliche Entwicklungen der Zeitreihe nachzuvollziehen. Umgekehrt bedeutet es jedoch auch, dass eine Basisfunktion mit kleinem Träger bei einer Lichtkurve mit großen Beobachtungslücken nicht definierbar

oder nur anhand weniger Beobachtungen modellierbar ist. Die Anwendung von Wavelets auf ungleichmäßig beobachtete Lichtkurven sind beispielsweise bei Szatmáry, Vinkó und Gál (1994) und Foster (1996) dokumentiert. In beiden Arbeiten wird festgestellt, dass die Analyse auf periodische Strukturen im Messzeitmuster reagiert. Analog zu der Nutzung des Deeming-Periodogramms als Fourier-Periodogramm in ungleichmäßigen Zeitreihen (vgl. Abschnitt 2.2.1) wird aber auch dort die Definition der Koeffizienten für gleichmäßige Messzeiten übernommen. Eine Anpassung an ungleichmäßige Messzeitmuster durch Anwendung von KQ-Regression geschieht bei Mathias et al. (2004). Andere Arbeiten, die ein analoges Vorgehen mit einer robusten Regressionstechnik beschreiben, konnten nicht gefunden werden. Hier besteht also noch Forschungspotential.

6.3.2. Gestaltänderung der periodischen Fluktuation

Der vorige Abschnitt konzentriert sich auf eine Veränderung der Fluktuationsperiode. In einem anderen Kontext, der Schätzung der heterogenen Struktur von Beton, wird die sich verändernde Gestalt einer periodischen Fluktuation bei bekannter Fluktuationsperiode und gleichmäßig beobachteten Daten untersucht.
In einem Experiment wird zu gleichmäßig abständigen Messzeiten $t_1, \ldots, t_n$ die Kraft $y_1, \ldots, y_n$ gemessen, die ein Bohrer benötigt, um mit gleichbleibender Geschwindigkeit in ein Werkstück aus Beton einzudringen. Der Bohrer bewegt sich dabei auf einer Kreisbahn und die Beobachtungen y_i und y_{i+p_f}, $i = 1, \ldots, n - p_f$, werden jeweils an der gleichen Phase (Position auf der Kreisbahn) $\varphi_{p_f}(t_i) = \varphi_{p_f}(t_{i+p_f})$ gemessen. Da die Geschwindigkeit des Bohrers und der Radius der Kreisbahn bekannt sind, ist auch p_f bekannt. Es liegen nur vollständige Samplingzyklen vor, das heißt p_f ist ein Teiler von n.
Beton ist ein poröser Verbundstoff. Die zum Bohren benötigte Kraft ändert sich jeweils sprunghaft an der Begrenzung zwischen zwei Bestandteilen, im Folgenden Kante genannt. Es wird daher eine lokal konstante Fluktuation $Y_{f;1}, \ldots, Y_{f;n}$ angenommen, die die Niveauänderungen im Kraftaufwand $Y_1, \ldots, Y_n$ erklärt. Ein einfaches Modell hierzu ist

$$Y_i = y_{f;i} + Y_{e;i}, i = 1, \ldots, n, \tag{6.6}$$

wobei $y_{f;i}$ deterministisch die durch die Materialheterogenität hervorgerufenen Änderungen im Kraftaufwand beschreibt, und $Y_{e;i}$ eine Fehlerkomponente ist. Sie umfasst sowohl Rausch-, als auch mögliche Ausreißerkomponenten.
Durch die Bewegung des Bohrers auf einer Kreisbahn ist $y_{f;i}$ periodisch mit Fluktuationsperiode p_f. Da es aber auch in vertikaler Richtung zu Kanten im Beton kommen kann, verändert sich die periodische Fluktuation im Laufe eines Bohrexperiments sprunghaft und jeweils nur in Phasenabschnitten. Ein Phasendiagramm eignet sich nicht zur Betrachtung einer solchen periodischen Fluktuation, da ihre zeitliche Entwicklung so nicht gut erkennbar ist. Es bietet sich an, die Messwerte y_i nicht nur in Abhängigkeit von ihrer Phase $\varphi_{p_f}(t_i)$, sondern auch in Abhängigkeit von ihrer Zykluszugehörigkeit $z_{p_f}(t_i)$ (der Bohrtiefe) zu betrachten. In Abbildung 6.11(b) ist die Überlegenheit einer solchen Darstellung gegenüber

einem einfachen Phasendiagramm in Abbildung 6.11(a) für ein künstliches Datenbeispiel zu erkennen.
Zur Simulation von Bohrexperimenten sollen periodische Fluktuationen verwendet werden, die aus experimentellen Daten geschätzt wurden. Dazu werden die Beobachtungen $y_1 \ldots, y_n$ mit Hilfe eines Medianfilters (Tukey 1977, S. 110) abwechselnd in Richtung der Kreisbahn (in Abbildung 6.11(b) horizontal) und in Richtung der Bohrtiefe (in Abbildung 6.11(b) vertikal) geglättet. Das Ergebnis einer solchen Filterung ist in Abbildung 6.11(c) dargestellt. Das Medianfilter ist einerseits robust gegenüber Ausreißern, konserviert andererseits Sprünge. Im Vergleich zu einem aufwändigeren, auf LTS-Regression basierenden Filter arbeitet es auf den vorliegenden Daten schneller und erhält Sprünge besser. Weitere Details zu diesem Verfahren inklusive notwendiger Datenvorverarbeitung sowie eine Einbettung dieses Teilaspekts in die Gesamtproblematik der Modellierung und Simulation von Betonbohrungen sind bei Raabe et al. (2012) zu finden.
Fried, Raabe und Thieler (2012) beschäftigen sich darauf aufbauend mit der Kantenerkennung in der periodischen Fluktuation $y_{f;1}, \ldots y_{f;n}$. Allgemein bedeutet eine Niveauänderung zwischen den Zeitpunkten i und $i+1$ in einem Fenster mit Beobachtungen $x_{i-m+1}, \ldots, x_{i+m}$ eine erhöhte mediane Differenz

$$\widehat{\Delta}_m(x_i) = \operatorname{med}\left(x_{i+j_r} - x_{i+j_l}\right)_{j_r=1,\ldots,m,\ j_l=-m+1,\ldots,0}.$$

Die medianen Differenzen werden sowohl in horizontale Richtung (angewendet auf die Zeitreihe $(y_{f;i})_{i=1,\ldots,n}$) als auch in vertikaler Richtung (angewendet auf die Zeitreihen $\left(y_{f;i\cdot p_f+1}\right)_{i=0,\ldots,n/p_f-1}, \left(y_{f;i\cdot p_f+2}\right)_{i=0,\ldots,n/p_f-1}, \ldots, \left(y_{f;i\cdot p_f+p_f}\right)_{i=0,\ldots,n/p_f-1}$) berechnet, jeweils unter Auslassung der Randbereiche, in denen $\widehat{\Delta}_m$ nicht definiert ist.
Da eine feine Körnung und somit viele Sprünge in der Gestalt der periodischen Fluktuation erwartet werden und da in vertikaler Richtung häufig nicht viele Beobachtungen für die gleiche Phase vorliegen, werden kleine Fensterbreiten m gewählt. Statt der Nutzung von $\widehat{\Delta}_m(x_i)$ schlagen Fried und Dehling (2011) auch andere Maße zur Niveauänderung vor, empfehlen jedoch den Median speziell für kleine Fensterbreiten.
Zur Detektion von Sprüngen werden die beiden Mengen der horizontal und der vertikal berechneten medianen Differenzen separat bezüglich Ausreißern untersucht. Das Verfahren ist dabei ähnlich wie bei der Detektion von auffälligen Periodogrammbalken in Abschnitt 2.6: Mittels Cramér-von-Mises-Distanz-Minimierung wird eine geeignete Verteilung angepasst und extreme Werte werden detektiert. Die Detektionen in horizontaler und vertikaler Richtung werden zu einer Detektionskarte kombiniert (vgl. Abbildung 6.11(d)). In künstlichen Beispielen erhalten Fried, Raabe und Thieler (2012) auf diese Weise breitere geschätzte Kantenbereiche als durch ad-hoc-Anwendung des in der Bildbearbeitung etablierten Canny-Detektors (Canny 1986, vgl. Abbildung 6.11(e)) auf die geschätzte periodische Fluktuation $\widehat{y}_f$, finden aber auch mehr der vorhandenen Kanten. Eine Weiterverarbeitung der Detektionskarten, eventuell mit den im Canny-Detektor implementierten kantenausdünnenden Prozeduren, könnte die Kantenschätzung noch verbessern.

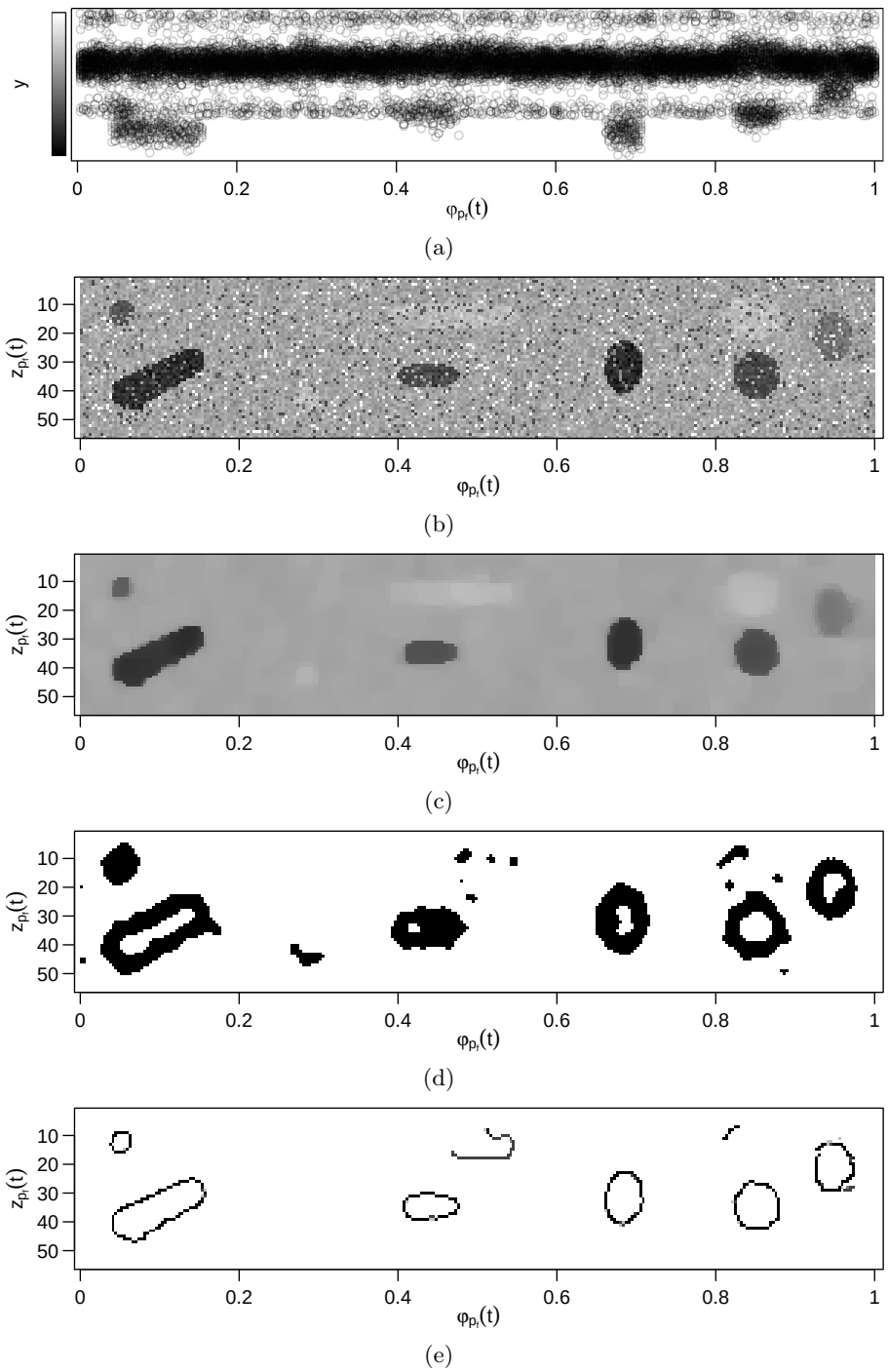


Abbildung 6.11.: Künstliches Beispiel aus Fried, Raabe und Thieler (2012) (für echte Daten vgl. ebendort und Raabe et al. 2012) für Bohrdaten: (a) Phasendiagramm zu bekannter Periode p_f mit Grauwertskala (Ordinate). (b)/(c) $y_i/\widehat{y}_{f;i}$ nach Phase $\varphi_{p_f}(t_i)$ (Abszisse) und Zyklus $z_{p_f}(t_i)$ (Ordinate) geordnet und als Grauwert entsprechend der Skala in (a) kodiert. (d)/(e) Durch mediane Differenzen / Canny-Detektor ermittelte Kantenbereiche (schwarz).

7. Zusammenfassung und Ausblick

Es kann festgestellt werden, dass robuste Regressionstechniken sinnvoll bei der Periodendetektion gestörter Lichtkurven einsetzbar sind. Die in der Astroteilchenphysik beliebte Stufenfunktion zu Anpassung kann gut mit robusten Techniken wie M-Huber- oder τ-Regression kombiniert werden, was im Rahmen dieser Arbeit und der dabei entstandenen Publikationen erstmals geschieht. Die ebenfalls beliebte Sinusfunktion schneidet in Simulationen vergleichsweise schlecht ab. Von der Berücksichtigung vorliegender Messfehler durch Gewichtung der Regression wird aufgrund der Simulationsergebnisse abgeraten. Das in dieser Arbeit entwickelte Detektionsprinzip für auffällig hohe Periodogrammbalken eignet sich in den betrachteten Szenarien besser als das unter Standardannahmen. Erste Ideen zum Umgang mit modifizierten Modellannahmen wie der Zulassung einer roten Rauschkomponente oder einer zeitveränderlichen Fluktuationsperiode werden präsentiert und erfolgreich auf kleine Datenbeispiele angewendet. Der Rest dieses Kapitels enthält einen ausführlicheren Überblick über die Ergebnisse dieser Arbeit und sich daraus ergebende Ansätze zu weiterer Forschung.

In dieser Arbeit werden Periodogramme betrachtet, die sich zur Analyse von Lichtkurven eignen. Lichtkurven sind spezielle Zeitreihen der Astroteilchenphysik. Die Beobachtung einer Lichtkurve besteht aus Messzeit, Messwert und Messfehler. Die Messzeiten liegen ungleichmäßig und weisen häufig ein periodisches Messzeitmuster auf. Die Messfehler geben an, wie genau die zugehörigen Messwerte gemessen werden konnten. Die Messwerte können neben einer Rauschkomponente auch eine periodische Fluktuation aufweisen, die es zu detektieren gilt. In der Praxis können zudem Störungen wie Ausreißer oder Intervallstörungen auftauchen. Mit Intervallstörungen sind stark erhöhte Messwerte gemeint, die zeitlich zusammenhängen (in einem Zeitintervall liegen).

Wegen des unregelmäßigen Messzeitmusters ist es nicht möglich, ein klassisches Fourier-Periodogramm für die Periodendetektion zu verwenden. Alternativen bilden Periodogramme, die durch Anpassung einer periodischen Funktion mit verschiedenen Testperioden berechnet werden. Am häufigsten wird hierbei eine Sinusfunktion mittels Kleinste-Quadrate-Regression angepasst, aber andere periodische Funktionen und Regressionstechniken sind möglich.

In dieser Arbeit wird der Nutzen robuster Regression beim Einsatz in Periodogramm- und darauf aufbauenden Detektionsmethoden mit einer Simulationsstudie untersucht. Eine Auswahl der besten Methoden der Simulationsstudie und einiger Standardmethoden werden zusätzlich auf echte Daten aus der Astroteilchenphysik, der Astrophysik und der Paläoklimatologie angewendet. Zur Anpassung der periodischen Funktion werden neben der Kleinste-Quadrate (KQ)-Regression die Kleinste-Beträge (L1)-Regression, die S-Regression, die τ-Regression, und die M-Regression unter Nutzung der Huber-Funktion (M-Huber) oder der Tukey-Funktion (M-Tukey) betrachtet. Als anzupassende periodische Funktionen werden

eine periodische Stufenfunktion mit zehn Stufen, die Sinusfunktion, Fouriersummen zweiten und dritten Grades und eine periodische kubische Splinefunktion mit vier Knotenpunkten pro Zyklus verwendet. Der Periodogrammbalken einer Testperiode ist identisch mit dem Bestimmtheitsmaß der entsprechenden Anpassung. Eine Berücksichtigung der Messfehler kann durch Gewichtung der Anpassung stattfinden und ist in der Astroteilchenphysik häufig zu finden. Dies bringt jedoch in der im Rahmen dieser Arbeit durchgeführten Simulationsstudie keine relevanten Vor-, häufig aber Nachteile. Es wird daher von der Berücksichtigung von Messfehlern mittels gewichteter Regression abgeraten.
Viele in der Vergangenheit vorgeschlagene Periodogramme funktionieren analog zu einer der hier betrachteten Methoden. Andere, etwa alle auf S- oder τ-Regression basierenden Periodogramme oder die auf gewichteter oder robuster Anpassung einer Stufenfunktion oder Fouriersumme basierende Periodogramme, werden in dieser Arbeit oder daraus entstandenen Publikationen erstmalig vorgeschlagen.
Zur Detektion von auffällig hohen Balken im Periodogramm, welche auf eine Periodizität hinweisen, wird eine Betaverteilung an die Balken angepasst. Es werden Periodogrammbalken detektiert, die höher als ein zuvor bestimmtes Quantil dieser Verteilung liegen. Die Betaverteilung wird dabei durch die Verteilung des Bestimmtheitsmaßes der KQ-Regression im Falle normalverteilter Messwerte motiviert. Da ein hoher Periodogrammbalken seine eigene Detektion durch Beeinflussung der Verteilungsschätzung nicht verhindern soll, wird die Verteilung robust mittels Cramér-von-Mises-Distanz-Minimierung angepasst.
In den Simulationsstudien kann beobachtet werden, dass mit robusten Regressionstechniken in Szenarien mit Intervallstörungen erheblich bessere Detektionsergebnisse erbracht werden können als mit Methoden, die auf KQ-Regression basieren. Am besten schneiden die Methoden ab, die auf τ-Anpassung oder M-Huber-Anpassung einer Stufenfunktion, oder auf M-Huber oder L1-Anpassung einer Fouriersumme dritten Grades basieren. Diese Methoden halten in der Simulationsstudie das Signifikanzniveau ein und detektieren die meisten periodischen Fluktuationen. Keine dieser vier Methoden wurde bisher außer in im Rahmen dieser Arbeit entstandenen Publikationen zur Periodogrammberechnung vorgeschlagen. Bei Anwendung dieser Methoden auf reale Datensätze und Vergleich mit den entsprechenden KQ-basierten Periodogrammen kann beobachtet werden, dass sich die auf robuster Regression basierenden Periodogrammmethoden bezüglich ihrer Gestalt robust gegenüber Ausreißern und Intervallstörungen verhalten.
In einigen der echten Lichtkurven können Hinweise auf Periodizität gefunden werden, die in der Literatur bisher keine Erwähnung finden. Besonders interessant ist hierbei eine Periode von 51 Tagen für Photonenemmissionen der Quelle Gro J1719–24. Sie kann mit verschiedenen auf robuster Regression basierenden Methoden, aber mit keiner auf KQ-Regression basierenden Methode gefunden werden. Detektionen sollten hierbei nicht als signifikant angesehen werden. Dafür ist zu wenig über die Deckung der gemachten Modellannahmen mit der Realität bekannt. Außerdem wurde eine hohe Anzahl Periodogramme betrachtet, die durch die Anwendung auf die gleichen Lichtkurven untereinander abhängig sind. Die hier gemachten Detektionen können aber als Hinweise angesehen werden und neue Beobachtungen können zukünftig für ihre Überprüfung genutzt werden.

Nicht alle robusten Regressionstechniken führen in der Simulationsstudie zu besseren Detektionsergebnissen. Für die LTS-Regression muss festgestellt werden, dass die Annahme von betaverteilten Periodogrammbalken im Anwendungsfall nicht haltbar ist. Die auf dieser Technik basierenden Detektionsmethoden erzielen meist schlechtere Detektionsraten als die anderen Methoden und halten das Signifikanzniveau nicht ein.
Ein Problem für die robusten Regressionstechniken stellen außerdem solche periodische Fluktuationen dar, die schlecht mit der anzupassenden periodischen Funktion nachzuzeichnen sind. Speziell auf S-Regression basierende Methoden erzielen schlechte Detektionsraten, wenn die anzupassende periodische Funktion nicht eine der tatsächlich vorliegenden periodischen Fluktuation sehr ähnliche Gestalt besitzt. Da viele auf S-Regression basierende Methoden gleichzeitig in Abwesenheit einer periodischen Fluktuation das Niveau einhalten, kann der Einsatz solcher Methoden bei genauer Kenntnis der Gestalt der potentiell vorhandenen periodischen Fluktuation dennoch sinnvoll sein. Die auf KQ-Regression basierenden Methoden sind am wenigsten abhängig von der Kenntnis der genauen Gestalt der potentiellen periodischen Fluktuation, können jedoch nur auf ungestörten Daten angewendet werden.
Eine Verbesserung der Detektionsergebnisse für S-, τ- und M-Regression kann eventuell durch den Einsatz anderer ρ-Funktionen erreicht werden. Die so entstehenden Methoden sind womöglich weniger robust, dafür toleranter gegenüber Abweichungen der Gestalt der anzupassenden Funktion von der der periodischen Fluktuation. Weiterhin können Periodogramme nach dem hier verfolgten Prinzip entwickelt werden, die auf anderen Regressionstechniken basieren.
Neben den bereits genannten Periodogrammen werden auch solche betrachtet, bei denen zu jeder Testperiode zwei Stufenfunktionen angepasst werden. Dabei hat die eine Stufenfunktion ihre Sprungstellen jeweils bei den Stufenmitten der anderen. Die separate Anpassung zweier solcher Funktionen wird untersucht, weil sie in der Astroteilchenphysik im beliebten Phase-Dispersion-Minimization-Periodogramm unter Verwendung von KQ-Regression durchgeführt wird. In der Simulationsstudie sind diese Periodogramme denen auf Anpassung jeweils einer einzelnen Stufenfunktion basierenden Periodogrammen nicht überlegen und erheblich rechenintensiver. Daher wird von ihrer Anwendung abgeraten.
Von den anzupassenden periodischen Funktionen zeigt sich die Sinusfunktion als eine der ungeeigneteren. Auf ihr basierende Methoden halten häufig das Signifikanzniveau nicht ein und zur Detektion von nicht-sinusförmigen Fluktuationen eigenen sich andere periodische Fluktuationen besser. Diese Feststellung ist besonders interessant, weil die KQ-Anpassung einer Sinusfunktion zusammen mit Anpassung einer Stufenfunktion bislang zu den beliebtesten Ansätzen in der Periodendetektion in Lichtkurven gehört. Andererseits zeigen sich auf der Anpassung hochparametrischer Funktionen basierende Detektionsmethoden weniger erfolgreich, wenn die zu Grunde liegende periodische Fluktuation eine einfachere Gestalt hat. Bei Anpassung einer zu flexiblen Funktion kommt es auch bei Mehrfachen der Fluktuationsperiode zu erhöhten Periodogrammbalken, was zur Nichtdetektion der Fluktuationsperiode führen kann.
Es lässt sich also nicht grundsätzlich sagen, welche periodische Funktion zur Periodendetektion bei Lichtkurven geeignet ist. Expertenwissen kann hier, ebenso wie bei der Wahl

