

Damit ist eine analytische Lösung von Gl. (5.30) nicht mehr möglich, und die Schätz-
aufgabe muß mittels Suchverfahren über die Beziehung

$$\frac{\partial Q}{\partial \hat{s}} \rightarrow 0 \quad (5.32)$$

erfolgen. Das Minimum entsprechend Gl. (5.32) wird bei allen Verfahren durch die
Ermittlung des geschätzten Parametervektors in der $(i+1)$ -ten Iteration nach der
Beziehung

$$\hat{s}^{i+1} = \hat{s}^i + \Delta \hat{s}^{i+1} \quad (5.33)$$

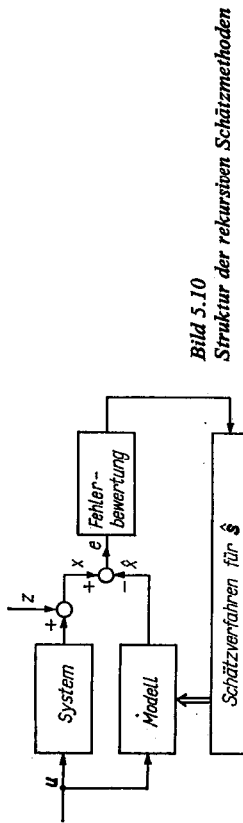
erhalten. Dabei unterscheiden sich die in Tafel 5.3 dargestellten Suchverfahren in der
Ermittlungsvorschrift für die Parameteränderung $\Delta \hat{s}^{i+1}$ zwischen der i -ten und $(i+1)$ -ten
Iteration.

5.1.5.3. Rekursive Lösung der Schätz Aufgabe

Die rekursive Lösung der Schätz Aufgabe besteht in der Ermittlung des Minimums der
Zielfunktion entsprechend Gl. (5.22) mittels Korrektur des alten Parametervektors $\hat{s}(k)$
zum k -ten Zeitpunkt durch Addition einer Parameteränderung $\Delta \hat{s}(k+1)$ nach der
Vorschrift

$$\hat{s}(k+1) = \hat{s}(k) + \Delta \hat{s}(k+1). \quad (5.34)$$

Damit ist eine laufende Modellanpassung und -nachführung leicht möglich. Das Grund-
schema der rekursiven Verfahren ist im Bild 5.10 dargestellt.



Erfolgt die Parametereinstellung kontinuierlich, so gilt folgende Parametereinstell-
vorschrift:

$$\dot{\hat{s}}(t) = -\gamma \mathbf{R}^T \nabla Q(\hat{s}, t); \quad (5.35)$$

γ Schrittweitenvektor
 $\mathbf{R}^T$ Richtungsmatrix.

Diese Strategien haben durch die breite Anwendung von digitaler Rechentechnik an
Bedeutung verloren und werden hier nur aus pädagogischen Gründen angeführt.

In Tafel 5.4 sind einige Verfahren dieser Gruppe vorgestellt, wobei der Zusammenhang
mit der iterativen Lösung sichtbar wird. Diese Gruppe von Verfahren ist besonders für
den Echtzeitbetrieb, die Verwendung von Mikrorechnern und bei adaptiven Konzepten
geeignet. Die Nachteile der direkten Lösungsmethoden, die in der Inversion und den
jedemal zu berechnenden Matrizen bei neuen Meßwerten bestehen, entfallen.

Tafel 5.4. Zusammenstellung von rekursiven Schätzverfahren

Verfahrens- gruppe	Lösungsgleichung für $\hat{s}$	Name des Verfahrens/ Literatur
Rekursive Lösung der Regression	$\hat{s}(k+1) = \hat{s}(k) + \Delta \hat{s}(k+1)$ $\hat{s}(k+1) = k(k+1) [x(k+1) - m^T(k+1) \hat{s}(k)]$	rekursive Regression [1.34, 1.38, 1.31] rekursive verallgemei- nerte Regression [1.34, 1.38, 5.20]
Gradientenverfahren - analog - diskret	$\dot{\hat{s}}(t) = -\gamma \mathbf{R}^T \nabla Q(\hat{s}, t)$ $\Delta \hat{s}(k+1) = -\gamma(k+1) \mathbf{R}(k+1) \nabla Q(k)$	Osborn [5.26] Raite [5.27] Relaxation [5.21, 5.22] steilster Abstieg [5.9, 5.11] stochastische Appro- ximation [1.41, 5.23, 5.24, 5.25]

5.2. Direkte Schätzverfahren

5.2.1. Voraussetzungen

Ausgehend von der im Abschnitt 5.1 formulierten Schätz Aufgabe ist auf der Grundlage
der zu den Zeitpunkten kT gewonnenen Meßwerte der Eingangsgrößen

$$\mathbf{m}^T(k) = [m_1(k), m_2(k), \dots, m_l(k)]$$

und der Ausgangsgröße

$$x(k)$$

sowie einer angenommenen oder a priori bekannten Struktur und einer gewählten Ziel-
funktion ein Modell in seinen Parametern und seiner Struktur so dem realen Prozeß an-
zupassen, daß die Zielfunktion minimal (maximal) wird. Für die Ableitungen in diesem
Abschnitt werden folgende Voraussetzungen gemacht:

- Die Werte der Signale $m_i(k)$ und $x(k)$ sind zeitlich unkorreliert, z. B.
 $E\{x(k)x(k+l)\} = 0; l \neq 0$.
- Die Eingangssignale $m_i(k)$ werden fehlerfrei gemessen.
- Zwischen den Eingangssignalen und der Störung besteht kein Zusammenhang, d. h.
 $E\{m_i(k)x(k)\} = 0$.
- Die Störung ist stationär mit $\mu_z = 0$ und $\sigma_z^2 = \text{const.}$
- Der Zusammenhang zwischen dem Fehler und den zu schätzenden Parametern ist
linear.

Es wird weiterhin von der Prozeßbeschreibung im Abschnitt 1.4, Gl. (1.3), ausge-
gangen und mit dem allgemeinen Eingangssignalvektor $\mathbf{m}$ gearbeitet. Diese Vorgehens-
weise wird gewählt, weil die in den folgenden Abschnitten abgeleiteten Verfahren als
Grundlage verschiedenster Schätz Aufgaben im Rahmen dieses Buches Anwendung finden.

Bei der Lösung der Schätzangabe für Modelle des statischen Systemverhaltens haben besonders Polynommodelle der Form

$$\hat{x} = \hat{a}_0 + \sum_{i=1}^l \hat{a}_i u_i + \sum_{i,j=1}^l \hat{a}_{ij} u_i u_j + \sum_{i=1}^l \hat{a}_{i1} u_i^2 + \dots \quad (5.36)$$

eine große Bedeutung erlangt, weil durch sie jeder analytische Zusammenhang durch eine geeignete Wahl der Ordnung hinreichend genau approximiert werden kann. Im einfachsten Fall hat der Eingangsgrößenvektor m die Gestalt

$$m^T = [1, u_1, \dots, u_l]. \quad (5.37)$$

In den jeweiligen Anwendungen der Verfahren im Rahmen dieses Buches wird die spezifische Bedeutung des allgemeinen Eingangssignalvektors m und des Parametervektors s angegeben.

5.2.2. Direkte Methode der Regression

Die *Methode der kleinsten Quadrate* (MKQ), die Regression, ist das älteste und zugleich robusteste Verfahren zur Schätzung der Parameter eines allgemeinen Modells der Form

$$\hat{x}(k) = m^T(k) s \quad (5.38)$$

mit

$$m^T(k) = [m_1(k), m_2(k), \dots, m_l(k)]$$

$$s^T = [s_1, s_2, \dots, s_l].$$

Für die Beschreibung des Systems gilt analog für eine Beobachtung

$$x(k) = m^T(k) s + z(k). \quad (5.39)$$

Zur Schätzung der unbekannten Parameter s sind n Beobachtungen ($n \geq l$) erforderlich, die das folgende Gleichungssystem für $k = j, j+1, \dots, j+(n-1)$ ergeben:

$$\begin{aligned} x(j) &= s_1 m_1(j) + \dots + s_l m_l(j) + z(j) \\ x(j+1) &= s_1 m_1(j+1) + \dots + s_l m_l(j+1) + z(j+1) \\ &\vdots \\ x(j+n-1) &= s_1 m_1(j+n-1) + \dots + s_l m_l(j+n-1) + z(j+n-1). \end{aligned} \quad (5.40)$$

In Vektorschreibweise kann Gl. (5.40) geschrieben werden:

$$x(n) = M(n) s + z(n) \quad (5.41)$$

mit den Komponenten

$$\begin{aligned} x^T(n) &= [x(j), x(j+1), \dots, x(j+n-1)] \\ z^T(n) &= [z(j), z(j+1), \dots, z(j+n-1)] \\ s^T &= [s_1, s_2, \dots, s_l] \end{aligned}$$

$$M(n) = \begin{bmatrix} m_1(j) & \dots & m_l(j) \\ \vdots & & \vdots \\ m_1(j+n-1) & \dots & m_l(j+n-1) \end{bmatrix}.$$

Entsprechend Gl. (5.41) gilt für die Modellgleichung

$$\hat{x}(n) = M(n) \hat{s} \quad (5.42)$$

und damit als Fehlergleichung

$$e(n) = x - M(n) \hat{s}. \quad (5.43)$$

Die Methode der kleinsten Fehlerquadrate beruht nun darauf, daß die Parameter $\hat{s}_i$ des Modells mit Hilfe des Gütekriteriums

$$Q = \|e(n)\|^2 = \sum_{k=j}^{j+n-1} e(k)^2 \quad (5.44)$$

so ermittelt werden, daß das Gütekriterium Q minimal wird (Bild 5.11). Durch Einsetzen von Gl. (5.43) in Gl. (5.44) erhält man die Beziehungen

$$\begin{aligned} Q &= (x(n) - M(n) \hat{s})^T (x(n) - M(n) \hat{s}) \\ Q &= x(n)^T x(n) - 2 \hat{s}^T M(n)^T x(n) + \hat{s}^T M(n)^T M(n) \hat{s}. \end{aligned} \quad (5.45)$$

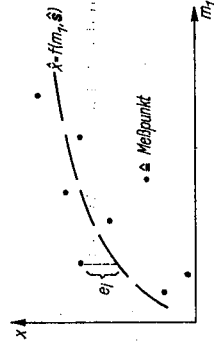


Bild 5.11
Lösung der Schätzangabe für ein System
mit einem Eingang und einem Ausgang

Das Minimum des Gütekriteriums Q wird durch Nullsetzen der ersten partiellen Ableitung

$$\frac{\partial Q}{\partial \hat{s}} = -2M(n)^T x(n) + 2M(n)^T M(n) \hat{s} = 0 \quad (5.46)$$

analytisch bestimmt.

Wenn die Inverse der Matrix $M(n)^T M(n)$ existiert, erhält man aus Gl. (5.46) als direkte analytische Lösung die Parameterschätzvorschrift der Regression zu

$$\hat{s} = [M(n)^T M(n)]^{-1} M(n)^T x(n). \quad (5.47)$$

Unter den Annahmen, daß

1. $E\{m(k)z(k+i)\} = 0$ für alle i ,
2. $E\{z(k)z(k+i)\} = 0$ für alle $i \neq 0$,
3. $E\{z(k)\} = 0$ für alle k ,

besitzt das Schätzverfahren die folgenden Eigenschaften:

Erwartungstreue, d.h. $E\{\hat{s}\} = s$

Beweis:

$$E\{\hat{s}\} = E\{[M(n)^T M(n)]^{-1} M(n)^T x(n)\}.$$

Mit Gl. (5.41) gilt

$$E\{\hat{s}\} = E\{[M(n)^T M(n)]^{-1} M(n)^T M(n) s + z(n)\}.$$

Mit den getroffenen Annahmen gilt somit

$$E\{\hat{s}\} = E\{s\} = s.$$

Beste lineare Schätzung, d.h. minimale Parametervarianz

Voraussetzungen: Annahme 1. bis 3.

$z(k)$ normalverteilt

$$\sigma_z^2 = \text{const}$$

Beweis s. Abschnitt 5.2.4.

Zur Beurteilung der Güte der Schätzung dienen folgende Kriterien:

1. Restquadratsumme

$$S_R = \sum_{k=j}^{J+n-1} (x(k) - \hat{x}(k))^2 \quad (5.48)$$

2. Reststreuung

$$s_R^2 = \frac{S_R}{n - (l + 1)} \quad (5.49)$$

3. Varianzen der geschätzten Parameter

Aus der Beziehung für die Kovarianzmatrix des geschätzten Parametervektors

$$\text{COV}\{\hat{s}\} = E\{(s - \hat{s})(s - \hat{s})^T\} \quad (5.50)$$

erhält man durch Einsetzen von Gl. (5.41) in Gl. (5.47)

$$\text{COV}\{\hat{s}\} = E\{[M(n)^T M(n)]^{-1} M(n)^T z(n)^T M(n) [M(n)^T M(n)]^{-1}\}.$$

Mit der Annahme, daß

$$E\{z(n) z(n)^T\} = \sigma_z^2 I$$

ist, ergibt sich für die Kovarianzmatrix des Parametervektors $\hat{s}$

$$\text{COV}\{\hat{s}\} = P(n) \sigma_z^2 \quad (5.51)$$

mit

$$\sigma_z^2 = E\{z^T(n) z(n)\}, P(n) = [M^T(n) M(n)]^{-1}.$$

Im einzelnen:

die Kovarianzen $\text{cov}\{\hat{s}_i, \hat{s}_j\} = p_{ij} \sigma_z^2$

die Varianzen $\text{var}\{\hat{s}_i\} = p_{ii} \sigma_z^2$.

4. Varianzen des geschätzten Ausgangs $\hat{x}$

Die Varianz des geschätzten Ausgangs ergibt sich aus

$$\begin{aligned} \text{var}\{\hat{x}\} &= E\{(x - \hat{x})^2\} \\ &= E\{[(s - \hat{s})^T m^2]\} \\ &= E\{m^T (s - \hat{s})(s - \hat{s})^T m\} \end{aligned}$$

mit den Beziehungen in Gl. (5.51) zu

$$\text{var}\{\hat{x}\} = \sigma_z^2 m^T [M^T M]^{-1} m. \quad (5.52)$$

Beispiel 5.1 Einfluß der Beobachtungszahl und der Störgröße auf die Parametergüte bei der Schätzung mit der MKQ

Gegeben: Systemgleichung

$$x = \sum_{i=1}^4 a_i u_i + z,$$

$$a^T = [1, -2, 4, -10],$$

Eingangssignale $u_i \in \text{NV}(0, 1)$,

Störung $z \in \text{NV}(0, \dots)$ (Störverhältnis $\sqrt{\sigma_z^2/\sigma_x^2} \approx 100$).

Gesucht: Schätzwerte der Parameter und die Reststreuung.

Ergebnis:

n	Stör- verhältnis %	Parameter				
		a_1	a_2	a_3	a_4	s_R^2
10	10	1,268	-2,206	3,930	9,457	3,576
	50	1,835	-2,144	3,391	12,256	29,613
100	10	1,045	-2,009	3,961	9,942	1,066
	50	1,134	-2,189	4,430	10,547	26,104
500	10	1,027	-2,013	4,017	9,940	1,082
	50	1,231	-2,032	4,200	10,445	23,139

Die erzielten Ergebnisse machen deutlich, daß

1. die Güte der Schätzung mit zunehmender Beobachtungszahl zunimmt,
2. die Güte der Schätzung mit wachsender Störung abnimmt,
3. der Fehler der kleinen Parameter gegenüber den größeren Parametern wesentlich stärker von der Störung abhängt.

5. Konfidenzintervall für s_i mit dem t -Test (s. Abschn. 2.3.4)

Mit dem Freiheitsgrad $f = n - l - 1$, der Irrtumswahrscheinlichkeit α und der Näherung von $\sigma_z^2 \approx s_R^2$ ergibt sich für das Konfidenzintervall

$$|s_i - \hat{s}_i| = t_{\alpha/2, f} s_R \sqrt{p_{ii}}. \quad (5.53)$$

6. Hypothesentest: $H_0: s_i = 0$ (t -Test; s. Abschn. 2.3.3)

Der geschätzte Parameter $\hat{s}_i$ ist mit der Irrtumswahrscheinlichkeit α statistisch ungleich null, wenn

$$|\hat{s}_i| \geq t_{\alpha/2, f} s_R \sqrt{p_{ii}}. \quad (5.54)$$

Das typische Verhalten der MKQ ist im Beispiel 5.1 bezüglich des Einflusses des Stichprobenumfangs und der Störampitude für einen Testfall dargestellt.

Zur Lösung der Schätzgleichung von Gl. (5.47) kommen verschiedene numerische Verfahren zur Anwendung. Ist die in der Gl. (5.47) zu invertierende Matrix $M(n)^T M(n)$ nichtsingulär und symmetrisch, ist die Anwendung des Cholesky-Verfahrens vorteilhaft. Wenn die zu invertierende Matrix schlecht konditioniert ist, kann die Inversion durch eine Umstellung der Gl. (5.47) in

$$M(n)^T M(n) \hat{s} = M(n)^T x(n) \quad (5.55)$$

umgangen werden. Für die Lösung dieses Gleichungssystems können dann die Gauß-Elimination, die Zerlegung in Dreiecksmatrizen nach *Gauß-Banachiewicz* oder die Spaltenpivotsuche nach *Gauß-Jordan* zur Anwendung (s. Tafel 5.2) kommen. Die Methode der Regression liefert unter den eingangs genannten Voraussetzungen gute und robuste Parameterschätzungen. In den folgenden Abschnitten sollen Methoden vorgestellt werden, die unter den Bedingungen einer starken Korrelation zwischen den Eingangsgrößen, der sog. Multikollinearität oder einer korrelierten Störung Schätzungen mit minimalen Varianzen der Parameter erlauben. Es sind dies die Methode der Kammlinien- und der verallgemeinerten Regression.

5.2.3. Methode der Kammlinienregression

Die Methode der Kammlinienregression (KLR), auch als Ridge-Regression bezeichnet, ist von *Hörl* ursprünglich zur Interpretation von Gütefeldern entwickelt worden [5.28]. Erst später wurde sie zur Parameterschätzung bei korrelierten Eingangsgrößen eingesetzt [5.24].

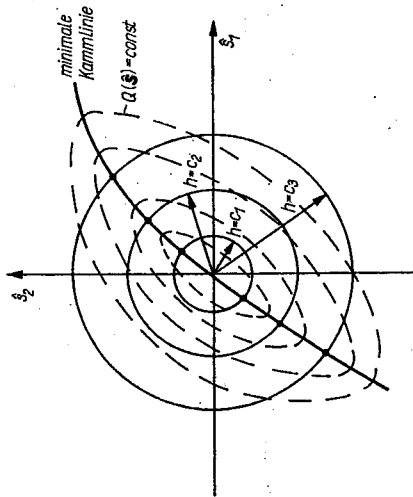


Bild 5.12
Gütefeldanalyse mit Hilfe
der Kammlinienregression

Die Methode geht von der Annahme aus, daß jedes Gütefeld durch eine Kammlinie gekennzeichnet werden kann, die bei vorgegebenen Radien h^2 im Parameterraum die Punkte minimaler Werte der Gütefunktion $Q(\xi)$ verbindet (Bild 5.12). Diese Kammlinie wird als minimale Kammlinie bezeichnet. Die Ermittlung der minimalen Kammlinie unter der Nebenbedingung konstanter Radien um das Zentrum des Gütefelds der Form

$$q(\xi) = h^2 - \xi^T \xi = \text{const} \quad (5.56)$$

mit

$$h = \|\xi\| = \sqrt{s_1^2 + \dots + s_i^2}$$

erfolgt durch eine Erweiterung der quadratischen Gütefunktion. Die in Gl. (5.56) formulierte Nebenbedingung wird über den Lagrange-Multiplikator λ in der erweiterten Gütefunktion

$$Q(\xi, \lambda) = Q(\xi) + \lambda q(\xi) \quad (5.57)$$

berücksichtigt.

Wird in Gl. (5.57) für

$$Q(\xi) = (x - M\xi)^T (x - M\xi)$$

eingesetzt, erhält man für

$$Q(\xi, \lambda) = (x - M\xi)^T (x - M\xi) + \lambda (h^2 - \xi^T \xi). \quad (5.58)$$

Aus Gl. (5.58) gewinnt man durch Ausmultiplizieren und Zusammenfassen als Ausgangsbeziehung für die Lösung der Minimierungsaufgabe:

$$Q(\xi, \lambda) = x^T x - 2\xi^T M^T x + \xi^T M^T M \xi + \lambda h^2 - \lambda \xi^T \xi. \quad (5.59)$$

Das Minimum der Gütefunktion $Q(\xi, \lambda)$ wird wiederum durch die Bildung der partiellen Ableitungen und anschließendes Nullsetzen ermittelt. Mit

$$\frac{\partial Q(\xi, \lambda)}{\partial \xi} = 0$$

und Gl. (5.59) erhält man

$$\frac{\partial Q(\xi, \lambda)}{\partial \xi} = 2 [M^T M \xi - \lambda \xi - M^T x] = 0. \quad (5.60)$$

Wenn die Inverse der Matrix $M^T M - D(\lambda)$ mit $D(\lambda) = \lambda I$ existiert, kann der Parametervektor nach der Beziehung

$$\xi = [M^T M - D(\lambda)]^{-1} M^T x \quad (5.61)$$

geschätzt werden. Die Vorschrift von Gl. (5.61) wird als *Kammlinienregression* bezeichnet. Die minimale Kammlinie ist nach [5.28] durch λ -Werte definiert, die kleiner sind als der minimale Eigenwert der Matrix $[M^T M]$. Da der minimale Eigenwert dieser Matrix bei Multikollinearität gegen Null strebt, d.h. $\lambda_{\min} \approx 0$, sind die λ -Werte in Gl. (5.61) von Null in Richtung negativer Werte, $\lambda < 0$, zu wählen. Als günstig hat sich bei vielen simulierten Testfällen der Bereich

$$0 > \lambda > -1$$

zur Suche eines zweckmäßigen λ -Wertes erwiesen. Die Anwendung der Kammlinienregression ist gegenüber der Regression nur sinnvoll, wenn

- zwei oder mehr Eingangsgrößen stark korreliert sind

$$\hat{\rho}(u_i, u_j) \geq 0,85,$$

- das System stärker gestört ist ($> 10\%$).

Unter diesen Bedingungen wird die Varianz der Parameter durch das Verfahren der Kammlinienregression stark reduziert. Im Beispiel 5.2 wird dieses Verhalten an ausgewählten Testfällen aufgezeigt.

Ein Problem bleibt das Kriterium zur Beurteilung der Modellgüte, d.h. die Festlegung eines λ -Wertes aus dem Kompromiß zwischen der Reduzierung der Parametervarianz und der Erhöhung der Restquadratsumme. Die minimale Restquadratsumme liefert das

Beispiel 5.2	Vergleich der Schätzverfahren MKQ und KLR bei verschiedener Korrelation von m
--------------	---

Gegeben:

Systemgleichung

$$x = \sum_{i=1}^5 a_i u_i + z,$$

$$a^T = [3, 2, -2, 1, -4],$$

Eingangssignale $u_i \in NV(0, 1)$ mit den Korrelationsmatrizen

Fall	u_1	u_2	u_3	u_4	u_5
I	u_1 0,947 u_2 0,024 u_3 -0,076 u_4 -0,110 u_5	1 0,023 -0,063 -0,201	1 0,160 -0,173	1 -0,125	1
II	u_1 0,820 u_2 0,937 u_3 0,960 u_4 0,986 u_5	1 0,968 0,947 0,903	1 0,997 0,982	1 0,993	1

Störung $z \in NV(0, \sigma_z)$, 10 %
(mit Störverhältnis $\sqrt{\sigma_z^2/\sigma_x^2} = 100$),

Modellgleichung

$$\hat{x} = \sum_{i=1}^5 \hat{a}_i u_i.$$

Gesucht: Schätzwerte der Parameter a_i .

Ergebnis: Für $n = 200$

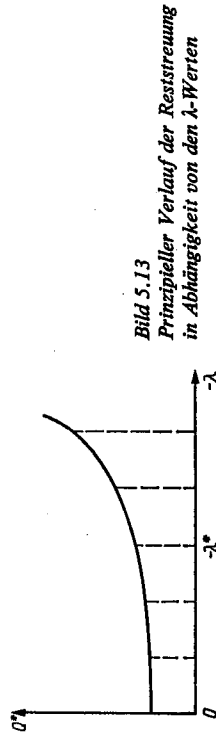
Testfall	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{a}_5$
I	MKQ 3,18 KLR 3,01 ($\lambda^* = -0,5$)	1,75 1,96	-2,06 -1,96	1,11 1,06	-4,01 -3,96
II	MKQ -60 KLR 3,24 ($\lambda^* = -0,25$)	169 1,15	-59 -0,81	554 2,68	268 -3,52

Die vorgelegten Ergebnisse zeigen, daß

1. der Parameterfehler im Testfall I, d.h. bei einer relativ geringen Korrelation zwischen den Komponenten des Meßwertvektors m (nur u_1 ist mit u_2 stark korreliert) der Methode der KLR nur geringfügig besser als der der MKQ ist,
2. bei einer starken Korrelation zwischen den Komponenten des Meßwertvektors m (Testfall II) das Versagen der MKQ deutlich wird; dagegen die Parameterwerte der KLR noch in Vorzeichen und Betrag sinnvoll sind.

Verfahren der Regression, d.h. $\lambda = 0$. Als günstig hat sich folgende heuristische Vorgehensweise bewährt [5.30]:

1. Schätzung mittels Regression ($\lambda = 0$)
2. Schätzung mittels Kammlinienregression bei verschiedenen λ_i
3. Wahl des negativsten λ_i -Wertes, bei dem noch kein wesentlicher Anstieg der Restquadratsumme vorliegt (s. Bild 5.13 und Beispiel 5.1).



Der rechentechnische Aufwand jeder einzelnen Schätzung ist bei der Kammlinienregression und der Regression nahezu gleich (Tafel 5.5). Zur Suche eines günstigen Kompromisses zwischen Reststreuung und Parametervarianz erhöht sich der Aufwand linear mit der Anzahl der gewählten λ -Werte. Weitere Wege zur iterativen Suche günstiger λ -Werte sind in [5.30] dargestellt.

Tafel 5.5
Vergleich rechentechnischer
Parameter direkter Schätzverfahren

Verfahren	Rechenzeit bestimmt durch	Speicherplatz
MKQ	Inversion $[M^T M]$ mit $l \times l$	$l^2 + l(1 + n) + n$
KLR	Inversion von $[M^T M]$ mit $l \times l$	$l^2 + l(2 + n) + n$
VMKQ	Inversion von $-W$ mit $n \times n$ $-[M^T W^{-1} M]$ mit $l \times l$	$l^2 + n^2 + n + l(1 + n)$

5.2.4. Direkte Methode der verallgemeinerten Regression/Markov-Schätzung

Die verallgemeinerte Methode der kleinsten Quadrate (VMKQ) geht von einer Zielfunktion

$$Q = e^T W^{-1} e \quad (5.62)$$

aus, bei der der Fehlervektor durch eine Wichtungsmatrix W^{-1} beeinflusst wird. Die Wichtung des Fehlers wird z.B. bei korrelierten Störungen z oder bei der Schätzung zeitvarianter Systeme angewendet. Analog zur MKQ gilt für die VMKQ die Fehlergleichung (5.43) und damit

$$Q = (x - M\hat{s})^T W^{-1} (x - M\hat{s}). \quad (5.63)$$

Über die Gleichung

$$Q = x^T W^{-1} x - 2\hat{s}^T M^T W^{-1} x + \hat{s}^T M^T W^{-1} M \hat{s} \quad (5.64)$$

erhält man das Minimum von Q aus der Beziehung

$$\frac{\partial Q}{\partial \hat{s}} = -2M^T W^{-1}x + 2M^T W^{-1}M\hat{s} = 0. \quad (5.65)$$

Wenn die Inverse der Matrix $M^T W^{-1}M$ existiert, gilt für die Schätzvorschrift der VMKQ

$$\hat{s} = [M^T W^{-1}M]^{-1} M^T W^{-1}x. \quad (5.66)$$

Das Schätzverfahren der VMKQ besitzt folgende Eigenschaften:

- Die Schätzung ist erwartungstreu unter Beachtung der Bedingungen von Abschnitt 5.2.1.
- Die Parametervarianzen sind minimal, da

$$\text{cov}\{\hat{s}_{\text{VMKQ}}\} = E\{[M^T W^{-1}M]^{-1} M^T W^{-1}z z^T W^{-1}M [M^T W^{-1}M]^{-1}\} \quad (5.67)$$

und somit

$$\text{cov}\{\hat{s}_{\text{VMKQ}}\} \leq \text{cov}\{\hat{s}_{\text{MKQ}}\}$$

gilt.

Das Hauptproblem der Methode besteht in der Festlegung der Elemente der Wichtungsmatrix W und ihrer Inversion. Die Dimension der Wichtungsmatrix ist $n \times n$ bei n Beobachtungen. Eine der leistungsfähigsten Methoden zur Festlegung der Wichtungsmatrix W ist die Verwendung der Kovarianzmatrix Z der Störung (s. Abschn. 5.1.2.3). Die mit dieser Festlegung erhaltene Schätzvorschrift wird als Markov-Schätzung bezeichnet und lautet:

$$\hat{s} = [M^T Z^{-1}M]^{-1} M^T Z^{-1}x. \quad (5.68)$$

Diese Schätzung besitzt auch bei korrelierten Störungen die für die VMKQ angegebenen Eigenschaften.

Da die Kovarianzmatrix Z der Störung eine doppelt symmetrische Matrix ist, ist durch eine Transformation der Meßwerte eine Rückführung auf die Regression möglich. Die Lösung erfolgt in den Schritten [1.31, 5.31]:

1. Cholesky-Zerlegung von Z in

$$Z = R^T R,$$

2. Einsetzen der Zerlegung in Gl. (5.68)

$$\hat{s} = [M^T (R^T R)^{-1} M]^{-1} M^T (R^T R)^{-1} x,$$

3. Transformation der Meßwerte

$$\tilde{M} = (R^{-1})^T M \quad (5.69)$$

$$\tilde{x} = (R^{-1})^T x,$$

4. Einsetzen der transformierten Meßwerte

$$\hat{s} = [\tilde{M}^T \tilde{M}]^{-1} \tilde{M}^T \tilde{x}. \quad (5.70)$$

Somit ist rechenstechnisch immer der gleiche Modul der Regression verwendbar. Da die Störung selten meßbar ist und häufig auch A-priori-Informationen über die Kovarianzmatrix fehlen, wurde neben Schätzverfahren eine Strategie erfolgreich erprobt, die die

Beispiel 5.3 Vergleich der Schätzverfahren MKQ und der Markov-Schätzung bei verschiedenen Störampplituden

Gegeben:

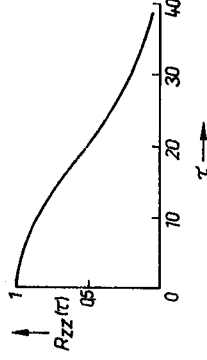
Systemgleichung

$$x = \sum_{i=1}^5 a_i u_i + z,$$

$$a^T = [1, -2, 3, 8, 10],$$

Eingangssignale $u_i \in \text{NV}(0, 1)$,

Störung $z \in \text{NV}(0, \dots)$,



Modellgleichung $\hat{x} = \sum_{i=1}^5 a_i u_i$.

Schätzwerte der Parameter a_i .

Für $n = 90$

Störung %		a_1	a_2	a_3	a_4	a_5
25	MKQ	1,42	-1,75	2,91	8,05	9,92
	Markov-Schätzung	1,01	-2,01	2,99	8,02	9,98
50	MKQ	1,78	-1,55	2,84	8,12	9,80
	Markov-Schätzung	1,03	-1,99	2,98	8,02	9,96
100	MKQ	2,58	-1,11	2,73	8,23	9,62
	Markov-Schätzung	1,05	-1,96	2,95	8,03	9,95
150	MKQ	3,42	-0,64	2,61	8,31	9,50
	Markov-Schätzung	1,15	-1,93	2,93	8,03	9,92

Die dargestellten Ergebnisse zeigen, daß

- die kleinsten Parameter mit dem größten relativen Fehler geschätzt werden,
- die Überlegenheit der Markov-Schätzung gegenüber der MKQ bei stärkeren Störungen noch zunimmt,
- erst bei einer relativ starken Korrelation der Störung die Überlegenheit der Markov-Schätzung merklich wird.

Kovarianzmatrix der Störung durch die des Fehlers annähert. Damit gilt die Annahme

$$Z \approx E = E\{ee^T\}. \quad (5.71)$$

Die Kovarianzmatrix des Fehlers kann durch eine zweistufige Schätzstrategie entsprechend dem im Bild 5.14 dargestellten Schema gewonnen und verwendet werden. Die erste Stufe besteht in einer MKQ-Schätzung. Aus dem Fehlervektor wird eine erste Schätzung der Kovarianzmatrix des Fehlers vorgenommen. In der zweiten Stufe wird dann die Markov-Schätzung durchgeführt, und über den neugebildeten Fehlervektor erfolgt in Iterationen eine weitere Anpassung, bis Abbruchschranken für die Parameterwerte und/oder Parametervarianzen erreicht sind. Weiterführende Untersuchungen sind in [5.31] enthalten.

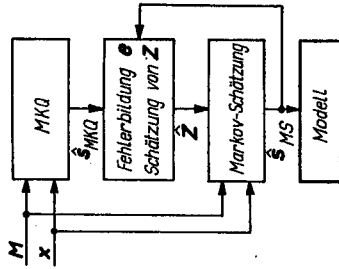


Bild 5.14
Schema der zweistufigen Markov-Schätzung

Im Vergleich zur MKQ liefert die Markov-Schätzung wesentlich bessere Schätzergebnisse für den Parametervektor, d. h. eine geringere Parametervarianz, wenn

1. die Störung stark korreliert ist, d. h. die Korrelationslänge größer als 5T bis 10T ist,
2. eine große Störampplitude (> 10%, bezogen auf die Ausgangsgröße x) vorhanden ist,
3. die aus der Stichprobe ermittelte oder für die Stichprobe geltende Schätzung der Kovarianzmatrix der Störung verwendet wird.

Die Reststreuung der Markov-Schätzung erhöht sich dagegen im Vergleich zur Regression. Diese Aussagen werden deutlich sichtbar im Beispiel 5.3.

Die Rechenzeit und der Speicherplatz werden bei der MKQ durch die Dimension der Matrizen M und $[M^T M]^{-1}$ dominierend bestimmt. Bei der Markov-Schätzung steigt mit wachsender Beobachtungszahl n der Rechenaufwand zusätzlich enorm an, da die Kovarianzmatrix Z als eine $n \times n$ -Matrix zu invertieren ist (Tafel 5.5).

5.3. Iterative Schätzverfahren

5.3.1. Allgemeine Grundstruktur

Ist der Zusammenhang zwischen dem Schätzfehler und den Parametern des Modells nichtlinear, so ist eine analytische Lösung der Schätz Aufgabe durch Nullsetzen der Ableitung der Gütefunktion direkt in einem Schritt nicht mehr möglich. Einen Ausweg bilden iterative Schätzverfahren der Grundstruktur (s. Abschn. 5.1.5.2)

$$s^{i+1} = s^i + \Delta s^{i+1}. \quad (5.72)$$

Von den verschiedenen möglichen Varianten der

- gradientenfreien Verfahren
- Gradientenverfahren (einschließlich der Benutzung der zweiten Ableitungen)
- Zufallssuchverfahren

und ihren Kombinationen (s. Tafel 5.3) sollen im Rahmen dieses Buches nur die Gradientenverfahren nach *Newton-Raphson* [5.9, 5.11] und nach *Marquardt-Levenberg* [5.13, 5.14, 5.15] vorgestellt werden. Sie haben im Rahmen der Parameterschätzung die breiteste Verwendung gefunden, besonders bei der Schätzung der Koeffizienten der Differenzengleichung nach der Maximum-Likelihood-Methode (s. Abschn. 6).

5.3.2. Newton-Raphson-Verfahren

Beim *Newton-Raphson-Verfahren* wird der Gradient der Gütefunktion

$$\nabla Q(s) = \partial Q(s) / \partial s \quad (5.73)$$

an der Stelle i des Parametervektors des Modells in eine Taylor-Reihe der Form

$$\nabla Q(s)|_{s=s^i + \Delta s} = \nabla Q(s)|_{s=s^i} + \frac{\partial}{\partial s} \nabla Q(s)|_{s=s^i} (s^{i+1} - s^i) + \dots \quad (5.74)$$

entwickelt. Wird in Gl. (5.74)

$$\nabla Q(s)|_{s=s^i + \Delta s} = 0$$

gesetzt und die Reihenentwicklung nach dem zweiten Glied abgebrochen, erhält man nach einer einfachen Umstellung die iterative Schätzvorschrift für die *Newton-Raphson-Verfahren* zu

$$s^{i+1} = s^i - \left[\frac{\partial}{\partial s} \nabla Q(s) \right]_{s=s^i}^{-1} \nabla Q(s)|_{s=s^i}. \quad (5.75)$$

Die partiellen Ableitungen ergeben sich zu

$$\frac{\partial Q(s)}{\partial s} = \begin{bmatrix} \frac{\partial Q(s)}{\partial s_1} \\ \vdots \\ \frac{\partial Q(s)}{\partial s_i} \end{bmatrix}, \quad \frac{\partial^2 Q(s)}{\partial s \partial s^T} = \begin{bmatrix} \frac{\partial^2 Q(s)}{\partial s_1 \partial s_1} & \dots & \frac{\partial^2 Q(s)}{\partial s_1 \partial s_i} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 Q(s)}{\partial s_i \partial s_1} & \dots & \frac{\partial^2 Q(s)}{\partial s_i \partial s_i} \end{bmatrix}$$

Mit der Gütefunktion

$$Q(s) = \frac{1}{2} \sum_{k=j}^{j+n-1} e^2(k)$$

erhält man für die ersten und zweiten Ableitungen folgende Ausdrücke:

$$\begin{aligned} \frac{\partial Q(s)}{\partial s_i} &= \sum_{k=j}^{j+n-1} e(k) \frac{\partial e(k)}{\partial s_i} \\ \frac{\partial^2 Q(s)}{\partial s_i \partial s_m} &= \sum_{k=j}^{j+n-1} \frac{\partial e(k)}{\partial s_i} \frac{\partial e(k)}{\partial s_m} + \sum_{k=j}^{j+n-1} \frac{\partial^2 e(k)}{\partial s_i \partial s_m} \end{aligned} \quad (5.76)$$

Die Realisierung der Schätzvorschrift in Gl. (5.75) erfolgt mit diesen Gleichungen in folgenden Schritten:

1. Wahl günstiger Startparameterwerte $\hat{s}^0$.
2. Berechnung der ersten und zweiten Ableitungen mit den Werten der Beobachtungen $e(k); k = j, \dots, j + n - 1$.
3. Ermittlung des neuen Parametervektors $\hat{s}^{i+1}$ nach Gl. (5.75).
4. Test, ob ein vorgegebenes Abbruchkriterium erfüllt ist.
5. Wenn Abbruchkriterium nicht erfüllt ist, Wiederholung ab Schritt 2.

