

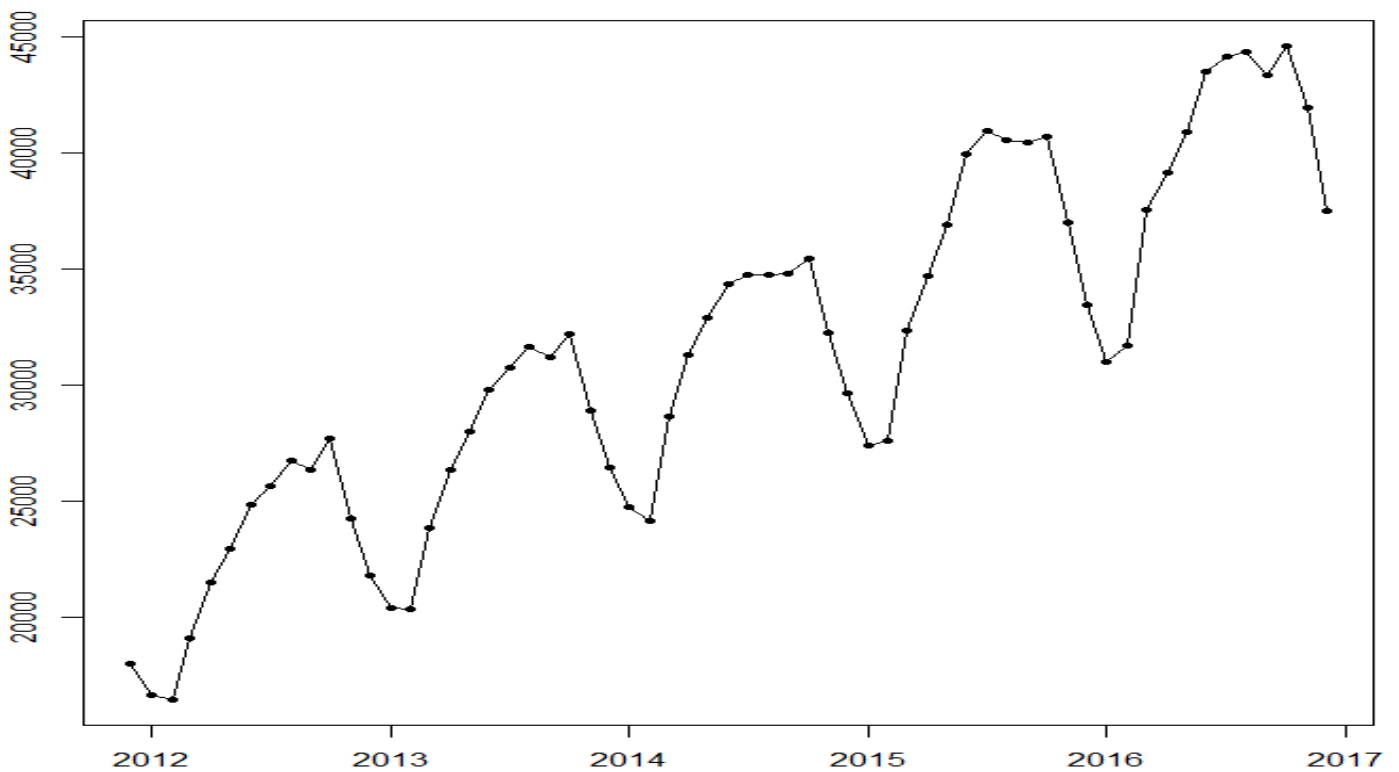
1. Saisonales Muster

Wir laden die privaten Wohnbauausgaben für die USA von Dezember 2011 bis Dezember 2016 von der Website <https://fred.stlouisfed.org/> der Federal Reserve Bank of St. Louis herunter und plotten sie mit R.

Total Private Construction Spending: Residential (Millions of Dollars, Seasonally Adjusted Annual Rate)

		01.01.2012	16614	01.01.2013	20385	01.01.2014	24704	01.01.2015	27387	01.01.2016	30972
		01.02.2012	16449	01.02.2013	20306	01.02.2014	24107	01.02.2015	27591	01.02.2016	31700
		01.03.2012	19082	01.03.2013	23846	01.03.2014	28619	01.03.2015	32341	01.03.2016	37529
		01.04.2012	21488	01.04.2013	26311	01.04.2014	31289	01.04.2015	34668	01.04.2016	39144
		01.05.2012	22904	01.05.2013	27973	01.05.2014	32892	01.05.2015	36868	01.05.2016	40859
		01.06.2012	24842	01.06.2013	29792	01.06.2014	34307	01.06.2015	39941	01.06.2016	43472
		01.07.2012	25621	01.07.2013	30751	01.07.2014	34722	01.07.2015	40951	01.07.2016	44136
		01.08.2012	26736	01.08.2013	31612	01.08.2014	34728	01.08.2015	40543	01.08.2016	44312
		01.09.2012	26315	01.09.2013	31160	01.09.2014	34791	01.09.2015	40416	01.09.2016	43306
		01.10.2012	27657	01.10.2013	32199	01.10.2014	35441	01.10.2015	40666	01.10.2016	44566
		01.11.2012	24237	01.11.2013	28902	01.11.2014	32246	01.11.2015	36974	01.11.2016	41923
01.12.2011	17960	01.12.2012	21785	01.12.2013	26403	01.12.2014	29615	01.12.2015	33424	01.12.2016	37496

```
X <- read.csv("PRRESCON.csv",header=TRUE) # first column: dates, second column: values  
D <- as.Date(X[,1]); Y <- X[,2]; plot(D,Y,type="o")
```



Abgesehen von einem in diesem Zeitraum annähernd linear ansteigenden Trend erkennt man ein ausgeprägtes saisonales Muster mit Tiefstwerten in den kalten Monaten Dezember, Jänner und Februar.

Im Folgenden seien Y_1, \dots, Y_{61} die Wohnbauausgaben von Dezember 2011 bis Dezember 2016. Für die theoretische Analyse macht es keinen Unterschied, ob die betrachtete Zeitreihe aus täglichen, wöchentlichen, monatlichen oder jährlichen Beobachtungen besteht. Wir interpretieren Y_1 als die Beobachtung zum Zeitpunkt 1 oder in Periode 1, Y_2 als die Beobachtung zum Zeitpunkt 2 oder in Periode 2 usw., egal um welche Zeiteinheit es sich handelt. Je nach Fragestellung werden wir beim Plotten als x-Werte einmal die Kalendermonate und ein anderes Mal die Zahlen 1, 2, 3, ... verwenden.

Die relative Änderung der Wohnbauausgaben Y_t in Periode t gegenüber der Vorperiode $t - 1$ ist gegeben durch

$$r_t = \frac{y_t - y_{t-1}}{y_{t-1}}$$

oder approximativ durch

$$y_t = \log(Y_t) - \log(Y_{t-1}).$$

Für die Änderung von Dezember 2011 auf Jänner 2012 erhält man

$$q_t = \frac{16614 - 17960}{17960} \approx -0.07494432$$

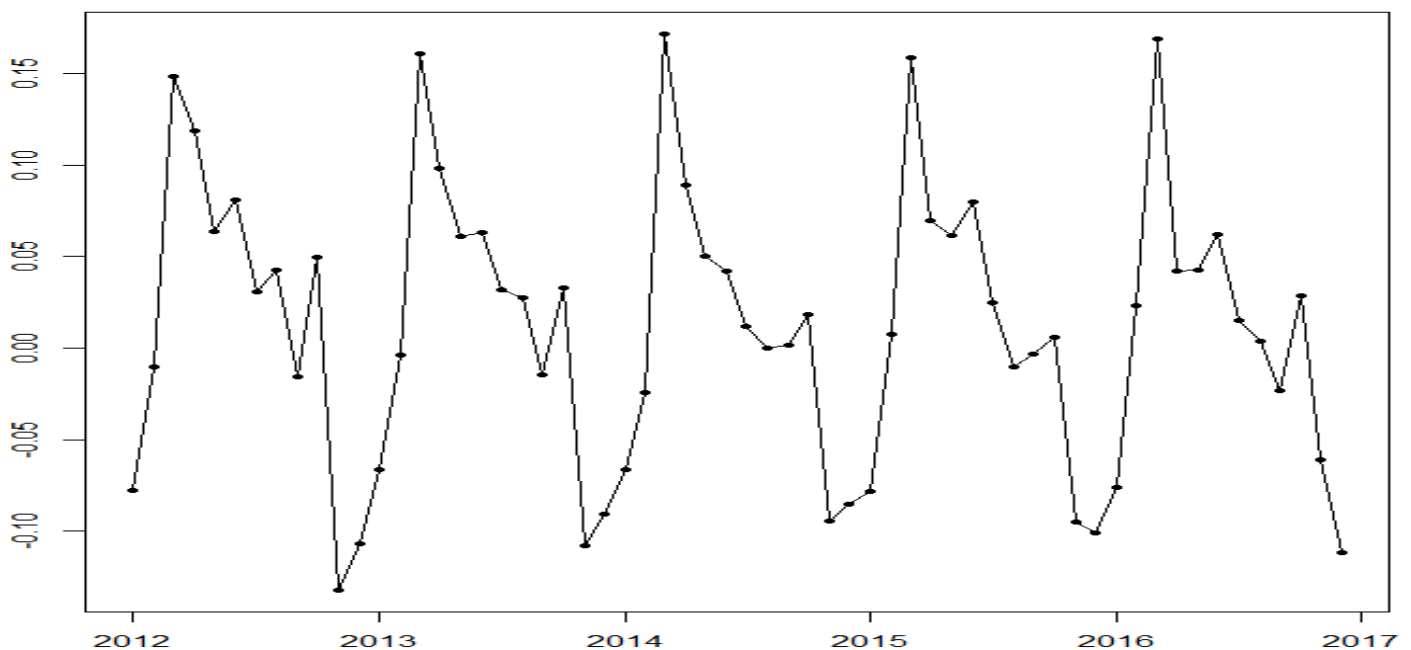
bzw.

$$y_t = \log(16614) - \log(17960) \approx -0.07790135.$$

Die Wohnbauausgaben sind also um mehr als 7% gesunken.

Wenig überraschend erkennt man auch in den (approximativen) relativen Änderungen y_1, \dots, y_{60} ein ausgeprägtes saisonales Muster mit Höchstwerten im März (wegen des steilen Anstiegs der Wohnbauausgaben von Februar auf März).

```
N <- length(Y); n <- N-1; y <- log(Y[2:N])-log(Y[1:n]); d <- D[-1]; plot(d,y,type="o")
```



Die Zeitreihe der Änderungen ist um einen Monat kürzer als die ursprüngliche Zeitreihe, weil es für den ersten Monat (Dezember 2011) keinen Vorwert gibt.