der Testperiodenmenge, sehr hilfreich sein. Eine Möglichkeit ist auch, einen Teil der vorliegenden Daten explorativ mit mehreren Periodogrammmethoden mit unterschiedlichen anzupassenden Funktionen zu untersuchen, um anschließend konfirmative Analysen auf dem anderen Teil durchzuführen. Für zukünftige Untersuchungen kann es auch lohnend sein, bei der periodischen Splinefunktion die Knotenanzahl pro Zyklus zu erhöhen, um der Funktion eine höhere Flexibilität zu geben. Auch die adaptive Bestimmung der Anzahl von Sprüngen bzw. Knoten der Stufen- bzw. Splinefunktion und ihrer Position weist Forschungspotential auf. Alternativ könnte sogar auf eine parametrische Funktion zur Anpassung der periodischen Fluktuation verzichtet und eine Funktion mittels Glättung anpasst werden. Bei der Entwicklung dieser adaptiven Ansätze muss jedoch bedacht werden, dass eine erhöhte Anpassungsfähigkeit der Funktion auch immer bedeutet, solche periodische Funktionen besser anpassen zu können, deren Periode nicht die gesuchte Fluktuationsperiode ist.
Neben der oben beschriebenen werden noch zwei andere Verfahren zur Detektion von auffällig hohen Periogrammbalken betrachtet. Bedingt durch die Betrachtung von ungleichmäßigen Messzeiten und beliebigen Testperiodenmengen kommt es zu unerwünschten Abhängigkeiten unter den Periodogrammbalken. In der Anwendung erweisen sich dabei vor allem die positive Abhängigkeit benachbarter Testperioden und die Abhängigkeit von Vielfachen der Fluktuationsperiode als störend. Um diese Abhängigkeiten zu reduzieren, werden auch zwei Möglichkeiten betrachtet, das Periodogramm vor Anpassung einer Betaverteilung auszudünnen. Die Detektionsmethoden ohne Ausdünnung des Periodogramms erweisen sich jedoch als erfolgreicher. Eventuell ist es bei zukünftigen Untersuchungen sinnvoll, nicht das Periodogramm selbst auszudünnen, wohl aber die Anzahl der beteiligten Periodogrammbalken für die Berechnung des Schwellwerts zu reduzieren. Dies könnte auch Detektionsschwierigkeiten entgegenwirken, die ohne Ausdünnung bei Nutzung einer zu großen Testperiodenmenge auftreten.
Zur Umsetzung der Simulationsstudie sowie der Anwendung auf echte Daten wird das R-Paket RobPer verwendet, welches im Rahmen dieser Arbeit entstanden ist. Die Kombination aller periodischen Funktionen, Regressionstechniken, ob gewichtet angepasst wird oder nicht und inwiefern das Periodogramm vor Anpassung einer Verteilung ausgedünnt wird, führt zu 252 Detektionsmethoden. Sie werden in der Simulationsstudie auf 20 verschiedene Lichtkurventypen angewendet. RobPer basiert teilweise auf bereits bestehenden R-Funktionen, von denen einige an die spezifischen Anforderungen der Periodogrammberechnung angepasst werden. Viele Teile des Pakets werden auch im Rahmen dieser Arbeit selbst implementiert. Die Funktion zur Periodogrammberechnung benötigt hierbei, je nach gewählter Methode, mit bis zu einer halben Stunde pro Periodogramm relativ viel Rechenzeit. Für die Zukunft kann hier eine Optimierung der Funktion, beispielsweise durch Parallelisierung der Anpassung verschiedener Testperioden, angedacht werden.
Die Abschwächung der Modellannahmen führt zu interessanten Forschungsfragen. Beispielsweise verursacht die Präsenz von sogenanntem roten Rauschen in den Messwerten zu einem Trend im Periodogramm. Für eine erfolgreiche Detektion kann dann ein zusätzlicher Bearbeitungsschritt notwendig werden. Ein auf Filterung basierender Ansatz wird in dieser Arbeit vorgeschlagen und erfolgreich in einigen simulierten und einem realen Datenbeispiel angewendet. Eine andere Lockerung in den Modellannahmen stellt die Zulassung einer

zeitabhängigen Fluktuationsperiode dar. Zum Umgang damit sind in der Vergangenheit Verfahren vorgeschlagen worden. Der Einsatz der in dieser Arbeit betrachteten Periodogramme in diesen Verfahren ist leicht umzusetzen und verspricht auch hier Robustheit gegen Ausreißer und Intervallstörungen. Eine andere Modifikation der Modellannahmen ist eine zeitveränderliche periodische Fluktuation. Beispiele und Anwendungen hierführ, wie sie im Maschinenbau zu finden sind, werden diskutiert.

A. Nachweis der Gleichungen (2.23) bis (2.28) aus Abschnitt 2.3

Gegeben sei die Lichtkurve $(t_i, y_i, s_i)_{i=1,\dots,n}$ und eine Testperiode p. Entsprechend seien die Phasen $\varphi_i = \varphi_p(t_i)$, $i = 1, \dots, n$ (vgl. Seite 9). Für alle Nachweise dieses Abschnittes werden Stufenfunktionen benötigt. Sei dazu stets $\mathcal{I}$ eine Partition des Intervalls [0,1[und $g : [0,1[\to \mathbb{R}$ eine Stufenfunktion so, dass g auf jeder Menge der Partition $\mathcal{I}$ konstant ist. Die Kleinste-Quadrate-Schätzung von g sei bezeichnet mit $\widehat{g}$ und SE sei gemäß Gleichung (2.12) auf Seite 11 definiert durch

$$\mathrm{SE} = \sum_{i=1}^{n} \left(y_i - \widehat{g}(\varphi_i)\right)^2 . \tag{A.1}$$

Für das Bestimmtheitsmaß R^2 gilt wie stets für Kleinste-Quadrate-Regression

$$R^2 = 1 - \frac{\mathrm{SE}}{\mathrm{SY}} = \frac{\mathrm{SR}}{\mathrm{SY}} \tag{A.2}$$

$$\text{mit } \mathrm{SR} = \sum_{i=1}^{n} \left(\widehat{g}(\varphi_i) - \bar{y}\right)^2 = \mathrm{SY} - \mathrm{SE}. \tag{A.3}$$

Die Mengen der Partition $\mathcal{I}$ seien bezeichnet mit $I_1, \dots, I_h$ und es gilt (als Spezialfall für Kleinste-Quadrate-Regression der in Abschnitt 2.5 eingeführten Notation)

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

$$\widehat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 , \tag{A.4}$$

$$n_l = \left|\{\varphi_1, \dots, \varphi_n\} \cap I_l\right|, \quad l = 1, \dots, h, \tag{A.5}$$

$$\bar{y}_l = \frac{1}{n_l} \sum_{i:\varphi_i \in I_l} y_i, \quad l = 1, \dots, h,$$

$$\widehat{\sigma}_l^2 = \frac{1}{n_l - 1} \sum_{i:\varphi_i \in I_l} \left(y_i - \bar{y}_l\right)^2 \quad l = 1, \dots, h. \tag{A.6}$$

Gemäß Gleichung 2.11 (Seite 11) gilt mit dieser Definition von $\widehat{\sigma}^2$ auch

$$\mathrm{SY} = (n-1)\widehat{\sigma}^2. \tag{A.7}$$

Da jede Stufe durch einen Kleinste-Quadrate-Lokationsschätzer aus den zughörigen Beobachtungen geschätzt werden kann, gilt außerdem für $\varphi \in [0,1[$

$$\widehat{g}(\varphi) = \bar{y}_l, \text{wobei } l \in \{1,\ldots,h\} \text{ so, dass } \varphi \in I_l \tag{A.8}$$

A.1. Epoch-Folding-Periodogramm

Das Epoch-Folding-Periodogramm von Leahy et al. (1983) ist für feste Testperiode p gegeben durch

$$\mathrm{Per}_{EF} = \sum_{l=1}^{m} (\bar{y}_l - \bar{y})^2 \frac{n_l}{\widehat{\sigma}^2},$$

wobei sich $\bar{y}_l$ und n_l auf die m Mengen der Partition $\mathcal{I} = \{[0,k_1[,[k_1,k_2[\ldots,[k_{m-1},1[\}$ mit $0 < k_1 < k_2 < \ldots < k_{m-1} < 1$ beziehen. Dann ist

$$\begin{aligned}
\mathrm{Per}_{EF} &= \sum_{l=1}^{m} (\bar{y}_l - \bar{y})^2 \frac{n_l}{\widehat{\sigma}^2} \\
&\overset{(A.7)}{=} \sum_{l=1}^{m} (\bar{y}_l - \bar{y})^2 \frac{n_l(n-1)}{\mathrm{SY}} \\
&\overset{(A.8)}{=} \frac{n-1}{\mathrm{SY}} \sum_{l=1}^{m} n_l \left(\widehat{g}(\varphi)_{\varphi\in I_l} - \bar{y}\right)^2 \\
&\overset{(A.5)}{=} \frac{n-1}{\mathrm{SY}} \sum_{i=1}^{n} \left(\widehat{g}(\varphi_i) - \bar{y}\right)^2 \\
&\overset{(A.3)}{=} (n-1)\frac{\mathrm{SR}}{\mathrm{SY}} \\
&\overset{(A.2)}{=} (n-1)R^2.
\end{aligned}$$

□

Dies ist die in Gleichung (2.23) auf Seite 20 gewählte Darstellung.

A.2. Analysis-of-Variance-Periodogramm

Das Analysis-of-Variance-Periodogramm von Schwarzenberg-Czerny (1989) ist auf der gleichen Partition $\mathcal{I}$ wie das Epoch-Folding-Periodogramm definiert durch

$$\mathrm{Per}_{AoV} = \frac{\frac{1}{m-1}\sum_{l=1}^{m} n_l \left(\bar{y}_l - \bar{y}\right)^2}{\frac{1}{n-m}\sum_{l=1}^{m}(n_l-1)\widehat{\sigma}_l^2}$$

und damit

$$\begin{aligned}
&\overset{(A.6)}{=} \frac{n-m}{m-1} \frac{\sum_{l=1}^{m} n_l \left(\bar{y}_l - \bar{y}\right)^2}{\sum_{l=1}^{m}(n_l-1)\frac{1}{n_l-1}\sum_{i:\varphi_i\in I_l}\left(y_i - \bar{y}_l\right)^2} \\
&\overset{(A.8)}{=} \frac{n-m}{m-1} \frac{\sum_{l=1}^{m} n_l \left(\widehat{g}(\varphi)_{\varphi\in I_l} - \bar{y}\right)^2}{\sum_{l=1}^{m}\sum_{i:\varphi_i\in I_l}\left(y_i - \widehat{g}(\varphi)_{\varphi\in I_l}\right)^2}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(A.5)}{=} \frac{n-m}{m-1} \frac{\sum_{i=1}^{n} \left(\widehat{g}(\varphi_i) - \bar{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \widehat{g}(\varphi_i)\right)^2} \\
&\overset{(A.1),(A.3)}{=} \frac{n-m}{m-1} \frac{\mathrm{SR}}{\mathrm{SE}} \\
&\overset{(A.3)}{=} \frac{n-m}{m-1} \frac{\mathrm{SY}-\mathrm{SE}}{\mathrm{SE}} \\
&= \frac{n-m}{m-1} \left(\frac{\mathrm{SY}}{\mathrm{SE}} - 1\right).
\end{aligned}
$$

□

Dies ist die in Gleichung (2.24) auf Seite 20 gewählte Darstellung.

A.3. Jurkevich-Periodogramm

Mit der gleichen Partition wie bisher ist das Periodogramm von Jurkevich (1971) definiert durch

$$\mathrm{Per}_{Jur} = \sum_{l=1}^{m} n_l \left(\bar{y}_l - \bar{y}\right)^2,$$

und damit

$$
\begin{aligned}
&\overset{(A.8)}{=} \sum_{l=1}^{m} n_l \left(\widehat{g}(\varphi)_{\varphi \in I_l} - \bar{y}\right)^2 \\
&\overset{(A.5)}{=} \sum_{l=1}^{n} \left(\widehat{g}(\varphi_i) - \bar{y}\right)^2 \\
&\overset{(A.3)}{=} \mathrm{SY} - \mathrm{SE}
\end{aligned}
$$

□

Dies ist die in Gleichung (2.25) auf Seite 20 gewählte Darstellung.

A.4. Phase-Dispersion-Minimization-Periodogramm

Für das Phase-Dispersion-Minimization-Periodogramm lässt Stellingwerf (1978) überlappende Klassen zu. Wir betrachten für $0 = k_0 < k_1 < k_2 < \ldots < k_{2m}$ einen Spezialfall. Sei dazu die Menge $\mathcal{I}^\star = \{I_1^\star, \ldots, I_{2m}^\star\}$ gegeben mit

$$
I_l^\star = \begin{cases}
[k_{2(l-1)}, k_{2l}[& l = 1, \ldots, m \\
[k_{2(l-m)-1}, k_{2(l-m)+1}[& l = m+1, \ldots, 2(m-1) \\
[k_{2m-1}, k_{2m}[\cup [k_0, k_1[& l = 2m
\end{cases} . \tag{A.9}
$$

Dabei bilden sowohl die Klassen

$$\begin{aligned} I_{1,1} = I_1^\star &= [k_0, k_2[, \\ I_{1,2} = I_2^\star &= [k_2, k_4[, \\ &\dots, \\ I_{1,m} = I_m^\star &= [k_{2(m-1)}, k_{2m}[\end{aligned}$$

als auch die Klassen

$$\begin{aligned} I_{2,1} = I_{m+1}^\star &= [k_1, k_3[, \\ I_{2,2} = I_{m+2}^\star &= [k_3, k_5[, \\ &\dots, \\ I_{2,m-1}^\star = I_{2m-1}^\star &= [k_{2m-3}, k_{2m-1}[, \\ I_{2,m}^\star = I_{2m}^\star &= [k_{2m-1}, k_{2m}[\cup [k_0, k_1[\end{aligned}$$

jeweils eine Partition des Intervalls [0,1[. Dieser Spezialfall der überlappenden Klassen findet sich zum Beispiel bei Reimann (1994), wo $m = 5$ und $k_l = 0{,}1 \cdot l$, $l = 1, \dots, 10$, gewählt wird. Sei zur Partition $\mathcal{I}_1 = \{I_{1,1}, \dots, I_{1,m}\}$ mit dazugehöriger Stufenfunktion g_1 die Kleinste-Quadrate-Schätzung von g_1 durch $\widehat{g}_1$ und die dazugehörigen Werte durch $\mathrm{SE}_1, \mathrm{SR}_1$ und R_1^2 bezeichnet. Analog seien g_2, $\widehat{g}_2$, SE_2, SR_2 und R_2^2 für die Partition $\mathcal{I}_2 = \{I_{2,1}, \dots, I_{2,m}\}$ definiert. Die Mächtigkeit der Menge $I_{k,l}$ sei für $k = 1, 2$ und $l = 1, \dots, m$ mit $n_{k,l}$ und das arithmetische Mittel der in die Menge fallenden Beobachtungen mit $\bar{y}_{k,l}$ bezeichnet. Die Bezeichnungen $n_1, \dots, n_{2m}$, $\widehat{\sigma}_1^2, \dots, \widehat{\sigma}_{2m}^2$ und $\bar{y}_1, \dots \bar{y}_{2m}$ beziehen sich auf $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$. Die Definition des Periodogramms nach Stellingwerf (1978) lautet

$$\mathrm{Per}_{PDM} = \frac{\frac{\sum_{l=1}^{2m}(n_l-1)\widehat{\sigma}_l^2}{\sum_{l=1}^{2m} n_l - 2m}}{\widehat{\sigma}^2}.$$

Unter separater Anpassung der Stufenfunktionen g_1 und g_2 ergibt sich daraus

$$\begin{aligned} &\overset{(A.6),(A.7)}{=} \frac{n-1}{\mathrm{SY}} \frac{\sum_{l=1}^{2m}(n_l-1)\frac{1}{n_l-1}\sum_{i:\varphi_i \in I_l^\star}(y_i - \bar{y}_l)^2}{\sum_{l=1}^{m} n_{1,l} + \sum_{l=1}^{m} n_{2,l} - 2m} \\ &= \frac{n-1}{\mathrm{SY}} \frac{\sum_{k=1}^{2}\sum_{l=1}^{m}\sum_{i:\varphi_i \in I_{k,l}}(y_i - \bar{y}_{k,l})^2}{2n-2m} \\ &\overset{(A.8)}{=} \frac{n-1}{2\,\mathrm{SY}(n-m)} \sum_{k=1}^{2}\sum_{l=1}^{m}\sum_{i:\varphi_i \in I_{k,l}} \left(y_i - \widehat{g}_k(\varphi)_{\varphi \in I_{k,l}}\right)^2 \\ &= \frac{n-1}{2\,\mathrm{SY}(n-m)} \sum_{k=1}^{2}\sum_{i=1}^{n}(y_i - \widehat{g}_k(\varphi_i))^2 \\ &\overset{(A.1)}{=} \frac{n-1}{n-m}\frac{1}{2}\frac{\sum_{k=1}^{2}\mathrm{SE}_k}{\mathrm{SY}} \end{aligned}$$

$$\underset{(A.2)}{=} \frac{n-1}{n-m}\left(1-\frac{R_1^2+R_2^2}{2}\right)$$

□

Dies ist die in Gleichung (2.26) auf Seite 20 gewählte Darstellung.

A.5. Lafler-Kinman-Periodogramm

Das Lafler-Kinman-Periodogramm ist nach Lafler und Kinman (1965) definiert als

$$\mathrm{Per}_{LK} = \frac{\sum_{i=1}^{n}\left(y_i^\star - y_{i+1}^\star\right)^2}{\sum_{i=1}^{n}\left(y_i^\star - \bar{y}\right)^2},$$

wobei $y_{n+1}^\star = y_1^\star$ und $y_i^\star$ für $i = 1,\dots n$ der Messwert y_j ist, der zur geordneten Phase $\varphi_{(i)}$ gehört, also mit $i = \mathcal{R}(\varphi_j)$, wobei $\mathcal{R}$ die Rangfunktion. Unter Verwendung ungewichteter Kleinste-Quadrate-Regression und mit $y_0^\star = y_n^\star$ kann dies umformuliert werden zu

$$\begin{aligned}
\mathrm{Per}_{LK} &= \frac{\sum_{i=1}^{n}\left(y_i^\star - y_{i+1}^\star\right)^2}{\mathrm{SY}} \\
&= \frac{1}{2\,\mathrm{SY}}\sum_{i=1}^{n}\left(y_i^\star - y_{i+1}^\star\right)^2 + \left(y_{i+1}^\star - y_i^\star\right)^2 \\
&= \frac{2}{\mathrm{SY}}\sum_{i=1}^{n}\left(\frac{y_i^\star - y_{i+1}^\star}{2}\right)^2 + \left(\frac{y_{i+1}^\star - y_i^\star}{2}\right)^2 \\
&= \frac{2}{\mathrm{SY}}\sum_{i=1}^{n}\left(\frac{y_i^\star - y_{i+1}^\star}{2}\right)^2 + \left(\frac{y_i^\star - y_{i-1}^\star}{2}\right)^2 \\
&= \frac{2}{\mathrm{SY}}\sum_{i=1}^{n}\left(y_i^\star - \frac{y_i^\star + y_{i+1}^\star}{2}\right)^2 + \left(y_i^\star - \frac{y_i^\star + y_{i-1}^\star}{2}\right)^2. \qquad \text{(A.10)}
\end{aligned}$$

Fall 1: n gerade

Für n gerade können die Stufenfunktionen g_1 und g_2 der Partitionen $\mathcal{I}_1$ und $\mathcal{I}_2$ angepasst werden, wobei

$$\begin{aligned}
\mathcal{I}_1 &= \left\{[0,\varphi_{(3)}[, [\varphi_{(3)},\varphi_{(5)}[, \dots, [\varphi_{(n-1)}, 1[\right\} \\
\text{und } \mathcal{I}_2 &= \left\{[\varphi_{(2)},\varphi_{(4)}[, [\varphi_{(4)},\varphi_{(6)}[, \dots, [\varphi_{(n-2)},\varphi_{(n)}[, [\varphi_{(n)}, 1[\cup[0,\varphi_{(1)}[\right\}.
\end{aligned}$$

Die Intervallgrenzen dieser Partitionen sind damit messzeitabhängig, die dazugehörigen Stufenfunktionen haben je $\frac{n}{2}$ Stufen und in jede Menge einer Partition fallen genau zwei beobachtete Phasen. Bei separater Kleinste-Quadrate-Anpassung von g_1 und g_2 mit den bisherigen Notationen ist dann

$$\widehat{g}_1(\varphi_{(1)}) = \widehat{g}_1(\varphi_{(2)}) = \frac{y_1^\star + y_2^\star}{2},$$

$$\begin{aligned}
\widehat{g}_1(\varphi_{(3)}) = \widehat{g}_1(\varphi_{(4)}) &= \frac{y_3^\star + y_4^\star}{2},\\
&\dots,\\
\widehat{g}_1(\varphi_{(n-1)}) = \widehat{g}_1(\varphi_{(n)}) &= \frac{y_{n-1}^\star + y_n^\star}{2},\\
\text{und } \widehat{g}_2(\varphi_{(2)}) = \widehat{g}_2(\varphi_{(3)}) &= \frac{y_2^\star + y_3^\star}{2},\\
\widehat{g}_2(\varphi_{(4)}) = \widehat{g}_2(\varphi_{(5)}) &= \frac{y_4^\star + y_5^\star}{2},\\
&\dots,\\
\widehat{g}_2(\varphi_{(n)}) = \widehat{g}_2(\varphi_{(1)}) &= \frac{y_n^\star + y_1^\star}{2}.
\end{aligned}$$

Für $i = 1, \dots, n$ gilt damit

$$\begin{aligned}
\frac{y_i^\star + y_{i+1}^\star}{2} &= \begin{cases} \widehat{g}_1(\varphi_{(i)}) & i \text{ ungerade} \\ \widehat{g}_2(\varphi_{(i)}) & i \text{ gerade} \end{cases},\\
\frac{y_i^\star + y_{i-1}^\star}{2} &= \begin{cases} \widehat{g}_2(\varphi_{(i)}) & i \text{ ungerade} \\ \widehat{g}_1(\varphi_{(i)}) & i \text{ gerade} \end{cases},
\end{aligned}$$

und damit sowohl für ungerade als auch für gerade i

$$\left(y_i^\star - \frac{y_i^\star + y_{i+1}^\star}{2}\right)^2 + \left(y_i^\star - \frac{y_i^\star + y_{i-1}^\star}{2}\right)^2 = \left(y_i^\star - \widehat{g}_1(\varphi_{(i)})\right)^2 + \left(y_i^\star - \widehat{g}_2(\varphi_{(i)})\right)^2$$

Mit (A.10) gilt dann

$$\begin{aligned}
\mathrm{Per}_{LK} &= \frac{2}{\mathrm{SY}} \sum_{i=1}^{n} \left(y_i^\star - \widehat{g}_1(\varphi_{(i)})\right)^2 + \left(y_i^\star - \widehat{g}_2(\varphi_{(i)})\right)^2\\
&\underset{\text{umordnen}}{=} \frac{2}{\mathrm{SY}} \sum_{i=1}^{n} \left(y_i - \widehat{g}_1(\varphi_i)\right)^2 + \left(y_i - \widehat{g}_2(\varphi_i)\right)^2\\
&\underset{(A.1)}{=} \frac{2}{\mathrm{SY}} \left(\mathrm{SE}_1 + \mathrm{SE}_2\right)
\end{aligned}$$

□

Fall 2: n ungerade

Sei im Fall n ungerade die Partition $\mathcal{I}_k$, $k = 1, \dots, n$, des Intervalls $[0,1[$ so definiert, dass $\varphi_{(k)}$ in ein eigenes Intervall fällt und sonst je zwei benachbarte Phasen in ein Intervall fallen. Die Partition enthält also $\frac{n+1}{2}$ Mengen. Präziser sei

$$\begin{aligned}
\text{für } k \text{ ungerade: } \mathcal{I}_k := \big\{ &\{[\varphi_{(r)}, \varphi_{(r+2)}[: r \text{ ungerade } < k\}\\
&\cup \{[\varphi_{(k)}, \varphi_{(k+1)}[\}\\
&\cup \{[\varphi_{(r)}, \varphi_{(r+2)}[: r \in \{k+1, \dots, n\}, r \text{ gerade}\}\big\}
\end{aligned}$$

$$\begin{aligned}\text{für } k \text{ gerade: } \mathcal{I}_k \coloneqq \big\{ &\{[\varphi_{(r)}, \varphi_{(r+2)}[: r \text{ gerade } < k\} \\ &\cup \{[\varphi_{(k)}, \varphi_{(k+1)}[\} \\ &\cup \{[\varphi_{(r)}, \varphi_{(r+2)}[: r \in \{k+1, \ldots, n\}, r \text{ ungerade}\}\big\},\end{aligned}$$

wobei im Falle von $i_1 < i_2$ und $i_2 > n$ mit $[\varphi_{(i_1)}, \varphi_{(i_2)}[$ die Menge $[\varphi_{(i_1)}, 1[\cup [0, \varphi_{(i_2-n)}[$ gemeint ist. Abbildung A.1 verdeutlicht die Partitionen für den Fall $n = 5$. Sei dann g_k, $k = 1, \ldots, n$, die zur Partition $\mathcal{I}_k$ gehörige Stufenfunktion und $\widehat{g}_k$ der zu g_k gehörige Kleinste-Quadrate-Schätzer. Gemäß Definition gilt

$$\widehat{g}_k(\varphi_{(i)}) = \begin{cases} \frac{y_i^\star + y_{i+1}^\star}{2}, & k \text{ gerade}, i \text{ gerade}, i < k \\ & \vee k \text{ gerade}, i \text{ ungerade}, i > k \\ & \vee k \text{ ungerade}, i \text{ ungerade}, i < k \\ & \vee k \text{ ungerade}, i \text{ gerade}, i > k, \\ \frac{y_i^\star + y_{i-1}^\star}{2}, & k \text{ gerade}, i \text{ gerade}, i > k \\ & \vee k \text{ gerade}, i \text{ ungerade}, i < k \\ & \vee k \text{ ungerade}, i \text{ ungerade}, i > k \\ & \vee k \text{ ungerade}, i \text{ gerade}, i < k, \\ y_i^\star, & i = k \end{cases} . \tag{A.11}$$

