

# MATHEMATIK-ARBEITSPAPIERE

A: MATHEMATISCHE FORSCHUNGSPAPIERE

SEMIPARAMETRIC ASSOCIATION MODELS:  
ESTIMATION AND ASYMPTOTIC INFERENCE

GERHARD OSIUS

No. 60

JUNE 2006



# MATHEMATIK-ARBEITSPAPIERE

A: MATHEMATISCHE FORSCHUNGSPAPIERE

SEMIPARAMETRIC ASSOCIATION MODELS:  
ESTIMATION AND ASYMPTOTIC INFERENCE

GERHARD OSIUS

No. 60

JUNE 2006

FACHBEREICH MATHEMATIK UND INFORMATIK  
UNIVERSITÄT BREMEN

Bibliotheksstraße  
D-28359 Bremen  
Germany

# Semiparametric Association Models: Estimation and Asymptotic Inference

Gerhard Osius

*Institut für Statistik, Fachbereich 3, Universität Bremen  
Bibliotheksstrasse, 28359 Bremen, Germany  
e-mail: osius@math.uni-bremen.de*

1. Introduction and outline
  2. The odds ratio function
  3. Association models
  4. Estimation
    - 4.1 Unconditional sampling
    - 4.2 Conditional sampling
    - 4.3 Log-bilinear association
  5. Conditional likelihood
  6. Asymptotics and consistency
  7. Asymptotic normality
  8. Discussion
- References  
Appendix: Proofs

## Abstract

The association between a pair of random elements  $X$  and  $Y$  (e.g. vectors) is completely determined by an odds ratio function, which can be specified up to an unknown parameter vector  $\theta$  (cf. Osius 2000, 2004). Using results on  $I$ -projections we first show - under weaker conditions than before - that a parametric odds ratio model is semiparametric in the sense, that it does not restrict the marginal distributions of  $X$  and  $Y$ . Inference for the odds ratio parameter  $\theta$  may be obtained from sampling either  $Y$  conditional on  $X$  or vice versa. Generalizing results from Prentice and Pyke (1979) and Weinberg and Wacholder (1993), we show that for samples of  $X$  conditional on  $Y$  the asymptotic inference for  $\theta$  may be obtained as if  $Y$  had been sampled conditional on  $X$  - including the consistency and asymptotic normality of the estimate. Common regression models - e.g. generalized linear models with canonical link or multivariate linear resp. logistic regression models - are recognized as odds ratio models where the regression parameter  $\beta$  is closely related (or even identical) to the odds ratio parameter  $\theta$ . For these models our results provides an alternative way of sampling ( $X$  conditional on  $Y$ ) in order to obtain asymptotic inference for the regression parameter  $\beta$  (e.g. testing a linear hypothesis) using standard statistical software.

*Keywords:* log-bilinear association, generalized linear model,  $I$ -projection, logistic regression, multivariate linear regression, odds ratio, semiparametric model.

## 1. Introduction and Outline

A common approach to describe the relationship between a random output variable  $Y$  of interest (e.g. a health status) and a random input vector  $X$  (e.g. consumption of tobacco, alcohol and other risk factors) is by means of a parametric regression model which specifies the conditional distribution of  $Y$  given  $X=x$  up to an unknown parameter vector. In the most simple case  $Y$  is an indicator (e.g. for the presence of a disease) taking values in  $\Omega_Y = \{0, 1\}$  and the conditional distribution is binomial  $B(1, p(x))$ . The popular logistic regression model relates the logistic transform of  $p(x)$  and a vector  $\mathbf{z} = h(x) \in \mathbb{R}^S$  of covariates - obtained from  $x$  by an suitable function  $h$  - through

$$\text{logit } p(x) := \log(p(x)/[1-p(x)]) = \gamma + \mathbf{z}^T \boldsymbol{\theta}$$

with parameters  $\gamma \in \mathbb{R}$  and  $\boldsymbol{\theta} \in \mathbb{R}^S$ . The appropriate sampling scheme for this model is to sample  $Y$  conditional  $X=x$  for specified values of  $x$ . In epidemiology this is called a *cohort study*, each of the  $J$  cohorts being determined by its value  $x$ . In contrast, the so called *case-control studies* are obtained by sampling  $X$  conditional on  $Y=1$  (cases) resp.  $Y=0$  (controls). An important result by Prentice and Pyke (1979) briefly states, that asymptotic inference for the parameter  $\boldsymbol{\theta}$  (but *not* for  $\gamma$ ) in a case-control study may be obtained as if the data came from a cohort study. Actually their work covers the multivariate logistic regression model (cf. example 3 later) for a random variable  $Y$  taking values in  $\{0, 1, \dots, K\}$ . Our aim is to generalize the results of Prentice and Pyke to the class of semiparametric *odds ratio models* for random elements (e.g. vectors)  $Y$  and  $X$  introduced in Osius (2000). The odds-ratio function  $OR(x, y)$  for the joint density  $p(x, y)$  of  $X$  and  $Y$  is defined as a cross-product ratio with respect to fixed reference values  $x^\circ$  and  $y^\circ$

$$OR(x, y) = \frac{p(x, y) \cdot p(x^\circ, y^\circ)}{p(x, y^\circ) \cdot p(x^\circ, y)}$$

An equivalent description is obtained by taking the corresponding cross-product ratio for the *conditional* density  $p(y|x)$  of  $Y$  given  $X$  - or vice versa. It has been shown in Osius (2000, 2004) that - under suitable assumptions - the joint distribution of  $(X, Y)$  is uniquely determined by the odds-ratio function and the marginal distributions of  $X$  and  $Y$ . And conversely, for any pair of marginal distributions for  $X$  and  $Y$  and a suitable odds-ratio function there exists a (unique) joint distribution having these properties. The odds-ratio function thus captures the

*complete* association structure of  $X$  and  $Y$  by ignoring the information contained in the marginal distributions. A parametric odds-ratio model specifies only the odds ratio function up to an unknown parameter vector  $\boldsymbol{\theta}$ , i.e.

$$\log OR(x, y) = \psi_{\boldsymbol{\theta}}(x, y).$$

This model is semiparametric in the sense that it does not restrict the marginal distributions of  $X$  and  $Y$ , but only the association structure. An important class are log-bilinear association models where the log-odds ratio function is a bilinear with respect to given transformations  $\mathbf{z} = h_X(x)$  and  $\mathbf{v} = h_Y(y)$ , i.e.

$$\log OR(x, y) = \mathbf{z}^T \boldsymbol{\theta} \mathbf{v}. \quad (1.1)$$

In fact, some widely used regression models like generalized linear models and multivariate linear resp. logistic regression models have a log-bilinear association structure. The assumptions concerning the conditional distribution of  $Y$  given  $X$  in these regression models may be removed by passing to the corresponding log-bilinear odds ratio model. One advantage of odds ratio models over regression models is that inference about the odds ratio parameter  $\boldsymbol{\theta}$  may be obtained from sampling  $X$  conditionally on  $Y$  or vice versa. To prove this, we first show that maximum likelihood estimation is invariant under both conditional sampling schemes, i.e. the estimate  $\hat{\boldsymbol{\theta}}$  maximizing the conditional likelihood  $L_{X|Y}$  for samples of  $X$  given  $Y$  also maximizes the corresponding conditional likelihood  $L_{Y|X}$  for samples of  $Y$  given  $X$  - and conversely. Generalizing the result in Prentice and Pike (1979), we show that the estimated asymptotic covariance matrix for  $\hat{\boldsymbol{\theta}}$  is invariant under both conditional sampling schemes, too. Hence asymptotic inference (e.g. tests) concerning the odds ratio parameter  $\boldsymbol{\theta}$  may be obtained from a sample  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , drawn conditionally on  $Y$  as if the sample was drawn conditional on  $X$ .

The paper is organized as follows. Section 2 deals with the fundamental result, that the joint distribution of  $(X, Y)$  is uniquely determined by its odds ratio function and the marginal distributions (uniqueness theorem), and that each of these three components can vary independently of another (existence theorem). The latter result will be proved under weaker assumptions than in Osigus (2000) using a different approach. Association models are introduced in section 3 and some widely used regression models are recognized having a log-bilinear association. Although log-bilinear association often is of primary interest, we derive the main results for more general odds ratio models determined by

$$\log OR(x, y) = G(\mathbf{z}, \mathbf{v}, \boldsymbol{\theta}), \quad (1.2)$$

where  $G$  is a given (sufficiently smooth) function. Maximum likelihood estimation is addressed in section 4, where we establish that the estimate  $\hat{\boldsymbol{\theta}}$  is invariant under the usual sampling schemes: unconditional or conditional on  $X$  resp.  $Y$ . For log-bilinear association models in particular, the likelihood to maximize is identified as the likelihood of a corresponding log-linear model for a suitable contingency table. Hence results on the existence and uniqueness as well as techniques to compute the estimate are already available.

Knowing that the estimate  $\hat{\boldsymbol{\theta}}$  is invariant under conditional sampling given either  $X$  or  $Y$ , we now establish in several steps our main result, that its estimated asymptotic normal distribution is invariant, too. More precisely, we consider sampling  $X$  conditional on  $Y$  and then maximize the "reverse" conditional log-likelihood  $\ell(\boldsymbol{\lambda})$  - arising for conditioning  $Y$  on  $X$  - with respect to  $\boldsymbol{\lambda} = (\boldsymbol{\theta}, \boldsymbol{\gamma}^*)$ , where  $\boldsymbol{\gamma}^*$  is a nuisance parameter vector. For the information matrix  $\mathbf{I}(\boldsymbol{\lambda}) = E(-D_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^2 \ell(\boldsymbol{\lambda}))$  we show in section 5, that the submatrix  $[\mathbf{I}^{-1}(\boldsymbol{\lambda})]_{\boldsymbol{\theta}\boldsymbol{\theta}}$  of  $\mathbf{I}^{-1}(\boldsymbol{\lambda})$  corresponding to  $\boldsymbol{\theta}$  is indeed the asymptotic covariance matrix of  $\hat{\boldsymbol{\theta}}$ . To establish the asymptotic normality of the estimate  $\hat{\boldsymbol{\lambda}}$ , we first prove its consistency in section 6. Our asymptotic approach applies to a *fixed* set  $\{y_0, \dots, y_K\}$  of values for  $Y$  to be conditioned upon and independent samples of size  $n_k$  from each conditional distribution of  $X$  given  $Y = y_k$ , such that  $n = \sum_k n_k$  tends to infinity while the ratios  $n_k/n$  remain *fixed*. In section 7 the asymptotic normality is shown more generally for *any* (weakly) consistent estimate  $\hat{\boldsymbol{\lambda}}$  which solves the estimating equation at least approximately, i.e.  $D_{\boldsymbol{\lambda}} \ell(\hat{\boldsymbol{\lambda}}) = o_p(\sqrt{n})$ . Using the observed information  $\mathbf{J}(\hat{\boldsymbol{\lambda}}) = -D_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^2 \ell(\hat{\boldsymbol{\lambda}})$  as a consistent estimate of  $\mathbf{I}(\boldsymbol{\lambda})$ , we finally obtain the asymptotic normality of the odds ratio estimate

$$\hat{\boldsymbol{\theta}} \underset{\text{as}}{\sim} N(\boldsymbol{\theta}, [\mathbf{J}^{-1}(\hat{\boldsymbol{\lambda}})]_{\boldsymbol{\theta}\boldsymbol{\theta}}).$$

The estimated asymptotic covariance matrix here is exactly the same as if sampling had been conditional on  $X$  for the observed  $x$ -values.

In our proofs we do not attempt to derive the results under the weakest possible assumptions - which typically are rather technical. Instead we use a few easily interpretable conditions to establish the asymptotic results above for the model (1.2). These conditions will be verified for the log-bilinear association model (1.1) under mild distributional assumptions.

Let us finally note, that the general approach adopted here is *symmetric* in  $X$  and  $Y$

so that interchanging  $X$  with  $Y$  in any concept or argument entails its *dual*. Most of the proofs as well as some preliminary results are given in the appendix.

## 2. The Odds Ratio Function

Consider arbitrary non-empty spaces  $\Omega_X$  resp.  $\Omega_Y$  with  $\sigma$ -algebras  $\mathcal{B}_X$  resp.  $\mathcal{B}_Y$  and denote the product  $\sigma$ -algebra on  $\Omega = \Omega_X \times \Omega_Y$  by  $\mathcal{B}$ . Let  $\mathcal{P}$  be the space of all probability measures  $P$  on  $(\Omega, \mathcal{B})$  and denote the marginal distributions of  $P$  on  $\Omega_X$  resp.  $\Omega_Y$  by  $P^X$  resp.  $P^Y$ . The definition of an odds ratio function for  $P$  requires a positive density with respect to a product measure and a natural (as well as most general) choice is the product  $P^{XY} = P^X \times P^Y$  of the marginals. This leads to the subspace  $\mathcal{P}_{\ll} \subset \mathcal{P}$  of probability measures  $P$  having a positive density with respect to  $P^{XY}$ , or equivalently, are dominated by and dominate  $P^{XY}$ , i.e.

$$\mathcal{P}_{\ll} = \{P \in \mathcal{P} \mid \frac{dP}{dP^{XY}} > 0\} = \{P \in \mathcal{P} \mid P \ll P^{XY} \ll P\}.$$

For any  $P \in \mathcal{P}_{\ll}$  with density  $p = dP/dP^{XY}$  its *odds ratio function*  $OR_p$  with respect to fixed reference values  $x^\circ \in \Omega_X$  and  $y^\circ \in \Omega_Y$  is defined on  $\Omega \times \Omega$  by

$$OR_p(x, y) = \frac{p(x, y) \cdot p(x^\circ, y^\circ)}{p(x, y^\circ) \cdot p(x^\circ, y)}. \quad (2.1)$$

The choice of the dominating product measure  $P^{XY}$  is not essential. More precisely, let  $\nu_X$  resp.  $\nu_Y$  be  $\sigma$ -finite measures on  $\Omega_X$  resp.  $\Omega_Y$  and  $\nu = \nu_X \times \nu_Y$ . If  $P \in \mathcal{P}$  has a *positive* density  $p_\nu = dP/d\nu$  then  $P^X$  and  $P^Y$  have positive densities  $p_X = dP^X/d\nu_X$  and  $p_Y = dP^Y/d\nu_Y$ . Hence  $p(x, y) = p_\nu(x, y) \cdot p_X(x) \cdot p_Y(y) > 0$  is a  $P^{XY}$ -density of  $P$  and the odds ratio function may also be expressed with  $p$  replaced by  $p_\nu$  in (2.1)

$$OR_p(x, y) = \frac{p_\nu(x, y) \cdot p_\nu(x^\circ, y^\circ)}{p_\nu(x, y^\circ) \cdot p_\nu(x^\circ, y)}.$$

Hence the odds ratio function of  $P$  is invariant with respect to the dominating product measure of  $P$ . - Since the density  $p$  of  $P$  is only unique up to almost-sure equality, the same holds for the odds ratio function  $OR_p$  of  $P$ , which nevertheless will also be denoted simply by  $OR(P)$ . The log odds ratio function may be written in terms of the log-density

$$\log OR_p(x, y) = \log p(x, y) + \log p(x^\circ, y^\circ) - \log p(x, y^\circ) - \log p(x^\circ, y). \quad (2.2)$$

It is convenient to view any  $P \in \mathcal{P}$  as a joint distribution of a pair  $(X, Y)$  of random elements defined on some probability space with values in  $\Omega$ . Following usual practice, we extend concepts defined for probability measures to random elements,

e.g.  $OR(X, Y) = OR(P)$  denotes the odds ratio function of  $(X, Y)$ .

To show that the odds ratio function completely characterizes the association between  $X$  and  $Y$ , we have to restrict the joint distribution  $P$  by requiring that the log-density  $\log p$  is  $P^{XY}$ -integrable, or equivalently, that the *Kullback-Leibler information* (Kullback 1959)

$$I(P^{XY} | P) = \int \log \left( \frac{dP^{XY}}{dP} \right) dP^{XY}$$

is *finite*. Within the subclass

$$\mathcal{P}_f = \left\{ P \in \mathcal{P}_{\ll} \mid \log \left( \frac{dP}{dP^{XY}} \right) \text{ is } P^{XY}\text{-integrable} \right\} = \left\{ P \in \mathcal{P}_{\ll} \mid I(P^{XY} | P) < \infty \right\}$$

we prove in **A1**, that  $P$  is uniquely determined by its marginal distributions and its odds ratio function.

**Theorem 1 (Uniqueness):** Any  $P_1, P_2 \in \mathcal{P}_f$  having the same marginals  $P_1^X = P_2^X$ ,  $P_1^Y = P_2^Y$  and the same odds ratio function  $OR(P_1) = OR(P_2)$  agree:  $P_1 = P_2$ .

Next we want to "construct" a distribution  $P$  on  $\Omega$  by specifying its marginal distributions and its (log) odds ratio function. For given distributions  $\pi_X$  on  $\Omega_X$  and  $\pi_Y$  on  $\Omega_Y$  and a measurable function  $\psi$  on  $\Omega$  we investigate under which conditions there exists a  $P \in \mathcal{P}_f$  with  $P^X = \pi_X$ ,  $P^Y = \pi_Y$  and  $\log OR(P) = \psi$ . First of all,  $\psi$  has to satisfy the obvious constraints

$$\text{(OR1)} \quad \psi(x, y^\circ) = 0, \quad \psi(x^\circ, y) = 0 \quad \text{for all } x, y.$$

Furthermore from  $P \in \mathcal{P}_f$  and (2.2) we obtain two *necessary* integrability conditions:

- (E1)**  $\psi$  is  $\pi_X \times \pi_Y$ -integrable.
- (E2)** There exists  $\pi_X$ -integrable  $\beta: \Omega_X \rightarrow \mathbb{R}$  and  $\pi_Y$ -integrable  $\gamma: \Omega_Y \rightarrow \mathbb{R}$  functions such that  $\exp(\psi - \beta - \gamma)$  is  $\pi_X \times \pi_Y$ -integrable.

Note that **(E1)** and **(E2)** are joint conditions on  $\pi_X$ ,  $\pi_Y$  and  $\psi$ . - The above conditions are also sufficient (cf. **A1**) for the existence of the wanted  $P \in \mathcal{P}_f$ .

**Theorem 2 (Existence):** For distributions  $\pi_X$  on  $\Omega_X$  and  $\pi_Y$  on  $\Omega_Y$  and a measurable function  $\psi$  on  $\Omega \times \Omega$  the following statements are equivalent:

- (a) There exists  $P \in \mathcal{P}_f$  with  $P^X = \pi_X$ ,  $P^Y = \pi_Y$  and  $\log OR(P) = \psi$ .
- (b) There exists  $P \in \mathcal{P}_f$  with  $\log OR(P) = \psi$ .
- (c)  $\psi$  satisfies the conditions **(OR1)**, **(E1)** and **(E2)**.

A few remarks on the conditions **(E1)** and **(E2)** are in order.

