Statistical Model Selection in Neural Networks

Ulrich Anders

Olaf Korn

Centre for European
Economic Research (ZEW)

Mannheim, November 1995

Kaiserring 14-16
Postfach 10 34 34
68163 Mannheim

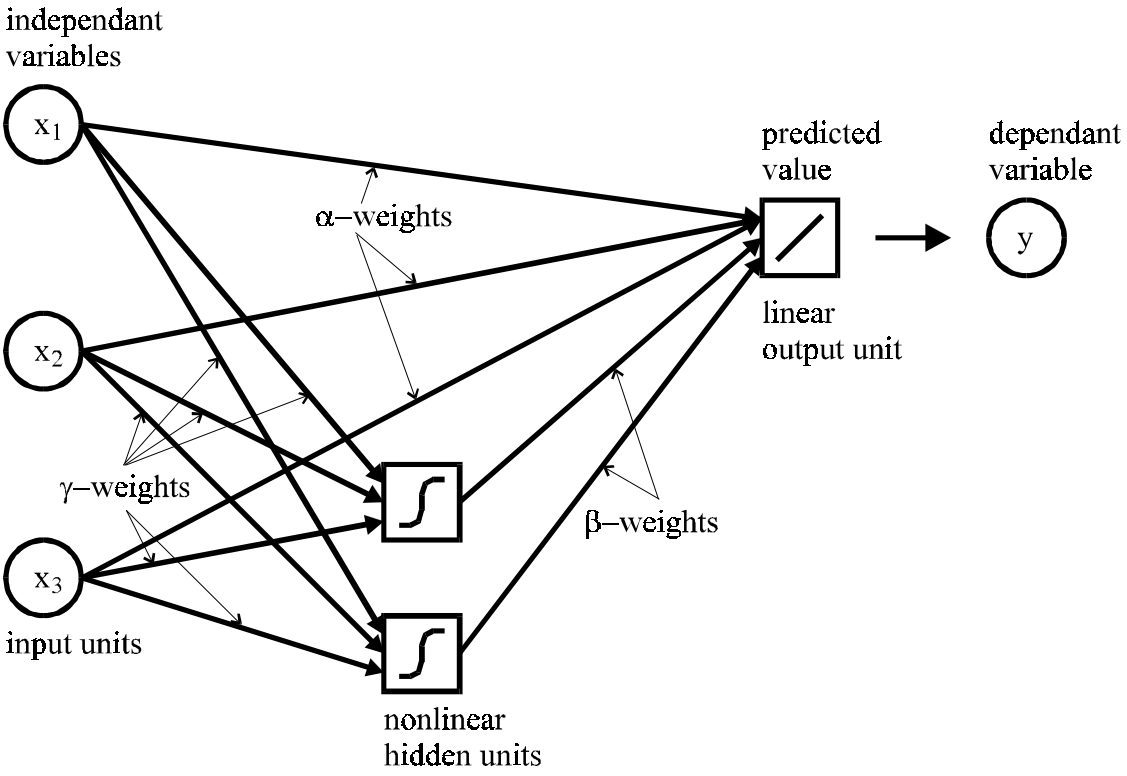
Tel: 0621/1235-141
Fax: 0621/1235-223
Email: anders@zew.de

---

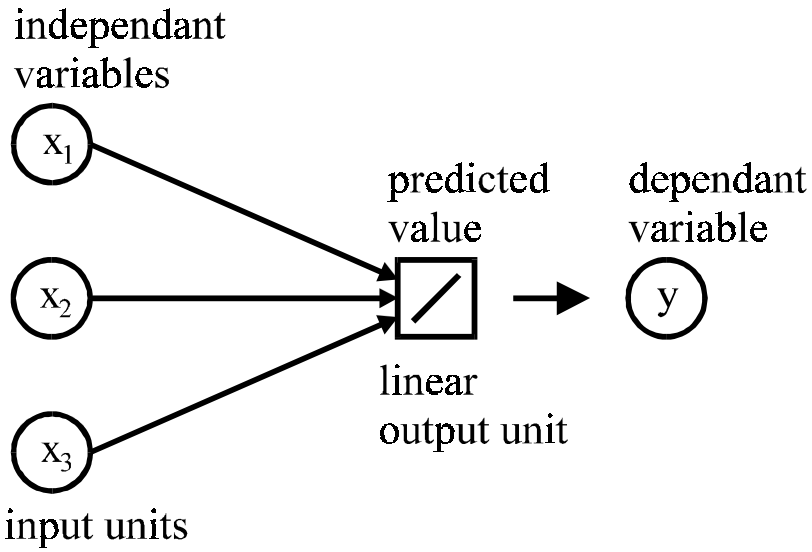
ftp.zew.de
login: anonymous
/home/ftp/pub/persons/Ulrich_Anders/vortraege/nnberlin.ps

