Statistische Grundlagen neuronaler Netzwerke

Ulrich Anders

Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim

März 1996

1 Einleitung

Nur wenigen Verfahren ist in den letzten Jahren so viel Aufmerksamkeit zuteil geworden wie neuronalen Netzwerken. Neuronale Netzwerken haben ihre Leistungsfähigkeit in den verschiedensten Anwendungsgebieten unter Beweis gestellt. Ihr Einsatz lohnt sich insbesondere bei komplexen Aufgabenstellungen, wie sie zum Beispiel die Prognose oder Erklärung von Finanzmarktdaten, also von Aktienkursen, Wechselkursen oder Optionspreisen darstellt. Selbst bei solch schwierigen Anwendungen konnten neuronale Netzwerke teilweise beachtliche Erfolge vorweisen.[1]

Dennoch wird in der Praxis das Potential neuronaler Netzwerke nur bedingt ausgeschöpft. Neuronale Netzwerke und insbesondere die in den meisten Fällen verwendeten *multilayer perceptron*–Netzwerke (MLP) sind nämlich nichts anderes als eine sehr flexible Klasse von statistischen Regressionsverfahren. Als solche bieten sie die Möglichkeit zu einer fundierten statistischen Analyse. Mittels einer solchen statistischen Analyse lassen sich beispielsweise geeignete Netzwerkarchitekturen bestimmen, die Signifikanz von einzelnen Variablen testen sowie bestimmte Datenprobleme erkennen.

Dieser Beitrag faßt die statistischen Grundlagen neuronaler Netzwerke zusammen. Dabei werden ausschließlich MLP–Netzwerke betrachtet (siehe Abbildung 1). Es eignen sich zwar auch andere Netzwerktypen für die Regressionsanalyse,[2] jedoch erlauben lediglich MLP–Netzwerke die Anwendung statistischer

Inferenztechniken. Die Verwendung von statistischen Methoden für neuronale Netzwerke wird im folgenden mit 'Neurometrie' bezeichnet.

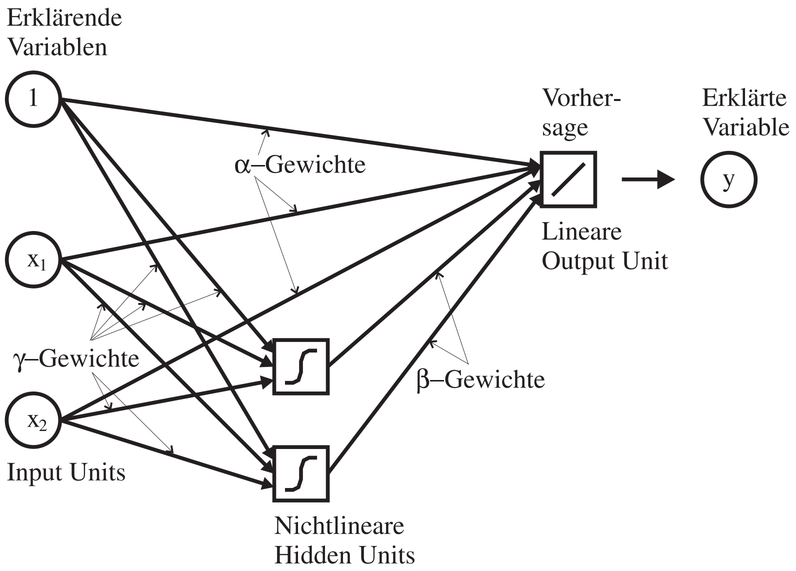

Abbildung 1: MLP–Netzwerk mit drei Schichten.

Jedes neuronale Netzwerk läßt sich durch eine Funktion der erklärenden Variablen $X = [x_0, x_1, \ldots, x_I]$ und der Netzwerkgewichte $w = (\alpha', \beta', \gamma')'$ beschreiben. x_0 ist kon-

[1] Vgl. Rehkugler/Zimmermann (1994) und Anders/Korn/Schmitt (1996).

[2] Z.B. die sogenannten *radial basis function*–Netzwerke (vgl. Poggio/Girosi, 1990).

stant und wird für alle Beobachtungen als $x_0 \equiv 1$ definiert. I bezeichnet die Anzahl der nichtkonstanten erklärenden Variablen, H die Anzahl der verwendeten *hidden units*. Für das in Abbildung 1 dargestellte Netzwerk ergibt sich damit die folgende funktionale Form $f(X, w)$:

$$f(X, w) = X\alpha + \sum_{h=1}^{H} \beta_h g\left(\sum_{i=0}^{I} \gamma_{hi} x_i\right) \tag{1}$$

Die Funktion $g(\cdot)$ stellt die sogenannte Transferfunktion eines Neurons dar und wird üblicherweise als die logistische oder die tangens hyperbolicus Funktion gewählt. Die tanh–Funktion bietet gegenüber der logistischen Funktion zwei Vorteile: erstens ist ihre Ableitung mit $\tanh(x)' = 1 - \tanh^2(x)$ leichter zu berechnen und zweitens ist sie symmetrisch zum Ursprung, so daß $\tanh(x) = -\tanh(-x)$ gilt.

2 Statistische neuronale Netzwerke

Neuronale Netzwerke lassen sich zur Lösung des nichtlinearen Regressionsproblems

$$y = F(X, w) + \varepsilon \tag{2}$$

einsetzen, wobei F einen unbekannten Zusammenhang und ε einen *iid*–Störterm mit $E[\varepsilon\varepsilon'] = \sigma I$, $E[\varepsilon] = 0$ und $E[\varepsilon|X] = 0$ bezeichnet. Neuronale Netzwerke können jede beliebige Funktion bis zu jeder gewünschten Genauigkeit approximieren, vorausgesetzt sie sind hinreichend groß dimensioniert.[3] Insbesondere sind sie also in der Lage, auch die unbekannte Funktion F beliebig genau abzubilden. Damit stellen neuronale Netzwerke also nichts anderes als nichtlineare Regressionsmodelle dar. Der Vorteil bei der Anwendung von neuronalen Netzwerken besteht darin, keine expliziten Annahmen über die funktionale Form des unbekannten Zusammenhangs machen zu müssen. Dennoch trifft man auch bei neuronalen Netzwerken eine Annahme: man unterstellt nämlich implizit, daß die gewählte Netzwerkarchitektur in der Lage ist, die unbekannte Funktion gut zu approximieren.