Es lässt sich auszählen (vergleiche Tabelle A.1), dass der Ausdruck $\widehat{g}_k(\varphi_i)$ für festes i und variables k in jeweils $\frac{n-1}{2}$ Fällen durch $\frac{y_i^\star + y_{i+1}^\star}{2}$ beziehungsweise $\frac{y_i^\star + y_{i-1}^\star}{2}$ und in einem Fall durch $y_i^\star$ ersetzt werden kann. Damit sind die folgenden Umformungen ausgehend von Gleichung (A.10) zulässig:

$$\begin{aligned}\mathrm{Per}_{LK} &= \frac{2}{\mathrm{SY}} \sum_{i=1}^{n} \left(y_i - \frac{y_i^\star + y_{i+1}^\star}{2} \right)^2 + \left(y_i^\star - \frac{y_i^\star + y_{i-1}^\star}{2} \right)^2 \\ &= \frac{2}{\mathrm{SY}} \frac{2}{n-1} \sum_{i=1}^{n} \frac{n-1}{2} \left(y_i^\star - \frac{y_i^\star + y_{i+1}^\star}{2} \right)^2 + \frac{n-1}{2} \left(y_i^\star - \frac{y_i^\star + y_{i-1}^\star}{2} \right)^2 + \underbrace{(y_i^\star - y_i^\star)^2}_{=0} \\ &\overset{\text{Tabelle A.1}}{=} \frac{4}{\mathrm{SY}(n-1)} \sum_{i=1}^{n} \sum_{k=1}^{n} \left(y_i^\star - \widehat{g}_k(\varphi_{(i)}) \right)^2 \\ &\overset{\text{umordnen}}{=} \frac{4}{\mathrm{SY}(n-1)} \sum_{i=1}^{n} \sum_{k=1}^{n} \left(y_i - \widehat{g}_k(\varphi_i) \right)^2 \\ &= \frac{4}{\mathrm{SY}(n-1)} \sum_{k=1}^{n} \sum_{i=1}^{n} \left(y_i - \widehat{g}_k(\varphi_i) \right)^2 \\ &\overset{(A.1)}{=} \frac{4}{\mathrm{SY}(n-1)} \sum_{k=1}^{n} \mathrm{SE}_k\end{aligned}$$

□

i	Term	Fälle	Anzahl Fälle bei festem i	Summe
gerade	$\frac{y_i^\star + y_{i+1}^\star}{2}$	k gerade, $i < k$	$\frac{n-i-1}{2}$	$\frac{n-1}{2}$
		k ungerade, $i > k$	$\frac{i}{2}$	
	$\frac{y_i^\star + y_{i-1}^\star}{2}$	k ungerade, $i < k$	$\frac{n-i+1}{2}$	$\frac{n-1}{2}$
		k gerade, $i > k$	$\frac{i-2}{2}$	
ungerade	$\frac{y_i^\star + y_{i+1}^\star}{2}$	k gerade, $i > k$	$\frac{i-1}{2}$	$\frac{n-1}{2}$
		k ungerade, $i < k$	$\frac{n-i}{2}$	
	$\frac{y_i^\star + y_{i-1}^\star}{2}$	k ungerade, $i > k$	$\frac{i-1}{2}$	$\frac{n-1}{2}$
		k gerade, $i < k$	$\frac{n-i}{2}$	
beliebig	$y_i^\star$	$k = i$	1	

Tabelle A.1.: Auszählen der Fälle, in denen $\widehat{g}_k(\varphi)$, $k = 1, \ldots, n$, die in Gleichung (A.11) aufgezählten Formen annimmt, bei festem $i \in \{1, \ldots, n\}$.

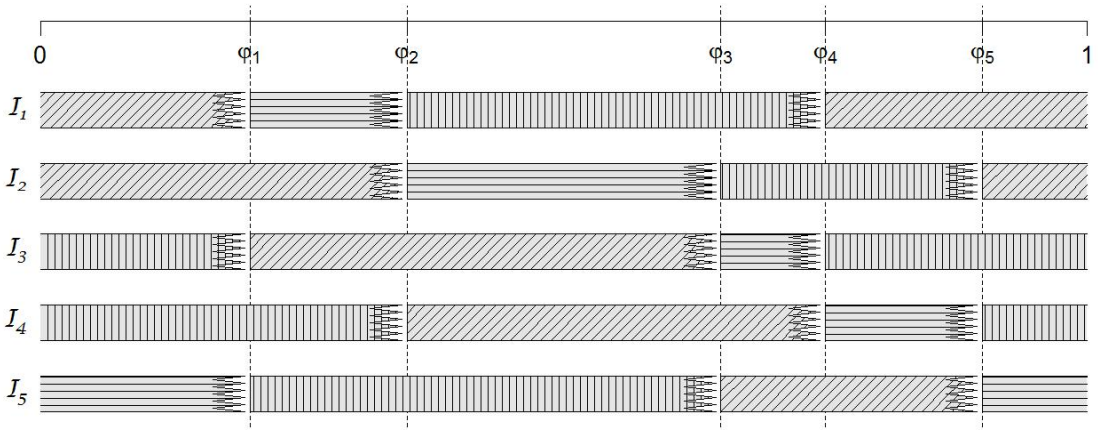


Abbildung A.1.: Partitionen für das Lafler-Kinman-Periodogramm für eine ungerade Anzahl Beobachtungen, hier $n = 5$: Fünf Partitionen definieren fünf dreistufige Stufenfunktionen. Der rechte gezackte Rand soll die nach rechts offenen Intervalle verdeutlichen

B. Herleitung der Basisfunktionen für die Splinefunktion in Abschnitt 2.3.4

Eine Funktion $g(\xi)$, $\xi \in [0,1]$, heißt gemäß Dierckx (1993, S. 3ff.) periodische Splinefunktion vom Grad $k > 0$ (Ordnung $k+1$) mit Knoten $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_{m+1} = 1$, wenn die folgenden Bedingungen erfüllt sind:
1. $g(\xi)$ lässt sich auf jedem Intervall $[\lambda_i, \lambda_{i+1}]$, $i = 0, \ldots, m$, als Polynom vom Grad k ausdrücken.
2. Die Funktion $g(\xi)$ ist auf [0,1] $(k-1)$-mal stetig differenzierbar.
3. Die Funktion $g(\xi)$ und ihre ersten $k-1$ Ableitungen nehmen jeweils auf den Intervallrändern den gleichen Wert an:

$$\left.\frac{\partial g}{\partial \xi^l}\right|_{\xi=0} = \left.\frac{\partial g}{\partial \xi^l}\right|_{\xi=1}, \qquad l = 0, \ldots, k-1.$$

Die Bedingungen 1 und 2 sind dabei Voraussetzung für alle Splinefunktionen auf dem Intervall $[0,1]$, Bedingung 3 stellt die Periodizität sicher.
Mit Hilfe zusätzlich konstruierter Knotenpunkte $\lambda_{-k} < \lambda_{-k+1} < \ldots < \lambda_{-1}$ und $\lambda_{m+1} < \lambda_{m+2} < \ldots < \lambda_{m+k}$ mit $\lambda_{-1} < \lambda_0$ und $\lambda_m < \lambda_{m+1}$, die sich außerhalb des Intervalls $[0,1]$ befinden, können Splinefunktionen als Linearkombinationen

$$g(\xi) = \sum_{i=-k}^{m-1} c_i N_{i,k+1}(\xi), \qquad c_{-k}, \ldots, c_{m-1} \in \mathbb{R}$$

von so genannten B-Spline-Funktionen

$$N_{i,k+1}(\xi) = (\lambda_{i+k+1} - \lambda_i) \sum_{j=0}^{k+1} \frac{\max(\lambda_{i+j} - \xi, 0)^k}{\prod_{\substack{l=0 \\ l \neq j}}^{k+1} (\lambda_{i+j} - \lambda_{i+l})}, \quad i = -k, \ldots, m-1$$

betrachtet werden, vgl. Dierckx (1993, S.8–11). Die B-Spline-Funktionen haben unter anderem die Eigenschaft

$$N_{i,k+1}(\xi) \geq 0 \forall \xi \in [\lambda_i, \lambda_{i+k+1}], \tag{B.1}$$

$$N_{i,k+1}(\xi) = 0 \forall \xi \notin [\lambda_i, \lambda_{i+k+1}]. \tag{B.2}$$

B. Herleitung der Basisfunktionen für die Splinefunktion in Abschnitt 2.3.4

Im Falle einer periodischen Splinefunktion gilt für die Knotenpunkte

$$\lambda_i = \begin{cases} \lambda_{m+i} - 1 & i = -k, \ldots, -1, \\ \text{gewählte Knotenpunkte im Intervall } [0,1] & i = 0, \ldots, m \\ \lambda_{i-m} + 1 & i = m+1, \ldots, m+k \end{cases}$$

und für die Skalare

$$c_i = c_{m+i} \text{ für } i = -k, \ldots, -1.$$

In dieser Arbeit werden periodische kubische Splinefunktionen ($k = 3$) mit $m + 1 = 5$ Knoten auf dem Intervall $[0, 1]$ angepasst. Die Knotenpunkte im Intervall $[0, 1]$ sind $\lambda_i = \frac{i}{4}, i = 0, \ldots, 4$. Mit den zusätzlichen Knotenpunkten außerhalb des Intervalls ergibt sich damit

$$\lambda_i = \left\{ \begin{array}{ll} \lambda_{4+i} - 1 = \frac{4+i}{4} - 1 & i = -3, \ldots, -1, \\ \frac{i}{4} & i = 0, \ldots, 4 \\ \lambda_{i-4} + 1 = \frac{i-4}{4} + 1 & i = 5, \ldots, 7 \end{array} \right\} = \frac{i}{4} \qquad \text{für } i = -3, \ldots, 7$$

und für die Skalare

$$c_i = c_{4+i} \text{ für } i = -3, \ldots, -1. \tag{B.3}$$

Wegen $N_{4,4}(\xi) = 0 \ \forall \xi < \lambda_4 = 1$ gemäß Gleichung (B.2) ergibt sich für eine periodische kubische Splinefunktion $g(\xi)$ mit den oben genannten Knoten auf dem Intervall [0,1[

$$g(\xi) = \sum_{i=-3}^{4} c_i N_{i,4}(\xi) \underset{(B.2)}{=} \sum_{i=-3}^{3} c_i N_{i,4}(\xi) \underset{(B.3)}{=} \sum_{i=0}^{3} c_i M_i(\xi), \qquad \xi \in [0,1], \tag{B.4}$$

mit

$$M_0(\xi) = N_{0,4}(\xi)$$
$$M_i(\xi) = N_{i-4,4}(\xi) + N_{i,4}(\xi) \qquad \text{für } i = 1, 2, 3.$$

Abbildung 2.10 (Seite 23) zeigt die Funktionen M_0, M_1, M_2 und M_3.

C. Nachweis von Formel (2.38)

Gegeben sei das Modell

$$y = X\beta + y_w, \quad y_{w;1}, \ldots, y_{w;n} \underset{u.i.v.}{\sim} \mathcal{N}(0, \sigma^2)$$

wobei $y \in \mathbb{R}^n$ der Vektor der Beobachtungen, $X \in \mathbb{R}^{n \times m}$ die Designmatrix, β ein Parametervektor und e ein Fehlervektor ist. Außerdem steht $\mathcal{N}$ für die Normal-, $\mathcal{F}$ für die F- und $\mathcal{B}$ für die Betaverteilung. Der Vektor β soll mittels Kleinste-Quadrate-Regression durch β_{KQ} geschätzt werden. Das Bestimmtheitsmaß ist dann (vergleiche auch Anhang A):

$$R^2 = \frac{\sum_{i=1}^{n} \left(X\widehat{\beta}_{KQ} - \bar{y}\right)_i^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2},$$

wobei $\bar{y}$ für das arithmetische Mittel steht. Für den Fall $\beta = 0$ gilt bei Anpassung des obigen Modells nach Seber und Lee (2003, S. 110):

$$\frac{R^2}{1 - R^2} \frac{n - m}{m - 1} \sim \mathcal{F}_{m-1,n-m}. \tag{C.1}$$

Nach Gupta und Nadarajah (2004, S. 51) gilt für jedes $U \sim \mathcal{F}_{\theta_1,\theta_2}$

$$\frac{\frac{\theta_1}{\theta_2} U}{1 + \frac{\theta_1}{\theta_2} U} \sim \mathcal{B}\left(\frac{\theta_1}{2}, \frac{\theta_2}{2}\right). \tag{C.2}$$

Mit $\theta_1 = m - 1$ und $\theta_2 = n - m$ und

$$Z = \frac{R^2}{1 - R^2} \frac{\theta_2}{\theta_1}$$

gilt

$$\begin{aligned}
& Z \underset{(C.1)}{\sim} \mathcal{F}_{\theta_1,\theta_2} \\
\Rightarrow R^2 = \frac{R^2}{1 - R^2} \frac{1}{\frac{1-R^2+R^2}{1-R^2}} &= \frac{\frac{R^2}{1-R^2}}{1 + \frac{R^2}{1-R^2}} = \frac{\frac{m-1}{n-m} \frac{R^2}{1-R^2} \frac{n-m}{m-1}}{1 + \frac{m-1}{n-m} \frac{R^2}{1-R^2} \frac{n-m}{m-1}} = \frac{\frac{\theta_1}{\theta_2} Z}{1 + \frac{\theta_1}{\theta_2} Z} \underset{(C.2)}{\sim} \mathcal{B}\left(\frac{\theta_1}{2}, \frac{\theta_2}{2}\right) \\
\Rightarrow R^2 &\sim \mathcal{B}\left(\frac{m-1}{2}, \frac{n-m}{2}\right).
\end{aligned}$$

□

D. Zur angenommenen Betaverteilung des Bestimmtheitsmaßes

Bei Anpassung eines linearen Modells an unabhängig identisch normalverteilte Messwerte ist das Bestimmtheitsmaß im Falle von Kleinste-Quadrate-Regression betaverteilt (vgl. Abschnitt 2.6.1). Mit der hier vorgestellten Simulation soll die Annahme des betaverteilten Bestimmtheitsmaßes für die anderen in dieser Arbeit verwendeten Regressiontechniken gerechtfertigt werden.
Sei $\mathcal{U}_{[0,1]}$ die Rechteckverteilung auf dem Intervall $[0,1]$, dann ist das Vorgehen das Folgende:

1. Generiere $\varphi_1, \ldots, \varphi_{200} \underset{u.i.v.}{\sim} \mathcal{U}_{[0,1]}$ und $y_1, \ldots, y_{200} \underset{u.i.v.}{\sim} \mathcal{N}(0,1)$. Passe eine periodische Funktion aus Abschnitt 2.3 mit einer Regressionstechnik aus Abschnitt 2.4 an und berechne das Bestimmtheitsmaß. Dazu wird die Implementierung aus Kapitel 3 mit den Voreinstellungen aus Kapitel 4 verwendet.

2. Wiederhole 10 000 Mal Schritt 1 und passe mittels Cramér-von-Mises (CvM) -Distanz-Minimierung eine Betaverteilung mit Parametern θ_1 und θ_2 an die Periodogrammbalken $R_1^2, \ldots, R_{10\,000}^2$ an. Führe einen Kolmogorov-Smirnov-Test (vgl. etwa Sachs und Hedderich 2006, S. 337) auf die angepasste Verteilung durch.

Abbildung D.1 zeigt Histogramme der 10000 Werte, die bei Anpassung des Doppelstufenmodells mit Kleinste-Quadrate-, Kleinste-Beträge-, S-, τ-, M-Huber- und M-Tukey- Regression erreicht wurden. Es ist zu erkennen, dass die Betaverteilung gut zur Häufigkeitsverteilung des Bestimmtheitsmaßes passt. In Tabelle D.1 sind die angepassten Parameter und die Ergebnisse der Kolmogorov-Smirnov-Tests angegeben. Der Kolmogorov-Smirnov-Wert lehnt zu einem Niveau von 5% nie die Nullhypothese der Betaverteilung ab. Für die hier genannten Regressionstechniken konnte dies auch für die anderen Modelle beobachtet werden. Auch wenn die Anpassung der Parameter θ_1 und θ_2 an die Werte $R_1^2, \ldots, R_{10000}^2$ statt durch CvM-Distanz-Minimierung in Schritt 2 durch Momentenschätzer

$$\widehat{\theta_1} = -\frac{\overline{\mathrm{Per}} \cdot \left(-\overline{\mathrm{Per}} + \overline{\mathrm{Per}}^2 + \mathrm{var}(\mathrm{Per})\right)}{\mathrm{var}(\mathrm{Per})}, \qquad \widehat{\theta_2} = \frac{\widehat{\theta_1}\left(1 - \overline{\mathrm{Per}}\right)}{\overline{\mathrm{Per}}} \tag{D.1}$$

geschieht, wobei var für die empirische Varianz steht, lehnt der Kolmogorov-Smirnov-Test nie ab.
Bei Verwendung von LTS-Regression passen die Periodogrammbalken nicht ganz so gut zu einer Betaverteilung. Wie den Werten in Tabelle D.2 zu entnehmen ist, lehnt der Kolmorogov-Smirnov-Test hier die Nullhypothese einer Betaverteilung für alle Modelle

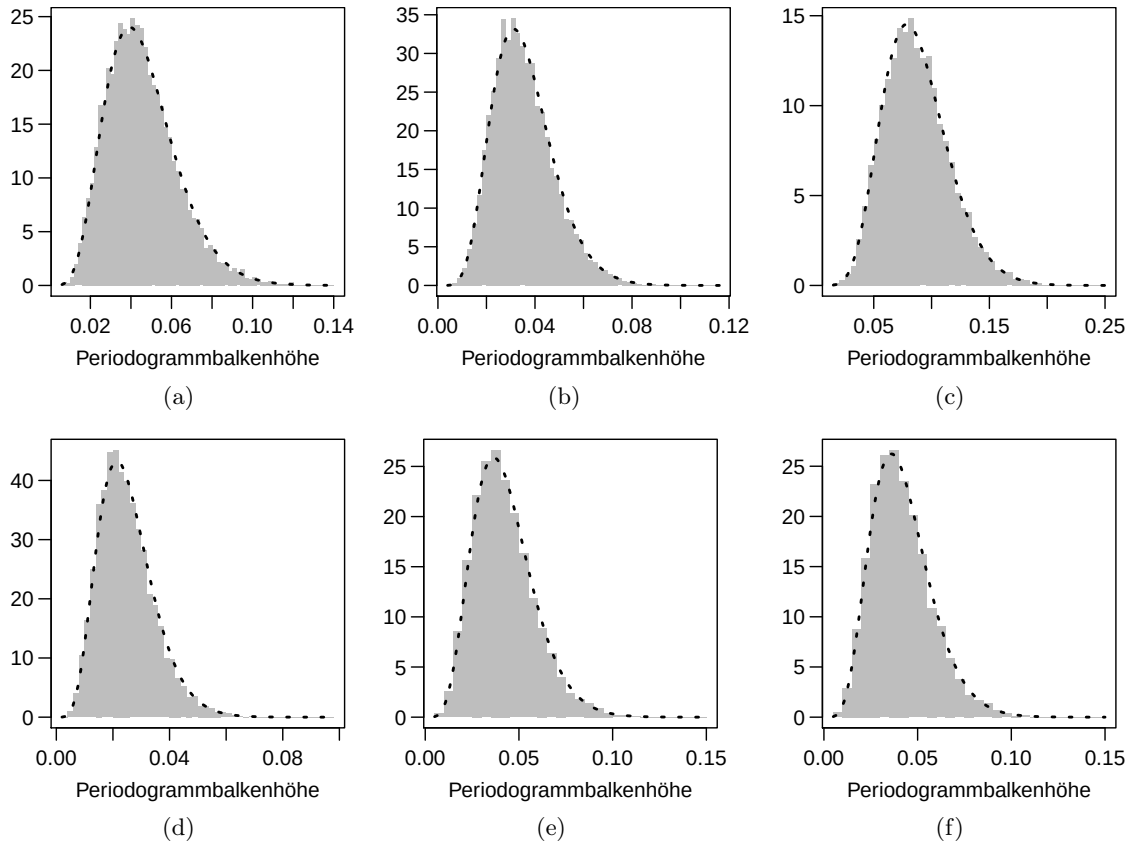


Abbildung D.1.: Histogramme von 10000 Bestimmtheitsmaßen bei Anpassung der Doppelstufenfunktion durch verschiedene Regressionstechniken: (a) KQ, (b) L1, (c) S, (d) τ, (e) M-Huber, (f) M-Tukey. Die Linien zeigen die mittels Cramér-von-Mises-Distanz-Minimierung angepassten Dichten einer Betaverteilung. Die Parameter und die Ergebnisse des Kolmogorov-Smirnov-Tests sind Tabelle D.1 zu entnehmen.

Regression	$\widehat{\theta_1}$	$\widehat{\theta_2}$	Teststatistik	p-Wert
L2	6,469	136,728	0,005	0,980
L1	7,623	209,243	0,006	0,868
S	8,568	89,566	0,006	0,869
τ	6,394	253,062	0,007	0,759
M-Huber	6,495	148,995	0,006	0,908
M-Tukey	6,490	150,917	0,007	0,777

Tabelle D.1.: Angepasste Betaverteilungen für die Szenarien aus Abbildung D.1: Angepasste Parameter $\widehat{\theta_1}$ und $\widehat{\theta_2}$, Teststatistik des Kolmogorov-Smirnov-Test, p-Wert des Tests.

außer den Stufenmodellen zum Niveau 5% ab. Bei Verwendung von Momentenschätzern statt der CvM-angepassten Parameter wird die Betaverteilung auch für das einfache Stufenmodell abgelehnt.
Bei Betrachtung der Histogramme für LTS-Anpassung verschiedener periodischer Funktionen (Abbildung D.2) scheint die Annahme einer Betaverteilung der Periodogrammbalken dennoch nicht völlig falsch zu sein. Hier ist zu beobachten, dass die Dichte den Großteil des Histogrammes ganz gut nachzeichnet, sehr kleine Beobachtungen jedoch überrepräsentiert sind und der Modalwert der angepassten Dichte zu niedrig liegt. Dies könnte mit der Implementierung der LTS-Regression erklärbar sein. Möglicherweise gibt es Fälle, in denen der Algorithmus keine gute Lösung des Regressionsproblems findet, weil die Zielfunktion viele lokale Optima hat und das globale Optimum dicht von lokalen Optima umgeben ist. Der Optimierer `genoud` ist so eingestellt, dass die Lösung für das Lokationsmodell stets auch Kandidat für das neue Modell ist (vergleiche Abschnitt 3.1.5). Dies stellt sicher, dass der Periodogrammbalken auf jeden Fall positiv ist. In Fällen, in denen das globale Optimum zwischen lokalen Optima „versteckt" ist, ist vorstellbar, dass die Optimierung bei einer Lösung endet, die nahe der Lokationsmodelllösung liegt. So ließe sich der unerwartet hohe Anteil niedriger Periodogrammbalken in Abbildung D.2 deuten. Diese Überlegung würde auch erklären, warum das Doppelstufenmodell die Betaverteilungsannahme eher erfüllt: Hier werden je Testperiode zwei Modelle angepasst und das arithmetische Mittel der Bestimmtheitsmaße als Periodogrammbalken verwendet. Eine gegebenenfalls fehlgeschlagene Optimierung kann so zumindest teilweise durch die zweite Anpassung ausgeglichen werden.
Möglicherweise beeinflusst die Überzahl kleiner Periodogrammbalken die Parameterschätzer so, dass die geschätzte Verteilung ein höheres Gewicht auf niedrige Werte legt und sich der Modalwert der Verteilung verringert. Dies führt jedoch nicht dazu, dass zu hohen Wahrscheinlichkeiten gehörende Quantile ebenfalls niedriger ausfallen. Beispielsweise liegt das 0,9-Quantil der angepassten Betaverteilung stets über dem 0,9-Quantil der simulierten Periodogrammbalken (vgl. Tabelle D.2). Dies gilt auch für höhere Quantile, beispielsweise das $\sqrt[200]{0{,}99}$-Quantil, dass man verwenden würde, wenn man in einem Periodogramm mit 200 Balken mit Niveau $\alpha = 0{,}01$ nach auffälligen Perioden suchen möchte. In dieser Hinsicht scheint ein Verfahren zur Detektion auffälliger Perioden unter Annahme betaverteilter Periodogrammbalken eher konservativ zu agieren und sollte das Niveau einhalten.
Mit den Ergebnissen der hier beschriebenen Simulation scheint es bei Vorliegen von ausschließlich unabhängig identisch normalverteilten Beobachtungen sinnvoll, für alle in dieser Arbeit verwendeten Bestimmtheitsmaße eine Betaverteilung anzunehmen.

