

**Measuring Demand Interdependencies
by Neural Networks**

Martin Natter and Christian Buchta

Forschungsbericht/
Research Memorandum No. 336
November 1993

Contents

1	Introduction	1
1.1	The Multivariate Logit Model	1
1.2	Estimation and Model Selection	2
2	Artificial Neural Networks	4
2.1	Selection of Network Architectures	5
2.2	Comparison to the MLM: a Network without HL	6
2.3	Networks with HL	8
3	Empirical Analysis	9
4	Assessment	10
5	References	12

Zusammenfassung

Neuere Untersuchungen zum Kaufverbund in Handelssortimenten verwenden ein multivariates Logitmodell (MLM) zur Messung der Kaufverbundenheit von Warengruppen und unterstellen symmetrische Verbundbeziehungen. Dabei werden in der Schätzung des MLM vorerst asymmetrische Verbundparameter zugelassen und im nachhinein die Symmetrie der Parameter hergestellt. Wir stellen in unserem Beitrag ein artifizielles neuronales Netzwerkmodell (ANN) zur Messung der Kaufverbundenheit vor und zeigen wie die Nebenbedingung der Symmetrie der Verbundbeziehungen dort berücksichtigt werden kann. Als Test auf Symmetrie verwenden wir einen Likelihood Ratio Test. Für das MLM ist Ridge Regression ein geeignetes Schätzverfahren, da Multikollinearität in den Daten zu erwarten ist, und für das ANN verwenden wir den Backpropagation Algorithmus. Soweit die MLM identifiziert werden konnten werden sie mit den ANN Modellen verglichen. 1669 Beobachtungen gemeinsamer Käufe 72 angebotener Warengruppen aus 32 Handelsfilialen wurden für die Messmodelle verwendet.

Abstract

Recent research efforts about demand interdependencies (DI) in assortments of retail firms are based on a probabilistic model of purchases of article groups, a multivariate logit model (MLM), and assume symmetric interdependencies. Usually the MLM is estimated with asymmetric parameters and symmetry is constructed ex post. We present an artificial neural network (ANN) approach for measuring DI between article groups and show how the restriction of symmetry can be taken into consideration during the ANN estimation process. A likelihood ratio test serves to test the assumption of symmetry. Backpropagation (bp) is an appropriate estimation technique for ANN and ridge regression (rr) is used for the MLM to cope with multicollinearity in the data. We compare the results of ANN and MLM as far as (rr) produced reasonable results. The data base we use consists of 1669 observations of joint purchases of 72 article groups offered in 32 retail outlets.

1 Introduction

Managers of retailing firms have to consider demand interdependencies (DI) when making decisions about their assortments as ignoring these effects results in inefficient assortments. Earlier work on DI (Böcker (1978), Merkle (1981)) is based on bivariate measures of joint purchases of article groups but fails to develop a multivariate model. Also these measures are more of descriptive nature and cannot serve as tool for proper statistical analysis of POS Scanner data.

Hruschka (1991) presented a model for measuring DI in assortments of retail firms based on the concept of contingency table analysis. The model he uses for estimation is a multivariate logit model (MLM) of conditional purchase probabilities. This model serves as a starting point for our investigation.

1.1 The Multivariate Logit Model

The probabilistic model for measuring DI is

$$P(Y_i | Y_{j \neq i}, \dots, Y_n) = \frac{1}{1 + e^{-z}}, \quad i = 1, \dots, n \quad (1)$$

$$z = u_i + \sum_{j \neq i} u_{ij} Y_j \quad (2)$$

$$s.t. \quad u_{ij} = u_{ji}, \forall \quad i, j. \quad (3)$$

Equations (1) and (2) describe a system of conditional logistic models (CLMs) with each of the article groups figuring as a dependent variable whence it is called a “multivariate” logit model (MLM). The Y_i and the Y_j 's are dichotomous purchase variables, i.e. $Y_i = 1$ if article group i is bought and $Y_i = 0$ otherwise.