1. **(E1)** and **(E2)** hold for *bounded*  $\psi$ , which already covers an important range of applications in practice, e.g. for *compact*  $\Omega$  and *continuous*  $\psi$ .
2. The integrability of  $\exp(\psi - \beta - \gamma)$  in **(E2)** obviously holds if  $\psi \leq \beta + \gamma$ . And if even  $|\psi| \leq \beta + \gamma$  then **(E1)** follows too.
3. For *finite*  $\Omega_Y$  (or  $\Omega_X$ ) condition **(E1)** implies **(E2)** for  $\beta(x) = \sum_y |\psi(x,y)|$  and  $\gamma = 0$ .

Although by theorem 1 the probability measure  $P$  in (a) is uniquely determined, there is no *explicit* formula for  $P$  available. In the proof  $P$  is given by an  $I$ -projection, which can only be obtained as a limit in an iterative procedure, e.g. as in the proof of theorem 2.1 in Csiszár (1975). An alternative algorithm is the iterative proportional fitting procedure (IPFP). However the convergence of the IPFP is known only under stronger assumptions than (b) or (c), see Rüschemdorff (1995) for details. More recently we have shown in Osius (2000, 2004) that the IPFP converges to the wanted  $P$ , if **(E2)** is replaced by the stronger condition:

**(E2)\*** There exists  $\pi_X$ -integrable  $\beta: \Omega_X \rightarrow \mathbb{R}$  and  $\pi_Y$ -integrable  $\gamma: \Omega_Y \rightarrow \mathbb{R}$  functions such that  $\exp(\psi - \beta)$  and  $\exp(\psi - \gamma)$  are  $\pi_X \times \pi_Y$ -integrable.

### 3. Association Models

An association model for the joint distribution  $P$  of  $(X, Y)$  only restricts the odds ratio function of  $P$  but leaves the marginal distributions  $P^X$  and  $P^Y$  of  $X$  and  $Y$  arbitrary. To formulate such a model we assume that  $P$  has a density with respect to a fixed product measure  $\nu = \nu_X \times \nu_Y$  of  $\sigma$ -finite measures  $\nu_X$  resp.  $\nu_Y$  on  $\Omega_X$  resp.  $\Omega_Y$ . In other words we restrict  $P$  to the class

$$\mathcal{P}^{XY} = \{P \in \mathcal{P} \mid P \ll \nu \ll P\} \subset \mathcal{P}_{\ll},$$

which also restricts the marginal distribution  $P^X$  of  $X$  to the class

$$\mathcal{P}^X = \{ \pi_X \text{ probability measure on } \Omega_X \mid \pi_X \ll \nu_X \ll \pi_X \},$$

and the marginal  $P^Y$  to the corresponding class  $\mathcal{P}^Y$ . From now on all densities on  $\Omega$  resp.  $\Omega_X, \Omega_Y$  are taken with respect to the dominating measure  $\nu$  resp.  $\nu_X, \nu_Y$ .

We consider parametric association models indexed by a parameter vector  $\theta \in \mathbb{R}^S$ . For any  $\theta$  let  $\psi_\theta$  be a measurable function on  $\Omega$  satisfying the constraints **(OR1)**. The corresponding parametric odds ratio model restricts the log odds ratio function of  $P$  to the form  $\log OR(P) = \psi_\theta$  for some  $\theta$ . To guarantee for any  $\psi_\theta$  and any marginals  $\pi_X, \pi_Y$  the existence of a joint distribution  $P$  with  $\psi_\theta = \log OR(P)$  and these marginals, we assume the following bounding condition.

**(OR2)** *There exist non-negative measurable functions  $\psi_X$  on  $\Omega_X$  and  $\psi_Y$  on  $\Omega_Y$  with*

$$|\psi_\theta(x, y)| \leq [\psi_X(x) + \psi_Y(y)] \cdot \|\theta\| \quad \text{for all } \theta, x, y.$$

And second, we restrict the marginal distribution  $\pi_X$  on  $\Omega_X$  to the class

$$\mathcal{P}_f^X = \{ \pi_X \in \mathcal{P}^X \mid \psi_X \text{ is } \pi_X\text{-integrable} \}$$

and  $\pi_Y$  to the corresponding class  $\mathcal{P}_f^Y$ . Then for any  $\pi_X \in \mathcal{P}_f^X, \pi_Y \in \mathcal{P}_f^Y$  and  $\theta$  the condition (c) in theorem 2 holds (cf. remark 2), and hence there exists a unique  $P \in \mathcal{P}_f$  with  $P^X = \pi_X, P^Y = \pi_Y$  and  $\log OR(P) = \psi_\theta$ . We have thus specified a **parametric association model** for distributions  $P$  in  $\mathcal{P}_f^{XY} = \mathcal{P}^{XY} \cap \mathcal{P}_f$  by the requirements

**(PAM)**  $\log OR(P) \in \{ \psi_\theta \mid \theta \in \mathbb{R}^S \}, \quad P^X \in \mathcal{P}_f^X, \quad P^Y \in \mathcal{P}_f^Y.$

This is a semi-parametric model for the joint distribution  $P$  since the marginals are not parametrized but only slightly restricted by integrability conditions.

By (2.2) the density  $p(X=x, Y=y)$  of  $P \in \mathcal{P}_f^{XY}$  satisfying **(PAM)** can be

parametrized as

$$\log p(X=x, Y=y) = \alpha + \beta(x) + \gamma(y) + \psi_{\boldsymbol{\theta}}(x, y) \quad (3.1)$$

with  $\alpha \in \mathbb{R}$  and integrable functions  $\beta$  and  $\gamma$ . Identifiability may be achieved through the constraints  $\beta(x^\circ) = 0$  and  $\gamma(y^\circ) = 0$ , which will be assumed here. Then  $\int p \, d\nu = 1$  already determines the integration constant  $\alpha$  via

$$\alpha = -\log \int \exp(\beta + \gamma + \psi_{\boldsymbol{\theta}}) \, d\nu. \quad (3.2)$$

The density  $p(X=x)$  of the marginal distribution  $P^X$  is given by

$$\begin{aligned} \log p(X=x) &= \alpha + \beta(x) + \delta(x), \\ \delta(x) &= \log \left[ \int \exp(\gamma(y) + \psi_{\boldsymbol{\theta}}(x, y)) \, d\nu_Y(y) \right]. \end{aligned}$$

The conditional distribution of  $Y$  given  $X=x$  belongs to the class  $\mathcal{P}^Y$  and the conditional density  $p(Y=y | X=x)$  satisfies

$$\log p(Y=y | X=x) = \gamma(y) + \psi_{\boldsymbol{\theta}}(x, y) - \delta(x). \quad (3.3)$$

The integration constant  $\delta(x)$  can be removed by passing to the density ratio

$$\log \frac{p(Y=y | X=x)}{p(Y=y^\circ | X=x)} = \gamma(y) + \psi_{\boldsymbol{\theta}}(x, y). \quad (3.4)$$

Equation (3.4) may be viewed as a "regression model". Conversely, suppose such a regression model for  $P$  is specified by (3.4) with an *arbitrary* integrable function  $\gamma$  and the parametric family  $\psi_{\boldsymbol{\theta}}$ . Then  $\log OR(P) = \psi_{\boldsymbol{\theta}}$  and hence the model (3.4) is semi-parametric in the sense that it does not restrict the marginal distributions  $P^X$  and  $P^Y$  - provided they belong to the class  $\mathcal{P}_f^X$  resp.  $\mathcal{P}_f^Y$ . In the latter case the regression model (3.4) with arbitrary  $\gamma$  is in fact equivalent to the association model (**PAM**). Note that for finite  $\Omega_Y$  and counting measure  $\nu_Y$  the integrability condition imposed by  $P^Y \in \mathcal{P}_f^Y$  always holds.

An important class of parametric association models are *log-bilinear association models* with respect to measurable maps  $h_X: \Omega_X \rightarrow \mathbb{R}^{K_X}$  and  $h_Y: \Omega_Y \rightarrow \mathbb{R}^{K_Y}$  which will always be chosen here such that  $h_X(x^\circ) = \mathbf{0}$  and  $h_Y(y^\circ) = \mathbf{0}$ . The parameter is a  $K_X \times K_Y$ -matrix  $\boldsymbol{\theta}$  and the log odds ratio function is a bilinear function in the transformed variables  $h_X(x)$  and  $h_Y(y)$

$$\text{(LBA)} \quad \psi_{\boldsymbol{\theta}}(x, y) = h_X(x)^T \boldsymbol{\theta} h_Y(y) \quad \text{for all } x, y.$$

The bounding condition (**OR2**) holds for

$$\psi_X(x) = \|h_X(x)\|^2, \quad \psi_Y(y) = \|h_Y(y)\|^2,$$

since  $|h_X(x)^T \boldsymbol{\theta} h_Y(y)| \leq \|h_X(x)\| \cdot \|h_Y(y)\| \cdot \|\boldsymbol{\theta}\| \leq [\psi_X(x) + \psi_Y(y)] \cdot \|\boldsymbol{\theta}\|$ .

And the integrability condition in  $\mathcal{P}_f^X$  and  $\mathcal{P}_f^Y$  state that the second moments  $E(\|h_X(X)\|^2)$  and  $E(\|h_Y(Y)\|^2)$  are finite.

Suppose a submodel of **(LBA)** is specified by a linear restriction of the form  $\boldsymbol{\theta} = \mathbf{A}^T \boldsymbol{\theta}^* \mathbf{B}$  with given matrices  $\mathbf{A}$ ,  $\mathbf{B}$  and an unknown parameter matrix  $\boldsymbol{\theta}^*$ . Then the submodel has a log-bilinear association too - with respect to  $h_X^* = \mathbf{A} h_X$ ,  $h_Y^* = \mathbf{B} h_Y$  and  $\boldsymbol{\theta}^*$ .

The following examples reveal, that the association structure of some widely used regression models is log-bilinear.

### *Example 1: Generalized Linear Models*

Let  $Y$  be a univariate random variable and  $X$  a random vector with values in  $\mathbb{R}^R$ . Suppose that the conditional density of  $Y$  given  $X=x$  belongs to the exponential family (cp. McCullagh and Nelder, 1989)

$$p(Y=y|X=x) = \exp \{ a(\phi)^{-1} [y \cdot \tau(x) - b(\tau(x))] + c(y, \phi) \}$$

with suitable functions  $a$ ,  $b$ ,  $c$ ,  $\tau$  and a (dispersion) parameter  $\phi$ . Then the log odds ratio function has the form

$$\psi(x, y) = a(\phi)^{-1} \cdot [\tau(x) - \tau(x^\circ)] \cdot [y - y^\circ].$$

Now  $\tau(x)$  is a strictly monotone function of the conditional expectation  $\mu(x) = E(Y|X=x)$ , namely  $\tau(x) = \lambda(\mu(x))$  where  $\lambda^{-1} = b' > 0$  is the derivative of  $b$ . And the conditional variance is  $\text{var}(Y|X=x) = b''(\tau(x)) \cdot a(\phi)$ . A **generalized linear model** specifies the conditional expectation via a link function  $g$

$$\text{(GLM)} \quad g(\mu(x)) = \alpha + \mathbf{z}^T \boldsymbol{\beta},$$

where  $\mathbf{z} = h_X(x) \in \mathbb{R}^S$  is a known vector of *formal* covariates (obtained from  $x$  by a given function  $h_X$ ) and  $\alpha \in \mathbb{R}$ ,  $\boldsymbol{\beta} \in \mathbb{R}^S$  are unknown parameters. The model is sometimes formulated with  $x$  instead of  $\mathbf{z}$ , but the distinction between the observed  $x$  and the formal covariate  $\mathbf{z}$  is useful. For example,  $\mathbf{z}$  may contain powers of continuous components of  $x$  as well as indicator variables for levels of discrete components of  $x$ .

For  $G = \lambda \circ g^{-1}$  and  $h_X(x^\circ) = \mathbf{0}$  the log odds ratio function under **(GLM)** becomes

$$\psi(x, y) = a(\phi)^{-1} \cdot [G(\alpha + \mathbf{z}^T \boldsymbol{\beta}) - G(\alpha)] \cdot [y - y^\circ].$$

If the *canonical link* is chosen, i.e.  $g = \lambda^{-1} = b'$ , then  $G$  is the identity and

$$\psi(x, y) = \mathbf{z}^T \boldsymbol{\theta} [y - y^\circ] \quad (3.5)$$

is of the form **(LBA)** with  $h_Y(y) = y - y^\circ$  and parameter  $\boldsymbol{\theta} = a(\phi)^{-1} \boldsymbol{\beta}$ . Note that the intercept  $\alpha$  from **(GLM)** is no longer present in (3.5). Starting with the log-bilinear association model (3.5) rather than with the generalized linear model **(GLM)** weakens the distributional assumption while still including the regression parameter  $\boldsymbol{\beta}$  up to a positive constant  $a(\phi)^{-1}$ . In particular a linear hypothesis  $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$  with a given matrix  $\mathbf{C}$  is equivalent to  $\mathbf{C}\boldsymbol{\theta} = \mathbf{0}$ , and a for a vector  $\mathbf{c} \in \mathbb{R}^S$  a one-sided hypothesis  $\mathbf{c}^T \boldsymbol{\beta} > 0$  is equivalent to  $\mathbf{c}^T \boldsymbol{\theta} > 0$ .

Let us now look at the most popular generalized linear models with canonical link. First, normal conditional distributions  $N(\mu(x), \sigma^2)$  of  $Y$  yield the classical linear model with  $a(\phi) = \sigma^2$ . Second, binomial conditional distributions  $B(\mu(x), 1)$  lead to (univariate) logistic regressions models. And finally, for poisson conditional distributions  $Pois(\mu(x))$  log-linear models are obtained. Note that for the latter two models we have  $a(\phi) = 1$  and hence  $\boldsymbol{\theta} = \boldsymbol{\beta}$ .  $\square$

### *Example 2: Multivariate Linear Logistic Regression*

Extending the univariate logistic regression above to the multivariate case, suppose  $Y$  (representing e.g. a disease status) takes values in  $\Omega_Y = \{0, 1, \dots, K\}$ ,  $K \geq 1$ , and  $X$  is a vector of *observed* covariates with values in  $\mathbb{R}^R$ . Then  $\mathcal{L}(Y | X = x)$  has a multinomial distribution  $M_{K+1}(1, \pi(x))$  with  $K+1$  classes and probabilities  $\pi_k(x) = P(Y = k | X = x) > 0$ . Using the multivariate logistic transformation  $\text{logit } \pi_k(x) = \log(\pi_k(x) / \pi_0(x))$  of  $\pi(x)$ , the *linear logistic regression model* is of the form

$$\text{(LLR)} \quad \text{logit } \pi_k(x) = \gamma_k + \mathbf{z}^T \boldsymbol{\theta}_k, \quad k = 1, \dots, K,$$

where  $\mathbf{z} = h_X(x) \in \mathbb{R}^S$  is as above a vector of *formal* covariates and  $\gamma_k \in \mathbb{R}$ ,  $\boldsymbol{\theta}_k \in \mathbb{R}^S$  are unknown parameters. Choosing  $y^\circ = 0$ , the log odds ratio function is

$$\text{(LLR)'} \quad \psi(x, k) = h_X(x)^T \boldsymbol{\theta}_k = h_X(x)^T \boldsymbol{\theta} h_Y(k),$$

where  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$  is an  $S \times K$  parameter matrix, and the function  $h_Y: \Omega_Y \rightarrow \mathbb{R}^K$  maps  $k > 0$  to the  $k$ -th unit vector  $\mathbf{e}_k$  and  $h_Y(0) = \mathbf{0}$ . Hence the linear logistic regression model is equivalent to the log-bilinear association model **(LLR)'** - provided  $E(\|h_X(X)\|^2)$  is finite. As mentioned above, this also holds for submodels given by linear constraints, e.g.  $\boldsymbol{\theta}_k = \boldsymbol{\theta}^*$  for all  $k > 0$ .

Replacing  $\mathbf{z}^T \boldsymbol{\theta}_k$  by an arbitrary function  $g(\mathbf{z}, \boldsymbol{\theta}_k)$  leads to the general logistic regression model

$$\text{(GLR)} \quad \text{logit } \pi_k(x) = \gamma_k + g(\mathbf{z}, \boldsymbol{\theta}_k), \quad k = 1, \dots, K,$$

which is equivalent to the log odds ratio model

$$\text{(GLR)'} \quad \psi(x, k) = g(h_X(x), \boldsymbol{\theta}_k) = g(h_X(x), \boldsymbol{\theta} h_Y(k)). \quad \square$$

### *Example 3: Multivariate Linear Regression*

Let  $Y$  and  $X$  be random vectors taking values in  $\mathbb{R}^K$  resp.  $\mathbb{R}^R$ , and suppose that the conditional distribution of  $Y$  given  $X$  is multivariate normal

$$\mathcal{L}(Y|X=x) = N_K(\mu_Y(x), \boldsymbol{\Sigma}), \quad (3.6)$$

such that the conditional covariance matrix  $\boldsymbol{\Sigma}$  is non-singular and does not depend on  $x$ . From the conditional log-density

$$\log p(Y=y|X=x) = -\frac{1}{2} \left[ \log[(2\pi)^K \det(\boldsymbol{\Sigma})] + [y - \mu_Y(x)]^T \boldsymbol{\Sigma}^{-1} [y - \mu_Y(x)] \right]$$

the log odds ratio function with respect to  $y^\circ = \mathbf{0}$  is

$$\psi(x, y) = [\mu_Y(x) - \mu_Y(x^\circ)]^T \boldsymbol{\Sigma}^{-1} y.$$

Then the multivariate linear regression model

$$\text{(MLR)} \quad \mu_Y(x) = \boldsymbol{\alpha} + \mathbf{z}^T \boldsymbol{\beta}$$

with covariates  $\mathbf{z} = h_X(x) \in \mathbb{R}^S$  has a log-bilinear association

$$\psi(x, y) = h_X(x)^T \boldsymbol{\theta} y \quad (3.7)$$

with parameter matrix  $\boldsymbol{\theta} = \boldsymbol{\beta} \boldsymbol{\Sigma}^{-1}$  - assuming  $h_X(x^\circ) = \mathbf{0}$ . Note that the regression parameter  $\boldsymbol{\beta}$  may only be recovered from  $\boldsymbol{\theta}$  if the covariance matrix  $\boldsymbol{\Sigma}$  is known. However any linear hypothesis  $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$  is equivalent to the corresponding hypothesis  $\mathbf{C}\boldsymbol{\theta} = \mathbf{0}$ , and the latter may be tested using the semiparametric association model (3.7) instead of the regression model (MLR) with the distributional assumption (3.6).