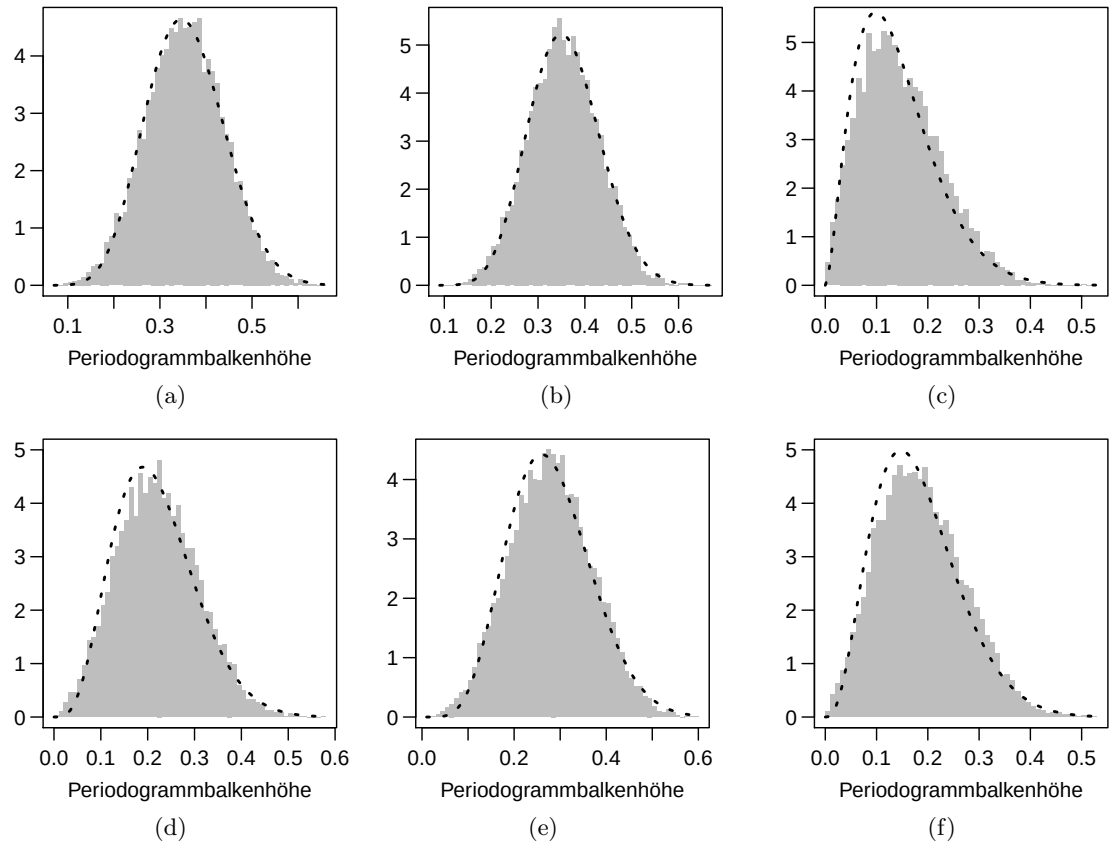


Abbildung D.2.: Histogramme von 10000 Bestimmtheitsmaßen für die LTS-Regression angewandt auf verschiedene Modelle: (a) Einfachstufenfunktion, (b) Doppelstufenfunktion, (c) Sinusfunktion, Fourierreihe (d) zweiten und (e) dritten Grades, periodische Splinefunktion. Die Linien zeigen die mittels Cramér-von-Mises-Distanz-Minimierung angepassten Dichten einer Betaverteilung. Die Parameter und die Ergebnisse des Kolmogorov-Smirnov-Tests sind Tabelle D.2 zu entnehmen.

Modell	$\widehat{\theta}_1$	$\widehat{\theta}_2$	Teststatistik	p-Wert	$b_{0,95}$	$\widehat{b}_{0,95}$	$b\,\sqrt[200]{0,99}$	$\widehat{b}\,\sqrt[200]{0,99}$
Einfachstufe	11,112	20,108	0,012	0,098	0,633	0,462	0,853	0,649
Doppelstufe	14,117	25,540	0,010	0,286	0,618	0,453	0,823	0,645
Sinus	2,699	15,209	0,015	0,017	0,777	0,257	0,988	0,516
Fourier(2)	4,809	16,634	0,019	0,002	0,706	0,335	0,950	0,553
Fourier(3)	6,808	17,272	0,017	0,005	0,672	0,398	0,913	0,590
Splines	3,775	16,061	0,016	0,014	0,735	0,301	0,970	0,518

Tabelle D.2.: Angepasste Betaverteilungen für die Szenarien aus Abbildung D.2: Angepasste Parameter $\widehat{\theta}_1$ und $\widehat{\theta}_2$, Teststatistik des Kolmogorov-Smirnov-Test, p-Wert des Tests, Quantile b der Stichprobe und $\widehat{b}$ der angepassten Verteilung.

E. Gipfelbreite bei vorliegender Periodizität

In Abschnitt 2.6.2 wird beschrieben, dass ein Gipfel um Periode p aus dem Periodogramm entfernt wird. Dazu wird eine Umgebung um p bestimmt und die zu den darin liegenden Perioden gehörenden Balken eliminiert.
An dieser Stelle soll genauer ausgeführt werden, wie diese Umgebung $[p_l, p_r]$ mit $p_l < p < p_r$ bestimmt wird. Das entsprechende Intervall auf Frequenzskala wird mit $[f_l, f_r]$ notiert, wobei $f_l = 1/p_r$ und $f_r = 1/p_l$ gilt.
Gemäß Zechmeister und Kürster (2009) hat der Gipfel um Frequenz $1/p$ auf Frequenzskala eine Breite von ungefähr $1/T$, wobei T die Dauer $t_n - t_1$ der Lichtkurve ist. Wegen der in Abschnitt 2.6.2 (speziell auch Abbildung 2.14) angestellten Überlegungen über Relationen auf Perioden- und Frequenzskala wird davon ausgegangen, dass der Gipfel auf der Frequenzskala symmetrisch ist. Daraus folgt

$$f_l = \frac{1}{p} - \frac{1}{2}\frac{1}{T} = \frac{2T-p}{2Tp} \quad \text{und} \quad f_r = \frac{1}{p} + \frac{1}{2}\frac{1}{T} = \frac{2T+p}{2Tp}.$$

Für p_r folgt damit

$$p_r = \frac{1}{f_l} = \frac{2Tp}{2T-p} = p\left(\frac{2T-p+p}{2T-p}\right) = p\left(1 + \frac{p}{2T-p}\right) = p + \frac{p^2}{2T-p}$$

und analog

$$p_l = p - \frac{p^2}{2T+p}.$$

Das so bestimmte Intervall $\left[p - \frac{p^2}{2T+p}, p + \frac{p^2}{2T-p}\right]$ hat eine Breite von

$$\begin{aligned} p_r - p_l &= \frac{p^2}{2T-p} + \frac{p^2}{2T+p} = p^2\left(\frac{2T+p+2T-p}{4T^2-p^2}\right) \\ &= \frac{p^2}{T}\left(1 + \frac{p^2}{4T^2-p^2}\right). \end{aligned}$$

E. Gipfelbreite bei vorliegender Periodizität

Gilt wie von Halpern, Leighly und Marshall (2003) empfohlen für alle Testperioden $p < T/10$, lässt sich der zweite Faktor nach oben abschätzen:

$$1 + \frac{p^2}{4T^2 - p^2} < 1 + \frac{\frac{T^2}{100}}{4T^2 - \frac{T^2}{100}} = 1 + \frac{1}{100} \cdot \frac{100}{399} \approx 1.0025.$$

Gleichzeitig gilt auch:

$$p^2 > 0$$

$$\text{und } 4T^2 - p^2 > 4T^2 - \frac{T^2}{100} > 0,$$

woraus folgt

$$1 + \frac{p^2}{4T^2 - p^2} > 1.$$

Das Intervall $[p_l, p_r]$ ist also etwas länger als p^2/T.

F. Aufbau der Funktion `RobPer`

In diesem Abschnitt wird der Aufbau der R-Funktion `RobPer` zur Periodogrammberechnung erläutert. Zur Ausführung der Funktion `RobPer` werden die folgenden R-Funktionen benötigt:

`FastS` Leicht modifizierte Version der Funktion `fast.s` aus Salibian-Barrera und Yohai (2006) mit folgenden Veränderungen:

1. Einige Kontrollparameter (`k` und `best.r`) werden umbenannt (zu `kk` und `tt`), um eine einheitlichere Notation für die Funktionen `FastS` und `FastTau` zu erreichen.
2. Einige Kontrollparameter wurden zu einer Liste zusammengefasst.
3. Ein Kandidat für den Regressionsparameter kann festgelegt werden (`RobPer` übergibt $\widehat{\beta}_\mu$, vgl. Gleichung 3.4, Seite 42).
4. Zur Auffindung einer Unterstichprobe in allgemeiner Lage werden Regressoren $x_{i\star}^\mathsf{T}$ aus der Menge in der Designmatrix X auftretenden Zeilen gezogen, ohne Berücksichtigung der Häufigkeit des Auftretens in der Designmatrix. Je Regressor $x_{i\star}^\mathsf{T}$ wird dann von den dazugehörigen Messwerten einer gezogen. Beim Stufenmodell wird für jede Stufe eine Beobachtung gezogen.
5. Wenn in 100 Versuchen keine Unterstichprobe in allgemeiner Lage gefunden werden kann, wird `NA` und eine Warnung ausgegeben. In der Ursprungsimplementierung versucht die Funktion stets weiter, ohne Abbruchkriterium, eine Unterstichprobe zu finden.
6. Wenn eine Intercept-Spalte zur Designmatrix hinzugefügt werden soll (im Falle `Scontrol$int=TRUE`, der aber in `RobPer` nicht auftaucht), wird erst anschließend die Spaltenanzahl der Designmatrix ermittelt.
7. Die Unterfunktionen `loss.S`, `re.s`, `f.w`, `scale1`, `our.solve` und `rho` werden nicht mehr bei jedem Aufruf von `FastS` neu definiert, sondern werden einmal global definiert.
8. Die Unterfunktion `norm` wird durch die R-Funktion `norm(...,"2")` mit gleicher Funktionalität aus dem R-Paket `base` ersetzt.
9. Die Ausgabe der Funktion wurde der Übersichtlichkeit halber mit erläuternden Bezeichnungen versehen.

`FastTau` Funktion zur τ-Regression aus Salibian-Barrera, Willems und Zamar (2008) mit leichten Modifikationen:

1. Modifikation analog zu Punkt 2 in `FastS`.

2. Modifikation analog zu Punkt 3 in FastS.
3. Modifikation analog zu Punkt 4 in FastS.
4. Wenn in 100 Versuchen keine Unterstichprobe in allgemeiner Lage gefunden werden kann, wird NA und eine Warnung ausgegeben. Da die Funktion nicht mehr mit der stop-Funktion beendet wird, können Funktionen, die FastTau verwenden, weiterlaufen und brechen nicht ab.
5. Ein mehrfach verwendeter Codeblock, der prüft, ob unter den Kandidaten für den Regressionsparameter einer mit bis dahin bestem Zielkriterium ist, wurde in die Unterfunktion checkbest ausgelagert.
6. Im verwendeten IRWLS-Algorithmus kommt es durch Rundungsungenauigkeiten teils zu betraglich sehr kleinen negativen Werten („negativen Nullen") als Gewicht. Durch Rundung der Gewichte auf acht Nachkommastellen kann dies verhindert werden.
7. Die Unterfunktion randomset erfüllt die gleiche Aufgabe wie die im Paket base implementierte Funktion sample und wurde durch sie ersetzt.

`genoud` (R-Paket rgenoud, Mebane Jr. und Sekhon 2011) optimiert $\hat{\beta}$ durch evolutionäre Optimierung (vgl. Abschnitt 3.1.5). Diese Funktion kommt bei Tukey-Regression und, falls gewünscht, bei LTS-Regression (LTSopt=TRUE) zum Einsatz. Ein Großteil der Voreinstellungen wird belassen. Lediglich die Einstellungen $\texttt{fn}(\beta) = \zeta\left(\tilde{y} - \tilde{X}\beta\right)$ und $\texttt{starting.values}=\{\widehat{\beta}, \widehat{\beta}_\mu\}$ (im Falle der LTS-Regression auch BFGS=FALSE) weichen von den Voreinstellungen ab, und solution.tolerance=tol, pop.size, wait.generations und max.generations werden unterschiedlich eingestellt.

`IRWLS` führt den Reweighted-Least-Squares-Schritt in der M-Regression durch, vgl. Abbildung F.4.

`lm` (R-Paket stats) führt einfache Kleinste-Quadrate-Regression durch.

`lmrob.S` (R-Paket robustbase, Rousseeuw et al. 2012) führt S-Regression durch.

`ltsReg` (R-Paket robustbase) führt LTS-Regression durch. Der Parameter alpha steuert das Trimming h. Dabei ist $h(\texttt{alpha}, n, m) = 2\left\lfloor \frac{n+m+1}{2} \right\rfloor - n - 2\left(n - \left\lfloor \frac{n+m+1}{2} \right\rfloor \cdot \texttt{alpha}\right)$ (vgl. Funktion h.alpha.n, R-Paket robustbase). Um $h(\texttt{alpha}, n, m) = \left\lfloor \frac{n}{2} \right\rfloor + \left\lfloor \frac{m+1}{2} \right\rfloor$ für einen asymptotischen Bruchpunkt von 0,5 zu erreichen (vgl. Abschnitt 2.4.1), muss $\texttt{alpha} \in I_{\texttt{alpha}}$ gewählt werden mit

$$I_{\texttt{alpha}} = \begin{cases} \left[\frac{1}{2} - \frac{1}{n-m-1}, \frac{1}{2}\right[, & n \text{ ungerade, } m \text{ gerade} \\ \left[\frac{1}{2}, \frac{1}{2} + \frac{1}{n-m-1}\right[, & n \text{ gerade, } m \text{ ungerade} \\ \left[\frac{1}{2} - \frac{1}{2(n-m)}, \frac{1}{2} + \frac{1}{2(n-m)}\right[, & n \text{ ungerade, } m \text{ ungerade} \\ \left[\frac{1}{2}, \frac{1}{2} + \frac{1}{n-m}\right[, & n \text{ gerade, } m \text{ gerade} \end{cases}.$$

Damit wird sichtbar, dass ltsReg für die Voreinstellung alpha=0,5 für n ungerade und m gerade nicht $h(m) = \left\lfloor \frac{n}{2} \right\rfloor + \left\lfloor \frac{m+1}{2} \right\rfloor$ verwendet, sondern $h(m) = \left\lfloor \frac{n+m+1}{2} \right\rfloor$. Beim

Lokationsmodell muss `alpha` so gewählt werden, dass $h(\texttt{alpha}, n, 1) = \left\lfloor \frac{n}{2} \right\rfloor + \left\lfloor \frac{m+1}{2} \right\rfloor$, dies ist äquivalent zu $\texttt{alpha} \in I_{\texttt{alpha},\mu}$ mit

$$I_{\texttt{alpha},\mu} = \begin{cases} \left[\frac{1}{2} + \frac{m-2}{2(n-1)}, \quad \frac{1}{2} + \frac{m}{2(n-1)}\right[, & n \text{ ungerade, } m \text{ gerade} \\ \left[\frac{1}{2} + \frac{m-1}{2(n-2)}, \quad \frac{1}{2} + \frac{m+1}{2(n-2)}\right[, & n \text{ gerade, } m \text{ ungerade} \\ \left[\frac{1}{2} + \frac{m-1}{2(n-1)}, \quad \frac{1}{2} + \frac{m+1}{2(n-1)}\right[, & n \text{ ungerade, } m \text{ ungerade} \\ \left[\frac{1}{2} + \frac{m-2}{2(n-2)}, \quad \frac{1}{2} + \frac{m}{2(n-2)}\right[, & n \text{ gerade, } m \text{ gerade} \end{cases}.$$

Für $m > 1$, $n > 1$, $n + 1 \geq m$ und $m + n > 3$ gilt $\frac{1}{2} + \frac{m-1}{2(n-2)} = \frac{n+m-3}{2(n-2)} \in I_{\texttt{alpha},\mu}$.

`rq` (R-Paket `quantreg`, Koenker 2012) führt Quantilsregression, speziell auch Median-Regression durch, die der Kleinste-Beträge-Regression entspricht.

`spline.des` (R-Paket `splines`) berechnet B-Splines (vgl. Anhang B) zu gegebenen Messzeiten.

`singlePerM` und `singlePernotM` sind Unterfunktionen von `RobPer` um einen einzelnen Periodogrammbalken zu berechnen. Sie sind im hier beschriebenen Struktogrammblock `singleFUN` (Abbildung F.3) zusammengefasst.

`Xgen` stellt je nach Modell `model` und Testperiode p_i die passende Designmatrix X auf und löscht Nullspalten (typischerweise bei Stufenfunktionen).

Der Aufbau von `RobPer` wird in diesem Kapitel als Nassi-Shneiderman-Diagramm (Struktogramm nach Norm DIN 66261) dargestellt. Abbildung F.1 enthält zunächst eine Leseanleitung für die verwendeten Strukturen. Die Eingabevariablen für `RobPer` sind Tabelle F.1 zu entnehmen. Das Struktogramm von `RobPer` wird in Abbildung F.2 dargestellt. Wie bisher (vgl. Abschnitt 2.4) sind die für die Kleinste-Quadrate(KQ)-, Least-Trimmed-Squares(LTS)-, Kleinste-Beträge(L1)-, M-Huber(MH)- und M-Tukey(MT)-Regression die folgenden mathematischen Funktionen definiert:

$$\zeta_{KQ}(r) = \sum_{i=1}^{n} r_i^2$$

$$\zeta_{LTS}(r) = \sum_{i=1}^{h(m)} r_{(i)}, \qquad h(m) = \left\lfloor \frac{n}{2} \right\rfloor + \left\lfloor \frac{m+1}{2} \right\rfloor,$$

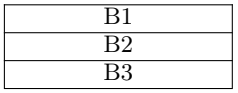
$$\zeta_{L1}(r) = \sum_{i=1}^{n} |r_i|,$$

$$\rho_{MH}(\nu) = \begin{cases} \nu^2 & |\nu| \leq k \\ 2k|\nu| - k^2 & |\nu| > k \end{cases}, \qquad \rho_{MT}(\nu) = \begin{cases} 1 - \left(1 - \left(\frac{\nu}{k}\right)^2\right)^3 & |\nu| \leq k \\ 1 & |r| > k \end{cases},$$

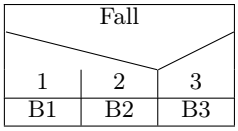
$$\zeta_{MH}(r) = \sum_{i=1}^{n} \rho_{MH}\left(\frac{r_i}{\widehat{\sigma}}\right), \qquad \zeta_{MT}(r) = \sum_{i=1}^{n} \rho_{MT}\left(\frac{r_i}{\widehat{\sigma}}\right),$$

$$W_{MH}(\nu) = \begin{cases} c_{MH} & |\nu| \leq k \\ c_{MH} \cdot \frac{k}{|\nu|} & |\nu| > k \end{cases}, \qquad W_{MT}(\nu) = \begin{cases} c_{MT} \cdot \left(1 - \left(\frac{\nu}{k}\right)^2\right)^2 & |\nu| \leq k \\ 0 & |\nu| > k \end{cases}.$$

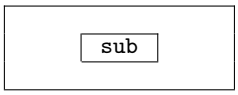
Gemäß Definition (3.3) (vgl. Seite 41) ist $c_{MH} = 2$ und $c_{MT} = \frac{6}{k^2}$, für die Implementierung kann wegen der Skaleninvarianz der Kleinste-Quadrate-Schätzung im Iteratively-Reweighted-Least-Squares(IRWLS)-Schritt $c_{MH} = c_{MT} = 1$ gesetzt werden.



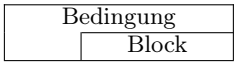
Führe erst B1 aus, dann B2, dann B3. Horizontale Abgrenzungen aufeinander folgender Blöcke werden aus Übersichtlichkeitsgründen teilweise weggelassen.



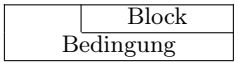
Wenn Fall 1, führe B1 aus, wenn Fall 2, führe B2 aus, wenn Fall 3, führe B3 aus.



Führe `sub` aus (ausgelagerter Code oder Funktion).

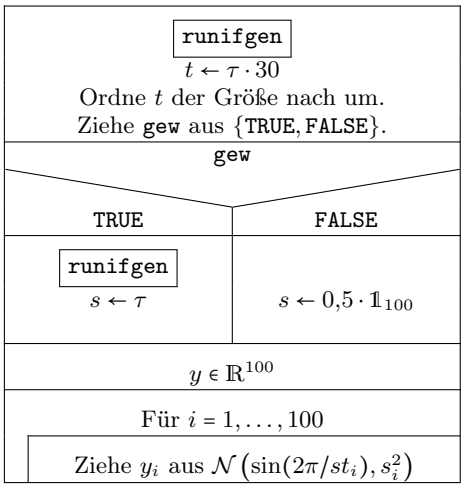


Wiederholung des Blocks mit vorgeschalteter Bedingungsprüfung (z.B. `for`-Schleife)



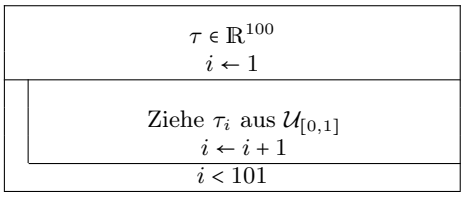
Wiederholung des Blocks mit nachgeschalteter Bedingungsprüfung (z.B. durch `if(!...)...break`)

(a)




```
eval(parse(text=runifgen))
t <- tau*30
t <- sort(t)
gew<- sample(c(TRUE, FALSE),1)
if(gew){
eval(parse(text=runifgen))
 s <- tau}
if(!gew){
 s<- rep(0.5, 100 )
 }
y <- numeric(100)
for(i in 1:100){
y[i]<- rnorm(1, mean=sin(2*pi/5*t[i]), sd=s)}
```

Block `runifgen`:




```
runifgen <- paste("
 tau <- numeric(100)
 i <- 1
 repeat{
 tau[i] <- runif(1)
 i <- i+1
 if(!i<101) break}
")
```

(b)

Abbildung F.1.: Leseanleitung zu den Struktogrammen: In (a) sind die zur Algorithmusdarstellung verwendeten Blöcke zu sehen. In (b) ist das Struktogramm (links) zu einem einfachen R-Code (rechts) zu sehen, der die Beobachtungen $(t_i, y_i, s_i)_{i=1,\ldots,100}$ einer einfachen Lichtkurve mit Fluktuationsperiode 5 generiert. Dieser R-Code dient nur Demonstrationszwecken und ist nicht effizient programmiert.