In the literature on contingency table analysis (Bishop et al. (1975), Hamerle and Tutz (1984)) the terms u_i and u_{ij} are referred to as main and first order interaction effects in analogy to analysis of variance. In the concept of DI the u_{ij} 's are interpreted as measuring the dependence between purchases of article groups i and j . Thus if $u_{ij} = 0$, Y_i and Y_j are said to be independent. If $u_{ij} > 0$, Y_i and Y_j are said to be complementary dependent and if $u_{ij} < 0$, to be substitutionally dependent.

Clearly this interpretation does not imply an effect of Y_i on Y_j although the CLM taken alone seem to support such an interpretation. It is more appropriate to interpret the MLM as a correlation model of binary variables.

In (3) symmetry of the interaction terms is imposed in order to obtain the same parameter estimates as in the log linear model of basket frequencies from which (1) - (2) is deduced. In this context a basket is the joint purchase of article groups of a consumer. The interested reader can find the relationship between MLM and the system of CLMs in (Hruschka (1991), Maddala (1985)).

1.2 Estimation and Model Selection

Since estimation of the MLM is easier if the CLMs are estimated independently, the symmetry restriction is dropped in the estimation and symmetric interaction parameters are constructed by some weighting ex post estimation scheme.

Least squares estimates for the CLM of the purchase of article group i as dependent variable are obtained by minimizing

$$SSE_i(u_i, u_{ij}) = \sum_{k=1}^N (Y_{ik} - P_{ik})^2 \rightarrow \min \quad ! \quad (4)$$

$Y_{ik} = \{0, 1\}$ is the purchase variable of article group i in basket k (k runs over N observations) and P_{ik} is the probability predicted by the i^{th} CLM of a purchase of article group i in basket k .

Minimization is usually done by using ridge regression (rr) in order to alleviate multicollinearity in the Y_j 's which figure as independent variables. Since P_{ik} is a nonlinear function of the u terms iterative techniques have to be used to obtain a solution to (4). A good introduction into such techniques can be found in (Gill et al. 1981) and into nonlinear regression models in (Gallant (1987)).

To restore the symmetry of the interaction parameters Hruschka (1991) suggests to use

$$u_{ij}^* = \frac{u_{ij} + u_{ji}}{2} \quad (5)$$

or alternatively

$$u_{ij}^* = \lambda u_{ij} + (1 - \lambda) u_{ji} \quad (6)$$

$$\lambda = \frac{\sigma_{ij}}{\sigma_{ij} + \sigma_{ji}} \quad (7)$$

$$\sigma_{i,j} = \sqrt{\sigma^2 f_{jj}} \quad (8)$$

$$\sigma^2 = SSE_i / N. \quad (9)$$

In (5) u_{ij}^* is simply the arithmetic mean of the corresponding parameters and in (6) - (9) u_{ij}^* is a variance weighted mean. σ_{ij} is an estimator of the variance of u_{ij} in the CLM for the purchase of article group i as a dependent variable. f_{jj} is the j^{th} diagonal element of the inverse of the Jacobian of (4) and σ^2 is an estimator of the variance of the CLM.

Following this approach model fit is being sacrificed if the CLMs are large (a lot of u_{ij} 's unequal zero) and if the u_{ij} terms in the pairwise corresponding CLMs differ considerably. So we expect a better fit if we obey the symmetry restriction in the estimation.

Up to this point measuring DI relies on standard techniques. Model selection is a more difficult task because the search space of the MLM is of the order $2^{n(n-1)}$ in the asymmetric case and $2^{n(n-1)/2}$ if the symmetry constraint is imposed. Usually n is in the range of 100 to 250 article groups for real scanner data.