If - instead of (3.6) - we allow the conditional covariance matrix to depend on  $x$ , i.e.

$$\mathcal{L}(Y|X=x) = N_K(\mu_Y(x), \boldsymbol{\Sigma}(x)),$$

then the model (MLR) leads to the log odds ratio function

$$\psi(x, y) = h_X(x)^T \boldsymbol{\beta} \boldsymbol{\Sigma}^{-1}(x) y,$$

which is not bilinear. □

The above examples reveal that important regression models may be viewed more generally as log-bilinear association models by ignoring the distributional

assumption for the conditional distribution of  $Y$  given  $X$ . Although log-bilinear association is a natural candidate we also consider the more general association model

$$(AM) \quad \psi_{\boldsymbol{\theta}}(x, y) = G(h_X(x), h_Y(y), \boldsymbol{\theta}) \quad \text{for all } x, y,$$

given by a fixed function  $G$  with  $G(\mathbf{0}, -, -) = G(-, \mathbf{0}, -) = 0$ . We assume throughout, that the function  $G$  satisfies the following regularity condition (although some results also hold under weaker assumptions):

$$(R1) \quad G(\mathbf{z}, \mathbf{v}, \boldsymbol{\theta}) \text{ is thrice continuously differentiable with respect to } \boldsymbol{\theta} \text{ for all } \mathbf{z} \in h_X[\Omega_X], \mathbf{v} \in h_Y[\Omega_Y] \text{ and the derivatives are continuous in } \mathbf{z} \text{ and } \mathbf{v}.$$

Further properties of the functions  $h_X$ ,  $h_Y$  and  $G$  will be assumed later, cf. (R2)' and (MC).

## 4. Estimation

For a given data set  $(x_i, y_i)$  with  $i = 1, \dots, n$  we want to estimate the association parameter  $\boldsymbol{\theta}$  of a parametric association model. Three important sampling schemes will be discussed: unconditional sampling from the joint distribution of  $(X, Y)$  and conditional sampling of  $Y$  given  $X$  or vice versa. We first address the question of how to *estimate* the association parameter  $\boldsymbol{\theta}$  and look at the properties of the estimate  $\hat{\boldsymbol{\theta}}$  later. It will turn out that an appropriate maximum likelihood estimate  $\hat{\boldsymbol{\theta}}$  under *any* of these three sampling schemes may be obtained as a solution the *same* estimating equation, thus making the estimation process invariant with respect to sampling.

### 4.1 Unconditional Sampling

For unconditional sampling the data set  $(x_i, y_i)$  is an independent sample from the joint distribution of  $(X, Y)$ . Suppose there are  $J+1$  different  $x$ -values and  $K+1$  different  $y$ -values observed and denote the corresponding subsets of  $\Omega_X$  and  $\Omega_Y$  by

$$\Omega_X^* = \{x_{(0)}, \dots, x_{(J)}\} = \{x_i \mid i = 1, \dots, n\}, \quad \Omega_Y^* = \{y_{(0)}, \dots, y_{(K)}\} = \{y_i \mid i = 1, \dots, n\}.$$

If  $r_{jk} = \#\{i = 1, \dots, n \mid x_i = x_{(j)}, y_i = y_{(k)}\}$  is the observed frequency of the pair  $(x_{(j)}, y_{(k)})$ , then the (unconditional) likelihood may be written

$$L_{XY} = \prod_{j=0}^J \prod_{k=0}^K p(X = x_{(j)}, Y = y_{(k)})^{r_{jk}} = L_{X|Y} \cdot L_Y$$

with the conditional and marginal likelihood

$$L_{X|Y} = \prod_{k=0}^K \prod_{j=0}^J p(X=x_{(j)} | Y=y_{(k)})^{r_{jk}}, \quad L_Y = \prod_{k=0}^K p(Y=y_{(k)})^{r_{+k}}, \quad (4.1)$$

where the subscript "+" indicates summation over the replaced index. Maximization of  $L_{XY}$  is equivalent to separate maximization of  $L_{X|Y}$  and  $L_Y$  because the latter two have no common parameters. Since the parametric association model does not restrict the marginal density of  $Y$  no direct maximization of  $L_Y$  is possible. However the usual nonparametric estimate for the marginal distribution  $P^Y$  is the empirical distribution of the  $y$ -values and hence the estimated density with respect to counting measure  $\nu_Y^*$  on  $\Omega_Y^*$  is given by

$$\hat{p}(Y=y_{(k)}) = \frac{1}{n} r_{+k} \quad \text{for } k = 0, \dots, K. \quad (4.2)$$

If we restrict the marginal distribution  $P^Y$  to the class  $\mathcal{P}_Y^*$  of all distributions with finite support  $\Omega_Y^*$ , then its  $\nu_Y^*$ -density satisfies  $\sum_k p(Y=y_{(k)}) = 1$ . Then  $L_Y$  is a multinomial likelihood which achieves its maximum for (4.2). Hence, for the purpose of estimating  $\theta$ , we may restrict the marginal  $P^Y$  to  $\mathcal{P}_Y^*$ .

Now we split the unconditional likelihood as  $L_{XY} = L_{Y|X} \cdot L_X$  with

$$L_{Y|X} = \prod_{j=0}^J \prod_{k=0}^K p(Y=y_{(k)} | X=x_{(j)})^{r_{jk}}, \quad L_X = \prod_{j=0}^J p(X=x_{(j)})^{r_{j+}}.$$

Repeating the above argument with  $Y$  replaced by  $X$  shows, that we may *additionally* restrict the marginal  $P^X$  to the class  $\mathcal{P}_X^*$  of all distributions with finite support  $\Omega_X^*$ . Under these restrictions for *both* marginals  $P^X$  and  $P^Y$  the likelihood  $L_{XY}$  is the multinomial likelihood for the observed  $(J+1) \times (K+1)$ -contingency table  $(r_{jk})$  since

$$\sum_j \sum_k p(X=x_{(j)}, Y=y_{(k)}) = \sum_j p(X=x_{(j)}) = 1.$$

Hence, estimation of  $\theta$  is reduced to estimating  $\theta$  in a multinomial model whose probabilities  $p_{jk} = p(X=x_{(j)}, Y=y_{(k)})$  satisfy the log odds ratio model

$$\log(p_{jk} p_{00} / p_{j0} p_{0k}) = \psi_{\theta}(x_{(j)}, y_{(k)}) =: \psi_{jk}(\theta) \quad \text{for all } j, \text{ and } k$$

with respect to the reference values  $x^\circ = x_{(0)}$  and  $y^\circ = y_{(0)}$ . The parametrization (3.1) of the density now involves only a *finite* number of parameters

$$\log p_{jk} = \beta_j + \gamma_k + \psi_{jk}(\theta) - \log(\sum_j \sum_k \exp[\beta_j + \gamma_k + \psi_{jk}(\theta)]), \quad (4.3)$$

namely  $\beta_j = \beta(x_{(j)})$ ,  $\gamma_k = \gamma(y_{(k)})$  and  $\theta$  with  $\beta_0 = \gamma_0 = 0$ . Instead of maximizing  $L_{XY}$  it is typically preferable to maximize the logarithm of either  $L_{Y|X}$  or  $L_{X|Y}$  using the

parametrization of the conditional probabilities  $p_{k|j} = p_{jk}/p_{j+}$  resp.  $p_{j|k} = p_{jk}/p_{+k}$  given by (3.3) resp. its dual

$$\log p_{k|j} = \gamma_k + \psi_{jk}(\boldsymbol{\theta}) - \delta_j \quad \text{resp.} \quad \log p_{j|k} = \beta_j + \psi_{jk}(\boldsymbol{\theta}) - \varepsilon_k,$$

where the parameters  $\delta_j$  resp.  $\varepsilon_k$  are determined by the remaining ones.

## 4.2 Conditional Sampling

When sampling is conditional on fixed values for  $Y$  taken from a set  $\{y_{(0)}, \dots, y_{(K)}\}$  then the data set  $(x_i, y_i)$  with  $i=1, \dots, n$  is partitioned into  $K+1$  independent subsamples given by  $I_k = \{i \mid y_i = y_{(k)}\}$ , such that each subsample  $(x_i)$  with  $i \in I_k$  is an independent sample from the conditional distribution  $\mathcal{L}(X \mid Y = y_{(k)})$ . Keeping the notation from unconditional sampling above, the appropriate (conditional) likelihood is now  $L_{X|Y}$  in (4.1). Using the empirical distribution (4.2) on  $\Omega_Y^*$  we define a joint distribution  $P^*$  on  $\Omega_X \times \Omega_Y^*$  via its density  $p^*$  with respect to  $\nu_X \times \nu_Y^*$  by

$$p^*(X = x, Y = y_{(k)}) = p(X = x \mid Y = y_{(k)}) \cdot \frac{1}{n} r_{+k} \quad \text{for all } x \text{ and } k.$$

If the conditional sampling had been from  $P^*$  instead of  $P$  then the resulting conditional likelihood  $L_{X|Y}$  would have been the same. Furthermore, instead of maximizing  $L_{X|Y}$  we can equivalently maximize the product  $L_{XY} = L_{X|Y} \cdot L_Y$  subject to the restriction  $P^Y \in \mathcal{P}_Y^*$ . This leads us back to the unconditional likelihood  $L_{XY}$  which - as observed above - can be maximized with respect to  $\boldsymbol{\theta}$  by restricting the marginal  $P^X$  to  $\mathcal{P}_X^*$ . Hence under conditional sampling instead of maximizing the conditional likelihood  $L_{X|Y}$  we may equivalently maximize the unconditional likelihood  $L_{XY}$ , or even the "reverse" conditional likelihood  $L_{Y|X}$ . Interchanging  $X$  with  $Y$  reveals, that this also holds if sampling is conditional on fixed values for  $X$  taken from a set  $\{x_{(0)}, \dots, x_{(J)}\}$ .

## 4.3 Log-bilinear Association

For the log-bilinear association model (**LBA**), the odds ratios may be written as

$$\psi_{jk}(\boldsymbol{\theta}) = \mathbf{z}_j^T \boldsymbol{\theta} \mathbf{v}_k, \quad \text{with} \quad \mathbf{z}_j = h_X(x_{(j)}) \in \mathbb{R}^{K_X}, \quad \mathbf{v}_k = h_Y(y_{(k)}) \in \mathbb{R}^{K_Y},$$

or in matrix notation

$$\boldsymbol{\psi}(\boldsymbol{\theta}) = \mathbf{Z} \boldsymbol{\theta} \mathbf{V}^T \in \mathbb{R}^{J \times K}, \quad \mathbf{Z} = (\mathbf{z}_j) \in \mathbb{R}^{J \times K_X}, \quad \mathbf{V} = (\mathbf{v}_k) \in \mathbb{R}^{K \times K_Y}.$$

Then (4.3) reduces to a log-linear model for the probabilities  $p_{jk}$

$$\log p_{jk} = \alpha + \beta_j + \gamma_k + \mathbf{z}_j^T \boldsymbol{\theta} \mathbf{v}_k \quad (4.4)$$

induced by the "covariates"  $\mathbf{z}_j$  and  $\mathbf{v}_k$ . Results by Haberman (1974) on the existence

and uniqueness of maximum likelihood estimates in log-linear models may be applied to the maximization of the unconditional likelihood  $L_{XY}$  for the model (4.4) and the marginal restrictions  $P^X \in \mathcal{P}_X^*$  and  $P^Y \in \mathcal{P}_Y^*$ . In particular the estimate  $\hat{\mathbf{p}} = (\hat{p}_{jk})$  is unique (provided it exists) and hence the estimate  $\hat{\boldsymbol{\theta}}$  is unique too, provided the parameter  $\boldsymbol{\theta}$  is identifiable.

For sampling conditional on  $Y$ , the values  $y_{(k)}$  should be chosen such that the rank condition below holds.

**(Rk)<sub>LBA</sub>** The  $K_Y \times K$ -matrix  $\mathbf{V}^T = (\mathbf{v}_1, \dots, \mathbf{v}_K)$  has rank  $K_Y$ .

This condition will be assumed whenever the log-bilinear association model is used. Then a convenient reparametrization is available

$$\psi_{jk}(\boldsymbol{\theta}) = \mathbf{z}_j^T \tilde{\boldsymbol{\theta}}_k, \quad \tilde{\boldsymbol{\theta}}_k = \boldsymbol{\theta} \mathbf{v}_k \in \mathbb{R}^{K_X}, \quad (4.6)$$

with a  $K_X \times K$  parameter-matrix  $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \dots, \tilde{\boldsymbol{\theta}}_K) = \boldsymbol{\theta} \mathbf{V}^T$ . - For the observed  $x$ -values the matrix  $\mathbf{Z}$  of covariates will typically have rank  $K_X$ . Then  $\boldsymbol{\theta}$  resp.  $\tilde{\boldsymbol{\theta}}$  is uniquely determined by  $\boldsymbol{\psi}(\boldsymbol{\theta}) = \mathbf{Z} \boldsymbol{\theta} \mathbf{V}^T = \mathbf{Z} \tilde{\boldsymbol{\theta}}$  and hence identifiable. More generally, identifiability of  $\boldsymbol{\theta}$  is guaranteed by assumption **(C3)** in section 6.

## 5. Conditional Likelihood

As seen above the maximum likelihood estimate  $\hat{\boldsymbol{\theta}}$  of the association parameter  $\boldsymbol{\theta}$  may be obtained by maximizing either the unconditional likelihood  $L_{XY}$  or any of the two conditional likelihoods. From a numerical point of view it is preferable to use the conditional likelihood  $L_{X|Y}$  resp.  $L_{Y|X}$  with fewer nuisance parameters  $\beta_j$  resp.  $\gamma_k$  - even if this is not the likelihood corresponding to the sampling scheme. We will now investigate this situation in detail and derive properties of the likelihood  $L_{Y|X}$  conditional on  $X$  under sampling conditional on  $Y$  (dual results are obtainable by interchanging  $X$  with  $Y$ ). An important example for *finite*  $\Omega_Y$  are case-control studies which may be analyzed using the likelihood of a cohort study and the (multivariate) logistic regression model, cf. Prentice and Pyke (1979). We want to extend their results extends to an *infinite* domain  $\Omega_Y$  as well as to the more general *association model (AM)*.

Instead of a fixed data set  $(x_i, y_i)$  considered in section 5, we now look at the underlying random variables. It is convenient to represent the sample as a vector of independent random variables  $\mathbf{X} = (X_{ki})$  indexed by  $k = 0, \dots, K$  and  $i = 1, \dots, n_k$ . Omitting the parenthesis in  $y_{(k)}$  and  $x_{(j)}$  here, each  $X_{ki}$  is distributed as

$X_k \sim \mathcal{L}(X|Y=y_k)$ . As above  $r_{jk}$  denotes the frequency of  $(x_j, y_k)$  in the sample  $(x_{ki}, y_k)$  and the empirical distribution on  $\Omega_Y^*$  is given by the proportions  $\bar{r}_k = n_k/n$  where  $n = n_+$  is the total sample size. The distribution  $P^*$  on  $\Omega_X \times \Omega_Y^*$  has the joint density

$$p^*(X=x, Y=y_k) = \bar{r}_k \cdot p(X=x|Y=y_k) \quad \text{for all } x, k,$$

and the marginal density for  $Y$  is

$$p^*(Y=y_k) = \bar{r}_k \quad \text{for } k = 0, \dots, K.$$

The marginal and conditional densities for  $X$  under  $P^*$  are

$$\begin{aligned} p^*(X=x) &= \sum_{k=0}^K \bar{r}_k \cdot p(X=x|Y=y_k), \\ p_k^*(x) := p^*(Y=y_k|X=x) &= \frac{\bar{r}_k \cdot p(X=x|Y=y_k)}{p^*(X=x)}. \end{aligned} \quad (5.1)$$

The conditional density may be parametrized as in (3.3)

$$\log p_k^*(x) = \gamma_k^* + \psi_{\boldsymbol{\theta}}(x, y_k) - \delta^*(x)$$

with nuisance parameters  $\gamma_k^* = \gamma^*(y_k)$  and

$$\delta^*(x) = \log \left[ \sum_{l=0}^K \exp(\gamma_l^* + \psi_{\boldsymbol{\theta}}(x, y_l)) \right].$$

Hence

$$p_k^*(x) = \frac{\exp[\gamma_k^* + \psi_{\boldsymbol{\theta}}(x, y_k)]}{\sum_l \exp[\gamma_l^* + \psi_{\boldsymbol{\theta}}(x, y_l)]}. \quad (5.2)$$

Choosing the reference value  $y^\circ = y_0$  we have  $\gamma_0^* = 0$ , and the nuisance parameter is  $\boldsymbol{\gamma}^* = (\gamma_1^*, \dots, \gamma_K^*) \in \mathbb{R}^K$ . Finally, the logarithm of the conditional likelihood  $L_{Y|X}$  may be written in terms of the compound parameter vector  $\boldsymbol{\lambda} := (\boldsymbol{\theta}, \boldsymbol{\gamma}^*) \in \mathbb{R}^{S+K}$

$$\begin{aligned} \ell(\boldsymbol{\lambda}) &:= \log L_{Y|X} = \sum_{k=0}^K \sum_{i=1}^{n_k} \log p_k^*(X_{ki}) \quad \text{with} \\ \log p_k^*(X_{ki}) &= \gamma_k^* + \psi_{\boldsymbol{\theta}}(X_{ki}, y_k) - \log \left[ \sum_{l=0}^K \exp(\gamma_l^* + \psi_{\boldsymbol{\theta}}(X_{ki}, y_l)) \right]. \end{aligned} \quad (5.3)$$

Notice that  $\ell(\boldsymbol{\lambda})$  is the log-likelihood of a multivariate logistic regression model

$$\text{logit } p_k^*(x) = \gamma_k^* + \psi_{\boldsymbol{\theta}}(x, y_k) \quad k = 1, \dots, K, \quad (5.4)$$

which however is *nonlinear* in general. The estimate  $\hat{\boldsymbol{\lambda}}$  maximizing  $\ell(\boldsymbol{\lambda})$  satisfies

$$D_{\boldsymbol{\lambda}} \ell(\boldsymbol{\lambda}) = \sum_{k=0}^K \sum_{i=1}^{n_k} D_{\boldsymbol{\lambda}} \log p_k^*(X_{ki}) = \mathbf{0}, \quad (5.5)$$

where  $D_{\lambda}$  denotes the differential operator with respect to  $\lambda$ . The basic stochastic properties of the solution of the estimating equation (5.5) depend on the moments of the estimating function  $D_{\lambda}\ell(\lambda)$  and its derivative. The first important property (proved in **A2**) is that its expectation is zero - which is not obvious since  $\ell(\lambda)$  is *not* the log-likelihood for the underlying sampling:

$$E[D_{\lambda}\ell(\lambda)] = \sum_k n_k \cdot E[D_{\lambda} \log p_k^*(X_k)] = \mathbf{0}. \quad (5.6)$$

Next, the components of the covariance matrix  $\Sigma(\lambda) := \text{Cov}(D_{\lambda}\ell(\lambda))$  are given by

$$\Sigma_{st}(\lambda) = \sum_k n_k \cdot \text{Cov}(D_{\lambda_s} \log p_k^*(X_k), D_{\lambda_t} \log p_k^*(X_k)), \quad (5.7)$$

and for the partial second derivatives we get

$$J_{st}(\lambda) := -D_{\lambda_s \lambda_t}^2 \ell(\lambda) = -\sum_k \sum_i D_{\lambda_s \lambda_t}^2 \log p_k^*(X_{ki}) \quad (5.8)$$

with expectation (cf. **A2**)

$$I_{st}(\lambda) := E(J_{st}(\lambda)) = \sum_k n_k \cdot E(D_{\lambda_s} \log p_k^*(X_k) \cdot D_{\lambda_t} \log p_k^*(X_k)). \quad (5.9)$$

Since  $\ell(\lambda)$  is not the log-likelihood for sampling conditional on  $X$ , the matrices  $\Sigma(\lambda)$  and  $\mathbf{I}(\lambda)$  need not be equal, but from (5.7) their difference is

$$I_{st}(\lambda) - \Sigma_{st}(\lambda) = \sum_k n_k \cdot E(D_{\lambda_s} \log p_k^*(X_k)) \cdot E(D_{\lambda_t} \log p_k^*(X_k)). \quad (5.10)$$

An essential assumption needed later is

**(R2)**  $\Sigma(\lambda) = \text{Cov}(D_{\lambda}\ell(\lambda))$  is positive definite for all  $\lambda$ .

Two equivalent formulations (cf. **A2**) are

**(R2)'**  $\mathbf{I}(\lambda)$  is positive definite for all  $\lambda$ .

**(R2)''** For all  $\theta$ , all  $\mathbf{s} \in \mathbb{R}^S$  and  $c_1, \dots, c_K \in \mathbb{R}$ :

$$D_{\theta} \psi_{\theta}(X, y_k) \cdot \mathbf{s} = c_k \quad \text{for } k = 1, \dots, K \quad \text{almost surely} \quad \Rightarrow \quad \mathbf{s} = \mathbf{0}.$$

In the last formulation - which does not include the nuisance parameter  $\gamma^*$  - we can replace  $X$  by  $X_k$  since their distributions belong to  $\mathcal{P}^X$  and hence dominate each other. From now on, condition **(R2)** will be assumed throughout without further notice.

