# The Association Between Two Random Elements: A Complete Characterization And Odds Ratio Models

**Gerhard Osius**

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

For random elements $X$ and $Y$ (e.g. vectors) a complete characterization of their association is given in terms of an odds ratio function. The main result establishes for any odds ratio function and any pre-specified marginals the unique existence of a corresponding joint distribution (the joint density is obtained as a limit of an iterative procedure of marginal fittings). Restricting only the odds ratio function but not the marginals leads to semi-parmetric *association models* for which statistical inference is available for samples drawn *conditionally on either $X$ or $Y$*. *Log-bilinear association models* for random vectors $X$ and $Y$ are introduced which generalize standard (regression) models by removing restrictions on the marginals. In particular, the logistic regression model is recognized as a log-bilinear association model. And the joint distribution of $X$ and $Y$ is shown to be *multivariate normal* if and only if both marginals are normal and the association is log-bilinear.

*Keywords:* association, marginal fitting, multivariate normal distribution, odds ratio, semi-parametric models.

## 1. Introduction and Outline

The question how a random output vector $Y$ of a system (e.g. the health status of a human) is associated to a random input vector $X$ (e.g. consumption of tobacco and alcohol, environmental pollution and other risk factors) is of major importance in statistical science. The standard approach is to consider the outcome $Y$ for given values of the input $X$ which amounts to specifying the conditional distribution $\mathscr{L}(Y \mid X)$ by means of an appropriate model - but leaving the marginal distribution $\mathscr{L}(X)$ of the input arbitrary. The corresponding sampling design is to collect a random sample $Y_i \sim \mathscr{L}(Y \mid X = x_i)$ for given values $x_i$ with $i = 1, ..., I$.

Sometimes however, the reverse approach may be more appropriate which asks what input $X$ has been responsible for given values of the output. This leads to a model for the conditional distribution of $\mathscr{L}(X \mid Y)$ and a sample $X_i \sim \mathscr{L}(X \mid Y = y_i)$ with fixed values $y_i$ $(i = 1, ..., I)$. A remarkable example are case-control studies in epidemiology, where $Y \in \{0,1\}$ is an indicator for a disease and the subpopulations $\{Y = 1\}$ resp. $\{Y = 0\}$ are called cases resp. controls (cf. Breslow and Day 1980).

Comparing both approaches for random vectors $X$ and $Y$ the question arises, which amount of information concerning the joint distribution of $(X,Y)$ is contained in both conditional distributions $\mathscr{L}(Y \mid X)$ and $\mathscr{L}(X \mid Y)$. Conditioning on $X$ resp. $Y$ removes the information on the marginal distribution $\mathscr{L}(X)$ resp. $\mathscr{L}(Y)$ from the joint distribution $\mathscr{L}(X,Y)$ and the question arises what *exactly* is left if we simultaneously remove the information on both marginal distributions from the joint distribution. Focussing on *densities* - on which likelihood analysis and much of Bayesian statistics are based - we are looking for an *association object* $Assoc(X,Y)$ derivable from *any* of the two conditional densities. The joint distribution $\mathscr{L}(X,Y)$ should be uniquely determined by the tripel $\big(\mathscr{L}(X), \mathscr{L}(Y), Assoc(X,Y)\big)$ and - more important - *any* such tripel should give rise to a (unique) joint distribution. This approach allows to define a model for the joint distribution by specifying separate models for the association structure $Assoc(X,Y)$ and the marginals. Any parametric model for the association $Assoc(X,Y)$ which leaves the marginals of $X$ and $Y$ completely arbitrary leads to an *association model*, which is *semi-parametric* with respect to the unspecified marginals. The advantage of an association model over a regression model (specifying the *conditionl* density of $Y$ given $X$) is that inference about the parameters of the association model can be based on samples from either $\mathscr{L}(Y|X)$ or $\mathscr{L}(X|Y)$, because the association structure $Assoc(X,Y)$ is a common part of both conditional densities.