Versucht man das saisonale Muster in den monatlichen Änderungen der Wohnbauausgaben mit einer periodischen Funktion zu beschreiben, dann müsste diese Funktion eine Periode von 12 haben. Die bekanntesten periodischen Funktionen, nämlich $\cos(t)$ und $\sin(t)$, sind daher für diesen Zweck nicht geeignet, weil sie eine Periode von 2π haben. Dieses Problem lässt sich aber leicht beheben. Wenn eine gegebene Funktion f periodisch mit Periode p ist, dann können wir leicht eine verwandte Funktion g finden, die periodisch mit einer neuen (gewünschten) Periode q ist.

Definieren wir g durch

$$g(t) := f\left(t \cdot \frac{p}{q}\right),$$

dann erhalten wir

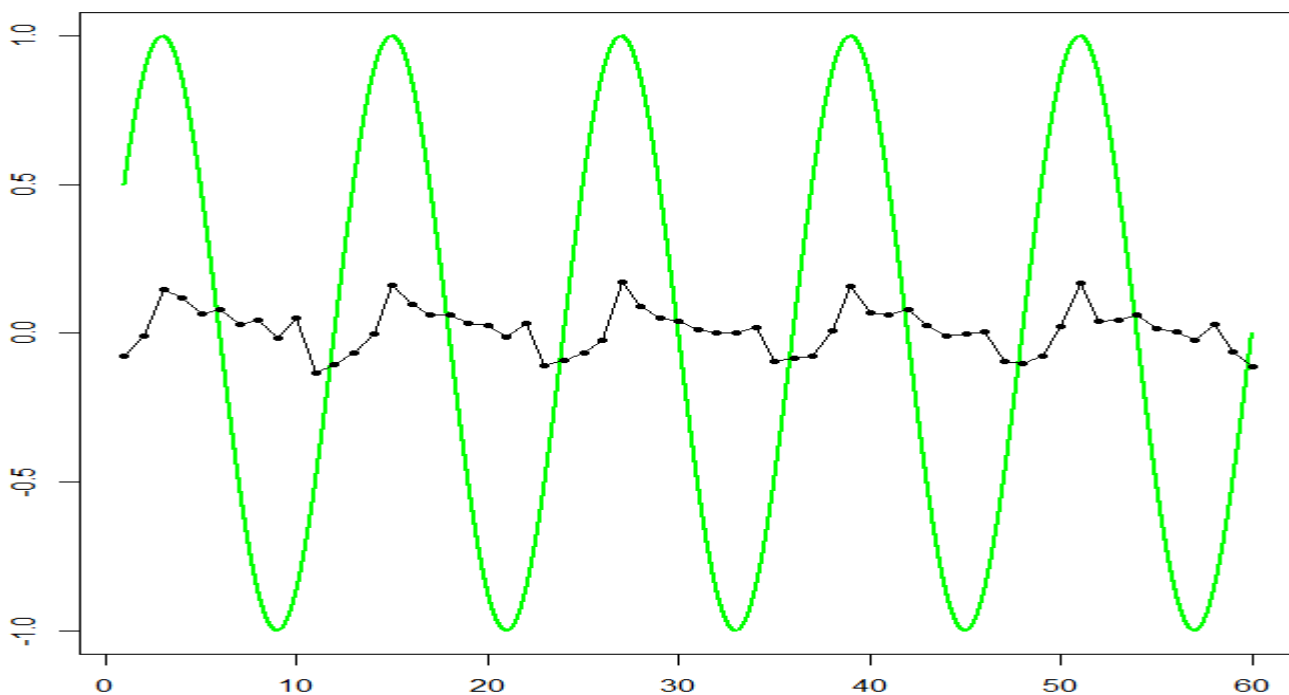
$$g(t + q) = f\left((t + q) \cdot \frac{p}{q}\right) = f\left(t \cdot \frac{p}{q} + p\right) = f\left(t \cdot \frac{p}{q}\right) = g(t).$$

Für unsere monatliche Zeitreihe könnten wir also die Funktion

$$h(t) := \sin\left(t \cdot \frac{2\pi}{12}\right)$$

versuchen. Dann stimmt zwar die Anzahl der Minima und Maxima der Funktion $h(t)$ mit der Anzahl der Hauptgipfel und Haupttäler unserer Zeitreihe überein, ihre Lagen müssen aber nicht notwendigerweise zueinander passen. Bei den Maxima ist die Übereinstimmung zufälligerweise gut, bei den Minima ist das nicht der Fall. Weiters bewegt sich die Funktion $h(t)$ in einem viel größeren Range (zwischen -1 und 1) als die Zeitreihe (zwischen -0.13 und 0.17).

```
u <- seq(1,60,0.001); plot(u,sin(u*2*pi/12),type="l",col="green",lwd=2); lines(y,type="o")
```



Die Anpassung an die gegebene Zeitreihe lässt sich weiter verbessern, indem man (i) eine Konstante ϕ zum Argument t der Funktion $g(t)$ addiert, was eine Verschiebung der ganzen Funktion nach links oder rechts ermöglicht, (ii) die Funktion mit einer positiven Konstante R multipliziert, wodurch der Range der Funktion vergrößert oder verkleinert werden kann, (iii) eine Konstante μ zur Funktion addiert, was eine Verschiebung der ganzen Funktion nach oben oder unten ermöglicht.

Eine Funktion der Form

$$g(t) = R \sin(\omega t + \phi)$$

heißt Sinusoid. Die Parameter R , ω und ϕ heißen Amplitude, Frequenz und Phase. Die Amplitude bestimmt den Range, in dem sich die Funktion bewegt, die Frequenz bestimmt, wie viele Minima und Maxima es in einer gegebenen Zeitspanne gibt, und die Phase bestimmt die Lage der Minima und Maxima.

Ein einfaches Modell zur Beschreibung einer periodischen Komponente in einer Zeitreihe y_1, \dots, y_n ist gegeben durch

$$y_t = \mu + R \sin(\omega t + \phi) + u_t.$$

Im Fall einer monatlichen Zeitreihe mit einer saisonalen Komponente ist klar, dass die Frequenz $\omega = 2\pi/12 = \pi/6$ gewählt werden muss. Die restlichen Parameter μ , R und ϕ müssen so gewählt werden, dass eine möglichst gute Anpassung an die gegebene Zeitreihe erreicht wird. Zu diesem Zweck formen wir das Modell mit Hilfe des Additionstheorems

$$\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)$$

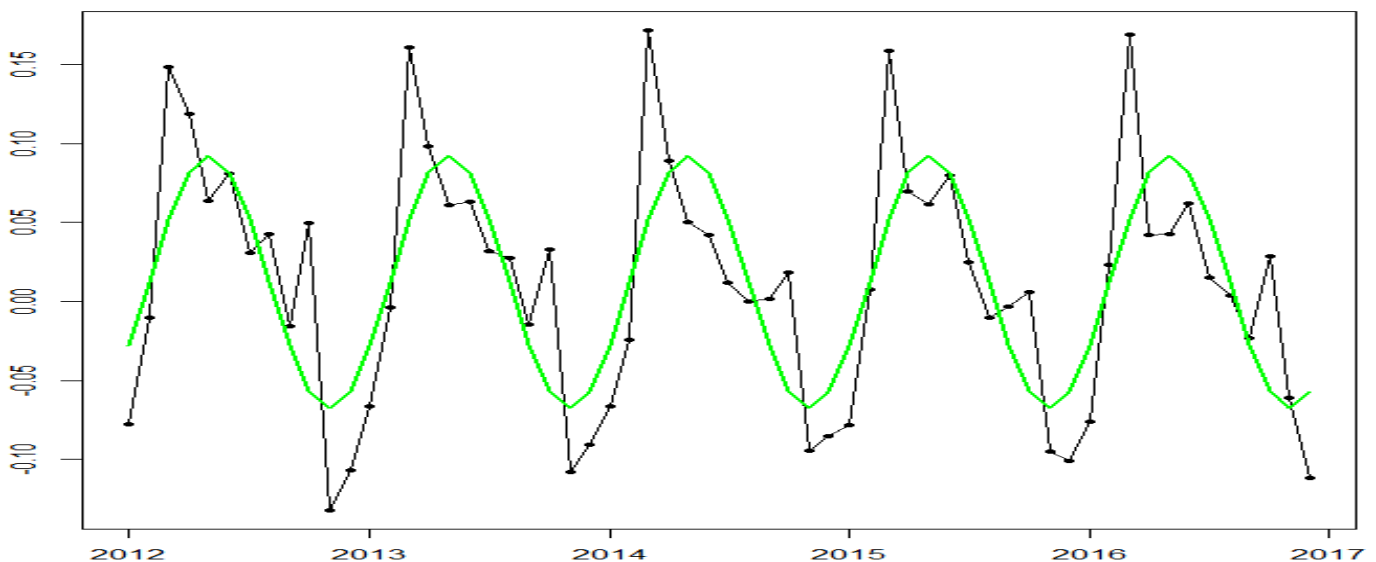
um und erhalten das lineare Regressionsmodell

$$\begin{aligned} y_t &= \mu + \underbrace{R \sin(\phi)}_A \underbrace{\cos(\omega t)}_{x_t} + \underbrace{R \cos(\phi)}_B \underbrace{\sin(\omega t)}_{z_t} + u_t. \\ &= \mu + Ax_t + Bz_t + u_t, \end{aligned}$$

dessen Regressoren x_t und z_t gegeben sind, wenn ω gegeben ist, und dessen Modellparameter μ , A und B mit der Methode der kleinsten Quadrate (least squares; LS) geschätzt werden können. Kennt man A und B , dann kann man die Amplitude R erhalten als Wurzel von

$$A^2 + B^2 = (R \sin(\phi))^2 + (R \cos(\phi))^2 = R^2(\sin^2(\phi) + \cos^2(\phi)) = R^2.$$

```
T <- 1:n; omega <- 2*pi/12; M <- lm(y~cos(omega*T)+sin(omega*T))  
plot(d,y,type="o"); lines(d,M$fitted.values,col="green",lwd=2)
```



Die Änderungen der Wohnbauausgaben steigen von einem Minimum sehr steil bis zum nächsten Maximum an und fallen dann relativ flach zum nächsten Minimum ab, weshalb das Maximum nicht in der Mitte zwischen den beiden Minima liegt. Mit einem einzelnen Sinusoid lässt sich eine solche Asymmetrie unmöglich beschreiben, sehr wohl aber mit einer Linearkombination von Sinusoiden mit verschiedenen Perioden bzw. Frequenzen.

