Flexible brand choice models which are not restricted to linear deterministic utility functions offer a way to prevent biased estimates due to nonlinear effects of predictors. One type of flexible choice models approximates deterministic utility by a multilayer perceptron. We describe several variants of this model type which differ with respect to assumptions about error terms and heterogeneity of coefficients. After dealing with appropriate estimation methods, model evaluation criteria and tools for interpreting estimation results, we give an overview of studies applying such models to brand choice data.

Keywords
Brand choice, neural net, multinomial logit, multinomial probit

1. Introduction

In marketing, stochastic choice models serve to investigate the effects of marketing variables on brand choice decisions of households. The majority of publications specifies deterministic utility as linear function. In other words, most authors postulate that deterministic utility is a linear combination of predictors like price, advertising, sales promotion or other marketing variables (e.g., Allenby/Ginter 1995; Allenby/Lenk 1994; Allenby/Lenk 1995; Chintagunta 1992; Gensch/Recker 1979; Guadagni/Little 1983; Gupta 1988; Kalwani et al. 1990; Lattin/Bucklin 1989; McCulloch/Rossi 1994; Winer 1986).

This commitment to the linear function leads to biased estimates if effects of predictors on deterministic utility are nonlinear. Nonlinearities could take the form of thresholds (i.e., utility changes only after a marketing instrument passes a minimum or maximum value), saturation (i.e., utility does not change if a marketing instrument is above a certain value) or interactions (e.g., the effect of an instrument depends on the value of another marketing instrument).

To avoid biases caused by sticking to linear functions, several authors investigated the use of more flexible models which are free from the assumption of a fixed parametric functional form. One finds different approaches based on general additive models (Abe 1999; Abe/Boztuğ/Hildebrandt 2004; Hildebrandt/Boztuğ 1999; Hruschka/Fettes/Probst 2004), linear splines (Schindler/Baumgartner/Hruschka 2007), multilayer perceptrons (Bentz/Merunka 1996; Bentz/Merunka 2000; Fish et al. 2004; Hruschka et al. 2002; Hruschka/Fettes/Probst 2004) or multivariate kernels (Briesch/Chintagunta/Matzkin 1996; Hildebrandt/Boztuğ 1999).

These publications are restricted to homogeneous models whose coefficients or functions are equal across households. A few publications introduce heterogeneous models, whose coefficients or functions differ between households. Briesch/Chintagunta/Matzkin (2002) specify heterogeneous brand constants in a multinomial logit (MNL) model whose deterministic utility also comprises a homogeneous general nonlinear function. Hruschka (2007) develops a multinomial probit (MNP) model which approximates deterministic utility by a multilayer perceptron with heterogeneous coefficients. In a similar manner Kim/Menzefricke/Feinberg (2007) use linear splines with heterogeneous knots in their MNP model.

This paper focuses on choice models which approximate deterministic utility by a multilayer perceptron (MLP) with a single layer of hidden units. We call these models MLP choice models in the following. The MLP represents the most frequently used type of artificial neural net. We do not consider studies which analyze consumers’ choices by multilayer perceptrons minimizing the error sum of squares as these studies do not obey the utility-maximizing choice modeling paradigm (Heimel et al. 1998; Hu/Shanker/Hung 1999; West et al. 1997).
2. MLP Choice Models

The data analyzed in brand choice studies refer to Ti purchases for each household i = 1,..., I. This data consists of the brand chosen out of a choice set of J brands and predictors like retail price, feature or display for each of the brands considered.

Stochastic choice models are based on the assumption that households decide on purchases by maximizing utility. At each purchase occasion t = 1,..., T, each household chooses the brand it perceives to have the highest stochastic utility of the brands belonging to the choice set. Stochastic utilities are formed by adding an error term to deterministic utility. Depending on the distribution of error terms we obtain different forms of choice models, of which the most popular ones in the marketing literature are the MNL and the MNP models.

2.1. Deterministic Utility

By collecting coefficients in (p x 1) vector $\beta_1^i$, (p x H) matrix $\beta_2^i$ and (H x 1) vector $\beta_3^i$ we can specify deterministic utilities computed by a MLP choice model with one layer of H hidden units as:

$$V_{ijt} = \alpha_{ijt} + X_{ijt}\beta_1^i + g(X_{ijt}\beta_2^i)\beta_3^i$$

(1)

As can be seen from expression (1), MLP extensions of choice models may have a linear part $\alpha_{ijt} + X_{ijt}\beta_1^i$ in addition to the proper nonlinear part $g(X_{ijt}\beta_2^i)\beta_3^i$. Coefficients connecting predictors to hidden units are contained in $\beta_2^i$, coefficients connecting hidden units to deterministic utility in $\beta_3^i$.