Für dieses Verfahren muß vorausgesetzt werden, daß die Gütefunktion $Q(s)$ in jedem Punkt $s \in S (S \in \mathbb{R}^n)$ stetige partielle Ableitungen 2. Ordnung besitzt. Da i. allg. nicht sichergestellt ist, daß $\hat{s}^{i+1} \in S$, wird der Algorithmus von Gl. (5.75) zur Sicherung der Konvergenz um einen Dämpfungsfaktor $\beta^i (\beta^i > 0)$ in der Form

$$\hat{s}^{i+1} = \hat{s}^i - \beta^i \left[\frac{\partial}{\partial \hat{s}} \nabla Q(\hat{s})^T \right]_{\hat{s}=\hat{s}^i}^{-1} \nabla Q(\hat{s})|_{\hat{s}=\hat{s}^i} \quad (5.77)$$

erweitert. Wenn der Faktor β^i genügend klein gewählt wird, ist nach [5.32] gesichert, daß der neue Parametervektor $\hat{s}^{i+1} \in S$ ist. Damit ist die Konvergenz garantiert. Der Rechenaufwand des Verfahrens ist durch die Bildung und Inversion der Matrix der zweiten Ableitungen, d. h. der Hesse-Matrix, relativ hoch. Für stärker gestörte Systeme ist dieses Verfahren ungeeignet.

5.3.3. Levenberg-Marquard-Verfahren

Um die Konvergenz unabhängig von den Startwerten $\hat{s}^0$ und der Größe der Störung zu sichern, wurde von *Levenberg-Marquard* [5.13, 5.33] der erweiterte Algorithmus von Gl. (5.78) entworfen. Er geht davon aus, daß die Suche des Minimums der Gütefunktion unter der Nebenbedingung konstanter Radien h^2 der Änderung des Parametervektors von Iteration zu Iteration erfolgt.

Das heißt, es gilt

$$\Delta \hat{s} = \hat{s}^{i+1} - \hat{s}^i \quad \text{und} \quad \Delta \hat{s}^T \Delta \hat{s} = h^2 = \text{const.}$$

Die Berücksichtigung der Nebenbedingung in Form einer Gleichungsbeschränkung erfolgt analog zu Abschnitt 5.2.3 über den Lagrange-Multiplikator λ . Die iterative Schätzvorschrift, die man über eine Reihe weiterer Ableitungen erhält (s. [5.13, 5.14]), hat folgende Gestalt:

$$\hat{s}^{i+1} = \hat{s}^i - \left[\frac{\partial}{\partial \hat{s}} \nabla Q(\hat{s})^T \right]_{\hat{s}=\hat{s}^i}^{-1} \nabla Q(\hat{s})|_{\hat{s}=\hat{s}^i} \quad (5.78)$$

Da für alle $\lambda < 0$ die zu invertierende Matrix positiv definit ist, ist damit auch für jeden Punkt $\hat{s} \in S$ die Konvergenz des Verfahrens gesichert. Durch eine geeignete Wahl des Faktors λ , der als ein Maß für die Schrittweite angesehen werden kann, ist das Konvergenzverhalten des Verfahrens gut zu steuern. Im Beispiel 5.4 ist für die Schätzung eines gestörten nichtlinearen Systems das Konvergenzverhalten des Newton-Raphson-Verfahrens mit dem des Levenberg-Marquard-Verfahrens verglichen worden.

Beispiel 5.4. Untersuchung des Konvergenzverhaltens des Newton-Raphson-Verfahrens mit dem Levenberg-Marquard-Verfahren

Gegeben:

Systemgleichung

$$x = 1u_1 + e^{+0.5u_2},$$

Eingangssignale NV (0, 1), unkorreliert,

Störsignal $z \approx NV(0; 1, 4)$,

Modellgleichung $\hat{x} = d_1 u_1 + e^{d_2 u_2}$,

Startparametervektor $\hat{d}(0) = 0$,

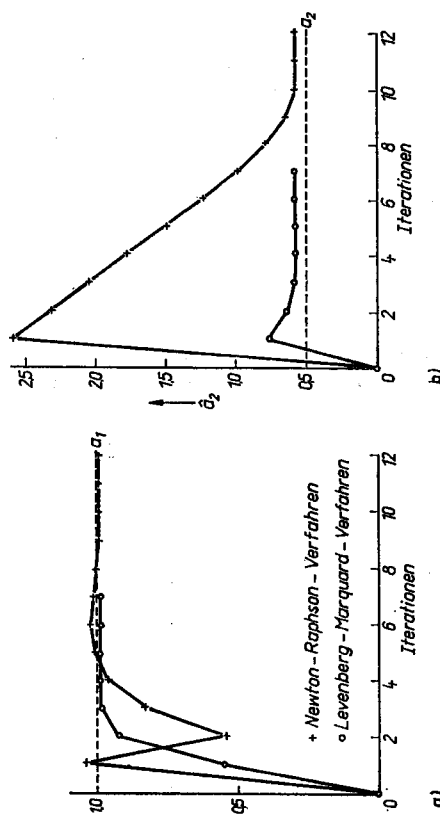
Abbruchgrenze $\varepsilon = 10^{-4}$.

Gesucht:

Schätzverlauf für die Parameter d_1 und d_2 .

Ergebnis:

Siehe Bild zu Beispiel 5.4.



Die Schätzverläufe für die Parameter d_1 und d_2 zeigen, daß bei einer geeigneten Wahl von λ die Konvergenz des Levenberg-Marquard-Verfahrens gegenüber dem Newton-Raphson-Verfahren besser ist. Beide Verfahren liefern gute Schätzergebnisse. Wird die Störampplitude noch größer, konvergiert das Newton-Raphson-Verfahren nicht mehr.

5.4. Indirekte/rekursive Schätzverfahren

5.4.1. Allgemeiner Ansatz

Die rekursiven Algorithmen haben besonders durch die Anwendung von Mikro- und Prozeßrechnern an Bedeutung gewonnen, weil sie einen geringen Speicherplatz benötigen, die Rechenzeit in einem Schritt gering ist und weil sie für die On-line-Datenverarbeitung und Steuerung sehr gut geeignet sind. Außerdem können sie durch geeignete Wahl der Korrekturfaktoren mit „Lerneigenschaften“ ausgerüstet werden, die eine aktuelle Parameter- und Zustandsschätzung zeitvarianter Systeme gestatten.

Sie beruhen darauf, daß die Zielfunktion Q schrittweise aufgrund neuer Beobachtungen

zum Minimum geführt wird. Die Vielfalt der entwickelten Verfahren kann durch die Grundstrukturen (s. Abschn. 5.1.5.3)

– für kontinuierliche Parameterschätzung

$$\dot{s}(t) = -\gamma R' \nabla Q(s, t) \quad (5.79)$$

– für diskontinuierliche Parameterschätzung

$$\begin{aligned} \text{neuer} \quad & s(k+1) = s(k) + k(k+1) [x(k+1) - m^T(k+1)s(k)] \\ \text{Parameter-} & \text{vektor} = \text{Parameter-} + \text{Korrektur-} \\ & \text{vektor} \quad \text{vektor} \quad \text{vektor} \quad \times \text{Vorhersage-} \\ & \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \text{fehler} \end{aligned} \quad (5.80)$$

beschrieben werden (s. Tafel 5.4).

Entsprechend der üblichen Schreibweise kennzeichnen die in den runden Klammern stehenden Argumente wieder den Zeitpunkt der Abtastung.

In den folgenden Abschnitten sollen die Grundideen wesentlicher Verfahren hergeleitet und das Konvergenzverhalten der Schätzung untersucht und verglichen werden. Dabei werden Verfahren ausgewählt, die sich bei einer Vielzahl von Modelltypen und in der Praxis bewährt haben. Bei der Ableitung der Algorithmen wird von den Gradientenverfahren ausgegangen und gezeigt, daß auch die rekursiven Varianten der MKQ und der VMKQ durch die in Gl. (5.80) angegebene Grundstruktur beschrieben werden können.

5.4.2. Gradientenmethoden

5.4.2.1. Allgemeine Grundstrukturen

Die Verfahren dieser Gruppe beruhen darauf, daß die Parameteranpassung und -nachführung durch die laufende Korrektur eines zeitlich zurückliegenden Parameterwerts mit Hilfe der Bewertung des Gradienten in Betrag und Richtung erfolgt (s. Tafel 5.4 und Abschn. 5.1.5.3). Im Rahmen dieses Buches werden nur Verfahren vorgestellt, die auf eine Richtungskorrektur des Gradienten verzichten, d. h., es gilt immer $R = I$. Weiterhin wird für die Schrittweitenmatrix Γ gesetzt:

$$\Gamma = \gamma I = \begin{bmatrix} \gamma_1 & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \gamma_l \end{bmatrix} \quad (5.81)$$

Werden die Schrittweiten γ_i für alle Parameter gleich groß gewählt, gilt

$$\Gamma = \gamma I. \quad (5.82)$$

Damit erhält man für die Einstellvorschriften der Gradientenverfahren im a) kontinuierlichen Fall

$$\dot{s}(t) = -\Gamma \nabla Q(s, t), \quad (5.83)$$

b) diskontinuierlichen Fall

$$s(k+1) = s(k) - \Gamma(k+1) \nabla Q(s, k). \quad (5.84)$$

Die kontinuierliche Ermittlung der Parameter erfolgt über eine Integration von Gl. (5.83), d. h. über die Vorschrift

$$\dot{s}(t) = -\Gamma \int_{-\infty}^t \nabla Q(s, \tau) d\tau. \quad (5.85)$$

Im Bild 5.15 ist das Konzept der kontinuierlichen Parametermittlung nach Gl. (5.85) für einen Parameter dargestellt. Es ist zu sehen, daß das Minimum der Zielfunktion Q und damit der richtige Parameter $s = s^*$ gefunden wird, indem die Parameteränderung entgegen der Richtung des Gradienten ∇Q und mit einer Schrittweite γ erfolgt. Ferner wird sichtbar, daß die Konvergenzgeschwindigkeit bei diesen Verfahren mit der Abnahme des Gradienten auch abnimmt. Das bedeutet, daß theoretisch erst für $t \rightarrow \infty$ das Minimum erreicht wird.

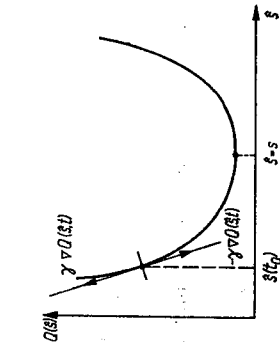


Bild 5.15. Schema der kontinuierlichen Parametermittlung

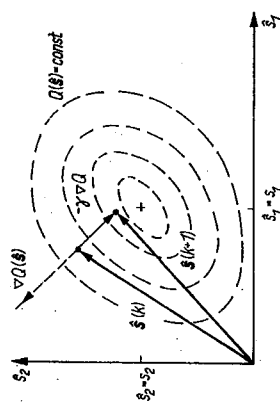


Bild 5.16. Schema der diskontinuierlichen Parametermittlung

Die diskontinuierliche Parametermittlung erfolgt direkt nach der Vorschrift in Gl. (5.84). Für die Ermittlung von zwei Parametern ist die Vorgehensweise im Bild 5.16 dargestellt. Typisch ist auch in diesem Fall, daß die Parameteranpassung, d. h. die Suche des Minimums im Gütefeld, von einem Startparametervektor ausgehend, entgegen der Richtung des Gradienten mit einer bestimmten Schrittweite erfolgt. Für die Ableitung der rekursiven Schätzverfahren und ihre einheitliche Darstellung ist es zweckmäßig, in Gl. (5.84) den Gradienten $\nabla Q(s, k)$ für eine vorgegebene Zielfunktion und eine Fehlergleichung analytisch zu berechnen.

Wird als Zielfunktion

$$Q(s, i) = \frac{1}{2} e(i)^2; \quad i = k, \dots, k+n \quad (5.86)$$

und als Fehlergleichung der Ausgangsfehler im $(k+1)$ -ten Schritt

$$e(k+1) = x(k+1) - m(k+1)^T s(k) \quad (5.87)$$

verwendet, erhält man für den Gradienten der Zielfunktion den Ausdruck

$$\nabla Q(s, k) = -m(k+1)^T [x(k+1) - m(k+1)^T s(k)]. \quad (5.88)$$

Unter der Annahme, daß die Schrittweiten in allen Richtungen des Gradienten gleich gewählt werden, d. h. $\gamma_1(k+1) = \gamma_2(k+1) = \dots = \gamma_l(k+1) = \gamma(k+1)$, kann mit Gl. (5.88) die Grundstruktur von Gl. (5.84) in die Beziehung

$$s(k+1) = s(k) + \gamma(k+1) m(k+1) [x(k+1) - m(k+1)^T s(k)] \quad (5.89)$$

als allgemeine Form der hier dargestellten Gradientenverfahren übergeführt werden. Der Zusammenhang zum Ansatz aller rekursiven Schätzverfahren – s. Gl. (5.80) – ist dann gegeben, wenn

$$\gamma(k+1)m(k+1) = k(k+1) \quad (5.90)$$

gesetzt wird.

Die in der Literatur vorgestellten reinen Gradientenverfahren unterscheiden sich dann nur noch in der Wahl der Schrittweite $\gamma(k+1)$. Aus diesem Grund werden in den weiteren Abschnitten typische Vertreter für die Wahl der Schrittweite vorgestellt.

5.4.2.2. Methode der Relaxation

Die Methode der *Relaxation* (REL) wurde zur Lösung determinierter Gleichungssysteme entwickelt [1.36, 5.21]. Bei Vorliegen einer quadratischen Zielfunktion Q gilt für die Konvergenz des Verfahrens folgende Bedingung für den Korrekturfaktor $\gamma(k+1)$:

$$0 < \gamma(k+1) < \frac{c}{\lambda_1} \quad (5.91)$$

Dabei ist λ_1 der größte Eigenwert der Matrix $M^* = mm^T$ und c eine wählbare Konstante. Mit der Abschätzung

$$\lambda_1 \leq \|m(k+1)m^T(k+1)\| \leq \|m(k+1)\|^2 \quad (5.92)$$

und

$$0 < \gamma(k+1) < \frac{c}{\|m(k+1)\|^2} \leq \frac{c}{\lambda_1}$$

folgt die Schätzgleichung für das Verfahren der Relaxation:

$$\hat{s}(k+1) = \hat{s}(k) + \frac{c}{\|m(k+1)\|^2} m(k+1) [x(k+1) - m^T(k+1)\hat{s}(k)]. \quad (5.93)$$

Die Konstante c liegt nach Schwarz [5.5] für bestimmte Eigenschaften der Matrix M^* im Bereich $0 < c < 2$. Nach Untersuchungen in [1.31] ist die Wahl im Bereich $0,9 < c < 1$ für viele Aufgaben der Identifikation ausreichend.

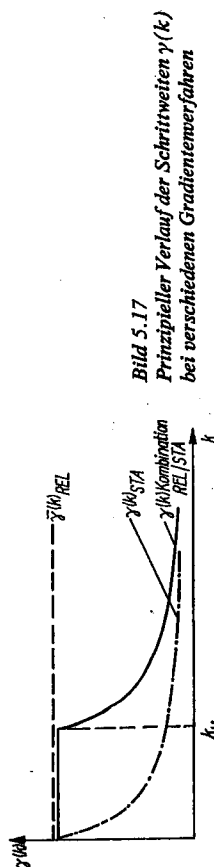


Bild 5.17
Prinzipieller Verlauf der Schrittweiten $\gamma(k)$
bei verschiedenen Gradientenverfahren

Wenn die Streuung der Eingangssignalwerte konstant ist, ist der Mittelwert $\bar{\gamma}(k)$ auch konstant (Bild 5.17). Dadurch ist das Konvergenzverhalten bei ungestörten bzw. gering gestörten Systemen gut. Die Parameterwerte sind bereits nach 51 bis 101 Beobachtungen praktisch konvergiert. Für gestörte Systeme ($z \geq 10\%$, bezogen auf die Ausgangsgröße) ist das Verfahren ungeeignet. In den Beispielen 5.5 und 5.6 sind diese Aussagen für einen Testfall dargestellt.

Beispiel 5.5 Vergleich des Konvergenzverhaltens von rekursiven Schätzverfahren – ungestörtes System

Gegeben:

Systemgleichung

$$x = \sum_{i=1}^5 a_i u_i$$

$$a^T = [1, 2, -3, -10, 20],$$

Eingangssignale $u_i \in \text{NV}(0, 1)$,

$$\text{Modellgleichung } \hat{x} = \sum_{i=1}^5 \hat{a}_i u_i,$$

Startparametervektor $\hat{a}(0) = 0$.

Gesucht:

Verlauf des Modellparameterfehlers

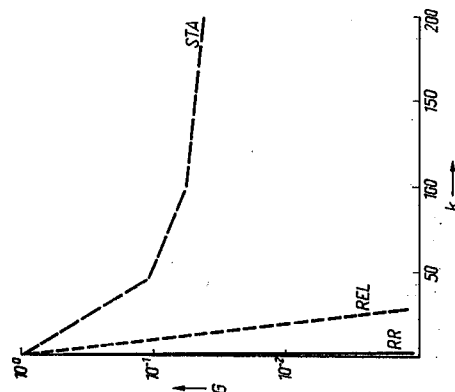
$$G = \frac{\|a - \hat{a}(k)\|}{\|a - \hat{a}(0)\|} \text{ als Funktion von } k = 1, \dots, n.$$

Ergebnis:

1. Parameterwerte bei $k = 20$

Verfahren	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{a}_5$
RR	1,00	2,00	-3,00	-10,00	20,00
STA	0,66	0,17	0,67	-10,93	19,29
REL	1,44	1,75	-1,91	-8,94	19,33

2. Verlauf des Modellfehlers G .



Die dargestellten Ergebnisse zeigen, daß

1. das Konvergenzverhalten des Verfahrens der rekursiven Regression (RR) allen anderen Verfahren überlegen ist,
2. das Konvergenzverhalten der Methode der Relaxation (REL) schlechter ist als das der rekursiven Regression, aber wesentlich besser als das der stochastischen Approximation (STA).

Beispiel 5.6	Vergleich des Konvergenzverhaltens von rekursiven Schätzverfahren - gestörtes System
--------------	--

Gegeben:

Systemgleichung

$$x = \sum_{i=1}^5 a_i u_i + z,$$

$$a^T = [1, 2, -3, -10, 20],$$

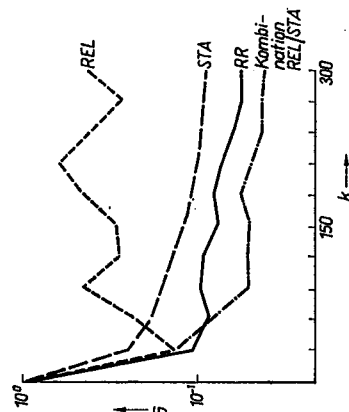
Eingangssignale $u_i \in NV(0, 1)$,Störung $z \in NV(0, \alpha)$; 20%,

$$\text{Modellgleichung } \hat{x} = \sum_{i=1}^5 \hat{a}_i u_i,$$

$$\text{Startparametervektor } \hat{a}(0) = 0.$$

Verlauf des Modellparameterfehlers G als Funktion von $k = 1, 2, \dots, n$.**Gesucht:****Ergebnis:**1. Parameterwerte $k = 200$

Verfahren	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{a}_5$
REL	4,37	4,44	-0,86	-11,25	24,08
STA	0,14	1,31	-1,48	-10,81	22,12
KOMB/REL-STA	0,19	2,34	-3,15	-10,32	20,64
$k_0 = 20$	1,18	1,98	-2,82	-10,90	22,07
RR					

2. Verlauf des Modellfehlers G .
Siehe Bild zu Beispiel 5.6.

Die Ergebnisse zeigen, daß

1. das Verfahren der Relaxation (REL) für gestörte Systeme ungeeignet ist,
2. die Methoden der rekursiven Regression (RR) und der Kombination Relaxation - stochastische Approximation (REL/STA) konvergieren,
3. das Verfahren der stochastischen Approximation (STA) langsam, aber sicher konvergiert.

Ferner ist das Verfahren der Relaxation bei ungestörten und gering gestörten Systemen in der Lage, zeitvariante Parameter nachzuführen. Ein typischer Schätzverlauf ist im Bild 5.18 dargestellt. Bei tiefrequenten Änderungen der Parameter erreicht das Verfahren fast die Güte der gewichteten rekursiven Regression.

Der rechentechnische Aufwand ist im Vergleich zu den direkten Verfahren und der Methode der rekursiven Regression sehr gering.

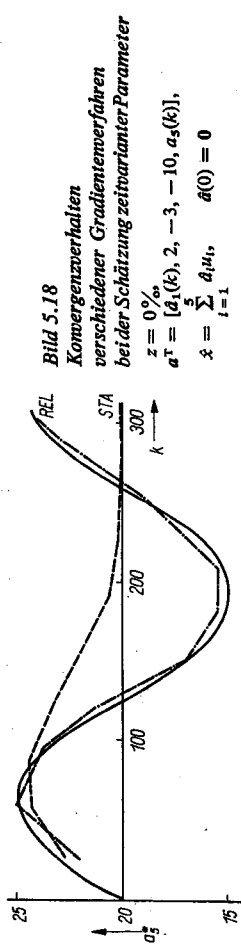


Bild 5.18

Konvergenzverhalten
verschiedener Gradientenverfahren
bei der Schätzung zeitvarianter Parameter

$$z = 0\%,$$

$$a^T = [\hat{a}_1(k), 2, -3, -10, a_5(k)],$$

$$\hat{x} = \sum_{i=1}^5 \hat{a}_i u_i, \quad \hat{a}(0) = 0$$

5.4.2.3. Methode der stochastischen Approximation

Die Methoden der *stochastischen Approximation* (STA) zählen zu den rekursiven Schätzverfahren und beruhen auf den Arbeiten von *Robbins-Monro* [5.23] und *Kiefer-Wolfowitz* [5.24]. Sie stellen die Anwendung einfacher Gradientenverfahren zur Bestimmung des Extremums einer Zielfunktion unter Verwendung der Gln. (5.84) und (5.86) dar. Damit erhält man als Grundgleichungen für die Verfahren der stochastischen Approximation

$$s(k+1) = s(k) + \Gamma(k+1)m(k+1)[x(k+1) - m(k+1)^T s(k)] \quad (5.94)$$

und bei gleicher Schrittweite $\gamma(k+1)$ für alle Parameter

$$s(k+1) = s(k) + \gamma(k+1)m(k+1)[x(k+1) - m(k+1)^T s(k)]. \quad (5.95)$$

Untersuchungen in [5.23, 5.24] haben ergeben, daß der in Gl. (5.94) dargestellte Algorithmus der stochastischen Approximation trotz stochastischer Störungen im Sinne des mittleren quadratischen Fehlers konvergiert, wenn die Schrittweite $\gamma(k)$ folgende Eigenschaften besitzt:

$$\sum_{k=1}^{\infty} \gamma(k) = \infty, \quad \sum_{k=1}^{\infty} \gamma(k)^2 < \infty \quad (5.96)$$

(Dvoretzky-Theorem [5.21]).

Eine allgemeine Berechnungsvorschrift für $\gamma(k)$ lautet:

$$\gamma(k) = \frac{C_1}{C_2 + C_3^{1-C_4}} \quad (5.97)$$

mit

$$C_1 \geq 1, \quad C_2 \geq 0, \quad C_3 > 0, \quad 0 \leq C_4 \leq 1/2.$$

Zumeist wird die in Gl. (5.97) auch enthaltene einfachste Form benutzt:

$$\gamma(k) = \frac{1}{k}.$$

Das Konvergenzverhalten der stochastischen Approximation ist sehr langsam, aber auch bei gestörten Systemen sicher (s. Beispiele 5.5 und 5.6). Die Ursache ist im Verlauf der Schrittweite $\gamma(k)$ als Funktion der bereits verarbeiteten Datenmenge zu sehen (s. Bild 5.17). Durch die Abnahme von $\gamma(k)$ mit Zunahme der Beobachtungsanzahl verliert der Algorithmus die Fähigkeit, Parameteränderungen des Systems zu erkennen und das Modell anzupassen. Daher ist das Verfahren für die Parameterschätzung zeitvarianter Systeme ungeeignet. Für einen Testfall ist dieses Verhalten im Bild 5.18 dargestellt.

5.4.2.4. Kombinierte Methode der Relaxation und der stochastischen Approximation

Aus der Analyse des Konvergenzverhaltens der Methode der Relaxation und der der stochastischen Approximation ist zu erkennen, daß eine Kombination beider Verfahren für die Parameterschätzung zeitinvarianter Systeme sinnvoll ist. Dabei wird folgende Strategie als kombinierte Methode vorgeschlagen [1.31]:

1. Beginn der Schätzung mit der Relaxation, d.h. $\gamma(k) = \text{const}$,
2. Umschalten auf die Methode der stochastischen Approximation nach 2l bis 4l Beobachtungen.

Damit erhält man den im Bild 5.17 dargestellten Verlauf der Schrittweite $\gamma(k)$. Als Schätzvorschrift gilt für die Kombination

$$\hat{s}(k+1) = \hat{s}(k) + \gamma'(k+1) \mathbf{m}(k+1) [x(k+1) - \mathbf{m}(k+1)^T \hat{s}] \quad (5.98)$$

mit

$$\gamma'(k+1) = \begin{cases} \frac{c}{\|\mathbf{m}(k+1)\|^2} & \text{für } k \leq k_u \\ c_u/k + 1 - k_u & \text{für } k > k_u; \end{cases}$$

k_u Umschaltzeitpunkt.

Die Schätzung konvergiert im Verhältnis zur normalen stochastischen Approximation bei kleiner und mittlerer Störung ($z < 50\%$) wesentlich schneller und erreicht in vielen Fällen die Modellgüte der rekursiven Regression (s. Beispiel 5.6). Für starke Störungen ($z > 100\%$) wird der günstige Umschaltzeitpunkt kleiner und die Güteverbesserung gegenüber der STA gering.

Insgesamt kann eingeschätzt werden, daß die Methode der Relaxation und der stochastischen Approximation sowie ihre Kombination nur dann angewendet werden sollten, wenn nicht genügend Rechenzeit und Speicherplatz für das leistungsstärkere Verfahren der rekursiven Regression zur Verfügung stehen.

5.4.3. Methode der rekursiven Regression

5.4.3.1. Ungewichtete rekursive Regression

Die rekursive Methode der Minimierung der kleinsten Quadrate oder die rekursive Regression (RR) ist eines der leistungsstärksten indirekten Parameterschätzverfahren. Ihre Ableitung geht von der Schätzgleichung der Regression Gl. (5.47) aus. Es wird angenommen, daß der Parametervektor $\hat{s}$ aus k Beobachtungen geschätzt wurde.

Damit gilt entsprechend Gl. (5.47) für die Schätzvorschrift zum k -ten Zeitpunkt

$$\hat{s}(k) = [\mathbf{M}(k)^T \mathbf{M}(k)]^{-1} \mathbf{M}(k)^T \mathbf{x}(k). \quad (5.99)$$

Steht eine $(k+1)$ -te Beobachtung, d.h. $\mathbf{m}(k+1)$ und $x(k+1)$, zur Verbesserung der Schätzung zur Verfügung, kann von Gl. (5.99) ausgehend für die Schätzvorschrift geschrieben werden:

$$\begin{aligned} \hat{s}(k+1) &= \left[[\mathbf{M}(k)^T \mathbf{m}(k+1)] \left[\frac{\mathbf{M}(k)}{\mathbf{m}^T(k+1)} \right] \right]^{-1} \\ &\quad \times [\mathbf{M}(k)^T \mathbf{m}(k+1)] \left[\frac{x(k)}{x(k+1)} \right]. \end{aligned} \quad (5.100)$$

Für die Matrix $\mathbf{P}(k+1)$ gilt dann analog

$$\begin{aligned} \mathbf{P}(k+1) &= [\mathbf{M}(k+1)^T \mathbf{M}(k+1)]^{-1} \\ &= [\mathbf{M}(k)^T \mathbf{M}(k) + \mathbf{m}(k+1) \mathbf{m}^T(k+1)]^{-1} \\ \mathbf{P}(k+1) &= [\mathbf{P}(k)^{-1} + \mathbf{m}(k+1) \mathbf{m}^T(k+1)]^{-1}. \end{aligned}$$

Unter Anwendung des Inversionslemmas für Matrizen

$$[\mathbf{A} + \mathbf{B} \mathbf{C} \mathbf{B}^T]^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{B} [\mathbf{C}^{-1} + \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B}]^{-1} \mathbf{B}^T \mathbf{A}^{-1}$$

kann Gl. (5.100) umgeformt werden in

$$\mathbf{P}(k+1) = \mathbf{P}(k) - \mathbf{k}(k+1) \mathbf{m}^T(k+1) \mathbf{P}(k) \quad (5.101)$$

mit

$$\mathbf{k}(k+1) = \mathbf{P}(k) \mathbf{m}(k+1) [1 + \mathbf{m}^T(k+1) \mathbf{P}(k) \mathbf{m}(k+1)]^{-1}.$$

Damit steht eine Vorschrift zur rekursiven Berechnung der Präzisionsmatrix $\mathbf{P}(k)$ zur Verfügung, die die rechtechnisch aufwendige Inversion in Gl. (5.99) umgeht. Es gilt damit für Gl. (5.100)

$$\hat{s}(k+1) = \mathbf{P}(k+1) [\mathbf{M}(k)^T \mathbf{x}(k) + \mathbf{m}(k+1) x(k+1)] \quad (5.102)$$

und mit den eingesetzten Beziehungen von Gl. (5.101)

$$\begin{aligned} \hat{s}(k+1) &= \mathbf{P}(k) - \mathbf{P}(k) \mathbf{m}(k+1) [1 + \mathbf{m}^T(k+1) \mathbf{P}(k) \mathbf{m}(k+1)]^{-1} \\ &\quad \times \mathbf{m}^T(k+1) \mathbf{P}(k) [\mathbf{M}(k)^T \mathbf{x}(k) + \mathbf{m}(k+1) x(k+1)]. \end{aligned} \quad (5.103)$$

Durch Ausmultiplizieren und einige Umformungen erhält man aus Gl. (5.103)

$$\begin{aligned} \hat{s}(k+1) &= \hat{s}(k) \\ &\quad - \mathbf{P}(k) \mathbf{m}(k+1) [1 + \mathbf{m}^T(k+1) \mathbf{P}(k) \mathbf{m}(k+1)]^{-1} \\ &\quad \times \mathbf{m}(k+1) \hat{s}(k) \\ &\quad + \mathbf{P}(k) \mathbf{m}(k+1) x(k+1) \\ &\quad - \mathbf{P}(k) \mathbf{m}(k+1) [1 + \mathbf{m}^T(k+1) \mathbf{P}(k) \mathbf{m}(k+1)]^{-1} \\ &\quad \times \mathbf{m}^T(k+1) \mathbf{P}(k) \mathbf{m}(k+1) x(k+1). \end{aligned} \quad (5.104)$$

Durch Ausklammern der Ausdrücke

$$\mathbf{P}(k) \mathbf{m}(k+1) x(k+1) \quad \text{und} \quad [1 + \mathbf{m}^T(k+1) \mathbf{P}(k) \mathbf{m}(k+1)]^{-1}$$

aus den letzten beiden Ausdrücken von Gl. (5.104) geht die Bestimmungsvorschrift für die Parameter in den Ausdruck

$$\begin{aligned} \hat{s}(k+1) &= \hat{s}(k) \\ &- P(k) m(k+1) [1 + m^T(k+1) P(k) m(k+1)]^{-1} m(k+1) \hat{s}(k) \\ &+ P(k) m(k+1) x(k+1) [1 + m^T(k+1) P(k) m(k+1)]^{-1} \end{aligned} \quad (5.105)$$

über. Mit den Vereinbarungen in Gl. (5.101) kann dieser Zusammenhang in die rekursive Schätzvorschrift – *rekursive Regression* – umgeschrieben werden:

$$\hat{s}(k+1) = \hat{s}(k) + k(k+1) [x(k+1) - m^T(k+1) \hat{s}(k)]. \quad (5.106)$$

Damit ergibt sich der neue Parametervektor $\hat{s}(k+1)$ des Modells aus dem alten Parametervektor $\hat{s}(k)$ plus der mit einem Bewertungsvektor $k(k+1)$ multiplizierten Differenz $x(k+1) - \hat{s}(k+1)$ von aktuellem und vorhergesagtem Ausgangssignalwert.

Die Schätzvorschrift der rekursiven Regression ist besonders für On-line-Aufgaben geeignet, da die anfallenden Beobachtungen sofort verarbeitet werden können und die rechentechnisch sehr aufwendige Inversion der Matrix $M(k)^T M(k)$ entfällt. Diesem Vorteil steht als Nachteil die Wahl geeigneter Startwerte für den Parametervektor $\hat{s}(0)$ und der Präzisionsmatrix $P(0)$ gegenüber, die die Konvergenzgeschwindigkeit der Schätzung beeinflussen. Bei Systemen mit zeitvariantem Verhalten konvergiert die Schätzung sicher, wenn

$$\det(M(k)^T M(k)) = \det(P(k)^{-1}) > 0. \quad (5.107)$$

Liegen keine A-priori-Informationen über den Prozeß vor, so ist es am zweckmäßigsten, daß auch die gesamte Information in der Matrix $[M(k)^T M(k)]$ null ist. Das bedeutet, daß alle Elemente dieser Matrix, nämlich

$$\sum_{i=1}^k m_i m_{ij}; \quad r, j = 1, 2, \dots, l,$$

zu null werden. Als Startmatrix $P(0)$ würde sich damit ergeben

$$P(0) = [M(0)^T M(0)]^{-1} = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}^{-1}. \quad (5.108)$$

Dieser Grenzfall ist nicht zu realisieren. Eine ausreichende Näherung ist gegeben, wenn als Startmatrix

$$P(0) = cI \quad (5.109)$$

mit einem hinreichend großen Wert für c gewählt wird. Umfangreiche Untersuchungen haben ergeben [1.31], daß bei einer Wahl des Faktors c im Bereich

$$10^{10} \leq c \leq 10^{15}$$

das Konvergenzverhalten im ungestörten Fall unabhängig vom Startparametervektor $\hat{s}(0)$ und bei l Parametern nach l Schritten die Schätzung konvergiert ist (s. Beispiel 5.6). Bei gestörten Systemen ist je nach der Stärke der Störung der Faktor c im Bereich

$$10^3 \leq c \leq 10^6$$

zu wählen. Dies gilt auch für Rechner mit einer geringen Genauigkeit. Bei dieser Gestaltung der Startbedingungen für die rekursive Regression ist das Verfahren in der Lage, die Modellparameter sicher an die Systemparameter anzupassen. Auch im gestörten Fall ist die rekursive Regression als eines der leistungsfähigsten Verfahren einzustufen (s. Beispiel 5.6).

Insgesamt ist damit für die rekursive Regression im Vergleich zur Relaxation und stochastischen Approximation folgende Wertung zu geben:

- wesentlich besseres Konvergenzverhalten,
- wesentlich höherer rechentechnischer Aufwand je Datengruppe.

5.4.3.2. Gewichtete rekursive Regression

Eine laufende Nachführung von Modellparametern $\hat{s}_i$ bei zeitvariantem Systemverhalten ist mit der in Gl. (5.106) angegebenen Schätzvorschrift der rekursiven Regression nicht möglich, da der Algorithmus mit Zunahme der verarbeiteten Datengruppen die Fähigkeit der Modellanpassung verliert (Bild 5.19). Die Ursache liegt im Verhalten der Elemente der Matrix $P(k)$, die mit wachsender Datenanzahl gegen Null streben. Ihr prinzipieller Verlauf ist im Bild 5.20 dargestellt. Damit werden auch die Elemente des Bewertungsvektors $k(k+1)$ zu null, und eine Korrektur des Parametervektors in Gl. (5.106) ist selbst bei einer erkannten Abweichung nicht mehr möglich.

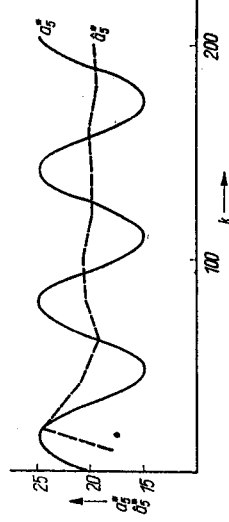


Bild 5.19. Parameternachführung der ungewichteten rekursiven Regression

$$\begin{aligned} z &= 0\%, & a^T &= [a_1(k), 2, -3, -10, a_5(k)], \\ \hat{s}(0) &= 0, & P(0) &= 10^{15}I \end{aligned}$$

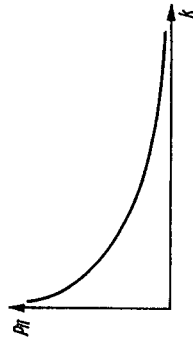


Bild 5.20. Verlauf der Elemente der Matrix P in Abhängigkeit von den verarbeiteten Datengruppen

Es besteht also die Aufgabe, das Absinken der Werte der Hauptdiagonalen der Matrix $P(k)$ zu verhindern. Als mögliche Lösungswege bieten sich an

- Division der Matrix $P(k)$ durch einen Wert c^2 ($0 < c < 1$) in jedem Schritt,

$$P^*(k) = \frac{1}{c^2} P(k), \quad (5.110)$$

- Addition einer Korrekturmatrix P^* zur Matrix $P(k)$ in jedem Schritt,

$$P^*(k) = P(k) + P^*, \quad (5.111)$$

mit

$$P^* = \begin{bmatrix} p_{11}^* & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & p_{ll}^* \end{bmatrix}; \quad p_{ii}^* > 0.$$

Beispiel 5.7 Einfluß der Änderungsgeschwindigkeit des Parameters auf den Folgefehler

Gegeben:

Systemgleichung

$$x = \sum_{i=1}^5 a_i u_i$$

$$a^T = [1, 2, -3, a_4^*, 20]$$

$$a_4^* = -30 (1 - 2 \cdot 10^{-4} k),$$

Eingangssignale $u_i \in NV(0, 1)$,

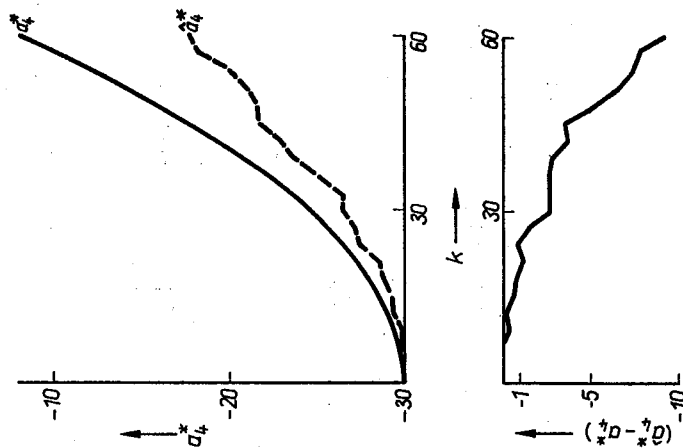
$$\text{Modellgleichung } \dot{x} = \sum_{i=1}^5 a_i u_i,$$

Wichtungsparameter $c^2 = 0,95$,Startparametervektor $\hat{a}(0) = 0$.Schätzverlauf für den zeitvarianten Parameter a_4^* .