Für die Beschreibung eines saisonalen Musters in einer monatlichen Zeitreihe können wir ein Sinusoid mit der Periode 12 verwenden, weil es in einem Jahr (12 Zeiteinheiten) einen kompletten Zyklus mit einem Minimum und einem Maximum durchläuft und deshalb zu Beginn eines Jahres wieder exakt denselben Wert annimmt wie zu Beginn des Vorjahres. Die Anpassung an das saisonale Muster ist daher in jedem Jahr gleich gut. Ein analoges Argument gilt aber auch für ein Sinusoid mit der Periode 6. Es durchläuft in einem Jahr 2 vollständige Zyklen mit je einem Minimum und einem Maximum und startet dann zu Beginn des Folgejahres wieder an derselben Stelle. Sinusoide mit den Perioden 4, 3, 2.4 und 2 durchlaufen 3, 4, 5 und 6 vollständige Zyklen in einem Jahr und eignen sich daher auch zur Beschreibung des saisonalen Musters. Mit 6 Zyklen in 12 Zeiteinheiten haben wir dann allerdings auch schon die Grenze der Auflösung erreicht. Für einen Zyklus brauchen wir zumindest zwei unterschiedlich große Werte, einen kleineren Wert und einen größeren Wert. Mehr als 6 Zyklen mit jeweils einem kleinen Wert und einem großen Wert können wir daher in 12 Zeiteinheiten nicht beobachten. Die kürzeste beobachtbare Periode ist also 2. Sie entspricht einer Frequenz von $2\pi/2 = \pi$. Diese Frequenz ist die höchste beobachtbare Frequenz und wird Nyquist-Frequenz genannt.

Wenn wir alle in Frage kommenden Frequenzen, also $\lambda_1 = 2\pi/12, \lambda_2 = 2\pi/6, \lambda_3 = 2\pi/4, \lambda_4 = 2\pi/3, \lambda_5 = 2\pi/2.4, \lambda_6 = 2\pi/2$, für die Beschreibung des saisonalen Musters in den Änderungen der Wohnbauausgaben verwenden wollen, dann müssen wir das Modell

$$y_t = \mu + \sum_{k=1}^6 R_k \sin(\lambda_k t + \phi_k) + u_t$$

schreiben als lineares Regressionsmodell

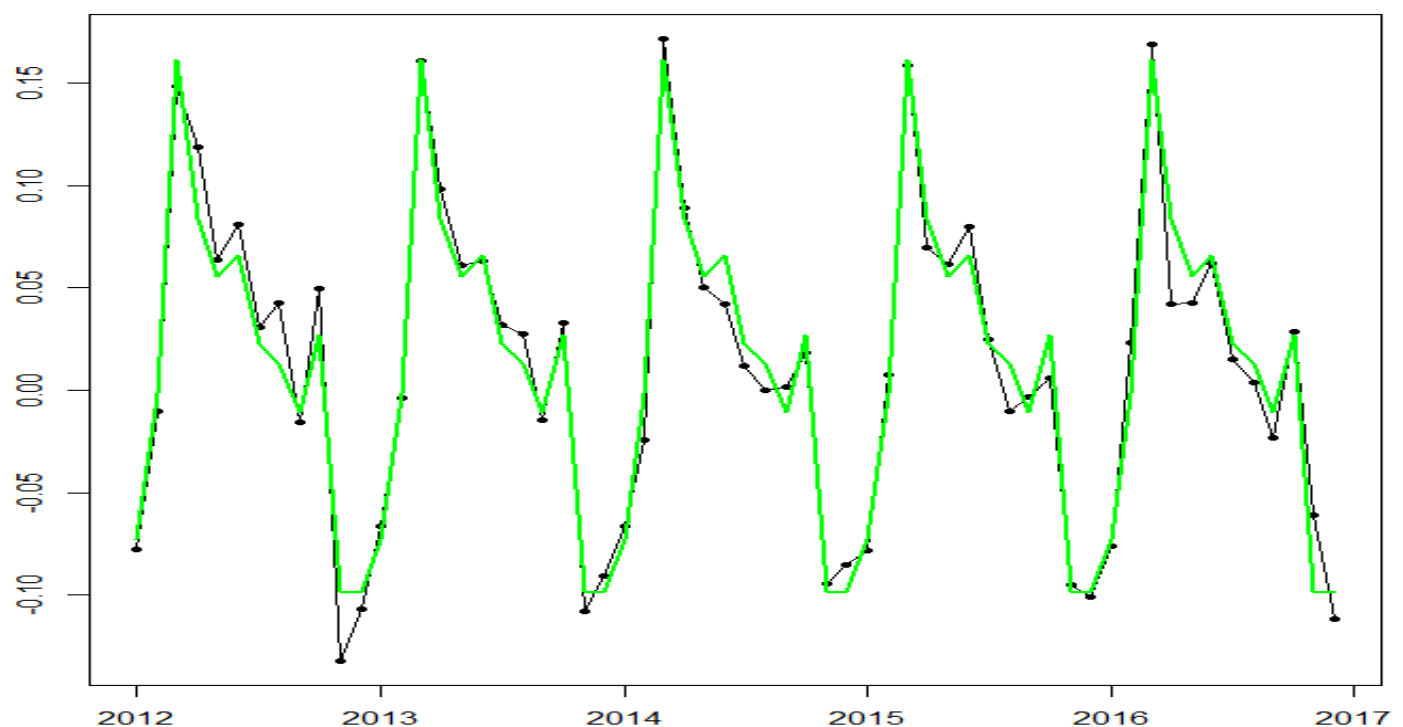
$$y_t = \mu + \sum_{k=1}^6 (A_k \cos(\lambda_k t) + B_k \sin(\lambda_k t)) + u_t$$

und die unbekannt Parameter schätzen mit der LS-Methode. Bei der praktischen Durchführung der Schätzung muss man den Regressor $\sin(\lambda_6 t)$ weglassen, da $\lambda_6 = \pi$ und $\sin(\pi t) = 0$.

```
lamda <- (1:6)*omega
```

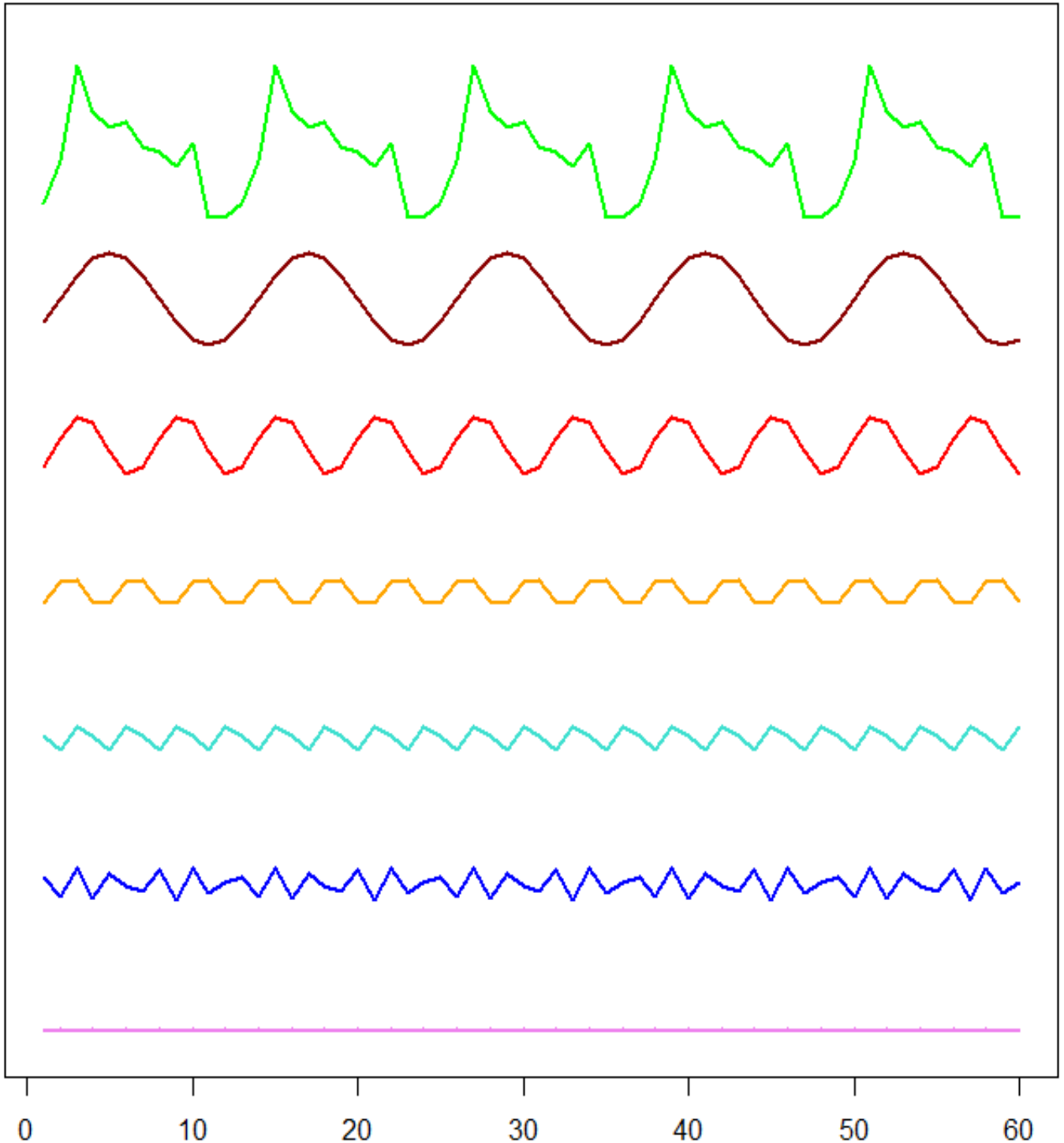
```
M6 <- lm(y~sin(lamda[1]*T)+cos(lamda[1]*T)+sin(lamda[2]*T)+cos(lamda[2]*T)
+ sin(lamda[3]*T)+cos(lamda[3]*T)+sin(lamda[4]*T)+cos(lamda[4]*T)
+ sin(lamda[5]*T)+cos(lamda[5]*T) +cos(lamda[6]*T))
```

```
plot(d,y,type="o"); lines(d,M6$fitted.values,col="green",lwd=2)
```



Die mit einer Kombination von sechs Sinusoiden erreichte Anpassung an die Daten ist sehr gut.