In marketing, so-called activation values of each of the H hidden units are usually computed by means of the non-polynomial logistic function. $g(Z)$ in expression (1) returns the logistic function $1/(1+exp(-z))$ for each element z of the (1 x H) vector Z.

Expression (1) describes the deterministic utility of a heterogeneous model, as coefficients are specific to each household. Coefficients of a homogeneous model do not differ across households, i.e. for all i $\alpha_{ijt} = \alpha_j$, $\beta_1^i = \beta_1^j$, $\beta_2^i = \beta_2^j$, $\beta_3^i = \beta_3^j$.

Note that adding an arbitrary constant to utilities leaves the distribution of observed choices unchanged, which implies an identification problem.

MLPs with one layer of hidden units approximate any continuous multivariate function and its derivatives to the desired level of precision given a sufficient number of hidden units with nonpolynomial activation functions (Lehno et al. 1993). MLPs are capable to reproduce interactions, threshold effects or concave relationships of predictors. These capabilities explain why MLP choice models include one layer of hidden units only. Mathematical proofs show that MLPs possess better approximation properties than other flexible models based on polynomial expansions, splines or kernels (Barron 1993; Hornik/Stinchcombe/White 1989).

2.2. Multinomial Logit Model

The MNL model results if errors follow an iid Gumbel distribution (Corstjens/Gautschi 1983; McFadden 1973; McFadden 1980). The MNL model is characterized by a closed form of the conditional choice probability $P_{ij}$ of household i for brand j at occasion t:

$$P_{ijt} = \frac{exp(V_{ijt})}{\sum_{i=1}^t exp(V_{ijt})}$$

(2)

The MNL model is prone to the so-called independence of irrelevant alternatives (IIA) property, which means that the ratio of choice probabilities of any two alternatives is unaffected by enlarging or reducing the choice set by another alternative.

Identification of the MNL model is attained if one of the brand constants is set to zero. We call the flexible multinomial logit model whose deterministic utilities are formed according to expression (1) MLP-MNL model.

2.3. Multinomial Probit Model

The MNP model results if errors follow a multivariate normal distribution $\epsilon_j \sim N(0,\Sigma)$ with $(J - 1 \times J - 1)$ covariance matrix $\Sigma$ (Haussman/Wise 1978; McCulloch/Rossi 1994). Contrary to the MNL model, the MNP model is not subject to the IIA property.

The identification problem mentioned in section 2.1 is solved by setting utility of reference brand J to zero and using utility differences w.r.t. this brand. The distribution of observed choices remains the same if stochastic utilities are multiplied by a positive constant. This identification problem is usually tackled by restricting element (1,1) of covariance matrix $\Sigma$ to one.

We call the flexible MNP whose deterministic utilities are formed according to expression (1) MLP-MNP model.

3. Estimation

Usually homogeneous MNL-MLP models are estimated by maximum likelihood. The loglikelihood function LL which runs over households, purchase occasions and brands can be written as:

$$LL = \sum_i \sum_{t=1}^T \sum_j Y_{ijt} \log P_{ijt}$$

(3)

Binary choice indicators $Y_{ijt}$ equal one for the brand chosen by household i at purchase occasion t and zero for the other brands.

Backpropagation is the most traditional and still most widespread estimation technique for MLP models. Most authors see backpropagation as combination of an effective technique to compute gradients (first order derivatives) of the log likelihood function and gradient descent. In stochastic or online backpropagation, observations are presented in random order and weights are updated after
each observation. This approach reduces the risk of getting trapped in a local optimum. Offline backpropagation on the other hand updates weights based on gradients for all observations of the estimation sample (for an excellent description of backpropagation and related estimation techniques see Bishop 1995).

Slow gradient descent can be replaced by faster nonlinear optimization techniques like scaled conjugate gradients (Möller 1993), BFGS (Saito/Nakano 1997) and Levenberg-Marquardt (Bishop 1995). Experience shows that these techniques often get stuck in local optima. This weakness can be alleviated by multiple random restarts or by hybrid algorithms. The latter use a stochastic method (e.g., stochastic backpropagation or a genetic algorithm) as first step, and fast optimization techniques as second step (e.g., Hruschka 2001).