Eingabe	Erläuterung
`ts` $\in \mathbb{R}^{n\times 3}$ oder $\mathbb{R}^{n\times 2}$	Lichtkurvenmatrix mit Zeilen (t_i, y_i, s_i), $i = 1, \ldots, n$; Bei `weighting=FALSE` genügt auch (t_i, y_i), $i = 1, \ldots, n$.
`weighting` $\in$ {`T`,`F`}	Ob Messfehler s_i durch gewichtete Regression einbezogen werden sollen.
`periods` $\in \mathbb{R}^q_{>0}$	Testperioden $p_1, \ldots, p_q$
`regression`	Regressionstechnik (vgl. Abschnitt 2.4), Einstellungen: `"L2"` (KQ), `"L1"`, `"LTS"`, `"S"`, `"huber"` (M-Huber), `"bisquare"` (M-Tukey), `"tau"` (τ)
`model`	Periodische Funktion g (vgl. Abschnitt 2.3), Einstellungen: `"step"` (Stufenmodell), `"2step"` (Doppelstufenfunktion), `"sine"` (Sinusfunktion), `"fourier(2)"` und `"fourier(3)"` (Fouriersummen zweiten und dritten Grades), `"splines"` (Splinefunktion)
`steps` $\in \mathbb{N}$	Anzahl Stufen des Modells Nur nötig wenn `model`$\in$`{"step", "2step"}`. Voreinstellung: 10
`var1` $\in$ {`T`,`F`}	Varianz auf 1 halten (vgl. Abschnitte 2.5 und 3.1.1)? Nur nötig wenn `regression`$\in$`{"huber", "bisquare"}`. Voreinstellung: `weighting`
`tol`$\in \mathbb{R}_{>0}$	Konvergenzschranke Nur nötig wenn `regression`$\in$`{"huber", "bisquare", "S"}` oder wenn `LTSopt=TRUE&regression="LTS"`. Voreinstellung: 10^{-3}
`genoudcontrol`$\in \mathbb{N}^3$	Einstellungen für `genoud`: `max.generations`, `wait.generations`, `pop.size` Je größer die Zahlen, umso langsamer und besser wird nachoptimiert, weitere Details bei Mebane Jr. und Sekhon (2011). Nur nötig bei `regression="bisquare"` oder wenn `LTSopt=TRUE&regression="LTS"`. Voreinstellung: {50,5,50}
`LTSopt` $\in$ {`T`,`F`}	LTS-Schätzung nachoptimieren (vgl. Abschnitte 3.1.4 und 3.1.5)? Nur nötig bei `regression="LTS"`. Voreinstellung: `TRUE` wenn `regression="LTS"`
`taucontrol`$\in \mathbb{N}^4 \times$ {`T`, `F`}	Einstellungen für τ-Regression: `N`, `kk`, `tt`, `rr`, `approximate` Je größer die natürlichen Zahlen, umso langsamer und präziser, weitere Details bei Salibian-Barrera, Willems und Zamar (2008). Nur nötig bei `regression="tau"`, `rr` nur nötig bei `approximate`=`TRUE`. Voreinstellung: {100, 2, 5, 2, `FALSE`}
`Scontrol` $\in \mathbb{N}^3 \times \mathbb{R}^2_{>0} \times \mathbb{N}$	Einstellung für S-Regression: `N`, `kk`, `tt`, `b`, `cc`, `seed` Um reproduzierbare Ergebnisse zu erhalten, wird `seed` festgesetzt und bleibt andernfalls undefiniert. Weitere Details bei Salibian-Barrera und Yohai (2006). Nur nötig bei `regression="S"`. Voreinstellung: {`N`,2,5,0,5,1.547,`NULL`} mit `N`=50 bei `weighting=F`, `N`=200 bei `weighting=T`.
Ausgabe	
`periodogram`$\in \mathbb{R}^q$	Vektor mit zu den Testperioden gehörende Periodogrammbalken
evtl. Warnungen	

Tabelle F.1.: Ein- und Ausgabeparameter der Funktion `RobPer`. {`T`,`F`} steht für {`TRUE`, `FALSE`}.

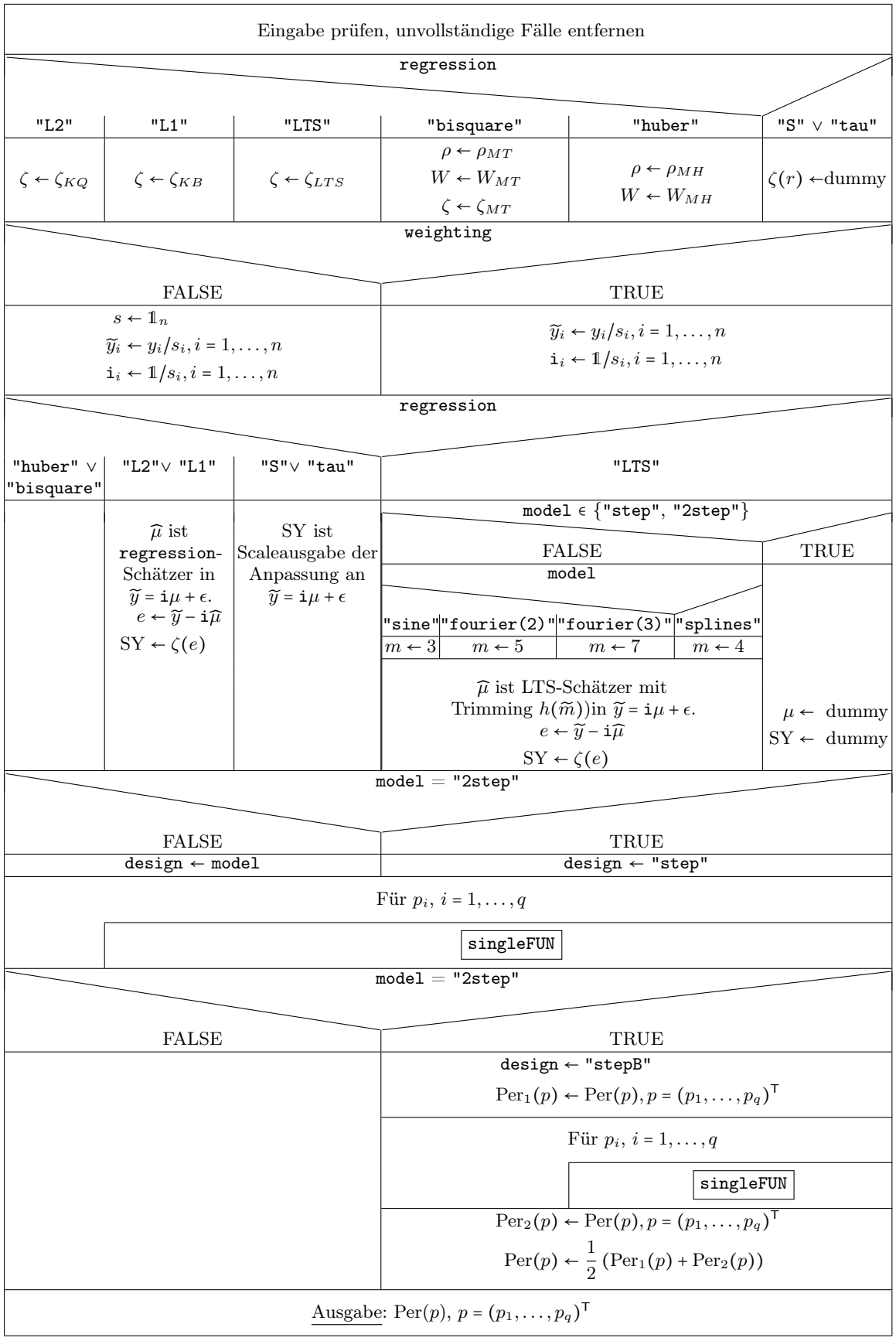


Abbildung F.2.: Struktogramm von `RobPer`. Der Block `singleFUN` wird in Abbildung F.3 detailliert beschrieben.

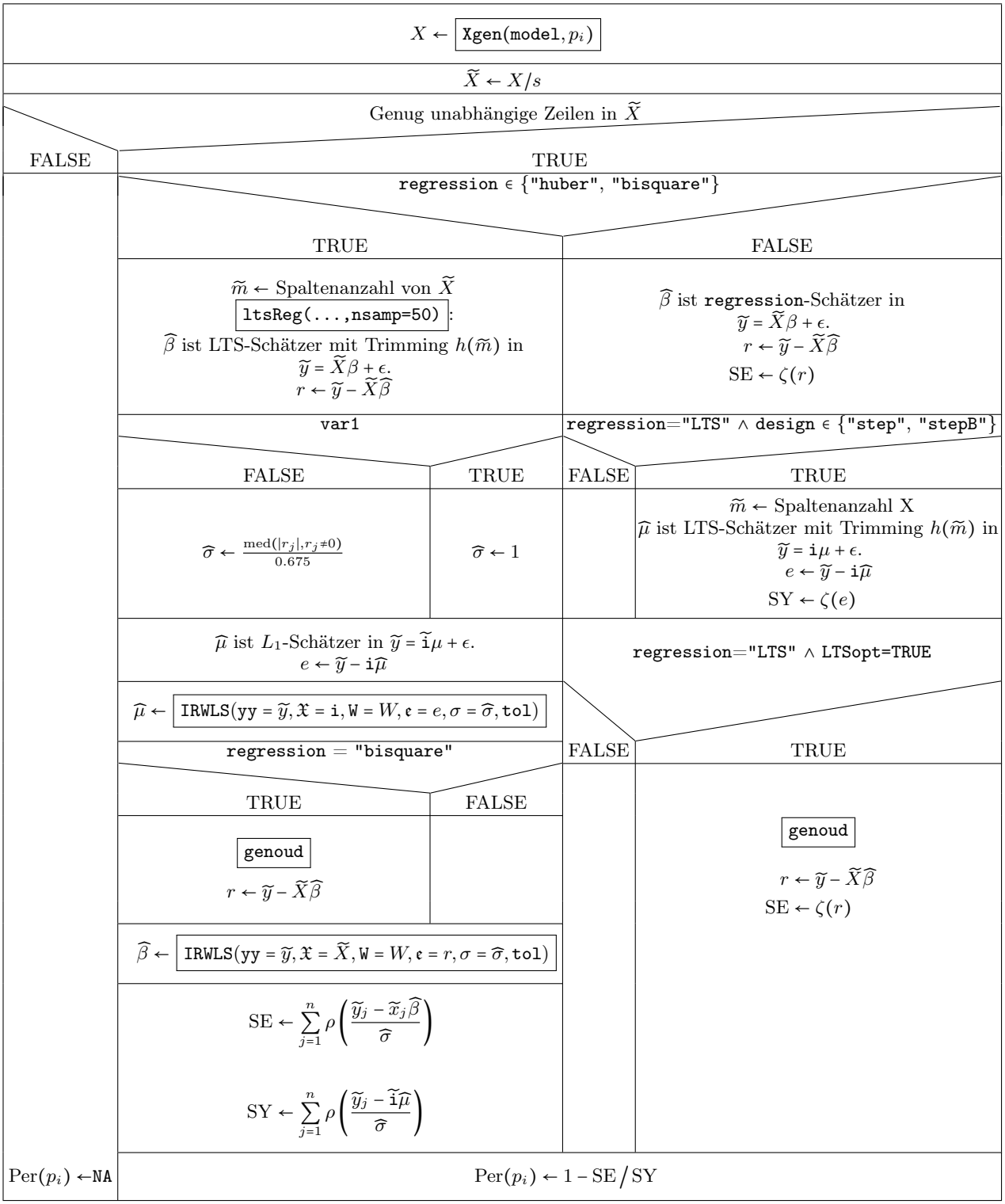


Abbildung F.3.: Struktogramm von singleFUN. NA ist der Indikator für einen fehlenden Wert. Der Block IRWLS wird in Abbildung F.4 detailliert beschrieben.

Eingabe	Symbol	Erläuterung
`yy` $\in \mathbb{R}^n$	`yy`	Messwerte
`matrix_` $\in \mathbb{R}^{n \times m}$	$\mathfrak{X}$	Designmatrix
`W`:$\mathbb{R} \to \mathbb{R}_{\geq 0}$	`W`	Gewichtungsfunktion
`residuals_` $\in \mathbb{R}^n$	$\mathfrak{e}$	Residualvektor
`scale_` $\in \mathbb{R}_{>0}$	σ	(Schätzer der) Standardabweichung
`tol` $\in \mathbb{R}_{>0}$	`tol`	Konvergenzschranke
Ausgabe		
`tempIRWLS$coeff`	$\widehat{b}$	Angepasster Parametervektor

Tabelle F.2.: Ein- und Ausgabeparameter der Funktion `IRWLS`.

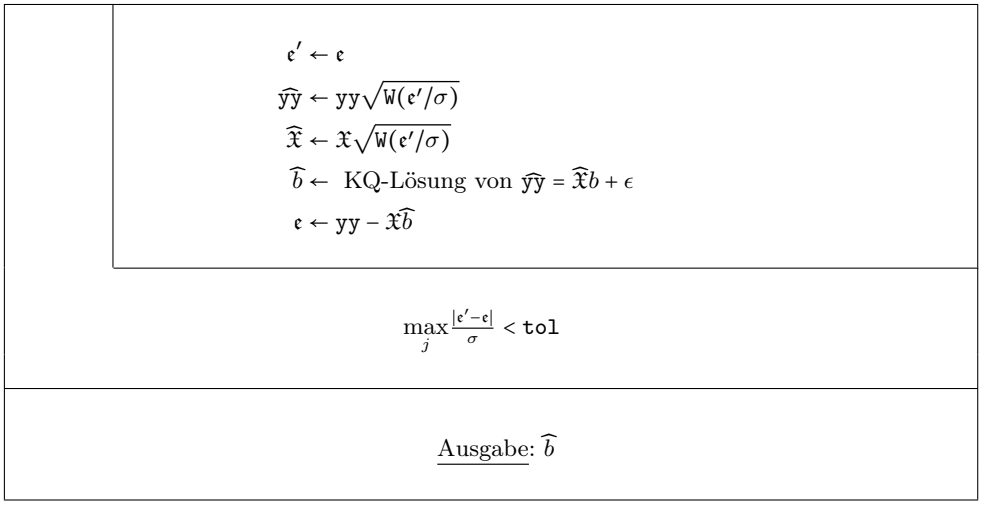


Abbildung F.4.: Struktogramm von `IRWLS` in `RobPer`

G. Eingabeparameter des Lichtkurvengenerators

Im Lichtkurvengenerator `tsgen` werden nacheinander die R-Funktionen `sampler`, `signalgen`, `lc_noise` und `disturber` aufgerufen. Die hierzu nötigen Eingabeparameter sind in den Tabellen G.1 bis G.4 aufgelistet. Bei grauer Unterlegung sind sie zugleich Eingabeparameter von `tsgen`. Details zu den Arbeitsweisen der Funktionen sind in Abschnitt 3.2 zu finden.

Parameter	Erläuterung
`ps`$\in \mathbb{R}_{>0}$	Samplingperiode p_s. Voreinstellung 1.
`Ncycles`$\in \mathbb{N}$	Anzahl n_s der Samplingzyklen.
`npoints`$\in \mathbb{N}$	Stichprobenumfang n
`ttype`	Verteilung $\mathcal{D}(p_s)$ der ungeordneten Messzeiten. Zur Auswahl stehen die Einstellungen `"equi"` (äquidistantes Sampling), `"unif"` (rechteckverteiltes Sampling), `"sine"` (mit Dichte d_{sin}) und `"trian"` (mit Dichte d_{trian}).

Tabelle G.1.: Eingabeparameter der Funktion `sampler`. Ausgabe ist ein Vektor $t \in \mathbb{R}^n$ geordneter Messzeiten. Grau unterlegte Parameter sind zugleich Eingabeparameter für die Mantelfunktion `tsgen`.

Parameter	Erläuterung
`tt`$\in \mathbb{R}^n$	Messzeiten $t_1, \ldots, t_n$, zum Beispiel die Ausgabe von `sampler`.
`pf`$\in \mathbb{R}_{>0}$	Fluktuationsperiode p_f. Voreinstellung 1.
`ytype`	Periodische Fluktuation f. Zur Auswahl stehen die Einstellungen `"const"` (konstante Funktion), `"sine"` (Sinusfunktion), `"trian"` (Dreiecksfunktion) und `"peak"` (Peakfunktion).

Tabelle G.2.: Eingabeparameter der Funktion `signalgen`. Ausgabe ist ein Vektor $y_f \in \mathbb{R}^n$ von Werten der periodischen Fluktuation. Grau unterlegte Parameter sind zugleich Eingabeparameter für die Mantelfunktion `tsgen`.

Parameter	Erläuterung
`tt`$\in \mathbb{R}^n$	Messzeiten $t_1, \ldots, t_n$
`sig`$\in \mathbb{R}^n$	Werte $y_{f;1}, \ldots, y_{f;n}$ der periodischen Fluktuation, z.B. Ausgabe von `signalgen`
`SNR`$\in \mathbb{R}_{>0}$	Verhältnis $\mathrm{var}(y_f)/\mathrm{var}(y_w + y_r)$
`redpart`$\in [0,1]$	Anteil $\mathrm{var}(y_r)/(\mathrm{var}(y_w) + \mathrm{var}(y_r))$ unbekannten Rauschens
`alpha`	Wird für $y_{r;1}, \ldots, y_{r;n} \underset{u.i.v.}{\sim} \mathcal{N}(0, \sigma^2)$ auf null gesetzt. Andere sinnvolle Einstellungen werden in Abschnitt 6.2 diskutiert.

Tabelle G.3.: Eingabeparameter der Funktion `lc_noise`. Ausgabe sind ein Vektor $y \in \mathbb{R}^n$ von Messwerten und ein Vektor $s \in \mathbb{R}^n$ von Messfehlern. Grau unterlegte Parameter sind zugleich Eingabeparameter für die Mantelfunktion `tsgen`.

Parameter	Erläuterung
`tt`$\in \mathbb{R}^n$	Messzeiten $t_1, \ldots, t_n$
`y`$\in \mathbb{R}^n$	Messwerte $y_1, \ldots, y_n$, z.B. Ausgabe von `lc_noise`
`s`$\in \mathbb{R}^n_{>0}$	Messfehler $s_1, \ldots, s_n$, z.B. Ausgabe von `lc_noise`
`ps`$\in \mathbb{R}_{>0}$	Samplingperiode p_s. Im Falle einer Intervallstörung überdeckt der ersetzende Ausschlag drei Samplingzyklen.
`s.outlier.fraction`$\in [0,1]$	Anteil der gestörten Messfehler
`interval`$\in \{$`TRUE`, `FALSE`$\}$	Ob eine Intervallstörung durchgeführt werden soll.

Tabelle G.4.: Eingabeparameter der Funktion `disturber`. Ausgabe sind ein Vektor $y \in \mathbb{R}^n$ von Messwerten und ein Vektor $s \in \mathbb{R}^n$ von Messfehlern. Grau unterlegte Parameter sind zugleich Eingabeparameter für die Mantelfunktion `tsgen`.

H. Eigenschaften der Peakfunktion $f_{peak:p_f}$

Sei

$$f_{peak:p_f}(t) = \begin{cases} 9\exp\left(-3p_f^2\left(\varphi_1(t)-\frac{2}{3}\right)^2\right) & 0 \le \varphi_1(t) \le \frac{2}{3} \\ 9\exp\left(-12p_f^2\left(\varphi_1(t)-\frac{2}{3}\right)^2\right) & \frac{2}{3} < \varphi_1(t) \le 1 \end{cases}$$

gegeben wie in Gleichung (3.9) auf Seite 48 mit $\varphi_1(t) = t - \lfloor t \rfloor$ (vgl. Gleichung 2.5, Seite 9). Dann gilt:
$f_{peak:p_f}(t)$ ist stetig in t:

$$\begin{aligned} \lim_{t\uparrow 0} f_{peak:p_f}(t) &= \lim_{t\uparrow 1} f_{peak:p_f}(t) = f_{peak:p_f}(1) = 9\exp\left(-12p_f^2\left(1-\frac{2}{3}\right)^2\right) \\ &= 9\exp\left(-3p_f^2\left(0-\frac{2}{3}\right)^2\right) = f_{peak:p_f}(0) = \lim_{t\downarrow 0} f_{peak:p_f}(t) \\ \lim_{t\uparrow \frac{2}{3}} f_{peak:p_f}(t) &= 9\exp\left(-3p_f^2\left(\frac{2}{3}-\frac{2}{3}\right)^2\right) \\ &= 9\exp\left(-12p_f^2\left(\frac{2}{3}-\frac{2}{3}\right)^2\right) = f_{peak:p_f}\left(\frac{2}{3}\right) = \lim_{t\downarrow \frac{2}{3}} f_{peak:p_f}(t) \end{aligned}$$

□

$f_{peak:p_f}(t)$ überschreitet für $p_f > 1$ in jedem Zyklus c für genau $\frac{1}{p_f}$ Zeiteinheiten ($t \in [c+\frac{2}{3}-\frac{2}{3p_f}, c+\frac{2}{3}+\frac{1}{3p_f}], c \in \mathbb{Z}$) den Wert $9\exp(-\frac{4}{3}) \approx 2.37$:
Sei o.B.d.A. $c = 0$. Mit

$$\varphi_1\left(\frac{2}{3}-\frac{2}{3p_f}\right) = \frac{2}{3}-\frac{2}{3p_f}-\left\lfloor \overbrace{\left(1-\frac{1}{p_f}\right)}^{<1}\overbrace{\frac{2}{3}}^{<1} \right\rfloor = \frac{2}{3}\left(1-\frac{1}{p_f}\right)$$

gilt

$$\begin{aligned} f_{peak:p_f}\left(\frac{2}{3}-\frac{2}{3p_f}\right) &= 9\exp\left(-3p_f^2\left(\varphi_1\left(\frac{2}{3}-\frac{2}{3p_f}\right)-\frac{2}{3}\right)^2\right) \\ &= 9\exp\left(-3p_f^2\left(\frac{2}{3}\left(1-\frac{1}{p_f}\right)-\frac{2}{3}\right)^2\right) \\ &= 9\exp\left(-3p_f^2\left(-\frac{2}{3p_f}\right)^2\right) = 9\exp\left(-\frac{4}{3}\right). \end{aligned}$$

H. Eigenschaften der Peakfunktion $f_{peak:p_f}$

Mit

$$\varphi_1\left(\frac{2}{3}+\frac{1}{3p_f}\right)=\frac{2}{3}+\frac{1}{3p_f}-\left\lfloor\overbrace{\frac{2}{3}+\frac{1}{3p_f}}^{<1}\right\rfloor=\frac{2}{3}+\frac{1}{3p_f}$$

gilt

$$\begin{aligned}f_{peak:p_f}\left(\frac{2}{3}+\frac{1}{3p_f}\right)&=9\exp\left(-12p_f^2\left(\varphi_1\left(\frac{2}{3}+\frac{1}{3p_f}\right)-\frac{2}{3}\right)^2\right)\\&=9\exp\left(-12p_f^2\left(\frac{2}{3}+\frac{1}{3p_f}-\frac{2}{3}\right)^2\right)\\&=9\exp\left(-12p_f^2\left(\frac{1}{3p_f}\right)^2\right)=9\exp\left(-\frac{4}{3}\right).\end{aligned}$$

Nach Definition ist $f_{peak:p_f}$ in t im Intervall $]0,\frac{2}{3}]$ streng monoton steigend und im Intervall $]\frac{2}{3},1[$ streng monoton fallend, damit gilt für $p_f>1$

$$f_{peak:p_f}(t)\begin{cases}\geq 9\exp\left(-\frac{4}{3}\right), t\in[\frac{2}{3}-\frac{2}{3p_f},\frac{2}{3}+\frac{1}{3p_f}]\\<9\exp\left(-\frac{4}{3}\right), t\notin[\frac{2}{3}-\frac{2}{3p_f},\frac{2}{3}+\frac{1}{3p_f}]\end{cases}.$$

□

I. Zusätzliche Details zur Simulationsstudie

Dieses Kapitel enthält detailliertere Informationen zu den in Kapitel 4 gewählten Kennzahl-Repräsentanten (Abschnitt I.1) und den gemessenen Rechenzeiten (Abschnitt I.2).

