Functional Flexibility, Latent Heterogeneity and Endogeneity in Aggregate Market Response Models

By Harald Hruschka

We focus on flexibility, latent heterogeneity and endogeneity in aggregate market response models which previous reviews have considered either incompletely or not at all. Ignoring these issues could lead to biased estimates of the effects of marketing variables and finally erroneous implications for marketing decision making. We recall the main characteristics of several more frequently applied parametric market response functions. In the next chapters we review relevant studies indicating both methods applied and results obtained. We start by presenting flexible aggregate market response models. Their nonparametric component is often specified as multilayer perceptron, spline regression, or kernel regression. We then deal with latent heterogeneity both across households and across retail chains or stores. In the next section we explain endogeneity in aggregate market response models focusing on instrumental variables estimation techniques. We finish by summarizing the implications of this overview and offering an outlook on open research problems.

Harald Hruschka is Professor of Marketing at the University of Regensburg, Universitätsstrasse 31, D-93053 Regensburg, Germany, Phone: +49/941 943 2279, Fax: +49/941 943 2828, E-Mail: harald.hruschka@wiwi.uni-regensburg.de.

Please note: The author would like to thank two anonymous referees for their helpful and constructive comments.

1. Introduction

Variables of aggregate market response models are defined as sums (e. g., of sales, of advertising budgets) or averages (e. g., of prices) across persons or households for a certain store, retail chain, region, etc. In this overview we consider as dependent variables sales or market shares of brands in one product category. We do not deal with models for sales at the category level or with models which include effects of marketing variables on other categories. Moreover, we focus on static response and ignore dynamic effects such as advertising goodwill. Independent variables of market response models typically consist of marketing variables for the brand whose sales or market share constitute the dependent variable as well as marketing variables for competing brands in the same category. To these independent variables other market or environmental variables may be added (Hruschka 1996; Hanssens et al. 2001).

We focus on flexibility, latent heterogeneity and endogeneity in aggregate market response models as previous reviews (e. g., Hanssens et al. 2001, Albers 2012) dealt with these issues incompletely or even not at all. We also think that consideration of these issues will gain in importance in the near future. Estimation of aggregate response models ignoring these issues could lead to biased estimates of the effects of marketing variables. Obviously, biased estimates entail erroneous implications for marketing decision making (e. g., the recommendation of prices which are too high if price effects are underestimated).

We start this review by discussing parametric aggregate response models and their properties. In section 3 we present the most frequently used flexible aggregate market response models. These models specify a nonparametric component by either a multilayer perceptron, a spline regression, or a kernel regression. Section 4 deals with continuous and finite mixture approaches by which aggregate response models were extended to take latent heterogeneity into account. In this regard we distinguish heterogeneity within regions or stores on one hand, and heterogeneity across retail chains or stores on the other hand. In section 5 we explain endogeneity in aggregate market response model. We focus on instrumental variables estimation techniques. We discuss instrumental variables that have been used in aggregate response mod-

els as well as instrument-free approaches. In each section we briefly present relevant studies and indicate both methods applied and results obtained. In the final section 6 we summarize the implications of this review and offer an outlook on open research problems or opportunities.

2. Parametric Market Response Models

In the following we discuss sensible characteristics of aggregate market response functions and refer to several appropriate parametric functions (Lilien at al. 1992; Hruschka 1996; Hanssens et al. 2001). *Tab. 1* contains the mathematical form and shape of these functions. *Fig. 1* illustrates shapes which these functions can assume.

Linear market response functions imply constant returns to scale, i. e., sales increases or decreases are the same for each additional input of a marketing variable, no matter how large or small the input already is. Linear functions are appropriate if the observed value range of marketing instruments is limited or interest is restricted to directional recommendations.

Nonlinear market response functions allow for varying returns to scale. For instruments like advertising, shelf space, sales effort, etc. concave functions are empirically supported. In a concave function each additional input leads to less additional sales. Semi-log, multiplicative, ADBUDG, quadratic, and modified exponential functions are able to generate a concave shape.

For appropriate parameter values the extended multiplicative function of Hruschka (1991b) can also lead to a concave shape, but values and increases of the dependent variable are lower at higher values of the marketing instrument compared to the basic multiplicative function. In the extended function the exponent is not constant, but a linear function of the log of the independent variable. Albers (2012) demonstrates that this function reproduces the concave shape of running means of sales with respect to sales calls as independent variable (please see section 3 for more details).

As a rule more incremental sales result at higher price decreases. In other words, sales response with respect to price has a convex shape which can be reproduced by multiplicative or exponential functions. The extended multiplicative function can also be used. In this case values of the dependent variable are higher and their decreases lower at higher values of the marketing instrument (e. g., price) compared to the basic multiplicative function.

S-shaped functions are composed of two parts, i. e., a convex followed by a concave part. At low values of a marketing variable returns to scale increase, at high values returns to scale decrease. S-shaped aggregate functions can be derived by theoretical arguments, but empirical support is limited except for shelf space. Log-reciprocal, ADBUDG, and logistic functions can assume a S-shape.

A lower threshold is at work, if marketing input below a certain value has no effect on sales or market share.

Name	Function	Shape
linear	$y_{jn} \mid a 2 b v_{jn}$	
semi-log	$y_{jn} \mid a 2 b \log v_{jn}$	concave
multiplicative	$y_{jn} \mid a v_{jn}^{b}$	concave for $0 < b < 1$ convex for $b < 0$
extended multiplicative	$y_{jn} \mid a v_{jn}^{b 2 c \log v_{jn}}$	concave for $0 \{ b 2 c \log v_{jn} < 1$ convex for $b 2 c \log v_{jn} < 0$
exponential	$y_{jn} \mid \exp(a 4 b v_{jn})$	convex
ADBUDG	$y_{jn} \mid a 2 / y_{max} 4 a 0 v_{jn}^{b} /(c 2 v_{jn}^{b})$	concave for $0 < b < 1$ S-shaped for $b > 1$
quadratic	$y_{jn} \mid a 2 b v_{jn} 4 c v_{jn}^{2}$	concave for $c < b$
log-reciprocal	$y_{jn} \mid \exp(a 2 b /(1 2 v_{jn}))$	S-shaped
logistic	$y_{jn} \mid y_{max} /(1 2 \exp(4(a 2 b v_{jn})))$	S-shaped
modified exponential	$y_{jn} \mid y_{max}(1 4 \exp(4 b v_{jn}))$	concave
attraction models	$y_{jn} \mid A_{jn} / \underset{j}{\text{——}} A_{j}$	
MNL	$A_{jn} \mid \exp(a_j 2 b_j v_{jn})$	S-shaped
MCI	$A_{jn} \mid \exp(a_j 2 b_j \log v_{jn})$	S-shaped

Notes: j brand index, *n* observation index, y_{jn} sales or market share, v_{jn} marketing variable, y_{max} saturation level.

Tab. 1: Parametric aggregate response functions for one marketing variable

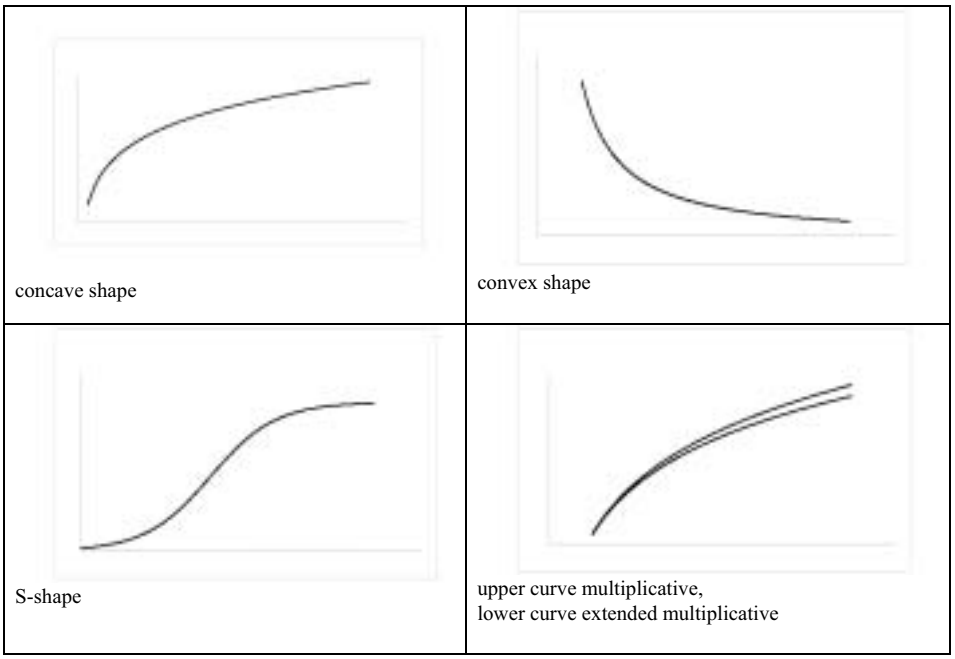

Fig. 1 Shape of parametric aggregate response functions

There is only limited empirical evidence of lower thresholds for advertising. We have to remark that lower thresholds are hard to distinguish from S-shaped functions. At saturation (upper threshold) sales or market share do not increase if the marketing input is greater than a certain value. Saturation can be reproduced by modified exponential, ADBUDG, and logistic functions.