Markov chain Monte Carlo (MCMC) techniques, which sample from the posterior of coefficients have been used in the neural network and statistics literatures (Ghosh et al. 2004; Lampinen/Vehati 2001; Lee 2000; Müller/Insua 1998; Neal 1996). These efforts were limited to homogeneous models.

Hruschka (2007) introduces the heterogeneous MLP-MNP model together with an appropriate MCMC algorithm for estimation. Coefficient vectors of households are assumed to be drawn from a multivariate normally distributed superpopulation with aggregate coefficients and their covariance matrix. The MCMC algorithm iterates over Gibbs-style draws of utility differences, error covariances, aggregate coefficients, their covariances and a Metropolis-Hastings step which samples household-specific coefficients. Sampling of household-specific coefficients of continuous predictors obeys monotonicity constraints so that, e.g., utility is monotonically decreasing w.r.t. the retail price of a brand.

4. Evaluating Models

Performance measures like log likelihood or hit rate (i.e. the proportion of correctly predicted brand choices) lead to an overly optimistic evaluation of a model if they are determined for the same data which were used for estimation. This problem usually aggravates because of the higher complexity of MLP choice models. Better estimates of performance measures are obtained by randomly dividing the whole data set into two parts, using one part for estimation and the other part, a holdout sample, to compute performance measures. But this procedure has the drawback that error measures are biased upwards (Ripley 1996).

An alternative method, K-fold cross-validation, randomly splits the whole data set into K exhaustive and disjoint subsets. For each subset, weights estimated from data of the remaining K-1 subsets are used to compute a performance measure. The arithmetic mean of performance measures across subsets serves to evaluate a model. The literature recommends 5 ≤ K ≤ 10 (Bishop 1995; Ripley 1996). For small data sets one can use leaving-one-out instead, which consists in predicting the response of each observation by a model which is estimated using the other I-1 observations (Bishop 1995).

Information criteria offer another way to both consider a model’s fit and its complexity, the latter measured by degrees of freedom (df). Given the same fit, a less complex model is preferred. \( AIC \) and \( BIC \) are information criteria often used to evaluate MLPs (Akaike 1974; Amemiya 1980; Schwartz 1979):

\[
AIC = -2LL + 2df, \quad BIC = -2LL + df \log(I)
\] (4)

As a rule, researchers have set the degrees of freedom of a homogeneous MLP equal to the number of its coefficients. Ingrassia/Morlini (2005) show that degrees of freedom correspond to the number of hidden units plus one for a MLP with one hidden layer if a constant term is used (if the MLP has a linear part one has to add the number of its coefficients). This result means that practically all researchers working with homogeneous MLPs may have overestimated their complexity and could explain why cross-validation often recommends a higher number of hidden units than information criteria.

Posterior model probabilities which also penalize models for complexity can be used to evaluate models. Assuming equal a priori model probabilities the posterior probability of a model \( m \) can be computed from marginal model densities \( f(y|m) \):

\[
p(y|m') = f(y|m') / \left( \sum_m f(y|m) \right)
\] (5)

Using MCMC draws marginal model densities may be determined by the harmonic mean estimator of Gelfand/Dey (1994). An alternative, but approximate way to compute posterior model probabilities starts from \( BIC \) values of models \( BIC_m \):

\[
p(y|m') = \exp(-\frac{1}{2}BIC_m) / \sum_m \exp(-\frac{1}{2}BIC_m)
\] (6)

5. Interpretation

As a rule, the effect of a predictor in MLP models cannot be seen by simply looking at one of its coefficients, especially if the model encompasses several hidden units. Partial dependence plots as well as elasticities enable interpretation of effects of predictors. Partial dependence plots show the effect of a predictor on deterministic utilities or utility differences across its value range for one or several fixed constellations of the other predictors (Hastie/Tibshirani/Friedman 2002; Hruschka/Fettes/Probst 2004). Fig. 1 and Fig. 2 contain examples of partial dependence plots.

In marketing, elasticities are very popular to measure the effect of predictors. Elasticities divide the relative change of a dependent variable by the relative change of
the respective predictor. The point elasticity $\varepsilon$ of the $p$th predictor $x_{pih}$ w.r.t. the choice probability $P_{pi}$ is:

$$\varepsilon = \frac{\partial P_{pi}}{\partial x_{pih}} \frac{x_{pih}}{P_{pi}}$$

(7)

For the MLP-MNL model, this point elasticity can be expressed in closed form:

$$\varepsilon = (1 - P_{pi})x_{pih} \left[ \beta_{pih}^{p} + \sum_{h=1}^{H} \beta_{pih}^{h} z_{ih} (1 - z_{ih}) \right]$$

(8)

$\beta_{pih}^{p}$ and $\beta_{pih}^{h}$ denote the household-specific coefficients for predictor $p$ in the linear part and the proper MLP part of the model, respectively. $\beta_{pih}$ symbolizes the household-specific coefficient of hidden unit $h$.