Deshalb stellt sich die Frage, welche Netzwerkarchitektur für ein gegebenes Problem optimal ist. Die Wahl einer geeigneten Netzwerkarchitektur ist eines der schwierigsten Probleme im Umgang mit neuronalen Netzwerken. Für die Approximation einer sehr unregelmäßig verlaufenden Funktion, muß ein neuronales Netzwerk eine entsprechend komplexere Architektur aufweisen, als wenn der zu approximierende Zusammenhang einen sehr regelmäßigen Verlauf besitzt. Ist ein Netzwerk zu klein dimensioniert, dann ist es nicht flexibel genug, die gesuchte Funktion gut zu approximieren. Die Prognosen von einem Netzwerk mit zu wenigen Parametern werden deshalb immer einen systematischen Fehler machen (sog. *Bias*). Ein zu großes Netzwerk läuft hingegen Gefahr, sich den Störfaktoren

[3] Vgl. Hornik/Stinchcombe/White (1989).

zu sehr anzupassen und deshalb die Approximation der Funktion F zu verwischen. Die Prognosen von einem Netzwerk mit zu vielen Parametern werden daher eine sehr hohe Unzuverläßigkeit (d.h. Varianz) aufweisen. Hinzu kommt, daß jeder geschätzte Parameter in einem Netzwerk mit Unsicherheit behaftet ist. Je mehr Parameter also ein Netzwerk besitzt, desto höher ist die dadurch in das System eingehende Unsicherheit. Redundante Parameter tragen daher zu einer unnötigen Verschlechterung der Prognosequalität bei.

Ziel der Modellbildung sollte also eine Netzwerkarchitektur sein, die gerade groß genug ist, die unbekannte Funktion hinreichend gut zu approximieren. Ein solches Netzwerk weist keinen systematischen Fehler bei der Prognose auf und besitzt die dann kleinstmögliche Prognosevarianz. Die Suche nach einem geeigneten Netzwerkmodell ist ein iterativer Prozeß, der im Abschnitt 3 beschrieben wird.

2.1 Nichtlineare kleinste Quadrate

Zur Bestimmung der Gewichte bzw. Parameter w eines Netzwerks wird in den meisten Anwendungen die Summe der quadrierten Residuen[4] (SSE) der Regression minimiert:[5]

$$\mathrm{SSE}(w) = [y - f(X, w)]'[y - f(X, w)] \to \mathrm{Min!} \tag{3}$$

Mit den Annahmen für das nichtlineare Regressionsmodell (2) und einigen unter gewissen Voraussetzungen erfüllten Regularitätsbedingungen für f — insbesondere müssen die Parameter des Modells eindeutig bestimmbar sein — kann nun bewiesen werden (White, 1989a), daß der Parameterschätzer $\hat{w}$ konsistent ist und eine asymptotische Normalverteilung besitzt. Dieses Wissen erlaubt im Prinzip die Anwendung der üblichen Hypothesentests für nichtlineare Modelle, wie z.B. den Wald–Test oder den LM–Test.

Da neuronale Netzwerke den gesuchten Zusammenhang in den meisten Fällen lediglich approximieren, die funktionale Form des unbekannten Zusammenhangs jedoch nur mit einer geringen Wahrscheinlichkeit in die Funktionsklasse neuronaler Netzwerke fällt, gehören letztere zur Klasse der fehlspezifizierten Modelle. Die zugehörige Theorie ist von White (1994) entwickelt worden. Er beweist, daß die Anwendung asymptotischer Standardtests zulässig ist, wenn man die Fehlspezifikation des Modells bei der Berechnung der Kovarianzmatrix der Parameter berücksichtigt. Die geschätzten Parameter $\hat{w}$ sind normalverteilt mit Erwartungswert w^* und Kovarianzmatrix $\frac{1}{T}C$. Der Parametervektor w^* entspricht der besten Projektion des fehlspezifizierten auf das wahre Modell. Damit ergibt sich:

$$\sqrt{T}(\hat{w} - w^*) \sim N(0, C), \tag{4}$$

[4] *sum of squared errors*

[5] Die kleinste–Quadrate–Schätzung ist eine Spezialfall der *Maximum Likelihood*–Schätzung, wenn die gemeinsame Verteilung der Residuen der multivariaten Normalverteilung mit homoskedastischer Varianz entspricht. Andersherum bedeutet dies, daß bei Verwendung der Methode der kleinsten Quadrate implizit immer die Annahme normalverteilter Residuen getätigt wird. Ist diese Annahme falsch, dann ist die Verwendung der kleinsten–Quadrate–Methode ineffizient.

wobei T die Anzahl der Beobachtungen bezeichnet. Nach der Theorie fehlspezifizierter Modelle berechnet sich die Kovarianzmatrix durch $\frac{1}{T}C = \frac{1}{T}A^{-1}BA^{-1}$. Die Matrizen A und B sind als $A \equiv E[\nabla^2 \mathrm{SE}_t]$ und als $B \equiv E[\nabla \mathrm{SE}_t \nabla \mathrm{SE}_t']$ definiert. SE_t bezeichnet den quadrierten Fehlerbeitrag[6] der t-ten Beobachtung und ∇ den Gradienten in Richtung der Gewichte.