Siehe Bild zu Beispiel 5.7.

Gesucht:

Ergebnis:



Der Schätzverlauf zeigt, daß bei einer konstanten Wichtung mit der Zunahme der Änderungsgeschwindigkeit der Folgefehler zunimmt. Damit wird sichtbar, daß im ungestörten Fall die Vergessensrate zu nehmen muß, wenn die Änderungsgeschwindigkeit der Parameter wächst.

Damit erhält man aus den Gln. (5.101) und (5.106) für den Algorithmus der gewichteten rekursiven Regression die Beziehungen

$$s(k+1) = \hat{s}(k) + k^*(k+1) [x(k+1) - m^T(k+1) \hat{s}(k)]$$

$$k^*(k+1) = P^*(k) m(k+1) [1 + m^T(k+1) P^*(k) m(k+1)]^{-1}$$

$$P^*(k) = \frac{1}{c^2} P(k), \quad P^*(k) = P(k) + P^*$$

$$P(k) = P^*(k-1) - P^*(k-1) m(k) [1 + m^T(k) P^*(k-1) m(k)]^{-1} \times m^T P^*(k-1). \quad (5.112)$$

Die Korrektur der Matrix $P(k)$ nach den angegebenen Strategien bedeutet, daß die am weitesten zurückliegenden Daten mit dem kleinsten Gewicht in die Schätzung eingehen sollen. Für die Vorgehensweise in Gl. (5.110) soll dies kurz aufgezeigt werden. Multipliziert man in jedem Tastschritt die Meßwertmatrix $M(k)$ mit der Konstanten c , so erhält man die gewichtete Matrix $M^*(k)$:

$$M^*(k) = \begin{bmatrix} c^k m_{11} & \dots & c^k m_{1n} \\ \vdots & & \vdots \\ c^k m_{1,k-1} & \dots & c^k m_{1,k-1} \\ c m_{1k} & \dots & c m_{1k} \end{bmatrix}. \quad (5.113)$$

Damit gilt für die Matrix $P^*(k)$:

$$P^*(k) = [c M(k)^T c M(k)]^{-1} = \frac{1}{c^2} P(k). \quad (5.114)$$

Die Wichtung der Datensätze kann durch sehr viele Formen vorgenommen werden [1.34, 1.31, 5.34]. Praktisch bewährt haben sich aufgrund der einfachen Realisierbarkeit und der hohen Leistungsfähigkeit die exponentielle, die blockweise und die gezielte Wichtung.

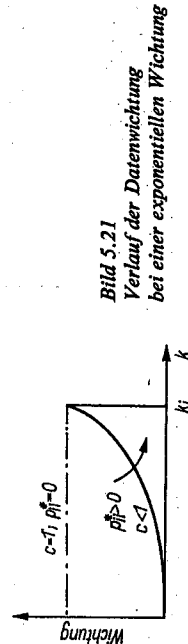


Bild 5.21
Verlauf der Datenwichtung
bei einer exponentiellen Wichtung

Exponentielle Wichtung. Wird der Wichtungsparameter c im Bereich $0 < c < 1$ bzw. werden die Werte $p_i^* > 0$ gewählt und in allen k Schritten konstant gehalten, ergibt sich der im Bild 5.21 dargestellte prinzipielle Verlauf für die Datenwichtung. Als Grenzwert ist der ungewichtete Fall (d.h. $c = 1$ oder $p_i^* = 0$) eingetragen. Je kleiner der Wert von c und je größer die Werte p_i^* sind, um so größer ist das „Vergessen“ der weiter zurückliegenden Daten und um so größer ist die Fähigkeit, Änderungen der Systemparameter zu erkennen und diesen zu folgen (s. Beispiel 5.7). Auf der anderen Seite nimmt die Fähigkeit, Störungen herauszumitteln, bei dieser Gestaltung der Wichtungsfaktoren ab. Nimmt die Wichtung der zurückliegenden Daten zu stark ab, kann der Algorithmus der rekursiven Regression instabil werden, weil zuwenig Datensätze zur Lösung des Gleichungssystems vorhanden sind. Die exponentielle Wichtung ist immer dann anzuwenden, wenn sich die Systemparameter laufend ändern oder der Zeitpunkt des Beginns der Änderung nicht bekannt ist.

Blockweise Wichtung. Eine blockweise Wichtung wird dadurch erreicht, daß für die Wichtungsparameter in Gl. (5.110) und in Gl. (5.111) gesetzt wird (Bild 5.22)

$$\begin{aligned} \text{a) } c &= \begin{cases} 0 & \text{für } k < k_i \\ 1 & \text{für } k \geq k_i, \end{cases} \\ \text{b) } p_{ii}^* &= \begin{cases} 0 & \text{für } k \neq k_i \\ 10^{15} \dots 10^6 & \text{für } k = k_i. \end{cases} \end{aligned} \quad (5.115)$$

In Gl. (5.115) bedeutet k_i den Zeitpunkt einer einmaligen Parameteränderung oder der letzten zyklischen Aktualisierung des Modells. Der prinzipielle Verlauf der Datenwichtung ist im Bild 5.22 dargestellt. Ihre Anwendung ist dann sinnvoll, wenn eine Modellanpassung zyklisch in größeren Abständen erfolgen oder nach Überschreiten von Toleranzbändern des Fehlers ausgelöst werden soll.

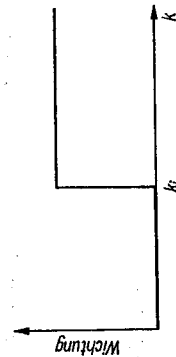


Bild 5.22. Verlauf der Datenwichtung bei einer blockweisen Wichtung

Bild 5.23. Parameterverlauf der Schätzung mit der gezielten Wichtung

$$\begin{aligned} z &= 0\%, \quad a^T = [a_1^*, a_2^*, a_3^*, -10, -20], \\ \hat{a}(0) &= 0, \quad a_1^* = 1 + 0,3k, \quad a_2^* = 2 - 0,04k, \\ a_3^* &= -3 - 0,3k, \quad P(0) = 10^{15} I, \\ p_{11}^* &= 10, \quad p_{22}^* = 0,1, \quad p_{33}^* = 1, \quad p_{44}^* = p_{55}^* = 0 \end{aligned}$$

Gezielte Wichtung. Die Datenwichtung entsprechend Gl. (5.111) ermöglicht eine sehr einfache und selektive Wahl der Wichtung und damit des Vergessens, entsprechend der Änderung jedes einzelnen Parameters. Über die Beziehung der Kovarianzmatrix für die Parameter von Gl. (5.51) existiert eine feste Zuordnung der Elemente p_{ii}^* und der Varianz der Parameter. Sind einige Parameter nicht zeitvariant, sind die Wichtungswerte p_{ii}^* null zu setzen, da dann die Gesamtvarianz minimal wird. Für die zeitvarianten Parameter sind die Wichtungsparameter p_{ii}^* in Gl. (5.111) gezielt entsprechend der Änderungsgeschwindigkeit zu wählen, wobei für hohe Änderungsgeschwindigkeiten größere Werte für die Elemente p_{ii}^* zu nehmen sind.

Dabei ist nach [1.31] der Wertebereich von $10^{-2} \leq p_{ii} \leq 10^2$ in den meisten Fällen ausreichend. Bild 5.23 zeigt aber auch, daß eine gewisse Unempfindlichkeit bei der Wahl der Wichtungsparameter besteht.

Dagegen wirken sich die gezielte und die ungezielte Wichtung enorm auf den Modellfehler aus. Im Beispiel 5.8 sind sowohl der Modellfehler G als auch die Parameterwerte der zeitvarianten Parameter in Abhängigkeit von der Wichtungsart dargestellt. Deutlich ist die Überlegenheit der gezielten Wichtung in beiden Darstellungen zu erkennen.

Beispiel 5.8

Gegeben:

Systemgleichung

$$x = \sum_{i=1}^5 a_i u_i + z,$$

$$a^T = [a_1^*, 2, -3, -10, a_5^*],$$

$$a_1^* = 1 + 0,25 \sin k/10, \quad a_5^* = 20 + 5 \sin k/10,$$

Eingangssignale $u_i \in NV(0, 1)$,

Störung $z \in NV(0, \sigma)$; 10%,

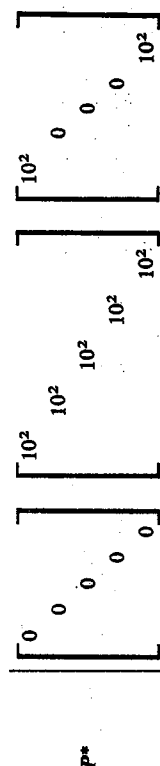
Modellgleichung $\hat{x} = \sum_{i=1}^5 \hat{a}_i u_i$,

Startparameter $\hat{a}(0) = 0$, $P(0) = 10^{15} I$.

Einfluß der Wichtungsmatrix P^* auf den Modellfehler und die Parameterwerte.

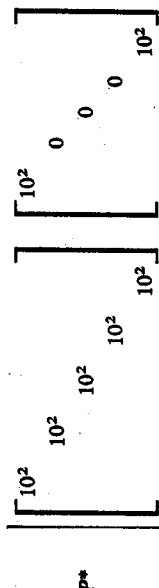
Ergebnis:

1. Modellfehler



Modellfehler G bei $k = 40$

2. Parameterwerte bei $k = 40$



d_2	1,77	1,99
d_3	-2,76	-3,00
d_4	-9,77	-10,00

Die Ergebnisse zeigen, daß die Modellgüte durch eine gezielte Wichtung wesentlich verbessert werden kann. Dies gilt besonders für die Güte der zeitvarianten Parameter.

5.4.4. Methode der rekursiven verallgemeinerten Regression

Die im Abschnitt 5.2.4 vorgestellte Methode der verallgemeinerten Regression und ihres Spezialfalls, der Markov-Schätzung, sind für eine Echtzeitparameterschätzung ungeeignet. Deshalb soll in diesem Abschnitt ein Weg zur rekursiven Lösung der Schätz Aufgabe vorgestellt werden. Den Ausgangspunkt der rekursiven Variante der Verfahren bildet die Schätzvorschrift

$$\hat{s} = [M^T W^{-1} M]^{-1} M^T W^{-1} x,$$

die unter der Annahme, daß die Matrix W eine doppelt symmetrische Gestalt besitzt, durch eine Cholesky-Zerlegung der Art

$$W = R^T R$$

in die Form

$$\hat{s} = [\hat{M}^T \hat{M}]^{-1} \hat{M}^T \hat{x} \quad (5.117)$$

übergeführt werden kann; s. Abschnitt 5.2.4, Gln. (5.68) bis (5.70). Aus Gl. (5.117) kann in Analogie zur Ableitung des Verfahrens der rekursiven Regression (s. Abschn. 5.3.3) die rekursive Form der verallgemeinerten Regression hergeleitet werden. Als Ergebnis erhält man für die Schätzvorschrift folgende Beziehungen:

$$\hat{s}(k+1) = \hat{s}(k) + \hat{k}(k+1) [\hat{x}(k+1) - \hat{m}(k+1)^T \hat{s}(k)] \quad (5.118)$$

$$\hat{k}(k+1) = P'(k) \hat{m}(k+1) [1 + \hat{m}(k+1)^T P'(k) \hat{m}(k+1)]^{-1}$$

$$P'(k+1) = P'(k) - \hat{k}(k+1) \hat{m}(k+1)^T P'(k).$$

Der rekursive Algorithmus ist in der Lage, das Modell dem System anzupassen und bei zusätzlicher entsprechender Wichtung der Daten durch eine der im Abschnitt 5.3.3.2 dargestellten Formen auch nachzuführen. Rechentechnisch hat er gegenüber der direkten Methode wesentliche Vorteile, weil eine Inversion entfällt. Problematisch bleibt die Ermittlung und die hohe Dimension der Wichtungsmatrix W .

Wird davon ausgegangen, daß die Wichtungsmatrix W durch die Kovarianzmatrix der Störung Z ersetzt wird (Markov-Schätzung), kann eine sehr günstige Strategie zur Ermittlung der Wichtungsmatrix gefunden werden [1.31, 5.31]. Wie bereits im Abschnitt 5.2.4 vorgestellt, ist es zweckmäßig, die häufig unbekannte Kovarianzmatrix der Störung Z durch die des Fehlers E zu approximieren. Der Fehler e steht bei den rekursiven Verfahren – s. Gl. (5.118) – in jedem Schritt zu Verfügung. Die Kovarianzmatrix des Fehlers hat dann folgende Gestalt:

$$E = \begin{bmatrix} \hat{R}_{ee}(0) & \dots & \hat{R}_{ee}(k) \\ \vdots & \ddots & \vdots \\ \hat{R}_{ee}(k) & \dots & \hat{R}_{ee}(0) \end{bmatrix}. \quad (5.119)$$

Die Elemente der Kovarianzmatrix werden ebenfalls rekursiv nach folgenden Beziehungen berechnet:

Schritt 1

$$\hat{R}'_{ee, k+1}(i) = \hat{R}'_{ee, k}(i) + e(k+1)e(k+1-i); \quad i = 0, 1, \dots, k. \quad (5.120)$$

Schritt 2

$$\hat{R}_{ee, k+1}(i) = \frac{\hat{R}'_{ee, k+1}(i)}{\hat{R}_{ee, k+1}(0)} \frac{k+1}{k+1-i}. \quad (5.121)$$

Als sehr erschwerend für den Echtzeitbetrieb bleibt nach wie vor die mit der Datenanzahl zunehmende Dimension der Kovarianzmatrix E des Fehlers. Wird davon ausgegangen, daß die Störung z keine periodischen Komponenten und keinen Gleichanteil enthält, werden die Elemente der Kovarianzmatrix $R_{ee}(i)$ für eine endliche Verschiebung $i_{\max}$

gegen Null streben, d. h. $R_{ee}(i_{\max}) \approx 0$. Dann geht die Matrix E in eine Bandmatrix der Form

$$E = \begin{bmatrix} \hat{R}_{ee}(0) & \dots & \hat{R}_{ee}(i_{\max}) & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \hat{R}_{ee}(i_{\max}) & \dots & \hat{R}_{ee}(0) & \dots & \hat{R}_{ee}(i_{\max}) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \hat{R}_{ee}(i_{\max}) & \dots & \hat{R}_{ee}(0) \end{bmatrix}$$

über. Damit ist die Möglichkeit gegeben, die A-priori-Information in einer Fehlerkovarianzmatrix E der Dimension $(i_{\max} + 1)(i_{\max} + 1)$ anstelle der Matrix Z der Dimension $(k \times k)$ näherungsweise zu erfassen. Der Aufbau der angenäherten Kovarianzmatrix und die Einschränkung auf die Dimension der Korrelationstiefe $i_{\max}$ führen aufgrund der Symmetrie der Matrix zu einfachen Prozeduren bei der Berechnung der für die Transformation benötigten Matrix $(R^{-1})^T$. Da die Transformationsmatrix auch die Dimension $(i_{\max} + 1)$ hat, werden zur Realisierung der Transformation – s. Gl. (5.69) – des k -ten Meßwertsatzes von m und x außer der Transformationsmatrix noch die $i_{\max}$ zurückliegenden Datensätze von x und m benötigt. Sie müssen in einem Meßwertpuffer gespeichert werden. Er arbeitet folgendermaßen:

1. Der älteste Meßwertsatz im Puffer wird durch den nächstältesten überschrieben.
2. Alle übrigen Meßwertsätze rücken nach.
3. Der Puffer wird mit dem jüngsten Meßwertsatz aufgefüllt.

Diese 3 Stufen wiederholen sich in jedem Schritt. Die Berechnung des k -ten transformierten Meßwertsatzes erfolgt entsprechend Gl. (5.69) durch die Multiplikation der $(i_{\max} + 1)$ -ten Zeile von $(R^{-1})^T$ mit den Spalten des Meßwertpuffers. Damit gilt allgemein mit $i_{\max} + 1 = j$

$$[\tilde{m}(k); \tilde{x}(k)] = \begin{bmatrix} r_{11} & r_{21} & \dots & r_{j1} \\ r_{22} & r_{32} & \dots & r_{j2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{j1} & r_{j2} & \dots & r_{jj} \end{bmatrix} \begin{bmatrix} m_{k-j+1,1} & \dots & m_{k-j+1,j} \\ \vdots & \ddots & \vdots \\ m_{k-1,1} & \dots & m_{k-1,j} \\ m_{k,1} & \dots & m_{k,j} \end{bmatrix} \begin{bmatrix} x_{k-j} \\ \vdots \\ x_{k-1} \\ x_k \end{bmatrix}. \quad (5.122)$$

Mit Gl. (5.122) wird deutlich, daß zur Realisierung der erforderlichen Transformation nur noch $i_{\max}$ zurückliegende Meßwerte anstelle der ständig wachsenden Meßwertzahl k benötigt werden. Außerdem ist zur Transformation von der Matrix $(R^{-1})^T$ nur die letzte Zeile erforderlich. Diese mit der Näherung der Kovarianzmatrix erzielten Vorteile gestatten den effektiven Einsatz der rekursiven verallgemeinerten Regression [1.31, 5.31].

Da vor allem in der Startphase durch große Parameterabweichungen des Modells von denen des Systems auch der Fehler stark von der Störung z abweicht, ist es günstig, durch eine Vorsuche mit der Regression oder der rekursiven Regression über $k \geq i_{\max}$ Beobachtungen ein Startmodell für die Parameter und die Kovarianzmatrix des Fehlers zu ermitteln. Danach erfolgt die Parameterschätzung nur mit der rekursiven verallgemeinerten Regression.

Das generelle Verhalten dieser Schätzmethode entspricht der rekursiven Regression. Bezüglich ihres Einsatzes gelten die im Abschnitt 5.2.4 zum Vergleich Regression-verallgemeinerte Regression getroffenen Aussagen.

Im Bild 5.24 ist der Schätzverlauf einiger Parameter für die rekursive Regression und für die rekursive verallgemeinerte Regression eines gestörten Systems dargestellt. Nach

relativ wenigen Beobachtungen ist die Schätzung der rekursiven verallgemeinerten Regression schon gut konvergiert; dagegen liegen noch relativ große Parameterfehler beim Verfahren der rekursiven Regression vor. Die gleiche Aussage kann für die Parameterschätzung zeitvarianter Systeme getroffen werden (Bild 5.25).

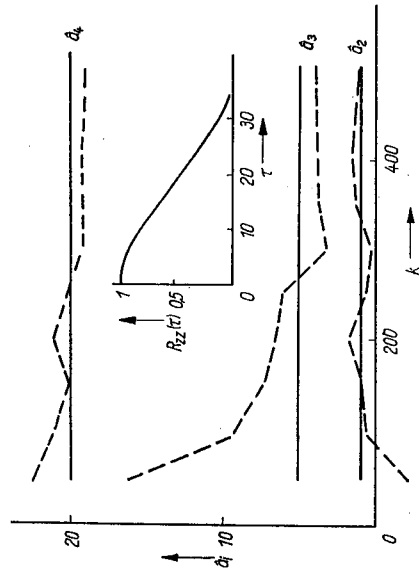


Bild 5.24
Schätzverlauf der Verfahren
rekursive Regression (---)
und rekursive verallgemeinerte
Regression (—)
 $a^* = [-3, 1, 5, 20]$,
 $z = 60\%$, stark korreliert,
 $a(0) = 0$

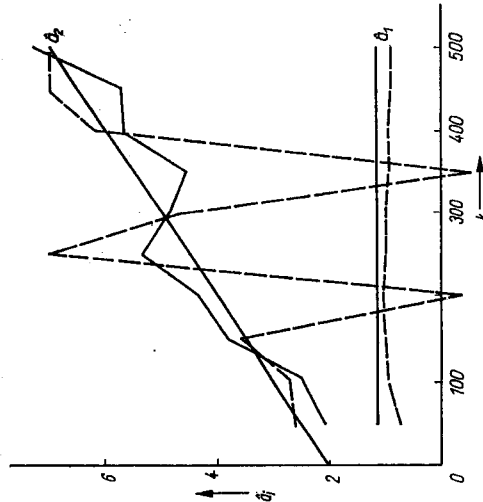


Bild 5.25
Verlauf der Parameterschätzung
der gewichteten rekursiven Regression (---)
und der gewichteten verallgemeinerten
Regression (—)
 $a^* = [1, 2 (1 + 0,005k)]$,
 $z = 60\%$, stark korreliert,
 $a(0) = 0$, $p_{22}^* = 10^3$

5.5. Methoden der optimalen Versuchsplanung

5.5.1. Zielstellung und ausgewählte Grundlagen

Die Methoden der optimalen Versuchsplanung [1.39, 1.42, 1.43, 1.48, 5.35 bis 5.42] zur Parameterschätzung statischer Systeme dienen dem Ziel, durch eine geeignete Wahl der Eingangsgrößen (Vorzeichen, Amplituden) die Güte der Schätzung weiter zu verbessern. Gleichzeitig soll durch die gezielte Planung der Experimente der Beobachtungs- und

Bearbeitungsaufwand minimiert werden. Als Optimalitätskriterien kommen somit in Frage

1. ein maximal nutzbarer Informationsgewinn,
2. eine möglichst geringe Anzahl von Experimenten,
3. eine statistische Unabhängigkeit der Parameterschätzungen,
4. ein geringer rechentechnischer Aufwand.

Es wird nun davon ausgegangen, daß als Parameterschätzverfahren die MKQ (s. Abschnitt 5.2) verwendet wird. Dieses Verfahren liefert unter bestimmten Bedingungen optimale Schätzungen für die Parameter. Wie bereits gezeigt wurde, hängt die zu erreichende Präzision der Schätzung wesentlich von der Wahl der Eingangsgrößen ab. Entsprechend den Betrachtungen im Abschnitt 5.2 ist der Zusammenhang zwischen der Modellgüte und den Werten der Eingangsgrößen durch die Beziehungen

$$\text{COV} \{\hat{s}\} = [M^T M]^{-1} \sigma_z^2 \quad (5.123)$$

und

$$\text{var} \{\hat{s}\} = \sigma_z^2 m^T [M^T M]^{-1} m \quad (5.124)$$

unmittelbar gegeben. Der Entwurf optimaler Werte der Eingangsgrößen erfolgt durch Ausnutzung einer der beiden Beziehungen u. a. mit dem Ziel, die Varianzen/Kovarianzen der Parameter und/oder Varianzen des geschätzten Ausgangs zu minimieren. Um das Problem des Entwurfs optimaler Versuchspläne lösen zu können, wird die Struktur des Systems als bekannt vorausgesetzt. Im allgemeinen wird das System hinreichend durch ein Polynom vom Grade d approximiert in der Form

$$\begin{aligned} \hat{x} &= \hat{a}_0 + \sum_{i=1}^I \hat{a}_i u_i + \sum_{i=1}^{I-1} \sum_{j=i+1}^I \hat{a}_{ij} u_i u_j \\ &\quad + \sum_{i=1}^I \hat{a}_{ii} u_i^2 + \dots + \sum_{i=1}^I \hat{a}_{ii} \dots u_i^d. \end{aligned} \quad (5.125)$$

In Vektorschreibweise und bei Verwendung der allgemeinen Systembeschreibung gilt somit

$$\hat{x} = m^T \hat{s}$$

mit

$$m^T = [1, u_1, \dots, u_I, u_1 u_2 \dots u_I^2, \dots, u_I^d]$$

$$\hat{s}^T = [\hat{a}_0, \hat{a}_1, \dots, \hat{a}_I, \hat{a}_{12} \dots \hat{a}_{II}, \dots, \hat{a}_{II} \dots \hat{a}_{II} \dots]$$

Von praktischer Bedeutung sind dabei folgende Strukturannahmen für das Modell:

- a) Polynome 1. Ordnung

$$\hat{x} = m^T \hat{s} \quad (5.126)$$

mit

$$m^T = [1, u_1, u_2, \dots, u_I]$$

$$\hat{s}^T = [\hat{a}_0, \hat{a}_1, \hat{a}_2, \dots, \hat{a}_I].$$

- b) Polynom 2. Ordnung

$$\hat{x} = m^T \hat{s}$$

mit

$$\mathbf{m}^T = [1, u_1, \dots, u_i; u_1 u_2, \dots, u_{i-1} u_i; u_1^2, \dots, u_i^2]$$

$$\mathbf{s}^T = [\hat{a}_0, \hat{a}_1, \dots, \hat{a}_i; \hat{a}_1 \hat{a}_2, \dots, \hat{a}_{i-1} \hat{a}_i; \hat{a}_{11}, \dots, \hat{a}_{ii}].$$

Ansätze höherer Ordnung werden selten benötigt und daher im Rahmen dieses Buches nicht weiter behandelt.

Bevor auf die Erstellung von Versuchsplänen für die in den Gln. (5.125) und (5.126) genannten Systemstrukturen eingegangen wird, soll auf einige Grundbegriffe der Versuchsplanung hingewiesen werden.

Unter einem Versuch wird im folgenden die Gewinnung eines Wertes der Ausgangsgröße x in einem vorgegebenen Versuchspunkt $\mathbf{u}_i$ bezeichnet. Alle Punkte $\mathbf{u}_i; i = 1, 2, \dots, n$, in denen Versuche realisiert werden können, werden zum Versuchsbereich V zusammengenommen. Die Gesamtheit aller zu realisierenden Versuche wird als Versuchsplan bezeichnet.

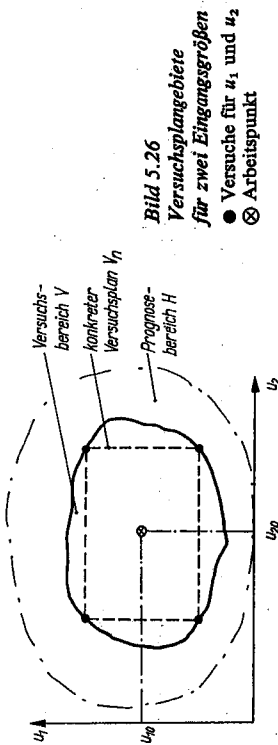


Bild 5.26
Versuchsplangebiete
für zwei Eingangsgrößen
● Versuche für u_1 und u_2
⊗ Arbeitspunkt

Im Bild 5.26 ist für zwei Eingangsgrößen ein möglicher Versuchsbereich dargestellt. Eine Auswahl von n voneinander verschiedenen Punkten $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ aus dem Bereich V , in denen je ein Versuch durchgeführt werden soll, nennt man einen konkreten Versuchsplan V_n vom Umfang n (s. Bild 5.26). Um die mit einem konkreten Versuchsplan V_n zu erzielende Parametertüte zu beschreiben, wurde in Anlehnung an die für normalverteilte Ausgangsgrößen x geltende *Fishersche Informationsmatrix*

$$\mathbf{J} = \sigma_x^2 (\mathbf{M}^T \mathbf{M}) \quad (5.127)$$

für die Schätzung $\hat{s}$ der Begriff der *Informationsmatrix* des konkreten Versuchsplans V_n eingeführt:

$$\mathbf{J}(V_n) = \frac{1}{n} \mathbf{M}^T \mathbf{M}. \quad (5.128)$$

Damit gilt für die Kovarianzmatrix der Parameter – s. Gl. (5.123) –

$$\text{COV} \{\hat{s}\} = \frac{\sigma_x^2}{n} \mathbf{J}(V_n)^{-1}. \quad (5.129)$$

Insgesamt wird deutlich, daß die für die Beurteilung der Güte des konkreten Versuchsplans verwendeten Beziehungen der Gln. (5.123), (5.124) und (5.129) nur von den Meßpunkten $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$, d.h. vom konkreten Versuchsplan V_n , abhängen. Jede Optimierung der Genauigkeit setzt damit voraus, daß eine Möglichkeit geschaffen wird, die Matrizen $\mathbf{M}^T \mathbf{M}$ oder ihre Inversen für verschiedene Versuchspläne miteinander zu vergleichen.

Die Möglichkeit ist durch die Verwendung von Funktionalen gegeben, mit denen die Matrix $\mathbf{M}^T \mathbf{M}$ auf die reelle Achse abgebildet wird. Bei n Versuchen des Planes ist die Informationsmatrix $\mathbf{J}(V_n)$ proportional zur Matrix $\mathbf{M}^T \mathbf{M}$. Die Optimierung muß sich damit über alle konkreten Versuchspläne V_n erstrecken. Die Menge der Pläne wird mit $\{V_n\}$ gekennzeichnet. Durch die Beziehung von $\mathbf{J}(V_n)^{-1}$ mit der Kovarianzmatrix $\text{COV} \{\hat{s}\}$ entsprechend Gl. (5.129) ergeben sich auch gute Möglichkeiten für eine Interpretation der Optimalitätskriterien.

Aus der Vielzahl der in der Literatur [1.39, 1.42, 1.43] vorgeschlagenen und untersuchten Kriterien werden im Rahmen dieses Buches wegen ihrer Praxisrelevanz folgende ausgewählt:

1. D-Optimalität

Ein konkreter Versuchsplan V_n^* heißt *D-optimal*, wenn die Beziehung

$$\min_{V_n \in \{V_n\}} \det (\mathbf{J}(V_n)^{-1}) = \det (\mathbf{J}(V_n^*)^{-1}) \quad (5.130)$$

erfüllt ist. Ist der Ausgangsvektor x normalverteilt, wird damit auch das Volumen des Streuungsellipsoids der Parameterschätzung $\hat{s}$ minimiert [1.42].

2. A-Optimalität

Ein konkreter Plan V_n^* heißt *A-optimal*, wenn gilt

$$\min_{V_n \in \{V_n\}} \text{Sp } \mathbf{J}(V_n)^{-1} = \text{Sp } \mathbf{J}(V_n^*)^{-1} \quad (5.132)$$

Die Elemente der Hauptdiagonalen von $\mathbf{J}(V_n)^{-1}$ sind bis auf einen konstanten Faktor die Varianzen der Schätzungen. Der *A-optimale* Versuchsplan minimiert die mittlere Varianz der Parameterschätzungen [1.42].

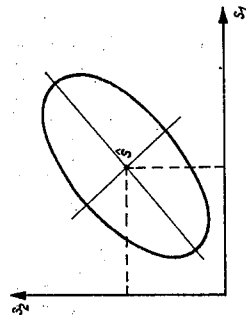


Bild 5.27
Streuungsellipsoid für die Schätzung
von zwei Parametern

Da die Spur auch gleich der Summe der Halbachsenlängen des Streuungsellipsoids ist, wird durch dieses Kriterium [1.42, 1.48] auch die mittlere Halbachsenlänge des Streuungsellipsoids der Parameterschätzungen $\hat{s}$ minimiert (Bild 5.27). Die genannten Kriterien beziehen sich auf die Schätzung des Parametervektors s und sind daher unabhängig vom vorgegebenen Prognosebereich H (Bereich, der alle $\mathbf{u}$ enthält, für die Schätzwerte $\hat{x}$ bzw. $\hat{s}$ betrachtet werden sollen; s. Bild 5.27).

Wird die Ausgangsgröße x betrachtet, spielt der Bereich H eine Rolle. Damit gelangt man zum Begriff der *G-Optimalität*.

3. G-Optimalität

Ein konkreter Versuchsplan V_n^* heißt *G-optimal*, wenn gilt [1.42]

$$\min_{V_n \in \{V_n\}} \max_{\mathbf{u} \in H} \mathbf{m}^T \mathbf{J}(V_n)^{-1} \mathbf{m} = \max_{\mathbf{u} \in H} \mathbf{m}^T \mathbf{J}(V_n^*)^{-1} \mathbf{m}. \quad (5.133)$$

Ein G -optimaler Versuchsplan minimiert den maximalen Wert der Varianzfunktion des geschätzten Ausgangs, d. h. $\text{var } \{x\}$ im Bereich H .

Aus der Literatur (s. z. B. [1.39, 1.42, 1.43]) sind eine Reihe weiterer Optimalitätskriterien für den Entwurf von konkreten Versuchsplänen bekannt, die jedoch hier nicht näher betrachtet werden sollen.

Die Menge aller möglichen Versuchspläne mit einem Umfang n ist häufig von sehr unübersichtlicher und komplizierter Struktur. Deshalb werden aus praktischen Gründen nur gewisse Teilmengen der möglichen Versuchspläne mit einfacherer Struktur verwendet. Die Auswahl erreicht man, indem man von den Plänen zusätzliche Eigenschaften fordert. Diese sind die Orthogonalität und die Drehbarkeit.

Orthogonalität. Ein konkreter Versuchsplan V_n soll *orthogonal* genannt werden, wenn seine Informationsmatrix $J(V_n)$ eine Diagonalmatrix ist. Die Zeilen von M^T sind dann paarweise orthogonale Vektoren. Die Schätzungen nach Gl. (5.47) sind für die einzelnen Koeffizienten des gewählten Ansatzes unkorreliert.

Drehbarkeit. Ein konkreter Versuchsplan V_n für einen zum Koordinatensprung symmetrischen Versuchsplan V soll *drehbar* genannt werden, wenn die Varianzfunktion $\text{var } \{x\}$ der Schätzung für x nur vom Abstand $\varrho = |x|$ des Punktes x vom Ursprung abhängt, d. h. es gilt [1.41]:

$$\text{var } \{x\} = \sigma_x^2 m^T [M^T M]^{-1} m = h(\varrho). \quad (5.134)$$

Die Realisierung der beiden genannten Eigenschaften führt zu sehr praktikablen Versuchsplänen. Die Güte der Parameterschätzung wird nicht verändert.

Der allgemeine Entwurf der Versuchspläne nach diesen Kriterien ist sehr aufwendig und der Spezialliteratur zu entnehmen (z. B. [1.42, 1.48, 5.36, 5.38, 5.43]). Im Rahmen dieses Buches wird nur der Entwurf von Versuchsplänen für Polynomansätze k -ter Ordnung, die auf den Arbeiten von Box und Hunter [5.44] beruhen, dargestellt.

Im folgenden Abschnitt werden deshalb orthogonale und drehbare Versuchspläne vorgestellt, die das Systemverhalten durch Polynomansätze 1. und 2. Ordnung – siehe Gl. (5.125) und (5.126) – annähern, weitere zusätzliche praktische Forderungen erfüllen und bezüglich eines oder mehrerer Kriterien (D -, A -, G -) optimal sind. Diese Versuchspläne werden als Mehrfaktorpläne bezeichnet.

5.5.2. Mehrfaktorpläne für lineare Modelle

5.5.2.1. Allgemeine Betrachtungen

Im Abschnitt 5.5.1 wurde ein wahres Modell vorausgesetzt. Bei einer praktischen Aufgabenstellung ist aber in den meisten Fällen die wahre Systemstruktur unbekannt. Daraus folgt, daß für diesen Fall nicht mehr die Aussage des Gauß–Markov-Theorems gilt, die MKQ sei die beste lineare erwartungstreue Schätzung. Es wird deshalb auch keine Optimierung der bisher betrachteten Kriterien durch den Versuchsplan mehr gefordert. Man versucht dagegen praktisch motivierte Forderungen an die verwendeten Versuchspläne zu erfüllen.

Wesentliche Forderungen sind [1.42, 5.44]:

1. Der Versuchsplan soll eine Schätzung des Modells – s. Gl. (5.125) und (5.126) – mit einer ausreichenden Genauigkeit im Prognosebereich H erlauben.
2. Die Güte der gefundenen Beschreibung soll über einen Test nachprüfbar sein. Ist das Modell nicht dem System adäquat, wird der Grad des Polynoms um 1 erhöht.

3. Jeder Versuchsplan soll bereits den Kern eines Versuchsplans für ein Modell nächsthöherer Ordnung bilden.

4. Die Anzahl der Versuchspunkte soll möglichst klein sein.

5. Die Versuchspläne sollen teilbar sein (Blockbildung), ohne daß wesentliche Eigenschaften verlorengehen.

Es besteht nun die Aufgabe, für das in Gl. (5.126) angenommene Modell einen l -dimensionalen Versuchsplan der Ordnung d zu entwerfen. Für den Entwurf wird angenommen, daß Versuchs- und Prognosebereich übereinstimmen sollen.

Da das allgemeinste Polynom d -ten Grades $\binom{l+d}{d}$ Koeffizienten besitzt, muß ein l -dimensionaler Versuchsplan der Ordnung d wenigstens $\binom{l+d}{d}$ Versuchspunkte enthalten, um jeden Koeffizienten schätzen zu können. Gleichzeitig muß jede einzelne Eingangsgröße – im weiteren entsprechend dem Sprachgebrauch der Theorie der optimalen Versuchsplanung mit dem Begriff „Faktor“ bezeichnet – im konkreten Versuchsplan V_n wenigstens $d + 1$ verschiedene Werte (Stufen/Niveaus) annehmen, um aus dem Plan die Koeffizienten eines Polynoms d -ten Grades ermitteln zu können.

Da stets $l \geq 2$ Einflußgrößen (Faktoren) betrachtet werden, wird in der Literatur von Mehrfaktorplänen gesprochen.

In einem derartigen Versuchsplan nehmen damit die l stetig veränderlichen Faktoren nur $\alpha_i; i = 1, \dots, l$, verschiedene feste Stufen an, die noch geeignet zu wählen sind. Durch die Festlegung der Niveaus wird der Versuchsbereich diskretisiert.

Ein Versuchsplan, der alle möglichen oder einen nach bestimmten Regeln festgelegten Teil (s. Abschn. 5.5.2.2) aller möglichen Kombinationen der endlich vielen Niveaus jedes Faktors ($l \geq 2$) enthält, wird als faktorieller Versuchsplan bezeichnet [1.42].

Bei einem Struktursatz vom Grade d in allen l Variablen muß stets $\alpha_i = d + 1; i = 1, \dots, l$, gewählt werden. Wird für alle Eingangsgrößen/Faktoren eine gleiche Anzahl von Niveaus verwendet, ist der Versuchsplan symmetrisch. Unsymmetrische Versuchspläne mit verschiedener Niveaunzahl für die einzelne Eingangsgröße werden verwendet, wenn a priori bekannt ist, daß der Struktursatz nicht für alle Eingangsgrößen gleich gewählt zu werden braucht.

5.5.2.2. Vollständige Faktorpläne

Wird das statische Verhalten des Systems durch ein Polynom 1. Ordnung – Gl. (5.126) – hinreichend beschrieben, ist die Wahl von zwei Stufen für jede Einflußgröße/Faktor hinreichend. Damit gilt $\alpha_i = 2; i = 1, 2, \dots, l$.

