Nichtlineare Dynamik und Zeitreihenanalyse

— 1. Übung —

1 Weißes Rauschen

Die Datenreihe *whitenoise.dat* wurde in das Plot– und Analyseprogramm Grace (Version 5.1.12 vom 19. April 2003, unter Debian GNU Linux) geladen (`xmgrace whitenoise.dat`) und über die Funktion *Edit/Data sets* die wichtigsten statistischen Eigenschaften der Zeitreihe ermittelt:

$$x_{\mathrm{min}} = -4.27899 \qquad \bar{x} = 0.01054$$
$$x_{\mathrm{max}} = 3.8567 \qquad \sigma_x = 0.98507$$

Schon die Werte für Mittelwert (*mean*) $\bar{x}$ und Standardabweichung (*standard deviation*) σ_x legen eine Gaußsche Normalverteilung der Daten nahe. Das wird durch die Histogramme

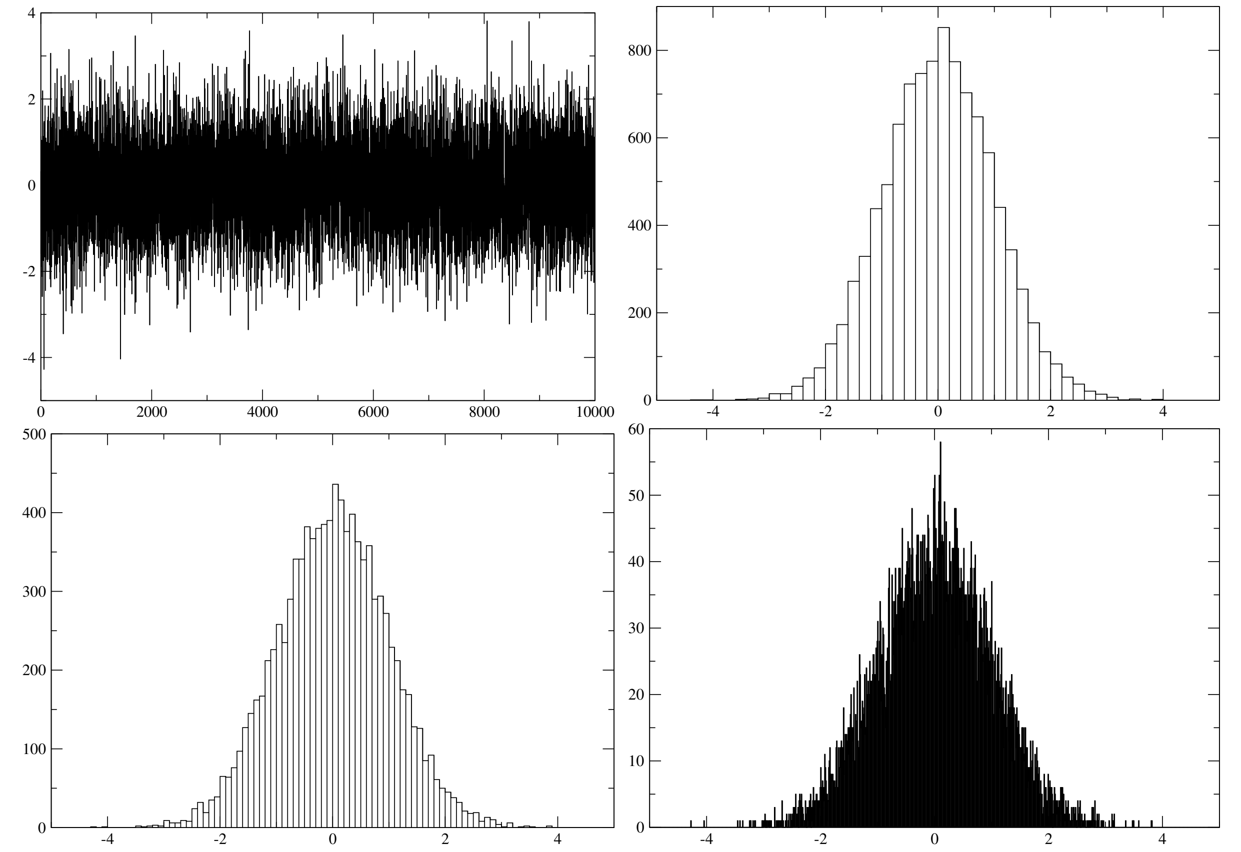

Abb. 1.1: Weißes Rauschen. Obere Zeile: Darstellung der Zeitreihe (x,t) (links) und Histogramm mit 50 bins (entsprechend einer “bin width” von 200 Samples). Untere Zeile: Histogramm mit 100 bins (bin width: 100 samples) und 1000 bins (bin width: 10). Aus allen drei Histogrammen läßt sich gut die den Daten zugrundeliegende Verteilung ablesen: Es handelt sich um eine Gaußsche Normalverteilung.