© Ulrich Anders

What are neural networks?



A single layer feedforward neural network.



The linear regression model
as neural network

© Ulrich Anders

Glossary

Neural network	Statistical model
Network architecture	Model specification
Unadequate net architecture	Over- or underparameterization
Activation function	Nonlinear transfer function
Inputs	Independant variables
Outputs	Dependant variables
Weights	Parameters
Training	Parameter estimation
Convergence	in-sample quality
Generalization	out-of-sample quality
Pattern	Observation

© Ulrich Anders

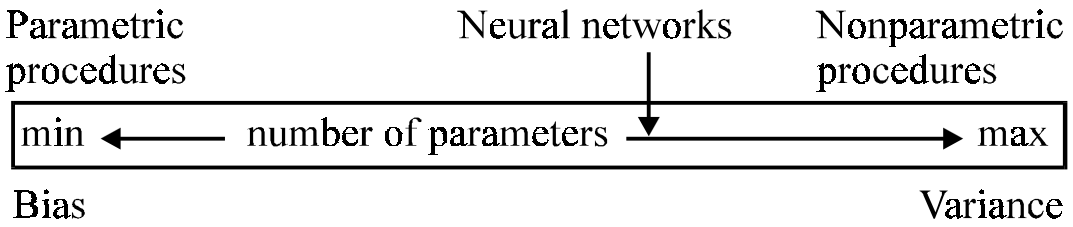
- Neural networks are regression models.
- Typical problem:

$$y = F(X) + \varepsilon \qquad E[\varepsilon|X] = 0,\ E[\varepsilon] = 0,\ \varepsilon \text{ is iid}$$

- Aim: approximation of *F(X)* through $f(X,w)$.

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

- Neural networks are universal approximators.
- Neural networks do not make an explicit assumption about the model structure of *F(X)*. Neural networks are 'atheoretical' models.
- Neural networks are parametric procedures with a potentially nonparametric interpretation.