Ein zu unseren Monatsdaten passendes saisonales Muster (grün) haben wir durch Kombination von sechs Sinusoiden mit den Perioden 12 (dunkelrot), 6 (rot), 4 (orange), 3 (türkis), 2.4 (blau) und 2 (violett) gefunden. Wenig überraschend hat das Sinusoid mit der Periode 12 bzw. der Frequenz $\lambda_1 = 2\pi/12$ die größte Amplitude, erklärt also den größten Teil des saisonalen Musters. Die anderen Frequenzen $\lambda_2 = 2\pi/6 = 2\lambda_1$, $\lambda_3 = 2\pi/4 = 3\lambda_1$, $\lambda_4 = 2\pi/3 = 4\lambda_1$, $\lambda_5 = 2\pi/2.4 = 5\lambda_1$ und $\lambda_6 = 2\pi/2 = 6\lambda_1$ sind ganzzahlige Vielfache der Grundfrequenz λ_1 , die dazugehörigen Sinusoide werden Oberschwingungen genannt. Hat das saisonale Muster eine perfekte Sinusform, dann benötigt man nur die Grundschwingung und keine Oberschwingungen. Und umgekehrt weisen größere Amplituden der Oberschwingungen auf größere Abweichungen von einer perfekten Sinusform hin.



Wegen der Orthogonalität der Regressoren

$$\begin{aligned} & (1, \dots, 1)', \\ & (\cos(\lambda_1 1), \dots, \cos(\lambda_1 n))', (\sin(\lambda_1 1), \dots, \sin(\lambda_1 n))', \\ & \quad \vdots \\ & (\cos(\lambda_5 1), \dots, \cos(\lambda_5 n))', (\sin(\lambda_5 1), \dots, \sin(\lambda_5 n))', \\ & (\cos(\lambda_6 1), \dots, \cos(\lambda_6 n))' \end{aligned}$$

erhält man exakt die gleichen LS-Schätzwerte für die 12 Parameter $\mu, A_1, B_1, \dots, A_5, B_5, A_6$, wenn man statt des großen Modells

$$y_t = \mu + \sum_{k=1}^5 (A_k \cos(\lambda_k t) + B_k \sin(\lambda_k t)) + A_6 \cos(\lambda_6 t) + u_t$$

die 12 kleinen Modelle

$$y_t = \mu + u_t, \quad y_t = A_k \cos(\lambda_k t) + u_t, \quad k = 1, \dots, 6, \quad y_t = A_k \sin(\lambda_k t) + u_t, \quad k = 1, \dots, 5,$$

schätzt.

Voraussetzung für die Orthogonalität ist, dass nur vollständige Zyklen auftreten. In unserer Anwendung ist $n = 60$. Im Fall von $\lambda_1 = 2\pi/12$ ist die Periode gleich 12 und es gibt somit 5 vollständige Zyklen in der gesamten Beobachtungsperiode. Analog haben wir im Fall der Frequenzen $\lambda_2 = 2\pi/6, \lambda_3 = 2\pi/4, \lambda_4 = 2\pi/3, \lambda_5 = 2\pi/2.4, \lambda_6 = 2\pi/2$ die Perioden 6, 4, 3, 2.4, 2 und somit 10, 15, 20, 25, 30 vollständige Zyklen in der gesamten Beobachtungsperiode. Um die Anzahl der vollständigen Zyklen hervorzuheben, schreiben wir unsere 6 saisonalen Frequenzen auch in der Form $\lambda_1 = 5 \cdot 2\pi/n, \lambda_2 = 10 \cdot 2\pi/n, \lambda_3 = 15 \cdot 2\pi/n, \lambda_4 = 20 \cdot 2\pi/n, \lambda_5 = 25 \cdot 2\pi/n, \lambda_6 = 30 \cdot 2\pi/n$. Allgemein wird eine Frequenz der Form $k \cdot 2\pi/n$ k -te Fourier-Frequenz genannt und mit ω_k bezeichnet.

Verwenden wir alle Fourier-Frequenzen zwischen 0 und π für unsere Regression und nicht nur die saisonalen Frequenzen $\lambda_1 = \omega_5, \lambda_2 = \omega_{10}, \lambda_3 = \omega_{15}, \lambda_4 = \omega_{20}, \lambda_5 = \omega_{25}, \lambda_6 = \omega_{30}$, dann bilden die Regressoren eine Orthogonalbasis. Jeder Vektor der Länge n , insbesondere also auch unser Datenvektor $(y_1, \dots, y_n)'$, kann also als Linearkombination der Basisvektoren dargestellt werden:

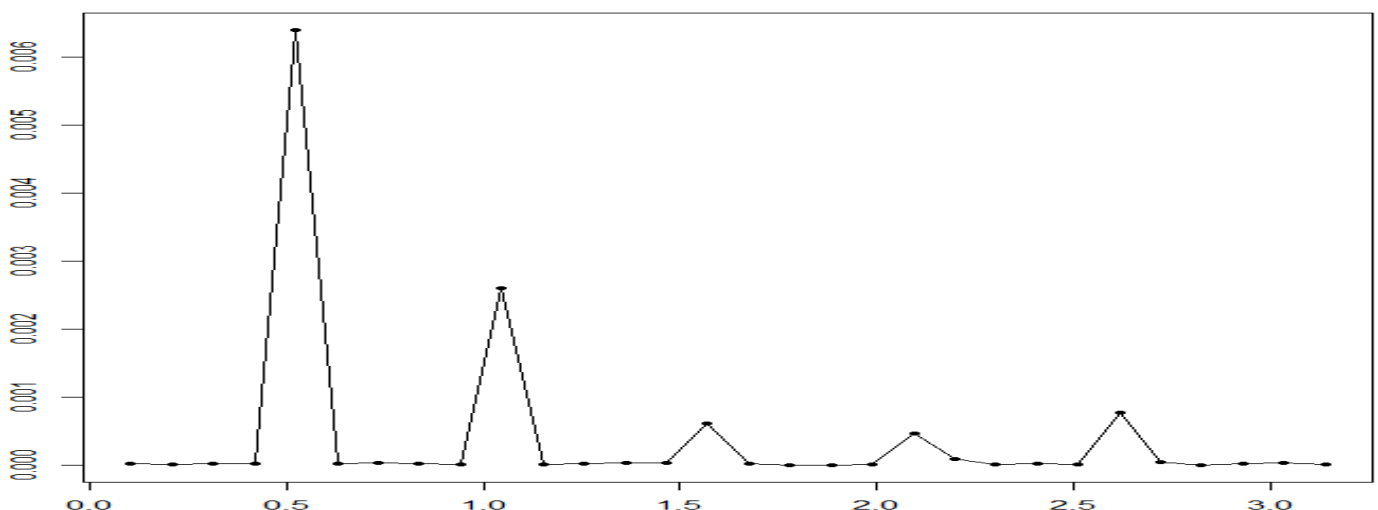
$$y_t = \mu + \sum_{k=1}^{n/2-1} (A_k \cos(\omega_k t) + B_k \sin(\omega_k t)) + A_{n/2} \cos(\omega_{n/2} t),$$

falls n gerade ist, und

$$y_t = \mu + \sum_{k=1}^{(n-1)/2} (A_k \cos(\omega_k t) + B_k \sin(\omega_k t)),$$

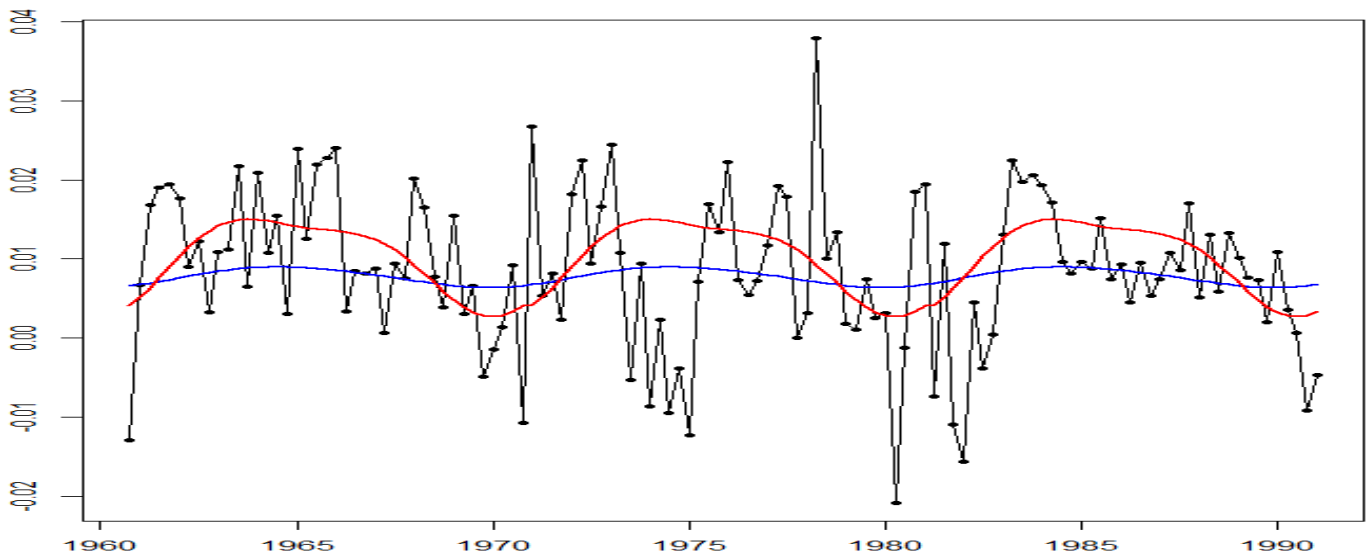
falls n ungerade ist.

Plottet man die LS-Schätzwerte $\hat{R}_k^2 = \hat{A}_k^2 + \hat{B}_k^2$ gegen die Fourier-Frequenzen ω_k , dann sieht man große Werte nur bei fünf saisonalen Frequenzen, das saisonale Muster ist also für den Großteil der Varianz in den Daten verantwortlich.