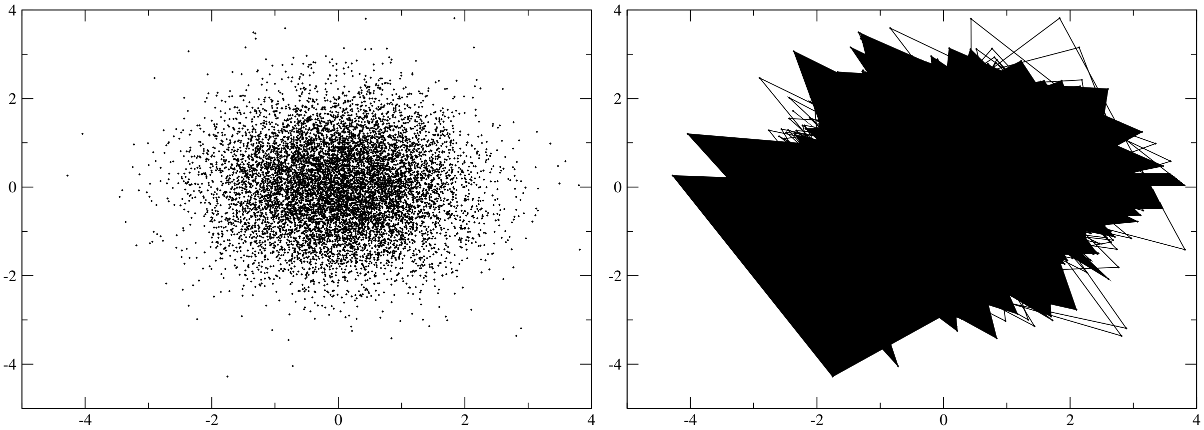

Abb. 1.2: Erste Iterierte des weißen Rauschens. Für eine Darstellung der ursprünglichen Zeitreihe vgl. Abb. 1.1. Links die reinen Datenpunkte, rechts die verbundenen Datenpunkte. Aus beiden Abbildungen geht deutlich hervor, daß aufeinanderfolgende Datenpunkte im Mitel nicht korreliert sind.

bestätigt (vgl. Abb. 1.1).

Ein weiteres Charakteristikum des weißen Rauschens und gaußsch normalverteilter Daten, die Unkorreliertheit aufeinanderfolgender Datenpunkte, wird durch die Darstellung der ersten iterierten Abbildung

$$x(t+1) = f[x(t)]$$

deutlich (Abb. 1.2).

2 Periodische Signale mit Rauschen

Die Datenreihen *sn1000.001.dat* bis *sn1000.10.dat* wurden in Grace geladen. Für eine vergleichende Darstellung der vier Zeitreihen vgl. Abb. 2.3. Alle vier Zeitreihen sind Sinuszeitreihen mit einer Frequenz von 1000 Hz und einer Abtastfrequenz von 44100 Hz mit verschieden starken Rauschanteilen. Die Standardabweichungen σ_R der gaußverteilten Zufallszahlen des Rauschanteils betrugen 0.01, 0.1, 0.5 und 1.

Die Signale wurden mittels eines sogenannten *Running average*–Filters geglättet. Die Ergebnisse der Filterung für verschiedene Fensterlängen sind für die erste Zeitreihe ($\sigma_\mathrm{R} = 0.01$) in Abb. 2.4, für die dritte Zeitreihe ($\sigma_\mathrm{R} = 0.5$) in Abb. 2.5 dargestellt.

Im jeweils linken Graphen sind vor dem Hintergrund der ursprünglichen Zeitreihe die Ergebnisse der Filterung mit einer Fensterbreite von 100 und 1000 dargestellt. Anhand von Mittelwert und Standardabweichung kann der Effekt der Fensterbreite aufgezeigt werden, hier für die dritte Zeitreihe mit einer Standardabweichung des Rauschens $\sigma_\mathrm{R} = 0.5$:

$$\begin{aligned}
\bar{x}(10, 0.5) &= 0.00436 & \sigma_x(10, 0.5) &= 0.66868\\
\bar{x}(20, 0.5) &= 0.00413 & \sigma_x(20, 0.5) &= 0.50464\\
\bar{x}(100, 0.5) &= 0.00346 & \sigma_x(100, 0.5) &= 0.08964\\
\bar{x}(1000, 0.5) &= 0.00417 & \sigma_x(1000, 0.5) &= 0.02119
\end{aligned}$$

Während der Mittelwert mit steigender Fensterbreite quasi unverändert bleibt (die Werte für $\bar{x}$ sind als nichtsignifikante Schankungen um null anzusehen), nimmt die Standardab-