Für die weiteren Betrachtungen werden die Aussteuerungen der Eingangssignale/Faktoren um den Arbeitspunkt durch die Beziehung

$$u_{iN} = \frac{u_i - u_{i0}}{|u_i - u_{i0}|} \quad (5.135)$$

normiert. Unter Beachtung dieser Normierung auf die Aussteuerung um den Arbeitspunkt wird in den weiteren Ausführungen die vereinfachte Schreibweise $u_i \cong u_{iN}$ beibehalten.

Als Versuchsbereich entsteht damit ein l -dimensionaler Würfel der Art

$$V: -1 \leq u_i \leq 1; \quad i = 1, \dots, l. \quad (5.136)$$

Für $l = 2$ und $l = 3$ ist der Versuchsbereich im Bild 5.28 dargestellt.

Da die Varianz der zu schätzenden Parameter für die Niveaus ± 1 minimal wird [1.42, 5.44], werden diese bei allen Plänen verwendet.

Die Versuchspunkte mit den Niveaus $+1$ und -1 entsprechen der maximal gewählten Aussteuerung und erzwingen somit die größte Parameterempfindlichkeit und den größten Störabstand. Damit sind die Versuchspunkte die Eckpunkte des l -dimensionalen Würfels V (Bild 5.28). In den weiteren Betrachtungen werden die Niveaus $+1$ und -1 symbolisch häufig mit $+$ und $-$ bezeichnet.

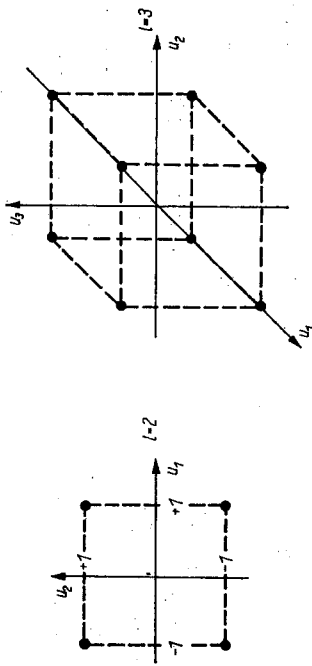


Bild 5.28
Vollständige Faktorpläne
für $l = 2$ und $l = 3$

Ein vollständiger faktorieller Versuchsplan des Typs 2^l – im weiteren durch den Ausdruck $VFV2^l$ abgekürzt – enthält somit 2^l Kombinationen der beiden Niveaus der l Einflußgrößen/Faktoren. Dabei wird die Basis des Exponentialausdrucks durch die Niveauezahl α_i und den Exponent durch die Anzahl der Faktoren l festgelegt. Somit ergibt sich die gesamte Versuchszahl n zu $n = 2^l$.

An Beispielen soll nun im weiteren der Entwurf vollständiger faktorieller Versuchspläne aufgezeigt werden.

Tafel 5.6
Vollständiger Faktorplan vom Typ 2^2

Versuchs-Nr.	u_0	u_1	u_2	$u_1 u_2$
1	+	-	-	+
2	+	+	-	-
3	+	-	+	-
4	+	+	+	+

Ausgangspunkt soll der Versuchsplan für zwei unabhängige Einflußgrößen/Faktoren ($l = 2$) sein. In Tafel 5.6 ist der $VFV2^2$ in einer für die Auswertung günstigen Form dargestellt. Dabei bilden die zweite und dritte Spalte den zu realisierenden Versuchsplan V_k . Es ist zu erkennen, daß neben den Koeffizienten des Arbeitspunktes ($\hat{a}_0$) und der Einflußgrößen ($\hat{a}_1, \hat{a}_2$) immer der Einfluß der Wechselwirkung der Faktoren u_1 und u_2 durch den Koeffizienten $\hat{a}_{12}$ ermittelt werden kann. Damit ist folgende Modellstruktur möglich:

$$\hat{x} = \hat{a}_0 u_0 + \hat{a}_1 u_1 + \hat{a}_2 u_2 + \hat{a}_{12} u_1 u_2; \quad u_0 = 1. \quad (5.137)$$

Wird ein Versuchsplan für zwei unabhängige Einflußgrößen ($l = 3$) entworfen und realisiert, kann folgendes Modell ermittelt werden:

$$\begin{aligned} \hat{x} = & \hat{a}_0 u_0 + \hat{a}_1 u_1 + \hat{a}_2 u_2 + \hat{a}_3 u_3 + \hat{a}_{12} u_1 u_2 \\ & + \hat{a}_{13} u_1 u_3 + \hat{a}_{23} u_2 u_3 + \hat{a}_{123} u_1 u_2 u_3. \end{aligned} \quad (5.138)$$

In Tafel 5.7 ist der entworfene $VFV2^3$ dargestellt, der sich als Erweiterung des $VFV2^2$ durch eine dritte Einflußgröße mit den Niveaus $+$ und $-$ ergibt.

Analog zu diesem Vorgehen werden für eine beliebige Anzahl von Einflußgrößen/Faktoren die $VFV2^l$ entworfen und realisiert.

Tafel 5.7
Vollständiger Faktorplan
vom Typ 2^3

Versuchs-Nr.	u_0	u_1	u_2	u_3	$u_1 u_2$	$u_1 u_3$	$u_2 u_3$	$u_1 u_2 u_3$
1	+	-	-	-	+	+	+	-
2	+	+	-	-	-	+	+	+
3	+	-	+	-	+	-	+	+
4	+	+	+	-	-	-	+	-
5	+	-	-	+	+	-	-	+
6	+	+	-	+	-	+	-	-
7	+	-	+	+	+	-	+	-
8	+	+	+	+	+	+	+	+

Aufgrund der Eigenschaften der $VFV2^l$ vereinfacht sich die Schätzvorschrift der Regression – s. Abschnitt 5.2, Gl. (5.47) – wesentlich, und die Güterwerte für die Parameterstreuungen werden optimal. Da die $VFV2^l$ orthogonale Versuchspläne sind, wird die Matrix $M^T M$ eine Diagonalmatrix, deren Diagonalelemente den Wert n (d.h. der Versuchszahl) besitzen. Damit ergibt sich für die in der MKQ benötigte Matrix

$$[M^T M]^{-1} = \frac{1}{n} I. \quad (5.139)$$

Unter Verwendung der Schätzvorschrift für die MKQ – s. Gl. (5.47) – folgt damit für die Schätzvorschrift der Parameter

$$\hat{\beta} = \frac{1}{n} M^T x \quad (5.140)$$

und für die Güterwerte der Parameter

$$\text{var} \{\hat{\beta}_i\} = \frac{1}{n} \sigma_x^2; \quad i = 1, 2, \dots, m$$

$$\text{cov} \{\hat{\beta}_i, \hat{\beta}_j\} = 0; \quad i \neq j, \quad i, j = 1, 2, \dots, m. \quad (5.141)$$

Damit wird deutlich, daß die Methode der Versuchsplanung garantiert, daß die Schätzung der Parameter des Modells unkorreliert ist. Als Probleme der vollständigen faktoriellen Versuchspläne sind zu nennen:

1. Die enorme Zunahme der Versuchszahl bei einer größer werdenden Anzahl von Einflußgrößen/Faktoren (z.B. $l = 6$, d.h. 64 Versuche; $l = 10$, d.h. 1024 Versuche). Einen praktikablen Ausweg stellt der Entwurf von unvollständigen Faktorplänen dar, die im Abschnitt 5.2.2.3 betrachtet werden.
2. Die häufig über zu große Zeiträume auf eines der beiden Niveaus ($+1, -1$) zu halten den Einflußgrößen bei Einhaltung der normalen Reihenfolge bedingten systematische Fehler oder gefährden das System. Der Ausweg besteht in einer zufälligen Gestaltung der Versuchsrufenfolge (der „Randomisierung“).

3. Die konsequente Fortführung des Konstruktionsprinzips für $VFV2^l$ für Systemmodelle d -ten Grades ($d \geq 2$) führt zu Plänen, die alle möglichen $(d+1)^L$ -Kombinationen der $(d+1)$ Niveaus der l Einflußgrößen/Faktoren enthalten. Ein vollständiger Versuchsplan für ein Modell 2. Grades ($d=2$) wäre somit vom Typ 3^4 . Diese Pläne werden aufgrund der sehr hohen Versuchsanzahl und ihrer ungünstigen statistischen Eigenschaften nicht verwendet. Mögliche Entwurfsstrategien für diesen Modelltyp, die die genannten Nachteile nicht besitzen, werden im Abschnitt 5.5.3 vorgestellt.

5.5.2.3. Teilfaktorpläne

Wie bereits im Abschnitt 5.5.2.2 gezeigt wurde, nimmt mit der Zunahme der Anzahl der Einflußgrößen/Faktoren die Anzahl der Versuche nach dem Gesetz $n = 2^l$ ebenfalls enorm zu. Der vollständige Versuchsplan ist erforderlich, wenn alle 2^l Parameter (d.h. das Absolutglied a_0 , die Parameter der einzelnen Einflußgrößen a_i ; $i = 1, 2, \dots, l$ und die Parameter aller Wechselwirkungen) ermittelt werden müssen. Dies ist in der Praxis selten der Fall. Auf der Grundlage von A-priori-Wissen wird deshalb immer davon ausgegangen, daß alle oder einige Wechselwirkungen nicht existieren. Damit wird es möglich, die Anzahl der notwendigen Versuche durch die Verwendung sog. teilweiser faktorieller Versuchspläne – im weiteren mit TFV bezeichnet – wesentlich zu reduzieren. Gleichzeitig sollen die TFV so gestaltet werden, daß die sehr guten statistischen Eigenschaften der VFV nicht verlorengehen.

Tafel 5.8
Teilfaktorplan vom Typ 2^{3-1}

Versuchs-Nr.	u_0	u_1	u_2	$u_3 (= u_1 u_2)$
1	+	-	-	+
2	+	+	-	-
3	+	-	+	-
4	+	+	+	+

Um das Grundprinzip des Entwurfs von TFV zu zeigen, soll von dem Beispiel in Gl. (5.138), d.h. von $l = 3$ Einflußgrößen, ausgegangen werden. Es sei a priori bekannt, daß alle Wechselwirkungen nicht existieren. Dann reduziert sich die Modellgleichung von Gl. (5.138) auf den Ausdruck

$$\hat{x} = \hat{a}_0 u_0 + \hat{a}_1 u_1 + \hat{a}_2 u_2 + \hat{a}_3 u_3. \quad (5.142)$$

Bei Verwendung des $VFV2^3$ würden die vier Parameter von Gl. (5.142) aus acht Versuchen ermittelt werden. Der Entwurf teilweiser faktorieller Versuchspläne geht nun davon aus, daß in VFV mit einer geringeren Anzahl von Einflußgrößen die Planungsniveaus der Wechselwirkungen für weitere Einflußgrößen zu verwenden sind. In dem hier gewählten Beispiel setzt man in einem $VFV2^2$ für die Wechselwirkung $u_1 u_2$ die unabhängige Einflußgröße u_3 ein. Der damit entstehende Versuchsplan für die Parameter des Modells ist, aufbauend auf den $VFV2^2$ (s. Tafel 5.6), in Tafel 5.8 dargestellt. Es ist zu erkennen, daß die Schätzung der Parameter ebenfalls entsprechend Gl. (5.140) erfolgen kann und daß sich die statistischen Gütewerte bezogen auf die Versuchsanzahl nicht verändern. Dagegen hat sich die Anzahl der Versuche bei diesem einfachen Beispiel bereits um die Hälfte verringert. Als allgemeine Bildungsvorschrift für Teilfaktorpläne $TFV2^{l-p}$ gilt folgendes:

Zur Konstruktion eines $TFV2^{l-p}$ wird von dem $(l-p)$ Einflußgrößen entsprechenden $VFV2^{l-p}$ ausgegangen. Dieser Plan wird durch p zusätzliche Einflußgrößen

ergänzt. Die zugefügten Spalten werden aus den elementweise gewonnenen Produkten von mindestens zwei und maximal $(l-p)$ Spalten des VFV gebildet. Häufig wird die Einführung des Begriffs „Generator“ vorgenommen, der angibt, welche der möglichen Produkte für die p Spalten des $TFV2^{l-p}$ verwendet wurden (in Tafel 5.8 wird das Produkt $u_1 u_2$ als Generator verwendet).

Tafel 5.9
Teilfaktorplan vom Typ 2^{4-1}

Versuchs-Nr.	u_0	u_1	u_2	u_3	$u_1 u_2$	$u_1 u_3$	$u_2 u_3$	$u_1 u_2 u_3$
1	+	-	-	-	+	+	+	-
2	+	+	-	-	-	-	+	+
3	+	-	+	-	+	+	-	+
4	+	+	+	-	-	-	-	-
5	+	-	-	+	+	-	+	+
6	+	+	-	+	-	+	-	-
7	+	-	+	+	+	-	+	+
8	+	+	+	+	+	+	+	+

Durch die Wahl verschiedener Generatoren für die zusätzlichen Einflußgrößen können verschiedene Varianten von TFV mit gleichen Eigenschaften entworfen werden. In Tafel 5.9 ist dies für ein $TFV2^{4-1}$, der auf ein $VFV2^3$ aufbaut, dargestellt. Es ist zu sehen, daß aus dem $VFV2^3$ mit den Generatoren

$$u_4 = \begin{cases} u_1 u_2 \\ u_1 u_3 \\ u_2 u_3 \\ u_1 u_2 u_3 \end{cases}$$

vier $TFV2^{4-1}$ entworfen werden können. Wie die VFV sind auch die TFV für lineare Modelle und Versuchspläne in der Form des Hyperwürfels mit den Versuchspunkten ± 1 D -, A - und G -optimal und gestatten somit die Schätzung der Parameter mit maximal möglicher Genauigkeit.

Als Problem bleibt bei den TFV der Einfluß von nicht beachteten Wechselwirkungen auf die geschätzten Parameter der Einflußgrößen. Betrachtet man die Meßwertmatrix des $TFV2^{3-1}$ in Tafel 5.8, so kann man feststellen, daß u_3 und $u_1 u_2$ identisch sind. Das bedeutet, daß dieser Plan nur die gemeinsame (vermengte) Schätzung von $\hat{a}_3$ und $\hat{a}_{12}$ ermöglicht. Der geschätzte Parameter $\hat{a}'_3$ des $TFV2^{3-1}$ ergibt sich damit zu

$$\hat{a}'_3 \rightarrow \hat{a}_3 + \hat{a}_{12}.$$

Hieraus ist zu ersehen, daß der Parameter der Einflußgröße u_3 nur im Fall der verschwindenden Wechselwirkung $u_1 u_2$ richtig geschätzt werden kann. Um das sog. Vermengungsgesetz anzugeben, das den Zusammenhang der Parameter der Einflußgrößen mit denen der Wechselwirkungen beschreibt, wurde in der Theorie der Versuchsplanung [1.42] der Begriff der definierenden Beziehung eingeführt. Sie ergibt sich aus den Produkten der Elemente der Matrix M , die die Elemente der ersten Spalte ergeben. Die Elemente der ersten Spalte sind immer 1 (Planung für u_0) und werden symbolisch mit 1 bezeichnet. Damit gilt für die definierende Beziehung des $TFV2^{3-1}$

$$I = u_1 u_2 u_3. \quad (5.143)$$

Aus der Gl. (5.143) ergeben sich durch Multiplikation mit den Einflußgrößen u_i ; $i = 1, 2, 3$, und bei Beachtung, daß $u_i^2 = 1$ ist, folgende Vermengungen:

$$\begin{aligned} u_1 &= u_2 u_3, & u_2 &= u_1 u_3, & u_3 &= u_1 u_2. \\ \hat{a}_0'' &\rightarrow \hat{a}_0 + \hat{a}_{123}, & \hat{a}_1' &\rightarrow \hat{a}_1 + \hat{a}_{23} \\ \hat{a}_2' &\rightarrow \hat{a}_2 + \hat{a}_{13}, & \hat{a}_3' &\rightarrow \hat{a}_3 + \hat{a}_{12}. \end{aligned} \quad (5.144)$$

Um die Wechselwirkungen erkennen und korrigieren zu können, wird ein alternativer $\text{TFV}2^{3-1}$ mit der definierenden Beziehung

$$I = -u_1 u_2 u_3$$

realisiert, der folgende vermengte Schätzwerte liefert:

$$\begin{aligned} \hat{a}_0'' &\rightarrow \hat{a}_0 - \hat{a}_{123}, & \hat{a}_1'' &\rightarrow \hat{a}_1 - \hat{a}_{23} \\ \hat{a}_2'' &\rightarrow \hat{a}_2 - \hat{a}_{13}, & \hat{a}_3'' &\rightarrow \hat{a}_3 - \hat{a}_{12}. \end{aligned} \quad (5.145)$$

Damit können aus beiden $\text{TFV}2^{3-1}$ die unvermengten Parameter der Einflußgrößen nach der Vorschrift

$$\hat{a}_i = \frac{\hat{a}_i' + \hat{a}_i''}{2}; \quad i = 0, 1, 2, 3,$$

und die der Wechselwirkungen nach der Vorschrift

$$\begin{aligned} \hat{a}_{123} &= \frac{\hat{a}_0' - \hat{a}_0''}{2}, & \hat{a}_{12} &= \frac{\hat{a}_3' - \hat{a}_3''}{2}, \\ \hat{a}_{13} &= \frac{\hat{a}_2' - \hat{a}_2''}{2}, & \hat{a}_{23} &= \frac{\hat{a}_1' - \hat{a}_1''}{2} \end{aligned}$$

ermittelt werden. Entsprechend dieser Strategie wird auch bei Versuchsplänen mit vier und mehr Einflußgrößen vorgegangen [1.42, 1.48, 5.39].

Von besonderem praktischem Interesse sind Versuchspläne, bei denen alle Wechselwirkungen der Einflußgrößen vernachlässigt und die Planungsniveaus durch neue Einflußgrößen belegt werden. Diese Versuchspläne, die als gesättigte Pläne bezeichnet werden, ermöglichen die Schätzung der Parameter aus $n = l + 1$ Versuchen. So ist es z. B. möglich, auf der Grundlage der $\text{VfV}2^3$ mit acht Versuchen die Parameter von sieben Einflußgrößen/Faktoren zu schätzen (Tafel 5.10). Der $\text{VfV}2^7$ würde dagegen 128 Versuche

Tafel 5.10. Gesättigter Teilfaktorplan für $l = 7$

Versuchs-Nr.	u_0	u_1	u_2	u_3	$u_4 (= u_1 u_2)$	$u_5 (= u_1 u_3)$	$u_6 (= u_2 u_3)$	$u_7 (= u_1 u_2 u_3)$
1	+	+	-	-	+	+	+	-
2	+	+	-	-	-	-	+	+
3	+	+	+	-	-	+	-	+
4	+	+	+	-	+	-	-	-
5	+	+	-	+	+	-	-	+
6	+	+	-	+	-	+	-	-
7	+	+	+	+	-	+	+	-
8	+	+	+	+	+	+	+	+

umfassen. Auf der Grundlage theoretischer Untersuchungen ist es *Plackett* und *Burman* [5.46] gelungen, für den Entwurf der gesättigten orthogonalen Versuchspläne eine einfach zu realisierende Strategie zu entwerfen, wenn die Anzahl der Versuche n ein Vielfaches von 4 ist.

In ihren Arbeiten sind Pläne von $n = 4$ bis $n = 100$ (außer 92) angegeben, die nach einer einfachen Vorschrift aus den in Tafel 5.11 angegebenen ersten Zeilen konstruiert werden können.

Tafel 5.11. Die erste Zeile der gesättigten Versuchspläne für den Entwurf nach *Plackett-Burman* für $n = 8$ bis $n = 84$ Versuche

n	1. Zeile
8	+
12	+
16	+
20	+
24	+
32	-
36	-
44	+
48	+
60	+
68	+
72	+
80	+
84	+

Die Versuchsmatrix M wird aus den in Tafel 5.11 angegebenen Zeilen nach folgender Vorschrift aufgebaut:

Ausgehend von der gewählten ersten Zeile wird die zweite Zeile durch Verschieben aller Elemente um eine Stelle nach rechts und Einsetzen des letzten Elements der ersten Zeile an die erste Stelle gewonnen. Diese Vorgehensweise wird bis zur $(n - 1)$ -ten Zeile zyklisch wiederholt. Die Versuchsmatrix M wird komplettiert, indem noch

1. eine n -te Zeile mit negativen Niveaus,
2. eine erste Spalte mit positiven Niveaus

Tafel 5.12. Gesättigter Versuchsplan für $l = 11$

Versuchs- Nr.	u_0	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}
1	+	+	+	-	+	+	+	-	-	-	+	-
2	+	-	+	+	-	+	+	+	-	-	-	+
3	+	+	-	+	+	-	+	+	+	-	-	-
4	+	-	+	-	+	+	-	+	+	+	-	-
5	+	-	-	+	+	+	+	-	+	+	+	-
6	+	+	-	-	+	+	+	+	+	+	+	+
7	+	+	+	-	-	+	+	-	+	+	+	+
8	+	+	+	-	-	+	+	-	+	+	+	+
9	+	+	+	+	-	-	-	+	+	+	+	-
10	+	+	-	+	+	+	-	-	+	+	+	+
11	+	+	+	+	+	+	-	-	+	+	+	+
12	+	-	-	-	-	-	-	-	-	-	-	-

hinzugefügt wird. In Tafel 5.12 ist diese Entwurfsstrategie für einen gesättigten Plan mit $l = 11$ Einflußgrößen dargestellt.

Bei der Realisierung aller vorgestellten Versuchspläne ist es zweckmäßig, in den n Versuchspunkten n' jeweils ϑ Versuche ($\vartheta = 2, 3, \dots$) zu realisieren, um aus den erhaltenen Ausgangsgrößen und den geschätzten Parametern Schätzwerte für die Streuung der Störung σ_z^2 und die Reststreuung σ_R^2 zu erhalten.

Beide Größen dienen als Ausgangspunkte für den Test auf Adäquatheit der gewählten Modellstruktur. Ist die Reststreuung groß gegenüber der Streuung der Störung, kann davon ausgegangen werden, daß die gewählte Modellstruktur nicht richtig ist.

Den Schätzwert für die Streuung der Störung erhält man aus den in den n Punkten des Versuchsplans zusätzlich durchgeführten Versuchen aus der Beziehung

$$s_z^2 = \frac{1}{n(\vartheta - 1)} \sum_{i=1}^n \sum_{j=1}^{\vartheta} (x^{ij} - \bar{x}^i)^2 \quad \text{mit} \quad \bar{x}^i = \frac{1}{\vartheta} \sum_{j=1}^{\vartheta} x^{ij} \quad (5.146)$$

Als Schätzwert für die Reststreuung σ_R^2 wird der Ausdruck

$$s_R^2 = \frac{1}{n - l - 1} \sum_{i=1}^n (x^i - \bar{x}^i)^2 \quad \text{mit} \quad \bar{x}^i = m^T \hat{a} \quad (5.147)$$

angesehen; s. auch Gl. (5.47).

Für den Vergleich der Reststreuung mit der Streuung der Störung wird der F -Test (s. Abschn. 2.3.3) in der Form

$$F = \frac{s_R^2}{s_z^2} \quad (5.148)$$

mit den Freiheitsgraden

$$f_1 = n - l - 1, \quad f_2 = n(\vartheta - 1)$$

verwendet. Ist der in Gl. (5.148) berechnete Wert kleiner als der für ein gewähltes Signifikanzniveau aus Tafel A.5 entnommene Wert F_{α, f_1, f_2} , wird davon ausgegangen, daß die gewählte Modellstruktur der vorhandenen Systemstruktur mit einer Wahrscheinlichkeit von $P = 1 - \alpha$ entspricht. Wird die Hypothese verworfen, ist z. B. von einer Modellstruktur 1. Grades zu der 2. Grades überzugehen und der Test zu wiederholen.

Im Beispiel 5.9 ist die Vorgehensweise für verschiedene Modellansätze dargestellt.

Beispiel 5.9 Ermittlung und Test von Modellen 1. und 2. Ordnung auf der Grundlage optimaler Versuchspläne

Gegeben:

Systemgleichung

$$x = a_0 + a_1 u_1 + a_2 u_2 + a_{12} u_1 u_2 + a_{11} u_1^2 + a_{22} u_2^2 + z,$$

$$a_0 = 2, \quad a_1 = a_2 = a_{12} = 1, \quad a_{11} = a_{22} = 0,5,$$

Störung $z \in NV(0, \dots)$.

Gesucht: Die Struktur und die Parameter des Systems auf der Grundlage der Methoden der optimalen Versuchsplanung.

Ergebnis:

Schritt 1

- Modellannahme

$$\hat{x} = \hat{a}_0 + \hat{a}_1 u_1 + \hat{a}_2 u_2$$

- Entwurf und Realisierung eines Versuchsplans vom Typ 2²

u_0	u_1	u_2	x^1
+1	-1	-1	2,05
+1	+1	-1	1,95
+1	-1	+1	2,025
+1	+1	+1	5,975

- Schätzung der Parameter nach Gl. (5.140)

$$\hat{a}^T = [3; 0,9625; 10],$$

- Test auf Adäquatheit der gewählten Modellstruktur (F -Test).

1. Die Berechnung der Reststreuung nach Gl. (5.146) ergibt $s_R^2 = 4,1$.

2. Die Berechnung der Streuung der Störung auf der Grundlage eines zweiten Versuchs ($\vartheta = 2$) in jedem Versuchspunkt

u_0	u_1	u_2	x^1	x^2
+1	-1	-1	2,025	1,975
+1	+1	-1	1,95	1,98
+1	-1	+1	2,05	2,02
+1	+1	+1	5,975	6,025

entsprechend Gl. (5.145) ergibt den Wert

$$s_z^2 = 8,5 \cdot 10^{-4}$$

3. Die Überprüfung der Adäquatheit des Modellansatzes erfolgt mit der Hypothese $H_0: \sigma_R^2 = \sigma_z^2$ für die Freiheitsgrade

$$f_z = n(\vartheta - 1) = 4, \quad f_R = n - l - 1 = 1$$

und den Tafelwert für ein Risiko von 5%

$$F_{0,5,4,1} = 7,71$$

(s. Anhang Tafel A.4). Aus dem Vergleich des berechneten F -Wertes entsprechend Gl. (5.147)

$$F = \frac{4,1}{0,00085} = 4823,5$$

mit dem Tafelwert $F_{0,5,4,1}$ wird ersichtlich, daß die Hypothese mit 95%iger Sicherheit abzulehnen ist. Damit ist der gewählte Ansatz zu verwerfen und die Modellordnung zu erhöhen.

Beispiel 5.9	Ermittlung und Test von Modellen 1. und 2. Ordnung auf der Grundlage optimaler Versuchspläne
-----------------	---

Schritt 2 (s. Abschn. 5.5.3)

– Neue Modellannahme

$$\hat{x} = \hat{a}_0 + \hat{a}_1 u_1 + \hat{a}_2 u_2 + \hat{a}_{12} u_1 u_2 + \hat{a}_{11} u_1^2 + \hat{a}_{22} u_2^2$$

– Entwurf und Realisierung eines orthogonal zentral zusammengesetzten Versuchsplans 2. Ordnung (Planungskern vom Typ 2²)

u_0	u_1	u_2	$u_1 u_2$	u_1^2	u_2^2	x
+1	-1	-1	+1	+1 - f	+1 - f	2,05
+1	+1	-1	-1	+1 - f	+1 - f	1,95
+1	-1	+1	-1	+1 - f	+1 - f	2,025
+1	+1	+1	+1	+1 - f	+1 - f	5,975
+1	+α	0	0	α ² - f	0 - f	3,53
+1	-α	0	0	α ² - f	0 - f	1,47
+1	0	+α	0	0 - f	α ² - f	3,46
+1	0	-α	0	0 - f	α ² - f	1,54
+1	0	0	0	0 - f	0 - f	2,05

mit α = 1 aus Gl. (5.154)

$$f = 0,66 \text{ aus Gl. (5.152)}$$

– Berechnung der Parameter des Modellansatzes mit Gl. (5.140) und Gl. (5.155) für $\hat{a}_0$ mit dem Ergebnis

$$\hat{a}_0^T = [2,02; 0,985; 0,986; 1,0125; 0,485; 0,485]$$

– Erneuter Test auf Adäquatheit der Modellstruktur mit dem F-Test

1. Die Berechnung der Reststreuung ergibt

$$s_R^2 = 3,2 \cdot 10^{-3}$$

2. Die Überprüfung der Hypothese $H_0: \sigma_R^2 = \sigma_f^2$ für die Freiheitsgrade

$$f_z = n(\theta - 1) = 3 \quad (\text{s. Schritt 1})$$

$$f_R = n - l - 1 = 3$$

und den Tafelwert für ein Risiko von 5%

$$F_{R, f_z, f_R} = 6,59$$

(s. Anhang Tafel A.4). Für den berechneten F-Wert erhält man

$$F = \frac{s_R^2}{s_f^2} = \frac{0,0032}{0,00085} = 3,76$$

Da in diesem Fall $F < F_{R, f_z, f_R}$ ist, wird die Hypothese angenommen. Dies bedeutet, daß kein Unterschied zwischen der Reststreuung und der Streuung der Störung mit 95%iger Sicherheit existiert. Damit ist die gewählte Modellstruktur der des Systems adäquat.

5.5.3. Mehrfaktorpläne für nichtlineare Modelle 2. Ordnung

5.5.3.1. Problemstellung

Ist die Beschreibung des statischen Verhaltens eines Systems durch ein Polynom 1. Ordnung – s. Gl. (5.126) – nicht hinreichend, muß eine Beschreibung mit einem Polynom höherer Ordnung versucht werden. Von praktischer Bedeutung und in sehr vielen Fällen ausreichend ist die Verwendung von Polynomen 2. Ordnung – s. Gl. (5.126). Auf diesen Modelltyp beschränken sich die weiteren Betrachtungen. Bei einem Modell 2. Ord-

nung sind mindestens drei Versuchspunkte für jede der l unabhängigen Eingangsgrößen erforderlich. Das bedeutet mindestens $n = 3^l$ Versuchspunkte. Im Bild 5.29 ist für eine Eingangsgröße u_1 , die Aussteuerung um den Arbeitspunkt dargestellt. Wird eine Normierung entsprechend Gl. (5.135) vorgenommen, erhält man als normierte Versuchspunkte die Werte $-1, 0, +1$. Für zwei Eingangsgrößen wird die Aussteuerung entsprechend diesen normierten Werten im Bild 5.30 gezeigt.

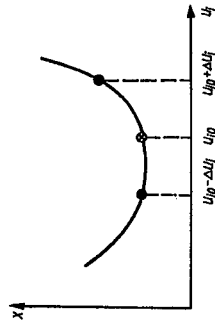


Bild 5.29. Aussteuerung einer nichtlinearen Kennlinie um den Arbeitspunkt

● Versuche um den Arbeitspunkt

⊗ Arbeitspunkt = Nullpunktversuch

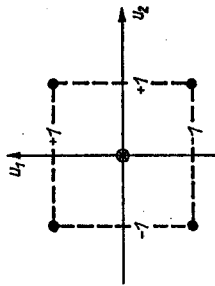


Bild 5.30. Versuchsplan des Typs 3²

Wird ein Versuchsplan nach diesen Prinzipien entworfen, erhält man z. B. für den Modellansatz

$$\hat{x} = \hat{a}_0 + \hat{a}_1 u_1 + \hat{a}_2 u_2 + \hat{a}_3 u_3 + \hat{a}_{11} u_1^2 + \hat{a}_{22} u_2^2 + \hat{a}_{33} u_3^2 \quad (5.149)$$

den in Tafel 5.13 angegebenen Versuchsplan. Eine Berechnung der Matrix $M^T M$ für diesen Plan ergibt

$$M^T M \neq cI. \quad (5.150)$$

Aus der nichtdiagonalen Matrix von Gl. (5.150) folgt, daß ein derartig gestalteter Versuchsplan folgende Eigenschaften hat:

- keine unabhängige Schätzung der Parameter (d. h. keine Orthogonalität)
- Varianz der Parameter nicht unabhängig von der Lage des Planes (d. h. keine Drehbarkeit des Planes).

Die genannten Nachteile dieser Entwurfsstrategie werden dadurch beseitigt, daß Versuchspläne entworfen werden, die orthogonal (orthogonale zentrale Versuchspläne) bzw. drehbar (drehbare zentrale Versuchspläne) sind. Diese erlauben unabhängige Parameterschätzungen mit einer von der Lage des Versuchsplans unabhängigen minimalen Varianz der Parameter.

Tafel 5.13

Versuchsplan vom Typ 3²
für das Modell von Gl. (5.149)

Versuchs-Nr.	u_0	u_1	u_2	u_3	u_1^2	u_2^2	u_3^2
1	1	1	1	1	1	1	1
2	1	1	1	0	1	1	0
3	1	1	1	-1	1	1	1
4	1	1	0	1	1	0	1
5	1	1	0	0	1	0	0
6	1	1	0	-1	1	0	1
7	1	1	-1	1	1	1	1
8	1	1	-1	0	1	1	0
9	1	1	-1	-1	1	1	1
10	1	0	1	1	0	1	1
11	1	0	1	0	0	1	0
12	1	0	1	-1	0	1	1
13	1	0	0	1	0	0	1
14	1	0	0	0	0	0	0
15	1	0	0	-1	0	0	1
16	1	-1	1	1	1	1	1
17	1	-1	1	0	1	1	0
18	1	-1	1	-1	1	1	1
19	1	-1	0	1	1	0	1
20	1	-1	0	0	1	0	0
21	1	-1	0	-1	1	0	1
22	1	-1	-1	1	1	1	1
23	1	-1	-1	0	1	1	0
24	1	-1	-1	-1	1	1	1
25	0	1	1	1	0	1	1
26	0	1	1	0	0	1	0
27	0	1	1	-1	0	1	1

5.5.3.2. Orthogonale zentrale Versuchspläne

Wie im Abschnitt 5.5.3.1 gezeigt wurde, liefern Versuchspläne vom Typ $n = 3^l$ nicht die gewünschten statistischen Eigenschaften. Deshalb wurden Pläne entworfen, die für jede Einflußgröße fünf Versuchspunkte vorsehen, mit denen die Bedingung der Orthogonalität eingehalten werden kann [5.44, 5.45, 5.47]. Der Versuchsplan wird aus folgenden Teilen zusammengesetzt:

- ein Planungskern vom Typ $2^l (2^{l-p})$
- 2^l Sternpunktversuche
- n_0 Nullpunktversuche ($n_0 = 1$).

Damit ergibt sich die Gesamtversuchszahl zu $n = 2^l + 2^{l+1}$. Im Bild 5.31 ist die Lage der einzelnen Komponenten des Planes und des Gesamtplans für zwei Einflußgrößen dargestellt.

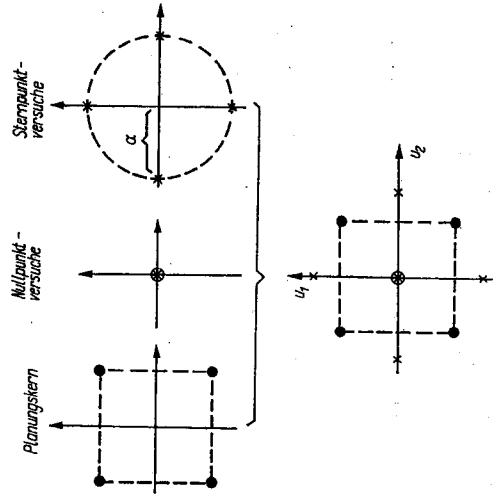


Bild 5.31
Entwurf von zusammengesetzten
zentralen Versuchsplänen
● Versuche des Planungskerns
⊙ Nullpunktversuche
× Sternpunktversuche

Die angestrebte Orthogonalität wird über die Normierung aller quadratischen Größen u_i^2 ; $i = 1, 2, \dots, l$, nach der Vorschrift

$$u_{iN}^2 = u_i^2 - f \quad (5.151)$$

mit

$$f = \overline{u_i^2} = \sum_{j=1}^n u_j^2 / n$$

und durch die Festlegung der Niveaus $\pm \alpha$ erzwingen. Für den in Gl. (5.149) angegebenen Ansatz ergibt sich damit der in Tafel 5.14 dargestellte zusammengesetzte Versuchsplan. Es ist ersichtlich, daß bei Verwendung von Planungskernen vom Typ 2^l der Wert für die Größe f allgemein aus der Beziehung

$$f = \frac{2^l + 2\alpha^2}{2^l + 2\alpha + 1} \quad (5.152)$$

bestimmbar ist. Der Wert für das Planungsniveau kann aus der Forderung abgeleitet werden, daß für orthogonale Versuchspläne die Summe der Produkte der Elemente

Tafel 5.14. Zusammengesetzter Versuchsplan mit einem Planungskern vom Typ 2^l für das Modell von Gl. (5.149)

Versuchs-Nr.	u_0	u_1	u_2	u_{1N}^2	u_{2N}^2	Planungskern
1	1	1	1	1 - f	1 - f	1 - f
2	1	1	-1	1 - f	1 - f	1 - f
3	1	-1	1	1 - f	1 - f	1 - f
4	1	-1	-1	1 - f	1 - f	1 - f
5	1	α	0	$\alpha^2 - f$	0 - f	Sternpunktversuche
6	1	$-\alpha$	0	$\alpha^2 - f$	0 - f	
7	1	0	α	0 - f	$\alpha^2 - f$	
8	1	0	$-\alpha$	0 - f	$\alpha^2 - f$	
9	1	0	0	0 - f	0 - f	Nullpunktversuch

zweier Spalten der quadrierten Größen u_{iN}^2 und z_{kN}^2 null sein muß:

$$\sum_{j=1}^n u_{iN,j}^2 u_{kN,j}^2 \stackrel{!}{=} 0. \quad (5.153)$$

Aus Gl. (5.153) erhält man damit den Ausdruck (s. auch Tafel 5.14)

$$2^n (1 - f)^2 - 2lf(\alpha^2 - f) + f^2 = 0.$$

Durch Einsetzen von Gl. (5.152) und einige Umformungen erhält man für einen Plan vom Typ 2^l das Planungsniveau der Sternpunktversuche zu

$$\alpha = \sqrt{\frac{1}{2} [\sqrt{2^l (2^l + 2l + 1)} - 2^l]}. \quad (5.154)$$

Ist der Planungskern vom Typ 2^{l-p} , so ist in Gl. (5.154) der Typ 2^l durch 2^{l-p} zu ersetzen. In Tafel 5.15 sind die Planungsniveauswerte als Funktion der Anzahl der Eingangsgrößen entsprechend Gl. (5.154) für Versuchsplanungskerne vom Typ 2^l und 2^{l-p} angegeben. Es ist bei der orthogonalen zentralen Versuchsplanung zu beachten, daß sich durch die Normierung der quadratischen Größen der richtige Schätzwert für den Parameter des Arbeitspunktes aus der Beziehung

$$\hat{a}_0 = \hat{a}'_0 - \hat{a}_{11}u_1^2 - \hat{a}_{22}u_2^2 - \dots \quad (5.155)$$

ergibt; s. Gl. (5.151). Die Erstellung und Auswertung dieser Versuchspläne sind im Beispiel 5.9 (s. Abschn. 5.2.3) dargestellt.