Two marketing variables interact if the effect of one variable depends on the value of the other variable. Technically speaking, second order derivatives of the response function with respect to both variables are different from zero. Of course, this property does not apply to linear functions without interaction terms. Nonlinear functions are more flexible in this regard, but may imply less obvious restrictions which are necessary to obtain economically plausible effects. In aggregate sales response functions with two marketing variables price and advertising only positive returns to scale are plausible for advertising. This requirement implies for a multiplicative sales response model that second order derivatives of advertising and price must be positive (Hruschka 1991b). This restriction can be eliminated by specifying the price coefficient as function of advertising. Such a specification allows negative interactions between price and advertising, though returns to scale of advertising remain positive.

The parametric functions mentioned so far can also be applied to investigate market shares as dependent variables. But market shares predicted on the basis of these models might not lie between zero and one. In addition, the sum of predicted market shares across all brands will not equal one. Attraction models avoid these inconsistencies by defining market share as ratio of a brand's attraction in the numerator and the sum of the attractions of all brands in the denominator (Cooper and Nakanishi 1988; Carpenter et al. 1988). Attractions are computed as exponential function of linear combinations of predictors. One can distinguish two basic variant of attraction models. The multinomial (MNL) attraction model is based on raw values of marketing variables. The multiplicative competitive interactions (MCI) model considers logs of marketing variables instead (Hanssens et al. 2001). This can be seen by taking the log of the multiplicative term $v_{jn}^{\beta_j}$ which gives $\beta_j \, log(v_{jn})$. The MNL attraction model allows decreasing returns to scale only if a brands market share is greater than 50 %. This condition represents a weakness of the MNL model, as it will be violated as a rule if a market consists of more than two brands.

3. Flexible Functions

In the previous section we dealt with aggregate response functions with a fixed parametric form. Hanssens et al. (2001) began to consider flexible approaches in the second edition of their book on aggregate market response modeling. Shape restrictions imposed by fixed parametric functions could be responsible for results which seem to indicate that aggregate response functions have no lower threshold or are not S-shaped. Using more flexible approaches allows greater opportunity to discover such effects.

Exploratory univariate smoothing techniques like running means, running medians or running line (Fahrmeier et al. 2007) may give first insights into the dependence of sales or market share on varying values of a marketing variable. A running mean equals the arithmetic mean of sales (market share) for a symmetrically defined neighborhood around any observed value of a predictor. Higher neighborhood sizes lead to more smoothing. Albers (2012) analyzes data from a pharmaceutical company by

running means with a neighborhood size of 250. The running means obtained in this manner indicate a concave relationship between sales and sales calls (both measured as averages per physician) in 5,007 sales territories.

Fully fledged flexible market response functions are as a rule semiparametric, i. e., they consist of a parametric (e. g., with one coefficient for each metric predictor or for dummy variables such as features, displays, seasons, etc.) and a nonparametric component $G(x_n)$. Then sales or market share can be written as:

$$y_{jn} = \beta' x_n + G(x_n) \quad (1)$$

x_n denotes a vector of predictors (regressors) which as a rule includes marketing variables of the same and of competing brands.

The inclusion of a parametric component in (1) ensures that the nonparametric component deals only with departures from the former. This specification usually leads to a less complex nonparametric component and faster convergence of estimates.

We find three types of nonparametric components in the literature on aggregate market response models, multilayer perceptrons (MLP), spline regressions, and kernel regressions (see *Tab. 2*). Both MLPs and spline regression specify the dependent variable as linear combination of basis functions, but differ with respect to the type of basis functions used. Basis functions of a MLP are S-shaped with respect to different linear combinations of predictors. In spline regression piecewise polynomials of predictors, which are joined together to form a curve, serve as basis functions.

Kernel regression is not based on a linear combination of basis functions in contrast to the MLP and spline regression. Kernel regression estimates values of the dependent variable as local weighted average of the dependent variable across all observations. The weight of an observation is proportional to its similarity to the respective vector of predictors. Similarity is measured by a kernel function. We give more details on the MLP, spline regression and kernel regression in the next three subsections.

Instead of MLPs, splines or kernels researchers might consider to use a polynomial function of high order, which according to the famous Stone-Weierstrass theorem is capable to approximate any continuous function. High order polynomials lead to excellent approximations driving residuals to zero. On the other hand, high order polynomials suffer from Runge's phenomenon, i. e., they produce interpolations which are erratic and far from the true function. In addition, this problem becomes more pronounced, the more observations are available (Hansen 2017). We emphasize that MLPs, spline regression and kernel regression are not subject to Runge's phenomenon.

Many marketing academics are hesitant to apply flexible models as they see the dangers of overfitting and also of economically implausible estimation results. Overfitting problems arise if model complexity is not carefully selected. Models which are too complex reproduce minor variations in the data which could simply be due to noise. Very wiggly curves or surfaces indicate overfitting. By averaging over larger neighborhoods (depending on the method applied by using less hidden units, less knots or a larger smoothing window) model complexity is lowered. Often information criteria or cross-validation serve to select model complexity as the papers discussed in the next three subsections show.

Even after careful selection of model complexity flexible techniques may provide economically implausible estimation results (e. g., non-monotone price response curves, convex or monotonically decreasing advertising response curves). These problems can be avoided by restricting the shape of function in an appropriate way, often related to first and second order derivatives.

3.1. Multilayer Perceptron

The MLP is the most popular type of artificial neural net in many application fields. In aggregate response modeling the MLP with one layer of H hidden units each with a S-shaped logistic activation function $f(c_h' x_n) = 1/(1 + \exp(-c_h' x_n))$ dominates (see Hruschka 1991a for an introduction to the use of artificial neural nets in marketing research in general). In relevant studies MLPs have been estimated by (stochastic) gradient descent, sometimes combined with faster nonlinear optimization methods like BFGS (Hruschka 2008).

Given a sufficient number of hidden units with S-shaped activation functions a MLP approximates any continuous

Multilayer perceptron	$G(x_n) = \sum_{h=1}^{H} b_h f(c_h' x_n)$
Spline regression	$G(x_n) = \sum_{p=1}^{P} g_p(x_{pn})$
Kernel regression	$G(x_n) = \sum_{m=1}^{N} K\left(\frac{x_m - x_n}{w}\right) \tilde{y}_{jm} \Big/ \sum_{m=1}^{N} K\left(\frac{x_m - x_n}{w}\right)$

Notes: H number of hidden units, $f()$ activation function, p predictor index, g_p general univariate function, $K()$ kernel, w bandwidth parameter, $\tilde{y}_{jm}$ nonparametric part of the dependent variable y_{jm}.

Tab. 2: Nonparametric component of semiparametric aggregate response models

Publication	Estimation Sample	Holdout Sample
	Sales as dependent variable	
Hruschka (1993)	15 % MSE	
Ainscough and Aronson (1999)	59 % MSE	
Pantelidaki and Bunn (2005)		lowest MAPE
	Market share as dependent variable	
Van Wezel and Baets (1995)	46 % RMSE	79 % RMSE
Wierenga and Kluytmans (1996)		68 % RMSE
Natter and Hruschka (1997)	small improvement of information criterion	
Gruca et al. (1998)	better MAPE in one of two categories	better MAPE in one of two categories
Hruschka (2001)	72 % MSE	

Notes: MSE mean squared error, RMSE root mean squared error, MAPE mean absolute percentage error
Reading help: 72 % MSE means that the MSE of the MLP amounts to 72 % of the parametric model

Tab. 3: Performance of multilayer perceptrons relative to parametric models

multivariate function and its derivatives with desired accuracy (Cybenko 1989; Hornik et al. 1989). The MLP is capable to discover interactions, thresholds and concave relationships of independent variables. The estimated function becomes less smooth the more hidden units are included.

In most studies which use MLPs to model aggregate market response we find prices and advertising (measured by expenditures or features) as predictors. One group of studies look at sales as dependent variables. As a rule MLPs improve performance in the estimation sample and also attain better predictions in holdout samples (see *Tab. 3*).

Most studies with market share as dependent variable specify a separate MLP for each brand. Only Hruschka (2001) estimates market shares of all brands of a product category by one MLP with brand-specific hidden units. Contrary to the conventional attraction model this model reproduces threshold effects of prices, i. e., market share changes are higher if prices are below or above a certain level. For one brand the MLP shows a weaker price effect except at very low prices. For another brand the MLP indicates higher effects at very high prices and lower effects over the remaining price range. Price response curves are economically plausible and remain monotonically decreasing over the range of observed prices due to a very low number of hidden units. The MLP in Hruschka (2001) implies higher optimal prices for two brands compared to those based on a conventional attraction model. These higher prices increase profits by more than 10 %.