Using the obvious block notation for an  $(S+K) \times (S+K)$  matrix, say

$$\Sigma = \begin{bmatrix} \Sigma_{\theta\theta} & \Sigma_{\theta\gamma} \\ \Sigma_{\gamma\theta} & \Sigma_{\gamma\gamma} \end{bmatrix},$$

the following fundamental representation can be derived (cf. **A2**) by adopting the method in Prentice and Pyke (1979).

**Theorem 3.** For any  $\lambda$

$$(a) \quad \mathbf{I}(\lambda) - \Sigma(\lambda) = \mathbf{I}(\lambda) \cdot \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix} \cdot \mathbf{I}(\lambda),$$

where the  $K \times K$ -matrix  $\mathbf{W}$  is the sum of the diagonal  $\text{diag}(n_1^{-1}, \dots, n_K^{-1})$  and the constant matrix  $(n_0^{-1})$ , i.e.  $W_{kl} = \Delta_{kl} n_k^{-1} + n_0^{-1}$  with the Kronecker's  $\Delta$ .

$$(b) \quad [\mathbf{I}^{-1}(\lambda)]_{\theta\theta} = [\mathbf{I}^{-1}(\lambda) \cdot \Sigma(\lambda) \cdot \mathbf{I}^{-1}(\lambda)]_{\theta\theta}.$$

The matrix in (b) will later turn out to be the asymptotic covariance matrix of the estimate  $\hat{\theta}$  (under appropriate conditions).

**Log-bilinear association:** Using the parametrization (4.6) and writing  $\theta$  instead of  $\tilde{\theta}$  the model states

$$\psi_{\theta}(x, y_k) = \mathbf{z}^T \theta_k, \quad \text{with} \quad \mathbf{z} = h_X(x), \quad \theta = (\theta_1, \dots, \theta_K) \in \mathbb{R}^{K_X \times K}, \quad (5.11)$$

and is equivalent to the linear logistic regression model given by (5.2), i.e.

$$\text{logit } p_k^*(x) = \gamma_k^* + \mathbf{z}^T \theta_k, \quad k = 1, \dots, K.$$

Condition **(R2)''** holds if  $h_X(X)$  is not concentrated on a hyperplane of  $\mathbb{R}^{K_X}$ , i.e. if the following condition is met (cf. **A2**):

$$(R2)_{\text{LBA}} \quad \text{For all } \mathbf{s} \in \mathbb{R}^{K_X}: \quad \mathbf{s}^T h_X(X) \text{ is constant almost surely} \quad \Rightarrow \quad \mathbf{s} = \mathbf{0}.$$

## 6. Asymptotics and Consistency

We now turn to the asymptotic properties of the estimate  $\hat{\lambda} = (\hat{\theta}, \hat{\gamma}^*)$  in the parametric association model **(AM)**. Our asymptotic approach assumes that set  $\Omega_Y^* = \{y_0, \dots, y_K\}$  of conditional values will remain *fixed* while all subsample sizes  $n_k$  tend to infinity with *fixed* sample ratios  $\bar{r}_k = n_k/n > 0$  for all  $n$  and  $k$ . Hence the nuisance parameter  $\gamma^*$ , the distribution  $P^*$  and its conditional densities  $p_k^*(x)$  do not vary with  $n$ . The true parameter will be denoted by  $\lambda^\circ = (\theta^\circ, \gamma^\circ)$ , and the notations  $E, P$  etc. refer to expectations, probabilities etc. with respect to  $\lambda^\circ$ .

The conditional log-likelihood  $\ell^{(n)}(\lambda)$  - the additional index  $n$  is supplied if

necessary - need not have a unique maximizing argument  $\hat{\lambda}$  for *every* sample. Concerning uniqueness, the strong law of large numbers (applied seperately to each subsample  $k$ ) yields for the matrix  $\mathbf{J}^{(n)}(\lambda) = -D_{\lambda\lambda}^2 \ell^{(n)}(\lambda)$  given in (5.8)

$$\frac{1}{n} \mathbf{J}^{(n)}(\lambda) \xrightarrow{n \rightarrow \infty} \bar{\mathbf{I}}(\lambda) := \sum_{k=0}^K \bar{r}_k \cdot E(-D_{\lambda\lambda}^2 \log p_k^*(X_k)) \quad \text{almost surely.} \quad (6.1)$$

The matrix  $\bar{\mathbf{I}}(\lambda) = \frac{1}{n} \mathbf{I}(\lambda)$  is positive-definite by **(R2)'** which implies

$$D_{\lambda\lambda}^2 \ell^{(n)}(\lambda) = -\mathbf{J}^{(n)}(\lambda) \text{ is negativ-definite for almost all } n \text{ almost surely.}$$

Hence - almost surely - the function  $\ell^{(n)}(\lambda)$  is strictly concave for almost all  $n$ , which implies that  $D_{\lambda} \ell^{(n)}(\lambda) = \mathbf{0}$  has at most one solution  $\hat{\lambda}$  which also maximizes  $\ell^{(n)}(\lambda)$ . Since the unique existence of a maximizing argument  $\hat{\lambda}$  of  $\ell^{(n)}(\lambda)$  is not guaranteed for *every*  $n$ , we consider *any* sequence of (measurable) functions  $\hat{\lambda}^{(n)}$  as estimators if the estimating condition is met:

$$\text{(C1)} \quad \text{If } \ell^{(n)}(\lambda) \text{ has a maximizing argument } \lambda, \text{ then:} \quad \ell^{(n)}(\hat{\lambda}^{(n)}) = \underset{\lambda}{\text{Max}} \ell^{(n)}(\lambda).$$

To establish the consistency of such a sequence of estimate  $\hat{\lambda}^{(n)}$  we assume an integrability condition

$$\text{(C2)} \quad E\{\psi_X(X_k)\} < \infty \text{ for all } k = 0, \dots, K,$$

and an identifiability condition

$$\text{(C3)} \quad \psi_{\theta_1}(X, y_k) = \psi_{\theta_2}(X, y_k) \quad \text{for } k = 1, \dots, K \text{ almost surely} \quad \Rightarrow \quad \theta_1 = \theta_2.$$

As in **(R2)''**, we can equivalently replace  $X$  by  $X_k$  in **(C3)**. From the above conditions we derive in **A3** the asymptotic (unique) existence and the consistency of the estimator as follows.

**Theorem 4 (Consistency).** Under **(C1)**, **(C2)** and **(C3)** the following properties hold almost surely

- (a) For almost all  $n$  there exists a unique  $\lambda$  maximizing  $\ell^{(n)}(\lambda)$ , namely  $\hat{\lambda}^{(n)}$ .
- (b) For almost all  $n$  there exists a unique solution  $\lambda$  of  $D_{\lambda} \ell^{(n)}(\lambda) = \mathbf{0}$ , namely  $\hat{\lambda}^{(n)}$ .
- (c)  $\hat{\lambda}^{(n)} = (\hat{\theta}^{(n)}, \gamma^{*(n)}) \xrightarrow{n \rightarrow \infty} \lambda^{\circ} = (\theta^{\circ}, \gamma^{\circ})$ .

**Log-bilinear association:** In view of  $\psi_X(x) = \|h_X(x)\|^2$  condition **(C2)** reduces to a moment condition for  $\mathbf{Z}_k = h_X(X_k)$ , namely

$$(C2)_{LBA} \quad E\{\|\mathbf{Z}_k\|^2\} < \infty \quad \text{for all } k = 0, \dots, K.$$

And - using the parametrization (5.11) - condition **(C3)** reduces to

$$h_X^T(X) \boldsymbol{\theta}_{k1} = h_X^T(X) \boldsymbol{\theta}_{k2} \text{ for } k = 1, \dots, K \text{ almost surely} \quad \Rightarrow \quad \boldsymbol{\theta}_{k1} = \boldsymbol{\theta}_{k2} \text{ for all } k,$$

which is implied by the stronger condition **(R2)**<sub>LBA</sub>.  $\square$

## 7. Asymptotic Normality

Let us finally establish the asymptotic normality for a sequence  $\hat{\lambda}^{(n)}$  of estimates. Instead of assuming the property **(C1)**, we derive the asymptotic distribution more generally for *any* weakly consistent sequence  $\hat{\lambda}^{(n)}$  solving the estimating equation at least approximately, i.e. we assume

$$(N1) \quad D_{\lambda} \ell^{(n)}(\hat{\lambda}^{(n)}) = o_p(\sqrt{n}) \quad \text{resp.} \quad n^{-1/2} \cdot D_{\lambda} \ell^{(n)}(\hat{\lambda}^{(n)}) \xrightarrow[n \rightarrow \infty]{P} \mathbf{0}.$$

$$(N2) \quad \hat{\lambda}^{(n)} \xrightarrow[n \rightarrow \infty]{P} \lambda^\circ.$$

Obviously both conditions hold under the assumptions **(C1)**, **(C2)** and **(C3)** of theorem 4. Furthermore we assume the following consistency results, which are derived later (theorem 6) from **(N2)** and additional moment conditions.

$$(N3) \quad \frac{1}{n} \int_0^1 \mathbf{J}^{(n)}(\lambda^\circ + t[\hat{\lambda}^{(n)} - \lambda^\circ]) dt \xrightarrow[n \rightarrow \infty]{P} \bar{\mathbf{I}}(\lambda^\circ).$$

$$(N4) \quad \frac{1}{n} \mathbf{J}^{(n)}(\hat{\lambda}^{(n)}) \xrightarrow[n \rightarrow \infty]{P} \bar{\mathbf{I}}(\lambda^\circ).$$

In **A4** we derive the asymptotic normality of the estimate as follows, where  $\mathbf{A}^-$  resp.  $\mathbf{A}^{1/2}$  denote the generalized Moore-Penrose inverse resp. the symmetric root of a positive semidefinite matrix  $\mathbf{A}$ , and  $\mathbb{I}$  is the identity matrix.

**Theorem 5 (Normality).** Any sequence  $\hat{\lambda}^{(n)}$  of estimators with (N1), (N2) and (N3) is asymptotic normal

$$(a) \quad \sqrt{n} [\hat{\lambda}^{(n)} - \lambda^\circ] \xrightarrow[n \rightarrow \infty]{\mathcal{L}} N(\mathbf{0}, \bar{\mathbf{I}}^{-1}(\lambda^\circ) \cdot \bar{\Sigma}(\lambda^\circ) \cdot \bar{\mathbf{I}}^{-1}(\lambda^\circ))$$

$$\text{with } \bar{\Sigma}(\lambda) := \sum_k \bar{r}_k \cdot \text{Cov}(D_\lambda \log p_k^*(X_k)),$$

$$(b) \quad \sqrt{n} [\hat{\theta}^{(n)} - \theta^\circ] \xrightarrow[n \rightarrow \infty]{\mathcal{L}} N(\mathbf{0}, [\bar{\mathbf{I}}^{-1}(\lambda^\circ)]_{\theta\theta}).$$

**Corollary.** If in addition (N4) holds, then

$$(c) \quad ([\mathbf{J}^{(n)}(\hat{\lambda}^{(n)})]_{\theta\theta}^{-1/2})^{-1} [\hat{\theta}^{(n)} - \theta^\circ] \xrightarrow[n \rightarrow \infty]{\mathcal{L}} N(\mathbf{0}, \mathbb{I}).$$

Replacing the matrices  $\bar{\mathbf{I}}$  resp.  $\bar{\Sigma}$  by  $\frac{1}{n} \mathbf{I}$  resp.  $\frac{1}{n} \Sigma$ , we get the asymptotic normality of the estimates  $\hat{\lambda}$  resp.  $\hat{\theta}$  in less formal notation as

$$\begin{aligned} \hat{\lambda} &\underset{\text{as}}{\sim} N(\lambda^\circ, \mathbf{I}^{-1}(\lambda^\circ) \cdot \Sigma(\lambda^\circ) \cdot \mathbf{I}^{-1}(\lambda^\circ)), \\ \hat{\theta} &\underset{\text{as}}{\sim} N(\theta, [\mathbf{I}^{-1}(\lambda^\circ)]_{\theta\theta}). \end{aligned}$$

The matrix  $\mathbf{J}(\hat{\lambda})$  is a consistent estimate of  $\mathbf{I}(\lambda^\circ)$  by (N4), and will be positive definite for almost all  $n$  (almost surely) by (6.1). In this case, (c) above may be written as

$$\hat{\theta} \underset{\text{as}}{\sim} N(\theta, [\mathbf{J}^{-1}(\hat{\lambda})]_{\theta\theta}). \quad (7.1)$$

Notice that for an observed data set, the estimated covariance matrix  $[\mathbf{J}^{-1}(\hat{\lambda})]_{\theta\theta}$  (where the random variables are replaced by observations) is *identical* to the corresponding matrix under sampling conditional on  $X$  (instead of  $Y$ ). And in this sense the estimate  $\hat{\theta}$  and its estimated asymptotic normal distribution are invariant under sampling conditional on either  $Y$  or  $X$ . Hence any asymptotic inference (i.e. tests or confidence regions) concerning the association parameter  $\theta$  which is based *only* on the asymptotic distribution (c) resp. (7.1) of the estimate  $\hat{\theta}$  are invariant under both conditional sampling schemes, too.

The above conditions (N3) and (N4) will now be derived from the consistency (N2) and additional stochastic properties of the function  $G$ . Defining the functions

$$\begin{aligned} H_r(\mathbf{z} | \theta) &= \sum_{k=0}^K |D_{\theta_r} G(\mathbf{z}, h_Y(y_k), \theta)|, \\ H_{rs}(\mathbf{z} | \theta) &= \sum_{k=0}^K |D_{\theta_r \theta_s}^2 G(\mathbf{z}, h_Y(y_k), \theta)|, \\ H_{rst}(\mathbf{z} | \theta) &= \sum_{k=0}^K |D_{\theta_r \theta_s \theta_t}^3 G(\mathbf{z}, h_Y(y_k), \theta)|, \end{aligned}$$

the following result is proved in **A5**.

**Theorem 6:** Conditions **(N3)** and **(N4)** follow from **(N2)** and the moment condition

**(MC)** There exists  $\varepsilon^\circ > 0$  such that for  $B(\boldsymbol{\theta}^\circ) = \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}^\circ\| \leq \varepsilon^\circ\}$  and all  $k = 0, \dots, K$  the following functions of  $\mathbf{Z}_k = h_X(X_k)$

$$\sup_{\boldsymbol{\theta} \in \bar{B}(\boldsymbol{\theta}^\circ)} H_r(\mathbf{Z}_k \mid \boldsymbol{\theta})^3, \quad \sup_{\boldsymbol{\theta} \in \bar{B}(\boldsymbol{\theta}^\circ)} H_{st}(\mathbf{Z}_k \mid \boldsymbol{\theta})^2, \quad \sup_{\boldsymbol{\theta} \in \bar{B}(\boldsymbol{\theta}^\circ)} H_{rst}(\mathbf{Z}_k \mid \boldsymbol{\theta})$$

have finite expectation for all  $r, s, t = 1, \dots, S$ .

Hence the requirement for the normality theorem are met if **(MC)** and the assumptions **(C1)**, **(C2)** and **(C3)** in theorem 4 hold.

**Log-bilinear association:** The log bilinear association model is based on the function  $G(\mathbf{z}, \mathbf{v}, \boldsymbol{\theta}) = \mathbf{z}^T \boldsymbol{\theta} \mathbf{v}$  with partial derivatives  $D_{\theta_{lm}} G(\mathbf{z}, \mathbf{v}, \boldsymbol{\theta}) = z_l v_m$  and vanishing higher derivatives. Hence **(MC)** holds provided condition **(C2)**<sub>LBA</sub> is strengthened to

$$\text{(MC)}_{\text{LBA}} \quad E\{\|\mathbf{Z}_k\|^3\} < \infty \quad \text{for all } k = 0, \dots, K. \quad \square$$

## 8. Discussion

Association models for a pair of random elements  $(X, Y)$  do not restrict the marginal distributions of  $X$  and  $Y$  but only their odds ratio function. We have looked at a rather general parametric association model **(AM)** which includes the important log-bilinear association models **(LBA)**. An advantage of these models is that inference about the odds ratio (or association) parameter vector  $\boldsymbol{\theta}$  may be obtained from sampling  $Y$  conditional on fixed values of  $X$  or vice versa. Moreover the maximum likelihood estimate  $\hat{\boldsymbol{\theta}}$  is the same for both conditional sampling schemes, i.e. it maximizes the conditional likelihood  $L_{Y|X}$  if and only if it maximizes the "reverse" conditional likelihood  $L_{X|Y}$ . Even more important is that asymptotic inference concerning  $\boldsymbol{\theta}$  (in a our asymptotic approach) is invariant with respect to sampling, too. More precisely, we have shown that for samples conditional on  $Y$ , the estimate  $\hat{\boldsymbol{\theta}}$  maximizing the "reverse" conditional likelihood  $L_{Y|X}$  is consistent, asymptotic normal and its estimated asymptotic covariance matrix is the same as if sampling had been conditional on  $X$ . These results have been obtained much earlier by Prentice and Pyke (1979) for *discrete*  $Y$  with *finite* range and the multivariate linear logistic regression model - which is equivalent to

a log-bilinear association. Weinberg, and Wacholder (1993) extended these results to the general logistic regression model (**GLR**) but require a *finite* range for  $X$  (and provide the proofs only for *binary*  $Y$ ). Our result are not limited neither to discrete distributions (with finite range) for  $Y$  or  $X$  nor to log-bilinear association. Furthermore, asymptotic inference for regression parameters is available when sampling is conditional on  $Y$  (instead of  $X$ ) as outlined below.