Continuum between parametric and nonparametric procedures

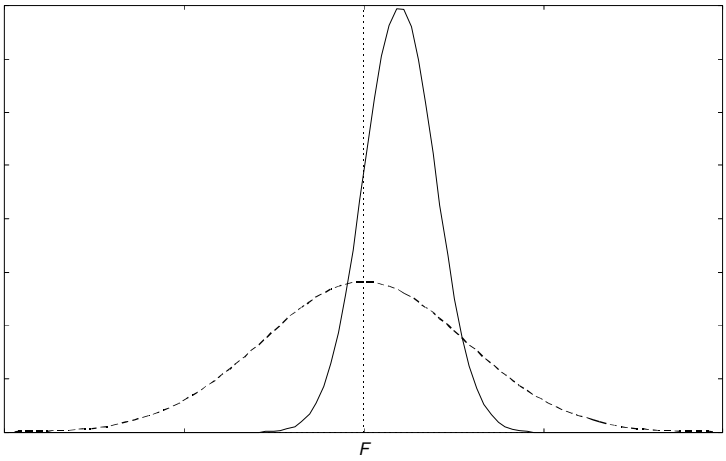
© Ulrich Anders

Excursus: The bias/variance-dilemma

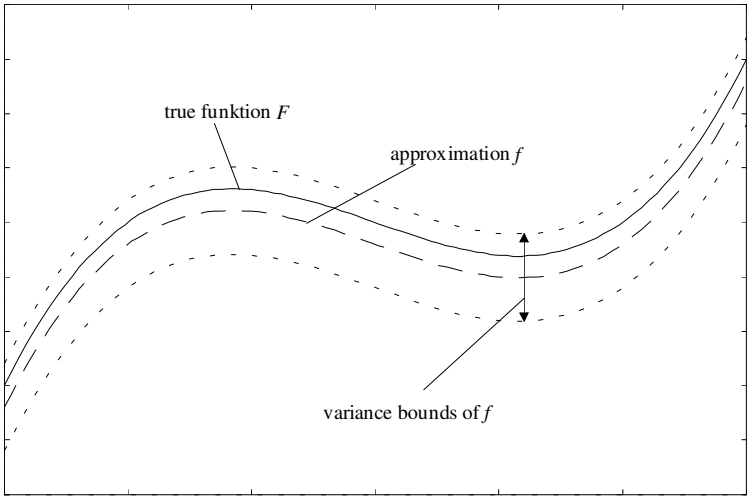
$$\begin{aligned}\mathrm{MSE} &= E[(y - f(X,w))^2] \\ &= E[(y - F(X))^2] + E[(F(X) - f(X,w))^2] \\ &= \mathrm{MSE}_u + \mathrm{MSE}_s\end{aligned}$$

$$\mathrm{MSE}_s = \mathrm{MSE}[f] = \mathrm{Bias}[f]^2 + \mathrm{Var}[f]$$

$$\mathrm{Bias}[f] = E[f] - F \qquad \mathrm{Var}[f] = \tfrac{1}{T}\sum_{t=1}^{T}(f_t - \bar{f})^2$$



Two possible distributions of an estimator f



Bias and standard deviation of a function estimator f

© Ulrich Anders

Training/parameter estimation

- Regression model

  $y = f(X,w) + \varepsilon \qquad \varepsilon \text{ ist } i.i.d. \qquad \text{Var}[\varepsilon] = \sigma \qquad E[\varepsilon] = 0$

- Non-linear least squares (NLS)

  $\text{SSE}(w) = [y - f(X,w)]'[y - f(X,w)] \rightarrow \text{Min!}$

- Maximum likelihood (ML)

  in general:

  $$\ln L(\theta) = \sum_{t=1}^{T} P[\text{E}_t = \varepsilon_t | \theta] \rightarrow \text{Max!}$$

  if errors are assumed to be normally distributed:

  $$\ln L(w,\Omega) = -\ln |\Omega| - [y - f(X,w)]' \Omega^{-1} [y - f(X,w)] \rightarrow \text{Max!}$$

- There is always an assumption about the distribution of the error term! Even in neural networks.

© Ulrich Anders

Asymptotics of neural networks (I)

- If the weights $\hat{w}$ are (locally) identified, then $\hat{w} \sim N(w^*, \frac{1}{T}C)$.

  Since neural networks are 'atheoretical' models, the theory of misspecified models applies (White, 1994) in which