3.2. Spline Regression

Splines are piecewise polynomials which are joined together to form a curve. Knots, i. e. values of a predictor, define these pieces or intervals. Smoothness is controlled either by the number of knots or by a smoothing parameter. Decreasing the number of knots or increasing the smoothing parameter makes the function smoother. Fahrmeier et al. (2007) give an excellent detailed introduction into spline regression and various estimation methods, e. g., backfitting or Markov chain Monte Carlo (MCMC) simulation.

Linear splines often lead to sharp kinks. Another disadvantage of linear splines is the fact that higher derivatives are not defined. That is why shape restrictions, which might be necessary to obtain economically plausible results, are not considered. The algorithm MARS (multivariate adaptive regression-splines) starts from linear splines at each observed input value and their interactions as basis functions which are backward- forward selected (Friedman 1991).

In contrast to linear splines first and second derivatives exist for cubic splines. Using the truncated power basis a cubic spline with knots $\kappa_1 \cdots, \kappa_L$ can be written as linear combination in the following way:

$$g_p(x_{pn}) = c_1 x_{pn}^2 + c_2 x_{pn}^3 + \sum_{l=1}^{L} c_{l+2} (x_{pn} - \kappa_l)_+^3 \quad (2)$$

The truncated cubic power function is defined as:

$$(x_{pm} - \kappa_l)_+^3 = \begin{cases} (x_{pn} - \kappa_l)^3 & if\ x_{pn} - \kappa_l > 0 \\ 0 & else \end{cases} \quad (3)$$

B-splines which provide exactly the same fit are numerically more stable than splines using the truncated power basis (de Boor 2001). Penalized or P-splines are B-splines at equally spaced knots with a roughness penalty. In Bayesian estimation approaches this penalty can be considered by means of first-order or second-order random walks of regression coefficients of adjacent splines. In Bayesian approaches the smoothness parameter can be estimated and is no longer required to be set in advance. It also possible to impose monotonic restrictions on effects (e. g., to prevent that sales of a brand increase at higher own prices or lower competitive prices if other effects are held constant).

Cubic smoothing splines place knots at all observations and shrink coefficients of the estimated function by regularization. Knot selection is replaced by setting a smooth-

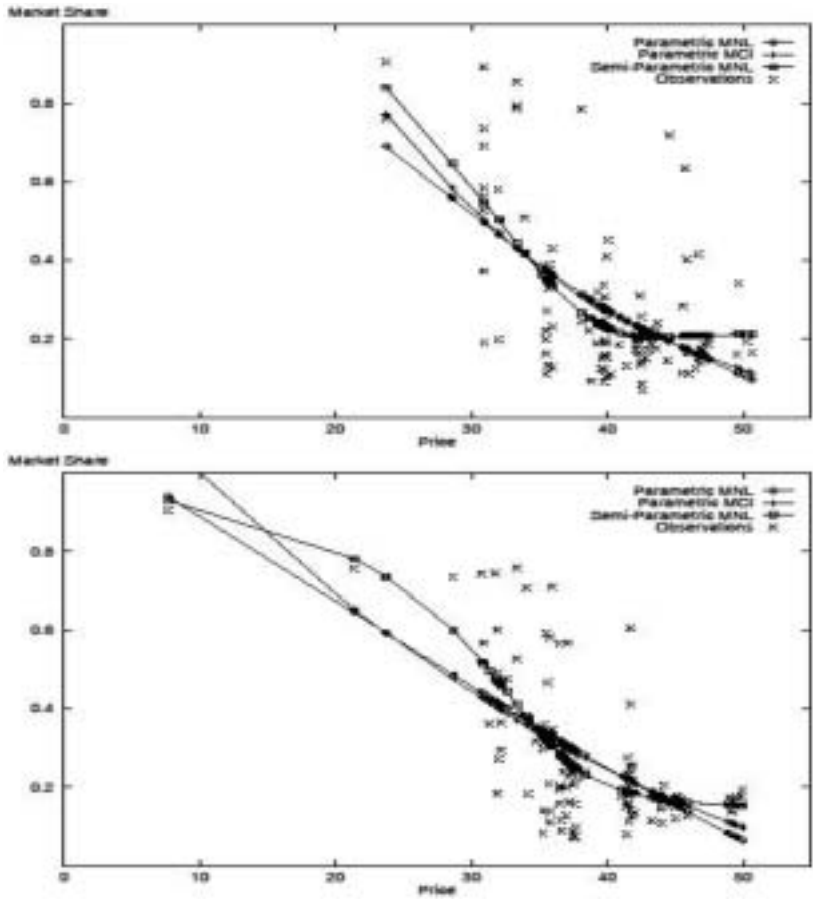

Fig. 2: Observed and estimated market shares of two brands

ing parameter or degrees of freedom. In addition cubic smoothing splines prohibit erratic behavior by forcing splines to be linear outside the boundaries enclosed by the smallest and highest knot (Fahrmeier et al. 2007).

Most aggregate response models with higher order splines follow the general additive modeling (GAM) framework which defines the nonparametric component as sum of flexible univariate functions $\sum_p g_p(x_{pn})$ (see *Tab. 2*) and usually ignore interactions (Hastie and Tibshirani 1986; 1990).

In Albers (2012) MARS indicates increasing returns to scale for higher values of sales calls values above a threshold which is not in agreement with theoretical and empirical knowledge. Kolsarici and Vakratsas (2011) analyze data for leading brands of, e. g., two motor vehicle categories. They apply MARS to determine sales response functions with advertising expenditures in different media as predictors. Main effects of advertising expenditures show sharp kinks. Several main effects even suggest that sales decrease at all values of advertising expenditures which is implausible as it implies that advertising should be avoided completely.

Kalyanam and Shively (1998) analyze sales and price data acquired in one store for five and four brands in two product categories. These authors estimate response models with cubic stochastic splines for own and competitive prices. Stochastic splines are characterized by slopes which follow a Wiener process. Based on results for one disaggregate price response experiment Kalyanam and Shively (1998) argue for irregular price effects. Even if one admits that such irregularities might occur, one may doubt that they remain at the aggregate level. The model with cubic stochastic splines leads to better adjusted R^2 values than a parametric exponential price response function. To our opinion a comparison to a regularized flexible model (e. g., with cubic smoothing splines) would be more convincing. For one brand sales even increase if its prices are higher than a certain threshold value. This result clearly contradicts economic theory.

Hruschka (2002) analyzes market shares of four brands. He adds cubic smoothing splines of own and competitive prices to the logarithmic attraction values of each brand for both a MCI and a MNL attraction model. To avoid overfitting problems, Hruschka (2002) evaluates models based on MSE averaged across 200 bootstrapped samples. *Fig. 2* shows observed market shares and their estimates obtained by parametric and semiparametric attractions models with splines. The overall best model is the semiparametric MNL attraction model with three degrees of freedom for each spline. It performs better than the two parametric models and also beats the semiparametric MCI model. In the semiparametric MNL attraction model all price effects are monotone. Marginal effects and elasticities differ from those implied by the two

parametric attraction models. Marginal price effects on market share are much smaller at higher prices. Based on the semiparametric MNL model one obtains lower optimal prices for three brands due to higher elasticities in the relevant range. For two brands these lower prices lead to profit increases of about 10 %.

Steiner et al. (2007) investigate sales of eight orange juice brands in 81 stores. Cubic P-splines serve to measure own and cross price effects for these brands. Steiner et al. compare to several parametric forms (e. g., exponential and multiplicative). They evaluate models by their predictive validity which is measured by nine-fold cross validated squared prediction errors. Among parametric models the multiplicative model has the highest predictive validity, but except for one brand it is outperformed by the semiparametric model. In contrast to the multiplicative model sales are more affected by price increases in the low price range according to the semiparametric model. Sales decreases due to higher prices become very small after an upper threshold value is exceeded.

Brezger and Steiner (2008) analyze the same data as Steiner et al. (2007). The models in this paper include three different cross price effects (with respect to all premium brands, all national brands and one store brand, respectively). The authors focus on comparing semiparametric models with and without monotonic restrictions. Models with monotonic restrictions attain a higher predictive validity. For one of the brands own and cross price effects turn out as follows. Own price effects show a reverse S-shape with higher marginal effects at medium prices. For very low prices an additional sales increase occurs. Cross price effects triggered by premium brands occur only if one of the premium brands is priced lower than a threshold. The cross price effect of national brands is S-shaped, but much smaller than the corresponding effect of premium brands. The store brand's cross price effect on the other hand is very small.

3.3. Kernel Regression

Kernel regression determines the estimated value of the dependent variable for each observed vector x_n of independent variables (e. g., of prices and advertising expenditures) as local weighted average of the dependent variable. The weight of any observation x_m is proportional to its similarity to the vector x_n. This similarity is measured by a kernel function $K(x_m - x_n)/w$. Higher values of window parameter w produce smoother functions. Kernel functions are probability density functions which are bounded and symmetric about zero (Härdle 1990).