Die vorgestellten zusammengesetzten Pläne sind orthogonal; die Varianz der geschätzten Parameter ist aber noch abhängig von der Lage des Planes.

Tafel 5.15
Planungsniveaus α
der Sternpunktversuche
für orthogonale zusammen-
gesetzte Versuchspläne

Typ 2^l	2^2	2^3	2^4	2^5	2^6	2^7	2^8
α	1,00	1,215	1,414	1,596	1,761	1,909	2,045
Typ 2^{l-p}	2^{5-1}	2^{6-1}	2^{7-1}	2^{8-1}	2^{9-1}	2^{10-1}	2^{11-1}
α	1,547	1,724	1,885	2,029			

5.5.3.3. Drehbare zentrale Versuchspläne

Durch das notwendige Hinzufügen der Sternpunktversuche zu den Versuchsplänen vom Typ 2^l oder 2^{l-p} geht die Drehbarkeit des normalen zusammengesetzten Planes verloren. Untersuchungen von *Box* und *Hunter* [5.44] haben ergeben, daß durch eine geeignete Wahl des Niveaus $\pm\alpha$ die Eigenschaft der Drehbarkeit auch bei zusammengesetzten Plänen erreicht werden kann. Wenn der Gesamtplan wieder aus

- dem Planungskern vom Typ 2^l (2^{l-p})
- den $2l$ Sternpunktversuchen
- n_0 Nullpunktversuchen; $u_0 = f(l)$

besteht, kann aus Bedingungen für die Drehbarkeit [1.39, 1.42, 1.43, 5.39] das Niveau der Sternpunktversuche nach Vorschrift

$$\alpha = \sqrt[4]{2^l} \quad \text{bei Kern vom Typ } 2^l$$

$$\alpha = \sqrt[4]{2^{l-p}} \quad \text{bei Kern vom Typ } 2^{l-p} \quad (5.156)$$

bestimmt werden. Die Anzahl der Nullpunktversuche n_0 wird am besten dadurch bestimmt, daß die Determinante der Informationsmatrix Gl. (5.128) maximiert wird [1.39]. In Tafel 5.16 sind für drehbare zusammengesetzte Versuchspläne die Planungsniveaus und die Anzahl der Nullpunktversuche angegeben. Für das in Gl. (1.49) gewählte Modell

Typ 2^l	2^2	2^3	2^4	2^5	2^6	2^7	2^8
α	1,414	1,682	2,00	2,378	2,828	3,364	4,00
n_0	8	9	12	17	24	35	52
Typ 2^{l-p}	2^{5-1}	2^{6-1}	2^{7-1}	2^{8-1}			
α	2,00	2,378	2,828	3,364			
n_0	10	15	22	33			

Tafel 5.16
Planungsniveaus α
der Sternpunktversuche
für drehbare orthogonale
zusammengesetzte
Versuchspläne

Versuchs-Nr.	u_0	u_1	u_2	u_1^2	u_2^2
1	1	1	1	1	1
2	1	1	-1	1	1
3	1	-1	1	1	1
4	1	-1	-1	1	1
5	1	1,414	0	1,999	0
6	1	-1,414	0	1,999	0
7	1	0	1,414	0	1,999
8	1	0	-1,414	0	1,999
9	1	0	0	0	0
10	1	0	0	0	0
:	:	:	:	:	:
16	1	0	0	0	0

Tafel 5.17
Drehbarer orthogonaler
Versuchsplan für ein Modell
mit zwei Eingangsgrößen
(Gl. (5.149))

ist in Tafel 5.17 der entsprechende Versuchsplan dieser Strategie dargestellt (s. auch Tafel 5.14).

Ist die Zahl der Einflußgrößen l hoch, wächst die Anzahl der Versuche stark gegenüber der Anzahl der zu schätzenden Parameter an. Wird auf die exakte Einhaltung der Bedingungen für die Orthogonalität und die Drehbarkeit verzichtet, können nach *Hartley* [5.48] Versuchspläne entworfen werden, die bei einer wesentlich geringeren Versuchszahl für viele Aufgaben der Praxis noch hinreichend genaue Modelle liefern.

5.5.3.4. Vereinfachte zentrale Versuchspläne

Von *Hartley* [5.48] wurde auf dem Grundprinzip des Entwurfs zentral zusammengesetzter Pläne untersucht, welche Typen von Teilfaktorplänen $TFV 2^{l-p}$ noch ausreichend sind, um eine nichtsinguläre Matrix M für die Versuchsplanung 2. Ordnung zu erhalten. Es gelang ihm, Pläne auf der Basis von $TFV 2^{l-p}$ als Kernversuche zu entwerfen, die unter der Bedingung, daß von den Wechselwirkungen nur die paarweisen Beziehungen beachtet werden ($u_i u_j$; $i \neq j$), die Ermittlung der Parameter

$$\hat{a}_0, \hat{a}_i, \hat{a}_{ii}; \quad i = 1, 2, \dots, l$$

$$\hat{a}_{ij}; \quad i, j = 1, \dots, l, \quad i \neq j$$

gestatten. Damit wird als allgemeines Systemmodell das in Gl. (5.126) angenommene Polynom 2. Ordnung verwendet. Die allgemeine Struktur dieser zusammengesetzten zentralen Pläne ist in Tafel 5.18 dargestellt. Neben den üblichen $2l$ Sternpunktversuchen wird in den Plänen von *Hartley* nur ein Nullpunktversuch ($n_0 = 1$) verwendet. Für den Entwurf der Pläne, deren Planungskern 2^{l-p} so gewählt werden muß, daß alle paarweisen Wechselwirkungen unverfälscht geschätzt werden können, ergeben sich für die Anzahl l der Einflußgrößen die in Tafel 5.19 angegebenen TFW als Kern des zusammengesetzten

Tafel 5.18. Allgemeine Form der Versuchsplanmatrix M für zusammengesetzte Pläne

Versuchsplan-komponenten	u_0	$u_1 \dots u_l$	$u_1^2 \dots u_l^2$	$u_1 u_2 \dots u_{l-1} u_l$
Planungskern Typ 2^{l-p}	1	1 ... 1	1 ... 1	1 ... 1
	1	-1 ... 1	1 ... 1	-1 ... 1
	:	:	:	:
Sternpunktversuche $2l$	1	$\alpha \dots 0$	$\alpha^2 \dots 0$	0 ... 0
	1	$-\alpha \dots 0$	$\alpha^2 \dots 0$	0 ... 0
	:	0 ... 0	0 ... 0	:
Nullpunktversuch	1	0 ... 0	0 ... 0	0 ... 0

Tafel 5.19
Versuchsplankerne für
vereinfachte zusammen-
gesetzte Pläne nach *Hartley*

Anzahl der Einflußgrößen l	3	4	5	6	7	8
Versuchsplan- typ 2^{l-p}	2^{3-1}	2^{4-1}	2^{5-1}	2^{6-2}	2^{7-2}	2^{8-2}
Versuchszahl n	11	17	27	29	47	81

Versuchsplans. Für drei Einflußgrößen ist der zu realisierende zentrale zusammengesetzte Plan 2. Ordnung nach *Hartley* in Tafel 5.20 entworfen worden. Aus dem Plan wird die Meßwertmatrix *M* entsprechend der Modellstruktur von Gl. (5.126) entworfen. Für die Wahl des Planungsniveaus der 2I Sternpunktversuche gibt es keine allgemeine Festlegung. Weit verbreitet ist die Berechnung der Werte aus Gl. (5.154) oder Gl. (5.156).

Tafel 5.20
Zusammengesetzter zentraler
Versuchsplan für drei Einflußgrößen
nach *Hartley*

Versuchs- Nr.	u_0	u_1	u_2	u_3
1	1	1	1	1
2	1	1	-1	-1
3	1	-1	1	-1
4	1	-1	-1	1
5	1	1,147	0	0
6	1	-1,147	0	0
7	1	0	1,147	0
8	1	0	-1,147	0
9	1	0	0	1,147
10	1	0	0	-1,147
11	1	0	0	0

Werden TFFV verwendet, die nicht auf die Zweierpotenz aufbauen (sie werden nicht-reguläre TFFV genannt), kann die Anzahl der Versuche noch weiter gesenkt werden, ohne daß wesentliche Güteverluste auftreten [5.49].

Insgesamt wird durch diese Versuchsplanstrategien die Anzahl der Experimente vor allem bei vielen Einflußgrößen stark reduziert (z.B. für $l = 6$ von 45 auf 29, für $l = 7$ von 79 auf 47).

5.6. Übungsaufgaben zum Abschnitt 5

Aufgabe 5.1

An einem System im stationären Zustand wurden die in Tafel 5.21 dargestellten Werte der Eingangs- und der Ausgangsgröße gemessen. Schätzen Sie mit dem Verfahren der Regression die Parameter $\hat{d}_i$ folgender Modellansätze:

a) $\hat{x} = \hat{d}_0 + \hat{d}_1 u$

b) $\hat{x} = \hat{d}_0 + \hat{d}_1 u + \hat{d}_2 u^2$

Ermitteln Sie für die Modellansätze die Restquadratsumme, die signifikanten Parameter und die Konfidenzintervalle für die Parameter mit einer Sicherheit von 95%! Interpretieren Sie das Gesamtergebnis, wenn Sie davon ausgehen, daß die Systemgleichung

$$x = 1 + 4u + z$$

ist!

Tafel 5.21
Meßwerte zu Aufgabe 5.1

Nr.	u	x
1	1,33	6,475
2	-0,32	-0,990
3	-0,60	-1,705
4	1,18	6,500
5	-1,09	-3,415
6	-0,49	-0,840

Aufgabe 5.2

Auf ein System wirken zwei Eingangsgrößen. Zur Ermittlung des Modellansatzes

$$\hat{x} = \hat{d}_1 u_1 + \hat{d}_2 u_2 + \hat{d}_3 u_2^2$$

wurden die in Tafel 5.22 dargestellten Werte der beiden Eingangsgrößen und der Ausgangsgröße gemessen. Schätzen Sie die Parameter des Modells, und ermitteln Sie ihre Varianzen!

Tafel 5.22
Meßwerte zu Aufgabe 5.2

Nr.	u_1	u_2	x
1	1	2	16,2
2	3	1	10,5
3	2	1	8,0
4	1	2	15,5
5	3	1	9,9
6	3	2	19,9

Aufgabe 5.3

An einem System wurden im stationären Zustand

- a) durch Beobachtung der Eingangs- und Ausgangsgrößen die in Tafel 5.23 dargestellten Werte,
b) durch die Realisierung eines vollständigen Versuchsplans die in Tafel 5.24 dargestellten Werte gewonnen!

Berechnen Sie die Schätzwerte für die Parameter sowie die Varianzen und Kovarianzen der Parameter für den Fall, daß die Streuung der Störung $\sigma_e^2 = 1$ ist!

Als Regressionsmodell wird gewählt

$$\hat{x} = \hat{d}_1 u_1 + \hat{d}_2 u_2$$

Tafel 5.23
Meßwerte zu Aufgabe 5.3a

Nr.	u_1	u_2	x
1	-1,5	-2	-9,7
2	-0,5	0	-0,1
3	0,5	0	2,0
4	1,5	2	7,8

Tafel 5.24
Meßwerte zu Aufgabe 5.3b

Nr.	u_1	u_2	x
1	-1	-1	-4
2	1	-1	-2
3	-1	1	0
4	1	1	6

Aufgabe 5.4

Gegeben sind die in Tafel 5.25 dargestellten Ergebnisse eines realisierten aktiven Experiments.

- a) Bestimmen Sie die Werte und die Varianzen der Parameter folgenden Modellansatzes:

$$\hat{x} = \hat{d}_0 + \hat{d}_1 u_1 + \hat{d}_2 u_2$$

Tafel 5.25. Meßwerte zu Aufgabe 5.4

Nr.	1	2	3	4	5	6	7	8
u_1	1	-1	1	-1	1	-1	1	-1
u_2	1	1	-1	-1	1	1	-1	-1
x	13,6	5,1	-3,5	-17,5	9,9	12,3	-9,3	6,9

b) Ermitteln Sie durch Anwendung des F-Tests, ob das Modell dem System mit einer Sicherheit von 95% äquivalent ist!
Zur Ermittlung der Streuung der Störung wurden die in Tafel 5.26 gegebenen Nullpunktversuche realisiert.

Tafel 5.26
Mehrwerte der Nullpunktversuche
zu Aufgabe 5.4

Nr.	x	Nr.	x
1	1,6	7	-1,3
2	-4,9	8	9,5
3	4,5	9	5,2
4	-7,5	10	-3,1
5	-2,1	11	-4,0
6	2,3	12	3,1

Aufgabe 5.5

Für eine Gewächshaussteuerung wird ein Modell benötigt, das den Zusammenhang zwischen der Innentemperatur des Gewächshauses (x), der Stellung der Lüftungsklappen (u_1), der Ventilstellung der Stehwandheizung (u_2) und dem Zustand der Heizlüfter (u_3) im stationären Zustand beschreibt.
Zu diesem Zweck wurde der in Tafel 5.27 angegebene Versuchsplan realisiert.

Tafel 5.27. Meßwerte zu Aufgabe 5.5

Nr.	Stellung Lüftungs- klappen	Betrieb Stehwand- heizung	Betrieb Heizlüfter	Luft- temperatur °C
1	+	-	+	20,32
2	+	-	-	18,60
3	+	+	+	22,04
4	+	+	-	19,20
5	-	-	+	24,60
6	-	-	-	20,45
7	-	+	+	26,59
8	-	+	-	23,64

Die Planungsniveaus bedeuten

- + maximaler Öffnungswinkel der Lüftungsklappen,
Ventil zur Dampfvorsorgung der Stehwandheizung voll geöffnet,
Heizlüfter in Betrieb;
- Lüftungsklappen geschlossen,
Heizungsventil geschlossen,
Heizlüfter außer Betrieb.

Schätzen Sie die Parameter eines linearen Modells ohne und mit den Wechselwirkungen $u_1 u_2$ und $u_1 u_3$!

6. Bestimmung des dynamischen Verhaltens gestörter Systeme

6.1. Zielstellung

Wie bereits im Abschnitt 1 gezeigt wurde, ist die Ermittlung von Modellen dynamischer Systeme zur Lösung vieler Aufgaben bei der Diagnose, Überwachung, Steuerung und/oder Vorhersage von Mengen- und Qualitätsparametern in technischen und nichttechnischen Prozessen erforderlich. In diesem Abschnitt sollen in Ergänzung der vorgestellten Verfahren zur Modellbestimmung (s. Abschn. 4) Identifikationsmethoden betrachtet werden, die eine Ermittlung dynamischer Modelle bei folgenden Systemeigenschaften gestatten:

1. Das System kann gestört sein.
2. Das Systemverhalten kann zeitvariant sein.
3. Das Systemverhalten ist linear, oder es kann durch einfache nichtlineare Strukturen beschrieben werden.
4. Das System ist ein Eingrößensystem oder ein Mehrgrößensystem mit p -kanonischer Struktur.
5. Das System besitzt konzentrierte Parameter.

Als Ein-/Ausgangs-Modelle werden das nichtparametrische Volterra-Reihen-Modell (Abschn. 6.2) und das parametrische Differenzgleichungsmodell (Abschn. 6.3) verwendet. Im Abschnitt 6.5 wird auf die Schätzung des Zustands mit Hilfe der sog. Kalman-Filter eingegangen. Die zweckmäßige Wahl des Modells hängt wesentlich von der späteren Verwendung und dem konkreten praktischen Einsatzfall ab. Eine allgemeine Richtlinie zur Modellauswahl kann nicht gegeben werden. Zur Ableitung der Schätzverfahren für die genannten Modelle werden in diesem Abschnitt aus folgenden Gesichtspunkten diskrete Beschreibungen für die kontinuierlichen Systeme bevorzugt:

1. Die notwendige rechen-technische Verarbeitung der gemessenen Eingangs- und Ausgangssignale erfolgt diskontinuierlich.
2. Der Reglerentwurf und die Simulationsverfahren verwenden häufig Modelle in diskreter Beschreibung.
3. Die Parameterschätzverfahren sind in diskreter Form einfacher und effektiver zu realisieren.

Aus diesen Gründen werden in den nachfolgenden Abschnitten die Signale überwiegend als abgetastete Signale betrachtet. Bei der Darstellung der Schätzverfahren wird ferner immer davon ausgegangen, daß das dynamische Modell des Systems in die allgemeine Struktur (s. Abschn. 1)

$$\dot{x} = m^T \hat{s} \tag{6.1}$$

übergeführt wird, um dadurch die Anwendung einiger bereits im Abschnitt 5 für die Parameterschätzung statischer Modelle hergeleiteter Schätzmethoden zu ermöglichen. Neben der Wahl der Modellbeschreibung und des Schätzverfahrens hat die Gestaltung

der Eingangssignale, d.h. der Entwurf optimaler Testsignale, eine große Bedeutung bei der Identifikation dynamischer Systeme. Auf dieses Problem wird im Abschnitt 6.4 näher eingegangen.

Auf einige weitere Probleme der Schätzung dynamischer Systeme wird im Abschnitt 6.6 hingewiesen.

6.2. Schätzung der Stützstellen der Kerne der Volterra-Reihe

6.2.1. Systembeschreibung durch die Volterra-Reihe

Bei der Beschreibung des Ein-/Ausgangs-Verhaltens eines Eingrößensystems durch eine Volterra-Reihe wird davon ausgegangen, daß die ungestörte Ausgangsgröße des Systems $x^0(t)$ zum Zeitpunkt $t = kT$ durch folgenden Zusammenhang charakterisiert ist:

$$x^0(kT) = f\{u(kT), u((k-1)T), \dots, u((k-n)T)\}. \quad (6.2)$$

Wird nun nach einem Vorschlag von Volterra [6.1] Gl. (6.2) in eine Taylor-Reihe um den Punkt $u(kT) = u((k-1)T) = \dots = u((k-n)T) = 0$ entwickelt, erhält man mit der verkürzten Schreibweise $x(kT) \triangleq x(k)$ usw. [1.46]:

$$\begin{aligned} x^0(k) \approx & f(0) + \left\{ u(k) \frac{\partial f(0)}{\partial u(k)} + \dots + u(k-n) \frac{\partial f(0)}{\partial u(k-n)} \right\} \\ & + \frac{1}{2} \left\{ u^2(k) \frac{\partial^2 f(0)}{\partial u^2(k)} + u(k) u(k-1) \frac{\partial^2 f(0)}{\partial u(k) \partial u(k-1)} + \dots \right. \\ & \left. + u^2(k-n) \frac{\partial^2 f(0)}{\partial u^2(k-n)} + \dots \right\} \end{aligned} \quad (6.3)$$

Wenn für $f(0) = f(0, 0, \dots, 0) = 0$ gesetzt wird, gilt für Gl. (6.3)

$$\begin{aligned} x^0(k) = & \{c_0 u(k) + \dots + c_n u(k-n)\} \\ & + \{c_{00} u^2(k) + c_{01} u(k) u(k-1) + \dots + c_{nn} u^2(k-n)\} + \dots \end{aligned} \quad (6.4)$$

und in Summenform

$$x^0(k) = \sum_{i=0}^n c_i u(k-i) + \sum_{i,j=0}^n c_{ij} u(k-i) u(k-j) + \dots$$

Ersetzt man die Parameter

$$c_i = g(iT) T$$

$$c_{ij} = g(iT, jT) T^2$$

und führt zur Verbesserung der Systembeschreibung die Grenzübergänge $T \rightarrow 0$ und $n \rightarrow \infty$ in Gl. (6.4) durch, gilt für das ungestörte Systemausgangssignal $x^0(t)$

$$\begin{aligned} x^0(t) = & \int_0^\infty g(\tau_1) u(t - \tau_1) d\tau_1 \\ & + \int_0^\infty \int_0^\infty g(\tau_1, \tau_2) u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2 + \dots \end{aligned} \quad (6.5)$$

Die in Gl. (6.5) dargestellte Beziehung wird als *Volterra-Reihe* bezeichnet. Die Zeitfunktionen $g(\tau_1), g(\tau_1, \tau_2), \dots$ nennt man Kerne der Volterra-Reihe. Sie stellen eine nichtparametrische Beschreibung eines nichtlinearen dynamischen Systems dar.

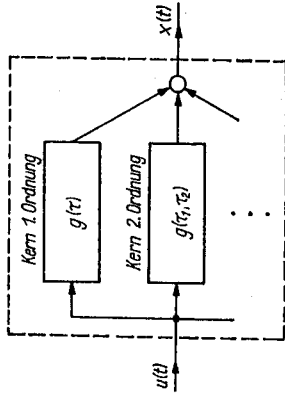


Bild 6.1
Beschreibung eines dynamischen Systems
durch eine Volterra-Reihe

Die Gesamtschreibung des Systems durch eine Volterra-Reihe besteht damit aus der Parallelanordnung eines linearen Teilsystems und nichtlinearer Teilsysteme beliebiger Ordnung (Bild 6.1). Jedes der Teilsysteme ist durch einen Kern der entsprechenden Ordnung gekennzeichnet. Die Schätzaufgabe besteht darin, Stützstellenwerte der Kerne bzw. Parameter sog. Basisfunktionen aus dem gemessenen Eingangs- und Ausgangssignal $u(t)$ und $x(t)$ zu ermitteln. Da die Beschreibung der Kerne durch Basisfunktionen keine praktische Bedeutung erlangt hat, wird sie nicht behandelt; der Leser sei auf die Literatur verwiesen [6.2, 6.3, 6.4]. Dem Vorteil der Allgemeingültigkeit der Beschreibung steht der Nachteil der großen Anzahl der zu ermittelnden Stützstellenwerte der Kerne gegenüber. Aus diesem Grund wird die Volterra-Reihe von Gl. (6.5) häufig nach dem quadratischen Glied (2. Ordnung) abgebrochen. Die Volterra-Reihe findet bevorzugt Anwendung bei der Lösung von Vorhersageproblemen und bei der Erstellung dynamischer Verhaltensmodelle in der Medizin, Biologie, Ökologie und Wasserwirtschaft. Weitere Vorteile der Beschreibung sind in der fehlenden Strukturbestimmung und der guten Anwendbarkeit von Methoden der optimalen Versuchsplanung zu sehen.

Als Sonderfall ist in Gl. (6.5) die lineare Systembeschreibung, das Faltungsintegral, enthalten. Für das gestörte System gilt dann

$$x(t) = \int_0^t g(\tau) u(t - \tau) d\tau + z(t). \quad (6.6)$$

Der Kern 1. Ordnung, die Gewichtsfunktion $g(\tau)$, wird durch eine Auswahl ihrer Ordnaten (Stützstellenwerte)

$$g(iT); \quad i = 0, 1, \dots, m \quad (6.7)$$

beschrieben. Die Schätzung der Stützstellenwerte setzt voraus, daß der Ausgang des Systems $x(t)$ mit einer Abtastzeit T zu diskreten Zeitpunkten gemessen wird und das Eingangssignal $u(t)$ im Zeitintervall T konstant sowie für $t < 0$ gleich null ist. Aus Gl. (6.6) folgt analog zu den Betrachtungen im Abschnitt 4.2.3 und in Anlehnung an die in [1.37, 3.1, 3.2] gewählte Schreibweise:

$$x(k) = \sum_{j=0}^{\infty} u(k-j) \int_{jT}^{(j+1)T} g(\tau) d\tau + z(k). \quad (6.8)$$

Setzt man in Gl. (6.8) als Näherung

$$\int_{JT}^{(J+1)T} g(\tau) d\tau = g(JT) T$$

ein, ergibt sich mit $g(JT) \doteq g(j)$

$$x(k) = T \sum_{j=0}^{\infty} g(j) u(k-j) + z(k). \quad (6.9)$$

In Gl. (6.9) ist die Zahl der Eingangswerte, die den Ausgang $x(k)$ beeinflussen, nicht konstant.

Beschränkt man sich jedoch auf Systeme mit Ausgleich, für die $g(\tau) \approx 0$ bei $\tau > T_s$ (T_s Übergangszeit des Systems), und wählt eine Beobachtungszeit $T_b = (m+1)T \approx T_s$, so führt ein Austausch der oberen Summationsgrenze in Gl. (6.9) von $k-1$ auf m zur Beziehung

$$x(k) = T \sum_{j=0}^m g(j) u(k-j) + z(k). \quad (6.10)$$

Wählt man als Beschreibung des Systems eine Volterra-Reihe 2. Ordnung – s. Gl. (6.5) –, werden die Kerne $g(\tau_1)$ und $g(\tau_1, \tau_2)$ durch eine Auswahl ihrer Ordinaten

$$g(iT), g(jT); \quad i, j = 0, 1, 2, \dots, m \quad (6.11)$$

gekennzeichnet.

Unter den bereits für das Faltungsintegral formulierten Bedingungen folgt aus Gl. (6.5) somit

$$x(k) = \sum_{j=0}^m u(k-j) \int_{JT}^{(j+1)T} g(\tau) d\tau + \sum_{i=0}^m \sum_{j=1}^m u(k-j) u(k-i) \int_{JT}^{(j+1)T} \int_{iT}^{(i+1)T} g(\tau_1, \tau_2) d\tau_1 d\tau_2 + z(k). \quad (6.12)$$

Setzt man in Gl. (6.12) für

$$\begin{aligned} \int_{JT}^{(j+1)T} g(\tau) d\tau &= g(jT) T, \\ \int_{JT}^{(j+1)T} \int_{iT}^{(i+1)T} g(\tau_1, \tau_2) d\tau_1 d\tau_2 &= g(iT, jT) T^2 \end{aligned}$$

ein, ergibt sich die Beziehung

$$x(k) = T \sum_{j=0}^m g(j) u(k-j) + T^2 \sum_{i=0}^m \sum_{j=1}^m g(i, j) u(k-j) u(k-i) + z(k). \quad (6.13)$$

Eine Vereinfachung der bisher abgeleiteten Modelle ist dadurch möglich, daß das nichtlineare dynamische Gesamtsystem als Reihenschaltung eines nichtlinearen statischen und

eines linearen dynamischen Übertragungsglieds dargestellt wird. Die Beschreibung des nichtlinearen Teilsystems erfolgt dabei gewöhnlich in Form eines Polynoms der Ordnung p

$$x_{NL}(k) = \sum_{i=0}^p r_i u_{NL}^i(k), \quad (6.14)$$

während das lineare Teilsystem durch die Gewichtsfunktion $g(\tau)$ dargestellt wird. In Abhängigkeit von der Reihenfolge der beiden Teilsysteme unterscheidet man das Wiener-Modell und das Hammerstein-Modell.

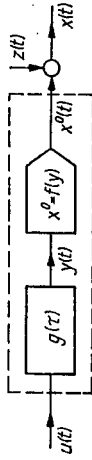


Bild 6.2. Struktur des Wiener-Modells (Gewichtsfunktionsmodell)

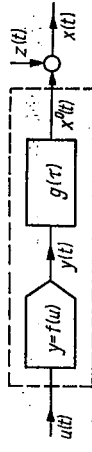


Bild 6.3. Struktur des Hammerstein-Modells (Gewichtsfunktionsmodell)

Das Wiener-Modell hat die im Bild 6.2 dargestellte Struktur. Daraus resultiert ein Zusammenhang zwischen Eingangs- und Ausgangssignal in der Form

$$x(t) = \sum_{i=0}^p r_i \left(\int_0^t g(\tau) u(t-\tau) d\tau \right)^i + z(t). \quad (6.15)$$

Unter den bereits für das Faltungsintegral angegebenen Diskretisierungsbedingungen ergibt sich aus Gl. (6.15)

$$x(k) = \sum_{i=0}^p r_i \left\{ \sum_{j=0}^m u(k-j) \int_{JT}^{(j+1)T} g(\tau) d\tau \right\}^i + z(k).$$

Setzt man wiederum für

$$\int_{JT}^{(j+1)T} g(\tau) d\tau = g(jT) T$$

ein, ergibt sich

$$x(k) = \sum_{i=0}^p r_i \left\{ T \sum_{j=0}^m g(j) u(k-j) \right\}^i + z(k). \quad (6.16)$$

Beschränkt man sich auf Nichtlinearitäten 2. Ordnung, gelangt man aus Gl. (6.16) zu

$$\begin{aligned} x(k) &= r_0 + r_1 T \sum_{j=0}^m g(j) u(k-j) \\ &\quad + r_2 T^2 \sum_{i=0}^m \sum_{j=1}^m g(i) g(j) u(k-i) u(k-j) + z(k). \end{aligned} \quad (6.17)$$

Das Hammerstein-Modell hat die im Bild 6.3 dargestellte Struktur. Daraus ergibt sich für den Zusammenhang zwischen dem Eingangs- und Ausgangssignal

$$x(t) = \sum_{i=0}^p r_i \int_0^t g(\tau) [u(t-\tau)]^i d\tau + z(k). \quad (6.18)$$

Unter den für alle bisher behandelten Modelle festgelegten Bedingungen ergibt sich aus Gl. (6.18)

$$x(k) = \sum_{i=0}^p r_i \sum_{j=0}^m u^i(k-j) \int_T^{(j+1)T} g(\tau) d\tau + z(k). \quad (6.19)$$

Setzt man in Gl. (6.19) die Beziehung

$$\int_T^{(j+1)T} g(\tau) d\tau = g(jT) T$$

ein, ergibt sich für das abgetastete Ausgangssignal

$$x(k) = \sum_{i=0}^p r_i T \sum_{j=0}^m u^i(k-j) g(j) + z(k). \quad (6.20)$$

Beschränkt man sich auf Nichtlinearitäten 2. Ordnung, gelangt man aus Gl. (6.20) zu

$$\begin{aligned} x(k) = & r_0 T \sum_{j=0}^m g(j) + r_1 T \sum_{j=0}^m u(k-j) g(j) \\ & + r_2 T \sum_{j=0}^m u^2(k-j) g(j) + z(k). \end{aligned} \quad (6.21)$$

Damit sind die wesentlichen Beschreibungsformen von Systemen durch Volterra-Reihen vorgestellt. Sie zeichnen sich alle dadurch aus, daß die zu ermittelnden Kennwerte, d. h. die Stützstellen der Kerne, linear mit dem Ausgangssignal zusammenhängen. Dadurch ist die Anwendung der im Abschnitt 5 vorgestellten Schätzverfahren der Regression möglich und in vielen Fällen ausreichend.

6.2.2. Schätzung der Stützstellenwerte der Gewichtsfunktion

6.2.2.1. Entfaltung mit Ausgleich bei linearen Systemen

Für lineare dynamische Eingrößensysteme wurde im Abschnitt 6.2.1 aus dem Faltungsintegral eine Beschreibung des Ein-/Ausgangs-Verhaltens bei Verwendung der Stützstellenwerte der Gewichtsfunktion entsprechend Gl. (6.9) abgeleitet:

$$x(k) = T \sum_{j=0}^m g(j) u(k-j) + z(k). \quad (6.22)$$

Gl. (6.22) läßt sich in eine vektorielle Form überführen, die die Anwendung des Regressionsverfahrens erleichtert. Zum k -ten Zeitpunkt gilt

$$x(k) = T m^T(k) s + z(k) \quad (6.23)$$

mit

$$m^T(k) = [u(k), u(k-1), \dots, u(k-m)]$$

$$s^T = [g(0), g(1), \dots, g(m)].$$

Für $n (n \geq m+1)$ Beobachtungen erhält man

$$x(k) = T m s + z(k) \quad (6.24)$$

mit

$$\begin{aligned} x^T(k) &= [x(k), x(k+1), \dots, x(k+n)] \\ M &= \begin{bmatrix} u(k) & \dots & u(k-m) \\ \vdots & & \vdots \\ u(k+n) & \dots & u(k+n-m) \end{bmatrix}. \end{aligned}$$

Weist die Störung $z(t)$ die Eigenschaften des normalverteilten „weißen Rauschens“ auf (s. Abschn. 5.2.2), liefert die Methode der kleinsten Quadrate eine erwartungstreue lineare Schätzung für die Stützstellen der Gewichtsfunktion der Form

$$\hat{s} = \frac{1}{T} [M^T M]^{-1} M^T x \quad (6.25)$$

mit

$$\hat{s}^T = [\hat{g}(0), \hat{g}(1), \dots, \hat{g}(m)].$$

Die Schätzung des Ausgangs ergibt sich zu

$$\hat{x}(k) = T m^T(k) \hat{s}. \quad (6.26)$$

Die Ermittlung statistischer Gütewerte für die Stützstellenwerte der Gewichtsfunktion erfolgt analog zu den Gleichungen im Abschnitt 5.2.2.

Neben der Methode der Regression wurden zur Schätzung der Stützstellenwerte der Gewichtsfunktion die rekursive Regression [1.31] und die stochastische Approximation [6.5, 6.6] verwendet. Da die Systemstruktur von Gl. (6.23) der des statischen Falles sehr ähnlich ist, zeigen auch die rekursiven Schätzverfahren die gleichen Konvergenzergebnisse, wie im Abschnitt 5.4.3 erläutert. Als rekursives Schätzverfahren hat sich die rekursive Regression für zeitvariante und zeitinvariante Systeme als das leistungsstärkste erwiesen [1.31].

Im Beispiel 6.1 ist die Ermittlung der Werte der zeitdiskreten Gewichtsfunktion eines gestörten linearen Systems mittels Regressionsverfahren dargestellt.

6.2.2.2. Korrelationsverfahren

Zur Beschreibung des statistischen Zusammenhangs zwischen dem Ausgangssignal $x(t)$ und dem Eingangssignal $u(t)$ eines gestörten linearen dynamischen Systems kann die Kreuzkorrelationsfunktion verwendet werden. Es gilt entsprechend Abschnitt 3

$$R_{ux}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T u(t) x(t+\tau) dt. \quad (6.27)$$

Berücksichtigt man in dieser Gleichung, daß der Zusammenhang zwischen Eingangs- und Ausgangssignal durch das Faltungsintegral entsprechend Gl. (6.28)

$$x(t) = \int_0^\infty g(\tau) u(t-\tau) d\tau + z(t) \quad (6.28)$$

bestimmt ist, ergibt sich unter der Voraussetzung, daß das Störsignal $z(t)$ nicht mit dem Eingangssignal $u(t)$ korreliert ist, die Beziehung zwischen der Gewichtsfunktion und der Kreuzkorrelationsfunktion in der bekannten Form (s. Abschn. 3.2.4) zu

$$R_{ux}(\tau) = \int_0^\infty g(t) R_{uu}(\tau-t) dt. \quad (6.29)$$

Gegeben: Systembeschreibung

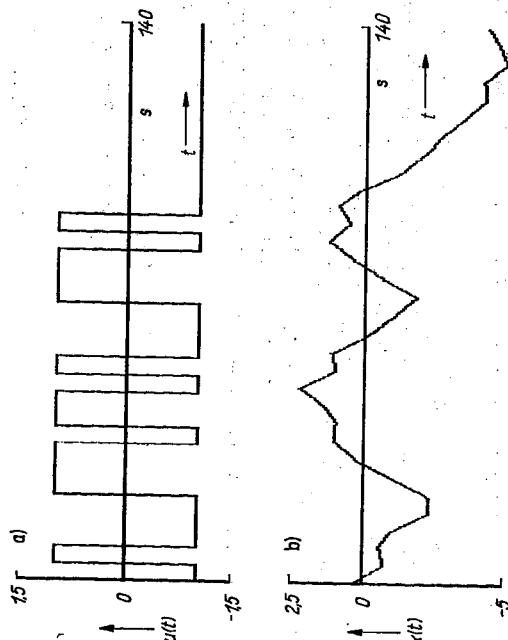
$$G(p) = \frac{5}{(1 + p \cdot 10 s)(1 + p \cdot 5 s)} \quad \text{bzw.} \quad g(\tau) = [e^{-\tau/10} - e^{-\tau/5}] s^{-1},$$

Abtastzeit $T = 4 s$,
Testsignal: Folge aus Plackett-Burman-Plan ($m + 1 = 11$, Amplitude ± 1 ,
Gesamtlänge 32T),

Störung $NV(0, \sigma_g^2)$; hochfrequent, Störverhältnis $\sqrt{\sigma_g^2/\sigma_x^2} = 0,13$.

Gesucht: Schätzwerte für die Stützstellen der Gewichtsfunktion und die Reststreuung s_R^2 .

Ergebnis: Aus den Meßwerten des Eingangs- und Ausgangssignals (Bild a und b)

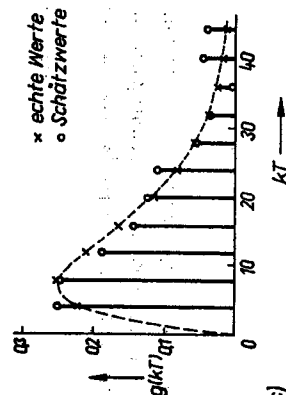


wurden folgende Stützstellenwerte geschätzt:

i	0	1	2	3	4	5
$\hat{g}(i)$	0,22099	0,24743	0,21048	0,16113	0,11701	0,08249
$\hat{g}(i)$	0,2502	0,2467	0,1873	0,1427	0,1242	0,1085
i	6	7	8	9	10	
$\hat{g}(i)$	0,05711	0,03910	0,02658	0,01793	0,01213	
$\hat{g}(i)$	0,0568	0,0346	0,0002	0,0454	0,0412	

Reststreuung $s_R^2 = 0,1105$.

Ein Vergleich der exakten und geschätzten Stützstellen zeigt, daß das Modell relativ gut ermittelt wird (Bild c).



Unter den bereits für das Faltungsintegral im Abschnitt 6.2.1 verwendeten Diskretisierungsbedingungen folgt aus Gl. (6.29) für den diskreten Fall

$$R_{ux}(m) = T \sum_{j=0}^{\infty} g(j) R_{uu}(m-j). \quad (6.30)$$

Beschränkt man sich auch hier auf Systeme mit Ausgleich, so daß für alle $j > m$ gilt $g(jT) = 0$, kann als obere Grenze für die Summation m eingesetzt werden.