2.2 Numerische Schätzverfahren

Im allgemeinen lassen sich die Parameter eines nichtlinearen Regressionsmodells nicht oder nur sehr schwer analytisch bestimmen. Deshalb ist es in den meisten Fällen notwendig, ein numerisches Verfahren einzusetzen, das die optimalen Parameter der Regressionsfunktion mit Hilfe eines Näherungsverfahrens bestimmen soll. Das Auffinden von Optima in nichtlinearen Funktionen ist ein Standardproblem in der numerischen Mathematik, und so gibt es eine Reihe von sehr effizienten Algorithmen, die für diesen Zweck geeignet sind (z.B. BFGS, DFP, Levenberg-Marquardt, etc.).[7] Der häufig eingesetzte *Backpropagation*–Algorithmus ist diesen Algorithmen aufgrund seiner simplen Funktionsweise in den meisten Fällen stark unterlegen und bräuchte daher keine Verwendung finden.[8]

Bei der Anwendung von numerischen Optimierungsverfahren empfiehlt es sich unbedingt, die Daten zu normieren, da dies die Konvergenz des numerischen Algorithmus in der Regel erheblich erleichtert. Als besonders geeignet hat sich eine Normierung aller Daten auf einen Mittelwert von Null und eine Standardabweichung von Eins herausgestellt. Zum einen ist es damit nicht mehr notwendig, eine Konstante für die *output unit* zu verwenden, da ihr Mittelwert auf Null normiert wurde. Man spart also einen Parameter. Zum anderen liegen fast alle Werte der normierten Variablen im Intervall [-3;3], ein Bereich, in dem die Funktionswerte des tangens hyperbolicus noch nicht saturiert sind.

3 Neurometrische Modellbildung

Die Suche nach einer möglichst guten Netzwerkarchitektur unterscheidet sich grundsätzlich nicht von dem Vorgehen, das üblicherweise in der Ökonometrie oder Statistik Anwendung findet. Die Suche ist ein iterativer Prozeß wie er in Abbildung 2 dargestellt ist. Er wird solange fortgeführt, bis die resultierende Netzwerkarchitektur für die zu lösende Aufgabe geeignet ist.

In der Literatur zu neuronalen Netzwerken wurde der Netzwerkspezifikation in der Vergangenheit viel zu wenig Aufmerksamkeit geschenkt. Das Resultat davon ist, daß eine große Anzahl von Untersuchungen auf einem überparametrisierten Modell basiert bzw. zu viele Inputvariablen verwendet.[9] Grundsätzlich muß die Anzahl der Beobachtungen die Anzahl der freien Modellparameter übersteigen. Wirklich zuverlässige Aussagen über ein Modell

[6] *squared error*
[7] Vgl. Press/Flannery/Teukolsky/Vetterling (1992).
[8] Vgl. Anders (1995)
[9] Z.B. Weigend/Huberman/Rumelhart (1990).

können nur getroffen werden, wenn die Anzahl der Beobachtungen die Parameterzahl um einen Faktor von mindestens zehn übertrifft (White, 1989a).

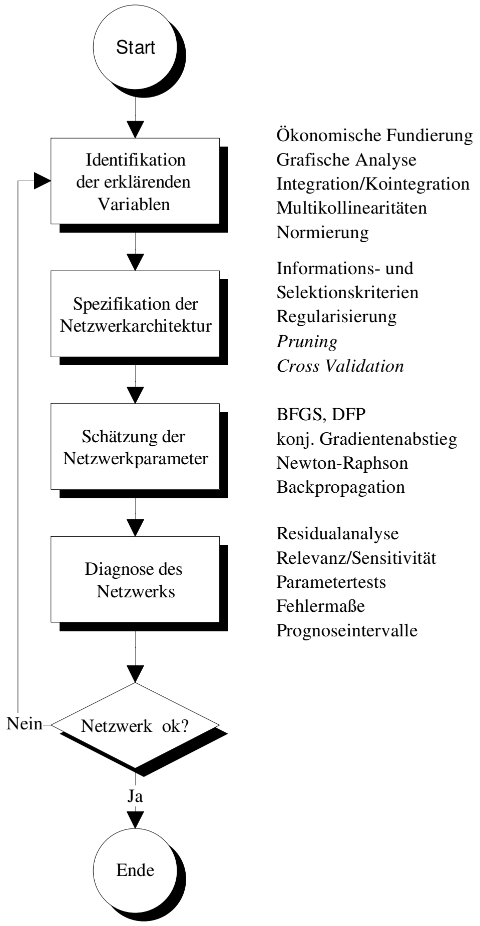

Abbildung 2: Der neurometrische Modellbildungsprozeß.

Die Auswahl der Daten erfolgt bei der Modellierung mit neuronalen Netzwerken genau wie bei herkömmlichen Modellierungsansätzen. Der erste Schritt für jeden Modellbilder sollte darin bestehen, sich über die Qualität seiner Daten zu informieren und die Daten um etwaige Problem zu bereinigen. Darüber hinaus sind der Integrationsgrad von Variablen sowie etwaige Multikollinearitäten in den Variablen zu beachten.

Ein guter Ansatz, die Modellierung eines unbekannten Zusammenhangs mit einem neuronalen Netzwerk zu beginnen, besteht darin, zunächst ein einfaches lineares Regressionsmodell aufzustellen. Die Analyse des linearen Regressionsmodells und der zugehörigen Residuen läßt erste Schlüsse über mögliche Heteroskedastizität oder Autokorrelation zu. Außerdem lassen sich eventuelle Modellinstabilitäten beispielsweise aufgrund eines oder mehrerer Strukturbrüche erkennen. Erst wenn man im linearen Regressionsmodell vernachlässigte Nichtlinearitäten erkennen kann, läßt sich der Einsatz eines neuronalen

Netzwerks rechtfertigen.[10] Dabei kann das lineare Regressionsmodell, wie in Abbildung 1 dargestellt, in das neuronale Netzwerk integriert werden.