Both studies dealing with aggregate response modeling which we discuss in the following use the Nadaraya-Watson estimator shown in *Tab. 2* with a Gaussian product kernel. The Gaussian product kernel corresponds to a multivariate standard normal distribution with independent components (Härdle 1990) and can be written as:

$$K\left(\frac{x_m - x_n}{w}\right) = (2\pi)^{0.5p} \prod_{p=1}^{P} \exp\left(-0.5\left(\frac{x_{mp} - x_{np}}{w}\right)^2\right) \quad (4)$$

Van Heerde et al. (2001) analyze sales of the main brands in each of three different product categories (tuna, beverages, food). The nonparametric component is related to logs of price indices for each brand. A price index is defined as actual price divided by regular price. A price index less than one therefore indicates a temporary price discount. The semiparametric model with kernels reduces mean squared errors in a split-half validation sample by 46.7 %, 8.9 % and 12.2 % over a parametric multiplicative model in the three product categories.

Van Heerde et al. (2001) also estimate a GAM which in contrast to the semiparametric model with kernels does not consider interactions of price indices. The paper is very parsimonious in providing information, e. g., it does not mention which type of univariate smoother is used for the univariate nonlinear functions. In the tuna category the GAM fits worse than the parametric multiplicative model for the estimation data. Presumably, the GAM does not include a parametric component identical to the multiplicative model, although this would be feasible. Otherwise the fit of the GAM should be at least as good as the fit of the multiplicative model.

Own price index effects of the semiparametric model with kernels estimated by van Heerde et al. (2001) have an (incomplete) S-shape for most brands. But several implausible non-monotonic sections occur in which sales increase at lower discounts, i. e., at higher prices. For one brand even an inverse u-shape is shown, i. e., medium price discounts are associated with higher sales than smaller or bigger price discounts. Cross price index effects are as a rule S-shaped.

In the paper already mentioned in section 2.2 Kolsarici and Vakratsas (2011) compare MARS to a completely nonparametric model with Gaussian kernels. The fact that the latter has no parametric component could explain why it is outperformed by MARS which attains lower mean squared errors for a split-half validation sample.

4. Latent Heterogeneity

We focus on latent or unobserved heterogeneity and ignore observed heterogeneity because effects of socio-economic variables aggregated for regions, catchment areas of stores etc. are usually low (Montgomery 1997; Lang et al. 2015). As a rule latent heterogeneity is taken into account by either finite mixtures (FM) or continuous mixtures (CM) of coefficients. The most popular continuous mixture is based on a multivariate normal prior distribution.

In aggregate response modeling researchers have considered two types of latent heterogeneity:

1. heterogeneity within regions or within stores across consumers
2. heterogeneity across retail chains or stores

4.1. Heterogeneity within Regions or Stores

It may surprise some readers that latent heterogeneity within regions or stores can be taken into account in response models whose variables are aggregated across consumers or households. These aggregate response models start from a choice model at the individual level. Aggregation is obtained by assuming that individual-level coefficients follow a continuous or, less frequently, a finite distribution. From now on we call these models heterogeneous choice models without reiterating that they are based on aggregate data.

Choice models, e. g., the multinomial logit model or the multinomial probit model, are nonlinear with respect to parameters. Explicit aggregation avoids the bias caused by the well-known fact that the average of a nonlinear function differs from the value of a nonlinear function at average values. The goal of parameter estimation in heterogeneous choice models consists in reproducing observed aggregate brand shares as good as possible. The typical approach for heterogeneous choice models starts by specifying mean utility across consumers as linear function. A generalized methods of moments (GMM) estimator can be used which as a rule includes simulation to compute market shares which are inverted by contraction mapping (for an excellent introduction to heterogeneous choice models and their estimation see Nevo 2000).

Just like their individual level relatives heterogeneous choice models are based on the assumption that consumers make at most one purchase at a store visit. The choice set usually consists of several purchase alternatives (brands) and one no-purchase option. In addition to observed characteristics (e. g., price, advertising, product attributes, brand constants) one unobserved characteristic which varies across brands and periods is considered as well. Such an unobserved characteristic may be due to missing variables or demand shocks.

Heterogeneous choice models need less parameters than many traditional sales response models as the latter often include cross effects for each pair of brands. Heterogeneous choice models are similar to attraction models of the MNL type without cross effects. But heterogeneous choice models differ from attraction models in several aspects, namely explicit aggregation, latent heterogeneity and unobserved characteristics. Except for brand constants coefficients of heterogeneous choice model are not brand-specific.

Heterogeneous choice models usually specify the indirect utility of a consumer *i* for a brand *j* at purchase occasion *t* as follows:

$$u_{ijt} = -\alpha_i p_{jt} + \beta_i' x_{jt} + \xi_{jt} + \varepsilon_{ijt} \quad (5)$$

p_{jt} is the observed price. x_{jt} denotes a vector of observed product characteristics which may include non-price marketing variables. ξ_{jt} symbolizes an unobserved characteristic which can be computed as residual of a regression of the part of the indirect utility which is constant across consumers on brand constants and other predictors. ε_{ijt} is an error term with zero expectation. For Gumbel distributed (multivariate normal distributed) error terms one obtains the heterogeneous logit (probit) model.

Most relevant studies apply the heterogeneous logit model which in contrast to the homogeneous logit model allows that cross-elasticities reflect different similarities of brand pairs. Of course, the heterogeneous probit model is capable to reproduce different similarities in a more direct manner. But for researchers who want to investigate at least a moderately high number of brands the heterogeneous logit model seems to be more viable due to its closed form expression for choice probabilities (e. g., Chintagunta 2001 estimates a heterogeneous probit model using aggregate data for three brands only).

Nevo (2001) analyzes data of 25 brands of one category in 65 cities by a CM logit model. He includes product characteristics, price, advertising expenditure, and brand constants as predictors. Nevo compares to a homogeneous logit model which corresponds to an attraction model without brand-specific coefficients except for brand constants. The CM logit model attains a better fit. Whereas its average price coefficients are similar to those of the homogeneous logit model, own price elasticities and especially cross elasticities are very different.

Besanko et al. (2003) investigate data of four brands acquired in nine stores of one retail chain. They consider price, feature, display, brand constants, and store constants as predictors. Besanko et al. (2003) follow a FM approach which suggests three segments. In agreement with expectations on the consequences of ignoring latent heterogeneity, they obtain lower price response parameters for the homogeneous logit model. To demonstrate a managerial implication Besanko et al. (2003) compare uniform pricing to prices differentiating between segments. Although consumers can only be imperfectly assigned to segments as only data of the current purchase are available, these authors show that price differentiation increases the retailer's profit by 11 %.

Chintagunta et al. (2003) estimate a CM logit model for nine and seven brands in the liquid detergents and refrigerated orange juice categories, respectively, using data from 83 stores. To save parameters these authors reduce the covariance matrix of brand constants to a two factor solution. Independent variables also include promotion incidence and package size. Chintagunta et al. (2003) also determine how much profits increase over uniform pricing at the chain level by alternative pricing policies. Store level pricing leads to profit increases of 9.6 % and 16.3 % for the two categories. Constrained store level pricing which prevents a reduction of customers' welfare leads to increases of 5.6 % and 7.4 %. Finally, pricing at the level of store clusters which are determined by a cluster analysis of store level prices increases profits by 3.8 % and 8.6 %.

In a simulation study Andrews et al. (2011) compare a homogeneous nested logit model to several related models which take latent heterogeneity into account by finite as well as continuous mixtures. These nested logit models have two branches, one for non-purchases and the other one for brand purchases. The homogeneous nested logit performs rather well with respect to fit and prediction. Only a nested logit model with continuous within-store heterogeneity based on store level data turns out to be better.

4.2. Heterogeneity across Retail Chains or Stores

We now deal with heterogeneity across retail chains, stores, or both. We can distinguish two extreme specifications to analyze such data. One extreme specifies one model for each retail chain or store. The other extreme consists of one homogeneous (pooled) model completely ignoring that responses might differ between chains, stores etc. The first alternative often provides implausible estimates (e. g., positive own or negative cross price effects). The second alternative leads to biased estimates. From a managerial point of view micro-marketing policies, e. g., setting prices which vary between stores, require heterogeneous estimates and cannot be derived from a homogeneous model. A compromise between the two extremes which produces estimates which are both plausible and less biased can be achieved by introducing mixture distributions on the coefficients.

CM approaches for aggregate response models which deal with heterogeneity across chains or stores are usually based on a multivariate normal prior which asserts that coefficients across stores are probably similar (Lancaster 2004). Placing this restriction on coefficients results in soft cross-dependencies across stores (chains). Coefficients differ from those obtained by separate regression models. They are closer to estimates of a homogeneous model (Lindley and Smith 1972) and less noisy compared to the estimates obtained by separate store- specific regression models (Hanssens et al. 2001). In addition (soft) sign restrictions may be set which guarantee plausible estimates with certainty (with high probability). Usually these models are estimated by MCMC simulation (for more details see, e. g., Greene 2003 or Lancaster 2004).