I.1. Detektionskurvennachweis

In den Tabellen I.1(a) und (b) ist verzeichnet, zu welchen Versuchen die Detektionskurven gehören, die in Abbildung 4.4 auf Seite 61 dargestellt sind. Die Versuche sind als Repräsentanten für Kurven mit entsprechender MDα oder MAAW zu verstehen, nicht als Repräsentanten für die entsprechenden Detektionsmethoden oder Datenszenarien.

	Wskt. δ	δ-Quantil	Regression	Modell	Auswertung	$\mathcal{D}(p_s)$	I.Störung
- - - - -	0	0,000	KQ	fourier(3)	Typ 1	d_{sin}	Ja
—•—	0,116	0,050	M-Tukey	fourier(3)	Typ 1	d_{sin}	Nein
·····	0,25	0,060	L1 (g)	step	Typ 2	d_{trian}	Nein
-·-·-·	0,5	0,820	M-Huber	sin	Typ 1	d_{sin}	Nein
———	0,75	0,108	M-Tukey (g)	2step	Typ 3	d_{trian}	Ja
———	1	0,799	LTS	sin	Typ 2	d_{sin}	Ja

(a) MDα

	Wskt. δ	δ-Quantil	Regression	Modell	Auswertung	$\mathcal{D}(p_s)$	I.Störung
- - - - -	0	0,002	M-Tukey (g)	step	Typ 2	d_{trian}	Nein
—•—	0,193	0,010	τ (g)	splines	Typ 1	d_{sin}	Ja
·····	0,25	0,013	M-Tukey (g)	fourier(3)	Typ 2	d_{trian}	Ja
-·-·-·	0,5	0,034	τ	fourier(2)	Typ 3	d_{sin}	Nein
———	0,75	0,058	M-Tukey (g)	2step	Typ 3	d_{trian}	Nein
———	1	0,749	LTS	sin	Typ 2	d_{sin}	Ja

(b) MAAW

Tabelle I.1.: Repräsentanten für verschiedene Quantile der Kennzahlen MDα und MAAW. Linie bezieht sich auf die in Abbildung 4.4 verwendeten Linientypen. Die verwendeten Abkürzungen sind in Abbildung 4.2, Seite 56, erläutert. Mit I.Störung ist eine Intervallstörung gemeint.

I.2. Rechenzeiten

Die Abbildungen I.1 bis I.6 stellen die in der Simulationsstudie benötigten Rechenzeiten dar. Es werden nicht nur, wie in Abschnitt 4.5, die einzelnen Regressionstechniken (hier farbig kodiert) und angepassten Modelle (je eine Abbildung) unterschieden, sondern auch, ob gewichtet oder ungewichtet (unterschieden mit (u) und (g)) angepasst wird, sowie gemäß den folgenden Datenszenarien:

A Keine Periodische Fluktuation, keine Intervallstörung,

B Sinusfluktuation der Periode 14, keine Intervallstörung,

C Sinusfluktuation der Periode 33, keine Intervallstörung,

D Peakfluktuation der Periode 14, keine Intervallstörung,

E Peakfluktuation der Periode 33, keine Intervallstörung,

F Keine Periodische Fluktuation, Intervallstörung,

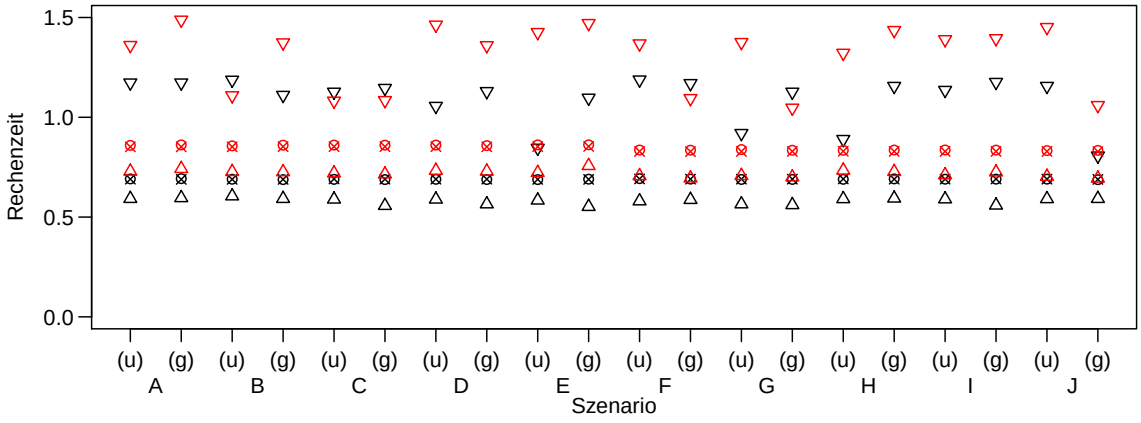
G Sinusfluktuation der Periode 14, Intervallstörung,

H Sinusfluktuation der Periode 33, Intervallstörung,

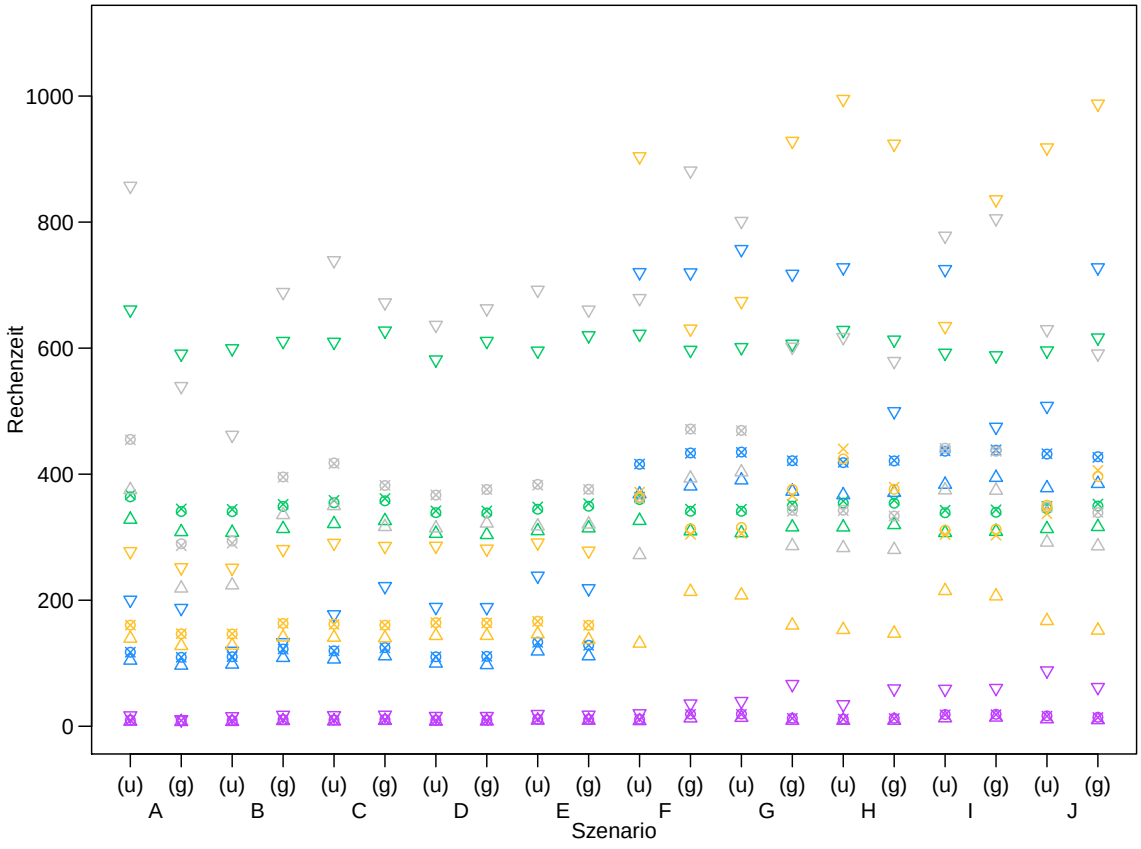
I Peakfluktuation der Periode 14, Intervallstörung,

J Peakfluktuation der Periode 33, Intervallstörung.

Es sind jeweils Minimum ($\triangle$), arithmetisches Mittel ($\times$), Median ($\circ$) und Maximum ($\triangledown$) der jeweiligen Rechenzeiten abgetragen.

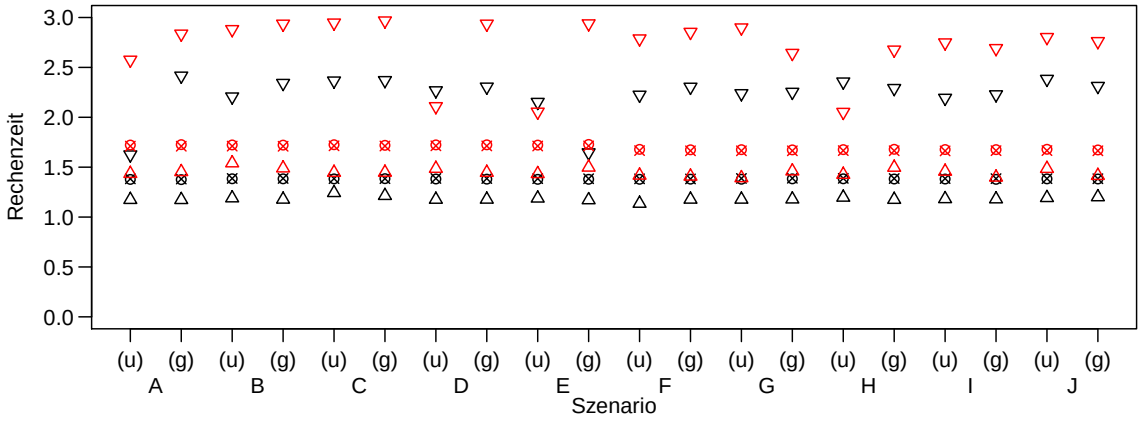


(a)

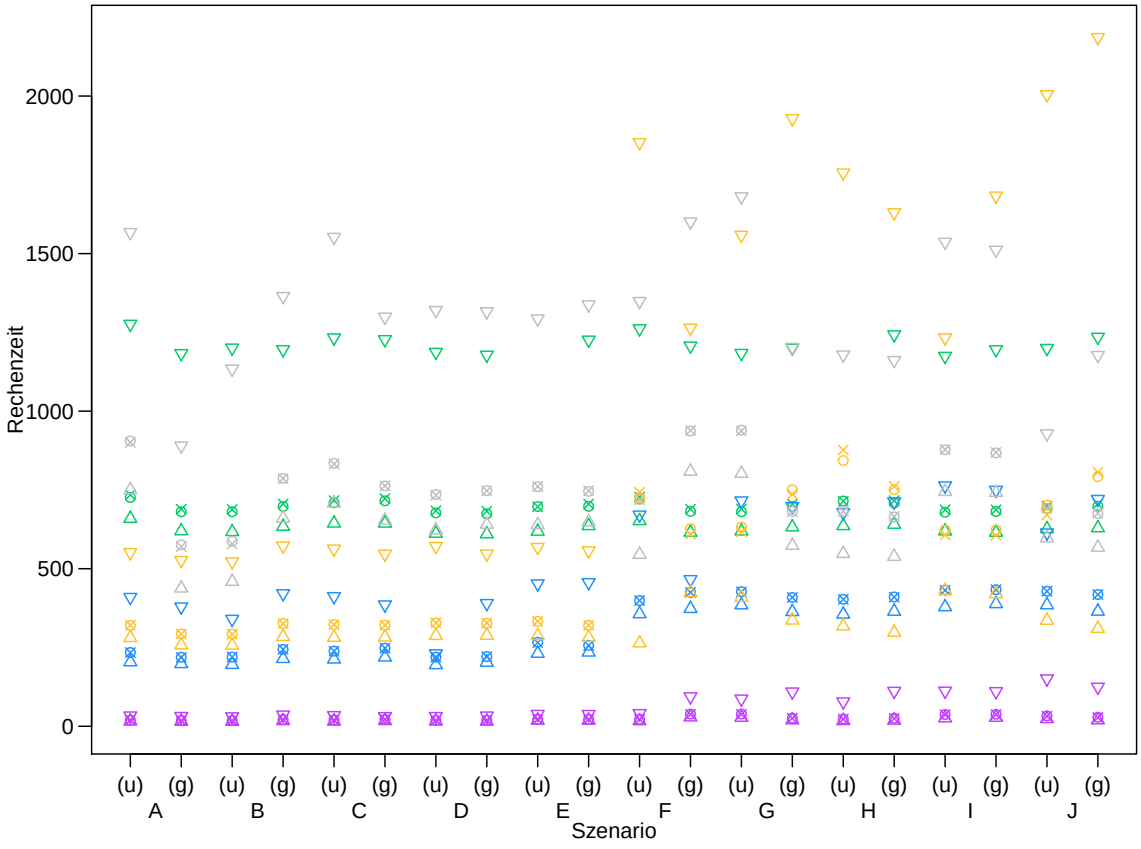


(b)

Abbildung I.1.: Rechenzeiten bei Anpassung der Einfachstufenfunktion in Sekunden: Für Erklärung der Symbole und Kodierung der Datenszenarien vgl. Seite 164. Regressionstechniken: (a) ● KQ und ● L1, (b) ● LTS, ● S, ● τ, ● M-Huber und ● M-Tukey.

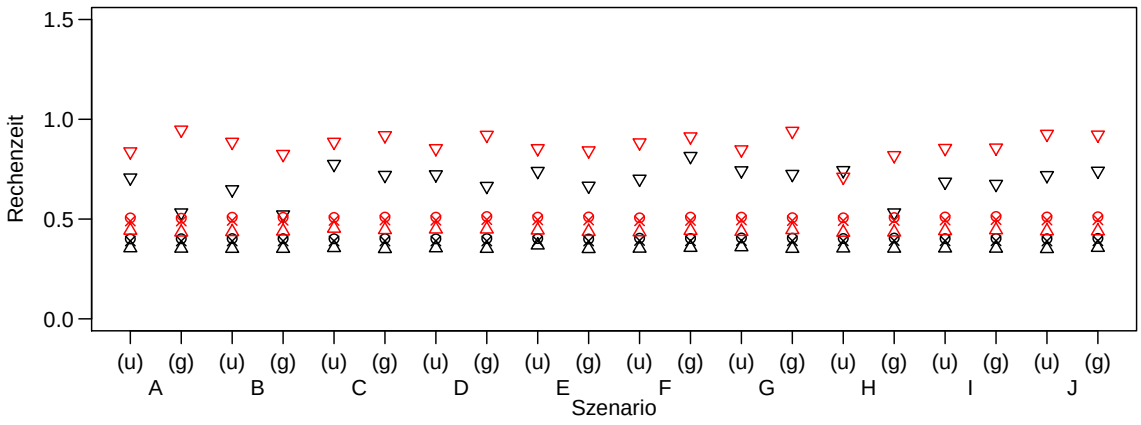


(a)

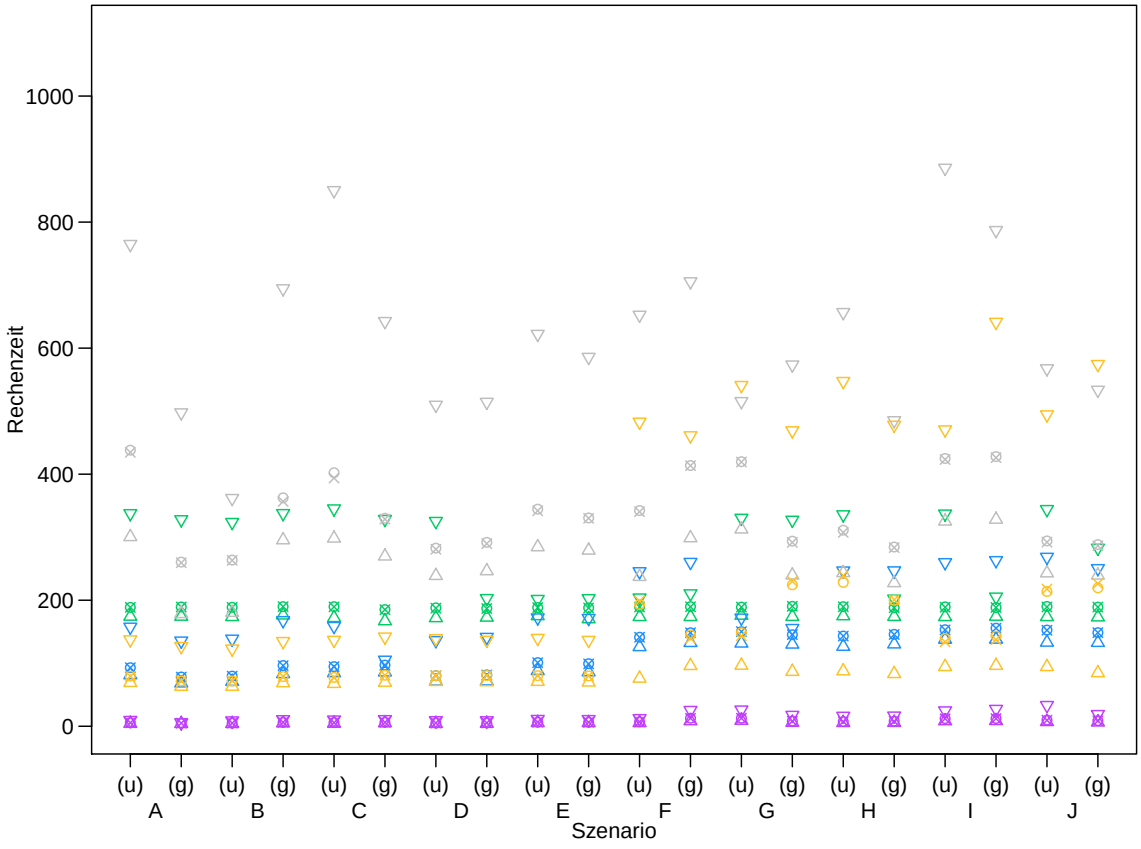


(b)

Abbildung I.2.: Rechenzeiten bei Anpassung der Doppelstufenfunktion in Sekunden: Für Erklärung der Symbole und Kodierung der Datenszenarien vgl. Seite 164. Regressionstechniken: (a) ● KQ und ● L1, (b) ● LTS, ● S, ● τ, ● M-Huber und ● M-Tukey.

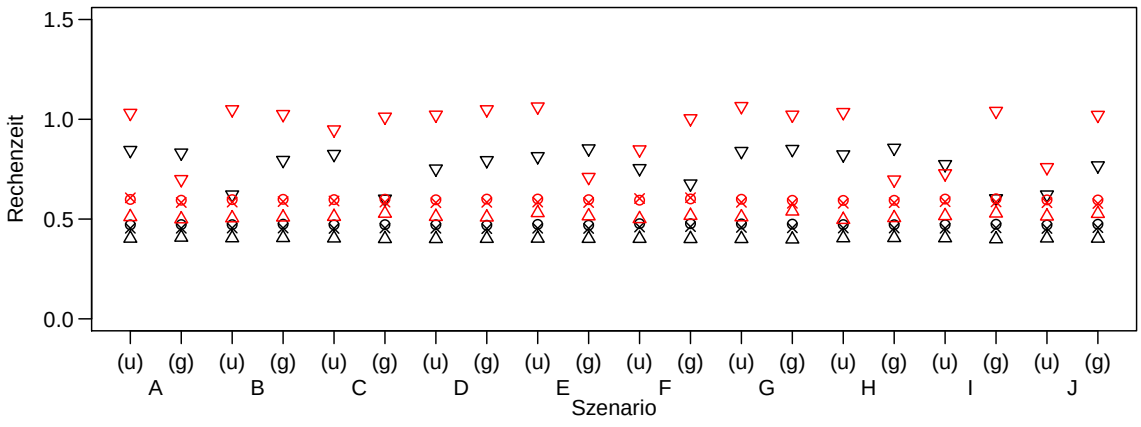


(a)

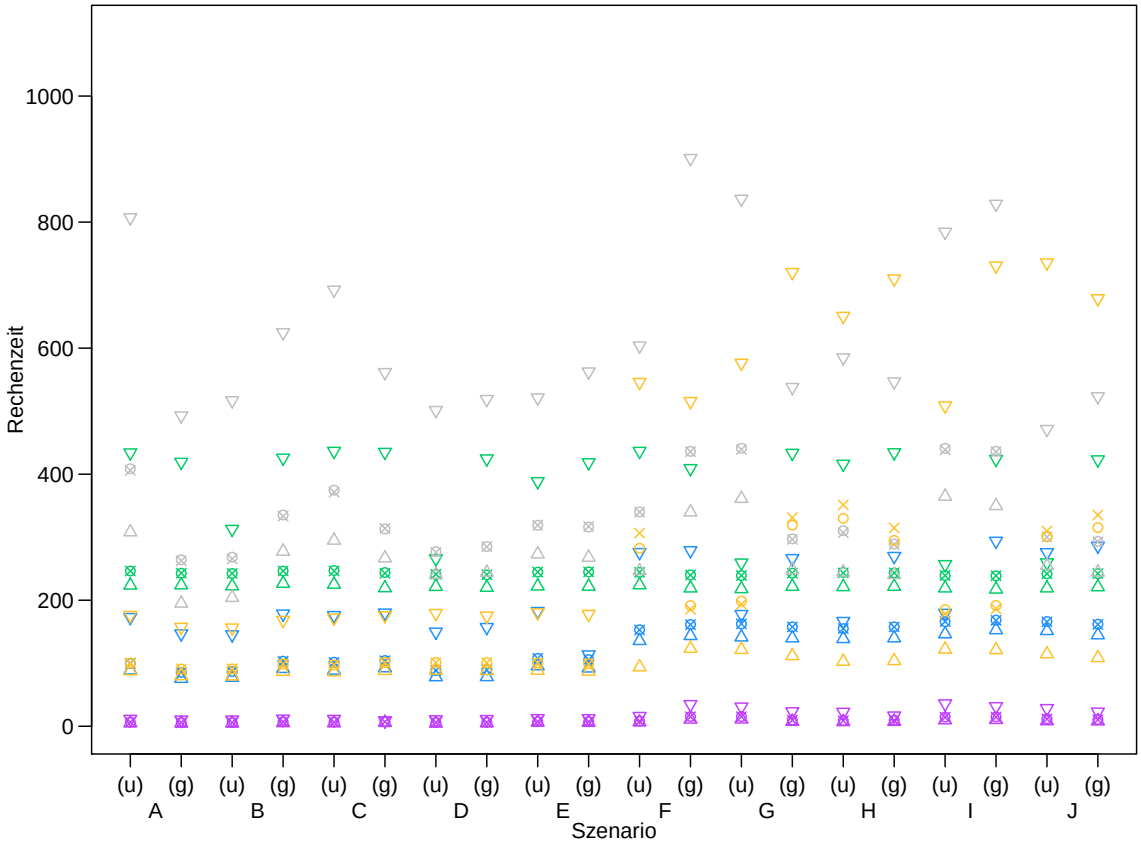


(b)

Abbildung I.3.: Rechenzeiten bei Anpassung der Sinusfunktion in Sekunden: Für Erklärung der Symbole und Kodierung der Datenszenarien vgl. Seite 164. Regressionstechniken: (a) ● KQ und ● L1, (b) ● LTS, ● S, ● τ, ● M-Huber und ● M-Tukey.