Für die Selektion einer geeigneten Netzwerkstruktur stehen sowohl die heuristischen Verfahren der Netzwerkliteratur, wie z.B. die Regularisierung oder das Pruning von neuronalen Netzwerken, als auch die Verfahren der Statistik und Ökonometrie, wie z.B. Hypothesentests auf Parametersignifikanz zur Verfügung. In den folgenden Abschnitten werden die wichtigsten statistischen Verfahren für neuronale Netzwerke dargestellt.

3.1 Signifikanztests für Parameter

Ein wichtiges Instrumentarium zur Diagnose von statistischen oder ökonometrischen Modellen sind Hypothesentests auf die Signifikanz von Parametern. In linearen Modellen ist die Berechnung sogenannter t-Werte bereits Standard, bei neuronalen Netzwerken wird sie jedoch in der Regel unterlassen. Die Berechnung der Parametersignifikanz ist eine große Hilfe bei der Wahl eines geeigneten Modells, denn nicht signifikante Parameter können im Prinzip aus dem betrachteten Modell entfernt werden. Dies gilt für lineare Modelle in gleicher Weise wie für neuronale Netzwerke. Die Errechnung von t-Werten ist für neuronalen Netzwerken jedoch ungleich komplizierter als für lineare Modelle.[11] Die t-Werte sind als

$$\frac{\hat{w}_k - r_k}{\hat{\sigma}_{\hat{w}_k}} \tag{5}$$

definiert, wobei r_k die zu testende Restriktion bezeichnet, die üblicherweise Null ist. Die geschätzte Standardabweichung $\hat{\sigma}_{\hat{w}_k}$ des Gewichts $\hat{w}_k$ steht auf der Hauptdiagonalen der Kovarianzmatrix $\hat{C}_T$. Diese kann mittels der Matrizen $\hat{A}_T$ und $\hat{B}_T$ konsistent geschätzt werden:

$$\hat{C}_T = \frac{1}{T}\hat{A}_T^{-1}\hat{B}_T\hat{A}_T^{-1}. \tag{6}$$

Für den Fall der kleinsten–Quadrate–Methode lassen sich die Matrizen A und B durch nachfolgende Gleichungen errechnen:

$$\hat{A}_T = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2\,\mathrm{SE}_t}{\partial w \partial w'} \text{ und } \hat{B}_T = \frac{1}{T}\sum_{t=1}^{T} u_t^2 \left(\frac{\partial f(X_t, w)}{\partial w}\right)\left(\frac{\partial f(X_t, w)}{\partial w}\right)' \tag{7}$$

[10] Vernachläßtigte Nichtlinearitäten lassen sich beispielsweise mit dem *Neural Network Test* (vgl. Lee/White/Granger, 1993) erkennen.

[11] Da die t-Werte in neuronalen Netzwerken üblicherweise nicht t-verteilt sind, werden sie von Davidson/McKinnon (1993) mit Pseudo–t-Werten bezeichnet.

Da aus Gleichung (4) bekannt ist, daß identifizierte Parameter eines Netzwerks asymptotisch normalverteilt sind, lassen sich jetzt unter Zuhilfenahme der geschätzten Kovarianzmatrix $\hat{C}_T$ Hypothesentests auf die Signfikanz einzelner Parameter eines Netzwerks durchführen. Dafür stehen nach der Theorie der fehlspezifizierten Modelle im wesentlichen ein Wald– und ein LM–Test zur Verfügung.

Vor der Anwendung der Tests muß allerdings sichergestellt sein, daß die Parameter — bis auf symmetrische Lösungen — tatsächlich eindeutig sind. Dies ist nicht der Fall, wenn sich in einem Netwerk irrelevante *hidden units* befinden. Da das β–Gewicht einer irrelevanten *hidden unit* Null ist, können die γ–Gewichte, die in diese Unit führen jeden beliebigen Wert annehmen und sind somit nicht mehr eindeutig bestimmbar. Die Schwierigkeit besteht darin, daß die Verteilung im Fall von nicht eindeutigen Gewichten nicht mehr multivariat normal ist, sondern einer viel allgemeineren Klasse von gemischten Gauß–Verteilungen angehört (Phillips, 1989). Damit ist eine auf der Normalverteilung beruhende Inferenz nicht mehr möglich.

Alle irrelevanten *hidden units* müssen also zunächst aus dem entsprechenden Netzwerk entfernt werden. Die Identifikation von irrelevanten *hidden units* ist jedoch äußert schwierig. Deshalb schlagen Anders/Korn (1996) vor, nur dann eine weitere *hidden unit* in ein neuronales Netzwerk aufzunehmen, wenn diese signifikant zur Erklärung der endogenen Variable beiträgt. Dafür läßt sich ein LM–Test verwenden, wie er von White (1989b) oder Teräsvirta (1993) entwickelt wurde.

Unter der Voraussetzung, daß sich keine irrelevanten *hidden units* in dem Netzwerk befinden, lassen sich beliebige Parameterrestriktionen für die γ–Gewichte eines Netzwerks testen. Besondere Beachtung verdient die Hypothese einer irrelevanten *input unit* $H_0 : R\hat{w} = 0$, wobei R unter den Gewichten $\hat{w}$ diejenigen $\hat{\gamma}$–Gewichte selektiert, die mit der zu testenden *input unit* verbunden sind.

Für das Testen dieser Hypothese verwendet man am besten einen Wald–Test, da dieser das Testen von Parameterrestriktionen im unrestringierten Modell erlaubt. Damit muß das neuronale Netzwerk also nur einmal geschätzt werden. Die Teststatistik errechnet sich durch:

$$(R\hat{w}_T)'(R\hat{C}_T R')^{-1}(R\hat{w}_T) \sim \chi^2_q, \tag{8}$$

wobei q die Anzahl der restringierten Parameter bezeichnet.