We start to overview studies which take latent heterogeneity into account and use parametric response functions. Blattberg and George (1991) specify an exponential sales response function. They consider price index (regular price divided by average regular competitive price) and deal discount as predictors amongst others. Estimating one individual sales response function for each of twelve chain-brand combinations by OLS provides implausible positive, but also extremely large (in absolute terms) negative coefficients for the price index. Moreover, coefficients for deal discounts vary strongly across combinations. CM models lead to more reasonable estimates and improve predictive performance (MSE in a holdout sample) compared to OLS models. In stark contrast pooled models suffer from high MSE values.

Montgomery (1997) estimates CM models with exponential sales response functions of eleven competing brands in one category from 83 stores of one retail chain. Each brand-specific sales response function includes the own price and prices of all competing brands as predictors. The CM models reduce MSE in the estimation sample by 45 % compared to the homogeneous models. In the validation sample MSE for the CM models is lower by 20 % and 30 % over store-specific models and the homogeneous models, respectively. Compared to uniform pricing profits increase by about 10 % by restricted store level pricing which allows only moderate increases of prices and revenues over current values.

Montgomery and Rossi (1999) investigate sales using the same data set as Montgomery (1997). They specify a demand system which also includes total sales summed over 28 categories as predictor. This demand system requires only 12=11+1 coefficients for one marketing instrument. A conventional sales response model on the other hand would need 121=11×11 coefficients. The authors also add soft sign restrictions and allow for residual correlations between brands. The developed CM approach dominates several alternatives (e. g., a homogeneous model and store-specific models) in a split-half validation sample. The CM model reduces MSE in the validation sample by 17 % compared to the homogeneous model.

Boatwright et al. (1999) specify a multiplicative sales response model for one brand. They include as predictors feature and display of the respective brand. They also consider shelf prices and regular prices of the respective as well as of two competitive brands. OLS estimation of individual multiplicative models for each of 77 retail chains leads to incorrectly signed coefficients, for example negative coefficients for display. To eliminate this problem Boatwright et al. estimate a CM model with sign restrictions. They demonstrate profit increases which can be achieved by allocating a total promotion budget to retailers based on response coefficients of the CM model over a conventional volume-based approach.

Hruschka (2006b) investigates sales of nine brands in one category from 81 stores. He estimates CM versions of several parametric sales response models (e. g., linear, multiplicative, exponential, logistic) with price of the respective brand and average price across competitors as predictors. Marginal model densities, i. e., likelihoods averaged over parameters with respect to the prior density, show that multiplicative models are vastly superior. The CM multiplicative models reduce MSE in the estimation sample by 39 % over the related homogeneous model. From a managerial perspective the pooled model is sufficient, if the retail chain sets prices which are uniform for all its stores. Optimal uniform prices based on the heterogeneous model do not achieve higher expected profits at a probability level of 95 %.

Andrews et al. (2008) analyze data of five brands in one category from 28 stores. These authors estimate brand-specific SCAN*PRO models which have a parametric multiplicative form and price indices (actual price divided by regular price) as predictors. Andrews et al. (2008) also investigate CM and FM extensions of this model. The maximum improvement of R^2 by a CM model in the estimation sample amounts to 7 %, whereas the FM extension does not lead to improvements. In a validation sample FM and CM perform better than homogeneous model for four brands, but differences are small.

Weber und Steiner (2012) estimate a CM multiplicative model for the same data as Steiner et al. (2007). This model includes cross effects of three price tiers (all premium brands, all national brands and one store brand). Nine-fold cross validation shows that the CM parametric model outperforms the homogeneous model in terms of MSE for two of eight brands only. The models in Weber et al. (2017) also allow for residual correlation between brands. The CM model has a higher log marginal model density, but the difference to the homogeneous model is small. In the validation sample the homogeneous model turns out to be the best parametric model.

We continue with semiparametric models which also take heterogeneity into account. Hruschka (2006a) estimates sales response models for nine brands in a CM framework using data from 81 stores. To achieve flexibility he adds a MLP component to the multiplicative model. Sign restrictions on coefficients guarantee monotone decreasing own price effects as well as monotone increasing cross price effects. For each brand the number of hidden units for each brand which maximizes the log marginal density is chosen. No sales response model has more than three hidden units. The heterogeneous flexible model vastly outperforms the heterogeneous multiplicative model in terms of log marginal densities for each brand. Moreover, price effects implied by the flexible model differ for eight brands, especially at high prices of competitive brands. As a rule, the heterogeneous flexible model expects higher sales at medium prices of competitors if the own price is not too high. At low prices of competitive brands the heterogeneous flexible model implies lower sales for four of the nine brands.

Hruschka (2007) investigates cluster level pricing based on the semiparametric models estimated in Hruschka (2006a). Store clusters and cluster level prices are simultaneously determined by a stochastic optimization algorithm. Between three and 83 clusters are investigated with the 83 cluster solution being equivalent to store level pricing. By forming eight clusters 90 % of the profits of store level pricing can be achieved.

In their paper already mentioned above Weber et al. (2017) also estimate several flexible model types with Bayesian P-splines for own and cross price effects which are subject to monotonicity restrictions. These authors take latent heterogeneity into account by either a FM or a CM approach. The flexible CM model performs best, followed by the homogeneous flexible model. These two models improve RMSE in the validation sample relative to the homogeneous multiplicative model by 10.48 and 5.73 %, respectively. In contrast, all the investigated heterogeneous multiplicative models lead to worse RMSE values. Weber et al. (2017) therefore conclude that functional flexibility is more important for predictive performance than latent heterogeneity (across stores). Their results also demonstrate that latent heterogeneity pays off only if functional flexibility is considered as well. Own and cross price response curves show that nonlinearities are not reproduced sufficiently by parametric models.

By means of an evolutionary algorithm Weber et al. (2017) determine optimal prices which are uniform for homogeneous models, cluster-specific for FM models and store-specific for CM models. Pricing based on parametric models leads to a lower expected category profit both for their homogeneous and heterogeneous variants. Profit is highest if pricing is based on the CM flexible models, but it is only about 0.6 % lower for the homogeneous flexible models. That is why Weber et al. (2017) conclude that uniform pricing is sufficient for the data analyzed.

5. Endogeneity

A marketing variable is endogenous if it is related to an unobserved factor which also determines the sales or market share. Depending on the situation variables like shelf space, preference changes, product characteristics, word-of-mouth might be unobservable. A marketing variable becomes endogenous if it is set by managers in accordance with an unobserved factor (e. g., if managers increase prices or advertising budgets to take advantage of a favorable preference change).

Models which do not take endogeneity into account provide biased estimates of the effects of marketing variables. The direction of bias depends on the correlation of the observed variable with the unobserved factor. We illustrate such biases by two examples. If management increases prices under a favorable preference change, one obtains a positive correlation between prices and this unobserved factor. This positive correlation entails underestimation of (absolute) price effects. Estimated absolute price effects are too low because the positive effects of the preference change are ignored. If management sets higher advertising budgets given favorable word-of-mouth, advertising and this unobservable factor are positively correlated. Here the positive correlation causes overestimation of advertising effects. In other words, word-of-mouth effects are erroneously ascribed to advertising.

Whereas Rossi (2014) admits that endogeneity may play a role for cross sectional data, he thinks that endogeneity becomes less important if panel or time-series are available. The latter expectation is based on the assumption that unobserved variables which influence sales change less frequently (e. g., not every week). This assumption,

of course, may not always be valid as several studies show (Besanko 2001; Chintagunta 2001; Chintagunta et al. 2003; Park and Gupta 2012; Andrews and Ebbes 2014). These studies provide evidence for endogeneity biases, though they analyze weekly data.

Endogeneity can be tackled by equilibrium models which are based on the assumption that firms follow a certain mechanism to set marketing variables. These models typically encompass both demand and cost functions (see Chintagunta et al. 2006b for an overview). Instrumental variables represent an alternative approach to take endogeneity into account. We focus on instrumental variable techniques as contrary to equilibrium models they need less data and do not depend on the correct specification of firms' decision making behavior.

The best known instrumental variable technique is two step estimation, but is restricted to linear models. Generalized Methods of Moments (GMM) estimation can be extended to nonlinear models to which the models with flexible functions discussed in section 2 belong (see, e. g., Cameron and Trivedi 2005). Related Bayesian estimation methods are increasingly used, especially if heterogeneity is considered as well (e. g., in Hruschka and Gerhardt 2012).

Several studies compare two variants of homogeneous logit models, one variant which ignores endogeneity, and another which takes it into account by an instrumental variable approach. The majority of these studies concludes that ignoring endogeneity leads to underestimation of price effects (Nevo 2001; Besanko 2003; Chintagunta et al. 2003, 2006a), but Andrews and Ebbes (2014) infer overestimation.