Nimmt die Verschiebung m verschiedene Werte an, so erhält man das Gleichungssystem

$$R_{ux} = T R_{uu} s \quad (6.31)$$

mit

$$R_{ux}^T = [R_{ux}(0), \dots, R_{ux}(m)]$$

$$R_{uu} = \begin{bmatrix} R_{uu}(0) & \dots & R_{uu}(m) \\ \vdots & & \vdots \\ R_{uu}(m) & \dots & R_{uu}(0) \end{bmatrix}$$

$$s^T = [g(0), \dots, g(m)].$$

Wenn die Inverse der Matrix R_{uu} existiert, kann aus Gl. (6.31) folgender Schätzalgorithmus unter Verwendung der empirisch ermittelten Korrelationsfunktionen $\hat{R}_{uu}$ und $\hat{R}_{ux}$ für die Werte an den Stützstellen abgeleitet werden:

$$\hat{s} = \frac{1}{T} [\hat{R}_{uu}]^{-1} \hat{R}_{ux} \quad (6.32)$$

mit

$$\hat{s}^T = [\hat{g}(0), \dots, \hat{g}(m)].$$

Das Problem des Korrelationsverfahrens liegt in der Ermittlung der Korrelationsfunktionen mit ausreichender Genauigkeit [1.37, 3.3, 4.9]. Werden die Korrelationsfunktionen aus gemessenen abgetasteten Werten des Eingangs- und Ausgangssignals gewonnen, sind die Schätzergebnisse von Gl. (6.32) mit denen von Gl. (6.25) identisch.

6.2.3. Schätzung der Stützstellenwerte der Kerne der Volterra-Reihe 2. Ordnung

6.2.3.1. Schätzung bei Hammerstein-Struktur des Systems

Für Systeme mit Hammerstein-Struktur wurde im Abschnitt 6.2.1 ein Modell zur Beschreibung des Ein-/Ausgangs-Verhaltens bei Darstellung der Gewichtsfunktion durch Stützstellen entsprechend Gl. (6.21) abgeleitet:

$$x(k) = r_0 T \sum_{j=0}^m g(j) + r_1 T \sum_{j=0}^m u(k-j) g(j) + r_2 T \sum_{j=0}^m u^2(k-j) g(j) + z(k). \quad (6.33)$$

Gl. (6.33) läßt sich ebenfalls in eine vektorielle Form überführen. Zum Zeitpunkt k gilt

$$x(k) = T m^T(k) s + z(k) \quad (6.34)$$

mit

$$\mathbf{m}^T = [1 : u(k), \dots, u(k-m) : u^2(k), \dots, u^2(k-m)]$$

$$\mathbf{s}^T = [g^0 : g^1(0), \dots, g^1(m) : g^2(0), \dots, g^2(m)].$$

Die einzelnen Komponenten des Parametervektors $\mathbf{s}$ haben dabei folgende Bedeutung:

$$g^0 = r_0 \sum_{j=0}^m g(j)$$

$$g^1(j) = r_1 g(j) \quad (6.35)$$

$$g^2(j) = r_2 g(j).$$

Für n Beobachtungen erhält man das Gleichungssystem

$$\mathbf{x}(k) = T\mathbf{M}\mathbf{s} + \mathbf{z}(k) \quad (6.36)$$

mit

$$\mathbf{x}^T(k) = [x(k), x(k+1), \dots, x(k+n)]$$

$$\mathbf{M} = \begin{bmatrix} 1 & u(k) & \dots & u(k-m) & u^2(k) & \dots & u^2(k+n) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & u(k+n) & \dots & u(k+n-m) & u^2(k+n) & \dots & u^2(k+n-m) \end{bmatrix}.$$

Die Schätzung für den unbekannten Vektor der Stützstellen ergibt sich nach der Methode der kleinsten Quadrate zu

$$\hat{\mathbf{s}} = \frac{1}{T} [\mathbf{M}^T \mathbf{M}]^{-1} \mathbf{M}^T \mathbf{x} \quad (6.37)$$

mit

$$\hat{\mathbf{s}}^T = [g^0, g^1(0), \dots, g^1(m), g^2(0), \dots, g^2(m)].$$

Wie aus Gl. (6.37) zu sehen ist, gelingt es für diese Modelldarstellung nicht, mit Hilfe der Regression eine getrennte Schätzung für die Parameter der beiden Teilsysteme zu erhalten. Eine Auffrennung ist jedoch zur Deutung vieler untersuchter praktischer Phänomene sinnvoll und möglich, wenn vorausgesetzt wird, daß das lineare dynamische Teilsystem einen Verstärkungsfaktor von $K = 1$ besitzt. Eine Schätzung für die Koeffizienten des Polynoms der Nichtlinearität ist dann durch folgende Beziehungen gegeben:

$$1. \text{ Aus } g^0 T = r_0' T \sum_{j=0}^m g'(j) \text{ mit } K = T \sum_{j=0}^m g'(j) = 1 \text{ folgt}$$

$$r_0' = g^0 T. \quad (6.38)$$

$$2. \text{ Aus } T \sum_{j=0}^m g^1(j) = r_1' T \sum_{j=0}^m g'(j) \text{ mit } K = T \sum_{j=0}^m g'(j) = 1 \text{ folgt}$$

$$r_1' = T \sum_{j=0}^m g^1(j). \quad (6.39)$$

$$3. \text{ Aus } T \sum_{j=0}^m g^2(j) = r_2' T \sum_{j=0}^m g'(j) \text{ mit } K = T \sum_{j=0}^m g'(j) = 1 \text{ folgt}$$

$$r_2' = T \sum_{j=0}^m g^2(j). \quad (6.40)$$

Beispiel 6.2

Parameterschätzung der Kerne der Volterra-Reihe 2. Ordnung (Hammerstein-Struktur) mittels Regression

Gegeben:

Systembeschreibung

– lineares dynamisches Teilsystem

$$G(p) = \frac{1}{(1+p \cdot 10 \text{ s})(1+p \cdot 5 \text{ s})} \quad \text{bzw.} \quad g(\tau) = 0,2 [e^{-\tau/10} - e^{-\tau/5}] \text{ s}^{-1}$$

– nichtlineares statisches Teilsystem

$$P(u) = r_0 + r_1 u + r_2 u^2$$

mit $r_0 = -1,0$, $r_1 = 2,0$, $r_2 = 0,5$.

Testsignal: Folge aus einem zentral zusammengesetzten Versuchsplan 2. Ordnung (Kern: Plackett-Burman-Plan für $m+1=11$, Amplituden ± 1 , $\pm \alpha = \pm 3,464$, Gesamtlänge 657).

Abtastzeit $T = 4 \text{ s}$,Störung NV $(0, \sigma^2)$: hochfrequent,Störverhältnis $\sqrt{\sigma^2/\sigma_{x,0}^2} = 0,15$.

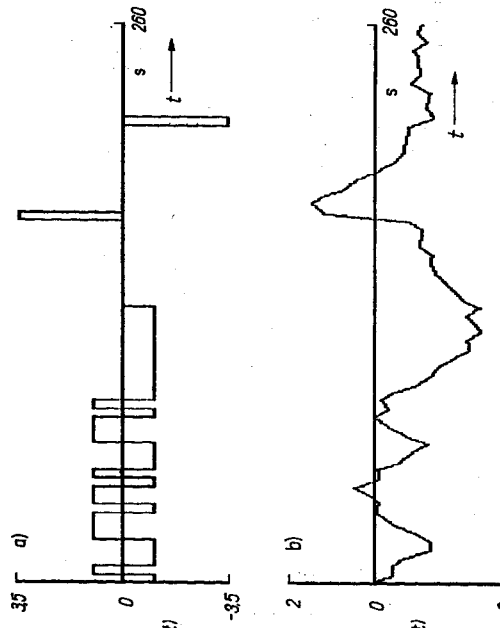
Gesucht:

– Schätzwerte für die Stützstellen der Volterra-Kerne 1. und 2. Ordnung

– Stützstellen des linearen dynamischen Teilsystems und Parameter der Nichtlinearität

– Reststreuung s_k^2 .

Ergebnis: Aus den Meßwerten des Eingangs- und Ausgangssignals (Bild a und b)



wurden folgende Stützstellenwerte für die Volterra-Kerne 1. und 2. Ordnung geschätzt:

i	$g^1(i)$	$g^1(i)$	$g^2(i)$	$g^2(i)$
0	0,08840	0,09614	0,02210	0,02146
1	0,09897	0,09764	0,02474	0,02561
2	0,08419	0,08335	0,02105	0,02234
3	0,06445	0,06521	0,01611	0,01574
4	0,04680	0,04582	0,01170	0,01784
5	0,03300	0,03682	0,00825	0,01084
6	0,02284	0,01935	0,00571	0,00720
7	0,01564	0,01182	0,00391	0,00498
8	0,01063	0,00306	0,00266	0,00571
9	0,00717	0,01627	0,00179	0,00170
10	0,00485	0,00958	0,00121	0,00452

Beispiel 6.2

Parameterschätzung der Kerne der Volterra-Reihe 2. Ordnung (Hammerstein-Struktur) mittels Regression

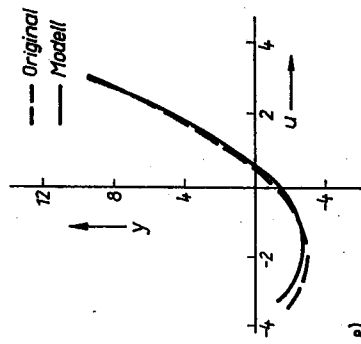
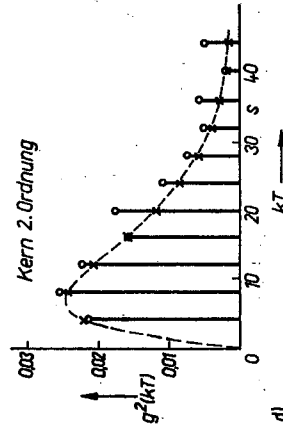
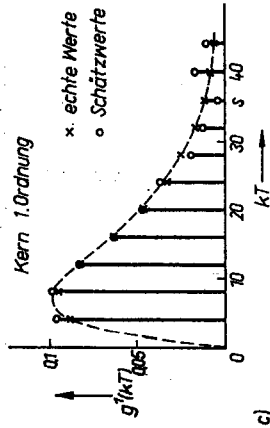
Für den Arbeitspunkt $g^0 = -1,0$ wurde als Schätzwert $g^0 = -1,0988$ erhalten. Daraus ergeben sich als Schätzwerte für die Koeffizienten der statischen Nichtlinearität

$$\hat{r}_0 = -1,0988, \quad \hat{r}_1 = 1,94269, \quad \hat{r}_2 = 0,5518$$

und für die Stützstellen der Gewichtsfunktion des linearen dynamischen Teilsystems

i	0	1	2	3	4	5
$g'(i)$	0,0442	0,0495	0,0421	0,0322	0,0234	0,0165
$g''(i)$	0,0495	0,0503	0,0429	0,0336	0,0236	0,0190

i	6	7	8	9	10
$g'(i)$	0,0114	0,0078	0,0053	0,0036	0,0024
$g''(i)$	0,0103	0,0061	0,0016	0,0084	0,0049



Die Reststreuung ist $s_k^2 = 0,0115$. Ein Vergleich der Modelle mit dem Originalsystem (Bild c bis e) zeigt, daß sowohl die Parameter des statischen Teilsystems als auch die Stützstellen beider Volterra-Kerne relativ gut ermittelt werden.

Für die normierten Werte der Stützstellen der Gewichtsfunktion gilt

$$g'_j = \frac{g^1(j)}{r_1} \quad \text{bzw.} \quad g'_j = \frac{g^2(j)}{r_2} \quad (6.41)$$

Die abgeleiteten Parameter r_i erlauben die Darstellung der statischen Nichtlinearität eines dynamischen Systems durch ein Polynom der Form

$$y = r'_0 + r'_1 u + r'_2 u^2. \quad (6.42)$$

Wie aus Gl. (6.35) zu entnehmen ist, stellen die für das Hammerstein-Modell zu schätzenden Stützstellen $g^0, g^1(j)$ und $g^2(j)$ Produkte der Koeffizienten der statischen Nichtlinearität r_i und der Abtastwerte der Gewichtsfunktion des linearen Teilmodells $g(j)$ dar. Die Gesamtzahl der zu schätzenden Parameter beträgt dadurch $N = p(m+1) + 1$, obwohl das Systemverhalten nur durch $(m+1)$ Stützstellen und p nichtlineare Koeffizienten bestimmt wird. Eine getrennte Schätzung der Parameter des nichtlinearen statischen Teilsystems und der Stützstellenwerte der Gewichtsfunktion und damit eine Parameterreduktion sind durch eine zweistufige iterative Gestaltung der Regression möglich. Die Methode ist in den Arbeiten [6.7, 6.8] dargestellt. Ihre Anwendung ist bei einer hohen Ordnung p der Nichtlinearität sinnvoll.

Die Ermittlung der Parameter eines Hammerstein-Modells für ein gestörtes System unter Verwendung der Methode der Regression ist im Beispiel 6.2 demonstriert.

6.2.3.2. Schätzung bei Wiener-Struktur des Systems

Für Systeme mit Wiener-Struktur wurde im Abschnitt 6.2.1 ein Modell zur Beschreibung des Ein-/Ausgangs-Verhaltens bei Darstellung der Gewichtsfunktion durch Stützstellen entsprechend Gl. (6.17) abgeleitet:

$$x(k) = r_0 + r_1 T \sum_{j=0}^m g(j) u(k-j) + r_2 T^2 \sum_{i=0}^m \sum_{j=1}^m g(i) g(j) u(k-i) u(k-j) + z(k). \quad (6.43)$$

Gl. (6.43) ist ebenfalls in eine vektorielle Form überführbar. Zum Zeitpunkt k gilt

$$x(k) = T^* m^T(k) s + z(k) \quad (6.44)$$

mit

$$m^T = [\underbrace{1, u(k), \dots, u(k-m)}_{\text{I}}, \underbrace{u^2(k), u(k)u(k-1), \dots, u^2(k-m)}_{\text{II}}]$$

$$T^* = \begin{cases} 1 & \text{für den Teilvektor I} \\ T & \text{für den Teilvektor II} \\ T^2 & \text{für den Teilvektor III} \end{cases}$$

$$s^T = [g^0, g^1(0), \dots, g^1(m), g^2(0, 0), g^2(0, 1), \dots, g^2(m, m)].$$

Die einzelnen Komponenten des Parametervektors s haben dabei folgende Bedeutung:

$$\begin{aligned} g^0 &= r_0 \\ g^1(j) &= r_1 g(j) \\ g^2(i, j) &= r_2 g(i) g(j). \end{aligned} \quad (6.45)$$

Für n Beobachtungen erhält man

$$\begin{aligned} x(k) &= T^* M s + z(k) \\ x^T(k) &= [x(k), x(k+1), \dots, x(k+m)] \end{aligned} \quad (6.46)$$

$M =$

$$\begin{bmatrix} 1u(k) & u(k-m) & u^2(k) & u(k)u(k-1) & \dots & u^2(k-m) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1u(k+n) & \dots & u(k+n-m) & u^2(k+n) & u(k+n-1) & \dots & u^2(k+n-m) \end{bmatrix} \quad \text{III}$$

I II

$$T^* = \begin{cases} 1 & \text{für die Teilmatrix I} \\ T & \text{für die Teilmatrix II} \\ T^2 & \text{für die Teilmatrix III.} \end{cases}$$

Die Schätzung für den unbekannten Vektor der Stützstellenwerte ergibt sich nach der Methode der kleinsten Quadrate zu

$$\hat{s} = \frac{1}{T^*} [M^T M]^{-1} M^T x \quad (6.47)$$

mit

$$\hat{s}^T = [\hat{g}^0, \hat{g}^1(0), \dots, \hat{g}^1(m), \hat{g}^2(0, 0), \hat{g}^2(0, 1), \dots, \hat{g}^2(m, m)].$$

Die Gln. (6.46) und (6.47) sind in gleicher Weise für die Schätzung des Volterra-Reihen-Modells 2. Ordnung gültig, da eine vektorielle Darstellung der Gl. (6.13) genau den Gln. (6.44) bzw. (6.46) entspricht – mit veränderter Bedeutung der $g(i, j)$.

Wie für das Hammerstein-Modell gelingt es entsprechend Gl. (6.45) auch für das Wiener-Modell nicht, eine getrennte Schätzung für die Parameter der beiden Teilsysteme zu erhalten. Eine nachträgliche Auftrennung ist jedoch ebenfalls möglich, wenn wiederum davon ausgegangen wird, daß das lineare dynamische Teilsystem einen Verstärkungsfaktor von $K = 1$ besitzt. Aus dieser Bedingung und den Beziehungen in Gl. (6.45) folgt für die dann gültigen normierten Koeffizienten r'_i :

$$1. \quad r_0 = g^0. \quad (6.48)$$

2. Aus

$$T \sum_{j=0}^m g^1(j) = r'_1 T \sum_{j=0}^m g'(j)$$

folgt mit

$$T \sum_{j=0}^m g'(j) = 1 \quad (6.49)$$

$$r'_1 = T \sum_{j=0}^m g^1(j).$$

3. Aus

$$T^2 \sum_{j=0}^m g^2(j, j) = r'_2 T^2 \sum_{j=0}^m g'(j)^2$$

folgt mit

$$\begin{aligned} g'(j) &= g^1(j)/r'_1 \\ r'_2 &= r'^2_2 \frac{\sum_{j=0}^m g^2(j, j)}{\sum_{j=0}^m [g^1(j)]^2} = T^2 \frac{\left[\sum_{j=0}^m g^1(j) \right]^2}{\sum_{j=0}^m [g^1(j)]^2}. \end{aligned} \quad (6.50)$$

Für die normierten Werte der Stützstellen der Gewichtsfunktion ergibt sich

$$g'(j) = g^1(j)/r'_1. \quad (6.51)$$

Die im Abschnitt 6.2.3.1 genannte Methode der zweistufigen Regression zur sofortigen getrennten Ermittlung der Parameter der beiden Teilsysteme ist auf das Wiener-Modell nicht anwendbar, da eine geschlossene vektorielle Darstellung des Modellausgangs in einer für das Verfahren erforderlichen Form nicht möglich ist [6.3].

6.2.4. Schätzung der Stützstellen der Gewichtsfunktionen in Mehrgrößensystemen

Wird eine p -kanonische Struktur des Mehrgrößensystems angenommen, kann die Schätzung der Stützstellen der Gewichtsfunktionen der l Teilsysteme, die auf den i -ten Ausgang wirken, in Analogie zu Gl. (6.20) erfolgen auf der Basis der Beziehung

$$x_i(k) = T \sum_{j=1}^{m_{si}} g_{si}(j) u_s(k-j) + z_i(k); \quad (6.52)$$

$g_{si}(j)$ j -te Stützstelle der Gewichtsfunktion $g_{si}(\tau)$ für die Teilstrecke zwischen dem θ -ten Eingang und dem i -ten Ausgang
($m_{si} + 1$) Anzahl der Stützstellen für die Gewichtsfunktion $g_{si}(\tau)$.

Gl. (6.52) ist ebenfalls in eine vektorielle Form überführbar. Zum Zeitpunkt k gilt

$$\begin{aligned} x_i(k) &= T m_i^T(k) s_i + z_i(k) \\ m_i^T(k) &= [u_1(k) \dots u_1(k-m_{1i}) : u_2(k) \dots u_2(k-m_{2i}) : \dots \\ &\quad \dots : u_l(k) \dots u_l(k-m_{li})] \end{aligned} \quad (6.53)$$

$$s_i^T = [g_{1i}(0) \dots g_{1i}(m_{1i}) : g_{2i}(0) \dots g_{2i}(m_{2i}) : \dots \\ \dots : g_{li}(0) \dots g_{li}(m_{li})].$$

Für n Beobachtungen erhält man

$$x_i(k) = T M_i s_i + z_i(k) \quad (6.54)$$

mit

$$x_i^T(k) = [x_i(k), x_i(k+1), \dots, x_i(k+n)]$$

$$M_i = \begin{bmatrix} u_1(k) & \dots & u_1(k-m_{1i}) & \dots & u_l(k) & \dots & u_l(k-m_{li}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_1(k+n) & \dots & u_1(k+n-m_{1i}) & \dots & u_l(k+n) & \dots & u_l(k+n-m_{li}) \end{bmatrix}$$

Mit der Methode der kleinsten Quadrate ergibt sich die Schätzung der Stützstellen der Gewichtsfunktionen zu

$$\hat{s}_i = \frac{1}{T} [M_i^T M_i]^{-1} M_i^T x_i \quad (6.55)$$

mit

$$s_i^T = [\hat{g}_{1i}(0) \dots \hat{g}_{1i}(m_{1i}) : \dots : \hat{g}_{ii}(0) \dots \hat{g}_{ii}(m_{ii})].$$

Damit wurde eine Zerlegung des Mehrgrößensystems mit l Eingängen und m Ausgängen in m Teilsysteme mit l Eingängen vorgenommen (s. Bilder 6.32 und 6.33).

Bisher wurde davon ausgegangen, daß für alle Teilstrecken die gleiche Abtastzeit T verwendet werden kann. Im Abschnitt 6.6.4 werden verschiedene Strategien zur Wahl der Abtastzeit T vorgestellt.

Für Mehrgrößensysteme p -kanonischer Struktur wird zunächst, ausgehend von diesen Strategien, eine günstige Abtastzeit T_{si} für jede Teilstrecke bestimmt. Durch Festlegung einer entsprechenden Stützstellenzahl ($m_{si} + 1$) für jede Teilstrecke sollte zur Vereinfachung der Versuchsdurchführung und der Auswertung versucht werden, eine gemeinsame Abtastzeit für alle Teilstrecken festzulegen. Dann gilt

$$T = T_{si}; \quad \phi = 1, \dots, l.$$

Unterscheiden sich die Abtastzeiten für die einzelnen Teilstrecken T_{si} wesentlich, führt die Verwendung einer gemeinsamen Abtastzeit T zu einer ungünstigen Anzahl von Stützstellen für die Gewichtsfunktionen der Teilstrecken. (Einzelne Teilstrecken haben eine zu niedrige Stützstellenzahl; andernfalls wird die Gesamtanzahl der Stützstellen für alle Teilstrecken sehr hoch.)

In einem solchen Fall muß von der Verwendung unterschiedlicher Abtastzeiten für die einzelnen Teilstrecken ausgegangen werden.

In den Gln. (6.54) und (6.55) ist dann die skalare Größe T durch eine Diagonalmatrix T_i der Form

$$T_i = \begin{bmatrix} T_{1i} & & & \\ & T_{2i} & & \\ & & \ddots & \\ & & & T_{ii} \end{bmatrix} \quad (6.56)$$

$\left. \begin{matrix} m_{1i} + 1 \\ m_{2i} + 1 \\ \vdots \\ m_{ii} + 1 \end{matrix} \right\} T_{ii}$

zu ersetzen. Als Schätzvorschrift für die Stützstellen der Gewichtsfunktionen für den i -ten Ausgang gilt dann

$$\hat{s}_i = T_i^{-1} [M_i^T M_i]^{-1} M_i^T x_i. \quad (6.57)$$

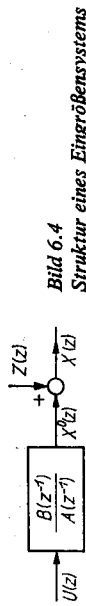
6.3. Schätzung der Koeffizienten der Differenzengleichung

6.3.1. Systembeschreibungen und Fehlergleichungen

6.3.1.1. Lineare Eingößensysteme

Eine weitere allgemeine Beschreibung des dynamischen Verhaltens eines diskreten Systems erfolgt durch die Differenzengleichung [1.34 bis 1.53, 3.2]. Für ein lineares zeitinvariantes ungestörtes Eingößensystem m -ter Ordnung ist sie gegeben durch

$$\begin{aligned} x^0(k) + a_1 x^0(k-1) + \dots + a_m x^0(k-m) \\ = b_0 u(k) + b_1 u(k-1) + \dots + b_m u(k-m). \end{aligned} \quad (6.58)$$



Dabei bedeuten $x^0(k)$ die zum Zeitpunkt kT abgetasteten Werte des ungestörten Ausgangssignals und $u(k)$ die abgetasteten Werte des Eingangssignals (Bild 6.4). Der Signalwert $x^0(k)$ ergibt sich aus Gl. (6.58) rekursiv zu

$$x^0(k) + \sum_{i=1}^m a_i x^0(k-i) = \sum_{i=0}^m b_i u(k-i). \quad (6.59)$$

In Vektorschreibweise ergibt sich aus Gl. (6.58)

$$x^0(k) = m^{0T}(k) s \quad (6.60)$$

mit

$$m^{0T}(k) = [-x^0(k-1) \dots -x^0(k-m) : u(k), u(k-1) \dots u(k-m)]$$

$$s^T = [a_1 \dots a_m : b_0, b_1 \dots b_m].$$

Für Systeme mit dominierendem Trägheitsverhalten (d.h. nichtsprungfähige Systeme) ist der Koeffizient $b_0 = 0$. Deshalb wird in vielen Arbeiten als Meßwertvektor

$$m^{0T}(k) = [-x^0(k-1) \dots -x^0(k-m) : u(k-1) \dots u(k-m)]$$

und als Parametervektor

$$s^T = [a_1 \dots a_m : b_1 \dots b_m]$$

verwendet.

Mit der Anwendung des Verschiebungssatzes der z -Transformation in der Form

$$Z\{u(k-i)\} = z^{-i} Z\{u(k)\} = z^{-i} U(z)$$

folgt aus Gl. (6.59)

$$X^0(z) [1 + a_1 z^{-1} + \dots + a_m z^{-m}] = U(z) [b_0 + b_1 z^{-1} + \dots + b_m z^{-m}]. \quad (6.61)$$

Wird für

$$[1 + a_1 z^{-1} + \dots + a_m z^{-m}] = A(z^{-1})$$

$$[b_0 + b_1 z^{-1} + \dots + b_m z^{-m}] = B(z^{-1})$$

gesetzt, erhält man für die Ausgangsgröße:

$$X^0(z) = \frac{B(z^{-1})}{A(z^{-1})} U(z). \quad (6.62)$$

Für das am Ausgang gestörte System gilt mit

$$X(z) = X^0(z) + Z(z) \quad (6.63)$$

und mit der Beschreibung des ungestörten Ausgangs $X^0(z)$

$$X(z) = \frac{B(z^{-1})}{A(z^{-1})} U(z) + Z(z). \quad (6.64)$$

Wird Gl. (6.64) analog zu den Gln. (6.59) und (6.60) in eine parameterlineare Form umgeformt, gilt

$$A(z^{-1}) X(z) = B(z^{-1}) U(z) + A(z^{-1}) Z(z) \quad (6.65)$$

$$x(k) = - \sum_{i=1}^m a_i x(k-i) + \sum_{i=0}^m b_i u(k-i) + \sum_{i=0}^m a_i z(k-i).$$

Als Systembeschreibung im Zeitbereich erhält man damit in Vektorschreibweise für Gl. (6.65)

$$x(k) = m^T(k) s + n^*(k) \quad (6.66)$$

mit

$$n^*(k) = z^T(k) a,$$

$$z^T = [z(k), \dots, z(k-m)]$$

$$a^T = [a_0, a_1, \dots, a_m].$$

Aus Gl. (6.66) ist zu ersehen, daß selbst bei unkorrelierter Störung $z(k)$ durch das Nennerpolynom $A(z^{-1})$ eine korrelierte Störung $n^*(k) = Z(z) A(z^{-1})$ entsteht.

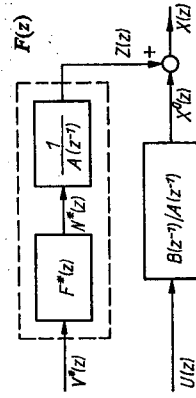


Bild 6.5
Erweiterte Struktur eines Eingangs-Systems

Da aber eine Korrelation zwischen der Störung $n^*(k)$ und dem Meßwertvektor $m(k)$ zu nicht erwartungstreuen Schätzungen führt (s. Abschn. 6.3.2 und [1.44, 1.47]), wurde von Aström [1.38] die Systembeschreibung um ein Filter $F(z)$ für die Störung erweitert. Dieses Filter hat die Aufgabe, die korrelierte Störung $n^*(k)$ in ein „weißes Rauschsignal“ $v^*(k)$ umzuwandeln. Dadurch wird wieder eine erwartungstreue Schätzung der Parameter möglich. Die Struktur der erweiterten Systembeschreibung ist im Bild 6.5 dargestellt. Damit gilt entsprechend Gl. (6.64)

$$X(z) = \frac{B(z^{-1})}{A(z^{-1})} U(z) + F(z) V^*(z). \quad (6.67)$$

Für $F(z)$ wird i. allg. folgender Ansatz gewählt:

$$F(z) = \frac{1}{A(z^{-1})} F^*(z) \quad (6.68)$$

mit

$$F^*(z) = C(z^{-1})/D(z^{-1})$$

$$C(z^{-1}) = [1 + c_1 z^{-1} + \dots + c_n z^{-n}]$$

$$D(z^{-1}) = [1 + d_1 z^{-1} + \dots + d_n z^{-n}].$$

Neben der Erweiterung der Beschreibung des Systems durch das Störfilter $F(z)$ spielt die Art der Fehlergleichung bei diesem Modelltyp eine wichtige Rolle. Wird der Ausgangs- oder der Eingangsfehler verwendet (s. Abschn. 5.1.4), erhält man Fehlergleichungen, in die ein Teil der Parameter nichtlinear eingeht. Es gilt für den Ausgangsfehler

$$E_x(z) = X(z) - \hat{X}(z) = X(z) - \frac{\hat{B}(z^{-1})}{\hat{A}(z^{-1})} U(z)$$

und für den Eingangsfehler

$$E_u(z) = U(z) - \hat{U}(z) = U(z) - \frac{\hat{A}(z^{-1})}{\hat{B}(z^{-1})} X(z).$$

Die Parameterschätzung für diese Fehler ist nur iterativ möglich (s. Abschn. 5.3). Einen Ausweg bildet die Anwendung des verallgemeinerten Fehlers (Gleichungsfehler). Für ihn gilt die Vorschrift

$$E_e(z) = X'(z) - U'(z) = \hat{A}(z^{-1}) X(z) - \hat{B}(z^{-1}) U(z).$$

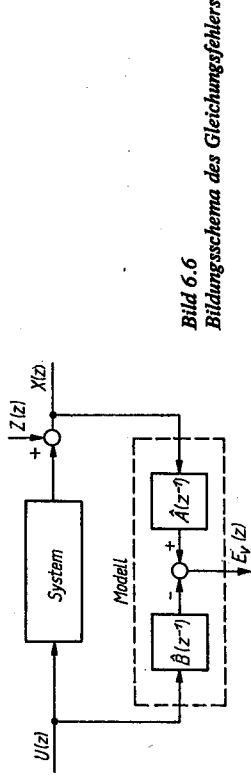


Bild 6.6
Bildungsschema des Gleichungsfehlers

Im Bild 6.6 ist das Schema dieser Fehlerbildung angegeben. Die Umwandlung des korrelierten Fehlers $e_e(k)$ – im folgenden mit $e(k)$ bezeichnet – in einen unkorrelierten Fehler $e(k)$ kann entweder über Filter für die Signale $u(k)$ und $x(k)$ (Bild 6.7a) oder für den Fehler (Bild 6.7b) erfolgen. Damit gilt mit der Beziehung Gl. (6.68)

$$E(z) = \hat{F}^*(z) e(z) \quad (6.69)$$

für den Gleichungsfehler

$$\hat{A}(z^{-1}) X(z) - \hat{B}(z^{-1}) U(z) = \hat{F}^*(z) e(z) \quad (6.70)$$

und für das vollständige Modell des Systems von Gl. (6.64)

$$X(z) = \frac{\hat{B}(z^{-1})}{\hat{A}(z^{-1})} U(z) + \frac{1}{\hat{A}(z^{-1})} \hat{F}^*(z) e(z). \quad (6.71)$$

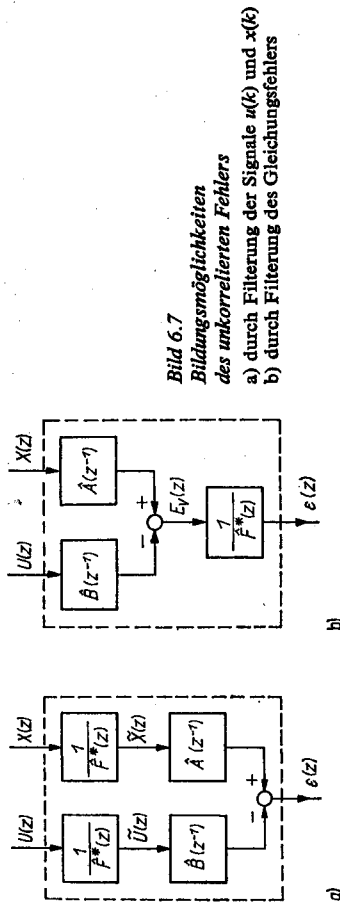


Bild 6.7
Bildungsmöglichkeiten
des unkorrelierten Fehlers
a) durch Filterung der Signale $u(k)$ und $x(k)$
b) durch Filterung des Gleichungsfehlers

Das durch Gl. (6.71) beschriebene Modell wird auch als ARMAX-Modell (autoregressive moving average with exogenous variable) bezeichnet. In der Praxis wird diese Modellform am häufigsten verwendet. Als Sonderfälle können aus diesem Modell die in Tafel 6.1 angegebenen Zuordnungen von Schätzverfahren zur entsprechenden Filterübertragungsfunktion gebildet werden.

Tafel 6.1. Schätzverfahren und ihre Formfilter $F^*(z)$

Schätzverfahren	$F^*(z)$	Literatur
Regression	1	[1.34, 1.38, 1.44, 1.45, 1.47]
Methode der Hilfsvariablen	1	[6.14, 6.15]
Verallgemeinerte Regression	$\frac{1}{D(z^{-1})}$	[6.16, 6.17]
Maximum-Likelihood-Schätzung	$\frac{1}{C(z^{-1})}$	[1.38, 6.19]

Um für das in Gl. (6.71) dargestellte allgemeine Modell eine Beschreibung zur Parameterschätzung mit den im Abschnitt 5 bereits abgeleiteten Verfahren zu erhalten, führt man eine Rücktransformation in den Zeitbereich durch. Damit erhält man die Differenzengleichung

$$x(k) + \sum_{i=1}^m \hat{a}_i x(k-i) - \sum_{i=0}^m \hat{b}_i u(k-i) = e(k) \quad (6.72)$$

mit

$$e(k) = - \sum_{i=1}^n \hat{d}_i e(k-i) + e(k) + \sum_{i=1}^n c_i e(k-i).$$

In Vektorschreibweise ergibt sich für den k -ten Zeitpunkt die Beziehung

$$x(k) = m^T(k) \hat{s} + e(k), \quad (6.73)$$

wobei

$$m^T(k) = [-x(k-1) \dots -x(k-m); u(k) \dots u(k-m); -e(k-1) \dots -e(k-n); e(k-1) \dots e(k-n)]$$

ein Meßwertvektor und

$$\hat{s}^T = [\hat{d}_1 \dots \hat{d}_m; \hat{b}_0 \dots \hat{b}_m; \hat{c}_1 \dots \hat{c}_n; \hat{d}_1 \dots \hat{d}_n]$$

der Parametervektor ist. Schwierigkeiten bereitet in dieser Beschreibung die Ermittlung der Signalwerte $e(k-i)$. Je nach Wahl der Übertragungsfunktion des Filters $F(z)$ können durch Gl. (6.73) verschiedene Modellstrukturen realisiert werden (s. Abschnitt 6.3.2 bis 6.3.5).

6.3.1.2. Einfache nichtlineare Eingrößensysteme

Neben den Beschreibungen des nichtlinearen dynamischen Verhaltens durch das nichtparametrische Volterra-Reihen-Modell (s. Abschn. 6.2) finden einfache nichtlineare Modelle, die aus einer Kombination eines linearen dynamischen Systems (mit L bezeichnet) und eines nichtlinearen statischen Systems (mit NL bezeichnet) bestehen, eine breite Anwendung. Dabei wird im folgenden davon ausgegangen, daß das nichtlineare statische Teilsystem durch folgende Gleichung hinreichend genau beschrieben wird:

$$x_{NL}(k) = \sum_{j=0}^p r_j \mu_{NL}^j(k). \quad (6.74)$$

Hammerstein-Modell

Wird entsprechend Bild 6.8 davon ausgegangen, daß sich die Nichtlinearität vor dem linearen dynamischen Teil befindet, wird diese Struktur Hammerstein-Modell genannt. Entsprechend der Beschreibung des linearen dynamischen Teilsystems von Gl. (6.58) gilt

$$x^0(k) = - \sum_{i=1}^m a_i x^0(k-i) + \sum_{i=0}^m b_i y(k-i).$$

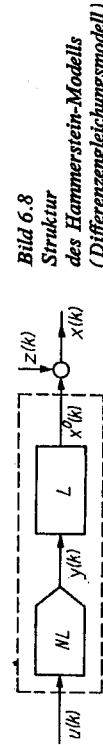


Bild 6.8
Struktur
des Hammerstein-Modells
(Differenzengleichungsmodell)

Das fiktive Zwischensignal $y(k)$ lautet bei Verwendung von Gl. (6.74)

$$y(k) = \sum_{j=0}^p r_j \mu^j(k). \quad (6.75)$$

Damit ergibt sich für die Beschreibung des Systems mit der Struktur eines Hammerstein Modells

$$x^0(k) = \sum_{i=1}^m a_i x^0(k-i) + \sum_{i=0}^m b_i \sum_{j=0}^p r_j \mu^j(k-i). \quad (6.76)$$

Wird der konstante Anteil der Signalfolge $x(k)$, der durch den Ausdruck

$$r_{00} = r_0 \sum_{i=0}^m b_i$$

gegeben ist, in Gl. (6.76) getrennt dargestellt, erhält man mit der Vereinbarung $b_i r_j \triangleq b_{ij}$

$$x^0(k) = r_{00} - \sum_{i=1}^m a_i x^0(k-i) + \sum_{j=1}^p \sum_{i=0}^m b_{ij} \mu^j(k-i). \quad (6.77)$$

Damit gilt in Vektorschreibweise

$$x^0(k) = m^T(k) s \quad (6.78)$$

mit

$$m^T(k) = [1, -x^0(k-1) \dots -x^0(k-m); u(k) \dots u(k-m); \\ \dots; u^p(k) \dots u^p(k-m)]$$

$$s^T = [r_{00}; a_1 \dots a_m; b_{10} \dots b_m; \\ \dots; b_p \dots b_{pm}].$$

In Gl. (6.78) wird sehr gut sichtbar, daß das nichtlineare Hammerstein-Modell als Spezialfall das lineare Systemverhalten mit enthält.

Tritt eine Störung $z(k)$ auf, ist analog der Vorgehensweise im linearen Fall die Beschreibung um ein Störfilter zu erweitern und in der Berechnungsvorschrift für den Gleichungsfehler – s. Gl. (6.72) – der entsprechende Ansatz aus Gl. (6.77) einzusetzen. Die Ordnung der statischen Nichtlinearität wird häufig auf $p = 2$ beschränkt. Da die Fehlergleichungen bei diesem nichtlinearen Modell auch parameterlinear sind, treten bei der Schätzung keine prinzipiellen neuen Probleme auf (s. Abschn. 6.3.5).