#### **Example 4: Linear Models**

Suppose in example 1 that  $\mathcal{L}(Y|X=x)$  has a normal distribution  $N(\mu(x), \sigma^2)$  and the linear model

$$\mu(x) = \alpha + \mathbf{z}^T \boldsymbol{\beta}$$

holds. This is a log-bilinear association model with parameter  $\boldsymbol{\theta} = \sigma^{-2} \boldsymbol{\beta}$ , and hence asymptotic inference for  $\boldsymbol{\theta}$  may be obtained from samples conditional on  $Y$ . This includes (asymptotic) tests of a linear hypothesis  $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$  - which is equivalent to  $\mathbf{C}\boldsymbol{\theta} = \mathbf{0}$ . However, confidence regions are only available for  $\boldsymbol{\theta}$  but not for  $\boldsymbol{\beta}$  - unless an estimate of  $\sigma^2$  from another source is at hand (e.g. from a historic sample). - These findings extend to the multivariate case (example 3) where the conditional distribution of  $Y$  is multivariate normal  $N_K(\mu(x), \boldsymbol{\Sigma})$  and the odds ratio parameter is given by  $\boldsymbol{\theta} = \boldsymbol{\beta} \boldsymbol{\Sigma}^{-1}$ .  $\square$

#### **Example 5: Log-linear Models**

Continuing example 1, let  $\mathcal{L}(Y|X=x)$  be a Poisson distribution satisfying the log-linear model

$$\log \mu(x) = \alpha + \mathbf{z}^T \boldsymbol{\beta},$$

which also represents a log-bilinear association model with parameter  $\boldsymbol{\theta} = \boldsymbol{\beta}$ . Hence asymptotic inference about the parameter  $\boldsymbol{\beta}$  is available when sampling is conditional on  $Y$ .  $\square$

The statistical analysis for  $\boldsymbol{\theta}$  in a *bilinear* association model and sampling conditional on  $Y$  is possible with standard software packages for multivariate linear logistic regression models (or the equivalent log-linear models). And for the more general association model (**AM**), any software for *nonlinear* logistic regression will do.

## Appendix: Proofs

### A1 Proof of the Results in Section 4

**Proof of theorem 1 (uniqueness):** The proof - already given in Osius (2000) under slightly weaker assumptions - is included here only for completeness. To establish  $P_1 = P_2$ , it suffices to show that the *Kullback-Leibler information*

$$I(P_1|P_2) = \int \log(p_1/p_2) dP_1$$

is zero. Using (2.2) and  $OR(P_1) = OR(P_2)$ , the difference of the log-densities  $\varphi_i = \log p_i$  of  $P_i$  may be written

$$(\varphi_1 - \varphi_2)(x, y) = \beta_1(x) + \gamma_1(y) \quad \text{with}$$

$$\beta_1(x) = (\varphi_1 - \varphi_2)(x, y^\circ), \quad \gamma_1(y) = (\varphi_1 - \varphi_2)(x^\circ, y) - (\varphi_1 - \varphi_2)(x^\circ, y^\circ).$$

For  $P^X = P_1^X = P_2^X$  and  $P^Y = P_1^Y = P_2^Y$  we get

$$I(P_1|P_2) = \int (\varphi_1 - \varphi_2) dP_1 = \int (\beta_1 + \gamma_1) dP_1 = \int \beta_1 dP^X + \int \gamma_1 dP^Y$$

and thus  $-\int \beta_1 dP^X \leq \int \gamma_1 dP^Y$ .

Switching the indices 1 and 2, we get the corresponding result

$$-\int \beta_2 dP^X \leq \int \gamma_2 dP^Y \quad \text{with} \quad \beta_2 = -\beta_1, \quad \gamma_2 = -\gamma_1.$$

Hence  $-\int \beta_1 dP^X = \int \gamma_1 dP^Y$  and  $I(P_1|P_2) = 0$ .  $\square$

**Proof of the theorem 2 (existence):** We have already shown, that (b) implies (c), and - since (a) obviously implies (b) - it remains derive (a) from (c). The proof uses the concept of an *I*-projection and heavily relies on results by Csiszár (1975) and Rüschemdorf & Thomsen (1993). Setting  $\pi = \pi_X \times \pi_Y$  we first conclude from **(E2)** the existence of  $R \in \mathcal{P}$  with  $\pi$ -density

$$r = \exp(\psi - \beta - \gamma - \alpha) > 0, \quad \alpha = \log \int \exp(\psi - \beta - \gamma) d\pi.$$

We will show, that the wanted  $P$  is the *I*-projection of  $R$  on the subset of  $\mathcal{P}$  with the given marginals

$$\mathcal{E} = \{P \in \mathcal{P} \mid P^X = \pi_X, P^Y = \pi_Y\}.$$

The integrability of  $\psi$ ,  $\beta$  and  $\gamma$  imply

$$I(\pi|R) = \int \log\left(\frac{1}{r}\right) d\pi = \int (\alpha + \beta + \gamma - \psi) d\pi < \infty,$$

and since  $\pi \in \mathcal{E}$ , we conclude from Csiszár (1975, Thm. 2.1), that  $R$  has an *I*-projection  $P$  on  $\mathcal{E}$ . Application of Csiszár (1975, Thm. 3.1) to the set

$$\mathcal{F} = \{f_X + f_Y \mid f_X \in \mathcal{L}_1(\pi_X), f_Y \in \mathcal{L}_1(\pi_Y)\} \subset \mathcal{L}_1(P)$$

yields that the  $R$ -density  $p_R$  of  $P$  satisfies  $p_R = \exp(h)$   $\pi$ -almost-surely, where  $h$  belongs to the closure  $\mathcal{F}^-$  of  $\mathcal{F}$  in  $\mathcal{L}_1(P)$ . Rüschemdorf & Thomsen (1993, Ex. 2) pointed out that  $\mathcal{F}$  need not be closed in  $\mathcal{L}_1(P)$  - which was claimed in the proof of Cor. 3.1 case (B) in Csiszár (1975). - Now  $R \ll \pi$  implies that  $\exp(h) > 0$  is an  $R$ -density of  $P$ , and hence  $R \ll P \ll R$ . Furthermore  $r > 0$  yields  $R \ll \pi \ll R$ , and hence  $P \in \mathcal{P}_{\ll}$ , since  $P^{XY} = \pi$ . From Csiszár (1975, Thm. 2.2) we obtain

$$I(\pi|P) + I(P|R) \leq I(\pi|R) < \infty,$$

which establishes  $P \in \mathcal{P}_f$ . Finally  $OR(P) = \psi$  remains to be shown. From  $P \ll P^{XY}$  and Rüschemdorf & Thomsen (1993, Prop. 2) we conclude the existence of measurable functions  $a: \Omega_X \rightarrow \mathbb{R}$  and  $b: \Omega_Y \rightarrow \mathbb{R}$ , such that  $h(x, y) = b(x) + c(y)$   $P$ -almost-surely, and hence  $R$ -almost-surely. Hence a  $\pi$ -density of  $p$  is given by

$$\frac{dP}{d\pi} = \frac{dP}{dR} \cdot \frac{dR}{d\pi} = \exp(b+c) \cdot r = \exp(b+c-\beta-\gamma-\alpha+\psi),$$

and a direct calculation yields  $\log OR(P) = \psi$  as required. - Note, that the proof only uses the integrability of the sum  $(\beta + \gamma - \psi)$  - but not the integrability of its components  $\beta, \gamma$  and  $\psi$ .  $\square$

## A2 Proof of the Results in Section 5

We start with some preliminary results. The derivatives of  $\log p_k^*$  are given by

$$\begin{aligned} D_{\lambda_s} \log p_k^*(x) &= \frac{D_{\lambda_s} p_k^*(x)}{p_k^*(x)} \\ D_{\lambda_s \lambda_t}^2 \log p_k^*(x) &= \frac{D_{\lambda_s \lambda_t}^2 p_k^*(x)}{p_k^*(x)} - D_{\lambda_s} \log p_k^*(x) \cdot D_{\lambda_t} \log p_k^*(x). \end{aligned} \quad (\text{A2.1})$$

For an arbitrary family of measurable functions  $G_k(x)$  we obtain from (5.1)

$$\begin{aligned} \sum_k \bar{r}_k \cdot E(G_k(X_k)) &= \sum_k \bar{r}_k \cdot E(G_k(X) \mid Y = y_k) \\ &= \sum_k \bar{r}_k \cdot \int G_k(x) \cdot p(X=x \mid Y=y_k) d\nu_X(x) \\ &= \int \sum_k G_k(x) \cdot p_k^*(x) \cdot p^*(X=x) d\nu_X(x) \\ &= E^* \left[ \sum_k G_k(X) \cdot p_k^*(X) \right], \end{aligned} \quad (\text{A2.2})$$

where  $E^*$  denotes expectation with respect to  $P^*$ .

In particular, we get for  $G_k(x) = H(x) \cdot D_{\lambda} \log p_k^*(x)$  and any measurable  $H(x)$

$$\begin{aligned} \sum_k \bar{r}_k \cdot E[H(X_k) \cdot D_{\lambda} \log p_k^*(X_k)] &= E^* \left[ \sum_k H(X) \cdot D_{\lambda} \log p_k^*(X) \cdot p_k^*(X) \right] \\ &= E^* \left[ H(X) \cdot \sum_k D_{\lambda} p_k^*(X) \right] \\ &= E^* \left[ H(X) \cdot D_{\lambda} p_+^*(X) \right] = \mathbf{0}, \end{aligned} \quad (\text{A2.3})$$

since  $p_+^*(x) = 1$ .

*Proof of (5.6):* Choose  $H(x) = 1$  in (A2.3).  $\square$

*Proof of (5.9):* Choosing  $G_k(X_k) = p_k^*(X_k)^{-1} \cdot D_{\lambda_s \lambda_t}^2 p_k^*(X_k)$  in (A2.2) yields

$$\begin{aligned} \sum_k \bar{r}_k \cdot E[p_k^*(X_k)^{-1} \cdot D_{\lambda_s \lambda_t}^2 p_k^*(X_k)] &= E^* \left[ \sum_k D_{\lambda_s \lambda_t}^2 p_k^*(X_k) \right] \\ &= E^* \left[ D_{\lambda_s \lambda_t}^2 p_+^*(X) \right] = \mathbf{0}, \end{aligned}$$

and (5.9) follows using (A2.1)

$$E(J_{st}(\lambda)) = n \cdot \sum_k \bar{r}_k \cdot E(D_{\lambda_s} \log p_k^*(X_k) \cdot D_{\lambda_t} \log p_k^*(X_k)). \quad \square$$

*Proof of (R2)  $\Leftrightarrow$  (R2)':*

By (5.10)  $\mathbf{I}(\lambda)$  is a sum of  $\Sigma(\lambda)$  and a positive semidefinite matrix. Hence  $\mathbf{I}(\lambda)$  is positive semidefinite, and even positive definite, provided (R2) holds. Conversely, let (R2)' hold. Then  $\mathbf{t}^T \Sigma(\lambda) \mathbf{t} = \text{Var}(\mathbf{t}^T D_{\lambda} \ell(\lambda)^T) = 0$  implies, that  $\mathbf{t}^T D_{\lambda} \ell(\lambda)^T$  is constant almost surely, and hence  $\mathbf{t}^T D_{\lambda \lambda}^2 \ell(\lambda) = D_{\lambda} [\mathbf{t}^T D_{\lambda} \ell(\lambda)^T] = \mathbf{0}$  almost surely. Thus  $\mathbf{t}^T \mathbf{I}(\lambda) = E(\mathbf{t}^T D_{\lambda \lambda}^2 \ell(\lambda)) = \mathbf{0}$ , which implies  $\mathbf{t} = \mathbf{0}$  by (R2)'. Hence (R2) holds.  $\square$

*Proof of (R2)'  $\Leftrightarrow$  (R2)'':*

$\mathbf{I}(\lambda)$  is positive semidefinite (as already observed) and hence (R2)' is equivalent to

$$\text{For all } \mathbf{t} \in \mathbb{R}^{S+K}: \quad \mathbf{t}^T \mathbf{I}(\lambda) \mathbf{t} = 0 \quad \Rightarrow \quad \mathbf{t} = \mathbf{0}. \quad (\text{A2.4})$$

For any  $\mathbf{t} \in \mathbb{R}^{S+K}$  we get from (5.9)

$$\begin{aligned} \mathbf{t}^T \mathbf{I}(\lambda) \mathbf{t} &= \sum_k n_k \cdot \mathbf{t}^T E(D_{\lambda} \log p_k^*(X_k)^T \cdot D_{\lambda} \log p_k^*(X_k)) \mathbf{t} \\ &= \sum_k n_k \cdot E(\|D_{\lambda} \log p_k^*(X_k) \cdot \mathbf{t}\|^2), \end{aligned}$$

and since the distributions of  $X_k$  and  $X$  dominate each other:

$$\mathbf{t}^T \mathbf{I}(\lambda) \mathbf{t} = 0 \quad \Leftrightarrow \quad D_{\lambda} \log p_k^*(X) \cdot \mathbf{t} = 0 \quad \text{for } k = 0, \dots, K \quad \text{almost surely.} \quad (\text{A2.5})$$

First we derive  $(\mathbf{R2})'$  from  $(\mathbf{R2})''$ . For any  $\mathbf{t}$  with  $\mathbf{t}^T \mathbf{I}(\lambda) \mathbf{t} = 0$  we have to show  $\mathbf{t} = \mathbf{0}$ . From (5.4) we get

$$\text{logit } p_k^*(X) = \log p_k^*(X) - \log p_0^*(X) = \gamma_k^* + \psi_{\boldsymbol{\theta}}(X, y_k). \quad (\text{A2.6})$$

Writing  $\mathbf{t} = (\mathbf{s}, -\mathbf{c})$  with  $\mathbf{s} \in \mathbb{R}^S$ ,  $\mathbf{c} = (c_1, \dots, c_K)$  we obtain from (A2.5)

$$\begin{aligned} 0 &= D_{\lambda} \text{logit } p_k^*(X) \cdot \mathbf{t} \\ &= D_{\boldsymbol{\theta}} \text{logit } p_k^*(X) \cdot \mathbf{s} - D_{\boldsymbol{\gamma}} \text{logit } p_k^*(X) \cdot \mathbf{c} \\ &= D_{\boldsymbol{\theta}} \psi_{\boldsymbol{\theta}}(X, y_k) \cdot \mathbf{s} - c_k \quad \text{for all } k = 1, \dots, K \quad \text{almost surely.} \end{aligned} \quad (\text{A2.7})$$

And from  $(\mathbf{R2})''$  we conclude  $\mathbf{s} = \mathbf{0}$  as well as  $c_k = 0$  for all  $k$ , and thus  $\mathbf{t} = \mathbf{0}$ .

Conversely, suppose  $(\mathbf{R2})'$  holds. To establish  $(\mathbf{R2})''$ , it suffices to show that (A2.7) implies  $\mathbf{s} = \mathbf{0}$ . From (5.2) and (5.4) we get

$$\begin{aligned} p_0^*(X) &= \left( \sum_l \exp[\text{logit } p_l^*(X, y_l)] \right)^{-1}, \\ D_{\lambda} \log p_0^*(X) \cdot \mathbf{t} &= p_0^*(X)^{-1} \sum_l \exp[\text{logit } p_l^*(X, y_l)] \cdot D_{\lambda} \text{logit } p_l^*(X, y_l) \cdot \mathbf{t}. \end{aligned}$$

Hence (A2.7) - and  $\text{logit } p_0^* = 0$  - imply  $D_{\lambda} \log p_0^*(X) \cdot \mathbf{t} = 0$  almost surely. From (A2.6) we get

$$D_{\lambda} \log p_k^*(X) \cdot \mathbf{t} = 0 \quad \text{for } k = 0, \dots, K \quad \text{almost surely,}$$

and (A2.5), (A2.4) establish  $\mathbf{t} = \mathbf{0}$  and hence  $\mathbf{s} = \mathbf{0}$ .  $\square$

**Proof of theorem 3:** Since  $\mathbf{I} = \mathbf{I}(\lambda)$  and  $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\lambda)$  are symmetric, (a) is equivalent to three equations

$$(a)_{\boldsymbol{\theta}\boldsymbol{\theta}} \quad \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\theta}} = \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\gamma}} \cdot \mathbf{W} \cdot \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\gamma}}^T,$$

$$(a)_{\boldsymbol{\theta}\boldsymbol{\gamma}} \quad \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\gamma}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\gamma}} = \mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\gamma}} \cdot \mathbf{W} \cdot \mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}},$$

$$(a)_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \quad \mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} - \boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} = \mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \cdot \mathbf{W} \cdot \mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}.$$

Some prerequisite results are derived first using the notations

$$\begin{aligned} b_{sk} &= E[D_{\theta_s} \log p_k^*(X_k)] \in \mathbb{R}, & \mathbf{b}_k &= (b_{1k}, \dots, b_{Sk}) \in \mathbb{R}^S, \\ c_{mk} &= E[D_{\gamma_m^*} \log p_k^*(X_k)] \in \mathbb{R}, & \mathbf{c}_k &= (c_{1k}, \dots, c_{Kk}) \in \mathbb{R}^K, \\ \mathbf{B} &= (\mathbf{b}_1, \dots, \mathbf{b}_K) \in \mathbb{R}^{S \times K}, & \bar{\mathbf{B}} &= (\mathbf{b}_0, \dots, \mathbf{b}_K) \in \mathbb{R}^{S \times (K+1)}, \\ \mathbf{C} &= (\mathbf{c}_1, \dots, \mathbf{c}_K) \in \mathbb{R}^{K \times K}, & \bar{\mathbf{C}} &= (\mathbf{c}_0, \dots, \mathbf{c}_K) \in \mathbb{R}^{K \times (K+1)}, \\ \mathbf{N} &= \text{diag}(n_1, \dots, n_K) \in \mathbb{R}^{K \times K}, & \bar{\mathbf{N}} &= \text{diag}(n_0, \dots, n_K) \in \mathbb{R}^{K \times (K+1)}. \end{aligned}$$

From (5.3) we obtain the partial derivatives

$$D_{\theta_s} \log p_k^*(x) = D_{\theta_s} \psi_{\boldsymbol{\theta}}(x, y_k) - \sum_l p_l^*(x) \cdot D_{\theta_s} \psi_{\boldsymbol{\theta}}(x, y_l),$$

$$D_{\gamma_m^*} \log p_k^*(x) = \Delta_{km} - p_m^*(x),$$

and (5.9) yields

$$\begin{aligned} I_{\lambda_t \gamma_m^*} &= n \sum_k \bar{r}_k \cdot E(D_{\theta_s} \log p_k^*(X_k) \cdot D_{\gamma_m^*} \log p_k^*(X_k)) \\ &= n_m \cdot E(D_{\theta_s} \log p_m^*(X_k)) - n \sum_k \bar{r}_k \cdot E(p_m^*(X_k) \cdot D_{\theta_s} \log p_k^*(X_k)) \\ &= n_m \cdot E(D_{\theta_s} \log p_m^*(X_k)) \quad \text{cf. (A2.3) for } H(x) = p_m^*(x). \end{aligned}$$

$$\begin{aligned} \text{Hence } I_{\theta_s \gamma_m^*} &= n_m \cdot b_{sm}, & I_{\gamma_l^* \gamma_m^*} &= n_m \cdot c_{lm}, & \text{or in matrix notation} \\ \mathbf{I}_{\theta\gamma} &= \mathbf{B} \cdot \mathbf{N}, & \mathbf{I}_{\gamma\gamma} &= \mathbf{C} \cdot \mathbf{N}. \end{aligned} \quad (\text{A2.8})$$