3.2 Informations- und Selektionskriterien

Ein alternatives Instrumentarium zur Modellselektion sind die sogenannten Informationskriterien. Sie setzen in der Regel die quadrierten Residuen eines Modells zu der Anzahl der freien Parameter dieses Modells in Beziehung. Die Intention besteht darin, den Fehler eines Modells gegen die Zahl seiner Parameter abzuwägen. Ein zusätzlicher Parameter sollte nur dann in ein Modell aufgenommen werden, wenn dadurch das entsprechende Informationskriterium einen geringeren Wert annimmt, d.h. die Summe der quadrierten

Residuen entsprechend stark reduziert wird. Unter verschiedenen miteinander konkurrierenden Modellen wird dasjenige gewählt, das den kleinsten Wert des entsprechenden Informationskriteriums besitzt.

Das bekannteste Informationskriterien ist das Akaike (1973) Informationskriterium (AIC)

$$\mathrm{AIC} = \ln(\frac{u'u}{T}) + \frac{2K}{T} \tag{9}$$

Dabei werden die Residuen der Regression mit u bezeichnet, K entspricht der Anzahl der freien Modellparamter und T der Anzahl von Beobachtungen. Dieses Kriterium darf aber bei fehlspezifizierten Modellen keine Anwendung finden. Deshalb wurde für neuronale Netzwerke ein eigenes Kriterium entwickelt (Murata/Yoshizawa/Amari, 1994), das Netzwerk Informationskriterium (NIC). Das NIC stellt eine Verallgemeinerung des AIC für den Fall dar, daß sich die wahre Funktion F mit dem gewählten Modell bzw. Netzwerk nicht exakt abbilden läßt.[12] Das NIC errechnet sich bei der kleinsten–Quadrate–Schätzung (3) aus der Summe des *mean squared error* (MSE) der Regression und einem Strafterm für die Anzahl der Parameter:

$$\mathrm{NIC} = \mathrm{MSE} + \frac{1}{T} \cdot \mathrm{tr}[BA^{-1}], \tag{10}$$

Für den Fall, daß das Netzwerk die wahre Funktion F exakt abbilden kann, also nicht fehlspezifiziert ist, läßt sich zwischen den Matrizen die asymptotische Beziehung $B = 2\sigma^2 A$ herleiten, so daß $\mathrm{tr}[BA^{-1}] = 2\sigma^2\,\mathrm{tr}[I] = 2\sigma^2 K$ gilt. Damit reduziert sich das NIC auf die von Amemiya (1980) angegebene Berechnungsformel des AIC:

$$\mathrm{AIC} = \mathrm{MSE} + \sigma^2 \frac{2K}{T} \tag{11}$$

An dieser Stelle soll darauf hingewiesen werden, daß die Anwendung von Informationskriterien auf neuronale Netzwerke theoretisch nur dann möglich ist, wenn die Netzwerke keine irrelevanten *hidden units* besitzen. Es läßt sich daher mit Hilfe des NIC nur entscheiden, welche γ–Gewichte in einem Netzwerk verbleiben sollen.

Dennoch läßt sich mittels des NIC eine Aussage über die Anzahl der *hidden units* treffen: Besitzt das Modell nämlich irrelevante *hidden units*, nimmt das NIC üblicherweise enorm große Werte an. Dies ist eine Konsequenz aus der Tatsache, daß überparametrisierte Modelle eine sehr schlechte Kovarianzmatrix aufweisen. Die Matrizen $\hat{A}$ und $\hat{B}$ weichen stark voneinander ab, wodurch sich die hohen Werte des NIC ergeben.

Neben den Informationskriterien gibt es eine Reihe weiterer Kriterien, die zur Modellselektion eingesetzt werden (z.B. FPE, PC, $\bar{R}^2$). Am bekanntesten ist das Bestimmtheitsmaß R^2, mit dem man versucht, den *Goodness of Fit* der Regression zu messen. Je näher der

[12] Genau wie das AIC ist das NIC kein konsistentes Kriterium.

Wert des Bestimmtheitsmaßes an Eins liegt, desto besser ist der *Fit* der Regression, je näher das Bestimmtheitsmaßes an Null liegt, desto schlechter ist ihr *Fit*.[13] Das Bestimmtheitsmaßes ist folgendermaßen definiert:

$$R^2 = \sum_{t=1}^{T} \frac{(y_t - \hat{y}_t)^2}{(y_t - \bar{y}_t)^2}. \tag{12}$$

3.3 Cross Validation

Ein häufig für neuronale Netzwerke und nichtparametrische Modelle eingesetztes Verfahren ist das *Cross Validation* (CV) (z.B. *v-leave-out*).[14] Mit Hilfe dieses Verfahrens versucht man, den zu erwartenden Fehler bezüglich unbekannter Daten zu prognostizieren. Unter mehreren konkurrierenden Modellen wird dasjenige gewählt, das den kleinsten Vorhersagefehler erwarten läßt. Zur Schätzung dieses Fehlers unterteilt man die Menge aller Beobachtungen in eine Anzahl von M disjunkten Teilmengen mit jeweils v Beobachtungen, wobei für den Parameter v häufig $v = 1$ gewählt wird. Man betrachtet nun sukzessive jeweils eine dieser Teilmengen als Validierungsmenge und schätzt die Parameter eines Modells, ohne die Beobachtungen dieser Menge zu benutzen. Die durchschnittlichen quadratischen Fehler[15] MSPE_m, die sich jeweils bezüglich der einzelnen Validierungsmengen ergeben, werden gemittelt und dienen als Schätzer für den erwarteten Fehler einer Vorhersage. Der *Cross Validation*–Fehler berechnet sich also durch

$$\text{CV} = \frac{1}{M} \sum_{m=1}^{M} \text{MSPE}_m \,. \tag{13}$$

Im Gegensatz zu den Informationskriterien ist das *Cross Validation*–Verfahren sehr rechenintensiv, denn die Parameterschätzungen des Modells müssen jeweils ohne die disjunkten Mengen wiederholt werden. Zudem können die durch dieses Verfahren gewählten Modelle sehr stark voneinander abweichen, je nachdem wie groß v gewählt wurde.