An association model may be obtained as a generalization from a common model (e.g. a regression model) by removing the restrictions imposed on the marginals. We briefly describe two examples to be studied in detail later. Consider first the *joint* multivariate normal distribution for random vectors $X$ and $Y$, which serves as a basic model in multivariate analysis. Identifying its association structure (which is *not* given by the correlation matrix, cf. **4.4**) allows generalizations of multivariate techniques to non-normal *joint* distributions with the same association structure but non-normal *marginals* (cf. van der Linde 2002a). And second, for $Y$ with finite range and a vector $X$ the widely used logistic regression model is equivalent to a corresponding association model (cf. **4.2**) and thus semi-parametric with respect to the marginals.

The decomposition of a joint distribution into the association and the marginals is also useful to distinguish measures of *association* from measures of *dependence*: the former should not depend on the marginals but the latter may. Furthermore, in Bayesian statistics a deeper understanding of the association between an observation $Y$ and a random parameter $\Theta$ (taking the role of $X$) is fundamental and can be studied through our approach (cf. van der Linde 2002b).

The purpose of this paper is to define the association structure *Assoc*$(X,Y)$, which turns out to be an odds ratio *function,* establish its basic properties and provide parametrizations of association models as a formal framework for statistical analysis. More detailed statistical applications of our results are given in van der Linde (2002ab) or van der Linde and Osius (2003). Before outlining the present work we look at two special cases which motivate the general approach.

For *simple* random variables $X \in \{0, ..., J\}$ and $Y \in \{0, ..., K\}$ (i.e. with finite range) the joint distribution is given by $\pi_{jk} = P(X=j\,,\; Y=k)$ for all $j$, $k$. And the $J{\times}K$ odds ratio (or cross product ratio) matrix $\theta$ with entries $\theta_{jk} = (\pi_{jk}\pi_{00})(\pi_{j0}\pi_{0k})^{-1}$ for $j, k \geq 1$ is known to be the wanted object *Assoc*$(X, Y)$, see e.g. Plackett (1974, Sec. 3.4). The result, that for any matrix $\theta$ with positive entries and given marginal distributions there exists a unique joint distribution $\pi$ may be obtained in two ways, both of which will be generalized here. First, the joint distribution $\pi$ may be derived by a method - due to Sinkhorn (1967) - as a limit of iteratively rescaled distributions $\pi_n$ with the given odds ratio matrix $\theta$. Second, the joint distribution $\pi$ may be obtained from results due to Haberman (1974, Theorem 2.6) as the unique argument maximizing a strictly concave function arising from the log-likelihood of Poisson distributions.

Now let $Y \in \{0,1\}$ be an indicator (e.g. for a disease) and $X$ a random vector. The *odds ratio* of an input value $x$ with respect to a reference value $x^\circ$

$$OR^\circ(x) \;=\; \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} \Bigg/ \frac{P(Y=1 \mid X=x^\circ)}{P(Y=0 \mid X=x^\circ)}$$

is a fundamental concept in epidemiology and we now briefly show that the corresponding *odds ratio function* $OR^\circ$ is the desired association object *Assoc*$(X,Y)$. Using the log odds ratio function $\psi^\circ = \log OR^\circ$, the conditional distribution $\mathcal{L}(Y \mid X=x)$ is determined by the conditional probability

$$P(Y=1 \mid X=x) \;=\; \{1 + a^{-1}\exp[-\psi^\circ(x)]\}^{-1} \qquad \text{resp.} \qquad (1)$$

$$\operatorname{logit} P(Y=1 \mid X=x) \;=\; \alpha + \psi^\circ(x)\,,$$

where $\operatorname{logit} p = \log\left[p/(1-p)\right]$ denotes the logistic transformation and $\alpha = \log a = \operatorname{logit} P(Y=1 \mid X=x^\circ)$. The marginal distribution of $Y$ is given by

$$P(Y=1) \;=\; E\big(\{1+a^{-1}\exp[-\,\psi^{\circ}(X)]\}^{-1}\big) \tag{2}$$

which is a strictly increasing function of $a$. Hence for $0 < P(Y=1) < 1$ there exists a unique $0 < a < \infty$ such that (2) holds. This shows that for fixed marginal distributions the joint distribution is uniquely determined by the log odds ratio function $\psi^{\circ}$. Furthermore, for fixed marginal distributions and a given log odds ratio function $\psi^{\circ}$ a joint distribution is defined by (1) with $a$ obtained from (2).