For the MLP-MNP model point elasticities do not have a closed form. As a rule they are computed by using a simulator for choice probabilities (e. g., the GHK simulator, see Train 2003) and numerically approximating the first derivative $\frac{\partial P_{pi}}{\partial x_{pih}}$ in expression (7) by:

$$P_{pi}(x_{pih} + \delta) - P_{pi}(x_{pih} - \delta)$$

$$2\delta$$

(9)

$\delta$ denotes a very small positive real number. $P_{pi}(x_{pih})$ indicates that choice probability is a function of predictor $x_{pih}$.

If the number of hidden units is not too high, one can also interpret MLP choice models by looking at the relationship between different constellations of predictor values and activation values of different hidden units given by $g(X, \beta^h)$ in expression (1). To provide an example, in Hruschka/Fettes/Probst (2004) one hidden unit assumes higher values if a brand is both featured and has a high reference price. In the same study another hidden unit assumes lower values if a brand has a high reference price no matter whether the brand is featured or not.


Tab. 1 describes several empirical studies of MLP choice models w.r.t. predictors, number of observations, number of hidden units, estimation method and the evaluation criterion, according to which the MLP choice model performs better than conventional models.

Bentz/Merunka (1996) introduce the MLP-MNL model into the marketing literature (for an English version of this paper see Bentz/Merunka 2000). In the simulation part of their paper, these authors generate brand choice data from MNL models with an interaction term of two predictors. For these artificial data, a MLP-MNL model with six hidden units outperforms a MNL model with linear utility in terms of log likelihood. In the empirical part they compare the MLP-MNL with four hidden units to a MNL model with linear deterministic utility. Both for the estimation and the 10 % holdout sample, the MLP-MNL model, which implies weak non-linearities (i. e. threshold effects for price reductions and an interaction of loyalty and price reductions), is slightly better.

Fish et al. (2004) estimate MLP-MNL models with four hidden units alternatively by backpropagation and a genetic algorithm. Hit rates for a holdout sample increase by 1 and 2 percentage points compared to the linear utility MNL model, if the MLP-MNL model is estimated by backpropagation and by the genetic algorithm respectively. The MLP-MNL model also predicts brand shares better than the MNL model with linear utility for a 30 % holdout period.


Both the latent class and the neural net approaches clearly outperform the homogeneous linear utility MNL model. The latent class extension of the MNL model leads to higher log likelihood values on estimation data. But for 10-fold cross-validation the MLP-MNL models achieve much better average log likelihood values than their latent class rivals. Elasticities differ significantly between the latent class and the NN-MNL models. Partial dependence plots for deterministic utilities for the

<table>
<thead>
<tr>
<th>Study</th>
<th>Predictors</th>
<th>Observations</th>
<th>Hidden units</th>
<th>Estimation method</th>
<th>MLP better w.r.t.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bentz/Merunka</td>
<td>price (reduction) brand and size loyalties</td>
<td>ca. 5,000</td>
<td>4</td>
<td>BP</td>
<td>LL</td>
</tr>
<tr>
<td>Fish et al.</td>
<td>price (reduction) promotion, lagged promotional</td>
<td>ca. 3,300</td>
<td>4</td>
<td>GA, BP</td>
<td>HR</td>
</tr>
<tr>
<td></td>
<td>purchases, loyalty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hruschka et al.</td>
<td>reference price, feature, display, brand loyalty</td>
<td>&gt;11,000</td>
<td>3, 7, 10</td>
<td>SBP + BFGS</td>
<td>LL</td>
</tr>
<tr>
<td>Hruschka/Fettes/Probst (2004)</td>
<td>reference price, feature, display, brand loyalty</td>
<td>&gt;11,000</td>
<td>4 or 5</td>
<td>SBP + BFGS</td>
<td>BIC</td>
</tr>
<tr>
<td>Hruschka</td>
<td>price, price deviation, feature, display</td>
<td>&gt;11,000</td>
<td>1 or 3</td>
<td>MCMC</td>
<td>MMD</td>
</tr>
</tbody>
</table>

BP backpropagation, SBP stochastic backpropagation, GA genetic algorithm
LL log likelihood, HR hit rate, MMD marginal model density

Table 1: Empirical Studies of MLP Choice Models

54 MARKETING·JRM·1/2010
MLP indicate interaction effects for loyalty and sales promotion, threshold effects and saturation effects for loyalty and an inverse S-shape for reference price.