Model selection consists of hypotheses testing and directing the search in model space according to the outcome of the tests. In our investigation we used a likelihood ratio test for nested models

$$\Lambda = N[\log(SSE_R) - \log(SSE_U)] \quad \text{with} \quad \chi_{\alpha}^2, |S_U| - |S_R|. \quad (10)$$

S_U is a set of u_{ij} terms not restricted to be zero. Corresponding to the model S_U is the hypothesis H_U that the u terms not included in the set S_U are zero. If H_R is a subset of H_U then H_R is nested in H_U and (10) serves as a test that the u_{ij} terms in $S_U \cap S_R^c$ are all zero. The model implied by S_U is said to be the unrestricted model and S_R to be the restricted model.

In terms of DI this means: when selecting interaction parameters for removal from a model the hypotheses of a substitutional or a complementary dependence of article groups i and j is tested against the hypothesis of independence of the article groups

$$H_0 : u_{ij} = 0, \quad H_c : u_{ij} > 0, \quad H_s : u_{ij} < 0. \quad (11)$$

A critical value for a candidate-for-removal u_{ij} term is

$$t = \frac{u_{ij}}{\sqrt{\sigma^2 f_{jj}}} \quad \text{with} \quad t \sim N(0, 1). \quad (12)$$

t is asymptotically normal, with σ^2 and f_{jj} as in (8). In the artificial neuronal network (ANN) approach (12) is made of different ingredients but has the same asymptotic properties.

Since we did not use the search strategy suggested by Hruschka (1991) we refer the interested reader to his article and defer the discussion of the search strategy until we have introduced the ANN approach.

Unfortunately rr was unstable for certain MLM when we tried to use the search strategy used in the ANN approach. Maybe this is due to over-parametrization at the beginning of the search because ANN starts with all the u_{ij} terms present. Since in these cases information for directing future search got lost we had ANN do the search work and restricted our investigation to estimating models identified by ANN with rr .

Similar problems arose if we tried to impose the symmetry constraint in ridge regression. For that reason we defer this topic to later investigation and make the comparisons with the constructed terms.

2 Artificial Neural Networks

This article does not provide a general survey of ANN. The interested reader may consult the literature. Rumelhart and McClelland (1986) or Hertz et al. (1991), for instance, give introductions to the theory of ANN.

In our analysis of DI we make use of a specific type of ANN, so called feed-forward neural networks, where nodes are divided into layers and connections only run one way: from input variables to output variables or from input variables to hidden variables and from hidden variables to output variables, respectively. Therefore layers are either called input layers (IL), hidden layers

(HL) or output layers (OL). The first model we consider consists of two layers (IL and OL), in the second model we use an additional hidden layer.

Linear Models without hidden units can only separate the input space into hyperplanes and are not able to learn simple functions like the exclusive or (xor, a binary function that returns true when one of its inputs is true, and false otherwise). These models are restricted to learn the sort of so-called linearly separable functions (Minsky and Papert (1969)).

Werbos (1974) provided an algorithm for training networks with hidden layers. This algorithm, back propagation (bp), uses squared error and steepest descent. The novel feature of bp is the use of (first order) dynamic feedback in combination with those two components. Dynamic feedback is a method for calculating the derivatives of some function of the inputs and outputs of a feedforward system, in a single pass through the system (Werbos (1988)).

It turned out that bp has some appealing properties: Denker et al.(1987) proved that a net with a hidden layer can compute any Boolean function and Hornik et al. (1989) and Cybenko (1989) even proved that feedforward neural networks with nonlinear transfer functions and one hidden layer are sufficient to approximate any continuous function to an arbitrary small error. However, we should not always use hidden layers or a large number of hidden units as this can potentially harm the ability of generalization. In that case the network can only reproduce the data it was trained on but is unable to produce similar outputs given similar inputs.

2.1 Selection of Network Architectures

The selection of the effective number of parameters (weights) in the network and an architecture that avoids overfitting has become an active area of research (Himmelblau (1991), Moody (1992) or Utans et al. (1991)). Cross-validatory assessment of models (Stone (1974)) as well as the Bayesian approach (MacKay (1991)), well proved statistical model selection methods, are also recommended for neural network model selection.

Determining an effective number of parameters for the networks we use for measuring the DI is equivalent to testing the hypotheses formulated in equation (11).