Only few studies look at the possible endogeneity of non-price variables in aggregate response models. Chintagunta et al. (2006a) demonstrate overestimation of the effect of non-price promotion. In Hruschka and Gerhardt (2012) the effects of two store attributes (e. g., store size) are overestimated if endogeneity is not taken into account. Managers who base their decision on this biased model set store size too high. Optimizing profits and assuming a quadratic cost function Hruschka and Gerhardt (2012) show that store size determined for the biased model exceeds the value inferred for the unbiased model by more than 64 %.

Instrumental variables partition the variation of endogenous predictors into two parts, one that is uncorrelated and another which may be correlated with the error term. Instrumental variables should fulfill two requirements. Firstly, they should be exogenous, in other words be unrelated to the error term. Secondly, instrumental variables should be related to the endogenous variable. If instrumental variables fail to fulfill only one of these two requirements, unreliable coefficient estimates with high standard errors result.

For linear models exogeneity equals zero correlation of instrumental variables with the error term. For nonlinear models (e. g., aggregate choice models) exogeneity corresponds to conditional independence, i. e., an instrumental variable should affect the dependent variable only indirectly via movement of the endogenous predictor. For nonlinear models any function of an exogenous instrument may also serve as instrument. We emphasize that exogeneity cannot be tested empirically. One has to resort to theoretical arguments which support the kind of indirect effect just explained.

In the literature on aggregate market response modeling one finds several types of instrumental variables, namely lagged prices, costs, wholesale price, prices of other markets (e. g., other regions, other stores). We emphasize that lagged prices are not exogenous if customers' inventories or reference prices affect sales or market share. We note that lagged prices are used as instrumental variables in Chintagunta (2001).

Studies which select costs as instrumental variables typically consider material and labour costs (e. g., Chintagunta 2001; Besanko et al. 2003). Frequently costs are not available and have to be replaced by proxy variables. E. g., Nevo (2001) introduces brand and regional dummies to this end. According to Rossi (2014) net wholesale prices are almost certainly not exogenous to demand shocks. In contrast to this opinion Chintagunta et al. (2003) argue that in their study wholesale prices are not related to sales because the retail firm has a market share of only about 25 % in the region.

Price data from other regions are exogenous if demand shocks or marketing policies are not common across regions (Nevo 2001). Consequently, prices from other regions should not be chosen as instrumental variables, if they are set by the same firm who sets prices in the regions to be investigated.

Rossi (2014) thinks that only costs of non-price variables such as advertising, promotion, detailing in the pharmaceutical industry are exogenous. Chintagunta et al. (2006a) use advertising levels from other regions and lagged promotions as instruments. Advertising levels of other regions are problematic if chosen by the same firm. Lagged promotions are not appropriate instrumental variables if they exert a direct effect on actual demand.

To the opinion of Rossi (2014) non-price variables are often set based on (partial) knowledge of the response. Therefore he recommends a different econometric approach, which links marketing variables to parameters of the response function, which so far has been mainly used in disaggregate response modeling (Manchanda et al. 2004; Hruschka 2010).

According to the second requirement mentioned above instrumental variables should be related to the endogenous variable. In linear models this requirement can be measured by the increase of R^2 obtained by adding instrumental variables to exogenous variables in the so-called first stage regression with the endogenous variable as regressand. In nonlinear models one has to determine

the conditional mean function of an endogenous regressor with instruments as inputs. This conditional mean function is basically unspecified, but can be approximated by a high order polynomial of the instruments. The residuals of the conditional mean function represent the portion of an endogenous regressor which is independent of the instruments (Rossi 2014).

Many suggested instrumental variables frequently fail to fulfill this second requirement. For example, costs as well their proxies often do not vary sufficiently (Nevo 2001). Similarly, wholesale list prices are often smoother than retail prices (Rossi 2014).

Andrews and Ebbes (2014) develop a procedure to deal with possible endogeneity in aggregate logit models. A common demand shock is inferred if the respective logit model is favored by a likelihood ratio test. At this step two forms of common demand shocks can be investigated, which are either constant across stores or constant across weeks. For the first case Andrews and Ebbes (2014) recommend store-centered, for the second case week-centered instrumental variables. Subsequently Andrews and Ebbes (2014) follow a control function approach (Petrin and Train 2010) by adding residuals which are determined by a first stage regression of prices on instruments and demand shocks as predictors to the logit model. This extended logit model is selected if preferred by a likelihood ratio test.

In view of the difficulties to find appropriate instruments researchers may turn to instrument-free alternatives (Ebbes et al. 2009), i. e., the Higher Moments (HM) approach (Erickson and Whited 2002; Lewbel 1997), the Identification through Heteroscedasticity (IH) estimator (Rigobon 2003), the Latent Instrumental Variables (LIV) method (Ebbes et al. 2005) and the copula-based method (Park and Gupta 2012). We emphasize that only the copula-based method can be extended to nonlinear models (Ebbes et al. 2005; Lewbel 1997).

In the HM approach instruments are based on higher order moments of observed variables. Errors of the first stage regression and the aggregate response model must be independent and have moments of every order, but are not restricted with respect to distribution. Hruschka and Gerhardt (2012) apply the HM approach to investigate sales of a cross section of gas stations by a finite mixture regression model. Store attributes as well as socio-economic and competitive regional profiles constitute the predictors. Constructed instrumental variables consist of products of mean centered sales and each mean centered endogenous variables on one hand, products of each mean centered endogenous variables and each mean centered exogenous variable on the other hand.

The IH estimator is also based on higher order moments. It needs an observable grouping variable which describes the heteroskedastic structure of the errors. The LIV approach approximates an unobserved instrument by a latent discrete variable. The model is not identified if the endogenous regressor is normally distributed. In an empirical application presented by Ebbes et al. (2005) the increase of R^2 in the first stage regression is much higher for the LIV approach compared to observed instruments which were derived from theoretical arguments.

The approach of Park and Gupta (2012) provides consistent parameter estimates by means of a Gaussian copula model which reproduces the correlation between endogenous regressors and the error term. Additive error terms are assumed to be normally distributed. Endogenous regressors are required to be non-normal just as in the LIV approach. Discrete regressors are allowed, if they are not binary. The essential characteristic of this approach consists in adding one artificial regressor to the model for each endogenous regressor. Such an artificial regressor v_{jn}^* equals the output of the inverse standard normal distribution function Φ^{-1} with the value of the empirical distribution function F for a given value v_{jn} of the respective regressor as argument:

$$v_{jn}^* = \Phi^{-1}(F(v_{jn})) \tag{6}$$

Park and Gupta demonstrate how this approach can be applied to linear regression models and to the heterogeneous logit models based on aggregate data which we discussed in section 4.1. The fact that the copula-based approach is appropriate for nonlinear models constitutes an important advantage compared to the other instrument-free methods.

6. Conclusion and Outlook

Let us summarize the main implications of this overview and also offer an outlook on open questions which on the other hand constitute opportunities for future research. Biased estimates of the effects of marketing variables can be avoided by functional flexibility, latent heterogeneity and by taking endogeneity into account. Avoiding biases is important if researchers want to derive implications for marketing decision making. If researchers are not interested in measuring the effects of marketing variables, but instead look only on overall predictive performance, they can safely ignore the endogeneity problem (Rossi 2014). On the other hand, many studies demonstrate that allowing for functional flexibility and to some degree for latent heterogeneity often lead to better predictions.

As semiparametric models attain usually (much) better performances even in validation samples than better known parametric alternatives, researchers should switch to these more flexible methods. Such a change seems to be quite feasible as the often heavy data limitations of previous decades have been overcome, at least in consumer goods marketing. Another motivation is offered by the fact that in many studies functional flexibility turns out to be more important than latent heterogeneity across stores.

In the following we present a summarizing juxtaposition of the three flexible approaches, MLP, splines regression and kernel regression. As we have found no empirical comparisons between these three methods for aggregate market response modeling, we rely on studies with related applications. Ahmed et al. (2010) evaluate the forecasting performance of several flexible methods for about one thousand business time series. In their study the MLP as a rule performs best and clearly beats kernel regression. For disaggregate brand choice data Hruschka et al. (2004) extend a conventional parametric model by either a GAM with cubic smoothing splines or a MLP. Ten-fold cross validated log likelihood values show that the MLP extension performs better for brand choice data in two different product categories.

Both kernel regression and MLP are easy to implement even for larger number of predictors. On the other hand, spline regression becomes cumbersome if it includes interaction terms. Spline regression and GAM models have the advantage over kernel regression that implementation of shape restrictions is less complex (Hansen 2017).

The MLP does relatively well with respect to computation time. Barron (1993) proves that for the MLP the rate of convergence does not decrease if the number of predictors increases. This property is in contrast to flexible models based on (truncated) polynomials which need more iterations the more predictors are considered. Kernel regression shares the same disadvantage with respect to the rate of convergence (Cameron and Trivedi 2005).