In another study Hruschka/Fettes/Probst (2004) compare the MLP-MNL model to the generalized additive MNL model of Abe (1999). Differences of BIC values show that approximate posterior probabilities of MLP-MNL models approach the maximum value of 1.0. The MLP-MNL model also performs better in terms of average log likelihood for 10-fold cross-validation.

The paper of Hruschka (2007) seems to be the first developing and applying a heterogeneous MLP choice model. All coefficients of the MLP-MNP model are household specific, whereas related studies with other flexible modelling approaches either restrict heterogeneity to brand constants (Briesch/Chintagunta/Matzkin 2007) or to one continuous predictor (Kim/Menzefricke/Feinberg 2007). The heterogeneous MLP-MNP models outperform the heterogeneous linear utility MNP models in terms of log marginal densities. Marginal model densities also serve to choose the number of hidden units. For most MLP-MNP models estimated in Hruschka (2007), the number of hidden units equals one, although as a rule these models improve log marginal densities drastically compared to the MNP model with linear utility.

Figure 1: Utility Difference vs. Price

Ketchup

Peanut Butter
The predictors in *Hruschka* (2007) consist of price, price deviation (the difference between price and reference price), feature and display. Reference prices are internal prices to which households compare observed prices (*Winer* 1988). The reference price model applied follows the concept of extrapolative expectations (*Winer* 1986) and defines reference price as one period forecast by a linear model with a brand-specific constant and prices lagged by three periods together with a time index as predictors. For the best performing model reference prices are computed by a heterogeneous reference price model whose coefficients vary across households.

Partial dependence plots of mean utility differences for prices and price deviations show nonlinear effects (i.e. changing marginal effects) for prices in both product categories. For ketchup, utility difference follows an inverse S-shape w.r.t. price, i.e. marginal effects for both higher and lower prices are weaker than those for medium prices (see *Fig. 1*). For peanut butter, marginal effects increase with price and two sections with lower or higher marginal effects may be distinguished. Utility difference decreases with price deviation for both product categories (see *Fig. 2*).

For ketchup, partial dependence plots of utility difference vs. price deviation show three sections for low,
medium and high price deviations. At medium and high prices marginal effects are strongest if price deviations are medium, at low prices marginal effects increase with price deviations. Compared to ketchup marginal effects of price deviations do not change much for peanut butter. Utility difference decreases with price for both product categories.

For ketchup, brands absolute price elasticities (4.56) according to the MLP-MNP model are lower than those inferred from the MNP model with linear utility (7.46) because the latter overestimates marginal price effects both in the high and the low price ranges. For peanut butter, absolute price elasticities (6.02) are higher than those inferred from the MNP model with linear utility (4.65) because the latter underestimates marginal price effects in the high price range.

7. Conclusions

MLPs are sometimes criticized for their black box character (e. g., Rangaswamy 1993). Tools like partial dependence plots or interpretation of hidden units which are described in section 5 offer a way to obtain insight about their internal working. Without any doubt, estimation of MLP choice models causes more effort in comparison to conventional choice models. Estimation of heterogeneous models requires some knowledge of an appropriate high-level programming language (e. g., GAUSS, Matlab, or R). On the other hand, homogeneous MLP choice models can be estimated more easily by using nonlinear optimization packages which are offered as add-ons of these widespread software tools. Because of the danger of local optima random restarts are strongly recommended. More research effort should be devoted to Bayesian estimation methods (e. g., MCMC methods) which already have turned out to be successful for more conventional models.

Studies in the marketing domain demonstrate as a rule that MLP choice models perform at least quite as good as more conventional statistical methods. For five of six relevant studies, shown in Tab. 1, MLP choice models lead to much better results in terms of information criteria or marginal model densities for real brand choice data. This fact indicates that deterministic utilities are frequently subject to important nonlinear effects of predictors.

The studies discussed in section 6 as a rule give evidence to changing marginal effects of prices, reference prices, or price reductions. Homogeneous MLP choice models indicate positive and decreasing marginal effects of loyalty on utility and interaction effects between prices and loyalty.

References


Hruschka, Neural Net Extensions of Stochastic Brand Choice Models