We start the model selection process by assuming interdependencies between all article groups. This is recommended (Hendry and Richard (1982)) because

this encompasses all rival models. We apply the bp-procedure to estimate the parameters of this model. To overcome multicollinearity we use the method of weight decay proposed by Hinton and Le Cun (1987). This corresponds to a quadratic complexity cost, known in the statistics community as ridge regression. There is a number of possibilities to fasten the bp algorithm. We use a momentum term to avoid oscillations in the estimation process.

Every weight of the network corresponds to one of the hypotheses about demand interdependencies. To test the hypotheses we calculate the Hessian matrix which consists of the second order derivatives of the error measure with respect to the weights in the network. Because the approximations of the Hessian (Le Cun et al. (1990)) may not be sufficiently accurate (Mackay (1991)) we exactly calculate the Hessian matrix (Bishop (1992)) to perform t-tests for all weights in the network. The critical value t for a weight w_{ij} is

$$t = \frac{w_{ij}}{\sqrt{2\sigma^2 h_{ii}}} \quad \text{with} \quad t \sim N(0, 1) \quad (13)$$

where σ^2 is the variance of the residuals and h_{ii} is the i^{th} diagonal element of the inverse of the Hessian for w_{ij} . We then eliminate all insignificant weights and use bp to get new estimates for the remaining network. Starting with a low level of significance we iteratively raise the significance level, get new estimates by bp and perform t-tests until all parameters are significant at the 5% level.

2.2 Comparison to the MLM: a Network without HL

Again purchases of article groups are binary coded. $Y_i = 1$ for a purchase of article group i and $Y_i = 0$ for a non-purchase. We train the network to get an estimate, \hat{Y}_i , for the conditional probabilities (Bergman et al.(1992)) of purchases of an article group i given purchases of the other article groups j . If we have n article groups this becomes:

$$\hat{Y}_i = P(Y_i | Y_{j \neq i}, \dots, Y_n) = f(w_{ii} + \sum_{j \neq i} w_{ij} Y_j) \quad (14)$$

where f represents the logistic transfer function. If z_i denotes the weighted sum of input variables this expression can be written as:

$$z_i = (w_{ii} + \sum_{j \neq i} w_{ij} Y_j) \quad (15)$$

$$\hat{Y}_i = P(Y_i | Y_{j \neq i}, \dots, Y_n) = \frac{1}{1 + e^{-z_i}} \quad (16)$$

As can be seen expressions (1) and (16) are identical. Like rr bp tries to minimize half of the sum of the squared differences between actual Y_i and the estimated purchase probability \hat{Y}_i over all observations N :

$$E_i = \frac{1}{2} \sum_{k=1}^N (Y_{ik} - \hat{Y}_{ik})^2 \rightarrow \min! \quad (17)$$

The structure of feedforward networks can be formulated as a set of equations or in form of a graph as in figure (1). For the case of three article groups we get the three equations:

$$\hat{Y}_1 = f(w_{11} + w_{12}Y_2 + w_{13}Y_3) \quad (18)$$

$$\hat{Y}_2 = f(w_{22} + w_{21}Y_1 + w_{23}Y_3) \quad (19)$$

$$\hat{Y}_3 = f(w_{33} + w_{31}Y_1 + w_{32}Y_2) \quad (20)$$

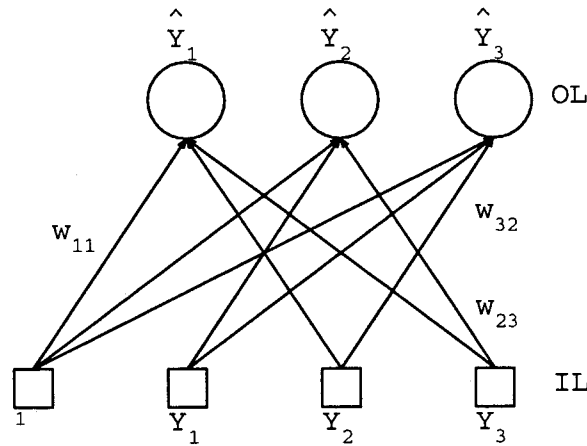


Figure 1: graph of the ANN model forecasting purchase probabilities

As we pointed out in the introduction Hruschka (1991) constructs symmetry ex post. This poses a problem: if, e.g., purchases of an article group i are dependent on purchases of article group j but purchases of j are independent of purchases of i , the resulting model must contain insignificant parameters. Because if we retain the significant parameter in, say, CLM i we have to retain