From (5.6) we have  $\sum_k n_k \cdot b_{sk} = 0$  and  $\sum_k n_k \cdot c_{mk} = 0$  or in matrix notation

$$\mathbf{0} = n_0 \mathbf{b}_0 + \mathbf{B} \mathbf{n}, \quad \mathbf{0} = n_0 \mathbf{c}_0 + \mathbf{C} \mathbf{n}, \quad \mathbf{n} = (n_1, \dots, n_K). \quad (\text{A2.9})$$

Using the constant vector  $\mathbf{e}_+ = (1)$  and constant matrix  $\mathbf{e}_+ \mathbf{e}_+^T = (1)$  we thus obtain

$$\begin{aligned} \mathbf{I}_{\theta\gamma} \cdot \mathbf{W} &= \mathbf{B} \cdot \mathbf{N} \cdot \mathbf{W} = \mathbf{B} \cdot \mathbf{N} [n_0^{-1} \mathbf{e}_+ \mathbf{e}_+^T + \mathbf{N}^{-1}] \\ &= n_0^{-1} \mathbf{B} \cdot \mathbf{N} \cdot \mathbf{e}_+ \mathbf{e}_+^T + \mathbf{B} \\ &= n_0^{-1} \mathbf{B} \cdot \mathbf{n} \cdot \mathbf{e}_+^T + \mathbf{B} \\ &= -\mathbf{b}_0 \cdot \mathbf{e}_+^T + \mathbf{B}, \end{aligned}$$

and similar with  $\mathbf{C}$  instead of  $\mathbf{B}$

$$\mathbf{I}_{\gamma\gamma} \cdot \mathbf{W} = \mathbf{C} \cdot \mathbf{N} \cdot \mathbf{W} = -\mathbf{c}_0 \cdot \mathbf{e}_+^T + \mathbf{C}.$$

We now get (a) <sub>$\theta\gamma$</sub>  as follows

$$\begin{aligned} \mathbf{I}_{\theta\gamma} \cdot \mathbf{W} \cdot \mathbf{I}_{\gamma\gamma}^T &= [\mathbf{B} - \mathbf{b}_0 \cdot \mathbf{e}_+^T] [\mathbf{C} \cdot \mathbf{N}]^T && \text{cf. (A2.8)} \\ &= \mathbf{B} \cdot \mathbf{N} \cdot \mathbf{C}^T - \mathbf{b}_0 \cdot [\mathbf{C} \cdot \mathbf{N} \cdot \mathbf{e}_+^T]^T \\ &= \dots \dots \dots - \mathbf{b}_0 \cdot [\mathbf{C} \cdot \mathbf{n}]^T \\ &= \dots \dots \dots + \mathbf{b}_0 \cdot n_0 \cdot \mathbf{c}_0^T && \text{cf. (A2.9)} \\ &= \bar{\mathbf{B}} \cdot \bar{\mathbf{N}} \cdot \bar{\mathbf{C}}^T \\ &= \mathbf{I}_{\theta\gamma} - \Sigma_{\theta\gamma} && \text{cf. (5.10)}. \end{aligned}$$

And (a) <sub>$\theta\theta$</sub>  resp. (a) <sub>$\gamma\gamma$</sub>  are established similarly (replace  $\mathbf{B}$  and  $\mathbf{b}_0$  by  $\mathbf{C}$  and  $\mathbf{c}_0$  resp. vice versa). Hence (a) holds, and multiplication from both sides with the inverse  $\mathbf{I}^{-1}(\lambda)$  yields (b).  $\square$

**Proof of (R2)<sub>LBA</sub> ⇒ (R2)'':**

Suppose for  $\mathbf{s} = (s_1, \dots, s_K) \in \mathbb{R}^{K \times K}$  and  $c_1, \dots, c_K \in \mathbb{R}$  we have for all  $k = 1, \dots, K$

$$c_k = D_{\boldsymbol{\theta}} \psi_{\boldsymbol{\theta}}(X, y_k) \cdot \mathbf{s} = \sum_l D_{\boldsymbol{\theta}_l} \psi_{\boldsymbol{\theta}}(X, y_k) \cdot s_l = h_X(X)^T \cdot \mathbf{s}_k \quad \text{almost surely.}$$

Then (R2)<sub>LBA</sub> implies  $\mathbf{s}_k = \mathbf{0}$  for all  $k$ , and hence  $\mathbf{s} = \mathbf{0}$ . □

### A3 Proof of Theorem 4 (Consistency) in Section 6

The proof is based on ideas from Wald (1949) and requires some preliminary results. The log odds ratio  $\psi_{\boldsymbol{\theta}}(x, y)$  in the model (AM) depends only on the vectors  $\mathbf{z} = h_X(x)$  and  $\mathbf{v} = h_Y(y)$ . Therefore we regard  $p_k^*(x) = \tilde{p}_k(\mathbf{z} | \boldsymbol{\lambda})$  as a function of  $\mathbf{z}$  and  $\boldsymbol{\lambda}$  using the notations

$$\begin{aligned} G_k(\mathbf{z}, \boldsymbol{\theta}) &:= G(\mathbf{z}, h_Y(y_k), \boldsymbol{\theta}) = \psi_{\boldsymbol{\theta}}(x, y_k), \\ \tilde{p}_k(\mathbf{z} | \boldsymbol{\lambda}) &:= \frac{\exp[\gamma_k^* + G_k(\mathbf{z}, \boldsymbol{\theta})]}{\sum_l \exp[\gamma_l^* + G_l(\mathbf{z}, \boldsymbol{\theta})]} = p_k^*(x), \\ \eta_k(\mathbf{z} | \boldsymbol{\lambda}) &:= \log \tilde{p}_k(\mathbf{z} | \boldsymbol{\lambda}) = \gamma_k^* + G_k(\mathbf{z}, \boldsymbol{\theta}) - \log(\sum_l \exp[\gamma_l^* + G_l(\mathbf{z}, \boldsymbol{\theta})]). \end{aligned}$$

We first show for  $\mathbf{Z}_k := h_X(X_k)$

$$E\{|\eta_k(\mathbf{Z}_k | \boldsymbol{\lambda})|\} < \infty \quad \text{for all } \boldsymbol{\lambda} \text{ and } k = 0, \dots, K. \quad (\text{A3.1})$$

From  $\gamma_0^* = 0 = G_0(\mathbf{z}, \boldsymbol{\theta})$  and  $\tilde{p}_0(\mathbf{z} | \boldsymbol{\lambda}) \leq 1$  we get

$$\begin{aligned} |\eta_0(\mathbf{z} | \boldsymbol{\lambda})| &= \log(\sum_l \exp[\gamma_l^* + G_l(\mathbf{z}, \boldsymbol{\theta})]) \\ &\leq \log((K+1) \cdot \text{Max}_l \exp[\gamma_l^* + G_l(\mathbf{z}, \boldsymbol{\theta})]) \\ &\leq \log(K+1) + \|\boldsymbol{\gamma}^*\| + \text{Max}_l |G_l(\mathbf{z}, \boldsymbol{\theta})|. \end{aligned}$$

And (OR2) yields

$$|G_l(\mathbf{z}, \boldsymbol{\theta})| \leq [\psi_X(x) + \psi_Y(y_l)] \cdot \|\boldsymbol{\theta}\|, \quad (\text{A3.2})$$

which in view of (C2) proves (A3.1) for  $k = 0$ . For  $k > 0$  we get

$$\begin{aligned} |\eta_k(\mathbf{z} | \boldsymbol{\lambda})| &= |\gamma_k^* + G_k(\mathbf{z}, \boldsymbol{\theta}) + \eta_0(\mathbf{z} | \boldsymbol{\lambda})| \\ &\leq \|\boldsymbol{\gamma}^*\| + |G_k(\mathbf{z}, \boldsymbol{\theta})| + |\eta_0(\mathbf{z} | \boldsymbol{\lambda})|. \end{aligned}$$

Hence (A3.2) and (C2) establish (A3.1). □

Next we prove three basic lemmas.

**Lemma 1.** For any  $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}^\circ$ : 
$$\sum_{k=0}^K \bar{r}_k \cdot E\{\eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}) - \eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}^\circ)\} < 0.$$

**Lemma 2.** For  $k = 0, \dots, K$  and any  $\lambda$ :

$$\lim_{\varepsilon \rightarrow 0} E \left\{ \sup_{\|\lambda' - \lambda\| \leq \varepsilon} \eta_k(\mathbf{Z}_k | \lambda') \right\} = E \{ \eta_k(\mathbf{Z}_k | \lambda) \}.$$

**Lemma 3.** For any compact set  $A \subset \mathbb{R}^K \times \mathbb{R}^S$  with  $\lambda^\circ \notin A$ :

$$P \left\{ \lim_{n \rightarrow \infty} \left[ \sup_{\lambda \in A} \ell^{(n)}(\lambda) - \ell^{(n)}(\lambda^\circ) \right] = -\infty \right\} = 1.$$

**Proof of lemma 1.** For  $k = 0, \dots, K$  the random variable  $U_k = \eta_k(\mathbf{Z}_k | \lambda) - \eta_k(\mathbf{Z}_k | \lambda^\circ)$  has finite expectation by (A3.1), and Jensen's inequality gives

$$E \{ U_k \} \leq \log E \{ \exp(U_k) \},$$

where equality holds if and only if there is a constant  $c_k > 0$  such that

$$U_k = \log c_k \quad \text{resp.} \quad \tilde{p}_k(\mathbf{Z}_k | \lambda) = c_k \cdot \tilde{p}_k(\mathbf{Z}_k | \lambda^\circ) \quad \text{almost surely.} \quad (\text{A3.3})$$

Hence

$$\begin{aligned} \sum_k \bar{r}_k \cdot E \{ U_k \} &\leq \sum_k \bar{r}_k \cdot \log E \{ \exp(U_k) \} \\ &\leq \log \left( \sum_k \bar{r}_k \cdot E \{ \exp(U_k) \} \right), \end{aligned} \quad (\text{A3.4})$$

where the second inequality is strict unless all expectations coincide, i.e.

$$E \{ \exp(U_0) \} = E \{ \exp(U_1) \} = \dots = E \{ \exp(U_K) \}. \quad (\text{A3.5})$$

Application of (A2.2) to  $G_k(X_k) = \exp(U_k) = \tilde{p}_k(\mathbf{Z}_k | \lambda) [\tilde{p}_k(\mathbf{Z}_k | \lambda^\circ)]^{-1}$  gives

$$\begin{aligned} \sum_k \bar{r}_k \cdot E \{ \exp(U_k) \} &= E^* \left\{ \sum_k \tilde{p}_k(\mathbf{Z} | \lambda) [\tilde{p}_k(\mathbf{Z} | \lambda^\circ)]^{-1} \cdot \tilde{p}_k(\mathbf{Z} | \lambda^\circ) \right\} \\ &= E^* \left\{ \sum_k \tilde{p}_k(\mathbf{Z} | \lambda) \right\} = 1 \end{aligned}$$

and (A3.4) implies  $\sum_k \bar{r}_k \cdot E \{ U_k \} \leq 0$ . It remains to show, that the last inequality is strict. We suppose, that this is not the case and derive a contradiction. Then (A3.3) holds for all  $k$ , and (A3.5) implies  $c_k = c$  for all  $k$ . From (A3.3) and  $\sum_k \tilde{p}_k = 1$  we get  $c = 1$ , and hence

$$\eta_k(\mathbf{Z}_k | \lambda) = \eta_k(\mathbf{Z}_k | \lambda^\circ) \quad \text{for all } k \quad \text{almost surely,} \quad (\text{A3.6})$$

and

$$\begin{aligned} \psi_{\boldsymbol{\theta}}(X_k, y_k) &= \eta_k(\mathbf{Z}_k | \lambda) + \eta_0(\mathbf{Z}_0 | \lambda) - \eta_0(\mathbf{Z}_k | \lambda) - \eta_k(\mathbf{Z}_0 | \lambda) \\ &= \eta_k(\mathbf{Z}_k | \lambda^\circ) + \eta_0(\mathbf{Z}_0 | \lambda^\circ) - \eta_0(\mathbf{Z}_k | \lambda^\circ) - \eta_k(\mathbf{Z}_0 | \lambda^\circ) \\ &= \psi_{\boldsymbol{\theta}^\circ}(X_k, y_k) \quad \text{for all } k \quad \text{almost surely.} \end{aligned}$$

Since the distributions of  $X_k$  and  $X$  dominate each other, we conclude

$$\psi_{\boldsymbol{\theta}}(X, y_k) = \psi_{\boldsymbol{\theta}^\circ}(X, y_k) \quad \text{for all } k \quad \text{almost surely,}$$

and (C3) yields  $\boldsymbol{\theta} = \boldsymbol{\theta}^\circ$ . Furthermore, (A3.6) gives for  $\lambda = (\boldsymbol{\theta}, \boldsymbol{\gamma}^*)$  almost surely

$$\begin{aligned}\gamma_k^* + G_k(\mathbf{Z}_k, \boldsymbol{\theta}) &= \eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}) - \eta_0(\mathbf{Z}_k | \boldsymbol{\lambda}) \\ &= \eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}^\circ) - \eta_0(\mathbf{Z}_k | \boldsymbol{\lambda}^\circ) = \gamma_k^\circ + G_k(\mathbf{Z}_k, \boldsymbol{\theta}^\circ)\end{aligned}\quad \text{for all } k,$$

and from  $\boldsymbol{\theta} = \boldsymbol{\theta}^\circ$  we conclude  $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^\circ$  which contradicts  $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}^\circ$ .  $\square$

**Proof of lemma 2.** For any positive sequence  $\varepsilon_n \rightarrow 0$  the continuity of  $\eta_k(\mathbf{z} | \boldsymbol{\lambda})$  with respect to  $\boldsymbol{\lambda}$  implies for any  $\mathbf{z}$

$$\sup_{\|\boldsymbol{\lambda}' - \boldsymbol{\lambda}\| \leq \varepsilon_n} \eta_k(\mathbf{z} | \boldsymbol{\lambda}') \xrightarrow{n \rightarrow \infty} \eta_k(\mathbf{z} | \boldsymbol{\lambda}).$$

$$\text{Since } \eta_k(\mathbf{z} | \boldsymbol{\lambda}) \leq \sup_{\|\boldsymbol{\lambda}' - \boldsymbol{\lambda}\| \leq \varepsilon_n} \eta_k(\mathbf{z} | \boldsymbol{\lambda}') \leq 0, \quad (\text{A3.7})$$

(A3.1) and the dominated convergence theorem yield

$$E\left\{ \sup_{\|\boldsymbol{\lambda}' - \boldsymbol{\lambda}\| \leq \varepsilon_n} \eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}') \right\} \xrightarrow{n \rightarrow \infty} E\left\{ \eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}) \right\}. \quad \square$$

**Proof of lemma 3.** For any  $\boldsymbol{\lambda}$  and  $\varepsilon > 0$  let

$$\begin{aligned}B(\boldsymbol{\lambda} | \varepsilon) &= \{ \boldsymbol{\lambda}' | \|\boldsymbol{\lambda}' - \boldsymbol{\lambda}\| \leq \varepsilon \}, & B^\circ(\boldsymbol{\lambda} | \varepsilon) &= \{ \boldsymbol{\lambda}' | \|\boldsymbol{\lambda}' - \boldsymbol{\lambda}\| < \varepsilon \}, \\ \eta_k(\mathbf{z} | \boldsymbol{\lambda}, \varepsilon) &= \sup_{\boldsymbol{\lambda}' \in B(\boldsymbol{\lambda} | \varepsilon)} \eta_k(\mathbf{z} | \boldsymbol{\lambda}').\end{aligned}$$

Then lemma 2 implies

$$\lim_{\varepsilon \rightarrow 0} \sum_k \bar{r}_k \cdot E\left\{ \eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}, \varepsilon) \right\} = \sum_k \bar{r}_k \cdot E\left\{ \eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}) \right\}.$$

For any  $\boldsymbol{\lambda} \in A$  lemma 1 gives

$$\sum_k \bar{r}_k \cdot E\left\{ \eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}) \right\} < \sum_k \bar{r}_k \cdot E\left\{ \eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}^\circ) \right\},$$

and hence there exists an  $\varepsilon_\lambda > 0$  such that

$$\sum_k \bar{r}_k \cdot E\left\{ \eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}, \varepsilon_\lambda) \right\} < \sum_k \bar{r}_k \cdot E\left\{ \eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}^\circ) \right\}. \quad (\text{A3.8})$$

Since  $A$  is compact, there exists a finite subset  $\{\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_M\} \subset A$  such that

$$A \subset \bigcup_{m=1}^M B^\circ(\boldsymbol{\lambda}_m | \varepsilon_{\boldsymbol{\lambda}_m}).$$

Hence for any  $\boldsymbol{\lambda} \in A$  we have  $\boldsymbol{\lambda} \in B^\circ(\boldsymbol{\lambda}_m | \varepsilon_{\boldsymbol{\lambda}_m})$  for at least one  $m \in \{1, \dots, M\}$  and thus

$$\eta_k(\mathbf{z} | \boldsymbol{\lambda}) \leq \eta_k(\mathbf{z} | \boldsymbol{\lambda}_m, \varepsilon_{\boldsymbol{\lambda}_m}),$$

$$\sup_{\boldsymbol{\lambda} \in A} \ell^{(n)}(\boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda} \in A} \sum_k \sum_i \eta_k(\mathbf{Z}_{ki} | \boldsymbol{\lambda}) \leq \text{Max}_m \sum_k \sum_i \eta_k(\mathbf{Z}_{ki} | \boldsymbol{\lambda}_m, \varepsilon_{\boldsymbol{\lambda}_m})$$

$$\text{and } \sup_{\boldsymbol{\lambda} \in A} \ell^{(n)}(\boldsymbol{\lambda}) - \ell^{(n)}(\boldsymbol{\lambda}^\circ) \leq \text{Max}_m \sum_k \sum_i \eta_k(\mathbf{Z}_{ki} | \boldsymbol{\lambda}_m, \varepsilon_{\boldsymbol{\lambda}_m}) - \ell^{(n)}(\boldsymbol{\lambda}^\circ).$$

The proof will be complete, if we show that right side tends to  $-\infty$  almost surely.