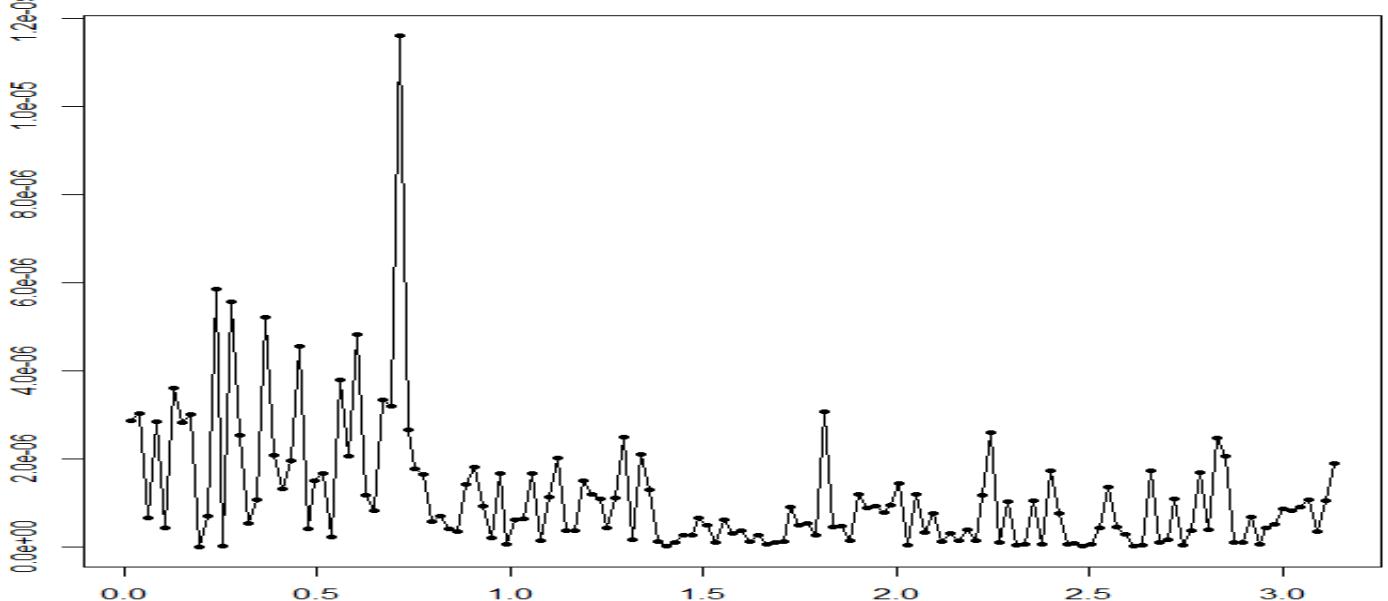
2. Das Periodogramm

Obwohl es beim saisonbereinigten amerikanischen Bruttoinlandsprodukt keine exakten Periodizitäten gibt, zeigt der Plot der (approximativen) Wachstumsraten von Oktober 1960 bis Jänner 1991 Zyklen von unterschiedlicher Stärke und Dauer.



Passt man ein Sinusoid mit der Frequenz $2\pi/40$ bzw. Periode 40 (bei Quartalsdaten sind das 10 Jahre) und eine Oberschwingung mit der doppelten Frequenz an das erste Drittel der Beobachtungsperiode an und setzt dieses Muster periodisch fort (rote Linie), dann erreicht man zwar etwas überraschend auch eine gute Anpassung im letzten Drittel, im mittleren Drittel fällt aber das Maximum der periodischen Kurve mit einem Minimum der Zeitreihe zusammen. Wächst die Beobachtungsperiode, dann ist zu erwarten, dass solche Diskrepanzen häufiger auftreten, weil es keine exakten Periodizitäten gibt. Erweitert man die Beobachtungsperiode auf den Zeitraum vom zweiten Quartal 1947 bis zum letzten Quartal 2019 und passt wiederum eine Kombination zweier Sinusoide mit den Frequenzen $2\pi/40$ und $2\pi/20$ an (blaue Linie), so erhält man in der Tat bereits eine deutlich kleinere Amplitude. In Abwesenheit von exakten Periodizitäten verschwinden alle Amplituden mit wachsendem Stichprobenumfang. Allerdings wächst dann auch die Anzahl der Fourier-Frequenzen und es ist durchaus möglich, dass die Amplituden in einem bestimmten Frequenzbereich immer viel größer sind als die in einem anderen Frequenzbereich, auch wenn alle Amplituden asymptotisch gegen null gehen.

Für die erweiterte Beobachtungsperiode ($n = 291$ Quartale) plotten wir die LS-Schätzwerte $\hat{R}_k^2 = \hat{A}_k^2 + \hat{B}_k^2$ gegen die Fourier-Frequenzen $\omega_k = 2\pi k/n$, $k = 1, \dots, [n/2] = 145$.



Das Sinusoid

$$g(t) = R_k \sin(\omega_k t + \phi_k) = \underbrace{R \sin(\phi)}_{A_k} \cos(\omega t) + \underbrace{R \cos(\phi)}_{B_k} \sin(\omega t)$$

mit der k -ten Fourier-Frequenz $\omega_k = 2\pi k/n$, der Amplitude R_k und der Phase λ_k ist periodisch mit der Periode n/k und durchläuft k vollständige Zyklen in der gesamten Beobachtungsperiode (n Zeiteinheiten).

Wegen der Orthogonalität der Regressoren $(\cos(\omega_k 1), \dots, \cos(\omega_k n))'$ und $(\sin(\omega_k 1), \dots, \sin(\omega_k n))'$ stimmen die LS-Schätzwerte für die Parameter A_k und B_k des linearen Regressionsmodells

$$y_t = A_k \cos(\omega_k t) + B_k \sin(\omega_k t) + u_t$$

mit den LS-Schätzwerten für die Parameter A_k und B_k der einfacheren Modelle

$$y_t = A_k \cos(\omega_k t) + u_t$$

und

$$y_t = B_k \sin(\omega_k t) + u_t$$

überein und sind daher gegeben durch

$$\hat{A}_k = \frac{\sum_{t=1}^n y_t \cos(\omega_k t)}{\sum_{t=1}^n \cos^2(\omega_k t)} = \frac{2}{n} \sum_{t=1}^n y_t \cos(\omega_k t)$$

und

$$\hat{B}_k = \frac{\sum_{t=1}^n y_t \sin(\omega_k t)}{\sum_{t=1}^n \sin^2(\omega_k t)} = \frac{2}{n} \sum_{t=1}^n y_t \sin(\omega_k t),$$

wobei wir verwendet haben, dass

$$\sum_{t=1}^n \cos^2(\omega_k t) = \sum_{t=1}^n \sin^2(\omega_k t) = \frac{n}{2}.$$

Letzteres gilt allerdings nicht immer. Falls n gerade ist, ist auch die Nyquist-Frequenz π eine Fourier-Frequenz, nämlich $\omega_{n/2} = 2\pi(n/2)/n = \pi$. In diesem Fall gilt wegen

$$\sum_{t=1}^n \cos^2(\pi t) = (-1)^2 + (1)^2 + (-1)^2 + (1)^2 + \dots + (-1)^2 + (1)^2 = n$$

ausnahmsweise

$$\hat{A}_{n/2} = \frac{\sum_{t=1}^n y_t \cos(\pi t)}{\sum_{t=1}^n \cos^2(\pi t)} = \frac{1}{n} \sum_{t=1}^n y_t \cos(\pi t)$$

sowie

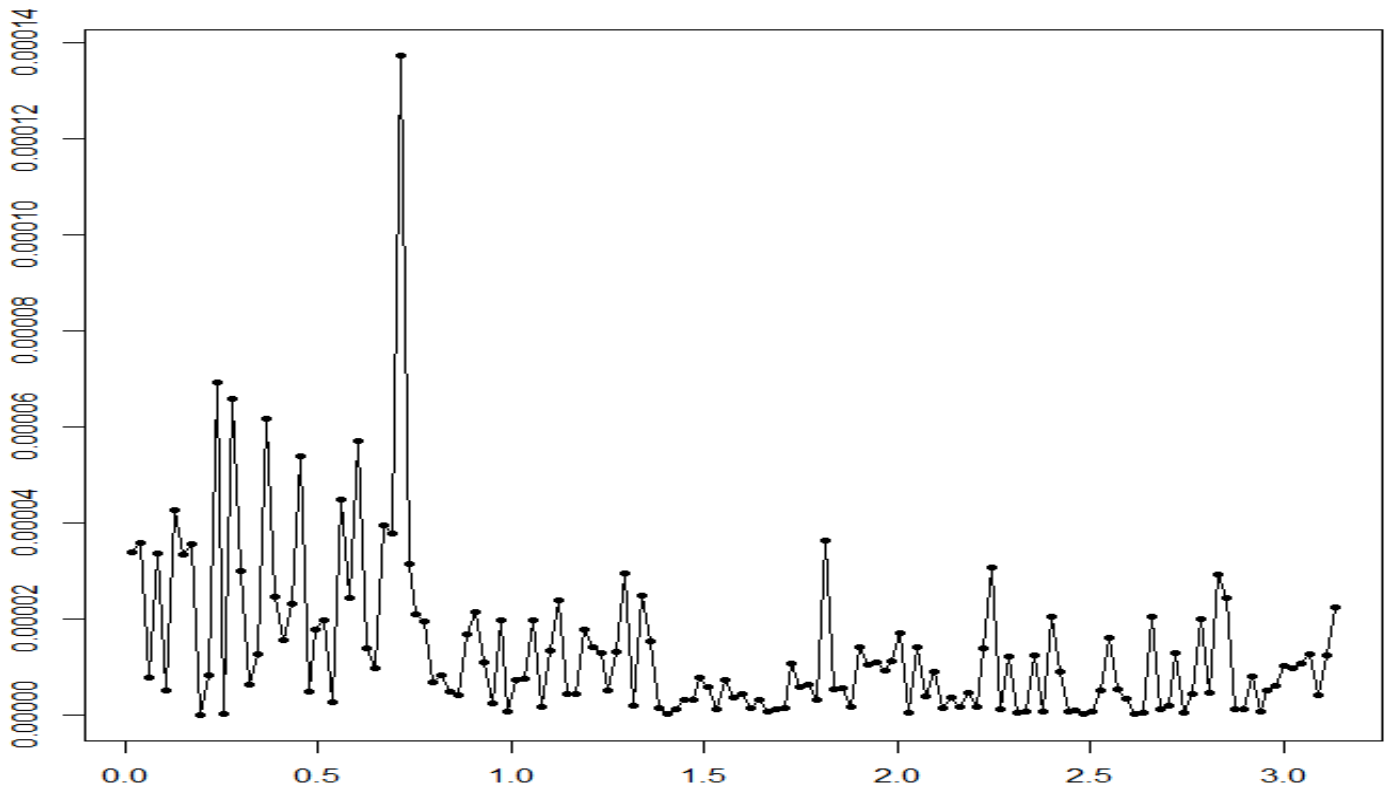
$$\hat{B}_{n/2} = 0.$$

Im Folgenden ignorieren wir allerdings diese Ausnahme und definieren das Periodogramm $I(\omega_k)$ für alle Fourier-Frequenzen durch

$$\begin{aligned} I(\omega_k) &= \frac{n}{8\pi} \left(\left(\frac{2}{n} \sum_{t=1}^n y_t \cos(\omega_k t) \right)^2 + \left(\frac{2}{n} \sum_{t=1}^n y_t \sin(\omega_k t) \right)^2 \right) \\ &= \frac{1}{2\pi n} \left(\left(\sum_{t=1}^n y_t \cos(\omega_k t) \right)^2 + \left(\sum_{t=1}^n y_t \sin(\omega_k t) \right)^2 \right) \\ &= \frac{1}{2\pi n} \left| \sum_{t=1}^n y_t \cos(\omega_k t) + i \sum_{t=1}^n y_t \sin(\omega_k t) \right|^2 \\ &= \frac{1}{2\pi n} \left| \sum_{t=1}^n y_t \exp(i\omega_k t) \right|^2. \end{aligned}$$