the insignificant parameter in CLM j . Clearly we could drop both parameters to the effect that there is no symmetric parameter for i and j but with a loss of fit of the CLM with the significant parameter. If there are a lot of such asymmetric relationships in the MLM this could have an impact on the reliability of a decision model which uses parameter estimates obtained by the symmetry constructing method.

This motivates us to test the symmetry of the weights. In testing we compare asymmetric to symmetric estimates of the weights. The adaptive nature of bp makes it easy to consider constraints on the parameters in multiple equation systems: After every iteration of the bp-algorithm we set the weights to:

$$w_{ij,new} = w_{ji,new} = \frac{w_{ij,old} + w_{ji,old}}{2} \quad (21)$$

Equivalently we could compute the new weights by adding half of the sum of the two gradients of each weight to the old weight. If the weights are asymmetric this leads to different gradients in the next iteration, but it drives the weights into the direction of the weight with the larger gradient. Weights change until they have the same absolute gradient. Bp terminates when the error measure E stabilizes. We test the results obtained by imposing the symmetry constraint against the results obtained by dropping it.

2.3 Networks with HL

As we pointed out above, the MLM and networks without hidden units are less powerful than networks with hidden units. To check if the MLM and ANN without hidden units are able to reproduce all the relevant information, we compare these restricted models with ANN with hidden units, applying the same model selection procedure as for the network without hidden units. DI are assumed to be more complicated functions in this case. Again Cross-validation or the Hessian can be used to determine the effective number of hidden units.

The structure of the network can be formulated in functional form:

$$\hat{Y}_o = f(w_{oo} + \sum_h w_{oh} f(w_{ii} + \sum_{i \neq h} w_{hi} Y_i)) \quad (22)$$

where o refers to output variables, i to input variables and h to hidden variables, or as a graph as shown in figure (2).

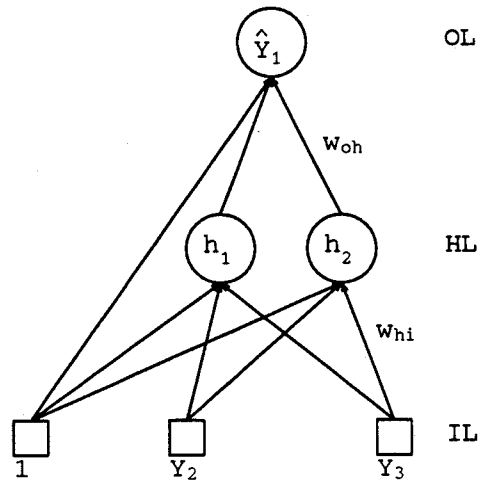


Figure 2: graph of the ANN model with 2 hidden units

3 Empirical Analysis

The sample we use for the different measurement models in our investigation consists of 1669 observations of joint purchases of article groups from 32 retail outlets. The number of article groups offered by the outlets was 72. After deletion of 15 article groups that were purchased in less than 35 cases 57 article groups remained for analysis.