Das Verfahren des *Cross Validation* darf aber nicht mit dem des *Stopped Training* verwechselt werden. Beim *Stopped Training* wird die zur Verfügung stehende Datenmenge in zwei Mengen, eine Trainingsmenge und eine Validierungsmenge, aufgespaltet. In dem sich anschließenden iterativen Schätzprozeß werden die quadratischen Abweichungen sowohl bezüglich der Trainings- als auch bezüglich der Validierungsmenge simultan berechnet. Die Iteration wird abgebrochen, wenn die Summe der quadratischen Fehler bzgl. der Validierungsmenge ihr Minimum erreicht hat.

Die Intention, die hinter *Stopped Training* steckt, besteht darin, das Training dann abzubrechen, wenn das Netzwerk offensichtlich beginnt, die Störterme zu approximieren,

[13] Im nichtlinearen Modell kann R^2 auch außerhalb des Intervals [0;1] liegen.

[14] Vgl. Craven/Wahba (1979) oder Härdle (1990).

[15] *mean squared prediction error*

d.h. das Rauschen erlernen zu wollen (*overfitting*). Obwohl diese Vorgehensweise auf den ersten Blick einsichtig erscheint, hat sie drei entscheidende Nachteile und ist daher wenig empfehlenswert. Zunächst ist klar, daß ein Netzwerk, das in der Lage ist, Rauschen zu approximieren, überparametrisiert ist. Um also die gewünschte Funktion approximieren zu können, sind vergleichsweise weniger Parameter notwendig als in dem Netzwerk verwendet wurden. Die Konsequenz daraus ist, die Anzahl der freien Parameter des Netzwerks zu reduzieren und nicht, den Schätzprozeß zu einem früheren Zeitpunkt, also an einem suboptimalen Punkt der Fehlerfunktion, abzubrechen. Der zweite erhebliche Nachteil besteht darin, daß der Zeitpunkt des Abbruchs sehr stark von der Wahl der Trainings- und Validierungsmenge abhängt, vor allen Dingen dann, wenn Trainings- und Validierungsmenge durch einen Zufallsmechanismus bestimmt worden sind. Der dritte Nachteil liegt darin, daß im Anschluß von *Stopped Training* keine statistische Inferenz mehr möglich ist. Zur Bestimmung der Kovarianzmatrix der Parameter mittels (6) muß die Hessematrix der Zielfunktion positiv definit sein. Dies ist üblicherweise aber nur dann der Fall, wenn ein lokales Minimum der Zielfunktion bestimmt worden ist. Beim *Stopped Training* wird die Konvergenz des Optimierungsverfahrens jedoch abgebrochen, bevor ein lokales Minimum der Zielfunktion erreicht wurde.

3.4 Prognosequalität

Die Qualität jedes ökonometrischen Modells muß sich daran messen, wie gut es zukünftige Entwicklungen vorhersagt. Entsprechend läßt sich der Einsatz eines neuronalen Netzwerks für die Prognose ökonomischer Variablen auch nur dann rechtfertigen, wenn der zu erwartende Prognosefehler kleiner ist als der einfacherer Modelle. Um darüber eine Aussage treffen zu können, ist die *out-of-sample*–Analyse extrem wichtig. Dennoch wird sie in der Literatur neuronaler Netzwerke ausgesprochen selten auf eine zulängliche Art und Weise durchgeführt. In Finanzmarktanwendungen findet man beispielsweise häufig, daß Trefferquoten ausgezählt werden oder daß D-Mark Beträge errechnet werden, die sich mit einer auf einem neuronalen Netzwerk aufbauenden Handelsstrategie hätten erwirtschaften lassen. Es läßt sich mit solchen Angaben jedoch nicht feststellen, ob die gute Modellqualität mehr oder minder nur „zufällig" erreicht worden ist, z.B. durch eine günstige Entwicklung der Daten in der betrachteten *out-of-sample*–Periode. Aus diesem Grund sollten immer auch alternative Modelle oder alternative Strategien betrachtet werden, um die relative Qualität eines Netzwerks beurteilen zu können.

Es existieren zwei Möglichkeiten der Prognose: die Vorhersage des Wertes, der in der Zukunft unter *ceteris paribus*–Annahmen am wahrscheinlichsten angenommen wird, und die Vorhersage eines Konfidenzintervalls, in dem der zukünftige Wert unter *ceteris paribus*–Annahmen mit großer Wahrscheinlichkeit liegen wird. Vorhersagen enthalten auch im korrekt spezifizierten und geschätzten Modell eine erhebliche Ungewißheit aufgrund der Störterme des Modells sowie der Varianz der Parameter. Die Unsicherheit, die jedem Modell immanent ist, stellt die Grundlage für die Konstruktion des Konfidenzintervalls der Prognose dar. Dabei hängt die Breite des Konfidenzintervalls sehr stark von der Parameterschätzung des Modells ab. Je effizienter die Schätzer, desto verläßlicher die Vorhersage des Modells.

3.4.1 Vorhersagefehler

Um den Vorhersagefehler im Erwartungswert zu quantifizieren, verwendet man eine Reihe verschiedener Statistiken, die im wesentlichen auf den geschätzten Residuen $\hat{u}$ basieren. Diese Maße sollten sowohl für die Trainingsmenge als auch für die Validierungsmenge berechnet und miteinander verglichen werden:

- der *Root Mean Squard Error* RMSE $= \sqrt{\frac{\hat{u}'\hat{u}}{T}}$
- Theils $U = \sqrt{(\frac{\hat{u}'\hat{u}}{y'y})}$
- der *Mean Error* ME $= \frac{1}{T}\sum_t \hat{u}_t$
- der *Mean Percentage Error* MPE $= \frac{100}{T}\sum_t \frac{\hat{u}_t}{y_t}$
- der *Mean Absolute Error* MAE $= \frac{1}{T}\sum_t |\hat{u}_t|$
- der *Mean Absolute Percentage Error* MAPE $= \frac{100}{T}\sum_t |\hat{u}_t|/|y_t|$

Die berechneten Statistiken helfen dabei, die Qualität des Modells zu evaluieren. Der RMSE ist der populärste unter den genannten Statistiken, da er mit dem Standardfehler der Regression, also der Standardabweichung der Residuen, unmittelbar vergleichbar ist. Jedoch hat er den Nachteil aller absoluten Maße: er kann nur interpretiert werden, wenn die Größenordnung der Variable y bekannt ist. Der Vorteil der relativen Maße besteht in ihrer Unabhängigkeit von der Größe der Variable y. Jedoch können sie nicht interpretiert werden, wenn einige y_t nahe an Null liegen, und nicht errechnet werden, wenn mindestens ein $y_t = 0$ ist.