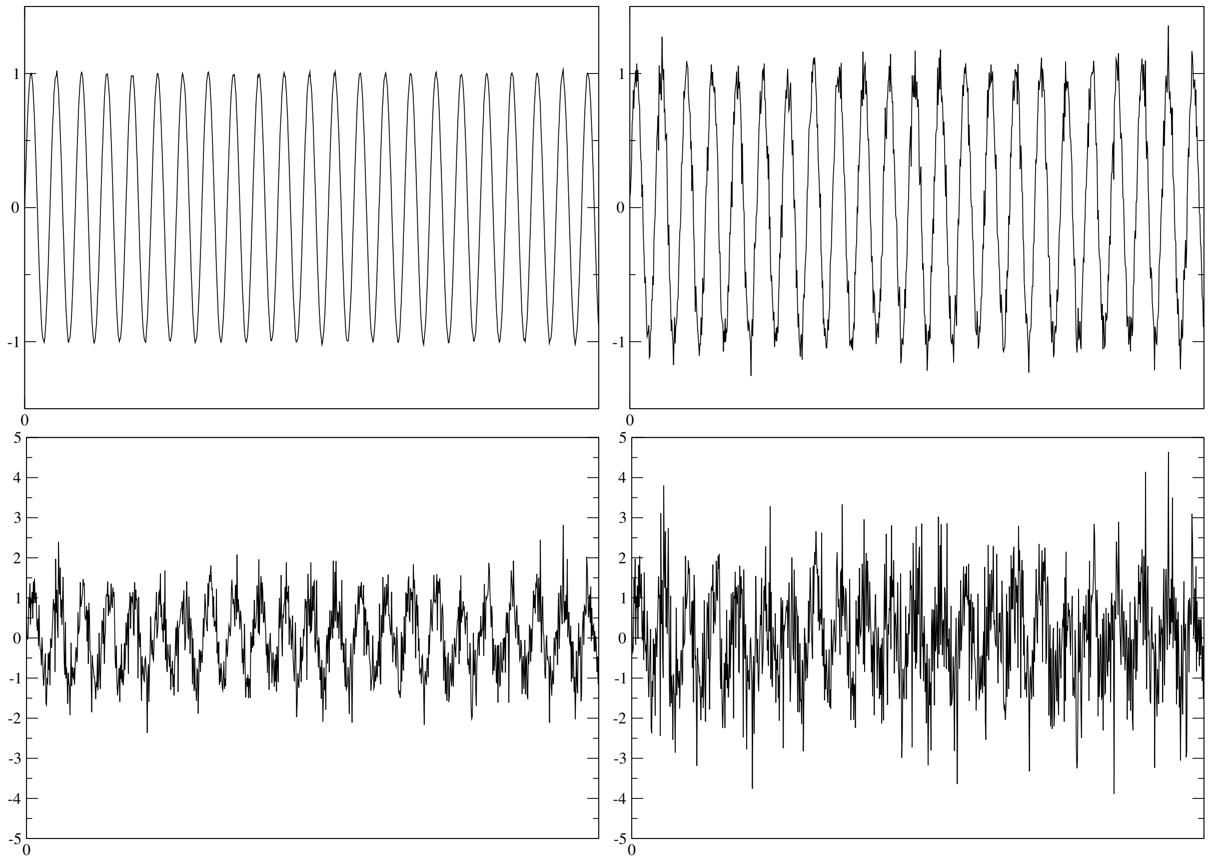

Abb. 2.3: Periodische Signale mit Rauschen. Alle vier Zeitreihen sind Sinuszeitreihen mit einer Frequenz von 1000 Hz und einer Abtastfrequenz von 44100 Hz mit verschieden starken Rauschanteilen. Der Anteil des Rauschens nimmt von links nach rechts und von oben nach unten zu. Die Standardabweichungen der gaußverteilten Zufallszahlen des Rauschanteils betrugen 0.01, 0.1, 0.5 und 1. Gezeigt ist aus Gründen der Auflösung nur das erste Zehntel der Zeitreihe.

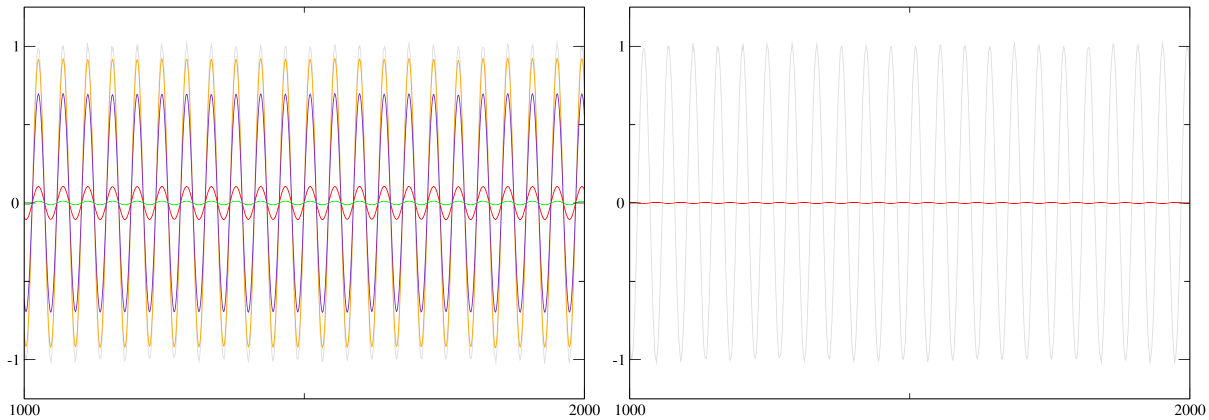

Abb. 2.4: Glättung des verrauschten periodischen Signals mittels *Running average*–Filter. Verwendet wurde die erste Zeitreihe aus Abb. 2.3 mit einer Standardabweichung des gaußschen Rauschens von 0.01. Im linken Graphen ist die Glättung mittels *Running average*–Filter der Fensterbreite 10, 20, 100 und 1000 zu sehen, im rechten Graphen der Effekt, wenn die Fensterbreite des Filters einem Vielfachen der Sinus–Periodendauer entspricht (Fensterbreite hier: 440). Im ersten Fall werden die Rauschanteile des Signals sehr gut geglättet, so daß der Sinus deutlich hervortritt, im zweiten Fall wird das Sinus–Signal selbst eliminiert. Im Hintergrund bei beiden Abbildungen das ursprüngliche Signal mit Rauschanteilen (große Amplitude). Gezeigt ist immer nur ein Ausschnitt aus der Zeitreihe.