  $\hat{C} = \hat{A}^{-1}\hat{B}\hat{A}^{-1}$ and

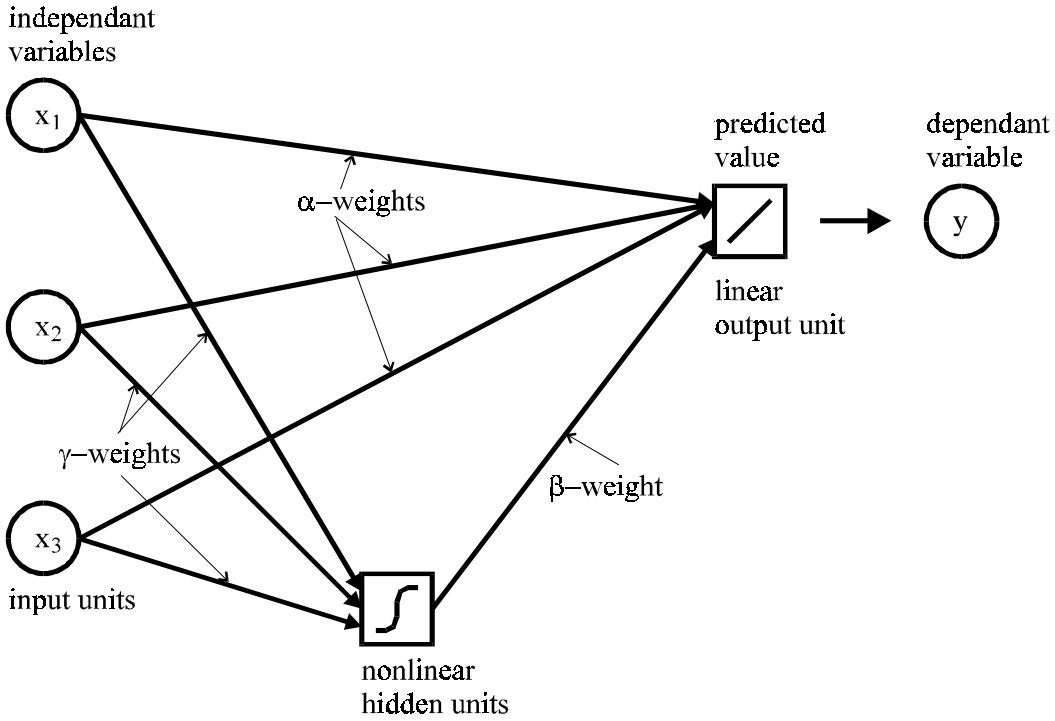
$$A = E\left[\frac{\partial^2 \ln L_t(w)}{\partial w \partial w'}\right] \quad B = E\left[\left(\frac{\partial \ln L_t(w)}{\partial w}\right)\left(\frac{\partial \ln L_t(w)}{\partial w}\right)'\right]$$

- Otherwise limiting distribution of $\hat{w}$ is 'mixed Gaussian' (Phillips, 1989) and further analysis is complicated.

© Ulrich Anders

Asymptotics of neural networks (II)

- Identification problem:



- Workaround by
  – White (1989): LM-test with random sampling
  – Teräsvirta (1993): LM-test with taylor expansion.

- Wald-test

$$T(R\hat{w}-r)'(R\hat{C}R')^{-1}(R\hat{w}-r)\rightarrow\chi^2_q$$

- Network Information Criterion (NIC)

$$\text{NIC} = -\frac{\log L}{T}+\frac{-tr\left[BA^{-1}\right]}{T}$$

© Ulrich Anders

Model selection strategies

1. Sequences of Hypotheses Tests:

   - Estimate ‘nested’ model.
   - LM-Test: Regress residuals on taylor expanded or "random" additional hidden unit.
   - If hidden unit is irrelevant, stop.
   - Else, enlarged model is reestimated and significance of input connections are tested by Wald-Test.

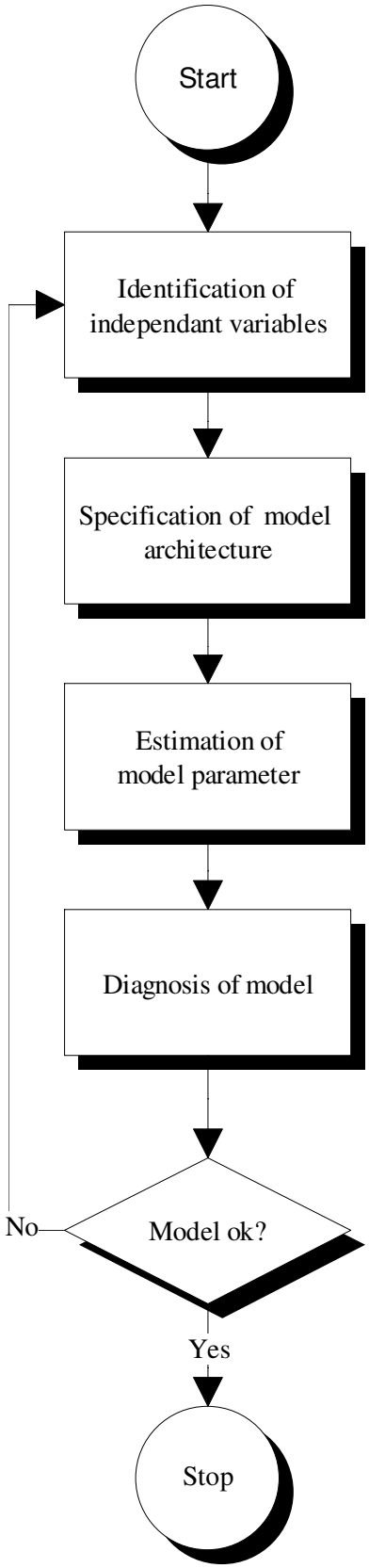
2. Information Criteria:

   - Estimate ‘nested’ model.
   - Regress residuals on taylor expanded additional hidden unit for all combinations of connections.
   - Select combination with lowest AIC, SIC, NIC.