Table 1: Model Comparison

Modell	P	SSE
asymmetric model NET_{IO}	497	8111
symmetric model NET_{IO}	277	8127
asymmetric MLM	497	8112
constructed symmetry MLM	277	8141
CLM for Y_{49}	13	219
NET_{IO} for Y_{49}	13	219
NET_{IHO} for Y_{49}	21	210

Table (1) shows the results for the best symmetric model (NET_{IO}) that can be identified by ANN. The sums of the squared errors (SSE) are rounded to the next integer. There are 277 significant parameters in this model and the sum of squared errors is 8127. If we drop the symmetry constraint in NET_{IO}

the sum of squared errors improve to 8111, but the chi-square statistic for 220 degrees of freedom of the likelihood ratio test shows that there is no difference between the two models at the 5% level. The degrees of freedom are the difference in the number of parameters of the two models, i.e half of the sum of the interaction parameters of the asymmetric model.

If we compare in a similar way the sum of squared errors of the asymmetric *MLM* and constructed symmetry *MLM*, we see that we have to reject the hypothesis that there is no significant difference between the two models at the 5% level. As we have pointed out construction of symmetric u_{ij} terms leads to worse results. Since the sum of squared errors of the unconstrained *MLM* and the *ANN* are nearly identical we think that imposition of the symmetry restriction on the *MLM* would produce a similar result as in the constraint *ANN*.

Finally we mention that six u_{ij} terms in the asymmetric *MLM* differ in their significance, i.e. in one model u_{ij} is significant in the other it is not.

The last three rows of table (1) show the sum of squared errors of the *CLM* of article group 49 as obtained by ridge regression, a single network without hidden layer and one with a hidden layer. Comparing ridge regression with the network without a hidden layer we see that they are practically identical. If a hidden layer is introduced squared errors improve only by about 5% at the cost of a more complex structure of the network, i.e. compared to the *CLM* there are eight additional parameters in the network. For other article groups we obtained similar results.

4 Assessment

Measuring DI offers a lot of difficult tasks. Estimation of encompassing models (probably an effect of correlated independent variables), the large number of rival models or the consideration of the symmetry restrictions are some of them. We presented a feedforward neural network without hidden units that has the same structure as a *MLM*. The network model is estimated by a variant of the bp algorithm, the *MLM* by ridge regression.

Starting with an encompassing model ridge regression has sincere problems with some of the logit equations whereas bp turns out to be more robust. Though once we found an efficient number of parameters by applying the iterative network model selection procedure (estimation by bp, calculation of the

Hessian and eliminating insignificant parameters, increasing the significance level) ridge regression was faster in producing similar results (SSE, parameter estimate and t-values).

Restrictions like symmetry can easily be considered in the network without HL. For the data at hand the results obtained by ANN showed that the hypothesis of symmetric DI cannot be refused.

Feedforward networks with hidden units are not limited to reproduce linearly separable functions like the MLM or networks without HL. DI between groups of articles are then more complex functions. The use of more than two hidden units was not appropriate (overfitting). For most of the equations the model selector refused hidden units at all. So our conclusion for the data at hand is that the problem of measuring DI is linearly separable and that it is sufficient to use networks without hidden layers as the appropriate measurement model.