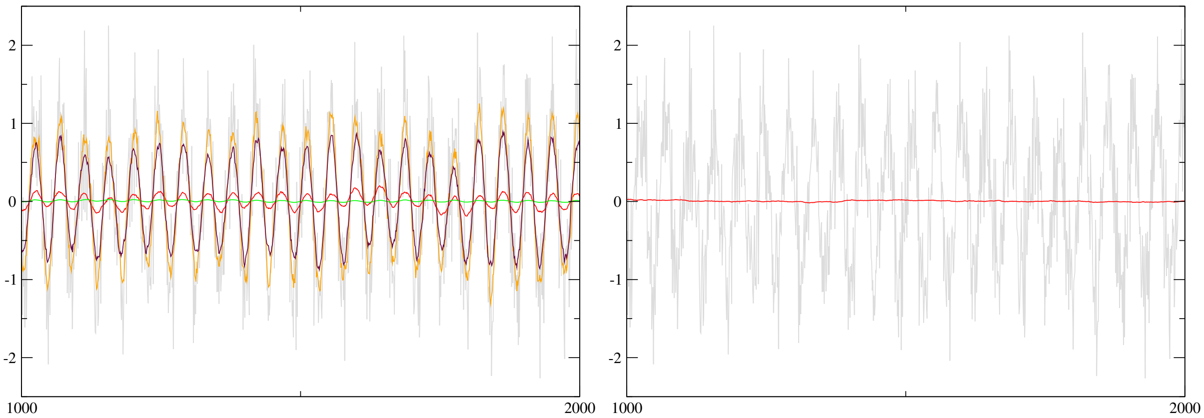

Abb. 2.5: Glättung des verrauschten periodischen Signals mittels *Running average*–Filter. Verwendet wurde die dritte Zeitreihe aus Abb. 2.3 mit einer Standardabweichung des gaußschen Rauschens von 0.5. Im linken Graphen ist die Glättung mittels *Running average*–Filter der Fensterbreite 10, 20, 100 und 1000 zu sehen, im rechten Graphen der Effekt, wenn die Fensterbreite des Filters einem Vielfachen der Sinus–Periodendauer entspricht (Fensterbreite hier: 440). Im Hintergrund bei beiden Abbildungen das ursprüngliche Signal mit Rauschanteilen (große Amplitude). Gezeigt ist immer nur ein Ausschnitt aus der Zeitreihe.

weichung ab. Ein besonders starker Abfall von σ_x ist zu verzeichnen, wenn die Fensterlänge des Filters größer ist als die Periodenlänge des Sinus–Signals.

Der rechte Graph gibt den Effekt wieder, der durch die Wahl der Fensterbreite bei einem Vielfachen der Periodendauer des Sinus–Signals entsteht. Gut zu sehen ist, daß diese Wahl der Fensterbreite des Filters das Sinus–Signal selbst "glättet" bzw. effektiv eliminiert.

Das eröffnet unterschiedliche Anwendungsmöglichkeiten von *Running average*–Filtern je nach dem gewünschten Effekt und den Anforderungen:

1. **Eliminierung von Rauschanteilen eines periodischen Signals**
   Soll ein periodisches Signal von Rauschanteilen befreit werden, sollte die Fensterbreite des *Running average*–Filters möglichst kleiner sein als die Periodenlänge des gewünschten periodischen Signals, keinesfalls jedoch ein Vielfaches derselben.

2. **Analyse der Langzeitentwicklung**
   Ist nicht die Periode des Signals von Interesse, sondern nur die Langzeitentwicklung der gesamten Zeitreihe (versteckte Trends, Perioden mit großer Periodendauer, ...), wird die Fensterbreite des *Running average*–Filters deutlich größer als die Periodendauer des periodischen Signals gewählt. Idealerweise wählt man ein Vielfaches der Periodendauer. Dadurch wird die Periode als solche eliminiert.

3 Trendelimination

Die Zeitreihe mit linearem Trend *trend.01.dat* wurde in Grace geladen und mehrmals numerisch differenziert. In Abb. 3.6 ist im linken Graph die ursprüngliche Zeitreihe mit linearem Trend und im rechten Graph zusätzlich noch die Zeitreihe nach einmaligem und zweimaligem Differenzieren zu sehen. Deutlich treten zwei Effekte des Differenzierens hervor:

1. Der Trend wird eliminiert.

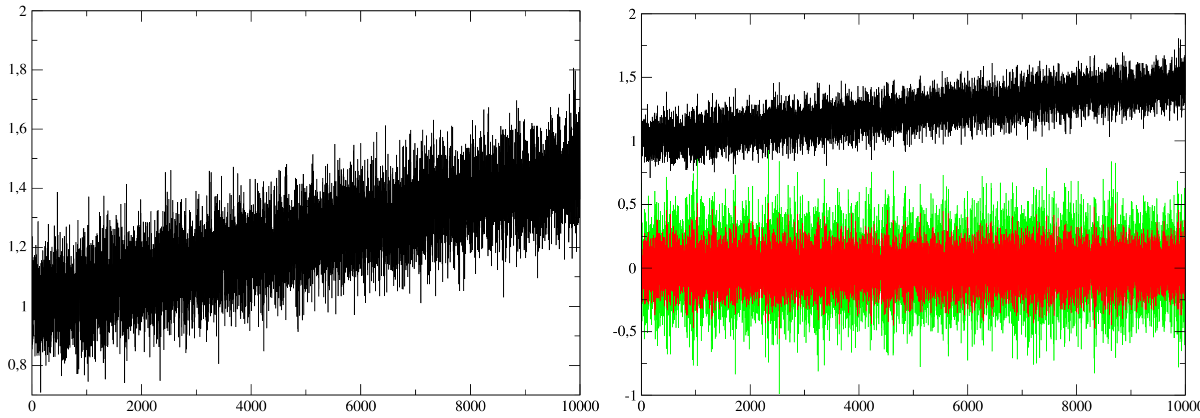

Abb. 3.6: Trendelimination einer Zeitreihe mit linearem Trend. Der linke Graph stellt die Ausgangs–Zeitreihe dar, im rechten Graph sind zusätzlich noch die erste (rot) und zweite (grün) Differenz dargestellt. Deutlich zu sehen ist, daß das Differenzieren den Trend eliminiert, aber gleichzeitig das Rauschen verstärkt.

2. Das Rauschen wird verstärkt.

Letztgenannter Effekt wird besonders deutlich sichtbar bei einer Betrachtung der Standardabweichungen:

$$
\begin{aligned}
\sigma_x &= 0.163196 & \sigma_x(D^3) &= 0.445591 \\
\sigma_x(D) &= 0.140207 & \sigma_x(D^4) &= 0.834969 \\
\sigma_x(D^2) &= 0.243570 & \sigma_x(D^5) &= 1.586320
\end{aligned}
$$

Die gegenüber der ursprünglichen Zeitreihe mit linearem Trend geringere Standardabweichung $\sigma_x(D)$ der Zeitreihe nach der ersten Differenzierung ist durch die Beseitigung des Trends bedingt. Daß Differenzieren die Standardabweichung der Zeitreihe vergrößert, effektiv also das Rauschen verstärkt, läßt sich auch mathematisch beweisen.

Beweis. Es sei z_i eine normalverteilte Zufallszahl, $z_i \in \mathbb{N}(0, \sigma)$. Dann gilt

$$
\begin{aligned}
\mathrm{Var}(z_i - z_{i-1}) &= \langle (z_i - z_{i-1})^2 \rangle - (\langle z_i \rangle - \langle z_{i-1} \rangle)^2 \\
&= \underbrace{\langle z_i^2 \rangle}_{\sigma^2} - 2 \underbrace{\langle z_i z_{i-1} \rangle}_{0} + \underbrace{\langle z_{i-1}^2 \rangle}_{\sigma^2} \\
&= 2\sigma^2
\end{aligned}
$$

□

In einem zweiten Schritt wurde die Zeitreihe mit quadratischem Trend *quadrtrend.001.dat* in Grace geladen und mittels verschiedener Regressionsverfahren (linear, quadratisch, kubisch, exponentiell) versucht, den Trend der Zeitreihe zu entfernen.

In den Abb. 3.7, 3.8, 3.9 und 3.10 sind jeweils im linken Plot die ursprüngliche Zeitreihe und die Regressionskurve dargestellt, im rechten Plot die Residuen der Regression.

Als Maß für die Güte der Regression wurden neben der Auftragung der Residuen (*residuals*) zum Test ihrer Linearität die Histogramme der Residuen aufgetragen (Abb. 3.11). Eine Übersicht über Mittelwerte $\bar{x}$ und Standardabweichungen σ_x jeweils sowohl für die Residuen als auch für die Histogramme gibt Tab. 3.1.

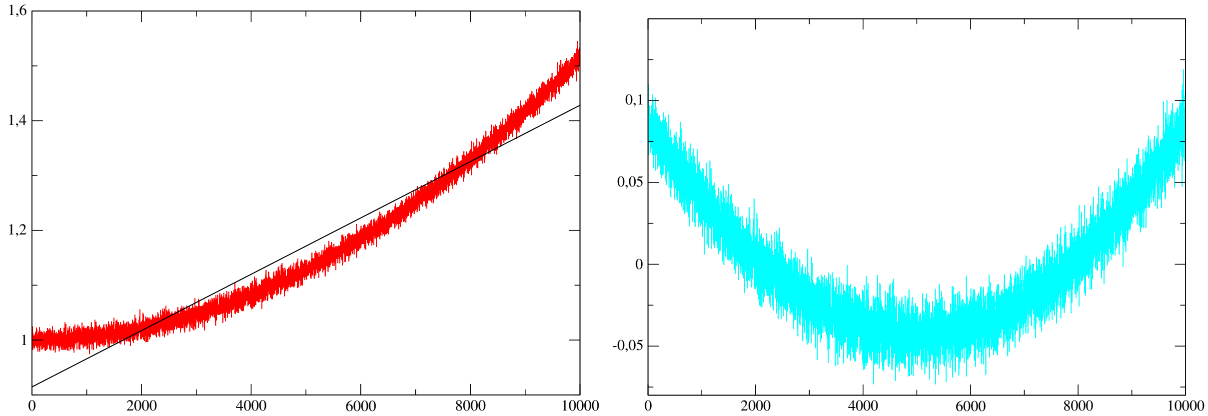

Abb. 3.7: Trendelimination einer Zeitreihe mit quadratischem Trend durch lineare Regression. Der linke Graph stellt die Ausgangs–Zeitreihe sowie die gefitteten Werte (*fitted values*, Regressionsgerade) dar, im rechten Graph sind die Residuen der Regression (*residuals*).