The MLP with logistic or other sigmoid activation functions for hidden units produces smooth curves, which asymptote rapidly to zero or one outside the data range and are well behaved if data are extrapolated (Denison et al. 2002). Based on the literature discussed in section 3 we recommend to use the MLP carefully selecting the number of hidden units to prevent overfitting. We also suggest to add monotone and other appropriate shape restrictions to guarantee that estimation results conform to economic theory which so far has been done in few studies applying flexible methods (Hruschka 2006a; Brezger and Steiner 2008; Weber et al. 2017).

We note that CM have been applied more frequently than FM approaches to deal with latent heterogeneity in aggregate response models. Moreover, studies which compare CM and FM as a rule indicate a better statistical performance of the former approach.

Most studies with parametric response functions which also allow for latent heterogeneity obtain a better statistical performance than related homogeneous models. Many of these studies also show that heterogeneous models lead to different implications for optimal pricing. This way these studies provide evidence that estimates of homogeneous parametric models are often biased. Therefore latent heterogeneity should be regularly investigated in parametric aggregate market response modeling.

There is a lack of studies which compare homogeneous and heterogeneous versions of semiparametric models. To our knowledge only Weber et al. (2017) investigate this question. In their study the CM version of a flexible model attains better statistical performance, but leads to only very small improvements for optimal pricing. This result seems to indicate that heterogeneity is less important than functional flexibility, but we have to emphasize that it has been obtained in one study only.

Latent heterogeneity within stores has been considered for the logit model (less frequently for the probit model) starting from a linear indirect utility specification. More flexible specifications have so far not been investigated. In this regard future research could start a from a homogeneous flexible choice model (which can be found in e. g., Briesch et al. 2002; Abe et al. 2004; Hruschka et al. 2004) and add heterogeneity of parameters in a manner similar to what has been done to arrive at the heterogeneous logit model. Of course, estimation of heterogeneous logit models with flexible indirect utility function will be more involved.

Extant studies consider either latent heterogeneity within stores or latent heterogeneity across stores. Combining these two types of latent heterogeneity in one model presents an obvious extension.

Tackling endogeneity by instrumental variable methods often leads to problems to find appropriate instrumental variables. The latter should be related to the endogenous variable, but also be unrelated to the error term of a model. Researchers may bypass these problems by turning to instrument-free methods of which the copula-based approach introduced by Park und Gupta (2012) seems to be most attractive if the relevant assumptions hold. If these assumptions are violated, linking marketing variables to parameters of flexible aggregate response functions (Manchanda et al. 2004; Hruschka 2010) constitutes a viable solution.

To our knowledge existing studies with flexible aggregate response functions have not taken endogeneity into account. The copula-based approach is appropriate for nonlinear models in principle, but its application to any of the semiparametric methods dealt with in section 3 has not been demonstrated.

References

Abe, M., Boztuğ, Y., & Hildebrandt, L. (2004). Investigating the Competitive Assumption of Multinomial Logit Models of Brand Choice by Nonparametric Modeling, *Computational Statistics*, 19(4), 635–657.

Ahmed, N. K., Atiya, A. F., El Gayar, N. & El-Shishiny, H. (2010). An Empirical Comparison of Machine Learning Models for Time Series Forecasting. *Econometric Reviews* 29(5–6), 591–621.

Ainscough, T. L., & Aronson, J. E. (1999). An Empirical Investigation and Comparison of Neural Networks and Regression for Scanner Data Analysis. *Journal of Retailing and Consumer Services*, 6(4), 205–217.

Albers, S. (2012). Optimizable and Implementable Aggregate Response Modeling for Marketing Decision Support. *International Journal of Research in Marketing*, 29(2), 111–122.

Andrews, R. L., & Ebbes, P. (2014). Properties of Instrumental Variables Estimation in Logit-based Demand Models Finite Sample Results. *Journal of Modelling in Management*, 9(3), 261–289.

Andrews, R. L., Currim, I., & Leeflang, P. S. H. (2011). A Comparison of Sales Response Predictions From Demand Models Applied to Store-Level versus Panel Data. *Journal of Business & Economic Statistics*, 29(2), 319–326.

Andrews, R., Currim, I., Leeflang, P., & Lim, J. (2008). Estimating the SCAN*PRO Model of Store Sales: HB, FM or just OLS? *International Journal of Research in Marketing*, 25(1), 22–33.

Barron, A. R. (1993). Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Transactions on Information Theory*, 39(3), 930–945.

Besanko, D., Dubé, J.-P., & Gupta, S. (2003). Competitive Price Discrimination Strategies in a Vertical Channel Using Aggregate Retail Data. *Management Science*, 49(9), 1121–1138.

Blattberg, R., & George, E. (1991). Shrinkage Estimation of Price and Promotional Elasticities. *Journal of the American Statistical Association*, 86(414), 304–315.

Boatwright, P., McCulloch, R., & Rossi, P. (1999). Account Level Modeling for Trade Promotion: An Application of a Constrained Parameter Hierarchical Model. *Journal of the American Statistical Association*, 94(448), 1063–1073.

Brezger, A., & Steiner, W. (2008). Monotonic Regression Based on Bayesian P-Splines: An Application to Estimating Price Response Functions from Store-Level Scanner Data. *Journal of Business & Economic Statistics*, 26(1), 90–104.

Briesch, R. A., Chintagunta, P., & Matzkin, R. L. (2002). Semiparametric Estimation of Brand Choice Behavior. *Journal of the American Statistical Association*, 97(460), 973–982.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics*, Cambridge: Cambridge University Press.

Carpenter, G. S., Cooper. L. G., Hanssens, D. M., Midgley, D. F. (1988). Modeling Asymmetric Competition. *Marketing Science*, 7(4), 393–412.

Chintagunta, P. K. (2001). Endogeneity and Heterogeneity in a Probit Demand Model: Estimation Using Aggregate Data. *Marketing Science*, 20(4), 442–456.

Chintagunta P. K., Dubé, J. P., & Singh, V. (2003). Balancing Profitability and Customer Welfare in a Supermarket Chain. *Quantitative Marketing and Economics*, 1(1), 111–147.

Chintagunta, P., Kadiyali, V., & Vilcassim, N. J., (2006a). Endogeneity and Simultaneity in Competitive Pricing and Advertising: A Logit Demand Analysis. *Journal of Business*, 79(6), 2761–2788.

Chintagunta, P. K., Erdem, T., Rossi, P., & Wedel, M. (2006b). Structural Modeling in Marketing: Review and Assessment. *Marketing Science*, 25(6) 604–616.

Cooper, L.G., Nakanishi, M. (1988*). Market Share Analysis*, Boston: Kluwer.

Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signal and Systems*, 2(4), 303–314.

de Boor, C. (2001*). A Practical Guide to Splines*, Berlin: Springer.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., & Smith, A. F. M (2001). *Bayesian Methods for Nonlinear Classification and Regression*, Chichester, England: John Wiley.

Ebbes, P., Wedel, M., & Böckenholt, U. (2009). Frugal IV Alternatives to Identify the Parameter for an Endogenous Regressor. *Journal of Applied Econometrics*, 24(3), 446–468.

Ebbes, P., Wedel, M., Böckenholt, U., & Steerneman, A. G. M. (2005). Solving and Testing for Regressor-Error (in)Dependence When no Instrumental Variables are Available: With New Evidence for the Effect of Education on Income. *Quantitative Marketing and Economics*, 3(4), 365–392.

Erickson, T., & Whited, T. M. (2002). Two-step GMM Estimation of the Errors-invariables Model Using High-order Moments. *Econometric Theory*, 18(3), 776–799.

Fahrmeier, L., Kneib, T., & Lang, S. (2007). *Regression. Modelle, Methoden und Anwendungen*, Berlin: Springer.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines (with discussion). *Annals of Statistics*, 19(1), 1–141.

Greene, W. H. (2003). *Econometric Analysis*, 5th Ed., Upper Saddle River, NJ: Prentice Hall.

Gruca, T. S., Klemz, B. R., & Petersen, E. A.. (1998). Mining Sales Data Using a Neural Network Model of Market Response. *SIGKDD Explorations*, 1(1), 39–43.

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

Hansen, B. E. (2017). *Econometrics*, http://www.ssc.wisc.edu/~bhansen/econometrics/(Accessed May 12, 2017).

Hanssens, D. M., Parsons, L. J., & Schultz, R. L. (2001). *Market Response Models. Econometric and Time Series Analysis*, 2nd Ed., Boston, MA: Kluwer Academic Publishers.

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Scienc*e, 1(3), 297–318.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models*, London: Chapman & Hall.

Hornik, K.., Stinchcombe, M., & White, H. (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5), 359–366.

Hruschka, H. (1991a). Einsatz künstlicher neuraler Netzwerke zur Datenanalyse im Marketing. *Marketing ZFP*, 13(4), 217–225.

Hruschka, H. (1991b). Marktreaktionsfunktionen mit Interaktionen zwischen Marketing-Instrumenten. *Zeitschrift für Betriebswirtschaft*, 61(3), 339–355.