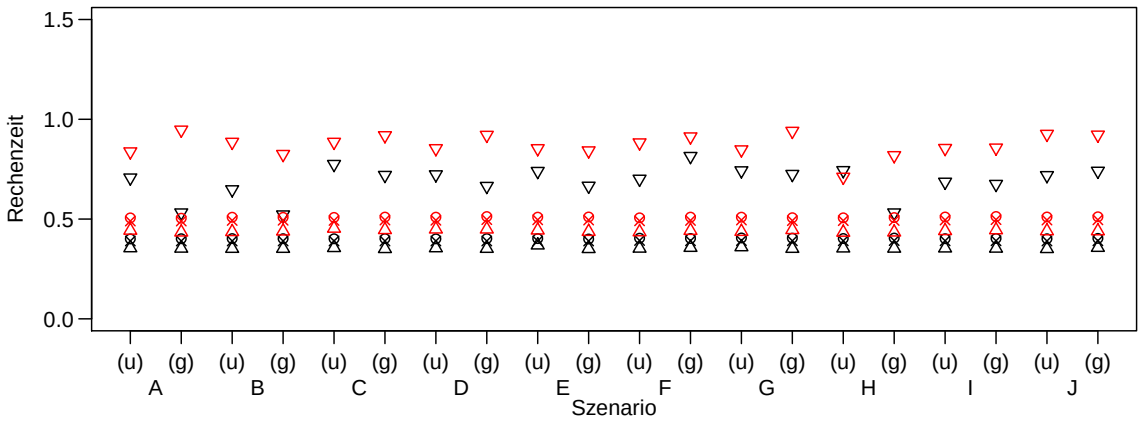


(a)

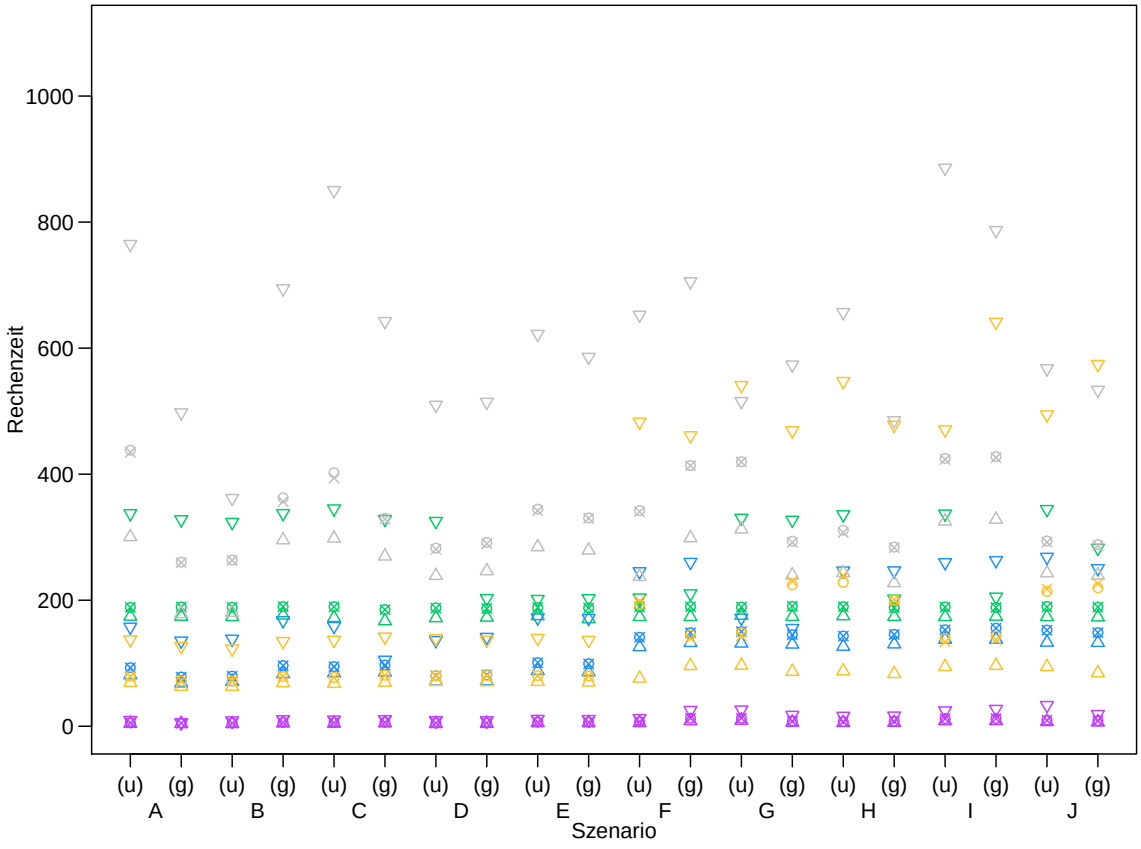


(b)

Abbildung I.4.: Rechenzeiten bei Anpassung der Fouriersumme zweiten Grades in Sekunden: Für Erklärung der Symbole und Kodierung der Datenszenarien vgl. Seite 164. Regressionstechniken: (a) ● KQ und ● L1, (b) ● LTS, ● S, ● τ, ● M-Huber und ● M-Tukey.

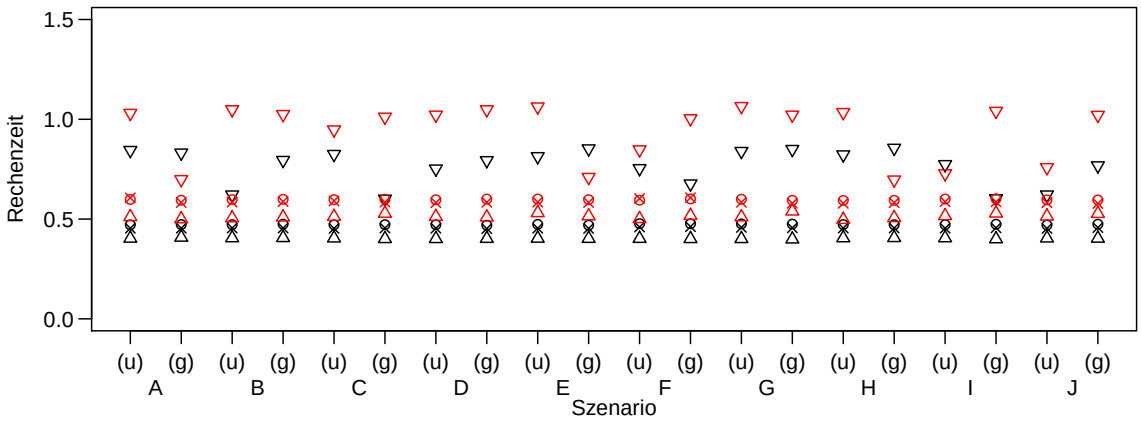


(a)

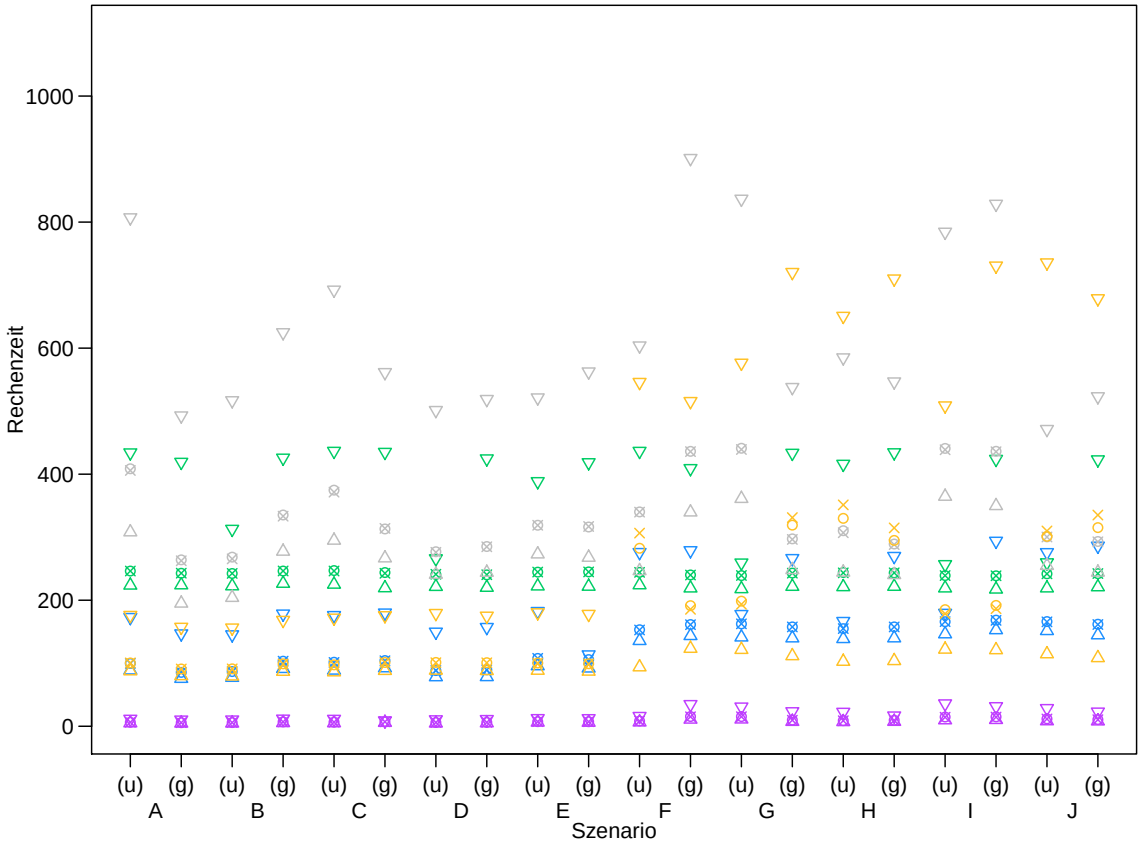


(b)

Abbildung I.5.: Rechenzeiten bei Anpassung der Fouriersumme dritten Grades in Sekunden: Für Erklärung der Symbole und Kodierung der Datenszenarien vgl. Seite 164. Regressionstechniken: (a) ● KQ und ● L1, (b) ● LTS, ● S, ● τ, ● M-Huber und ● M-Tukey.



(a)



(b)

Abbildung I.6.: Rechenzeiten bei Anpassung der Splinefunktion in Sekunden: Für Erklärung der Symbole und Kodierung der Datenszenarien vgl. Seite 164. Regressionstechniken: (a) ● KQ und ● L1, (b) ● LTS, ● S, ● τ, ● M-Huber und ● M-Tukey.

Symbolverzeichnis

Mathematische Zeichen

$[a_1, a_2], [a_1, a_2[$	Unten geschlossene Intervalle von a_1 bis a_2
$]a_1, a_2],]a_1, a_2[$	Unten offene Intervalle von a_1 bis a_2
$[a_1, a_2],]a_1, a_2]$	Oben geschlossene Intervalle von a_1 bis a_2
$[a_1, a_2[,]a_1, a_2[$	Oben offene Intervalle von a_1 bis a_2
$\lfloor \ \rfloor$	Abrundende Gaußklammer
$\langle \ \rangle$	Kaufmännisches Runden
$\|\|u\|\|$	Euklidische Norm des Vektors u
$\|M\|$	Betrag der Zahl M oder Mächtigkeit der Menge M
$()^\mathsf{T}$	Transponiert
$0_n, 1_n$	$(0, \ldots, 0)^\mathsf{T}, (1, \ldots, 1)^\mathsf{T} \in \mathbb{R}^n$
$u_{(i)}$	i-ter geordneter Wert des Vektors u
$\sim$	Verteilt
$\bar{u}$	Arithmetisches Mittel der Einträge von Vektor u
$\widehat{g}$	Schätzung der Funktion oder des Wertes g
$\widehat{u}^{(k)}$	k-ter Schätzung von u in iterativem Schätzverfahren ($\widehat{u}^{(0)}$ Initialschätzer)
u_k	k-te Komponente im Vektor u

Mehrbuchstabige Abkürzungen

i.v.	identisch verteilt
Per	Periodogrammbalkenfunktion
range(y)	Spannweite des Vektors y
SE	Minimiertes Zielkriterium im Lokationsmodell
SY	Minimiertes Zielkriterium im vollen Modell
u.i.v.	unabhängig identisch verteilt
var(u)	empirische Varianz des Vektors u

Lateinische Buchstaben

a	Kontextabhängige Bedeutung
A	Amplitude
$b_\alpha(\theta_1, \theta_2)$	α-Quantil der $\mathcal{B}(\theta_1, \theta_2)$-Verteilung
c	Kontextabhängige Bedeutung
d	Dichtefunktion
e_i	Fehlerterm in einem Modell

E	Erwartungswert
f	Funktion, die die periodische Fluktuation bestimmt
F_n	Empirische Verteilungsfunktion
F_θ	Theoretische Verteilungsfunktion mit Parameter(-vektor) θ
g	Periodische Funktion, die angepasst wird
h, i	Kontextabhängige Bedeutung
I	Intervall, Menge
j, J	Kontextabhängige Bedeutung
k	Kontextabhängige Bedeutung
k_i	Sprungstellen einer Stufenfunktion
l	Kontextabhängige Bedeutung
m	Dimension des angepassten Modells
M	Kontextabhängige Bedeutung
n	Stichprobenumfang
$n_1, \ldots, n_m$	Stichprobenumfänge in Teilstichproben
n_s	Anzahl Samplingzyklen
N	Kontextabhängige Bedeutung
p	Periode
$p_1, \ldots, p_q$	Testperioden
p_f	Fluktuationsperiode
p_s	Samplingperiode
$P(M)$	Wahrscheinlichkeit von M
q	Anzahl Testperioden
r_i, r	Residuum (Residuenvektor)
R^2	Bestimmtheitsmaß
s_i, s	Messfehler (-vektor)
$\dot{s}$	Wert der s-Ausreißer
t_i, t	Messzeit (-vektor)
$t_i^\star$	i-ter Messzeit der nach Phase geordneten Lichtkuve
T	Länge der Lichtkurve
T_i	Zufallsvariable zu t_i
u, U	Kontextabhängige Bedeutung
v	Kontextabhängige Bedeutung
V	Autokovarianz
W	Spektralfenster
$x_1, x_2, \ldots$	Beobachtungen einer beliebigen äquidistanten Zeitreihe
X_i	Zeile der Designmatrix X
X	Designmatrix
y_i, y	Messwert (-vektor)
$y_i^\star$	i-ter Messwert der nach Phase geordneten Lichtkurve
$y_{f;i}$	Signalkomponente des Messwertes
$y_{w;i}$	Von s_i abhängiges normalverteiltes Rauschen im Messwert (weißes Rauschen)

$y_{r;i}$	Von s_i unabhängige Rauschkomponente (u.a. rotes Rauschen)
$Y_i, Y_{f;i}, Y_{w;i}, Y_{r;i}$	Zufallsvariablen zu den entsprechenden kleingeschriebenen Werten
z_p	Zyklusfunktion
z_i	Samplingzyklus von Beobachtung i
Z	Kontextabhängige Bedeutung

Griechische Buchstaben

α	Kontextabhängige Bedeutung
β	Regressionskoeffizientenvektor
$\Gamma(\theta_1, \theta_2)$	Gammaverteilung mit Erwartungswert $\frac{\theta_1}{\theta_2}$ und Varianz $\frac{\theta_1}{\theta_2^2}$
$\Gamma(u)$	Wert der Gammafunktion an der Stelle $u \in \mathbb{R}$
δ	Kontextabhängige Bedeutung
Δ_{CvM}	Cramér-von-Mises-Distanz
ϵ	Fehlerterm
ζ	Zielfunktion einer Regressionstechnik
η	Parameter zur Steuerung der Rauschstärke in `lc_noise`
θ	Parameter(-vektor) einer Verteilung
λ	kontextabhängige Bedeutung
μ	Lokationsparameter
ν	Residuum
ξ	Zeitpunkt auf genormter Zeitachse
π	Kreiszahl
ρ	komponentenweise Minimierungsfunktion der M-Regression
σ	Standardabweichung
τ	Als Bestandteil der Bezeichnung "τ-Regression"
φ	Phasenanteil einer umgeklappten Beobachtung
Φ	Filterfunktion
ϕ	Phase als Parameter der Sinusfunktion
ψ	steuert den Anteils nicht durch s_i erklärten Rauschens in `lc_noise`

Andere Schriftarten

$\mathbb{1}$	Indikatorfunktion
$\mathcal{B}(\theta_1, \theta_2)$	Betaverteilung mit Parametern θ_1 und θ_2
∂	Ableitungsoperator
$\mathcal{D}(p)$	Periodische Verteilung mit Periode p
$\mathfrak{F}_n$	Diskrete Fouriertransformierte
$\mathfrak{F}$	Fouriertransformierte
$\mathcal{F}_{\theta_1,\theta_2}$	F-Verteilung mit Parametern θ_1 und θ_2
$\mathfrak{i}$	Imaginäre Einheit $=\sqrt{-1}$
$\mathbb{I}_n$	$n \times n$-Einheitsmatrix
$\mathcal{I}$	Partition des Intervalls $[0, 1[$
$\mathbb{L}$	Menge der Lichtkurven
$\mathbb{N}$	Menge der natürlichen Zahlen ohne null

$\mathbb{N}_0$	Menge der natürlichen Zahlen mit null
$\mathcal{N}(\theta_1, \theta_2^2)$	Normalverteilung mit Erwartungswert θ_1 und Varianz θ_2^2
$\mathbb{P}$	Menge der Testperioden
$\mathbb{R}$	Menge der reellen Zahlen
$\mathbb{R}_{\geq 0}$	Menge der nicht negativen reellen Zahlen
$\mathbb{R}_{>0}$	Menge der streng positiven reellen Zahlen
$\mathcal{R}(u)$	Rangvektor des Vektors u
$\mathcal{R}(u)_i$	Rang der Beobachtung u_i
$\mathcal{U}_M$	Gleichverteilung auf M (je nach M stetig oder disktret)
$\mathcal{X}$	Regressorbildende Funktion
$\mathfrak{X}$	Designmatrix-Platzhalter in IRWLS-Struktogrammdarstellung
$\mathfrak{Y}, \mathfrak{Z}$	Lichtkurve
$\mathbb{Z}$	Menge der ganzen Zahlen

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Anmerkung: In der Physik sind Artikel mit mehr als 20, 30 oder 100 Autoren keine Seltenheit. Daher wird in diesem Verzeichnis abweichend von der in der statistischen Literatur gängigen Praxis darauf verzichtet, stets alle Autoren zu nennen. Es werden nur die ersten zehn Autoren ausgeschrieben.

Abdo, A. A., **Ackermann, M.**, **Ajello, M.**, **Baldini, L.**, **Ballet, J.**, **Barbiellini, G.**, **Bastieri, D.**, **Bechtol, K.**, **Bellazzini, R.**, **Berenji, B.** et al. **(2011):** Fermi large area telescope observations of Markarian 421: The missing piece of its spectral energy distribution. The Astrophysical Journal, 736 Nr. 2 131–152

Ahdesmäki, M., **Fokianos, K.** und **Strimmer, K. (2012):** GeneCycle: identification of periodically expressed genes. R-Paket, Version 1.1.2 ⟨URL: `http://CRAN.R-project.org/package=GeneCycle`⟩

Ahdesmäki, M., **Lähdesmäki, H.**, **Gracey, A.**, **Shmulevich, I.** und **Yli-Harja, O. (2007):** Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data. BMC Bioinformatics, 8 Nr. 1 233–248

Akerlof, C., **Alcock, C.**, **Allsman, R.**, **Axelrod, T.**, **Bennett, D. P.**, **Cook, K. H.**, **Freeman, K.**, **Griest, K.**, **Marshall, S.**, **Park, H. S.** et al. **(1994):** Application of cubic splines to the spectral analysis of unequally spaced data. The Astrophysical Journal, 436 Nr. 2 787–794

Albert, J., **Aliu, E.**, **Anderhub, H.**, **Antonelli, L. A.**, **Antoranz, P.**, **Backes, M.**, **Baixeras, C.**, **Barrio, J. A.**, **Bartko, H.**, **Bastieri, D.** et al. **(2009):** Periodic very high energy γ-ray emission from LS I+ 61 303 observed with the MAGIC telescope. The Astrophysical Journal, 693 Nr. 1 303–310

Albert, J., **Aliu, E.**, **Anderhub, H.**, **Antoranz, P.**, **Armada, A.**, **Asensio, M.**, **Baixeras, C.**, **Barrio, J. A.**, **Bartelt, M.**, **Bartko, H.** et al. **(2006):** Variable very-high-energy gamma-ray emission from the microquasar LS I+ 61 303. Science, 312 Nr. 5781 1771–1773

Albert, J., **Aliu, E.**, **Anderhub, H.**, **Antoranz, P.**, **Armada, A.**, **Baixeras, C.**, **Barrio, J. A.**, **Bartko, H.**, **Bastieri, D.**, **Becker, J. K.** et al. **(2008):** VHE γ-Ray observation of the Crab Nebula and its pulsar with the MAGIC telescope. The Astrophysical Journal, 674 Nr. 2 1037–1055

Aleksić, J., **Alvarez, E. A.**, **Antonelli, L. A.**, **Antoranz, P.**, **Asensio, M.**, **Backes, M.**, **Barrio, J. A.**, **Bastieri, D.**, **Becerra González, J.**, **Bednarek, W.** et al. **(2012):** Performance of the MAGIC stereo system obtained with Crab Nebula data. Astroparticle Physics, 35 Nr. 7 435–448

Andrew, B. H., **MacLeod, J. M.**, **Locke, J. L.**, **Medd, W. J.** und **Purton, C. R. (1969):** Rapid radio variations in BL Lac. Nature, 223 Nr. 5206 598–599

Baisch, S. und **Bokelmann, G. H. R. (1999):** Spectral analysis with incomplete time series: an example from seismology. Computers and Geosciences, 25 Nr. 7 739–750

Becker, C. (2010): Der Compute-Cluster des ITMC: LiDOng. IT und Medien Update, TU Dortmund, Nr. 11 5

Benlloch, S., **Wilms, J.**, **Edelson, R.**, **Yaqoob, T.** und **Staubert, R. (2001):** Quasi-periodic oscillation in Seyfert galaxies: Significance levels. The case of Markarian 766. Astrophysical Journal, 562 Nr. 2 L121–L124

Benlloch García, S. (2003): Long-term X-ray variability of Active Galactic Nuclei and X-ray Binaries. Dissertation, Eberhard-Karls-Universität Tübingen

Bergh, J., **Ekstedt, F.** und **Lindberg, M. (2007):** Wavelets mit Anwendungen in Signal-und Bildverarbeitung. Berlin, Heidelberg: Springer

Bjornstad, O. N. (2013): nlts: (non)linear time series analysis. R-Paket, Version 0.2-0 ⟨URL: `http://CRAN.R-project.org/package=nlts`⟩

Bloomfield, P. (2000): Fourier analysis of time series. 2. Auflage. New York: Wiley

Bourguignon, S., **Carfantan, H.** und **Idier, J. (2007):** A sparsity-based method for the estimation of spectral lines from irregularly sampled data. IEEE Journal of Selected Topics in Signal Processing, 1 Nr. 4 575–585

Bühlmann, P. und **Geer, S. van der (2011):** Statistics for high-dimensional data. Berlin: Springer

Busch, H. (2004): Pattern formation and synchronization in excitable systems under the influence of spatiotemporal colored noise. Dissertation, Technische Universität Darmstadt

Canny, J. (1986): A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 8 Nr. 6 679–714

Chadid, M., **Perini, C.**, **Bono, G.**, **Auvergne, M.**, **Baglin, A.**, **Weiss, W. W.** und **Deboscher, J. (2011):** CoRoT light curves of Blazhko RR Lyrae stars. Astronomy and Astrophysics, 527 ID A146 (9 Seiten)

Chatfield, C. (2003): The analysis of time series: An introduction. 6. Auflage. Boca Raton, London, New York, Washington, D.C.: Chapman & Hall/CRC

Clarke, B. R., **McKinnon, P. L.** und **Riley, G. (2012):** A fast robust method for fitting gamma distributions. Statistical Papers, 53 Nr. 4 1001–1014

Clarkson, W. I., **Charles, P. A.**, **Coe, M. J.**, **Laycock, S.**, **Tout, M. D.** und **Wilson, C. A. (2003):** Long-term properties of accretion discs in X-ray binaries–I. The variable third period in SMC X-1. Monthly Notices of the Royal Astronomical Society, 339 Nr. 2 447–454

Croux, C. und **Dehon, C. (2003):** Estimators of the multiple correlation coefficient: local robustness and confidence intervals. Statistical Papers, 44 Nr. 3 315–334

Croux, C., **Rousseeuw, P. J.** und **Hössjer, O. (1994):** Generalized S-estimators. Journal of the American Statistical Association, 89 Nr. 428 1271–1281

Cumming, A. (2004): Detectability of extrasolar planets in radial velocity surveys. Monthly Notices of the Royal Astronomical Society, 354 Nr. 4 1165–1176

Cumming, A., **Marcy, G. W.** und **Butler, R. P. (1999):** The lick planet search: detectability and mass thresholds. The Astrophysical Journal, 526 Nr. 2 890–915

Davies, L. und **Gather, U. (1993):** The identification of multiple outliers. Journal of the American Statistical Association, 88 Nr. 423 782–792