Das Periodogramm $I(\omega_k)$ unterscheidet sich vom Amplitudenquadrat $\hat{R}_k^2 = \hat{A}_k^2 + \hat{B}_k^2$ nur durch den konstanten Skalierungsfaktor $n/(8\pi)$. Abgesehen von der Skalierung der y-Achse und einer möglichen Diskrepanz bei der Nyquist-Frequenz π (im Fall einer geraden Anzahl von Beobachtungen) sieht ein Plot des Periodogramms also genauso aus wie ein Plot der Amplitudenquadrate.

```
h <- spec.pgram(y,taper=0,detrend=FALSE,fast=FALSE,plot=FALSE)
plot(2*pi*h$freq,h$spec/(2*pi),type="o")
```

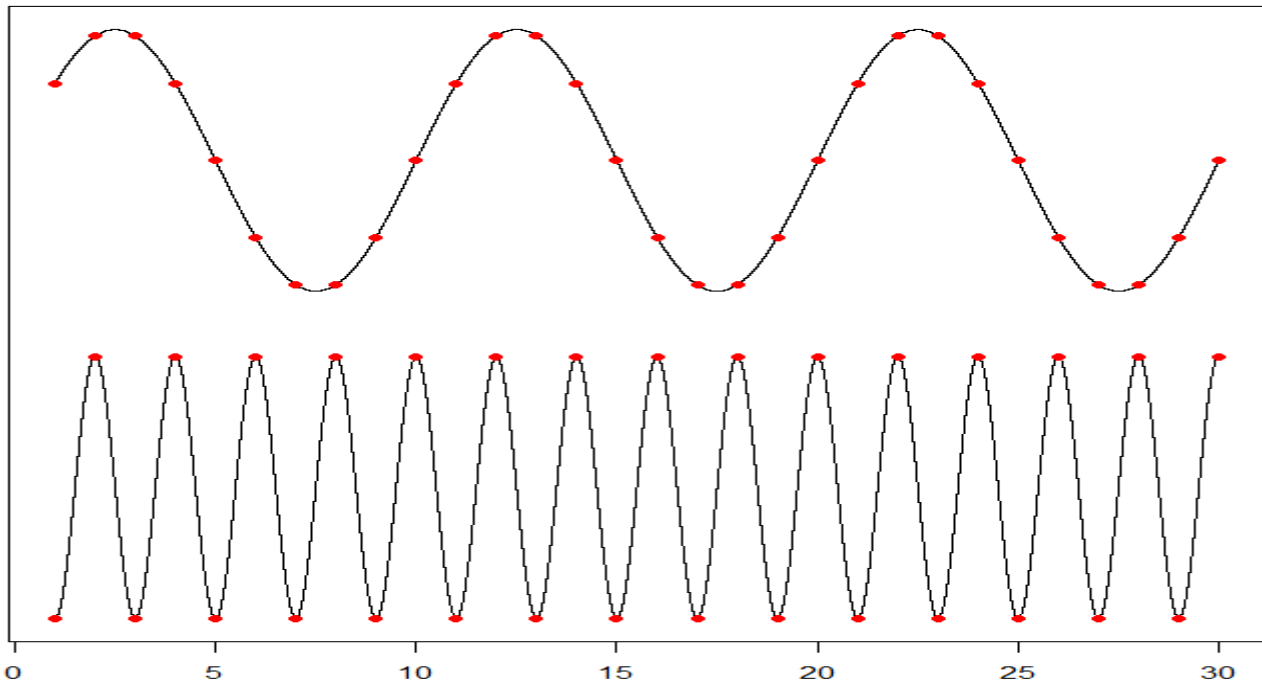


Niedrigfrequente Schwingungen bis ungefähr zur 36. Fourier-Frequenz (die einer Periode von $2\pi/\omega_{36} \approx 8.08$ Quartalen entspricht) spielen offenbar eine viel größere Rolle als höherfrequente Schwingungen. Wenn Makroökonomen von Konjunkturzyklen sprechen, meinen sie solche mehrjährigen Zyklen. Erwartungsgemäß gibt es bei unserer saisonbereinigten Zeitreihe keine Spitzen bei der saisonalen Frequenz oder bei Vielfachen dieser Frequenz. Im Fall von Quartalsdaten wäre die Periode eines saisonalen Musters gleich 4. Die dazugehörige Frequenz wäre $2\pi/4 = \pi/2$. Die einzig mögliche Oberschwingung hätte die Frequenz $2 \cdot \pi/2 = \pi$. In unserem Fall ist die Zeitreihenlänge $n = 291$ weder durch 4 noch durch 2 teilbar, sodass weder $\pi/2$ noch π Fourier-Frequenzen sind. Man könnte aber erwarten, dass etwaige saisonale Effekte in abgeschwächter Form auch bei den benachbarten Fourier-Frequenzen zu erkennen sind. Würden wir aber eine nichtsaisonbereinigte Quartalszeitreihe analysieren und nicht eine saisonbereinigte, dann sollten wir besser die ersten 3 Beobachtungen weglassen, um n auf 288 zu reduzieren.

3. Autokorrelation

Periode und Frequenz sind indirekt proportional: $\omega = 2\pi/p$, $p = 2\pi/\omega$

Niedrigfrequente Schwingungen haben eine lange Periode und hochfrequente Schwingungen eine kurze.



In dieser Abbildung folgen große Werte der niedrigfrequenten Schwingung typischerweise auf große Werte und kleine auf kleine. Im Gegensatz dazu folgen große Werte der hochfrequenten Schwingung auf kleine Werte und umgekehrt. Im Fall von Daten würde man bei der niedrigfrequenten Schwingung von einer positiven Korrelation zwischen aufeinanderfolgenden Beobachtungen sprechen und bei der hochfrequenten Schwingung von einer negativen Korrelation.

Sind n Paare $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ gegeben, dann kann man den (linearen) Zusammenhang zwischen der ersten und der zweiten Komponente messen durch die Stichprobenkovarianz

$$s_{xy} = \frac{1}{n} \sum_{t=1}^n (x_t - \bar{x})(y_t - \bar{y}).$$

Wenn wir für eine Zeitreihe y_1, y_2, \dots, y_n den (linearen) Zusammenhang zwischen der Beobachtung zum Zeitpunkt t und der Beobachtung zum nächsten Zeitpunkt $t + 1$ messen wollen, dann können wir hierfür die $n - 1$ Paare $(y_1, y_2), (y_2, y_3), \dots, (y_{n-1}, y_n)$ verwenden. Analog können wir die $n - 3$ Paare $(y_1, y_3), (y_2, y_4), \dots, (y_{n-3}, y_n)$ verwenden, wenn wir uns für den (linearen) Zusammenhang zwischen der Beobachtung zum Zeitpunkt t und der Beobachtung zum übernächsten Zeitpunkt $t + 2$ interessieren. Allgemein definieren wir die Stichproben-Autokovarianz zum Lag k durch

$$\hat{\gamma}(k) = \frac{1}{n} \sum_{t=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y}),$$

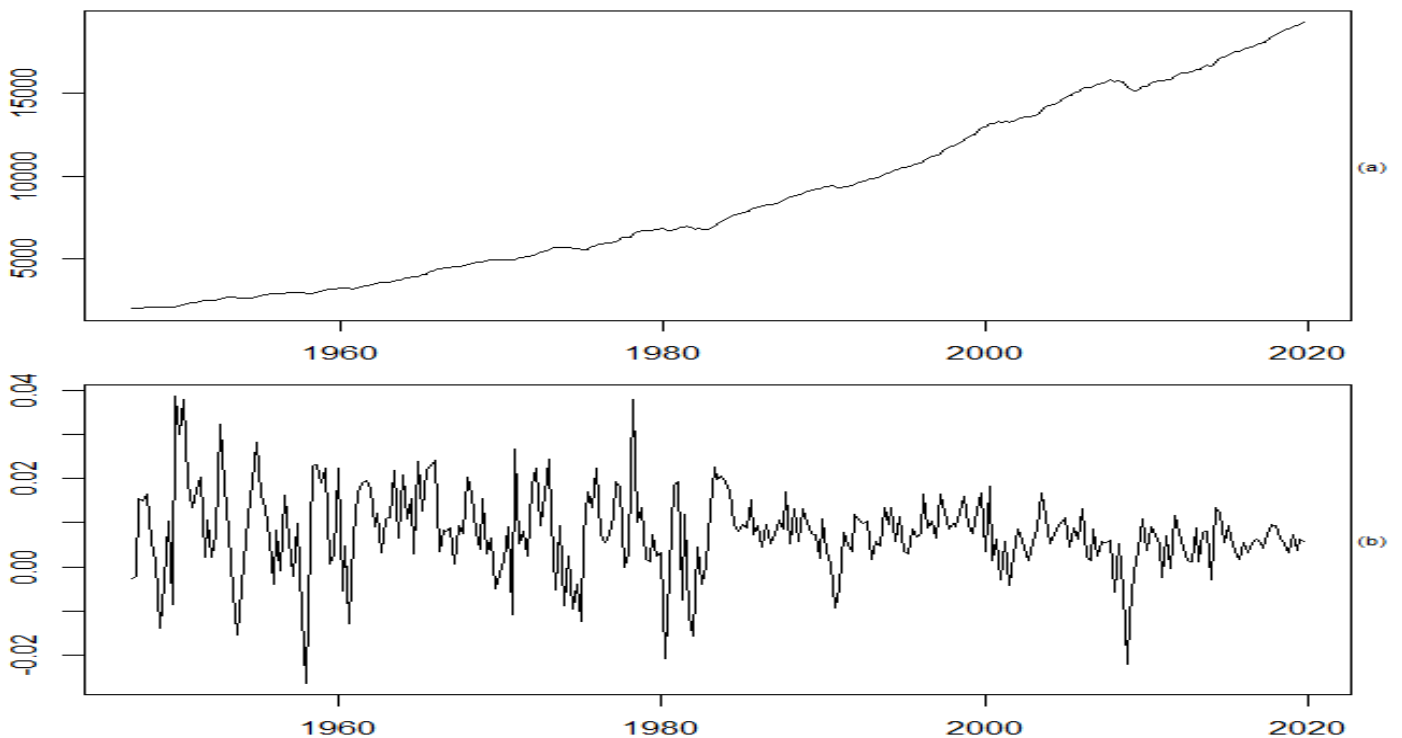
wobei \bar{y} immer von der ganzen Zeitreihe berechnet wird und unabhängig von der Anzahl der Summanden auch immer durch n dividiert wird. Während wir für $k = 0$ tatsächlich n Summanden haben und für $k = 1$ und $k = 2$ immerhin noch $n - 1$ bzw. $n - 2$, gibt es im Extremfall $k = n - 1$ nur noch einen einzigen Summanden. Die Division durch n führt mit wachsendem k zu einer immer stärkeren Schrumpfung des Schätzwerts. Dieser Effekt ist aber durchaus gewollt, weil ein auf einer kleineren Anzahl von Summanden basierender Schätzwert auch weniger vertrauenswürdig ist und deshalb ein kleineres Gewicht haben sollte.