Hence the logistic regression model (1) does not restrict the marginal distribution of $Y$. Thus any model for the odds ratio function, e.g. the $\psi^{\circ}(x) = (x - x^{\circ})^T \theta$ with parameter vector $\theta$, represents a semi-parametric distribution model in the sense above (the superscript "$T$" denotes the transpose). As a consequence - which is well known in epidemiology - statistical inference about the odds ratio parameter $\theta$ may be based either on cohort studies (where sampling is conditional on $X$) or on case-control studies (with sampling conditional on $Y$), whichever are easier to conduct in the particular application. These results will be extended to the logistic regression model for (non-binary) $Y$ with finite range (cf. **4.4**) but the proofs are substantially more complicated. However the major statistical benefit of our approach lies in a further generalization to semi-parametric association models for random *vectors* - or even arbitrary *random elements* - $X$ and $Y$.

Although our applications in section **4** deal only with random *vectors*, we will derive the main results for arbitrary *random elements* (including e.g. random *functions*) $X$ and $Y$ for two reasons. First, our arguments do not exploit specific properties of *finite-dimensional* euklidean space and second, a restriction to random *vectors* would not even simplify the proofs. However the reader may prefer to think of familiar random vectors instead of random elements.

We now outline the basic concepts and results for the general case, where $X$ and $Y$ are random elements. The wanted association object $Assoc(X,Y)$ turns out to be a straightforward generalization of the above odds ratio function. Given a *positive* density $p$ of the joint distribution with respect to some dominating product measure $\nu_X \times \mu_Y$ of $\sigma$-finite measures the odds ratio function with respect to a reference pair $(x^{\circ}, y^{\circ})$ is defined as

$$OR^{\circ}(x,y) \;=\; \frac{p(X=x,\,Y=y\,)}{p(X=x,\,Y=y^{\circ})} \;\Big/\; \frac{p(X=x^{\circ},\,Y=y\,)}{p(X=x^{\circ},\,Y=y^{\circ})}$$

where the *joint* density $p$ can equivalently be replaced by the *conditional* density of either $Y$ given $X$ or conversely. Furthermore the function $OR^{\circ}$ is invariant under a change of the dominating measure and a natural choice for the dominating measure is the product of the marginal distributions. The formal definition of the odds ratio function and its elementary properties are given in section **2**. Within the important class of joint distributions having *integrable* log-densities a modified log odds ratio function $\psi$ (which does not refer to reference values $x^\circ$ and $y^\circ$) is defined as a projection of the log-density into an appropriate association space.

In section **3** we show that the odds ratio function characterizes the association and hence may be taken as the desired association object $Assoc(X,Y)$. Using the Kullback-Leibler information we first prove (under mild integrability conditions) that the joint distribution is *uniquely determined* by the marginal distributions and the odds ratio function. To establish the *existence* of a joint distribution with given fixed marginals and an odds ratio function requires a greater effort and stronger additional assumptions. As our main result the wanted joint distribution will be obtained from a convergence theorem as a limit of a marginal fitting sequence of densities, which generalizes the sequence used by Sinkhorn (1967) for simple random variables. Furthermore the joint density may also be obtained by maximizing a strictly concave function which corresponds to the log-likelihood used by Haberman (1974) and is related to a Kullback-Leibler information. Although the concepts here are straightforward generalizations derived from the work of Sinkhorn and Haberman, their approaches exploit unique features of *finite* distributions which are not available for non-simple random variables.

The benefit of the theoretical results for statistical modelling is adressed in section **4**. First *association models* are introduced which only specify the odds ratio function, but not the marginal distributions. Explicit representations of the joint and the conditional densities are given in terms of odds ratios and nuisance parameters. From the existence theorem we conclude that these associaton models are *semi-parametric*, because any set of marginal distributions for $X$ and $Y$ may be achieved by a proper choice for the nuisance parameters. Removing restrictions on the marginals from standard (regression) models yields an imporant class of *log-bilinear association models*. This is illustrated for joint and conditional *normal* distribution models used in multivariate analysis for random vectors $X$ and $Y$. Furthermore we establish that joint multivariate normality holds if and only if both marginal distributions are normal and the association is *log-bilinear*. Hence log-bilinear association models are semi-parametric generalizations of multivariate normal distribution models. Statistical aspects of these results are pursued further in van der Linde (2002a) and van der Linde and Osius (2002).