5 References

- Bishop, C.** (1992): Exact Calculation of the Hessian Matrix for the Multilayer Perceptron. In: *Neural Computation* 4, 494-501.
- Bishop, Y.M.M., Fienberg, S.E., Holland, P.** (1975): *Discrete Multivariate Analysis*, MIT Press.
- Bergman, A., Grassberger, P., Meyer, T.P.** (1991): Forecasting Probabilities with Neural Networks. In: *Nonlinear Modeling And Forecasting*, editors: Casdagli, M. and Eubank, S., Addison Wesley, 305-315.
- Böcker, F.** (1978): *Die Bestimmung der Kaufverbundenheit von Produkten*.
- Cybenko, G.** (1989): Continuous Value Neural Networks with Two Hidden Layers are Sufficient. In: *Mathematics of Control, Signal and Systems*, 303-314.
- Denker, J., Schwartz, D., Wittner, Solla, S., Howard, R., Jackal, L., Hopfield, J.** (1987): Large Automatic Learning, Rule Extraction and Generalization. In: *Complex Systems*, 1, 877-922.
- Gallant, A.R.** (1987): *Nonlinear Statistical Models*. Wiley; New York, Chichester, Brisbane, Toronto, Singapore.
- Gill, P.E., Murray, W., Wright, M.H.** (1981): *Practical Optimization*. Academic Press; London, New York, Toronto, Sydney, San Francisco.
- Hamerle, A., Tutz, G.** (1983): Zusammenhangsanalyse in Mehrdimensionalen Kontingenztabellen. In: *Multivariate Statistische Verfahren*, ed. Fahrmeier, Hamerle, Walter de Gruyter; Berlin, New York, 473-574.
- Hendry, D.F., Richard, J.** (1982): On the Formulation of Empirical Models in Dynamic Econometrics. In: *Journal of Econometrics* 20. North-Holland.
- Hertz, J.A., Palmer, R.G., Krogh, A.S.** (1991): *Introduction to the theory of neural computation*. Lecture Notes Volume I, Addison Wesley.
- Himmelblau, D.M.** (1990): Introducing Efficient Second Order Effects into Back Propagation Learning. In: *Proceedings of the International Joint Conference on Neural Networks*, Volume I; ed. Caudill, Maureen; Hillsdale, New Jersey, Hove, London, Lawrence Erlbaum Associates.

- Hornik, K., Stinchcombe, M., White, H. (1989):** Multilayer Feedforward Networks are Universal Approximators. In: *Neural Networks*, 2, 359–366.
- Hruschka, H. (1991):** Bestimmung der Kaufverbundenheit mit Hilfe eines probabilistischen Meßmodells. In: *Zeitschrift für betriebswirtschaftliche Forschung* 43 (5/1991).
- Le Cun, Y., Denker, J.S., Solla, S. (1990):** Optimal Brain Damage. In: *Advances in Neural Information Processing Systems*, Vol. 2; Touretzky, D.S. ed., Morgan Kaufmann, San Mateo, CA, 598–605.
- MacKay, D.J.C. (1991):** A Practical Bayesian Framework for Backprop. Networks. In: *Neural Computation*. 4(3), 448–472.
- MacKay, D.J.C. (1991):** Bayesian Interpolation. In: *Neural Computation*. 4(3), 415–447.
- Maddala, G.S. (1983):** *Limited-dependent and qualitative variables in econometrics*, Cambridge University Press, 103–108.
- Merckle, E. (1981):** Die Erfassung und Nutzung von Informationen über den Sortimentsverbund in Handelsbetrieben.
- Minsky, M.L., Papert, S. (1969):** *Perceptrons*, Cambridge, MA:MIT Press.
- Moody, J.E. (1992):** The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems. Appears in: Moody, J.E., Hanson, S.J., and Lippmann, R.P., editors, *Advances in Neural Information Processing Systems 4*, Morgan Kaufmann Publishers, San Mateo, CA.
- Rumelhart, D.E., McClelland, J. (1986):** *The PDP Research Group: Parallel Distributed Processing: Explorations in the Microstructure of Cognition; Volume 1: Foundations*; Cambridge, MA: MIT Press.
- Stone, M. (1974):** Cross-Validatory Choice and Assessment of Statistical Predictions. In: *Journal of the Royal Statistical Society, Series B*, 111–133.
- Utans, J., Moody, J. (1991):** Selecting Neural Network Architectures via the Prediction Risk: Application to Corporate Bond Rating Prediction. First International Conference on Artificial Intelligence Applications on Wall Street; *IEEE Computer Society Press*, Los Alamitos, CA.
- Werbos, P. (1974):** *Beyond Regression: New tools for prediction and analysis in the Behavioral Sciences*. Ph.D. thesis, Harvard University.

Werbos, P. (1988): Generalization of Backpropagation with Application to a Recurrent Gas Market Model. In: *Neural Networks*, Vol 1, 339–356.

White, H. (1989): *Connectionist Nonparametric Regression. Multilayer Feed-forward Networks Can Learn Arbitrary Mappings.* Working Paper, Department of Economics, University of California San Diego.