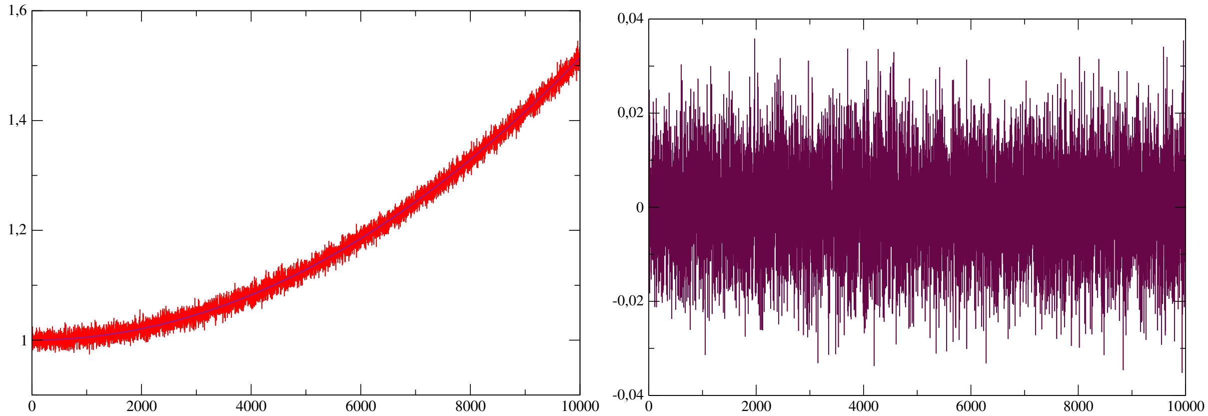

Abb. 3.8: Trendelimination einer Zeitreihe mit quadratischem Trend mittels quadratischer Regression. Der linke Graph stellt die Ausgangs–Zeitreihe sowie die gefitteten Werte (*fitted values*, Regressionskurve) dar, im rechten Graph sind die Residuen der Regression (*residuals*).

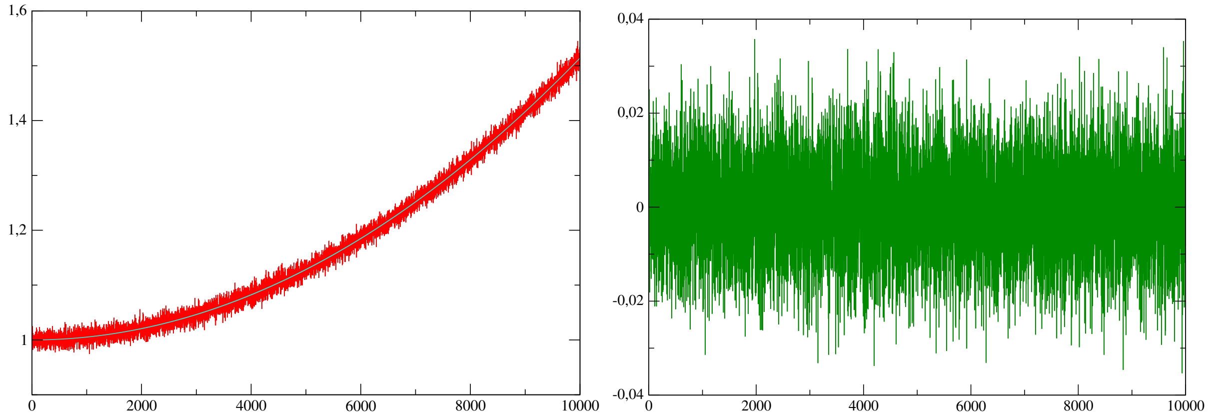

Abb. 3.9: Trendelimination einer Zeitreihe mit quadratischem Trend mittels kubischer Regression. Der linke Graph stellt die Ausgangs–Zeitreihe sowie die gefitteten Werte (*fitted values*, Regressionskurve) dar, im rechten Graph sind die Residuen der Regression (*residuals*).

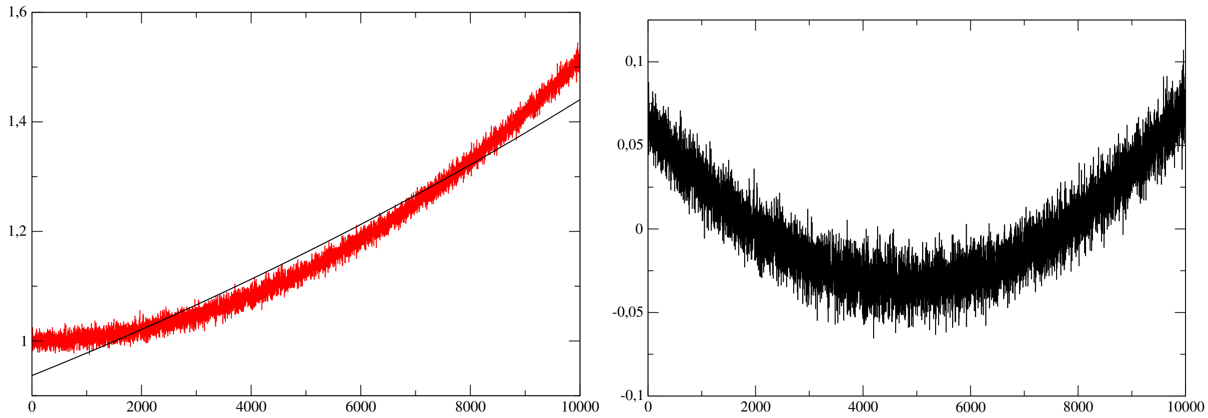

Abb. 3.10: Trendelimination einer Zeitreihe mit quadratischem Trend mittels exponentieller Regression. Der linke Graph stellt die Ausgangs–Zeitreihe sowie die gefitteten Werte (*fitted values*, Regressionskurve) dar, im rechten Graph sind die Residuen der Regression (*residuals*).