Dawson, R. I. und **Fabrycky, D. C. (2010):** Radial velocity planets de-aliased: A new, short period for super-earth 55 Cnc e. The Astrophysical Journal, 722 Nr. 1 937–953

Deeming, T. J. (1975): Fourier analysis with unequally-spaced data. Astrophysics and Space Science, 36 Nr. 1 137–158

Dierckx, P. (1993): Curve and surface fitting with splines. Oxford: Clarendon

Djurović, I., **Katkovnik, V.** und **Stanković, L. J. (2001):** Instantaneous frequency estimation based on the robust spectrogram. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Band 6, 3517–3520

Do, T., **Ghez, A. M.**, **Morris, M. R.**, **Yelda, S.**, **Meyer, L.**, **Lu, J. R.**, **Hornstein, S. D.** und **Matthews, K. (2009):** A near-infrared variability study of the galactic black hole: A red noise source with no detected periodicity. Astrophysical Journal, 691 Nr. 2 1021–1034

Drake, A. J., **Djorgovski, S. G.**, **Mahabal, A.**, **Beshore, E.**, **Larson, S.**, **Graham, M. J.**, **Williams, R.**, **Christensen, E.**, **Catelan, M.**, **Boattini., A.** et al. **(2009):** First results from the Catalina real-time transient survey. The Astrophysical Journal, 696 Nr. 1 870–884

Dupuy, D. L. und **Hoffman, G. A. (1985):** A Jurkevich period search program. International Amateur-Professional Photoelectric Photometry Communications, 20 1–17

Dworetsky, M. M. (1983): A period-finding method for sparse randomly spaced observations or 'How long is a piece of string?'. Monthly Notices of the Royal Astronomical Society, 203 917–924

Eyer, L. und **Genton, M. G. (1999):** Characterization of variable stars by robust wave variograms: An application to Hipparcos mission. Astronomy and Astrophysics Supplement Series, 136 Nr. 2 421–428

Ferraz-Mello, S. (1981): Estimation of periods from unequally spaced observations. The Astronomical Journal, 86 Nr. 4 619–624

Fishman, G. J., **Meegan, C. A.**, **Wilson, R. B.**, **Paciesas, W. S.**, **Parnell, T. A.**, **Austin, R. W.**, **Rehage, J. R.**, **Matteson, J. L.**, **Teegarden, B. J.**, **Cline, T. L.** et al. **(1989):** BATSE: The Burst And Transient Source Experiment on the Gamma Ray Observatory. In **Johnson, W. N.** (Hrsg.) Proceedings of the Gamma Ray Observatory Science Workshop 10-12 April 1989. 2:39–2:50

Fortin, P. und **Mackey, M. C. (1999):** Periodic chronic myelogenous leukaemia: spectral analysis of blood cell counts and aetiological implications. British Journal of Haematology, 104 Nr. 2 336–345

Foster, G. (1996): Wavelets for period analysis of unevenly sampled time series. The Astronomical Journal, 112 Nr. 4 1709–1729

Fried, R. und **Dehling, H.. (2011):** Robust nonparametric tests for the two sample location problem. Statistical Methods and Applications, 20 Nr. 4 409–422

Fried, R., **Raabe, N.** und **Thieler, A. M. (2012):** On the robust analysis of periodic nonstationary time series. In **Colubi, A.** (Hrsg.) 20th International Conference on Computational Statistics. 245–257

Glynn, E. F., **Chen, J.** und **Mushegian, A. R. (2006):** Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms. Bioinformatics, 22 Nr. 3 310–316

Graham, M. J., **Drake, A. J.**, **Djorgovski, S. G.**, **Mahabal, A. A.**, **Donalek, C.**, **Duan, V.** und **Maker, A. (2013):** A comparison of period finding algorithms. Monthly Notices of the Royal Astronomical Society, 434 Nr. 4 3423–3444

Grupen, C. (2000): Astroteilchenphysik: Das Universum im Licht der kosmischen Strahlung. 1. Auflage. Braunschweig, Wiesbaden: Vieweg

Gupta, A. K. und **Nadarajah, S. (2004):** Handbook of beta distribution and its applications. New York, Basel: Dekker

Hackbusch, W., **Schwarz, H. R.** und **Zeidler, E.; Zeidler, E.** (Hrsg.) **(2003):** Teubner-Taschenbuch der Mathematik. 2. Auflage. Wiesbaden: Teubner

Hall, P. und **Li, M. (2006):** Using the periodogram to estimate period in nonparametric regression. Biometrika, 93 Nr. 2 411–424

Hall, P., **Reimann, J.** und **Rice, J. (2000):** Nonparametric estimation of a periodic function. Biometrika, 87 Nr. 3 545–557

Hall, P. und **Yin, J. (2003):** Nonparametric methods for deconvolving multiperiodic functions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65 Nr. 4 869–886

Halpern, J. P., **Leighly, K. M.** und **Marshall, H. L. (2003):** An extreme ultraviolet explorer atlas of Seyfert galaxy light curves: Search for periodicity. The Astrophysical Journal, 585 Nr. 2 665–676

Höfler, A. (1913): Didaktik der Himmelskunde und der astronomischen Geographie. Leipzig: Teubner

Huijse, P., **Estevez, P. A.**, **Zegers, P.**, **Principe, J. C.** und **Protopapas, P. (2011):** Period estimation in astronomical time series using slotted correntropy. IEEE Signal Processing Letters, 18 Nr. 6 371–374

Islam, M. S. (2011): Testing periodicity in short series and application to gene expression data. Communications in Statistics - Simulation and Computation, 40 Nr. 4 561–573

Ismailov, N. Z. und **Adygezalzade, A. N. (2012):** Light curve analysis for RY Tau. Astronomy Reports, 56 Nr. 2 131–137

Israel, G. L. und **Stella, L. (1996):** A new technique for the detection of periodic signals in"coloured"power spectra. The Astrophysical Journal, 468 369–379

Johnson, I. M. (2011): robustreg: Robust regression functions. R package version 0.1-3

Jurkevich, I. (1971): A method of computing periods of cyclic phenomena. Astrophysics and Space Science, 13 Nr. 1 154–167

Kasdin, N. J. (1995): Discrete simulation of colored noise and stochastic processes and $1/f^{\alpha}$ power law noise generation. Proceedings of the IEEE, 83 Nr. 5 802–827

Kirchner, J. W., **Feng, X.** und **Neal, C. (2000):** Fractal stream chemistry and its implications for contaminant transport in catchments. Nature, 403 Nr. 6769 524–527

Koenker, R. (2012): quantreg: Quantile regression. R package version 4.90

Kong, A. K. H., **Charles, P. A.**, **Kuulkers, E.** und **Kitamoto, S. (2002):** Long-term X-ray variability and state transition of GX 339-4. Monthly Notices of the Royal Astronomical Society, 329 Nr. 3 588–596

Kudryavtseva, N. A. und **Pyatunina, T. B. (2006):** A search for periodicity in the light curves of selected blazars. Astronomy Reports, 50 Nr. 1 1–11

Lafler, J. und **Kinman, T. D. (1965):** An RR Lyrae star survey with the Lick 20-INCH Astrograph II. The calculation of RR Lyrae periods by electronic computer. The Astrophysical Journal Supplement Series, 11 216–222

Laguna, P., **Moody, G. B.** und **Mark, R. G. (1998):** Power spectral density of unevenly sampled data by least-square analysis: performance and application to heart rate signals. IEEE Transactions on Biomedical Engineering, 45 Nr. 6 698–715

Leahy, D. A., **Darbro, W.**, **Elsner, R. F.**, **Weisskopf, M. C.**, **Kahn, S.**, **Sutherland, P. G.** und **Grindlay, J. E. (1983):** On searches for pulsed emission with application to four globular cluster X-ray sources: NGC 1851, 6441, 6624, and 6712. The Astrophysical Journal, 266 Nr. 1 160–170

Li, T. H. (2009): A robust spectral analyzer for one-dimensional and multi-dimensional data analysis. US Patent Application 2009/0112954 A1

Li, T. H. (2010): A nonlinear method for robust spectral analysis. IEEE Transactions on Signal Processing, 58 Nr. 5 2466–2474

Li, T. H. (2012): On robust spectral analysis by least absolute deviations. Journal of Time Series Analysis, 33 Nr. 2 298–303

Liu, Y., **Fan, J. H.**, **Wang, H. G.** und **Deng, G. G. (2011):** Methods for the quasi-periodic variability analysis in blazars. Journal of Astrophysics and Astronomy, 32 Nr. 1–2 79–86

Lyutyi, V. M. und **Oknyanskii, V. L. (1987):** Amplitude–time characteristics of the optical variability of NGC 4151 in 1906–1984. Soviet Astronomy, 31 Nr. 3 245–250

Mann, M. E. und **Lees, J. M. (1996):** Robust estimation of background noise and signal detection in climatic time series. Climatic Change, 33 Nr. 3 409–445

Marazzi, A. (2011): robeth: R functions for robust statistics. R package version 2.2

Maronna, R. A., **Martin, R. D.** und **Yohai, V. J. (2006):** Robust statistics: Theory and methods. Chichester: Wiley

Mathias, A., **Grond, F.**, **Guardans, R.**, **Seese, D.**, **Canela, M.** und **Diebner, H.H. (2004):** Algorithms for spectral analysis of irregularly sampled time series. Journal of Statistical Software, 11 Nr. 2 1–30

Mattes, A., **Witte, K.**, **Hohmann, W.** und **Lemmer, B. (1991):** PHARMFIT – a nonlinear fitting program for pharmacology. Chronobiology International, 8 Nr. 6 460–476

McDonald, J. (1986): Periodic smoothing of time series. SIAM Journal on Scientific and Statistical Computing, 7 Nr. 2 665–688

Mebane Jr., W. R. und **Sekhon, J. S. (2011):** Genetic optimization using derivatives: The rgenoud Package for R. Journal of Statistical Software, 42 Nr. 11 1–26

Milotti, E. (2006): PLNoise: A package for exact numerical simulation of power-law noises. Computer Physics Communications, 175 Nr. 3 212–225

Mojón, A., **Fernández, J. R.** und **Hermida, R. C. (1992):** Chronolab: An interactive software package for chronobiologic time series analysis written for the Macintosh computer. Chronobiology International, 9 Nr. 6 403–412

Morell, O. (2012): On nonparametric methods for robust jump-preserving smoothing and trend detection. Dissertation, TU Dortmund

Mudelsee, M., **Scholz, D.**, **Röthlisberger, R.**, **Fleitmann, D.**, **Mangini, A.** und **Wolff, E. W. (2009):** Climate spectrum estimation in the presence of timescale errors. Nonlinear Processes in Geophysics, 16 Nr. 1 43–56

Norm DIN 66261 (1985): Sinnbilder für Struktogramme nach Nassi-Shneiderman. Berlin, Wien, Zürich: Beuth

Oda, M., **Doi, K.**, **Ogawara, Y.**, **Takagishi, K.** und **Wada, M. (1976):** Short-term variability of Cyg X-1. Astrophysics and Space Science, 42 Nr. 1 223–244

Oh, H. S., **Nychka, D.**, **Brown, T.** und **Charbonneau, P. (2004):** Period analysis of variable stars by robust smoothing. Journal of the Royal Statistical Society: Series C (Applied Statistics), 53 Nr. 1 15–30

Palmer, D. M. (2009): A fast chi-squared technique for period search of irregularly sampled data. The Astrophysical Journal, 695 Nr. 1 496–502

Paltani, S., **Courvoisier, T. J. L.**, **Blecha, A.** und **Bratschi, P. (1997):** Very rapid optical variability of PKS 2155-304. Astronomy and Astrophysics, 327 Nr. 2 539–549

Parsons, A. M., **Gehrels, N.**, **Paciesas, W. S.**, **Harmon, B. A.**, **Fishman, G. J.**, **Wilson, C. A.** und **Zhang, S. N. (1998):**

Multiyear BATSE earth occultation monitoring of NGC 4151. The Astrophysical Journal, 501 Nr. 2 608–615

Pearson, R. K., **Lähdesmäki, H.**, **Huttunen, H.** und **Yli-Harja, O. (2003):** Detecting periodicity in nonideal datasets. In Proceedings of the third SIAM International Conference on Data Mining. 274–278

Petit, J. R., **Jouzel, J.**, **Raynaud, D.**, **Barkov, N. I.**, **Barnola, J.-M.**, **Basile, I.**, **Bender, M.**, **Chappellaz, J.**, **Davis, M.**, **Delaygue, G.** et al. **(1999):** Climate and atmospheric history of the past 420 000 years from the Vostok ice core, Antarctica. Nature, 399 Nr. 6735 429–436

Pojmanski, G. (1997): The All Sky Automated Survey. Acta Astronomica, 47 Nr. 4 467–481

Pojmanski, G. und **Maciejewski, G. (2004):** The All Sky Automated Survey. Catalog of variable stars. III. 12h-18h quarter of the southern hemisphere. Acta Astronomica, 54 Nr. 2 153–179

Press, W. H. (1978): Flicker noises in astronomy and elsewhere. Comments on Astrophysics, 7 Nr. 4 103–119

R Core Team (2013): R: A Language and Environment for Statistical Computing. Wien: R Foundation for Statistical Computing, 2013

Raabe, N., **Thieler, A. M.**, **Weihs, C.**, **Fried, R.**, **Rautert, C.** und **Biermann, D. (2012):** Modeling material heterogeneity by Gaussian random fields for the simulation of inhomogeneous mineral subsoil machining. In **Diene, P.** und **Lorenz, P.** (Hrsg.) SIMUL 2012: The Fourth International Conference on Advances in System Simulation. 97–102

Rathjens, J. (2012): Robuste Regression im Sinusmodell. Bachelorarbeit TU Dortmund

Rödig, C., **Burkart, T.**, **Elbracht, O.** und **Spanier, F. (2009):** Multiwavelength periodicity study of Markarian 501. Astronomy and Astrophysics, 501 Nr. 3 925–932

Reegen, P. (2007): SigSpec – I. Frequency- and phase-resolved significance in Fourier space. Astronomy and Astrophysics, 467 Nr. 3 1353–1371

Reimann, J. D. (1994): Frequency estimation using unequally-spaced astronomical data. Dissertation, University of California at Berkeley

Renson, P. (1978): Méthode de recherche des périodes des étoiles variables. Astronomy and Astrophysics, 63 Nr. 1–2 125–129

Rieger, F. M. und **Mannheim, K. (2000):** Implications of a possible 23 day periodicity for binary black hole models in Mkn 501. Astronomy and Astrophysics, 359 948–952

Ripley, R. M., **Snijders, T. A. B.** und **Preciado, P. (2012):** Manual for RSiena. Technischer Bericht vom Department of Statistics, University of Oxford

Roelant, E., **Van Aelst, S.** und **Willems, G. (2011):** FRB: Fast and Robust Bootstrap. R package version 1.7

Rojo-Álvarez, J. L., **García-Alberola, A.**, **Martínez-Ramón, M.**, **Valdes, M.**, **Figueiras-Vidal, A. R.** und **Artes-Rodriguez, A. (2002):** Support vector robust algorithms for non-parametric spectral analysis. In **Dorronsoro, J.** (Hrsg.) Artificial neural networks – ICANN 2002. Band 2415, Berlin, Heidelberg: Springer 1100–1105

Rousseeuw, P., **Croux, C.**, **Todorov, V.**, **Ruckstuhl, A.**, **Salibian-Barrera, M.**, **Verbeke, T.**, **Koller, M.** und **Maechler, M. (2012):** robustbase: Basic robust statistics. R package version 0.9-3

Rousseeuw, P. J. (1985): Multivariate estimation with high breakdown point. In **Grossmann, W.**, **Pflug, G.**, **Vincze, I.** und **Werty, W.** (Hrsg.) Mathematical Statistics and Applications. Band B, Dordrecht: Reidel 283–297

Rousseeuw, P. J. und **Croux, C. (1993):** Alternatives to the median absolute deviation.

Journal of the American Statistical Association, 88 Nr. 424 1273–1283

Rousseeuw, P. J. und **Hubert, M. (1999):** Regression depth. Journal of the American Statistical Association, 94 Nr. 446 388–402

Rousseeuw, P. J. und **Leroy, A. M. (1987):** Robust regression and outlier detection. New York: Wiley

Rousseeuw, P. J. und **Yohai, V. J. (1984):** Robust regression by means of S-estimators. In **Franke, J.**, **Härdle, W.** und **Martin, D.** (Hrsg.) Robust and nonlinear time series analysis. Berlin, New York: Springer, Lecture Notes in Statistics No. 26 256–272

Ruf, T. (1999): The Lomb-Scargle periodogram in biological rhythm research: Analysis of incomplete and unequally spaced time-series. Biological Rhythm Research, 30 Nr. 2 178–201

Ruskin, D. N., **Bergstrom, D. A.**, **Kaneoke, Y.**, **Patel, B. N.**, **Twery, M. J.** und **Walters, J. R. (1999):** Multisecond oscillations in firing rate in the basal ganglia: Robust modulation by dopamine receptor activation and anesthesia. Journal of Neurophysiology, 81 Nr. 5 2046–2055

Sachs, L. und **Hedderich, J. (2006):** Angewandte Statistik. 11. Auflage. Berlin, Heidelberg: Springer

Salibian-Barrera, M., **Willems, G.** und **Zamar, R. (2008):** The fast-τ estimator for regression. Journal of Computational and Graphical Statistics, 17 Nr. 3 659–682

Salibian-Barrera, M. und **Yohai, V. J. (2006):** A fast algorithm for S-regression estimates. Journal of Computational and Graphical Statistics, 15 Nr. 2 414–427

Scargle, J. D. (1982): Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal, 263 835–853

Schulz, M. und **Mudelsee, M. (2002):** REDFIT: Estimating red-noise spectra directly from unevenly spaced paleoclimatic time series* 1. Computers and Geosciences, 28 Nr. 3 421–426

Schwarzenberg-Czerny, A. (1989): On the advantage of using analysis of variance for period search. Monthly Notices of the Royal Astronomical Society, 241 153–165

Schwarzenberg-Czerny, A. (1998a): Period search in large datasets. Baltic Astronomy, 7 Nr. 1 43–69

Schwarzenberg-Czerny, A. (1998b): The distribution of empirical periodograms: Lomb-Scargle and PDM spectra. Monthly Notices of the Royal Astronomical Society, 301 Nr. 3 831–840

Seber, G. A. F. und **Lee, A. J. (2003):** Linear regression analysis. 2. Auflage. Hoboken, New Jersey: Wiley

Siegel, A. F. (1982): Robust regression using repeated median. Biometrika, 69 Nr. 1 242–244

Simonetti, J. H., **Cordes, J. M.** und **Heeschen, D. S. (1985):** Flicker of extragalactic radio sources at two frequencies. The Astrophysical Journal, 296 Nr. 1 46–59

Skrøvseth, S. O. und **Godtliebsen, F. (2011):** Scale space methods for analysis of type 2 diabetes patients' blood glucose values. Computational and Mathematical Methods in Medicine, ID 672039 (7 Seiten)

Someren, E. J. W. Van, **Swaab, D. F.**, **Colenda, C. C.**, **Cohen, W.**, **McCall, W. V.** und **Rosenquist, P. B. (1999):** Bright light therapy: Improved sensitivity to its effects on rest-activity rhythms in Alzheimer patients by application of nonparametric methods. Chronobiology International, 16 Nr. 4 505–518

Stellingwerf, R. F. (1978): Period determination using phase dispersion minimization. The Astrophysical Journal, 224 953–960

Sturrock, P. A. und **Scargle, J. D. (2010):** False-alarm probability in relation to oversampled power spectra, with application to superkamiokande solar neutrino data. The Astrophysical Journal, 718 Nr. 1 527–529

Szatmáry, K., **Vinkó, J.** und **Gál, J. (1994):** Application of wavelet analysis in variable star research. I. Properties of the wavelet map of simulated variable star light curves. Astronomy and Astrophysics Supplement Series, 108 377–394

Thieler, A. M. (2011): Periodicity detection in irregularly sampled light curves by robust regression and outlier detection. In **Morik, K.** und **Rhode, W.** (Hrsg.) Technical report for Collaborative Research Center SFB 876 – Graduate School. Technischer Bericht vom Sonderforschungsbereich 876, TU Dortmund, Nr. 4 154–157

Thieler, A. M. (2012): Robust and weighted regression to calculate periodograms for disturbed irregularly sampled time series. In **Morik, K.** und **Rhode, W.** (Hrsg.) Technical report for Collaborative Research Center SFB 876 – Graduate School. Technischer Bericht vom Sonderforschungsbereich 876, TU Dortmund, Nr. 2 194–197

Thieler, A. M., **Backes, M.**, **Fried, R.** und **Rhode, W. (2013):** Periodicity detection in irregularly sampled light curves by robust regression and outlier detection. Statistical Analysis and Data Mining, 6 Nr. 1 73–89

Thieler, A. M., **Fried, R.** und **Rathjens, J. (2013):** RobPer: An R package to calculate periodograms for light curves based on robust regression. Technischer Bericht vom Sonderforschungsbereich 876, TU Dortmund, Nr. 2 (eingereicht)

Timmer, J. und **König, M. (1995):** On generating power law noise. Astronomy and Astrophysics, 300 707–710

Tluczykont, M., **Bernardini, E.**, **Satalecka, K.**, **Clavero, R.**, **Shayduk, M.** und **Kalekin, O. (2010):** Long-term lightcurves from combined unified very high energy gamma-ray data. Astronomy and Astrophysics, 524 ID A48 (6 Seiten)

Tukey, J. W. (1977): Exploratory data analysis. Reading: Addison-Wesley

Uttley, P., **McHardy, I. M.** und **Papadakis, I. E. (2002):** Measuring the broad-band power spectra of active galactic nuclei with RXTE. Monthly Notices of the Royal Astronomical Society, 332 Nr. 1 231–250

Varadhan, R. und **Gilbert, P. (2009):** BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. Journal of Statistical Software, 32 Nr. 4 1–26

Vaughan, S., **Edelson, R.**, **Warwick, R. S.** und **Uttley, P. (2003):** On characterizing the variability properties of X-ray light curves from active galaxies. Monthly Notices of the Royal Astronomical Society, 345 Nr. 4 1271–1284

Venables, W. N. und **Ripley, B. D. (2002):** Modern applied statistics with S. 4. Auflage. New York: Springer

Víšek, J. Á. (2000): Regression with high breakdown point. In Proceedings of ROBUST 2000. 324–356

Wang, J., **Zamar, R.**, **Marazzi, A.**, **Yohai, V.**, **Salibian-Barrera, M.**, **Maronna, R.**, **Zivot, E.**, **Rocke, D.**, **Martin, D.**, **Maechler, M.** et al. **(2012):** robust: Insightful robust library. R package version 0.3-19

Wang, Z. (2013): cts: an R package for continuous time autoregressive models via Kalman filter. Journal of Statistical Software, 53 Nr. 5 1–19

Warner, B. D. (2006): Lightcurve photometry and analysis. New York: Springer

Water Survey of Canada (1992): Historical water levels summary: Ontario to 1990. Ottawa: Environment Canada

Webb, J. R., **Smith, A. G.**, **Leacock, R. J.**, **Fitzgibbons, G. L.**, **Gombola, P. P.** und **Shepherd, D. W. (1988):** Optical observations of 22 violently variable extragalactic sources-1968-1986. The Astronomical Journal, 95 Nr. 2 374–397

Wilms, J., **Nowak, M. A.**, **Pottschmidt, K.**, **Heindl, W. A.**, **Dove, J. B.** und **Begelman, M. C. (2001):** Discovery of recurring soft-to-hard state transitions in LMC X-3. Monthly Notices of the Royal Astronomical Society, 320 Nr. 3 327–340

Yohai, V. J. und **Zamar, R. H. (1988):** High breakdown-point estimates of regression by means of the minimization of an efficient scale. Journal of the American Statistical Association, 83 Nr. 402 406–413

Zechmeister, M. und **Kürster, M. (2009):** The generalised Lomb-Scargle periodogram. A new formalism for the floating-mean and Keplerian periodograms. Astronomy and Astrophysics, 496 Nr. 2 577–584

Zelo, I. (2013): Periodenfindung in rot verrauschten Zeitreihen. Masterarbeit (in Bearbeitung) TU Dortmund

Zhang, Z. und **Chan, S. C. (2005):** Robust adaptive Lomb periodogram for time-frequency analysis of signals with sinusoidal and transient components. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Band 4, IEEE 493–496