Eine Trennung der gemeinsam geschätzten Parameter $\hat{b}_i, \hat{p}_j \triangleq \hat{b}_{ij}$ der statischen und dynamischen Teilsysteme ist unter der Annahme, daß der Verstärkungsfaktor des linearen dynamischen Teilsystems $K_L = 1$ ist, möglich. Mit Hilfe des Zusammenhangs

$$K_L = \sum_{i=0}^m \frac{\hat{b}_i}{1 + \sum_{i=1}^m \hat{a}_i} = 1 \quad (6.79)$$

können die Parameter aus folgenden Beziehungen ermittelt werden:

$$\hat{p}_j = \frac{\hat{b}_{ij}}{\hat{b}_i} = \frac{\sum_{i=0}^m \hat{b}_{ij}}{1 + \sum_{i=1}^m \hat{a}_i} = \frac{\hat{p}_{00}}{1 + \sum_{i=1}^m \hat{a}_i} \quad (6.80)$$

$$\hat{b}_i = \hat{b}_{ij}/\hat{p}_j.$$

Die Trennung der Parameter ist bei rein integralem oder differentiellem Verhalten nicht möglich, da die Verstärkung gegen Unendlich bzw. gegen Null strebt.

Wiener-Modell

Die Struktur eines Wiener-Modells ist im Bild 6.9 dargestellt. Die Beschreibung ist dadurch gekennzeichnet, daß das statische nichtlineare Teilsystem nach dem linearen dynamischen Teilsystem angeordnet ist. Für die Beschreibung gelten mit den bereits getroffenen Vereinbarungen

$$y(k) = - \sum_{i=1}^m a_i y(k-i) + \sum_{i=0}^m b_i u(k-i) \quad (6.81)$$

$$x^0(k) = \sum_{j=0}^p r_j y^j(k).$$

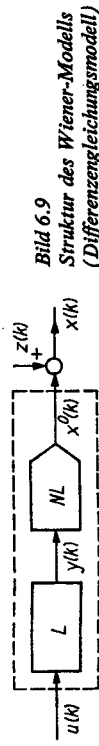


Bild 6.9
Struktur des Wiener-Modells
(Differenzgleichungsmodell)

Für die Gesamtbeschreibung ergibt sich damit

$$x^0(k) = \sum_{j=0}^p r_j \left[- \sum_{i=1}^m a_i y(k-i) + \sum_{i=0}^m b_i u(k-i) \right].$$

Diese Beziehung kann wieder in eine Vektorbeziehung der Form

$$x^0(k) = m^T(k) s$$

umgeschrieben werden, wobei der Meßwertvektor $m(k)$ abgetastete Signalwerte des Eingangs- und des Zwischensignals in linearer und potenziert Form, Wechselwirkungen zu gleichen und unterschiedlichen Zeitpunkten und zeitlich verschobene Wechselwirkungen der Signale selbst enthält. Die Anzahl der zu ermittelnden Parameter ist dadurch gegenüber dem Hammerstein-Modell sehr groß. Außerdem erfordert die Eliminierung des fiktiven Zwischensignals $y(k)$, die nur im Fall einer eindeutig umkehrbaren Kennlinie des statischen Teilsystems möglich ist, eine iterative Lösungsstrategie [6.9, 6.10, 6.11]. Aus diesen Gründen ist die Anwendung dieser Beschreibung in der Praxis nicht so häufig.

In den nächsten Abschnitten werden einige wesentliche Schätzverfahren unter Beachtung folgender Voraussetzungen näher betrachtet:

- Die abgetasteten Signalwerte der Eingänge $u(k)$ sind ungestört.
- Die Störung $z(k)$ greift additiv ein, und für den Erwartungswert gilt $E\{z(k)\} = 0$.
- Die Ordnung m und die Totzeit $T_t = dT$ des Systems sind bekannt.
- Die Werte der Eingangsgrößen und der Störung sind unkorreliert, d. h., es gilt $E\{z(k)u(k)\} = 0$.

6.3.2. Methoden der Regression

6.3.2.1. Direkte Methode der Regression

Bei der Herleitung der direkten Methode der Regression für die Parameterschätzung der linearen Differenzengleichung wird von der Fehlergleichung entsprechend Gl. (6.71) unter der Annahme, daß die Filterübertragungsfunktion $\hat{F}^*(z) = 1$ ist, ausgegangen. Damit gilt

$$\varepsilon(z) = \hat{A}(z^{-1}) X(z) + \hat{B}(z^{-1}) U(z). \quad (6.82)$$

Dies bedeutet, daß das Fehlersignal $\varepsilon(k) = \varepsilon(k)$ und damit ein „weißes Rauschsignal“ ist. Für den k -ten Beobachtungszeitpunkt erhält man für Gl. (6.82) im Zeitbereich

$$\varepsilon(k) = x(k) - m^T(k) s$$

mit

$$m^T(k) = [-x(k-1) \dots -x(k-m); u(k-1) \dots u(k-m)]$$

$$s^T = [a_1 \dots a_m; b_1 \dots b_m].$$

Bei Verwendung der bekannten Zielfunktion für die Methode der Regression,

$$Q = \varepsilon^T(k) \varepsilon(k) = \sum_{k=j}^{j+n-1} \varepsilon^2(k), \quad (6.83)$$

der Meßwertmatrix $M(n)$ für n Beobachtungen,

$$M(n) = \begin{bmatrix} -x(j-1) & \dots & -x(j-m) & u(j-1) & \dots & u(j-m) \\ \vdots & & \vdots & \vdots & & \vdots \\ -x(j-1+n-1) & \dots & -x(j-m+n-1) & u(j-1+n-1) & \dots & u(j-m+n-1) \end{bmatrix},$$

und des gemessenen Ausgangssignalvektors $x(k)$ für n Beobachtungen,

$$x^T(n) = [x(j), x(j+1), \dots, x(j+n-1)],$$

ergibt sich analog zu den Ableitungen im Abschnitt 5.2.2 die Regressionsvorschrift zu

$$\hat{s} = [M^T(n) M(n)]^{-1} M^T(n) x(n). \quad (6.84)$$

Die Eigenschaften der Schätzvorschrift und die zu verwendenden numerischen Lösungsverfahren entsprechen den im Abschnitt 5.2.2 angegebenen.

Schwierigkeiten bei der Inversion der Matrix $[M^T(n) M(n)]$ treten nur auf, wenn sich das Eingangssignal von Tastpunkt zu Tastpunkt zu gering ändert und dadurch das System nicht ausreichend erregt wird. Einen Ausweg bilden auch hier die Methoden der aktiven Versuchsplanung (s. Abschn. 6.4).

Ein generelles Problem der Schätzung der Parameter der Differenzengleichung mit der Methode der Regression stellt die Erwartungstreue dar. Unter den oben angenommenen Voraussetzungen gilt für die Schätzung

$$E\{\hat{s} - s\} = 0, \quad (6.85)$$

d.h. Erwartungstreue. Allgemein ist dies aber nicht der Fall. Entsprechend den Betrachtungen im Abschnitt 5.2.2 gilt für die Erwartungstreue bei Schätzung der Koeffizienten der linearen Differenzengleichung für eine hinreichend große Anzahl von Abtastwerten $u(k)$ und $x(k)$ mit $k = j, \dots, j+n-1$ bei Verwendung von (6.84) und (6.65)

$$E\{\hat{s}\} = E\{[M^T(n) M(n)]^{-1} M^T(n) [M(n)s + n^*(n)]\}. \quad (6.86)$$

Durch weitere Umformungen erhält man den für die Betrachtungen günstigen Ausdruck

$$E\{\hat{s} - s\} = E\{[M^T(n) M(n)]^{-1}\} E\{M^T(n) n^*(n)\}. \quad (6.87)$$

Da der erste Term in Gl. (6.87) nicht null werden kann, ist Gl. (6.85) nur zu erfüllen, wenn der Ausdruck

$$E\{M^T(n) n^*(n)\} = 0 \quad (6.88)$$

ist. Werden die einzelnen Meßwerte in die Gl. (6.88) eingesetzt, erhält man

$$E\{M^T(n) n^*(n)\} = E \left\{ \begin{bmatrix} u(j) & \dots & u(j+n-1) & n^*(j) \\ \vdots & & \vdots & \vdots \\ u(j-m) & \dots & u(j-m+n-1) & \vdots \\ x(j-1) & \dots & x(j-1+n-1) & \vdots \\ x(j-m) & \dots & x(j-m+n-1) & n^*(j+n-1) \end{bmatrix} \right\}. \quad (6.89)$$

Aus Gl. (6.89) ist zu entnehmen, daß der obere Teil des Vektors Elemente mit dem Wert null ergibt, da die Störungswerte $z(k)$ und damit auch $n^*(k)$ mit den Werten der Eingangsgrößen $u(k)$ laut Voraussetzung unkorreliert sein sollen. Im unteren Teil des Vektors kommen Ausdrücke der Form

$$E\{x(j-1)n^*(j)\} = E\left\{x(j-1)\left[z(j) + \sum_{i=1}^m a_{iz}(j-i)\right]\right\} \quad (6.90)$$

vor. Wird in Gl. (6.90) für die abgetastete gestörte Ausgangsgröße $x(k) = x^0(k) + z(k)$ entsprechend Gl. (6.62) gesetzt, erhält man

$$E\{x(j-1)n^*(j)\} = E\left\{[x^0(j-1) + z(j-1)]\left[z(j) + \sum_{i=1}^m a_{iz}(j-i)\right]\right\}. \quad (6.91)$$

Der Erwartungswert der Beziehung in Gl. (6.91) wird null und damit die Schätzung erwartungstreu, wenn

- die abgetasteten Störungswerte $z(k)$ null sind oder
- bei vorhandener Störung $z(k)$ die abgetasteten Störungswerte unkorreliert sind („weisses Rauschen“) und die durch die Beschreibung verursachte Korrelation durch eine Filterung mit $F(z) = 1/A(z^{-1})$ beseitigt wird, d.h. der Gleichungsfehler $\varepsilon(k)$ unkorreliert ist.

Sind die abgetasteten Signalwerte der Störung $z(k)$ korreliert, liefert die Methode der Regression verschobene Schätzungen. Einen Ausweg, unter diesen Bedingungen unverschobene Schätzungen zu erhalten, besteht in der Verwendung erweiterter Systemmodelle (s. Abschnitte 6.3.4 und 6.3.5) und der Verwendung von sog. „Hilfsvariablen“ (s. Abschn. 6.3.3).

Ein weiteres Problem besteht in der Ermittlung der Kovarianzmatrix

$$\text{COV}\{\hat{s}\} = E\{(\hat{s} - s)(\hat{s} - s)^T\} \quad (6.92)$$

für diesen Modelltyp. Analog zu den Betrachtungen im Abschnitt 5.2.2 gilt für Gl. (6.92)

$$\text{COV}\{\hat{s}\} = E\{[M^T(n) M(n)]^{-1} M^T(n) n^*(n) n^{*T}(n) M(n) [M^T(n) M(n)]^{-1}\}. \quad (6.93)$$

Da in der Meßwertmatrix $M(n)$ neben den ungestörten Signalwerten $u(k)$ auch gestörte Signalwerte $x(k)$ enthalten sind, ist, selbst wenn

$$E\{n^*(n) n^{*T}(n)\} = \sigma_n^2 I \quad (6.94)$$

gilt, eine Vereinfachung analog zu Gl. (5.31) nicht mehr möglich. In den Arbeiten [6.12, 6.13] wurde aus Gl. (6.93) eine Abschätzung der Form

$$\text{COV}\{\hat{s}\} \approx \sigma_n^2 [M^T(n) M(n)]^{-1} \quad (6.95)$$

vorgeschlagen. Ihre Aussagekraft ist aber aufgrund der vorgenommenen Vereinfachung in vielen Fällen nicht ausreichend (s. Abschn. 6.4).

Aus Gl. (6.95) wird auch ersichtlich, daß eine Filterung des Ausgangssignals zur Beseitigung des Einflusses der Störung immer sinnvoll ist und die Güte der Schätzung verbessert. Ist die Störung tieffrequent (hochfrequent), wird in einer sog. „Vorfilterung“ durch einen Hochpaß (Tiefpaß) (s. Abschn. 3.4) ihr Einfluß reduziert. Diese Vorfilterung

Beispiel 6.3 Parameterschätzung eines linearen Differenzgleichungsmodells mittels der Regression

Gegeben:

Systembeschreibung

$$G(p) = \frac{5}{(1 + p \cdot 10 s)(1 + p \cdot 5 s)}$$

Mit $T = 2$ folgt

$$G(z) = \frac{0,164z^{-1} + 0,134z^{-2}}{1 - 1,489z^{-1} + 0,548z^{-2}}$$

Testsignal: PRBS-Folge (Länge 157, Amplitude $\pm 2,5$, Gesamtlänge 607),

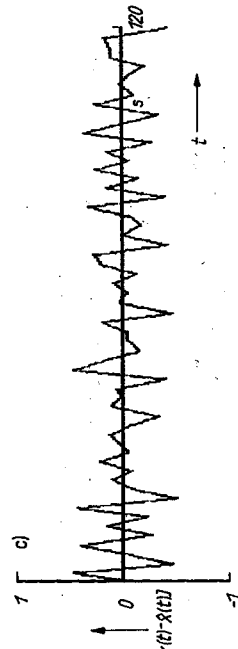
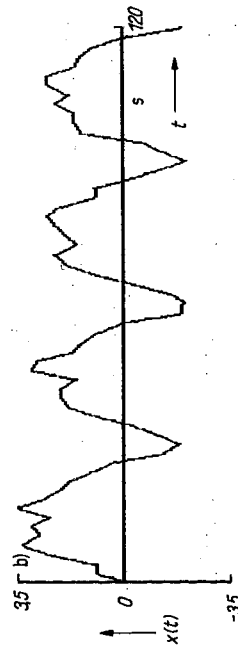
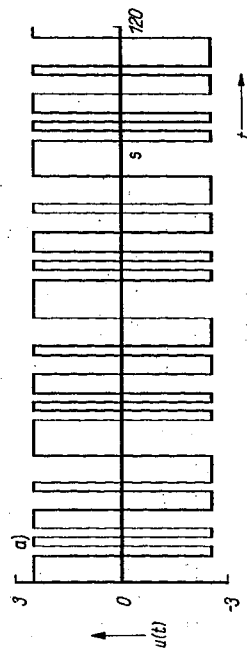
Störung NV ($0, \sigma_2^2$); hochfrequent, Störverhältnis $\sqrt{\sigma_2^2/\sigma_{\hat{x}_0}^2} = 0,16$.

Gesucht: Schätzwerte der Koeffizienten der diskreten Übertragungsfunktion und die Reststreuung s_e^2 .

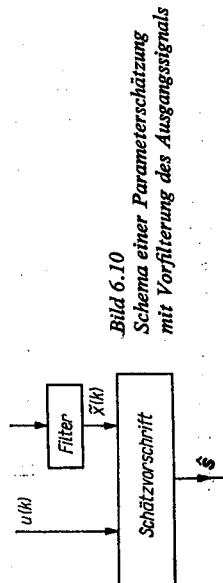
Ergebnis: Aus den Meßwerten des Eingangs- und Ausgangssignals wurden folgende Parameter geschätzt:

i	a_i	$\hat{a}_i$	b_i	$\hat{b}_i$
1	-1,489	-1,108	0,164	0,162
2	0,548	0,162	0,134	0,216

Die Reststreuung ist $s_e^2 = 0,264$.



Das Schätzergebnis zeigt, daß besonders bei den Parametern $\hat{a}_i$ bei Anwendung der Methode der Regression ein relativ großer Parameterfehler auftritt. Er ist bedingt durch die Länge der Stichprobe und die durch die Störung verursachte Nichterwartungstreue der Schätzung bei diesem Verfahren. Dagegen ist die Reststreuung s_e^2 relativ klein, was auch am Verlauf der Differenz zwischen System- und Modellausgang deutlich wird.



6.3.2.2. Rekursive Methode der Regression

Wie bereits bei der Schätzung der Parameter für statische Modelle im Abschnitt 5.4.3 besprochen, hat die direkte Methode der Regression einige entscheidende Nachteile:

- Der Speicheraufwand für die gemessenen n Meßwertpaare $u(k)$ und $x(k)$ ist hoch.
- Bei neuen Meßwerten muß die Gl. (6.84) jeweils erneut vollständig gelöst werden.
- Der Rechenaufwand zur Inversion der Matrix $[M^T M]$ ist hoch.
- Die Eignung für die Parameterschätzung zeitvarianter Systeme ist schlecht.

Diese und andere Nachteile können durch eine rekursive Lösung des Schätzproblems beseitigt werden. Von den verschiedenen rekursiven Schätzverfahren wird für die Schätzung der Koeffizienten der Differenzgleichung die Methode der rekursiven Regression betrachtet, weil sie sich bereits bei der Erstellung statischer Modelle gegenüber der Relaxation und der stochastischen Approximation als überlegen gezeigt hat. Da die ausführliche Herleitung des Verfahrens bereits im Abschnitt 5.4.3 erfolgte, wird an dieser Stelle nur noch auf die für dieses Modell typischen Probleme eingegangen. Bezüglich der rekursiven Schätzverfahren die in den runden Klammern stehenden Argumente bei den Punkt k der Abtastung und nicht, wie bisher verwendet, die Anzahl der Beobachtungen.

Wie bereits im Abschnitt 5.4.3 ausgeführt, besteht der Grundgedanke der rekursiven Regression darin, daß durch eine neue Beobachtung zum $(k+1)$ -ten Zeitpunkt die Meßwertsätze

$$m^T(k+1) = [-x(k) \dots -x(k-m+1); u(k) \dots u(k-m+1)] \text{ und } x(k+1)$$

für die Verbesserung der Parameterschätzung zur Verfügung stehen.

Auf der Basis der Schätzgleichung von Gl. (6.84) zum k -ten Zeitpunkt

$$\hat{s}(k) = [M^T(n)M(k)]^{-1} M^T(k)x(k) \quad (6.96)$$

mit

$$\hat{s}^T(k) = [\hat{a}_1(k) \dots \hat{a}_m(k); \hat{b}_1(k) \dots \hat{b}_m(k)]$$

$$x^T(k) = [x(k-l) \dots x(k)]$$

$$M^T(k) = \begin{bmatrix} m^T(k-l) \\ \vdots \\ m^T(k) \end{bmatrix}$$

erhält man unter Einbeziehung der neuen Information zum $(k+1)$ -ten Zeitpunkt entsprechend Gl. (6.96)

$$\hat{s}(k+1) = [M^T(k+1)M(k+1)]^{-1} M^T(k+1)x(k+1) \quad (6.97)$$

mit

$$\hat{s}^T(k+1) = [\hat{a}_1(k+1) \dots \hat{a}_m(k+1) : \hat{b}_1(k+1) \dots \hat{b}_m(k+1)]$$

$$x^T(k+1) = [x(k-l) \dots x(k) x(k+1)]$$

$$M^T(k+1) = \begin{bmatrix} M^T(k) \\ m^T(k+1) \end{bmatrix}$$

$$m^T(k+1) = [-x(k) \dots -x(k-m+1) : u(k) \dots u(k-m+1)].$$

Mit diesen für das Modell der Differenzengleichung getroffenen Vereinbarungen kann analog zu den Ableitungen im Abschnitt 5.4.3 die Schätzvorschrift der rekursiven Regression hergeleitet werden. Entsprechend den Gln. (5.82), (5.78) und der im Abschnitt 5.4.3 dargestellten Lösungsmethode lautet die rekursive Schätzvorschrift der Regression

$$\hat{s}(k+1) = \hat{s}(k) + k(k+1)[x(k+1) - m^T(k+1)\hat{s}(k)] \quad (6.98)$$

mit

$$k(k+1) = P(k)m(k+1)[1 + m^T(k+1)P(k)m(k+1)]^{-1}$$

$$P(k+1) = P(k) - k(k+1)m^T(k+1)P(k)$$

$$P(k) = [M^T(k)M(k)]^{-1}.$$

Damit wird wiederum deutlich, daß sich der neue geschätzte Parametervektor $\hat{s}(k+1)$ aus dem alten Schätzwert $\hat{s}(k)$ und der Addition des bewerteten sog. „Prädiktionsfehlers“ $e(k+1)$ ergibt. Der Prädiktionsfehler $e(k+1)$, der sich aus

$$e(k+1) = [x(k+1) - m^T(k+1)\hat{s}(k)] \quad (6.99)$$

berechnet, stellt den Vergleich des abgetasteten Ausgangssignals zum $(k+1)$ -ten Zeitpunkt mit dessen Vorherberechnung auf der Grundlage des alten Parametervektors dar.

Wie bereits im Abschnitt 5.4.3 ausführlich dargelegt, steht dem Vorteil der fehlenden Inversion bei der rekursiven Regression der Nachteil der geeigneten Wahl der Startparameter $\hat{s}(0)$ und $P(0)$ gegenüber. Grundsätzlich bestehen wiederum die Möglichkeiten:

- Abschätzung aus einer Schätzung mit Hilfe der direkten Regression aus wenigen Meßwertsätzen,
- Vorgabe von Startwerten aus der theoretischen Prozeßanalyse oder vorhergegangenen Schätzungen,
- Wahl der Startparameter der Matrix P entsprechend Gl. (5.109)

a) bei geringer Störung: $P(0) = cI$, $10^{10} \leq c \leq 10^{15}$;

b) bei größerer Störung: $P(0) = cI$, $10^3 \leq c \leq 10^6$.

Dann ist die Konvergenz der Parameterschätzung relativ unabhängig vom Startparametervektor $\hat{s}(0)$, der deshalb häufig $\hat{s}(0) = 0$ gesetzt wird.

Sollen die Parameter eines zeitvarianten Systems aktuell ermittelt werden, sind in die Schätzvorschrift der rekursiven Regression von Gl. (6.98) entsprechend der Vorgehensweise im Abschnitt 5.4.3.2 die Matrizen $P(k)$ durch die gewichteten Matrizen $P^*(k)$, die

entsprechend den Gln. (5.110) und (5.111) berechnet werden, zu ersetzen. Die Wichtigkeit und die Größe der Vergessensrate sind entsprechend den Empfehlungen im Abschnitt 5.4.3.2 zu wählen.

Insgesamt kann eingeschätzt werden, daß die Methode der rekursiven Regression dann angewendet werden sollte, wenn

- das System gering gestört ist oder
- die Voraussetzungen für eine erwartungstreue Schätzung mit Hilfe eines Regressionsverfahrens gegeben sind oder
- auf eine erwartungstreue Schätzung verzichtet werden kann, z. B. bei der Realisierung adaptiver Steuerungskonzepte.

6.3.3. Methoden der Hilfsvariablen

6.3.3.1. Direkte Methode der Hilfsvariablen

Wie bereits im Abschnitt 6.3.2 gezeigt wurde, ist mit der Methode der Regression bei einem korrelierten Gleichungsfehler keine erwartungstreue Schätzung der Parameter der Differenzengleichung möglich. Um die Vorteile der Regression zu nutzen und gleichzeitig eine erwartungstreue Schätzung zu erzwingen, wurde die Methode der Hilfsvariablen (instrumentellen Variablen) entwickelt [1.49, 6.14, 6.15].

Diese Methode beruht auf der Verwendung der Fehlergleichung für den korrelierten Fall in der Form - s. Gl. (6.68) -

$$E(z) = \hat{A}(z^{-1})X(z) - \hat{B}(z^{-1})U(z) \quad (6.100)$$

und in Vektorschreibweise für n Beobachtungen

$$e(n) = x(n) - M^T(n)\hat{s}. \quad (6.101)$$

Die unter Verwendung des Gütekriteriums $Q = e^T(k)e(k)$ analog zu Abschnitt 6.3.2 ableitbare Schätzgleichung für die Parameter

$$\hat{s} = [M^T(n)M(n)]^{-1} M^T(n)x(n) \quad (6.102)$$

liefert bekanntlich keine erwartungstreuen Schätzungen. Damit nun bei Verwendung der in Gl. (6.101) angegebenen Beschreibung die Erwartungstreue erzwungen wird, ersetzt man in Gl. (6.102) die Meßwertmatrix $M^T(n)$ durch die Hilfsvariablenmatrix $H^T(n)$. Die Matrix $H^T(n)$ ist dann eine Hilfsvariablenmatrix, wenn sie folgende Bedingungen erfüllt:

$$E\{H^T(n)e(n)\} = 0 \quad (6.103)$$

$$E\{H^T(n)M(n)\} \text{ positiv definit.}$$

Setzt man die Matrix der Hilfsvariablen in die Schätzgleichung der Regression von Gl. (6.102) ein, ergibt sich die Schätzvorschrift der Methode der Hilfsvariablen zu

$$\hat{s} = [H^T(n)M(n)]^{-1} H^T(n)x(n). \quad (6.104)$$

Es ist leicht zu zeigen, daß unter den in Gl. (6.103) angegebenen Bedingungen für die Wahl der Hilfsvariablenmatrix $H^T(n)$ die Schätzung auch bei einem korrelierten Gleichungsfehler $e(n)$ erwartungstreu ist. Analog zu den Betrachtungen im Abschnitt 6.3.2 gilt

$$E\{\hat{s} - s\} = E\{[H^T(n)M(n)]^{-1}\} E\{H^T(n)e(n)\} \quad (6.105)$$

und unter Beachtung der Gl. (6.103)

$$E\{s - \hat{s}\} = 0.$$

Dem Vorteil der geringen Anzahl der zu ermittelnden Parameter und der günstigen numerischen Lösung von Gl. (6.104) steht das schwierige Problem der Wahl der Elemente der Matrix der Hilfsvariablen gegenüber. Allgemein sollten zur Erfüllung der Bedingungen (6.103) die Elemente der Hilfsvariablenmatrix so gewählt werden, daß sie stark mit den ungestörten Signalwerten in der Matrix M und nicht mit den Werten der Störung z korrelieren. Da in der Meßwertmatrix M die Signalwerte der Eingangsgröße bereits ungestört sind, besteht das Problem der Wahl der Elemente der Matrix der Hilfsvariablen vor allem in der Festlegung bzw. dem Ersatz der Signalwerte des gestörten Ausgangs. Folgende Varianten wurden für dieses Problem vorgeschlagen:

Variante 1: Verwendung verschobener Eingangssignale
Für den Vektor der Hilfsvariablen ergibt sich dann

$$h^T(k) = [-u(k-r-1) \dots -u(k-r-m); u(k-1) \dots u(k-m)]. \quad (6.106)$$

Variante 2: Verwendung verschobener Ausgangssignale

Bei dieser Variante wird der Vektor der Hilfsvariablen durch die Verwendung von um p Taktpunkte in die Vergangenheit verschobenen Ausgangssignalen gebildet. Damit gilt

$$h^T(k) = [-x(k-p-1) \dots -x(k-p-m); u(k-1) \dots u(k-m)]. \quad (6.107)$$

Als Intervall für die Verschiebung ist $m \leq p \leq 3m$ sinnvoll.

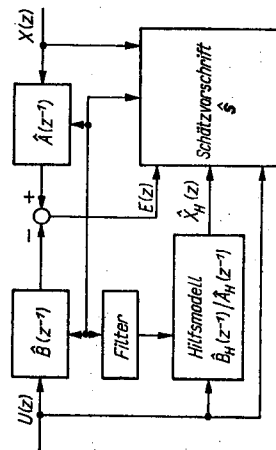


Bild 6.11
Schema der Methode der Hilfsvariablen
bei Verwendung eines Hilfsmodells

Variante 3: Verwendung geschätzter Werte des Ausgangssignals eines Hilfsmodells
Mit den geschätzten Werten $\hat{x}_H(k-1)$ der Ausgangsgröße eines Hilfsmodells (Bild 6.11) erhält man

$$h^T(k) = [-\hat{x}_H(k-1) \dots -\hat{x}_H(k-m); u(k-1) \dots u(k-m)]. \quad (6.108)$$

Die Werte $\hat{x}_H(k-i)$ werden nach der Vorschrift

$$\hat{x}_H(k) = m^T(k) \hat{s}_H \quad (6.109)$$

mit

$$m^T(k) = [-\hat{x}_H(k-1) \dots -\hat{x}_H(k-m); u(k-1) \dots u(k-m)]$$

$$\hat{s}_H^T = [\hat{a}_{1H} \dots \hat{a}_{mH}; \hat{b}_{1H} \dots \hat{b}_{mH}]$$

ermittelt. Der Parametervektor des Hilfsmodells $\hat{s}_H$ sollte bei Anwendung der rekursiven Schätzvorschrift (s. Abschn. 6.3.3.2) aus dem geschätzten Parametervektor $\hat{s}$ durch eine

Gegeben:

Systembeschreibung

$$G(p) = \frac{5}{(1+p \cdot 10s)(1+p \cdot 5s)}$$

Mit $T = 2s$ folgt

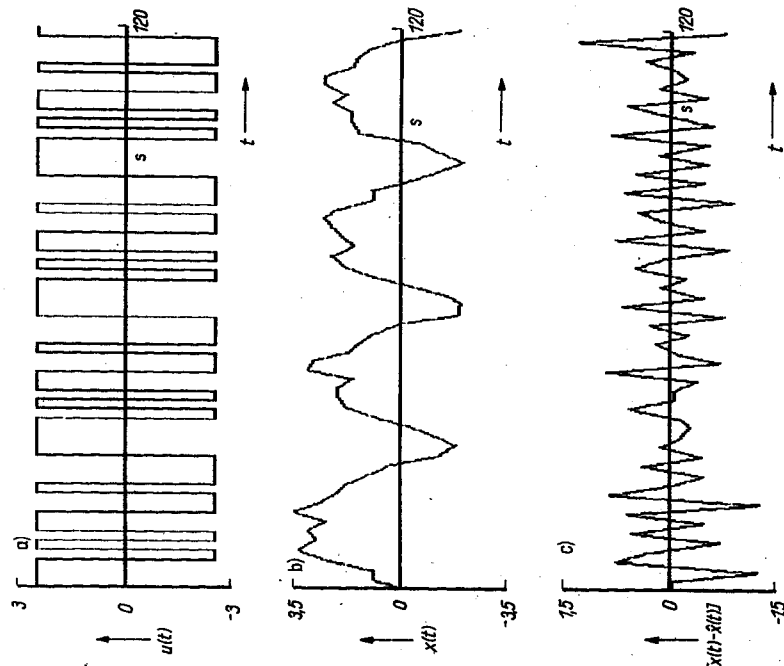
$$G(z) = \frac{0,164z^{-1} + 0,134z^{-2}}{1 - 1,489z^{-1} + 0,548z^{-2}}$$

Testsignal: PRBS-Folge (Länge 157, Amplitude $\pm 2,5$, Gesamtlänge 607),
Störung NV ($0, \sigma_z^2$); hochfrequent,
Störverhältnis $\sqrt{\sigma_z^2/\sigma_{z0}^2} = 0,16$.

Gesucht: Schätzwerte der Koeffizienten der diskreten Übertragungsfunktion und die Reststreuung s_z^2 .

Ergebnis: Aus den Meßwerten des Eingangs- und Ausgangssignals wurden die folgenden Parameter im angegebenen Iterationsschritt geschätzt:

Iterations-schritt	a_1	a_2	b_1	b_2	s_z^2
1	-1,11	0,162	0,162	0,216	0,264
5	-1,20	0,255	0,162	0,203	0,270
10	-1,28	0,333	0,163	0,191	0,283
15	-1,33	0,382	0,163	0,184	0,295
20	-1,36	0,413	0,164	0,180	0,304



Beispiel 6.4 Parameterschätzung eines linearen Differenzgleichungsmodells mit der Methode der Hilfsvariablen

Das Ergebnis zeigt, daß der Parameterfehler bei der Methode der Hilfsvariablen von Iterationsschritt zu Iterationsschritt kleiner wird. Die Ergebnisse im ersten Iterationsschritt sind mit denen der direkten Regression (s. Beispiel 6.3) identisch. Gleichzeitig wird deutlich, daß mit sich verbessernder Parameter-treue die Reststreuung der Schätzung zunimmt. Dies ist auch am Verlauf der Differenz zwischen Modell- und Systemausgang zu sehen. Damit ist das Verhalten der Methode der Hilfsvariablen gegenüber der Regression unter Beweis gestellt. Der durch die notwendigen Iterationen erhöhte Rechenaufwand der Methode der Hilfsvariablen kann durch die Anwendung der rekursiven Variante wesentlich reduziert werden.

Filterung (s. Abschn. 3.4.4.2) der Form

$$\hat{s}_H(k+1) = (1-\alpha)\hat{s}_H(k) + \alpha(k+1) \quad (6.110)$$

mit

$$0,01 \leq \alpha \leq 0,1$$

gewonnen werden, um starke Schwankungen des Parametervektors zu Beginn der Schätzung zu vermeiden. Damit erhält man bei dieser Variante eine Hilfsvariablenmatrix $H(n)$ folgender Form für $k = j, \dots, j+n-1$ Beobachtungen:

$$H(n) = \begin{bmatrix} -\hat{x}_H(j-1) & \dots & -\hat{x}_H(j-m) & u(j-1) & \dots & u(j-m) \\ \vdots & & \vdots & \vdots & & \vdots \\ -\hat{x}_H(j-1+n-1) & \dots & -\hat{x}_H(j-m+n-1) & u(j-1+n-1) & \dots & u(j-m+n-1) \end{bmatrix} \quad (6.111)$$

Die Parameterschätzung erfolgt iterativ in Stufen

- Schätzung des Parametervektors $\hat{s}$ aus n Beobachtungen mit der Regression Gl. (6.84),
- Ermittlung der Hilfsvariablenmatrix $H(n)$ nach Gl. (6.109),
- Schätzung des Parametervektors $\hat{s}$ aus n Beobachtungen mit der Methode der Hilfsvariablen Gl. (6.104),
- Test, ob die Parameteränderung von Iteration zu Iteration $\Delta\hat{s} \leq \text{Toleranzschranke}$ erfüllt ist; wenn nicht, werden erneut die letzten drei Stufen durchlaufen.

Diese Iteration ist immer erforderlich, weil zu Beginn der Schätzung die Störung über die geschätzten Parameter $\hat{s}$ die Hilfsvariablen beeinflusst. Damit sind die Hilfsvariablen mit dem Störsignal und somit auch mit dem Fehlersignal korreliert. Dies hat zur Folge, daß die erste Beziehung von Gl. (6.103) zu Beginn der Schätzung nicht erfüllt ist.

Von den drei vorgestellten Varianten hat sich die dritte in der Praxis aufgrund ihrer Leistungsfähigkeit durchgesetzt. Die direkte Methode der Hilfsvariablen ist bei vorliegenden korrelierten Gleichungsfehlern eines der leistungsstärksten Parameterschätzverfahren, das erwartungstreue Schätzungen absichert bei einem gegenüber der Regression gering erhöhten Aufwand an Rechenzeit und Speicherplatz.

Im Beispiel 6.4 ist die Schätzung der Koeffizienten eines Differenzgleichungsmodells 2. Ordnung mittels der direkten Methode der Hilfsvariablen dargestellt.

6.3.3.2. Rekursive Methode der Hilfsvariablen

Aus den im Abschnitt 6.3.2.2 genannten Gründen und durch analoge Ableitungen kann man aus der Schätzvorschrift der direkten Methode der Hilfsvariablen die rekursive Form der Hilfsvariablenschätzung gewinnen. Entsprechend den Vereinbarungen und Schätzgleichungen für die rekursive Regression von Gl. (6.98) ergibt sich mit Ersatz des Meßwertvektors $m^T(k+1)$ durch den Vektor der Hilfsvariablen $h^T(k+1)$ in der Form

$$h^T(k+1) = [-\hat{x}_H(k) \dots -\hat{x}_H(k-m+1); u(k) \dots u(k-m+1)] \quad (6.112)$$

die rekursive Form der Hilfsvariablenschätzung zu

$$\hat{s}(k+1) = \hat{s}(k) + k(k+1)[x(k+1) - m^T(k+1)\hat{s}(k)] \quad (6.113)$$

mit

$$k(k+1) = P(k)h(k+1)[1 + m^T(k+1)P(k)h^T(k+1)]^{-1}$$

$$P(k+1) = P(k) - k(k+1)m^T(k+1)P(k).$$

Für die Wahl der die Konvergenzgeschwindigkeit der Schätzung beeinflussenden Startwerte $\hat{s}(0)$ und $P(0)$ gelten die für die rekursive Regression getroffenen Aussagen. Das gleiche gilt auch für die Gestaltung der Wichtungsfaktoren bei der Schätzung zeitvarianter Systeme.

Neben der Wahl der Startwerte für $\hat{s}$ und P wird die Konvergenz von der Wahl der Hilfsvariablen bestimmt. In Gl. (6.112) wurden bereits als Hilfsvariablenwerte die aus einem Hilfsmodell entsprechend Gl. (6.109) berechneten ungestörten Ausgangssignalwerte $\hat{x}_H(k-i)$ verwendet (s. Bild 6.11). Sie stellen angenäherte optimale Hilfsvariablen dar [1.45]. Zu Beginn der Schätzung ist es günstig, im Bereich $k = 2m$ den Meßwertvektor $m^T(k+1)$ anstelle des Vektors der Hilfsvariablen $h^T(k+1)$ zu setzen, d.h., die Methode der Regression zu verwenden. Danach wird mit dem dann vorliegenden Schätzwert für den Parametervektor der Hilfsvariablenvektor ermittelt und die Schätzung mit der rekursiven Methode der Hilfsvariablen fortgesetzt. Wird dieser Hinweis nicht beachtet, kann es zu instabilen Schätzungen kommen.

Außerdem ist es vorteilhaft, den geschätzten Parametervektor $\hat{s}(k)$ entsprechend Gl. (6.110) tiefrequent zu filtern, um den Einfluß der Störung $z(k)$ auf die Hilfsvariablen $\hat{x}_H(k)$ zu verringern. Diese Filterung reduziert auch die Korrelation zwischen den Hilfsvariablen und dem Gleichungsfehler. Sie wirkt ähnlich wie der Vorschlag, eine Totzeit um d Tastpunkte zwischen dem geschätzten Parametervektor $\hat{s}(k)$ und dem des Hilfsmodells $\hat{x}_H(k-d)$ vorzusehen [1.51]. Theoretisch muß zur Erlangung der Erwartungstreue die Totzeit gleich der Korrelationstiefe des Fehlers bzw. der Störung gewählt werden. Da das digitale Tiefpaßfilter in den meisten Fällen bereits zur Beseitigung der Korrelation ausreicht, hat sich dieser in [1.52] unterbreitete Vorschlag praktisch nicht durchgesetzt.

Insgesamt kann eingeschätzt werden, daß die rekursive Methode der Hilfsvariablenschätzung gegenüber der rekursiven Regression immer dann angewendet werden sollte, wenn der Gleichungsfehler korreliert ist und erwartungstreue Schätzungen gefordert werden. Der rechenstechnisch geringe Mehraufwand besteht in der Berechnung der Hilfsvariablen aus dem Hilfsmodell entsprechend Gl. (6.109). Gegenüber der direkten Methode der Hilfsvariablen besitzt die rekursive neben den rechenstechnischen Vorteilen die Eigenschaft zum On-line-Betrieb und zur Parameterschätzung zeitvarianter Systeme.