The strong law of large numbers gives almost surely for each  $k$  and  $m$

$$\lim_{n_k \rightarrow \infty} \frac{1}{n_k} \sum_i [\eta_k(\mathbf{Z}_{ki} | \boldsymbol{\lambda}_m, \varepsilon_{\boldsymbol{\lambda}_m}) - \eta_k(\mathbf{Z}_{ki} | \boldsymbol{\lambda}^\circ)] = E\{\eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}_m, \varepsilon_{\boldsymbol{\lambda}_m})\} - E\{\eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}^\circ)\},$$

whith finite expectations by (A3.1) and (A3.7). Thus

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_k \sum_i [\eta_k(\mathbf{Z}_{ki} | \boldsymbol{\lambda}_m, \varepsilon_{\boldsymbol{\lambda}_m}) - \eta_k(\mathbf{Z}_{ki} | \boldsymbol{\lambda}^\circ)] &= \\ \sum_k \bar{r}_k \cdot [E\{\eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}_m, \varepsilon_{\boldsymbol{\lambda}_m})\} - E\{\eta_k(\mathbf{Z}_k | \boldsymbol{\lambda}^\circ)\}] &< 0, \quad \text{cf. (A3.8)}. \end{aligned}$$

Hence, we have almost surely for each  $m$

$$\lim_{n \rightarrow \infty} \sum_k \sum_i [\eta_k(\mathbf{Z}_{ki} | \boldsymbol{\lambda}_m, \varepsilon_{\boldsymbol{\lambda}_m}) - \eta_k(\mathbf{Z}_{ki} | \boldsymbol{\lambda}^\circ)] = -\infty,$$

and thus 
$$\lim_{n \rightarrow \infty} \left[ \text{Max}_m \sum_k \sum_i \eta_k(\mathbf{Z}_{ki} | \boldsymbol{\lambda}_m, \varepsilon_{\boldsymbol{\lambda}_m}) - \ell^{(n)}(\boldsymbol{\lambda}^\circ) \right] =$$

$$\lim_{n \rightarrow \infty} \text{Max}_m \sum_k \sum_i [\eta_k(\mathbf{Z}_{ki} | \boldsymbol{\lambda}_m, \varepsilon_{\boldsymbol{\lambda}_m}) - \eta_k(\mathbf{Z}_{ki} | \boldsymbol{\lambda}^\circ)] = -\infty. \quad \square$$

**Proof of Theorem 4 (Consistency).** For any  $\varepsilon > 0$ , the function  $\ell^{(n)}(\boldsymbol{\lambda})$  attains its maximum on the compact ball  $B(\boldsymbol{\lambda}^\circ | \varepsilon)$ . We show first that (almost surely) the maximizing argument lies (for almost all  $n$ ) in the *open* ball  $B^\circ(\boldsymbol{\lambda}^\circ | \varepsilon)$ , and hence is a solution of  $D_\lambda \ell^{(n)}(\boldsymbol{\lambda}) = \mathbf{0}$ . Applying lemma 3 to the compact boundary  $A_\varepsilon = \partial B(\boldsymbol{\lambda}^\circ | \varepsilon)$  yields that the following statements hold almost surely for almost all  $n$

(i) 
$$\sup_{\boldsymbol{\lambda} \in A_\varepsilon} \ell^{(n)}(\boldsymbol{\lambda}) < \ell^{(n)}(\boldsymbol{\lambda}^\circ),$$

(ii) 
$$\sup_{\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\circ\| \leq \varepsilon} \ell^{(n)}(\boldsymbol{\lambda}) \leq \sup_{\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\circ\| < \varepsilon} \ell^{(n)}(\boldsymbol{\lambda}),$$

(iii) there exists  $\tilde{\boldsymbol{\lambda}}^{(n)} \in B^\circ(\boldsymbol{\lambda}^\circ | \varepsilon)$  with  $D_\lambda \ell^{(n)}(\tilde{\boldsymbol{\lambda}}^{(n)}) = \mathbf{0}$ ,

(iv)  $\ell^{(n)}(\boldsymbol{\lambda})$  ist strictly concave, cf. (6.1), (R2)'

(v) there is a unique  $\tilde{\boldsymbol{\lambda}}^{(n)} \in B^\circ(\boldsymbol{\lambda}^\circ | \varepsilon)$  maximizing  $\ell^{(n)}(\boldsymbol{\lambda})$ ,

(vi)  $\tilde{\boldsymbol{\lambda}}^{(n)} = \hat{\boldsymbol{\lambda}}^{(n)}$ , cf. (C1).

This proves (a), (b) and also (c), since  $\varepsilon$  was arbitrary. □

#### A4 Proof of the Results in Section 7

**Proof of Theorem 5 (Normality).** The proof follows the usual pattern, and therefore not all details are given. First we decompose the score function  $\mathbf{U} = D_\lambda \ell^T$  according to the subsamples

$$\mathbf{U}^{(n)}(\boldsymbol{\lambda}) := [D_{\boldsymbol{\lambda}} \ell^{(n)}(\boldsymbol{\lambda})]^T = \sum_{k=0}^K \mathbf{U}_k^{(n)}(\boldsymbol{\lambda}), \quad \mathbf{U}_k^{(n)}(\boldsymbol{\lambda}) := \sum_{i=1}^{n_k} D_{\boldsymbol{\lambda}} \log p_k^*(X_{ki}).$$

From the central limit theorem and Cramér-Wold's device we obtain that each  $\mathbf{U}_k^{(n)}(\boldsymbol{\lambda})$  is asymptotic normal

$$n_k^{-1/2} [\mathbf{U}_k^{(n)}(\boldsymbol{\lambda}) - n_k \cdot \boldsymbol{\mu}_k(\boldsymbol{\lambda})] \xrightarrow[n \rightarrow \infty]{\mathcal{L}} N(\mathbf{0}, \boldsymbol{\Sigma}_k(\boldsymbol{\lambda})) \quad \text{with}$$

$$\boldsymbol{\mu}_k(\boldsymbol{\lambda}) := E(D_{\boldsymbol{\lambda}} \log p_k^*(X_k)), \quad \boldsymbol{\Sigma}_k(\boldsymbol{\lambda}) := \text{Cov}(D_{\boldsymbol{\lambda}} \log p_k^*(X_k)).$$

Since  $\sum_k n_k \cdot \boldsymbol{\mu}_k(\boldsymbol{\lambda}) = \mathbf{0}$  by (5.6), we obtain the asymptotic normality of  $\mathbf{U}(\boldsymbol{\lambda})$

$$n^{-1/2} \mathbf{U}^{(n)}(\boldsymbol{\lambda}) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} N(\mathbf{0}, \bar{\boldsymbol{\Sigma}}(\boldsymbol{\lambda})). \quad (\text{A4.1})$$

The covariance matrix  $\bar{\boldsymbol{\Sigma}}(\boldsymbol{\lambda}) = \frac{1}{n} \boldsymbol{\Sigma}^{(n)}(\boldsymbol{\lambda})$  is positive-definite by **(R2)**. A first-order Taylor expansion about  $\boldsymbol{\lambda}^\circ$  gives

$$n^{-1/2} \mathbf{U}^{(n)}(\hat{\boldsymbol{\lambda}}^{(n)}) = n^{-1/2} \mathbf{U}^{(n)}(\boldsymbol{\lambda}^\circ) + \mathbf{D}_n \cdot \sqrt{n} [\hat{\boldsymbol{\lambda}}^{(n)} - \boldsymbol{\lambda}^\circ] \quad \text{with}$$

$$\mathbf{D}_n := \frac{1}{n} \int_0^1 D_{\boldsymbol{\lambda}} \mathbf{U}^{(n)}(\boldsymbol{\lambda}^\circ + t[\hat{\boldsymbol{\lambda}}^{(n)} - \boldsymbol{\lambda}^\circ]) dt = -\frac{1}{n} \int_0^1 \mathbf{J}^{(n)}(\boldsymbol{\lambda}^\circ + t[\hat{\boldsymbol{\lambda}}^{(n)} - \boldsymbol{\lambda}^\circ]) dt,$$

and **(N1)** implies

$$\mathbf{D}_n \cdot \sqrt{n} [\hat{\boldsymbol{\lambda}}^{(n)} - \boldsymbol{\lambda}^\circ] + n^{-1/2} \mathbf{U}^{(n)}(\boldsymbol{\lambda}^\circ) \xrightarrow[n \rightarrow \infty]{P} \mathbf{0}. \quad (\text{A4.2})$$

We show next, that  $\mathbf{D}_n$  can be replaced in (A4.2) by its limit  $-\bar{\mathbf{I}}(\boldsymbol{\lambda}^\circ)$  from **(N3)**, i.e.

$$\sqrt{n} [\hat{\boldsymbol{\lambda}}^{(n)} - \boldsymbol{\lambda}^\circ] - n^{-1/2} \bar{\mathbf{I}}^{-1}(\boldsymbol{\lambda}^\circ) \mathbf{U}^{(n)}(\boldsymbol{\lambda}^\circ) \xrightarrow[n \rightarrow \infty]{P} \mathbf{0}, \quad (\text{A4.3})$$

which together with (A4.1) - applied to  $\boldsymbol{\lambda} = \boldsymbol{\lambda}^\circ$  - establishes (a). Since the set  $\mathcal{S}$  of all invertible  $(S+K) \times (S+K)$ -matrices is open, we conclude from **(N3)** and  $\bar{\mathbf{I}}(\boldsymbol{\lambda}^\circ) \in \mathcal{S}$

$$P\{\mathbf{D}_n \in \mathcal{S}\} \xrightarrow[n \rightarrow \infty]{P} 1.$$

Hence  $P\{\mathbf{D}_n \in \mathcal{S}\} \leq P\{[\mathbb{I} - \mathbf{D}_n^- \mathbf{D}_n] \sqrt{n} [\hat{\boldsymbol{\lambda}}^{(n)} - \boldsymbol{\lambda}^\circ] = \mathbf{0}\}$

yields  $[\mathbb{I} - \mathbf{D}_n^- \mathbf{D}_n] \sqrt{n} [\hat{\boldsymbol{\lambda}}^{(n)} - \boldsymbol{\lambda}^\circ] \xrightarrow[n \rightarrow \infty]{P} \mathbf{0}$ . (A4.4)

Furthermore, **(N3)** implies

$$[\mathbf{D}_n^- + \bar{\mathbf{I}}^{-1}(\boldsymbol{\lambda}^\circ)] \xrightarrow[n \rightarrow \infty]{P} \mathbf{0},$$

and from (A4.1) with  $\boldsymbol{\lambda} = \boldsymbol{\lambda}^\circ$  we get

$$n^{-1/2} [\mathbf{D}_n^- + \bar{\mathbf{I}}^{-1}(\boldsymbol{\lambda}^\circ)] \mathbf{U}^{(n)}(\boldsymbol{\lambda}^\circ) \xrightarrow[n \rightarrow \infty]{P} \mathbf{0}. \quad (\text{A4.5})$$

Multiplying (A4.2) with  $\mathbf{D}_n^- \xrightarrow{P} -\bar{\mathbf{I}}^{-1}(\boldsymbol{\lambda}^\circ)$  yields

$$\mathbf{D}_n^- \mathbf{D}_n \cdot \sqrt{n} [\hat{\boldsymbol{\lambda}}^{(n)} - \boldsymbol{\lambda}^\circ] + n^{-1/2} \mathbf{D}_n^- \mathbf{U}^{(n)}(\boldsymbol{\lambda}^\circ) \xrightarrow[n \rightarrow \infty]{P} \mathbf{0},$$

hence  $\sqrt{n} [\hat{\boldsymbol{\lambda}}^{(n)} - \boldsymbol{\lambda}^\circ] + n^{-1/2} \mathbf{D}_n^- \mathbf{U}^{(n)}(\boldsymbol{\lambda}^\circ) \xrightarrow[n \rightarrow \infty]{P} \mathbf{0}$ , by (A4.4)

which implies (A4.3) using (A4.5). This completes the proof of (a), which implies (b) in view of theorem 3 (b). Since  $\mathbf{A}^-$  and  $\mathbf{A}^{1/2}$  are continuous operators, (N4) yields

$$\frac{1}{\sqrt{n}}([\mathbf{J}^{(n)}(\hat{\lambda}^{(n)})^-]_{\theta\theta}^{1/2})^- = ([\frac{1}{n}\mathbf{J}^{(n)}(\hat{\lambda}^{(n)})^-]_{\theta\theta}^{1/2})^- \xrightarrow[n \rightarrow \infty]{P} ([\bar{\mathbf{I}}^{-1}(\lambda^\circ)]_{\theta\theta}^{1/2})^-$$

and multiplication with (b) establishes (c).  $\square$

**Proof of Theorem 6.** Keeping the notations from A3, the partial derivatives of

$$\eta_k(\mathbf{z} | \lambda) = \log \tilde{p}_k(\mathbf{z} | \lambda) = \gamma_k^* + G_k(\mathbf{z}, \theta) - \log(\sum_l \exp[\gamma_l^* + G_l(\mathbf{z}, \theta)])$$

up to order three are given by

$$\begin{aligned} D_{\gamma_m^*} \eta_k(\mathbf{z} | \lambda) &= \Delta_{km} - \tilde{p}_m(\mathbf{z} | \lambda), \\ D_{\gamma_m^* \lambda_s}^2 \eta_k(\mathbf{z} | \lambda) &= -D_{\lambda_s} \tilde{p}_k(\mathbf{z} | \lambda), \\ D_{\gamma_m^* \lambda_s \lambda_t}^3 \eta_k(\mathbf{z} | \lambda) &= -D_{\lambda_s \lambda_t}^2 \tilde{p}_k(\mathbf{z} | \lambda), \\ D_{\theta_r} \eta_k(\mathbf{z} | \lambda) &= \sum_l [\Delta_{kl} - \tilde{p}_l(\mathbf{z} | \lambda)] \cdot D_{\theta_r} G_l(\mathbf{z}, \theta), \\ D_{\theta_r \theta_s}^2 \eta_k(\mathbf{z} | \lambda) &= \sum_l ([\Delta_{kl} - \tilde{p}_l(\mathbf{z} | \lambda)] \cdot D_{\theta_r \theta_s}^2 G_l(\mathbf{z}, \theta) - D_{\theta_s} \tilde{p}_l(\mathbf{z} | \lambda) \cdot D_{\theta_r} G_l(\mathbf{z}, \theta)), \\ D_{\theta_r \theta_s \theta_t}^3 \eta_k(\mathbf{z} | \lambda) &= \sum_l ([\Delta_{kl} - \tilde{p}_l(\mathbf{z} | \lambda)] \cdot D_{\theta_r \theta_s \theta_t}^3 G_l(\mathbf{z}, \theta) - D_{\theta_t} \tilde{p}_l(\mathbf{z} | \lambda) \cdot D_{\theta_r \theta_s}^2 G_l(\mathbf{z}, \theta) \\ &\quad - D_{\theta_s} \tilde{p}_l(\mathbf{z} | \lambda) \cdot D_{\theta_r \theta_t}^2 G_l(\mathbf{z}, \theta) - D_{\theta_s \theta_t}^2 \tilde{p}_l(\mathbf{z} | \lambda) \cdot D_{\theta_r} G_l(\mathbf{z}, \theta)), \end{aligned}$$

with partial derivatives - cf. (A2.1)

$$\begin{aligned} D_{\lambda_s} \tilde{p}_k(\mathbf{z} | \lambda) &= \tilde{p}_k(\mathbf{z} | \lambda) \cdot D_{\lambda_s} \eta_k(\mathbf{z} | \lambda), \\ D_{\lambda_s \lambda_t}^2 \tilde{p}_k(\mathbf{z} | \lambda) &= \tilde{p}_k(\mathbf{z} | \lambda) [D_{\lambda_s \lambda_t}^2 \eta_k(\mathbf{z} | \lambda) - D_{\lambda_s} \eta_k(\mathbf{z} | \lambda) \cdot D_{\lambda_t} \eta_k(\mathbf{z} | \lambda)]. \end{aligned}$$

Next we deduce from (MC) a weaker moment condition, from which (N3) and (N4) will be derived (cf. lemma 4).

(MC) $\sim$  There exists  $\varepsilon^\circ > 0$  such that for  $B(\lambda^\circ) = \{\lambda \mid \|\lambda - \lambda^\circ\| \leq \varepsilon^\circ\}$  and all  $k = 0, \dots, K$  the following functions

$$\sup_{\lambda \in B(\lambda^\circ)} |D_{\lambda_r \lambda_s \lambda_t}^3 \eta_l(\mathbf{Z}_k | \lambda)| \quad \text{with} \quad \mathbf{Z}_k = h_X(X_k)$$

have finite expectation for all  $r, s, t = 1, \dots, S$  and  $l = 0, \dots, K$ .