An important feature of the approach here is a *symmetry* in presentation between $X$ and $Y$: by interchanging $X$ with $Y$ any concept or argument entails its *dual*. - Most of the proofs, and further results needed for them, are given in the appendix.

## 2. The Odds Ratio Function

To formalize the introductary discussion we consider arbitrary non-empty probability spaces $(\Omega_X, \mathscr{B}_X, \pi_X)$ and $(\Omega_Y, \mathscr{B}_Y, \pi_Y)$ as well as their product $(\Omega, \mathscr{B}, \pi)$, i.e. the set $\Omega = \Omega_X \times \Omega_Y$ equipped with the product measure $\pi = \pi_X \times \pi_Y$. Let $\mathscr{P}$ denote the class of probability measures $P$ on $(\Omega, \mathscr{B})$ which have a *positive* density $f > 0$ with respect to $\pi$, i.e. $P$ is dominated by $\pi$ and dominates $\pi$: $P \ll \pi \ll P$. Further let $\mathscr{F}$ be the class of corresponding densities, i.e. the Radon-Nikodym derivatives $f = dP/d\pi$ for any $P \in \mathscr{P}$, and let $\Phi = \{ \log f \mid f \in \mathscr{F} \}$ be the corresponding class of log-densities.

The restriction to *positive* densities is essential for the definition of the odds ratio function and only rules out less interesting joint distributions. For *simple* random variables $X$ and $Y$ the condition $P \in \mathscr{P}$ is equivalent to

$$P(X = j), P(Y = k) > 0 \quad \Rightarrow \quad P(X = j, Y = k) > 0 \qquad \text{for all } j, k.$$

And if, for example, $P$ is a multivariate *normal* distribution $N_k(\mu, \Sigma)$, then the condition $P \in \mathscr{P}$ holds if and only if the rank of the covariance matrix $\Sigma$ equals the sum of the ranks for both marginal covariance matrices (cf. **4.4**).

### 2.1 Odds and Log Odds Ratio Function

For any density $f \in \mathscr{F}$ the *odds ratio function* $OR_f$ or $OR(f)$ is a map $\Omega \times \Omega \longrightarrow (0, \infty)$ defined by

$$OR_f(x, y \mid x', y') \;=\; [f(x, y) \cdot f(x', y')] \cdot [f(x, y') \cdot f(x', y)]^{-1}$$

The *log odds ratio function* $\log OR_f$ depends only on the log-density $\varphi = \log f \in \Phi$ and will be denoted by $\psi_\varphi = \log OR_f$, i.e.

$$\psi_\varphi(x, y \mid x', y') \;=\; \varphi(x, y) + \varphi(x', y') - \varphi(x, y') - \varphi(x', y)\,.$$

For any pair of *fixed reference values* $x^\circ \in \Omega_X$ and $y^\circ \in \Omega_Y$ the odds resp. log odds ratio function is already determined by its partial function $OR_f^\circ = OR_f(\text{-}, \text{-} \mid x^\circ, y^\circ) : \Omega \longrightarrow (0, \infty)$ resp. $\psi_\varphi^\circ = \log OR_f^\circ$ since

$$\psi_\varphi(x,y\,|\,x',y') \;=\; \psi^\circ_\varphi(x,y) + \psi^\circ_\varphi(x',y') - \psi^\circ_\varphi(x,y') - \psi^\circ_\varphi(x',y)\;.$$

Now for any $P \in \mathscr{P}$ its odds resp. log odds ratio function is defined as $OR_f$ resp. $\psi_\varphi$ for $f = dP/d\pi \in \mathscr{F}$ resp. $\varphi = \log dP/d\pi \in \Psi$. Strictly speaking, the odds ratio function of $P$ is only unique modulo the equivalence relation $\underset{\pi}{\sim}$ of $\pi$-almost-sure equality, i.e. $g \underset{\pi}{\sim} h$ if and only if $g = h$ $\pi$-almost surely.