3.4.2 Prognoseintervalle

Die Konstruktion eines Vorhersageintervalls basiert auf der Varianz des Vorhersagefehlers. Im linearen Modell werden für die Berechnung des Konfidenzintervalls üblicherweise die Störterme der Regression sowie die Parameterunsicherheit berücksichtigt. Der Vorhersagefehler $\hat{u}_{T+1} = y_{T+1} - \hat{y}_{T+1}$ setzt sich also aus

$$\hat{u}_{T+1} = y_{T+1} - \hat{y}_{T+1} = X_{T+1}(\beta - \hat{\beta}) + u_{T+1} \tag{14}$$

zusammen. Damit ergibt sich die Varianz des Vorhersagefehlers durch:

$$\begin{aligned} \sigma^2_{\hat{u}_{T+1}} &= \mathrm{Var}[X_{T+1}(\beta - \hat{\beta})] + \sigma^2_u \\ &= \mathrm{Var}[\hat{y}_{T+1}] + \sigma^2_u \\ &= X'_{T+1} C_{\hat{\beta}} X_{T+1} + \sigma^2_u, \end{aligned} \tag{15}$$

wobei $C_{\hat{\beta}} = \sigma^2(X'X)^{-1}$ die Kovarianzmatrix der Parameter des linearen Modells darstellt. Das $(1-\alpha)$–Vorhersageintervall der exogenen Variable errechnet sich jetzt durch

$$y_{T+1} \in [\hat{y}_{T+1} - \varphi_{\alpha/2}\sigma_{\hat{u}_{T+1}}; \hat{y}_{T+1} + \varphi_{\alpha/2}\sigma_{\hat{u}_{T+1}}], \tag{16}$$

wobei $\varphi_{\alpha/2}$ den Wert der Standardnormalverteilung am $\alpha/2$–Wahrscheinlichkeitsniveau bezeichnet.[16]

Im nichtlinearen Fall wird die Unsicherheit der Prognose aufgrund der Parameterunsicherheit in vielen Fällen ignoriert. Das dann angegebene Intervall unterschätzt also die wahre Intervallbreite. Eine Berechnung der Varianz von $\hat{y}_{T+1}$ ist jedoch mittels einer linearen Taylorreihenentwicklung um w möglich

$$f(X_{T+1}, \hat{w}) \approx f(X_{T+1}, w) - \nabla f(X_{T+1}, w) \cdot (\hat{w} - w), \tag{17}$$

wobei ∇f den Gradienten der nichtlinearen Regressionsfunktion f bezeichnet.[17] Damit ergibt sich die Varianz des Vorhersagefehlers analog zu Gleichung (15) mit

$$\sigma^2_{\hat{u}_{T+1}} \quad = \quad \nabla f(X_{T+1}, w)' C_{\hat{w}} \nabla f(X_{T+1}, w) + \sigma^2_u \tag{18}$$

und das Prognoseintervall entsprechend der Gleichung (16). Die Kovarianzmatrix $C_{\hat{w}}$ der Gewichte läßt sich entsprechend der Gleichung (6) schätzen.

An dieser Stelle läßt sich noch einmal leicht erkennen, daß es ein Ziel jeder Modellierung neuronaler Netzwerke sein sollte, diese so sparsam wie möglich zu parametrisieren. Redundante Parameter sorgen nämlich für eine unnötige Vergrößerung der Kovarianzmatrix $\hat{C}$, wodurch das Konfidenzintervall des Prognosewertes breiter als notwendig wird.

3.4.3 Bootstrapping

Anstatt die Prognosesvarianz eines Modells zu schätzen, kann sie auch simuliert werden. Ein dazu geeignetes Verfahren ist das *Bootstrapping*.[18] Das Verfahren besteht darin, aus der zur Verfügung stehenden Stichprobe eine Anzahl von (beispielsweise 200) *Bootstrap*–Stichproben zu generieren. Die *Bootstrap*–Stichproben werden dabei mittels eines Zufallsgenerators aus der ursprünglichen Stichprobe „mit Zurücklegen“ gezogen, bis sie dieselbe Größe wie die ursprüngliche Stichprobe erreicht haben. Anschließend wird das zu untersuchende Modell auf der Basis aller *Bootstrap*–Stichproben jeweils neu geschätzt und jeder interessierende Wert berechnet. Insbesondere lassen sich nun auch die verschiedenen Prognosewerte simulieren, woraus sich das Vorhersageintervall der Prognosen ergibt.[19]

[16] Beim üblichen 95%–Wahrscheinlichkeitsniveau ist $\varphi = 1,96$.
[17] Vgl. Seber/Wild (1989).
[18] Vgl. Efron/Tibishirani (1993).
[19] Vgl. Efron/Tibishirani (1986).

Die Verteilung (bzw. das Histogramm) des entsprechenden Schätzwertes erlaubt eine Aussage über Erwartungswert und Varianz dieses Wertes. Damit besteht der große Vorteil des *Bootstrapping*–Verfahrens darin, daß es eine *Bias*– sowie Varianz–Korrektur von Schätzwerten ermöglicht. Leider ist *Bootstrapping* sehr rechenintensiv, da jeweils so viele Modelle neu geschätzt werden müssen wie *Bootstrap*–Stichproben vereinbart wurden.