For the above derivatives we successively get the following bounds, where the fixed argument  $\mathbf{z}$  is omitted:

$$\begin{aligned} |D_{\gamma_m^*} \eta_k(\lambda)| &\leq 1, & |D_{\theta_r} \eta_k(\lambda)| &\leq H_+(\theta), \\ |D_{\lambda_r} \eta_k(\lambda)| &\leq H_+^*(\theta) := 1 + H_+(\theta), \end{aligned}$$

$$\begin{aligned}
|D_{\gamma_m^* \lambda_s}^2 \eta_k(\boldsymbol{\lambda})| &= |D_{\lambda_s} \tilde{p}_k(\boldsymbol{\lambda})| \leq |D_{\lambda_s} \eta_k(\boldsymbol{\lambda})| \leq H_+^*(\boldsymbol{\theta}), \\
|D_{\theta_r \theta_s}^2 \eta_k(\boldsymbol{\lambda})| &\leq H_{++}(\boldsymbol{\theta}) + H_+(\boldsymbol{\theta})^2, \\
|D_{\lambda_s \lambda_t}^2 \eta_k(\boldsymbol{\lambda})| &\leq H_+^*(\boldsymbol{\theta})^2 + H_{++}(\boldsymbol{\theta}), \\
|D_{\lambda_s \lambda_t}^2 \tilde{p}_k(\boldsymbol{\lambda})| &\leq 2H_+^*(\boldsymbol{\theta})^2 + H_{++}(\boldsymbol{\theta}), \\
|D_{\gamma_m^* \lambda_s \lambda_t}^3 \eta_k(\boldsymbol{\lambda})| &= |D_{\lambda_s \lambda_t}^2 \tilde{p}_k(\boldsymbol{\lambda})| \leq 2H_+^*(\boldsymbol{\theta})^2 + H_{++}(\boldsymbol{\theta}), \\
|D_{\theta_r \theta_s \theta_t}^3 \eta_k(\boldsymbol{\lambda})| &\leq H_{+++}(\boldsymbol{\theta}) + 3H_+^*(\boldsymbol{\theta})H_{++}(\boldsymbol{\theta}) + 2H_+^*(\boldsymbol{\theta})^3.
\end{aligned}$$

Taking (for fixed  $\mathbf{z}$ ) the supremum over the ball  $B(\boldsymbol{\theta}^\circ)$  gives

$$\begin{aligned}
\sup H_+^* &\leq 1 + \sum_r \sup H_r, \\
\sup H_+^{*2} &\leq 1 + 2 \sum_s \sup H_s + \sum_s \sum_t \sup H_{st}, \\
\sup H_+^{*3} &\leq 1 + 3 \sum_r \sup H_r + 3 \sum_r \sum_s \sup H_r H_s + \sum_r \sum_s \sum_t \sup H_r H_s H_t, \\
\sup H_{++} &\leq \sum_s \sum_t \sup H_{st}, \\
\sup H_{+++} &\leq \sum_r \sum_s \sum_t \sup H_{rst}, \\
\sup H_+^* \cdot H_{++} &\leq \sum_r \sum_s \sum_t [\sup H_{st} + \sup H_r H_{st}].
\end{aligned}$$

Condition (MC) obviously implies that for  $i = 1, 2$

$$\sup_{\boldsymbol{\theta} \in B(\boldsymbol{\theta}^\circ)} H_r(\mathbf{Z}_k | \boldsymbol{\theta})^i, \quad \sup_{\boldsymbol{\theta} \in B(\boldsymbol{\theta}^\circ)} H_r(\mathbf{Z}_k | \boldsymbol{\theta}) \cdot H_{st}(\mathbf{Z}_k | \boldsymbol{\theta})$$

have finite expectation too. Hence

$$\sup_{\boldsymbol{\gamma}} \sup_{\boldsymbol{\theta} \in B(\boldsymbol{\theta}^\circ)} |D_{\lambda_r \lambda_s \lambda_t}^3 \eta_l(\mathbf{Z}_k | \boldsymbol{\theta}, \boldsymbol{\gamma})|$$

has finite expectation for any  $r, s, t$  and any  $k, l$ . This proves (MC) $^\sim$ , since  $\boldsymbol{\lambda} = (\boldsymbol{\theta}, \boldsymbol{\gamma}) \in B(\boldsymbol{\lambda}^\circ)$  implies  $\boldsymbol{\theta} \in B(\boldsymbol{\theta}^\circ)$ . And the next lemma establishes the theorem.  $\square$

**Lemma 4:** (N2) and (MC) $^\sim$  imply (N3) and (N4).

**Proof:** Note that if almost sure convergence  $\hat{\boldsymbol{\lambda}}^{(n)} \rightarrow \boldsymbol{\lambda}^\circ$  (strong consistency) is assumed instead of (N2), then following proofs establish almost sure convergence in (N3) and (N4), too. - From (6.1) we get for  $\boldsymbol{\lambda} = \boldsymbol{\lambda}^\circ$

$$\frac{1}{n} \mathbf{J}^{(n)}(\boldsymbol{\lambda}^\circ) \xrightarrow[n \rightarrow \infty]{P} \bar{\mathbf{I}}(\boldsymbol{\lambda}^\circ), \tag{A4.6}$$

and hence (N3) will hold, if we show for any  $s$  and  $t$

$$\frac{1}{n} \int_0^1 [J_{st}^{(n)}(\lambda^\circ + t[\hat{\lambda}^{(n)} - \lambda^\circ]) - J_{st}^{(n)}(\lambda^\circ)] dt \xrightarrow[n \rightarrow \infty]{P} 0. \quad (\text{A4.7})$$

Now for any  $\varepsilon > 0$  we get

$$\begin{aligned} \|\hat{\lambda}^{(n)} - \lambda^\circ\| < \varepsilon \quad \Rightarrow \quad & \left| \frac{1}{n} \int_0^1 [J_{st}^{(n)}(\lambda^\circ + t[\hat{\lambda}^{(n)} - \lambda^\circ]) - J_{st}^{(n)}(\lambda^\circ)] dt \right| \leq \\ & H_{st}^{(n)}(\varepsilon) := \frac{1}{n} \sup_{\|\lambda - \lambda^\circ\| \leq \varepsilon} |J_{st}^{(n)}(\lambda) - J_{st}(\lambda^\circ)|, \end{aligned} \quad (\text{A4.8})$$

and for  $\mathbf{Z}_{ki} = h_X(X_{ki})$  we have

$$J_{st}^{(n)}(\lambda) - J_{st}^{(n)}(\lambda^\circ) = \sum_k \sum_i [D_{\lambda_s \lambda_t}^2 \eta_k(\mathbf{Z}_{ki} | \lambda^\circ) - D_{\lambda_s \lambda_t}^2 \eta_k(\mathbf{Z}_{ki} | \lambda)].$$

A first-order Taylor expansion around  $\lambda^\circ$  gives

$$D_{\lambda_s \lambda_t}^2 \eta_k(\mathbf{z} | \lambda^\circ) - D_{\lambda_s \lambda_t}^2 \eta_k(\mathbf{z} | \lambda) = \left[ \int_0^1 D_{\lambda_s \lambda_t}^3 \eta_k(\mathbf{z} | \lambda^\circ + t[\lambda - \lambda^\circ]) dt \right] (\lambda^\circ - \lambda).$$

Hence

$$\|\lambda - \lambda^\circ\| < \varepsilon \quad \Rightarrow \quad |J_{st}^{(n)}(\lambda) - J_{st}(\lambda^\circ)| \leq \sum_k \sum_i \sup_{\|\lambda' - \lambda^\circ\| \leq \varepsilon} \|D_{\lambda_s \lambda_t}^3 \eta_k(\mathbf{Z}_{ki} | \lambda')\| \cdot \varepsilon$$

and

$$H_{st}^{(n)}(\varepsilon) \leq \varepsilon \frac{1}{n} \sum_k \sum_i \sup_{\|\lambda' - \lambda^\circ\| \leq \varepsilon} \|D_{\lambda_s \lambda_t}^3 \eta_k(\mathbf{Z}_{ki} | \lambda')\|. \quad (\text{A4.9})$$

For any  $k$  and  $0 < \varepsilon \leq \varepsilon^\circ$  the strong law of large numbers and **(MC)** $\sim$  gives

$$\begin{aligned} S_k^{(n)}(\varepsilon) &:= \frac{1}{n_k} \sum_i \sup_{\|\lambda' - \lambda^\circ\| \leq \varepsilon} \|D_{\lambda_s \lambda_t}^3 \eta_k(\mathbf{Z}_{ki} | \lambda')\| \xrightarrow[n \rightarrow \infty]{} \\ s_k(\varepsilon) &:= E \left( \sup_{\|\lambda' - \lambda^\circ\| \leq \varepsilon} \|D_{\lambda_s \lambda_t}^3 \eta_k(\mathbf{Z}_k | \lambda')\| \right) \quad \text{almost surely,} \end{aligned}$$

and hence

$$\begin{aligned} \bar{S}^{(n)}(\varepsilon) &:= \sum_k \bar{r}_k S_k^{(n)}(\varepsilon) = \frac{1}{n} \sum_k \sum_i \sup_{\|\lambda' - \lambda^\circ\| \leq \varepsilon} \|D_{\lambda_s \lambda_t}^3 \eta_k(\mathbf{Z}_{ki} | \lambda')\| \\ &\xrightarrow[n \rightarrow \infty]{} \bar{s}(\varepsilon) := \sum_k \bar{r}_k s_k(\varepsilon) \quad \text{almost surely.} \end{aligned} \quad (\text{A4.10})$$

From (A4.8) and (A4.9) we get for  $\varepsilon \leq \varepsilon^\circ$

$$\|\hat{\lambda}^{(n)} - \lambda^\circ\| < \varepsilon \quad \Rightarrow \quad \left| \frac{1}{n} \int_0^1 [J_{st}^{(n)}(\lambda^\circ + t[\hat{\lambda}^{(n)} - \lambda^\circ]) - J_{st}^{(n)}(\lambda^\circ)] dt \right| \leq H_{st}^{(n)}(\varepsilon) \leq \varepsilon \bar{S}^{(n)}(\varepsilon),$$

which - together with **(N2)** and (A4.10) - implies (A4.7).

To derive **(N4)**, we first obtain from (A4.9) for any  $\varepsilon \leq \varepsilon^\circ$

$$\|\hat{\lambda}^{(n)} - \lambda^\circ\| < \varepsilon \quad \Rightarrow \quad \frac{1}{n} |J_{st}^{(n)}(\hat{\lambda}^{(n)}) - J_{st}^{(n)}(\lambda^\circ)| \leq H_{st}^{(n)}(\varepsilon) \leq \varepsilon \bar{S}^{(n)}(\varepsilon),$$

which - together with (N2) and (A4.10) - implies

$$\frac{1}{n} [\mathbf{J}^{(n)}(\hat{\lambda}^{(n)}) - \mathbf{J}^{(n)}(\lambda^\circ)] \xrightarrow[n \rightarrow \infty]{P} \mathbf{0}.$$

And from (A4.6) we conclude (N4). □

## References

- Csiszár, I. (1975) *I-divergence geometry of probability distributions and minimization problems*. Annals of Probability 3, 146-158.
- Haberman S.J. (1974) *The Analysis of Frequency Data*. The University of Chicago Press, Chicago.
- Kullback S. (1959) *Information Theory and Statistics*. Mineola, New York: Dover Publications (1968).
- McCullagh P. and Nelder J.A. (1989): *Generalized Linear Models (Second Edition)*. London: Chapman & Hall.
- Osius, G. (2000) *The association between two random elements: A complete characterization in terms of odds ratios*. Mathematik-Arbeitspapiere 53, Universität Bremen (<http://www.math.uni-bremen.de/~osius>).
- Osius, G. (2004) *The association between two random elements: A complete characterization and odds ratio models*. Metrika 60, 261-277.
- Prentice, R.L. and Pyke, R. (1979) *Logistic disease incidence models and case-control studies*. Biometrika 66, 403 - 411.
- Rüschendorf, L. and Thomsen, W. (1993) *Note on the Schrödinger equation and I-projections*. Statistics and Prob. Letters 17, 369-375.
- Rüschendorf, L. (1995) *Convergence of the iterative proportional fitting procedure*. Annals of Statistics 23, 1160-1174.
- Wald, A. (1949) *Note on the consistency of the maximum likelihood estimate*. Annals of Mathematical Statistics 20, 595-601.
- Weinberg, C.R. and Wacholder, S. (1993) *Prospective analysis of case-control data under general multiplicative-intercept risk models*. Biometrika 80, 461-465.

Date: 23-June-2006 (printed edition)

5-Oct-2006 (PDF-version with corrected misprints)

This paper is available for download at: <http://www.math.uni-bremen.de/~osius>

Vertrieb der Hefte 4, 14, 23, 26 durch Universitätsbuchhandlung, Bibliothekstr. 3, D-28359 Bremen.  
Vertrieb der übrigen Hefte (soweit nicht vergriffen) durch die Autoren oder FB 3 Mathematik/Informatik  
Universität Bremen, Postfach 330440, D-28334 Bremen.

1. Ulrich Krause (1976): Strukturen in unendlichdimensionalen konvexen Mengen, 74 S.
2. Fritz Colonius, Diederich Hinrichsen (1976): Optimal control of hereditary differential systems. Part I, 66 S.
3. Günter Matthiessen (1976): Theorie der heterogenen Algebren, 88 S.
4. H. Wolfgang Fischer, Jens Gamst, Klaus Horneffer (1976): Skript zur Analysis, Band 1 (13. Auflage 2004), 286 S.
5. Wolfgang Schröder (1977): Operator-algebraische Ergodentheorie für Quantensysteme, 59 S.
6. Rolf Röhrig, Michael Unterstein (1977): Analyse multivariabler Systeme mit Hilfe komplexer Matrixfunktionen, 216 S.
7. Horst Herrlich, Hans-Eberhard Porst, Rudolf-Eberhard Hoffmann, Manfred Bernd Wischnewsky (1976): Nordwestdeutsches Kategorienseminar, 193 S.
8. Fritz Colonius, Diederich Hinrichsen (1977): Optimal Control of Hereditary Differential Systems. Part II: Differential State Space Description, 36 S.
9. Ludwig Arnold (1977): Differentialgleichungen und Regelungstheorie, 185 S.
10. Rudolf Lorenz (1977): Iterative Verfahren zur Lösung großer, dünnbesetzter symmetrischer Eigenwertprobleme, 104 S.
11. Konrad Behnen, Hans-Peter Kinder, Gerhard Osius, Rüdiger Schäfer, Jürgen Timm (1977): Dose-Response-Analysis, 206 S.
12. Hans-Friedrich Münzner, Dieter Prätzel-Wolters (1978): Minimalbasen polynomialer Moduln, Strukturindizes und BRUNOVSKY-Transformationen, 53 S.
13. Konrad Behnen (1978): Vorzeichen-Rangtests mit Nullen und Bindungen, 53 S.
14. H. Wolfgang Fischer, Jens Gamst, Klaus Horneffer, Eberhard Oeljeklaus (1978): Skript zur Linearen Algebra, Band 1 (15. Auflage 2004), 249 S.
15. Günter Ludyk (1978): Abtastregelung zeitvarianter Einfach- und Mehrfachsysteme, 54 S.
16. Momme Johs Thomsen (1977): Zur Theorie der Fastalgebren, 146 S.
17. Klaus Horneffer, Horst Diehl (1978): Modellrechnungen zur anaeroben Reduktionskinetik des Cytochroms P-450, 34 S.
18. Horst Herrlich, Rudolf-Eberhard Hoffmann, Hans-Eberhard Porst, Manfred Bernd Wischnewsky (1979): Structure of Topological Categories, 252 S.
19. Hans-Friedrich Münzner, Dieter Prätzel-Wolters (1979): Geometric and moduletheoretic approach to linear systems. Part I: Basic categories and functors, 28 S.
20. Hans-Friedrich Münzner, Dieter Prätzel-Wolters (1979): Geometric and moduletheoretic approach to linear systems. Part II: Moduletheoretic characterization and reachability, 28 S.
21. Eckart Beutler, Hans Kaiser, Günter Matthiessen, Jürgen Timm (1979): Biduale Algebren, 165 S.
22. Horst Diehl, Detlef Harbach, Jürgen Timm (1980): Planung und Auswertung von Atomabsorptions-Spektrometrie-Untersuchungen mit der Additionsmethode, 44 S.
23. H. Wolfgang Fischer, Jens Gamst, Klaus Horneffer (1981): Skript zur Analysis, Band 2 (8. Auflage 2003), 299 S.

24. Horst Herrlich (1981): *Categorical Topology 1971-1981*, 105 S.
25. Horst Herrlich, Rudolf-Eberhard Hoffmann, Hans-Eberhard Porst, Manfred Bernd Wischnewsky (1981): *Special Topics in Topology and Category Theory*, 108 S.
26. H. Wolfgang Fischer, Jens Gamst, Klaus Horneffer (1984): *Skript zur Linearen Algebra, Band 2* (9. Auflage 2005), 257 S.
27. Rudolf-Eberhard Hoffmann (1982): *Continuous Lattices and Related Topics*, 314 S.
28. Horst Herrlich, Rudolf-Eberhard Hoffmann, Hans-Eberhard Porst (1987): *Workshop on Category Theory*, 169 S.
29. Harald Boehme (1987): *Zur Berufspraxis des Diplommathematikers*, 16 S.
30. Jürgen Timm (1986): *Mathematische Modelle der Dosis-Wirkungsanalyse bei den experimentellen Untersuchungen der Arbeitsgruppe zur karzinogenen Belastung des Menschen durch Luftverunreinigung*, 65 S.
31. Dieter Denneberg (1988): *Mathematik für Wirtschaftswissenschaftler. I. Lineare Algebra*, 97 S.
32. Peter E. Crouch, Diederich Hinrichsen, Anthony J. Pritchard, Dietmar Salamon (1988, previous edition University of Warwick 1981): *Introduction to Mathematical Systems Theory*, 244 S.
33. Gerhard Osius (1989): *Some Results on Convergence of Moments and Convergence in Distribution with Applications in Statistics*, 27 S.
34. Dieter Denneberg (1989): *Verzerrte Wahrscheinlichkeiten in der Versicherungsmathematik, Quantilsabhängige Prämienprinzipien*, 24 S.
35. Eberhard Oeljeklaus (1989): *Birational splitting of homogeneous Albanese bundles*, 30 S.
36. Gerhard Osius, Dieter Rojek (1989): *Normal Goodness-of-Fit Tests for Parametric Multinomial Models with Large Degrees of Freedom*, 38 S.
37. Dieter Denneberg (1990): *Mathematik zur Wirtschaftswissenschaft. II. Analysis*, 59 S.
38. Ulrich Krause, Cornelia Zahlten (1990): *Arithmetik in Krull monoids and the cross number of divisor class groups*, 29 S.
39. Dieter Denneberg (1990): *Subadditive Measure and Integral*, 39 S.
40. Ulrich Krause, Peter Ranft (1991): *A limit set trichotomy for monotone nonlinear dynamical systems*, 31 S.
41. Angelika van der Linde (1992): *Statistical analyses with splines: are they well defined?* 22 S.
42. Dieter Denneberg (1992): *Lectures on non-additive measure and integral (new edition: Non-additive measure and integral. TDLB 27, Kluwer Academic, Dordrecht (1994))*, 114 S.
43. Gerhard Osius (1993): *Separating Agreement from Association in Log-linear Models for Square Contingency Tables With Applications*, 23 S.
44. Hans-Peter Kinder, Friedrich Liese (1995): *Bremen-Rostock Statistik Seminar, 5. - 7. März 1992*, 110 S.
45. Dieter Denneberg (1995): *Extension of a measurable space and linear representation of the Choquet Integral*, 30 S.
46. Dieter Denneberg, Michael Grabisch (1996): *Shapley value and interaction index*, 20 S.
47. Angelika Bunse-Gerstner, Heike Faßbender (1996): *A Jacobi-like method for solving algebraic Riccati equations on parallel computers*, 24 S.
48. Hans-Eberhard Porst editor (1997): *Categorical methods in algebra and topology - a collection of papers in honour of Horst Herrlich*, 498 S.
49. Angelika van der Linde, Gerhard Osius (1997): *Estimation of nonparametric risk functions In matched case-control studies*, 28 S.

50. Angelika van der Linde (1997): Estimating the smoothing parameter in generalized spline-based regression, 46 S.
51. Ursula Müller, Gerhard Osius (1998): Asymptotic normality of goodness-of-fit statistics for sparse Poisson data, 15 S.
52. Ursula Müller (1999): Nonparametric regression for threshold data, 18 S.
53. Gerhard Osius (2000): The association between two random elements – A complete characterization in terms of odds ratios, 32 S.
54. Horst Herrlich, Hans-E. Porst (2000): CatMAT 2000, Proceedings of the Conference: Categorical Methods in Algebra and Topology, 490 S.
55. Gerhard Osius (2001): A formal derivation of the conditional likelihood for matched case-control studies, 30 S.
56. Angelika van der Linde (2002): Dimension reduction and linear discriminant functions based on an odds ratio parameterization, 46 S.
57. Angelika van der Linde (2002): On the association between a random parameter and an observable, 26 S.
58. Walter Schill, Pascal Wild (2002): Design Optimisation for Logistic Regression in Two-Phase Studies, 53 S.
59. Gerhard Osius (2002): Statistik in den Naturwissenschaften (Digitale Neuauflage als PDF: März 2006), 291 S.
60. Gerhard Osius (2006): Semiparametric association models: estimation and asymptotic inference, 38 S.