4. Stationarität

Einen wichtigen Spezialfall der Stichproben-Autokovarianz erhält man für $k = 0$, nämlich die Stichproben-Varianz

$$\hat{\gamma}(0) = \frac{1}{n} \sum_{t=1}^n (y_t - \bar{y})^2.$$

Diese ist gegeben durch das arithmetische Mittel der quadrierten Abweichungen von \bar{y} , dem arithmetischen Mittel aller Beobachtungen. Haben alle Beobachtungen den gleichen Erwartungswert, dann ist \bar{y} ein sinnvoller Schätzer für diesen gemeinsamen Erwartungswert. Haben die Beobachtungen zusätzlich auch die gleiche Varianz, dann ist $\hat{\gamma}(0)$ ein sinnvoller Schätzer für diese gemeinsame Varianz. Im Unterschied zum klassischen Fall von Zufallsstichproben kann man im Fall von Zeitreihen aber nicht davon ausgehen, dass diese beiden Annahmen erfüllt sind. Beispielsweise wachsen das Bruttoinlandsprodukt (siehe Abbildung a) und viele andere makroökonomische Zeitreihen langfristig. Der Erwartungswert im Jahr 2020 ist also wesentlich größer als der im Jahr 1950. Man kann natürlich versuchen, einen offensichtlichen Trend zu eliminieren, beispielsweise durch Anpassung eines Polynoms an die gegebene Zeitreihe mit Hilfe der LS-Methode. Durch Subtraktion des angepassten Trends von der Zeitreihe erhält man die Trendresiduen. Auch wenn die Trendresiduen einen annähernd konstanten Erwartungswert haben, ist ihre weitere statistische Analyse sehr problematisch, da sie durch eine Reihe subjektiver Entscheidungen aus der ursprünglichen Zeitreihe hervorgegangen sind. Die Eigenschaften der Trendresiduen hängen entscheidend ab von (i) der Wahl der Modellklasse (z.B. Polynome), die zur Modellierung des Trends verwendet wird, (ii) der Wahl eines Tuning-Parameters (z.B. Grad des Polynoms), der die Glattheit des angepassten Trends bestimmt, und (iii) der Anpassungsmethode (z.B. LS-Methode), die zur Schätzung der Modellparameter (z.B. Koeffizienten des Polynoms) verwendet wird. Viel unproblematischer im Hinblick auf mögliche Inferenzstatistik ist die Verwendung der (approximativen) Wachstumsraten anstatt der Trendresiduen. Das Wachstum hat sich zwar im Laufe der Zeit etwas abgeschwächt (insbesondere nach der Ölkrise in den frühen 70er Jahren und der Immobilienkrise seit 2007), in Relation zur Schwankungsbreite sind die Unterschiede aber nicht dramatisch (siehe Abbildung b). Etwas gravierender ist der Rückgang der Volatilität seit den frühen 80er Jahren (Great Moderation). Wir gehen im Folgenden aber trotzdem davon aus, dass sowohl der Erwartungswert als auch die Varianz annähernd konstant sind. Wir nennen eine Zeitreihe stationär, wenn nicht nur $E(y_t)$ und $Var(y_t)$ konstant sind, also nicht von t abhängen, sondern darüber hinaus auch die Autokovarianzen $\gamma(k) = E(y_t - \mu)(y_{t+k} - \mu)$ nur vom Lag k und nicht vom Zeitpunkt t abhängen. Wie allgemein üblich bezeichnen wir den gemeinsamen Erwartungswert mit μ .



5. Die Spektraldichte

Methoden zur Analyse von stationären Zeitreihen basieren typischerweise auf Autokovarianzen zu verschiedenen Lags oder auf Schwingungen mit verschiedenen Frequenzen. Im ersten Fall spricht man von Zeitbereichsmethoden und im zweiten Fall von Frequenzbereichsmethoden. Beispielsweise sind die Stichproben-Autokovarianzen Zeitbereichsschätzer und das Periodogramm ist ein Frequenzbereichsschätzer. Die Stichproben-Autokovarianzen $\hat{\gamma}(k)$ sind die empirischen Gegenstücke zu den theoretischen Autokovarianzen $\gamma(k)$, aber was ist das theoretische Gegenstück zum Periodogramm? Diese Frage lässt sich leichter beantworten, wenn man das Periodogramm umschreibt als

$$\begin{aligned} I(\omega_k) &= \frac{1}{2\pi n} |\sum_{t=1}^n y_t \exp(i\omega_k t)|^2 = \frac{1}{2\pi n} |\sum_{t=1}^n y_t \exp(i\omega_k t) - \underbrace{\bar{y} \sum_{t=1}^n \exp(i\omega_k t)}_{=0}|^2 \\ &= \frac{1}{2\pi n} |\sum_{t=1}^n (y_t - \bar{y}) \exp(i\omega_k t)|^2 = \frac{1}{2\pi n} \sum_{t=1}^n (y_t - \bar{y}) \exp(i\omega_k t) \overline{\sum_{t=1}^n (y_t - \bar{y}) \exp(i\omega_k t)} \\ &= \frac{1}{2\pi n} (\sum_{t=1}^n (y_t - \bar{y})^2 \exp(i\omega_k 0) + \sum_{t=1}^{n-1} (y_t - \bar{y})(y_{t+1} - \bar{y})(\exp(i\omega_k 1) + \exp(-i\omega_k 1)) + \dots) \\ &= \frac{1}{2\pi} (\hat{\gamma}(0) + 2 \sum_{j=1}^{n-1} \hat{\gamma}(j) \cos(\omega_k j)). \end{aligned}$$

Ersetzt man in dieser Darstellung die Stichproben-Autokovarianzen $\hat{\gamma}(j)$ durch die theoretischen Autokovarianzen $\gamma(j)$ und hebt die Beschränkung durch den Stichprobenumfang n auf, so erhält man die Spektraldichte

$$f(\omega) = \frac{1}{2\pi} (\gamma(0) + 2 \sum_{j=1}^{\infty} \gamma(j) \cos(\omega j)).$$

Ein stationärer Prozess mit unkorrelierten Beobachtungen (d.h., $\gamma(j) = 0$, fall $j \neq 0$) heißt weißes Rauschen. Die Spektraldichte von weißem Rauschen ist konstant:

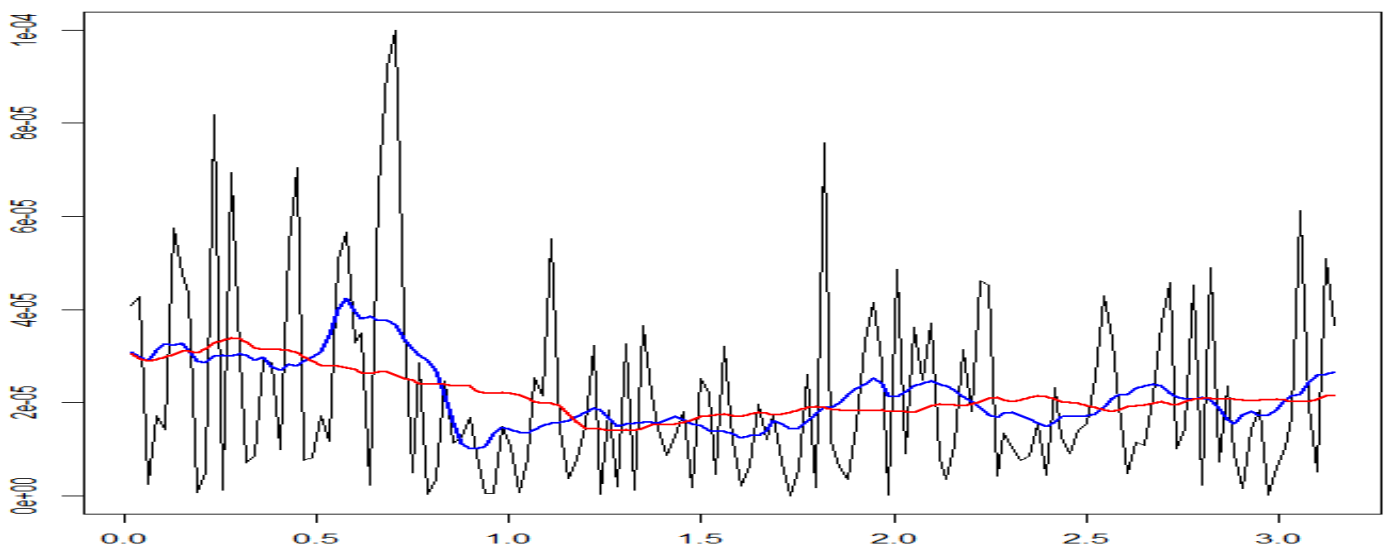
$$f(\omega) = \frac{1}{2\pi} \gamma(0)$$

Im Allgemeinen ist die Spektraldichte unbekannt und muss geschätzt werden, beispielsweise indem man über benachbarte Periodogrammordinaten mittelt:

$$\hat{f}(\omega_k) = \frac{1}{2k+1} (I(\omega_{k-m}) + \dots + I(\omega_{k-1}) + I(\omega_k) + I(\omega_{k+1}) + \dots + I(\omega_{k+m}))$$

Der Tuning Parameter m steuert die Glattheit der Schätzung. Mit wachsendem m nimmt die Glattheit zu.