4 Zusammenfassung

Neuronale Netzwerke sind eine sehr flexible Klasse von statistischen Regressionsverfahren. Der wesentliche Vorteil bei der Anwendung von neuronalen Netzwerken besteht darin, keine expliziten Annahmen über die funktionale Form des unbekannten Zusammenhangs machen zu müssen.

Da neuronale Netzwerke nichts anderes sind als nichtlineare statistsiche Verfahren bieten sie die Möglichkeit zu einer fundierten statistischen Analyse. Die Anwendung statistischer Verfahren auf neuronale Netzwerke wird in dieser Arbeit mit 'Neurometrie' bezeichnet.

Eine neurometrische Modellierung neuronaler Netzwerke ist ausgesprochen hilfreich, wenn man für ein gegebenes Problem eine geeigneten Netzwerkarchitektur auswählen will. Ziel ist die Konstruktion eines neuronalen Netzwerks, das gerade groß genug ist, die unbekannte Funktion zu approximieren. Ein solches Netzwerk weist keine systematische Verzerrung auf und bietet die kleinstmögliche Varianz bei einer Prognose.

Die für die Modellselektion verwendbaren statistischen Verfahren sind: Signifikanztests für Parameter, Informations- und Selektionskriterien und *Cross Validation*. In dieser Arbeit wurde dargestellt, daß die Methode des *Stopped Training* als Methode zur Bestimmung von Netzwerkparametern nicht sinnvoll ist.

Die umfangreiche Diagnose eines trainierten neuronalen Netzwerkes ist außerordentlich wichtig, um festzustellen, ob die Netzwerkarchitektur geeignet ist, einen gesuchten funktionalen Zusammenhang erwartungstreu wiederzugeben. Dazu lassen sich insbesondere die Diagnoseverfahren der Ökonometrie verwenden (z.B. die Residualanalyse).

Die Prognosequalität neuronaler Netzwerke läßt sich mit Hilfe verschiedener Fehlermaße quantifizieren, die die Abweichung im Erwartungswert wiedergeben. Zusätzlich lassen sich Prognoseintervalle angeben, die sich entweder analytisch errechnen lassen oder mittels des *bootstrapping* simuliert werden können.

5 Literatur

Akaike H. (1973): *Information Theory and an Extension of the Maximum Likelihood Principle.* In Petrov B., Csake F. (eds): *2nd International Symposium on Information Theory.* Budapest.

Amemiya T. (1980): *Selection of Regressors.* International Economic Review, 21, 331–354.

Anders U. (1996): *Was neuronale Netze wirklich leisten.* Die Bank, 3, 162–165.

Anders U. (1995): *Neuronale Netzwerke in der Ökonometrie.* ZEW Discussion Paper 95-26.

Anders U., Korn O. (1996): *Model Selection in Neural Networks.* ZEW Discussion Paper, bevorstehend.

Anders U., Korn O., Schmitt C. (1996): *Improving the Pricing of Options — A Neural Network Approach.* ZEW Discussion Paper, 96-04.

Arminger G. (1993): *Ökonometrische Schätzverfahren für neuronale Netze.* In Bol G., Nakhaeizadeh G., Vollmer K.-H. (Hrsg): *Finanzmarktanwendungen Neuronaler Netze und ökonometrischer Verfahren.* Physica–Verlag, 25–39.

Craven P., Wahba G. (1979): *Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation.* Numerische Mathematik, 31, 377–403.

Davidson R., MacKinnon J.G. (1993): *Estimation and Inference in Econometrics.* Oxford University Press.

Efron B., Tibshirani R. (1986): *Bootstrap Methods for Standard Errors Confidence Intervals, and Other Measures of Statistical Accuracy.* Statistical Science, 1(1), 54–77.

Efron B., Tibshirani R.J. (1993): *An Introduction to the Bootstrap.* Chapman & Hall.

Härdle W. (1990): *Applied Nonparametric Regression.* Cambridge University Press.

Hornik K., Stinchcombe M., White H. (1989): *Multilayer Feedforward Networks are Universal Approximators.* Neural Networks, 2, 359–366.

Lee T.-H., White H., Granger C.W.J. (1993): *Testing for Neglected Nonlinearity in Time Series Models.* Journal of Econometrics, 56, 269–290.

Murata N., Yoshizawa S., Amari S. (1994): *Network Information Criterion Determining the Number of Hidden Units for Artificial Neural Network Models.* IEEE Trans. Neural Networks, 5, 865–872.

Phillips P.C.B. (1989): *Partially Identified Econometric Models.* Econometric Theory, 5, 181–240.

Poggio T., Girosi F. (1990): *Networks for Approximation and Learning.* Proceedings of the IEEE, 78(9), 91–106.

Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T. (1992): *Numerical Recipes in C.* Cambridge University Press.

Rehkugler H., Zimmermann H.G. (1994): *Neuronale Netze in der Ökonomie.* Vahlen.

Seber G.A.F., Wild C.J. (1989): *Nonlinear Regression.* John Wiley & Sons.

Teräsvirta T., Lin C.-F., Granger C.W. (1993): *Power of the Neural Network Linearity Test.* Journal of Time Series Analysis, 14(2), 209–220.

Weigend A.S., Hubermann B.A., Rummelhart D.E. (1990): *Predicting the Future: a Connectionist Approach.* International Journal of Neural Systems, 1, 193–209.

White H. (1989a): *Learning in Neural Networks: A Statistical Perspective.* Neural Computation, 1, 425–464.

White H. (1989b): *An Additional Hidden Unit Test for Neglected Non-linearity in Multilayer Feedforward Networks.* Proceedings of the International Joint Conference on Neural Networks, Washington, DC. San Diego: SOS Printing, II, 451–455.

White H. (1994): *Estimation, Inference and Specification Analysis.* Cambridge University Press.