6.3.4. Weitere Parameterschätzverfahren

Neben der Methode der Regression und der Hilfsvariablen sind die „verallgemeinerte Regression“ und die „Maximum-Likelihood-Schätzung“ von besonderer Bedeutung. Beide Methoden versuchen, eine erwartungstreue Schätzung durch eine Filterung des korrelierten Fehlers bzw. durch Verwendung eines erweiterten Modellansatzes zu erreichen.

6.3.4.1. Methode der verallgemeinerten Regression

Bei der Ableitung der *verallgemeinerten Regression* wird von der Fehlergleichung für den korrelierten Gleichungsfehler

$$E(z) = \hat{A}(z^{-1}) X(z) - \hat{B}(z^{-1}) U(z) \quad (6.114)$$

ausgegangen. Der korrelierte Fehler $E(z)$ soll über ein Filter

$$\hat{F}^*(z) = 1/\hat{D}(z^{-1})$$

in den unkorrelierten Fehler $\varepsilon(z)$ umgewandelt werden (Bild 6.12). Damit gilt

$$\varepsilon(z) = \hat{A}(z^{-1}) \tilde{X}(z) - \hat{B}(z^{-1}) \tilde{U}(z) \quad (6.115)$$

mit

$$\tilde{X}(z) = X(z) \hat{D}(z^{-1})$$

$$\tilde{U}(z) = U(z) \hat{D}(z^{-1}).$$

Mit den gefilterten Werten $\tilde{x}(k)$ und $\tilde{u}(k)$ kann dann wieder eine Regression durchgeführt werden. Das Problem ist die Ermittlung des Filters $\hat{F}^*(z)$. Bevor auf die Bestimmung eines geeigneten Filters eingegangen werden soll, werden die grundsätzlichen Beziehungen der verallgemeinerten Regression, basierend auf den Ableitungen im Abschnitt 5.2.4, näher betrachtet.

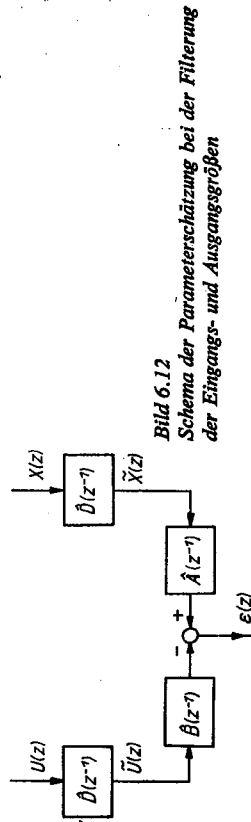


Bild 6.12
Schema der Parameterschätzung bei der Filterung
der Eingangs- und Ausgangsgrößen

Als Gütekriterium wird die bekannte Beziehung

$$Q = e^T(k) W^{-1} e(k); \quad (6.116)$$

W^{-1} Bewertungsmatrix, Filtermatrix,

verwendet, wobei die Bewertungsmatrix so zu wählen ist, daß der korrelierte Fehler $e(k)$ in den unkorrelierten Fehler $\varepsilon(k)$ umgeformt wird. Damit gilt

$$Q = e^T(k) W^{-1} e(k) = \varepsilon^T(k) \varepsilon(k). \quad (6.117)$$

Mit den im Abschnitt 6.3.1 angegebenen System-, Modell- und Fehlergleichungen ergibt sich als Schätzvorschrift für n Beobachtungen (s. Abschn. 5.2.4)

$$\hat{s} = [M^T(n) W^{-1} M(n)]^{-1} M^T(n) W^{-1} x(n). \quad (6.118)$$

Wenn die Bewertungsmatrix W^{-1} in das Produkt zweier Dreiecksmatrizen F^T und F

$$W^{-1} = F^T F$$

zerlegt werden kann, erhält man aus Gl. (6.118)

$$\hat{s} = [\tilde{M}^T(n) \tilde{M}(n)]^{-1} \tilde{M}^T(n) \tilde{x}(n) \quad (6.119)$$

mit

$$\tilde{M}(n) = FM(n), \quad \tilde{x}(n) = Fx(n).$$

Gleichung (6.119) entspricht formal dem Algorithmus der Regression, nur daß gefilterte Meßwerte verwendet werden; s. Gl. (6.84). Da außer den Systemparametern auch noch die Parameter des Filters bestimmt werden müssen, erfolgt die Schätzung im Wechsel von System- und Filterparametern in mehreren Schritten mit der sog. Mehrschritt-methode, weil die Elemente der Wichtungsmatrix W^{-1} (damit auch der Matrix F) in der Regel nicht a priori bekannt sind [6.7, 6.16]. Die Bewertungsmatrix F in Gl. (6.119) stellt ein Filter dar, das die Werte der Eingangs- und Ausgangssignale so transformiert, daß aus dem korrelierten Gleichungsfehler $e(k)$ der unkorrelierte Gleichungsfehler $\varepsilon(k)$ entsteht. Von den vielen Möglichkeiten zur Wahl der Filterübertragungsfunktion wird entsprechend Gl. (6.115)

$$\hat{F}(z) = \frac{1}{\hat{A}(z^{-1}) \hat{D}(z^{-1})} \quad (6.120)$$

für das Verfahren der verallgemeinerten Regression gewählt. Die Schätzung der Koeffizienten a_i und b_i des Systems und der Koeffizienten d_i des Filters erfolgt in den Schritten (Bild 6.13):

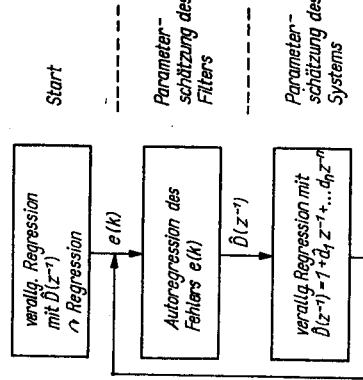


Bild 6.13
Struktur der Mehrschritt-methode bei dem Verfahren
der verallgemeinerten Regression

Schritt 1. Schätzung der Systemparameter a_i und b_i aus den ungefilterten Eingangs- und Ausgangssignalwerten, d.h. $\hat{D}(z^{-1}) = 1$, für n Beobachtungen ($k = j, j+1, \dots, j+n-1$). Die Schätzung entspricht der Methode der Regression.

Schritt 2. Ermittlung der Koeffizienten d_i des Filters auf der Grundlage des mit den geschätzten Parametern $\hat{a}_i$ und $\hat{b}_i$ nach Gl. (6.72) berechneten Fehlers $e(k)$. Für den Zu-

sammenhang zwischen dem korrelierten und dem unkorrelierten Gleichungsfehler gilt im Zeitbereich für die k -te Beobachtung

$$\varepsilon(k) = e(k) + \hat{d}_1 e(k-1) + \dots + \hat{d}_r e(k-r). \quad (6.121)$$

Damit kann aus Gl. (6.121) die Fehlergleichung

$$\varepsilon(k) = e(k) - [-\hat{d}_1 e(k-1) - \dots - \hat{d}_r e(k-r)]$$

für die k -te Beobachtung und das Gleichungssystem für n Beobachtungen in der Form

$$\begin{aligned} \varepsilon(j) &= e(j) & - [-\hat{d}_1 e(j-1) & \dots - \hat{d}_r e(j-r)] \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon(j+n-1) &= e(j+n-1) & - [-\hat{d}_1 e(j+n-1) & \dots - \hat{d}_r e(j+n-1-r)] \end{aligned}$$

aufgebaut werden. In Vektorschreibweise erhält man somit für das Gleichungssystem die Beziehung

$$\varepsilon(n) = e(n) - E^{*T}(n) \hat{d}. \quad (6.122)$$

Bei Verwendung der Methode der Regression erhält man als Schätzvorschrift für die Parameter $\hat{d}$ des Filters aus den Werten des korrelierten Fehlers $e(k)$

$$\hat{d} = [E^{*T}(n) E^*(n)]^{-1} E^{*T}(n) e(n) \quad (6.123)$$

mit

$$\begin{aligned} \hat{d}^T &= [\hat{d}_1, \hat{d}_2, \dots, \hat{d}_r] \\ e^T(n) &= [e(j), e(j+1), \dots, e(j+n-1)] \\ E^{*T}(n) &= \begin{bmatrix} -e(j-1) & \dots & -e(j-r) \\ \vdots & \ddots & \vdots \\ -e(j-1+n-1) & \dots & -e(j-r+n-1) \end{bmatrix}. \end{aligned}$$

Schritt 3. Filterung der Eingangs- und Ausgangssignale mit den neuen Filterparametern entsprechend den Vorschriften von Gl. (6.115) und Schätzung der Parameter $\hat{a}_1$ und $\hat{b}_1$ entsprechend der Beziehung von Gl. (6.119).

Mit diesen neu erhaltenen Schätzungen für die Systemparameter wird die Schätzstrategie ab dem zweiten Schritt wiederholt, bis das Fehlerkriterium von Gl. (6.117) ein Minimum hat.

Werden die Eingangs- und Ausgangssignale optimal gefiltert, wird der Gleichungsfehler in ein „weißes Rauschsignal“ übergehen, und die Parameterschätzung ist erwartungstreu (s. Abschn. 6.3.2). Der relativ hohe rechentechnische Aufwand durch die Inversionen in (6.119) und (6.123) kann wiederum durch eine rekursive Gestaltung der Schätzvorschriften von (6.119) und (6.123) reduziert werden. Im Gegensatz zur blockweisen Abarbeitung der Meßwerte wird das Filter in jedem Schritt neu berechnet und so jeder Meßwert mit den aktuellen Filterparametern nach den Beziehungen

$$\tilde{x}(k) = x^T(k) \hat{d}(k), \quad \tilde{u}(k) = u^T(k) \hat{d}(k) \quad (6.124)$$

mit

$$\begin{aligned} x^T(k) &= [x(k), x(k-1), \dots, x(k-r)] \\ u^T(k) &= [u(k), u(k-1), \dots, u(k-r)] \\ \hat{d}^T(k) &= [1, \hat{d}_1, \dots, \hat{d}_r] \end{aligned}$$

gefiltert. Die Mehrschrittstrategie bleibt auch bei der rekursiven Version erhalten. Für die Schätzvorschrift der Systemparameter gilt analog zu den Beziehungen im Abschnitt 6.3.3.2

$$\hat{s}(k+1) = \hat{s}(k) + k(k+1) [x(k+1) - m^T(k+1) \hat{s}(k)] \quad (6.125)$$

mit

$$k(k+1) = P(k) m(k+1) [1 + m^T(k+1) P(k) m(k+1)]^{-1}$$

$$P(k+1) = P(k) - k(k+1) m^T(k+1) P(k).$$

Die Filterparameter $\hat{d}_i$ werden analog nach der Vorschrift

$$\hat{d}(k+1) = \hat{d}(k) + k'(k+1) [e(k+1) - e^T(k) \hat{d}(k)] \quad (6.126)$$

mit

$$k'(k+1) = \psi(k) e(k+1) [1 + e^T(k+1) \psi(k) e(k+1)]^{-1}$$

$$\psi(k+1) = \psi(k) - k'(k+1) e^T(k+1) \psi(k)$$

$$\psi(k) = [E^{*T}(k) E^*(k)]^{-1}$$

bestimmt. Die mit den Gln. (6.125) und (6.126) realisierte rekursive Schätzung ist in der Regel aber nicht immer asymptotisch erwartungstreu [6.18]. Dem Vorteil der fehlenden Inversion von Matrizen steht die Wahl der Anfangswerte für die System- und Filterparameter sowie der Matrizen $\tilde{P}(k)$ und $\psi(k)$ als Nachteil gegenüber.

Insgesamt kann eingeschätzt werden, daß sich der rechentechnische Aufwand gegenüber den Methoden der Regression und der Hilfsvariablen beachtlich erhöht und dazu geführt hat, daß die praktische Anwendung gering ist.

6.3.4.2. Maximum-Likelihood-Methode

Die Methode beruht auf der Maximierung der Wahrscheinlichkeitsdichtefunktion des unkorrelierten Fehlers $\varepsilon(k)$ unter der Annahme, daß sein Verhalten einer Normalverteilung NV $(0, \sigma_\varepsilon^2)$ entspricht. Für n Beobachtungen gilt damit für die Likelihood-Funktion (s. Abschn. 5.1)

$$L = p(\varepsilon(k) | \hat{s}, u) = \frac{1}{(2\pi)^{j/2} \sigma_\varepsilon^j} \exp \left(-\frac{1}{2\sigma_\varepsilon^2} \sum_{k=j}^{j+n-1} \varepsilon^2(k) \right). \quad (6.127)$$

Das Maximum der Likelihood-Funktion L wird erzielt, wenn der Ausdruck

$$Q = \frac{1}{2} \sum_{k=j}^{j+n-1} \varepsilon^2(k) \quad (6.128)$$

minimal wird.

Für die weitere Betrachtung ist die Strategie der Umwandlung des korrelierten Fehlers $e(n)$ in den unkorrelierten Fehler $\varepsilon(k)$ festzulegen. Die von *Åström* und *Bohlin* [1.38, 6.19] vorgeschlagene Methode geht von der Fehlergleichung entsprechend Gl. (6.68) aus. Der korrelierte Fehler $E(z)$ wird bei dieser Methode über die Beziehung

$$E(z) = \hat{C}(z^{-1}) \varepsilon(z) \quad (6.129)$$

in den unkorrelierten Fehler $\varepsilon(z)$ umgeformt. Damit gilt für die Fehlergleichung

$$\hat{C}(z^{-1}) \varepsilon(z) = \hat{A}(z^{-1}) X(z) - \hat{B}(z^{-1}) U(z) \quad (6.130)$$

oder

$$\varepsilon(z) = \frac{\hat{A}(z^{-1})}{\hat{C}(z^{-1})} X(z) - \frac{\hat{B}(z^{-1})}{\hat{C}(z^{-1})} U(z). \quad (6.131)$$

Aus Gl. (6.131) ist sofort ersichtlich, daß die Parameter $\hat{a}_i$ und $\hat{b}_i$; $i = 1, \dots, m$, linear und die Parameter $\hat{c}_i$; $i = 1, \dots, n$, nichtlinear in den Fehler $\varepsilon(k)$ eingehen. Damit kann das Gütekriterium von Gl. (6.83) nicht mehr direkt analytisch gelöst werden. Eine Möglichkeit besteht in der Anwendung der im Abschnitt 5.3 vorgestellten iterativen Schätzverfahren. Wird als Schätzverfahren für die Parametervektoren

$$\hat{s}^T = [\hat{a}_1, \dots, \hat{a}_m; \hat{b}_1, \dots, \hat{b}_m; \hat{c}_1, \dots, \hat{c}_n] \quad (6.132)$$

entsprechend den Ableitungen im Abschnitt 5.3 die Newton-Raphson-Methode verwendet, gilt analog zu Gl. (5.75)

$$\hat{s}^{i+1} = \hat{s}^i - \left[\frac{\partial}{\partial \hat{s}} \nabla Q(\hat{s}) \right]_{\hat{s}=\hat{s}^i}^{-1} \nabla Q(\hat{s})|_{\hat{s}=\hat{s}^i}. \quad (6.133)$$

Die erforderlichen ersten und zweiten Ableitungen des Gütekriteriums werden unter Beachtung des für dieses Modell geltenden Parametervektors Gl. (6.132) nach den in Gl. (5.76) angegebenen Beziehungen ermittelt. Die iterative Lösung erfolgt ebenfalls nach der im Abschnitt 5.3.2 angegebenen Strategie.

Die Methode der Maximum-Likelihood-Schätzung zeichnet sich durch eine geringe Empfindlichkeit gegen Störungen aus und hat sich im Vergleich zu den anderen vorgestellten Verfahren bei korrelierten Fehlern in vielen Testfällen als überlegen erwiesen [1.1, 1.44, 1.45, 6.20]. Nachteilig wirkt sich der erhöhte Rechenaufwand aus. Die vorgeschlagene rekursive Version der Maximum-Likelihood-Schätzung [1.45] hat sich aufgrund der vielen Näherungen nicht durchgesetzt.

In der Literatur sind eine Vielzahl weiterer Verfahren entworfen und erprobt worden, die teilweise unter Beachtung weiterer Voraussetzungen an die auf das System wirkenden Signale ebenfalls sehr gute und effektive Schätzungen der Parameter der Differenzengleichung gestatten. Genannt seien an dieser Stelle

- Erweiterte Matrizen-Methoden auf der Basis der Gl. (6.73), [1.44, 1.45],
- Tally-Methode [6.21],
- Prior-Knowledge-Fitting-Methode [6.22],
- zweistufige Schätzungen in Form der Ermittlung des parametrischen Modells aus dem nichtparametrischen Modell der Gewichtsfunktion [1.44].

Ihre Anwendungsbreite ist jedoch gegenüber den ausführlicher vorgestellten¹ Verfahren wesentlich geringer.

6.3.5. Parameterschätzung einfacher nichtlinearer Systeme

Wie bereits im Abschnitt 6.3.1.2 erläutert, können viele technische und nichttechnische Prozesse auch in der Umgebung des Arbeitspunktes nicht mehr hinreichend durch lineare Modelle beschrieben werden. Gegenüber den im Abschnitt 6.2 vorgestellten nichtpara-

metrischen Modellen in Form der Volterra-Reihen haben die im Abschnitt 6.3.1.2 genannten nichtlinearen Differenzgleichungsmodelle folgende Vorteile [6.22]:

- Nichtlineare Differenzgleichungsmodelle können direkt oder durch eine Approximation aus einer theoretischen Prozeßanalyse gewonnen werden.
- Die Anzahl der Parameter ist wesentlich geringer als beim Volterra-Reihen-Modell.
- Das Modell ist numerisch leicht zu handhaben.
- Die parametrischen nichtlinearen Modelle sind für die Lösung von Vorhersage- und Steuerungsaufgaben gut geeignet.

Auf der Basis der im Abschnitt 6.3.1.2 angegebenen Modellstrukturen werden in den folgenden Abschnitten die direkte, die zweistufige und die iterative Schätzstrategie zur Ermittlung der Parameter vorgestellt. Bei den Ausführungen wird vorwiegend das Hammerstein-Modell betrachtet, weil es relativ einfache Lösungen ermöglicht und in der Praxis die breiteste Anwendung gefunden hat. Die Lösungen für das Wiener-Modell sind in der weiterführenden Literatur enthalten [6.9].

6.3.5.1. Direkte Schätzstrategien

Als direkte Schätzstrategie soll die Ermittlung aller Parameter der vorgestellten Systembeschreibungen durch eine Schätzgleichung bezeichnet werden. Auf der Grundlage der im Abschnitt 6.3.1.2 abgeleiteten Beziehungen für das Hammerstein-Modell gilt entsprechend Gl. (6.78)

$$\hat{x}(k) = m^T(k) \hat{s} \quad (6.134)$$

mit

$$m^T(k) = [1; -x(k-1) \dots -x(k-m); u(k) \dots u(k-m); \\ \dots; u^p(k) \dots u^p(k-m)]$$

$$\hat{s}^T = [\hat{a}_{00}; \hat{a}_1 \dots \hat{a}_m; \hat{b}_0 \dots \hat{b}_m; \\ \dots; \hat{b}_{p0} \dots \hat{b}_{pm}].$$

Mit diesem umgeformten Modellansatz können die für die Parameterschätzung linearer dynamischer Systeme entwickelten Verfahren direkt angewendet werden, weil die zu schätzenden Parameter linear in die Fehlergleichung eingehen. Als Schätzverfahren können die Methode der Regression entsprechend Gl. (6.84), die Methode der Hilfsvariablen entsprechend Gl. (6.104), die Methode der verallgemeinerten Regression entsprechend Gl. (6.119) und die Maximum-Likelihood-Schätzung entsprechend Gl. (6.133) verwendet werden [1.44, 1.49, 6.9, 6.10, 6.11]. Die in den Abschnitten 6.3.2 bis 6.3.4 getroffenen Einschätzungen zur Leistungsfähigkeit der Verfahren gelten auch für diesen Modellansatz.

Im Beispiel 6.5 ist die direkte Schätzung der Parameter eines nichtlinearen dynamischen Systems mit einer Hammerstein-Struktur und der Methode der Hilfsvariablen dargestellt.

6.3.5.2. Mehrstufige Schätzstrategien

Die mehrstufigen Schätzstrategien gehen davon aus, daß es durch die Gestaltung der auf das System aufgetragten Testsignale möglich ist, die Parameter des linearen dynamischen Teilsystems und die des nichtlinearen statischen Teilsystems getrennt zu schätzen.

Beispiel 6.5	Parameterschätzung eines nichtlinearen Differenzengleichungsmodells mit der Methode der Hilfsvariablen
--------------	--

Gegeben:

a) statisches Teilsystem

$$y = -1 + 2u + 0,5u^2,$$

b) dynamisches Teilsystem

$$G(p) = \frac{1}{(1 + p \cdot 10 s)(1 + p \cdot 5 s)}.$$

Mit $T = 2 s$ folgt

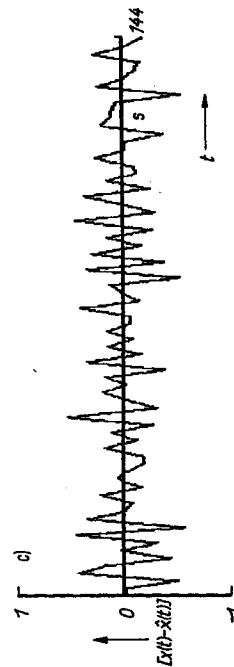
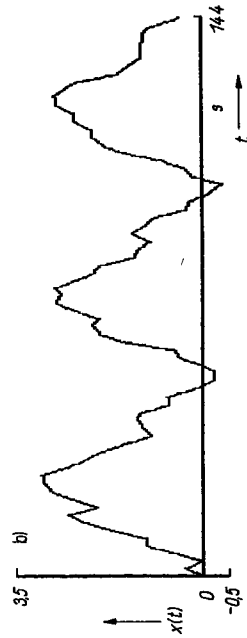
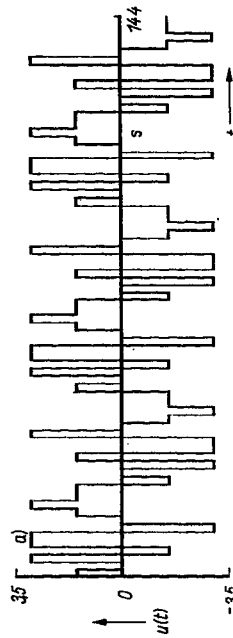
$$G(z) = \frac{0,0328 z^{-1} + 0,0269 z^{-2}}{1 - 1,489 z^{-1} + 0,5488 z^{-2}}.$$

Testsignal: Fünfwertige m -Folge (Länge 247, maximale Amplitude ± 3 , Gesamtlänge 727),

Störung NV $(0, \sigma_z^2)$: hochfrequent,

Störverhältnis $\sqrt{\sigma_z^2 / \sigma_x^2} = 0,13$.

Gesucht: Schätzwerte der Koeffizienten für die Beschreibungen des statischen und dynamischen Teilsystems sowie die Reststreuung s_R^2 .



Beispiel 6.5	Parameterschätzung eines nichtlinearen Differenzengleichungsmodells mit der Methode der Hilfsvariablen
--------------	--

Ergebnis: Aus den Meßwerten des Eingangs- und Ausgangssignals wurden folgende Parameter nach der Auftrennung in die Teilmodelle im jeweiligen Iterationsschritt ermittelt:

Parameterwert	Iterationsschritt 1	Iterationsschritt 20
$\hat{r}_0$	-2,787	-4,402
$\hat{r}_1$	1,864	2,060
$\hat{r}_2$	0,3880	0,527
$\hat{a}_1$	-1,069	-1,424
$\hat{a}_2$	0,123	0,464
$\hat{b}_1$	0,0249	0,0218
$\hat{b}_2$	0,0291	0,0182
s_R^2	0,063	0,077

Die Ergebnisse der Schätzung zeigen, daß die Methode der Hilfsvariablen gegenüber der Regression (erster Iterationsschritt) den Schätzfehler der meisten Parameter reduziert. Der große Fehler des Parameters des Arbeitspunktes $\hat{r}_0$ kann reduziert werden, wenn das Testsignal über längere Perioden den Wert null hat. Am Verlauf der Differenz zwischen System- und Modellausgang ist wiederum zu erkennen, daß bereits eine gute Nachbildung des Systemverhaltens erfolgt.

Die erste Strategie besteht in der Ermittlung der Teilsysteme in folgender Reihenfolge:

Nichtlinearität – Linearität

Diese Vorgehensweise ist für das Hammerstein-Modell besonders geeignet. Für das Wiener-Modell muß eine eindeutig umkehrbare Kennlinie der Nichtlinearität vorausgesetzt werden. Die Parameterschätzung erfolgt in zwei Stufen:

Stufe 1. Ermittlung der statischen Modellparameter r_j aus n Versuchen um den Arbeitspunkt mittels Regression; s. Gl. (5.47).

Stufe 2. Ermittlung der Modellparameter $\hat{a}_i$ und $\hat{b}_i$ für das lineare dynamische Teilsystem aus den mit dem nichtlinearen Teilmodell aus Stufe 1 berechneten Zwischensignalen $\hat{y}(k)$ und den gemessenen Ausgangssignalen $x(k)$. Als Schätzverfahren können die Regression – Gl. (6.84) – und die Hilfsvariablenschätzung – Gl. (6.104) – verwendet werden. Die zu verwendenden Systemgleichungen haben folgendes Aussehen:

a) Hammerstein-Modell

$$x(k) = m^T s + n^*(k)$$

(6.135)

mit

$$m^T(k) = [-x(k-1) \dots -x(k-m) : y(k) \dots y(k-m)]$$

$$s^T = [a_1 \dots a_m : b_0 \dots b_m];$$

b) Wiener-Modell

$$y(k) = m^T(k) s + n^*(k)$$

(6.136)

mit

$$m^T(k) = [-y(k-1) \dots -y(k-m) : u(k) \dots u(k-m)]$$

$$s^T = [a_1 \dots a_m : b_0 \dots b_m].$$

Die zweite Strategie ermittelt die Parameter der Teilsysteme in folgender Reihenfolge:

Linearität – Nichtlinearität

Diese Strategie setzt einen linearen Bereich der statischen Kennlinie voraus und ist dann für das Hammerstein- und Wiener-Modell gleichermaßen geeignet. Die Schätzung wird wieder in zwei Stufen durchgeführt:

Stufe 1. Ermittlung der Parameter $\hat{a}_i$ und $\hat{b}_i$ für das dynamische Teilmodell mit einem Testsignal, das das System nur im linearen Bereich ansteuert. Hierzu werden die Regression oder die Methode der Hilfsvariablen verwendet.

Stufe 2. Ermittlung der Parameter $\hat{p}_j$ aus den berechneten Zwischensignalwerten $\hat{y}(k)$ sowie den Eingangssignalwerten $u(k)$ beim Hammerstein-Modell und den Ausgangssignalwerten beim Wiener-Modell mit der Methode der Regression.

Wie umfangreiche Untersuchungen in [6.10] gezeigt haben, sind für beide Modellstrukturen mit diesen Strategien gute Ergebnisse möglich. Beachtet werden sollte, daß es sinnvoll ist, die Verstärkung des linearen dynamischen Teilsystems auf 1 zu normieren, da die Verstärkung des Gesamtsystems den Teilsystemen nicht eindeutig zugeordnet ist.

6.3.5.3. Iterative Schätzstrategien

Bei den iterativen Schätzstrategien wird davon ausgegangen, daß die Parameter der Teilsysteme auf der Grundlage von n beobachteten Werten der Eingangs- und Ausgangsgrößen $u(k)$ und $x(k)$ abwechselnd geschätzt werden [6.9, 6.10, 6.11]. Das Zwischensignal $y(k)$ wird in die Schätzung einbezogen, und im Gegensatz zu den zweistufigen Verfahren sind keine gesonderten Versuche notwendig, wenn die an dem System anliegenden Signalewerte von $u(k)$ die Amplituden- und Frequenzbereiche des Systems ausreichend aussteuern bzw. anregen. Die iterative Schätzstrategie hat sich besonders für das Hammerstein-Modell bewährt [6.9, 6.11]. Beim Wiener-Modell treten bei Störungen am Ausgang Probleme auf [6.10]. Die Strategie wird in zwei Schritten realisiert:

Schritt 1. Schätzung der Parameter $\hat{a}_i$ und $\hat{b}_i$ des linearen dynamischen Teilsystems aus Ausgangssignalwerten $x(k)$ und den aus den Eingangssignalwerten $u(k)$ und den Parametern des nichtlinearen statischen Teilsystems $\hat{p}_j$ berechneten Zwischensignalen $\hat{y}(k)$:

$$\hat{y}(k) = \sum_{j=0}^p \hat{p}_j u^j(k). \quad (6.137)$$

Als Startwerte werden die Signalwerte $y(k) = u(k)$ verwendet. Damit ergibt sich als Schätzvorschrift analog zu Gl. (6.84)

$$\hat{s}_1 = [M_1^T(n) M_1(n)]^{-1} M_1^T(n) x(n) \quad (6.138)$$

mit

$$m_1^T(k) = [-x(k-1) \dots -x(k-m) : \hat{y}(k) \dots \hat{y}(k-m)]$$

$$s_1^T = [\hat{a}_1 \dots \hat{a}_m : \hat{b}_0 \dots \hat{b}_m].$$

Schritt 2. Schätzung der Parameter $\hat{p}_j$ des nichtlinearen statischen Teilsystems auf der Grundlage der im ersten Schritt ermittelten Parameter des dynamischen Teilsystems. Entsprechend Gl. (6.76) gilt mit den geschätzten Parametern $\hat{a}_i$ und $\hat{b}_i$ für die k -te Beobachtung

$$x(k) + \hat{a}^T x(k-1) = \hat{b}^T U(k) \hat{p} \quad (6.139)$$

mit

$$x^T(k-1) = [x(k-1) \dots x(k-m)]$$

$$U(k) = \begin{bmatrix} 1 & u(k) & u^2(k) & \dots & u^p(k) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u(k-m) & u^2(k-m) & \dots & u^p(k-m) \end{bmatrix}.$$

Mit Anwendung der Methode der Regression erhält man für n Beobachtungen die Schätzvorschrift für die Parameter der statischen Nichtlinearität zu

$$\hat{p} = [M_2^T(n) M_2(n)]^{-1} M_2^T(n) x_2(n) \quad (6.140)$$

mit

$$m_2^T(k) = [\hat{b}_0, \dots, \hat{b}_m : \hat{b}_0 u(k) \dots \hat{b}_m u(k-m) :$$

$$\dots : \hat{b}_0 u^p(k) \dots \hat{b}_m u^p(k-m)]$$

$$x_2(k) = [x(k) + \hat{a}^T x(k-1)].$$

Die Schätzergebnisse werden wieder im ersten Schritt eingesetzt und die Iterationen so lange fortgesetzt, bis Abbruchschranken für die Parameteränderungen bzw. für das Gütekriterium erfüllt sind.

Die Vorteile der iterativen Schätzstrategie bestehen

- in der geringeren Parameteranzahl
- in der fehlenden Parametertransformation
- in der Möglichkeit der sehr guten Einbindung von A-priori-Informationen.

Probleme bestehen bei der Echtzeitmodellbildung und dem Nachweis der Konvergenz [6.9, 6.10].

6.4. Entwurf von Testsignalfolgen

6.4.1. Notwendigkeit und Ziele

Wenn es möglich ist, zur Ermittlung des dynamischen Verhaltens eines Systems die Eingangssignale gezielt zu verändern, muß die Frage nach dem Entwurf von Testsignalfolgen beantwortet werden. In der Vergangenheit haben sich zwei grundsätzliche Wege des Entwurfs herausgebildet.

Der erste Weg geht davon aus, daß für ein gewähltes Modell und ein Schätzverfahren auf der Grundlage von Optimalitätskriterien der Versuchsplanung Testsignalfolgen entworfen werden [6.2, 6.3, 6.4, 6.12, 6.24, 6.25, 6.26].

Der zweite Weg basiert auf der Überlegung, daß das Leistungsdichtespektrum des Testsignals so zu entwerfen ist, daß das Signal den interessierenden Frequenzbereich des Systems ausreichend anregt. Da dieser Bereich selten bekannt ist, wurden Testsignale entworfen, die die Eigenschaften eines weißen Rauschens gut annähern und gleichzeitig einfach zu realisieren sind [1.37, 1.46, 1.47, 6.27 bis 6.29, 3.3, 4.10]. Zu dieser Gruppe von Signalen gehören die pseudostochastischen binären Testsignale (PRBS-Folgen) für lineare dynamische Systeme bzw. die pseudostochastischen mehrwertigen Testsignale (m -Folgen) für nichtlineare dynamische Systeme. Auf ihre Erzeugung und ihrer Anwendung soll nun zunächst eingegangen werden.

6.4.2. Pseudostochastische binäre Testsignale

Zu den in der Vergangenheit am häufigsten angewendeten Testsignalfolgen bei der Identifikation dynamischer Systeme gehören die PRBS-Folgen (pseudo-random-binary-signals). Eine pseudozufällige Binärfolge ist eine zweiwertige Folge, deren Werte nur zu bestimmten, äquidistanten Zeitpunkten wechseln dürfen und die sich nach einer bestimmten Zeit $T_p = LT$ (Periodendauer) wiederholen. Dabei bedeuten L Anzahl der Takte je Periode; T Taktzeit, Abtastintervall.

Sie wurden mit dem Ziel erstellt, diskrete Signale zu erzeugen, die in bestimmten Bereichen die Eigenschaften des weißen Rauschens nachbilden. Sie können unter Verwendung eines rückgekoppelten Schieberegisters, das aus NR Stufen mit den Zuständen

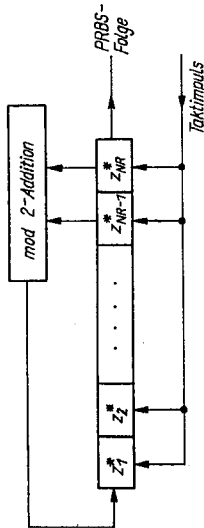


Bild 6.14 Grundstruktur eines Schieberegisters zur Erzeugung von PRBS-Folgen

$z_i^* = (0, 1)$; $i = 1, \dots, NR$, besteht, erzeugt werden. Die Grundstruktur eines solchen Registers ist im Bild 6.14 dargestellt. Für NR Registerstufen erhält man eine PRBS-Folge der Periodenlänge $L = 2^{NR} - 1$. (Der Zustand „Null“ in allen Registerstufen muß ausgeschlossen werden.)

Tafel 6.2 zeigt eine Auswahl in der Praxis häufig verwendeter PRBS-Folgen mit den verwendeten Stufenzahlen NR , den Rückkopplungsfaktoren c_i und der je Periodendauer T_p auftretenden Anzahl verschiedener Binärkombinationen bzw. der Folgenlänge L .

Tafel 6.2 Bedingungen zur Erzeugung einliger PRBS-Folgen mit Schieberegistern

Stufenzahl NR	Rückkopplungsfaktoren c_i	$L = 2^{NR} - 1$
2	1 1	3
3	1 0 1	7
4	1 0 0 1	15
5	0 1 0 0 1	31
6	1 0 0 0 0 1	63
7	0 0 1 0 0 0 1	127
9	0 0 0 1 0 0 0 0 1	511

Die z. B. mit dem vierstufigen Register bei den angegebenen Rückkopplungen und den Anfangsbelegungen der Registerstufen 0 1 1 1 erzeugte Folge hat den im Bild 6.15 dargestellten Verlauf.

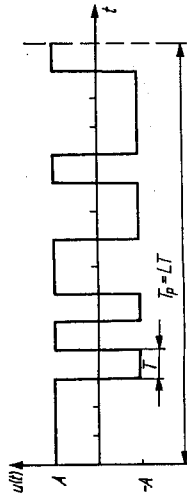


Bild 6.15 PRBS-Folge aus einem vierstufigen Schieberegister

Um den arithmetischen Mittelwert näherungsweise zu null zu machen, wurden die zwei Zustände der durch das Schieberegister erzeugten binären Folge mit $+A$ und $-A$ angenommen. Dann gilt

$$\bar{u} = \frac{1}{L} \sum_{k=1}^L u(k) = \frac{A}{L} \quad (6.141)$$

Für den quadratischen Mittelwert folgt

$$\bar{u}^2 = \frac{1}{L} \sum_{k=1}^L u^2(k) = A^2. \quad (6.142)$$

Die Autokorrelationsfunktion der Folge ist durch die Beziehung

$$R_{uu}(m) = \frac{1}{L} \sum_{i=0}^L u(i) u(i+m) \quad (6.143)$$

definiert und hat den im Bild 6.16 angegebenen prinzipiellen Verlauf.

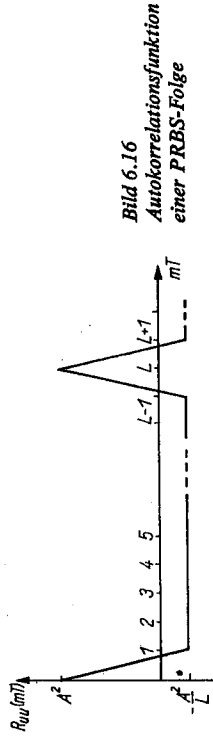


Bild 6.16 Autokorrelationsfunktion einer PRBS-Folge

Die dargestellte Funktion ist periodisch mit der Periodenlänge L . Es gilt

$$R_{uu}(m) = \begin{cases} A^2 & \text{für } m = iL; \quad i = 0, 1, 2, \dots \\ -\frac{A^2}{L} & \text{sonst.} \end{cases} \quad (6.144)$$

Für große Folgenlängen L wird die Funktion immer besser an die Korrelationsfunktion des idealen weißen Rauschens angenähert. Ähnlich verhält es sich mit dem zur Autokorrelationsfunktion im Frequenzbereich äquivalenten Leistungsspektrum. Es ergibt sich aus der Wiener-Chinschinschen Beziehung (s. Abschn. 3.2.3) zu

$$S_{uu}(\omega) = \frac{1}{\pi} \int_{-\infty}^{+\infty} R_{uu}(\tau) e^{-j\omega\tau} d\tau \quad (6.145)$$

und nimmt für PRBS-Folgen folgende Form an:

$$S_{uu}(\omega) = \frac{2\pi A^2}{L^2} \left[\delta(\omega) + (L+1) \sum_{n=-\infty}^{\infty} \sin^2\left(\frac{n\pi}{L}\right) \delta(\omega - n\omega_0) \right]; \quad (6.146)$$

$\delta(\omega)$ Dirac-Impuls.

Gl. (6.146) beschreibt ein Linienspektrum, wobei die einzelnen Spektrallinien an den Stellen $\omega = n\omega_0 = n2\pi/LT$ auftreten.