Hruschka, H. (1993). Determining Market Response Functions by Neural Network Modeling. A Comparison to Econometric Techniques. *European Journal of Operational Research*, 66(1), 27–35.

Hruschka, H. (1996). *Marketing-Entscheidungen*, München: Vahlen.

Hruschka, H. (2001). An Artificial Neural Net Attraction Model (ANNAM) to Analyze Market Share Effects of Marketing Instruments. *Schmalenbach Business Review – zfbf*, 53(1), 27–40.

Hruschka, H. (2002). Market Share Analysis Using Semi-Parametric Attraction Models. *European Journal of Operational Research*, 138(1), 212–225.

Hruschka H (2006a). Relevance of Functional Flexibility for Heterogeneous Sales Response Models: a Comparison of Parametric and Semi-nonparametric Models. *European Journal of Operational Research*, 174(2), 1009–1020.

Hruschka H (2006b). Statistical and Managerial Relevance of Aggregation Level and Heterogeneity in Sales Response Models. *Marketing ZFP – JRM*, 28(2), 94–102.

Hruschka, H. (2007). Clusterwise Pricing in Stores of a Retail Chain. *OR Spectrum*, 29(4), 579–595.

Hruschka, H. (2008). Neural Nets and Genetic Algorithms in Marketing. In B. Wierenga, (Ed.), *Handbook of Marketing Decision Models*, New York: Springer, 399–433.

Hruschka, H. (2010). Considering Endogeneity for Optimal Catalog Allocation in Direct Marketing. *European Journal of Operational Research*, 206(1), 239–247.

Hruschka, H., & Gerhardt, R. G. (2012). Endogeneity of Store Attributes in Heterogeneous Store-Level Sales Response Models. *OR Spectrum*, 34(1), 199–214.

Hruschka, H., Fettes, W., & Probst, M. (2004). An Empirical Comparison of the Validity of a Neural Net Based Multinomial Logit Choice Model to Alternative Model Specifications. *European Journal of Operational Research*, 159(1), 166–180.

Kalyanam, K., & Shively, T. S. (1998). Estimating Irregular Pricing Effects: a Stochastic Spline Regression Approach. *Journal of Marketing Research*, 35, 16–29.

Kolsarici, C., & Vakratsas, D. (2011). The Complexity of Multi-Media Effects. *Marketing Science Institute Working Paper Series*, Cambridge, MA.0

Lancaster, T. (2004). *An Introduction to Modern Bayesian Econometrics*, Oxford, UK: Blackwell Publishing.

Lang, S., Steiner, W., Weber, A., & Wechselberger, P. (2015). Accommodating Heterogeneity and Nonlinearity in Price Effects for Predicting Brand Sales and Profits. *European Journal of Operational Research*, 246(1), 232–241.

Lewbel A. (1997). Constructing Instruments for Regressions with Measurement Error When no Additional Data are Available, with an Application to Patents and R&D. *Econometrica*, 65(5), 1201–1213.

Lilien, G. L., Kotler, P., Moorthy, K. S., (1992). *Marketing Models*, Englewood Cliffs, NJ Prentice-Hall.

Lindley, D. V. & Smith, A. F. M. (1972): Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society B*, 34, 1–14.

Manchanda, P., Rossi P. E., & Chintagunta, P. K. (2004). Response Modeling with Nonrandom Marketing-Mix Variables. *Journal of Marketing Researc*h, 41(4), 467–478.

Montgomery, A. L. (1997). Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data. *Marketing Science*, 16(4), 315–337.

Montgomery, A. L., & Rossi P. E. (1999). Estimating Price Elasticities with Theory-based Priors. *Journal of Marketing Research*, 36(4), 413–423.

Natter, M., & Hruschka, H. (1997). Ankerpreise als Erwartungen oder dynamische latente Variablen in Marktreaktionsmodellen. *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung*, 49(9), 747–764.

Nevo, A. (2000). A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand. *Journal of Economics & Management Strategy*, 9(4), 513–548.

Nevo, A. (2001). Measuring Market Power in the Ready-to-Eat Cereal Industry. *Econometrica*, 69(2), 307–342.

Pantelidaki, S., & Bunn, D. (2005). Development of a Multifunctional Sales Response Model with the Diagnostic Aid of Artificial Neural Networks. *Journal of Forecasting*, 24(7), 505–521.

Park, S., Gupta, S. (2012). Handling Endogenous Regressors by Joint Estimation Using Copulas. *Marketing Science, 31(4)*, 567–586.

Petrin, A., & Train, K. (2010). A Control Function Approach to Endogeneity in Consumer Choice Models. *Journal of Marketing Research*, 4(1), 3–13.

Rigobon, R. (2003). Identification Through Heteroskedasticity. *The Review of Economics and Statistics*, 85(4), 777–792.

Rossi, P. E. (2014). Even the Rich Can Make Themselves Poor: A Critical Examination of IV Methods in Marketing Applications. *Marketing Science*, 33(5), 655–672.

Steiner, W., Brezger, A., & Belitz, C. (2007). Flexible Estimation of Price Response Functions Using Retail Scanner Data. *Journal of Retailing and Consumer Services*, 14(6), 383–393.

van Heerde, H., Leeflang, P. S. H., & Wittink, D. R. (2001). Semiparametric Analysis to Estimate the Deal Effect Curve. *Journal of Marketing Research*, 38(2), 197–215.

van Wezel, M. C., & Baets, W. R. J. (1995). Predicting Market Responses with a Neural Network: the Case of Fast Moving Consumer Goods. *Marketing Intelligence & Planning*, 13(7), 23–30.

Weber, A., & Steiner, W. (2012). Zur Berücksichtigung von Heterogenität versus funktionaler Flexibilität in Absatzreaktionsmodellen: Eine empirische Studie auf Basis von Handelsdaten. *Zeitschrift für Betriebswirtschaft*, 82(12), 1337–1365.

Weber, A., Steiner, W., & Lang, S. (2017). A Comparison of Semiparametric and Heterogeneous Store Sales Models for Optimal Category Pricing. *OR Spectrum*, 39(2), 403–445.

Wierenga, B., & Kluytmans, J. (1996). Prediction with Neural Nets in Marketing Times Series Data. *Management Report Series*, Erasmus Universiteit Rotterdam.

Keywords

Market Response Models, Flexible Functions, Latent Heterogeneity, Endogeneity.

MARKETING
ZFP – Journal of Research and Management

Editor-in-Chief: Prof. Dr. Bernhard Swoboda, Chair for Marketing and Retailing, University of Trier, Universitätsring 15, D-54296 Trier, Phone: +49 651 201 3050, Fax: +49 651 201 4165,
Email: editor@marketing-zfp.de

Senior Editors: Prof. Dr. Heribert Gierl, Universityof Augsburg, Prof. Dr. Andrea Gröppel-Klein, Saarland University, Prof. Dr. Lutz Hildebrandt, Humboldt-University of Berlin, Prof. Dr. Hans Mühlbacher, International University of Monaco, Prof. Dr. Henrik Sattler, University of Hamburg, Prof. Dr. Udo Wagner, University of Vienna.

Manuscripts: We ask all authors who would like to submit a paper to send this paper to the editor-in-chief Bernhard Swoboda via mail: editor@marketing-zfp.de. Neither the publisher nor the editors assume any liability for unsolicited manuscripts. Unsolicited manuscripts will only be returned if accompanied by return postage. The accteptance of a contribution has to be in writing.

Publisher: C.H.BECK oHG, Wilhelmstr. 9, 80801 München; postal address: P.O.Box 40 03 40, 80703 München; phone: +49 89 38189 0; fax: +49 89 38189 398, Bank account: Postbank München IBAN: DE82 7001 0080 0006 2298 02, BIC: PBNKDEFFXXX.

The publisher is a *offene Handelsgesellschaft* under German law with Dr. Hans Dieter Beck and Dr. h.c. Wolfgang Beck as partners.

Subscription: An annual subscription to the journal comprises four issues.

Subscription rates 2017: € 219 (VAT incl.) annual subscription rate, discounted rate for students (related subject, proof needed) € 135 (VAT incl.), campus licence € 349 (VAT incl.). Single Issue: € 61,50 (VAT incl.), shipping charges have to be added to the rates. Subscription and rate include print issue and a licence for the online archive. The components cannot be cancelled separately. Complaints about copies not received must be lodged within 6 weeks starting at the end of the quarter.

Subscription service: Please order with either the publisher or any book shop.

CustomerServiceCenter: Phone: +49 89 38189 750, Fax: +49 89 38189 358, E-Mail: kundenservice@beck.de

Cancellation: The subscription may be cancelled in writing 6 weeks before the end of a calendar year.

Citation: Marketing ZFP – Journal of Research and Management, number of volume(number of issue), year, page.

Typesetting: FotoSatz Pfeifer GmbH, 82152 Krailling.

Printing: Kessler Druck und Medien GmbH & Co. KG, Michael-Schäffer-Straße 1, 86399 Bobingen.