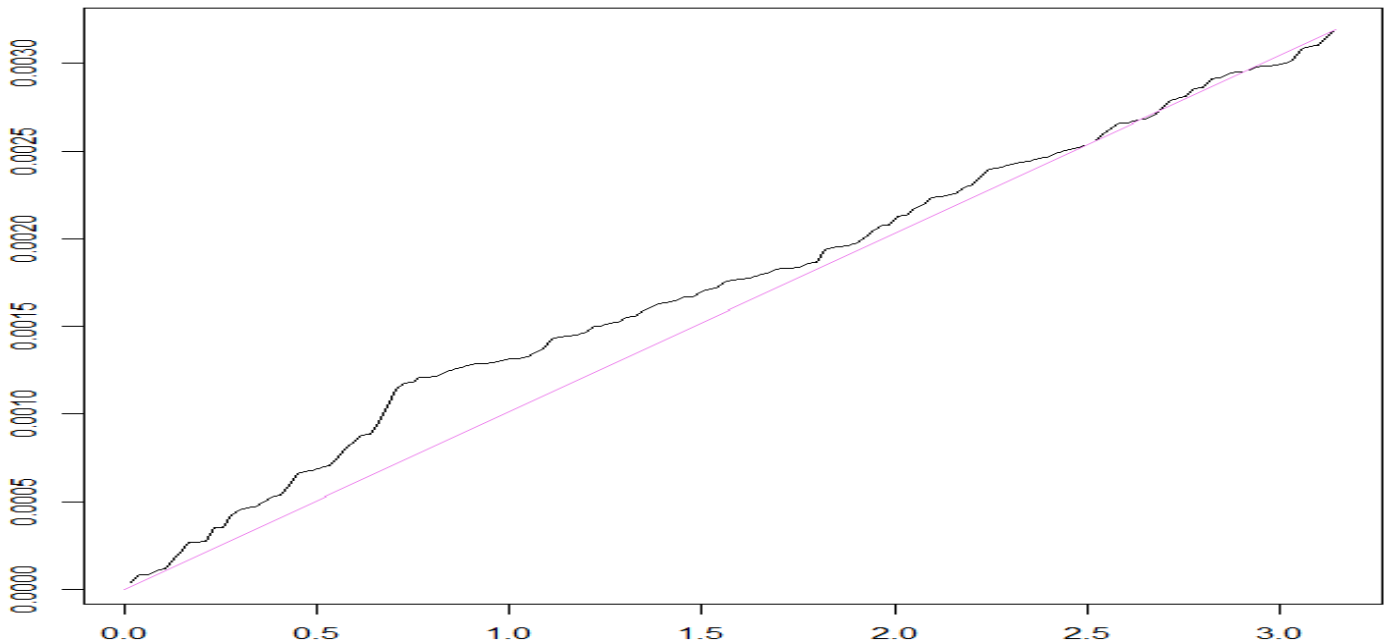
```
h <- spec.pgram(y,taper=0,detrend=FALSE,fast=FALSE,plot=FALSE)
plot(2*pi*h$freq,h$spec/(2*pi),type="l")
h <- spec.pgram(...,spans=15); lines(2*pi*h$freq,h$spec/(2*pi),col="blue",lwd=2)
h <- spec.pgram(...,spans=45); lines(2*pi*h$freq,h$spec/(2*pi),col="red",lwd=2)
```



6. Tests

Ein flaches, uncharakteristisches Periodogramm ohne ausgeprägte Gifel bzw. Täler deutet auf eine konstante Spektraldichte hin, also auf weißes Rauschen. Wegen der starken Volatilität des Periodogramms ist das aber schwierig zu beurteilen. Eine dramatische Reduktion der Volatilität erreicht man, indem man die Periodogrammordinaten kumulativ aufsummiert. Im kumulativen Periodogramm würde sich eine konstante Spektraldichte als linearer Anstieg widerspiegeln.

```
h <- spec.pgram(y,...); plot(2*pi*h$freq,cumsum(h$spec/(2*pi)),type="l")  
lines(c(0,max(2*pi*h$freq)),c(0,sum(h$spec/(2*pi))),col="violet")
```



Der Vergleich mit einer passenden Vergleichslinie (violett) zeigt, dass der Anstieg des kumulativen Periodogramms definitiv nicht linear ist. Im niedrigfrequenten Bereich ist der Anstieg deutlich steiler als bei den höheren Frequenzen. Bei den (approximativen) Wachstumsraten handelt es sich also sicher nicht um weißes Rauschen. Für eine etwas formale Überprüfung dieser Einschätzung müssten wir eine geeignete Teststatistik definieren und die Verteilung dieser Teststatistik unter der Nullhypothese von weißem Rauschen herleiten. Zeitbereichstests basieren typischerweise auf den Stichproben-Autokovarianzen und verwerfen die Nullhypothese, wenn eine oder mehrere Stichproben-Autokovarianzen sich deutlich von Null unterscheiden. Frequenzbereichstests basieren auf dem Periodogramm und verwerfen die Nullhypothese, wenn eine oder mehrere Periodogrammordinaten zu groß sind.

Unter der Nullhypothese von weißem Rauschen gilt für die LS-Schätzer $\hat{A}_k, \hat{B}_k, 0 < k < n/2$,

$$E(\hat{A}_k) = \frac{2}{n} \sum_{t=1}^n E(y_t) \cos(\omega_k t) = \frac{2}{n} \mu \underbrace{\sum_{t=1}^n \cos(\omega_k t)}_{=0},$$

$$\text{Var}(\hat{A}_k) = \frac{4}{n^2} \sum_{t=1}^n \text{Var}(y_t) \cos^2(\omega_k t) = \frac{4}{n^2} \gamma(0) \underbrace{\sum_{t=1}^n \cos^2(\omega_k t)}_{=\frac{n}{2}} = \frac{2}{n} \gamma(0),$$

$$\begin{aligned} \text{Cov}(\hat{A}_k, \hat{B}_k) &= \frac{4}{n^2} \sum_{s=1}^n \sum_{t=1}^n \text{Cov}(y_s, y_t) \sin(\omega_k s) \cos(\omega_k t) = \frac{4}{n^2} \sum_{t=1}^n \text{Var}(y_t) \sin(\omega_k t) \cos(\omega_k t) \\ &= \frac{4}{n^2} \gamma(0) \underbrace{\sum_{t=1}^n \sin(\omega_k t) \cos(\omega_k t)}_{=0} = 0 \end{aligned}$$

und analog

$$E(\hat{B}_k) = 0, \text{Var}(\hat{B}_k) = \frac{2}{n} \gamma(0)$$

sowie

$$\text{Cov}(\hat{A}_j, \hat{A}_k) = \text{Cov}(\hat{B}_j, \hat{B}_k) = 0,$$

falls $j \neq k$. Unter der zusätzlichen Annahme, dass die Beobachtungen y_1, \dots, y_n gemeinsam normalverteilt sind, sind auch die LS-Schätzer als Linearkombinationen der Beobachtungen normalverteilt. Darüber hinaus sind sie auch unabhängig, weil im Fall der Normalverteilung Unkorreliertheit gleichbedeutend ist mit Unabhängigkeit. In weiterer Folge sind dann auch die quadrierten LS-Schätzer \hat{A}_k^2, \hat{B}_k^2 unabhängig und ebenso die quadrierten standardisierten LS-Schätzer $((\hat{A}_k - 0)/(2\gamma(0)/n))^2, ((\hat{B}_k - 0)/(2\gamma(0)/n))^2$. Letztere sind als Quadrate von standardnormalverteilten Zufallsvariablen $\chi^2(1)$ -verteilt, woraus folgt, dass

$$\frac{n}{2\gamma(0)} \hat{R}_k^2 = \frac{n}{2\gamma(0)} (\hat{A}_k^2 + \hat{B}_k^2) = \frac{n}{2\gamma(0)} \frac{8\pi}{n} \frac{n}{8\pi} (\hat{A}_k^2 + \hat{B}_k^2) \underset{=I(\omega_k)}{\sim} \chi^2(2)$$

und

$$\frac{4\pi}{\gamma(0)} \sum_{k=1}^K \hat{R}_k^2 \sim \chi^2(2m).$$

Die Normalverteilungsannahme ist übrigens nicht wirklich nötig, weil die LS-Schätzer Summen sind und daher bei großem n ohnehin annähernd normalverteilt sind, auch wenn die ursprünglichen Beobachtungen nicht normalverteilt sind. Diese Robustheit gegenüber Ausreißern ist nicht nur wichtig im Fall von Aktien>Returns (wegen gelegentlicher Crashes usw.), sondern spätestens seit 2020 auch für makroökonomische Anwendungen (wegen des Corona-bedingten Wachstumseinbruchs). Allerdings steht einer praktischen Verwendung dieser Statistiken im Wege, dass die Varianz $\gamma(0)$ unbekannt ist. Wenn n hinreichend groß ist, sollte es aber keinen großen Unterschied machen, wenn wir sie durch die Stichprobenvarianz $\hat{\gamma}(0)$ ersetzen. Es sollte auch egal sein, ob wir bei der Berechnung der Stichprobenvarianz den Nenner n oder (wie in der R-function `var`) den Nenner $n - 1$ verwenden.

```
h <- spec.pgram(y,taper=0,detrend=FALSE,fast=FALSE,plot=FALSE); h <- h$spec/(2*pi)
```

```
m <- 20; sum(h[1:m])*4*pi/var(y); qchisq(0.95,2*m)
```

```
[1] 54.63176
```

```
[1] 55.75848
```

```
m <- 30; sum(h[1:m])*4*pi/var(y); qchisq(0.95,2*m)
```

```
[1] 81.61181
```

```
[1] 79.08194
```

```
m <- 40; sum(h[1:m])*4*pi/var(y); qchisq(0.95,2*m)
```

```
[1] 115.5604
```

```
[1] 101.8795
```

Wenn wir die Summe der ersten 20 Periodogrammordinaten verwenden, erhalten wir einen Wert der Teststatistik, der kleiner ist als das 95% Quantil der χ^2 -Verteilung mit $2 \cdot 20$ Freiheitsgraden. Hingegen erhalten wir ein signifikantes Testergebnis auf dem 5% Signifikanzniveau, wenn wir 30 bzw. 40 Periodogrammordinaten verwenden. Das passt zum Ergebnis unserer visuellen Analyse des kumulativen Periodogramms. Die Abweichungen von der linearen Vergleichslinie waren bei den Fourier-Frequenzen $I(\omega_{30}) \approx 0.64$ und $I(\omega_{40}) \approx 0.85$ viel größer als bei der Fourier-Frequenz $I(\omega_{20}) \approx 0.43$.