Aus allen drei Darstellungsarten (*fitted values*, *residuals*, Histogramme) geht eindeutig hervor, daß sowohl die lineare als auch die exponentielle Regression ungeeignet sind für die Trendelimination des quadratischen Trends der Zeitreihe. Die quadratische und kubische Regression hingegen können den Trend gut eliminieren.

Gleiches können wir den statistischen Eigenschaften der Residuen und Histogramme (Tab. 3.1) entnehmen: Die Mittelwerte sind jeweils nicht signifikant von Null verschieden, die Unterschiede zwischen den einzelnen Regressionsmethoden schlagen sich nur in den zweiten Momenten (hier sind statt der Varianzen σ_x^2 die Standardabweichungen σ dargestellt) nieder. Die Standardabweichungen der linearen und exponentiellen Regression sind signifikant erhöht gegenüber den entsprechenden Werten für die quadratische und kubische Regression.

Interessant ist, daß die Mittelwerte der Histogramme auch für die Residuen der linearen und exponentiellen Regression quasi null sind, obwohl die Verteilung eindeutig schief ist. Zur quantitativen Beschreibung der Schiefe können die dritten Momente der Verteilung (Schiefe, s) verwendet werden:

$$s = \frac{\langle (x_i - \langle x_i \rangle)^3 \rangle}{\sigma^3}$$

Aus den Histogrammen läßt sich für die quadratische und kubische Regression eindeutig ablesen, daß die Residuen gaußisch normalverteilt sind. Das ist ein Zeichen für die Güte der Regression.

Ein weiterer Aspekt bei Regressionen, der hier zutage tritt, ist, daß ein durch Polynome eliminierbarer Trend durch Polynome beliebiger Ordnung gefittet werden kann. Allerdings ändert sich die Güte der Regression nicht mehr, wenn die Ordnung der Polynome die des Trends übersteigt.

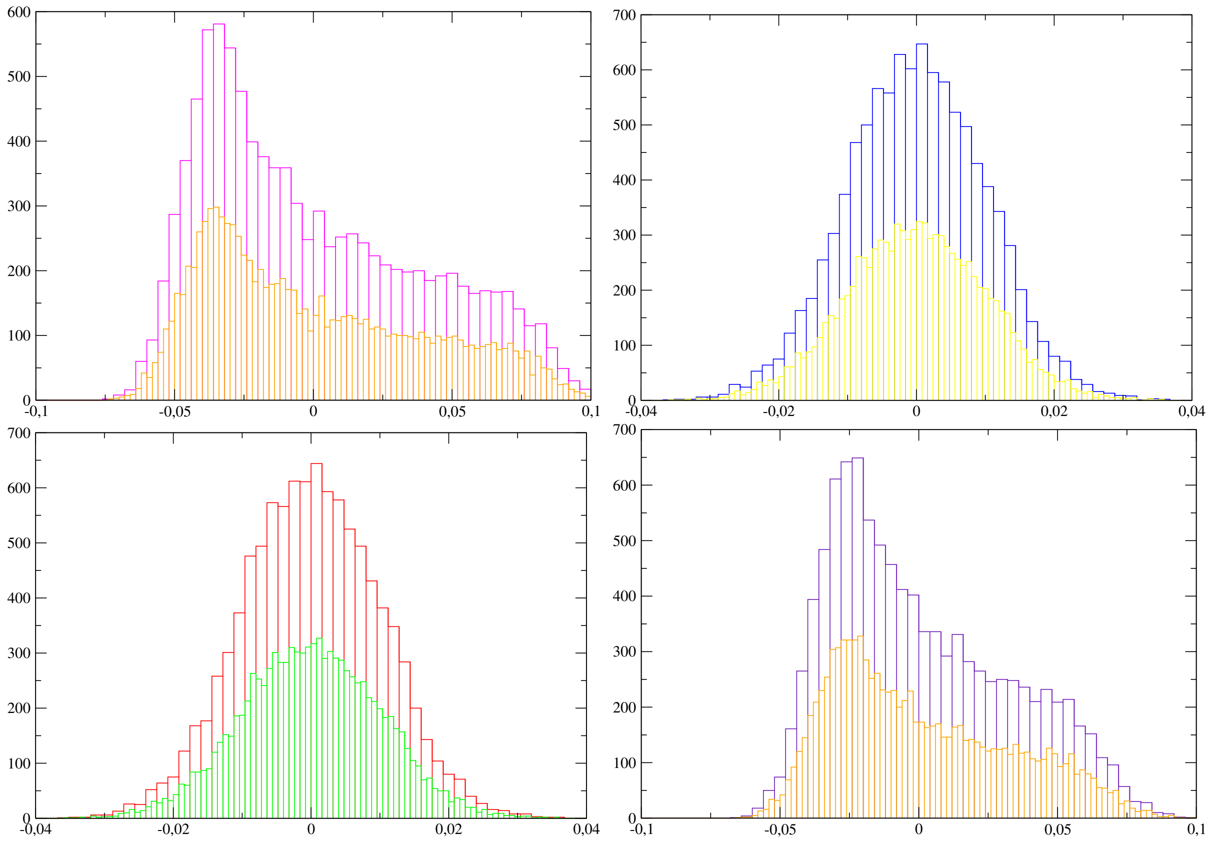

Abb. 3.11: Vergleich der Histogramme der Residuen der einzelnen Regressionsverfahren. Aufgetragen wurden jeweils Histogramme mit einer *bin size* von 100 und 200.