3. Cross Validation

© Ulrich Anders

Model building with neural networks



Theoretical foundation
Graphical analysis
Integration/Cointegration
Multicollinearities
Omitted variables
Normalization

NIC, AIC
Cross validation
Regularization
Pruning
Stopped training

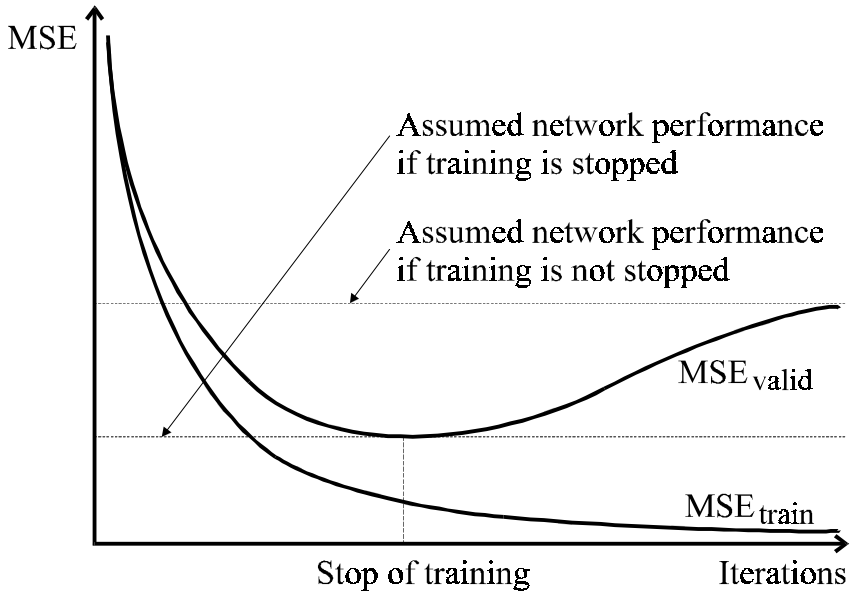
Quasi-Newton (BFGS, DFP)
Conj. gradient descent
Newton-Raphson
Backpropagation

Residual analysis
Two-step estimation
Input relevance/sensitivity
Parameter tests
Measurement of errors
Confidence intervals
Bootstrapping

© Ulrich Anders

Procedures without statistical basis

- Regularization
- Pruning
- Stopped Training



The method of Stopped Training.

© Ulrich Anders

Summary

- Neural networks are a new, very flexible class of statistical models.
- Neural networks do not assumme a 'true' model structure.
- The theory of misspecified models applies.
- Model selection can be supported with parameter inference.
- Modelling of neural networks is an iterative approach in four steps:

    1. Identification of data
    2. Specification of model architecture
    3. Estimation of model parameters
    4. Model diagnosis

- There is a need for simulation studies to give reliable result for modelling strategies.

© Ulrich Anders

Literature

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

Murata N., Yoshizawa S., Amari S. (1994): *A Criterion for Determining the Number of Parameters in an Artificial Neural Network Model.* IEEE Trans. Neural Networks, Vol 5, No 6, 865-872.

Phillips P.C.B. (1989): *Partially Identified Econometric Modells.* Econometric Theory, Vol 5, 181-240.

Teräsvirta T., Lin C.-F., Granger C.W. (1993): *Power of the Neural Network Linearity Test.* Journal of Time Series Analysis, Vol 14, No 2, 209-220.

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

White H. (1992): *Artifical Neural Networks: Approximation and Learning Theory.* Blackwell.

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

© Ulrich Anders