Tab. 3.1: Statistische Eigenschaften der Residuen und der Histogramme der Residuen der verschiedenen Regressionen. Bei den Histogrammen wurden die Werte für die Histogramme mit einer *bin size* von 200 Werten ($\triangleq$ 50 *bins*) gewählt. Der Unterschied der Histogramme mit verschiedener *bin size* schlägt sich in der Standardabweichung nieder, ist bei dieser Betrachtung aber irrelevant, da nur Histogramme mit identischer *bin size* miteinander verglichen werden.

	Residuen		Histogramme	
	$\bar{x}$	σ_x	$\bar{x}$	σ_x
linear	$-1.29937e{-12}$	0.0396067	$1.63268e{-18}$	0.0594643
quadratisch	$8.90551e{-13}$	0.0100176	$6.80284e{-19}$	0.0237857
kubisch	$2.8258e{-13}$	0.0100175	$6.80284e{-19}$	0.0237857
exponentiell	$6.41377e{-3}$	0.0320639	$1.63268e{-18}$	0.0594643

Tab. 3.2: Lineare Regression der Zeitreihe zur Trendelimination. Ausgabe des Programmes Grace. Für eine Darstellung der Zeitreihe mit der Regressionsgeraden sowie der Residuen der Regression vgl. Abb. 3.7.

```
Number of observations                  = 10000
Mean of independent variable            = 4999.5
Mean of dependent variable              = 1.171298
Standard dev. of ind. variable          = 2886.896
Standard dev. of dep. variable          = 0.1535123
Correlation coefficient                 = 0.966144
Regression coefficient (SLOPE)          = 5.137526e-05
Standard error of coefficient           = 1.372086e-07
t - value for coefficient               = 374.4319
Regression constant (INTERCEPT)         = 0.914447
Standard error of constant              = 0.0007921147
t - value for constant                  = 1154.438

Analysis of variance
Source           d.f     Sum of squares  Mean Square     F
Regression         1     219.9515        219.9515        140199.2
Residual         9998    15.68536        0.001568849
Total            9999    235.6368

y = 0.91445 + 5.1375e-05 x
```

Tab. 3.3: Quadratische Regression der Zeitreihe zur Trendelimination. Ausgabe des Programmes Grace. Für eine Darstellung der Zeitreihe mit der Regressionskurve sowie der Residuen der Regression vgl. Abb. 3.8.

```
SSerr =           1.00343

Variable order:
    0    1    2
Vector Beta      :
 0: 1
 1: -2.725e-08
 2: 5.141e-09
Vector d         :
 0: 1e+04
 1: 8.333e+10
 2: 5.556e+17
rbar matrix:
        4999.5   3.33283e+07
                        9999
Vector thetab    :
 0: 1.171
 1: 5.138e-05
 2: 5.141e-09

y = 1.0001 - 2.725e-08 x + 5.1408e-09 x^2
```

Tab. 3.4: Kubische Regression der Zeitreihe zur Trendelimination. Ausgabe des Programmes Grace. Für eine Darstellung der Zeitreihe mit der Regressionskurve sowie der Residuen der Regression vgl. Abb. 3.9.

```
SSerr =           1.00341

Variable order:
    0    1    2    3
Vector Beta      :
 0: 1
 1: 1.337e-07
 2: 5.101e-09
 3: 2.683e-15
Vector d         :
 0: 1e+04
 1: 8.333e+10
 2: 5.556e+17
 3: 3.571e+24
rbar matrix:
        4999.5   3.33283e+07    2.4995e+11
                        9999    8.9985e+07
                                   14998.5
Vector thetab    :
 0: 1.171
 1: 5.138e-05
 2: 5.141e-09
 3: 2.683e-15

y = 0.99997 + 1.337e-07 x + 5.1005e-09 x^2 + 2.6832e-15 x^3
```

Tab. 3.5: Exponentielle Regression der Zeitreihe zur Trendelimination. Ausgabe des Programmes Grace. Für eine Darstellung der Zeitreihe mit der Regressionskurve sowie der Residuen der Regression vgl. Abb. 3.10.

```
Number of observations                   = 10000
Mean of independent variable             = 4999.5
Mean of dependent variable               = 0.1498607
Standard dev. of ind. variable           = 2886.896
Standard dev. of dep. variable           = 0.127229
Correlation coefficient                  = 0.9764241
Regression coefficient (SLOPE)           = 4.303219e-05
Standard error of coefficient            = 9.514207e-08
t - value for coefficient                = 452.294
Regression constant (INTERCEPT)          = -0.06527874
Standard error of constant               = 0.0005492618
t - value for constant                   = -118.8481

Analysis of variance
Source           d.f     Sum of squares  Mean Square     F
Regression         1    154.3141        154.3141        204569.9
Residual        9998    7.541836        0.0007543345
Total           9999    161.8559

y = 0.93681 exp( 4.3032e-05 x )
```