It is convenient to view any $P \in \mathscr{P}$ as a *joint distribution* of a pair $(X,Y)$ of random elements defined on a probability space $(\Omega_0, \mathscr{B}_0, P_0)$ with values in $\Omega$, i.e. $P$ is the image measure of $P_0$ under the mapping $(X,Y)\colon \Omega_0 \longrightarrow \Omega$. Following usual practice, we extend concepts defined for *probability measures* to *random elements* via their distribution, e.g. the odds ratio function for $(X,Y)$ is that of their joint distribution: $OR(X,Y) = OR(P)$.

## 2.2 Marginal and Conditional Distributions

If $f \in \mathscr{F}$ is the *joint* density of $(X,Y)$ then the *marginal* densities of $X$ and $Y$ are denoted by $f^X(x) = \int f(x,y)\, d\pi_Y(y) > 0$ and $f^Y(y) = \int f(x,y)\, d\pi_X(x) > 0$. Since $f^X$ and $f^Y$ are finite almost surely we may choose $f$ such that $f^X$ and $f^Y$ are both finite. The *conditional* density $f^{|X} \in \mathscr{F}$ of $Y$ given $X$ - defined by $f^{|X}(y\,|\,x) = f(x,y)/f^X(x)$ - is positive and evidently has the *same* odds ratio function as the unconditional density $f$. This also holds for the *conditional* density $f^{|Y} \in \mathscr{F}$ of $X$ given $Y$, and hence $OR(f^{|X}) = OR(f) = OR(f^{|Y})$.

## 2.3 Change of Dominating Measures

An important property of the odds ratio function is its invariance with respect to dominating $\sigma$-finite measures. More precisely, suppose that the marginal distribution $\pi_X$ resp. $\pi_Y$ is dominated by a $\sigma$-finite measure $\nu_X$ resp. $\nu_Y$ with positive density $\delta_X = d\pi_X/d\nu_X$ resp. $\delta_Y = d\pi_Y/d\nu_Y$. Then $P \in \mathscr{P}$ is dominated by the product measure $\nu = \nu_X \times \nu_Y$ and $f_\delta(x,y) = f(x,y)\, \delta_X(x)\, \delta_Y(y)$ defines a $\nu$-density of $P$. Hence the odds ratio function may also be expressed with $f$ replaced by $f_\delta$

$$OR_f(x,y\,|\,x',y') \;=\; [f_\delta(x,y)\cdot f_\delta(x',y')]\cdot[f_\delta(x,y')\cdot f_\delta(x',y)]^{-1}$$

For common dominating measures - e.g. Lebesgue's resp. the counting measure for continuous resp. simple random variables - this representation is typically used to *define* the odds ratio. However the product measure $\pi$ of the marginal distributions is a *canonical* choice for a dominating measure and moreover allows a definition of the odds ratio in situations where no densities with respect to the above standard measures are available (e.g. for a multivariate normal distribution with singular covariance matrix, cf. **4.4**).

## 2.4 One-to-one Transformations

The odds ratio of one-to-one transformations $U = g(X)$ and $V = h(Y)$ are easily obtained as the odds ratio of $X$ and $Y$ evaluated at the corresponding inverse points. More precisely, let $g : \Omega_X \longrightarrow \Omega_U$ resp. $h : \Omega_Y \longrightarrow \Omega_V$ be measurable (with respect to some $\sigma$-algebras $\mathscr{B}_U$ resp. $\mathscr{B}_V$) having measurable inverses $g^{-1}$ resp. $h^{-1}$. Consider the space $\Omega' = \Omega_U \times \Omega_V$ equipped with the product $\pi' = \pi_U \times \pi_V$ of the distributions $\pi_U$ and $\pi_V$ of $U$ and $V$ and the corresponding product $\sigma$-algebra $\mathscr{B}'$. Then a positive $\pi'$-density of the joint distribution $P'$ of $(U, V)$ is given by $f'(u,v) = f(g^{-1}(u), h^{-1}(v))$, and hence the odds ratio function of $(U, V)$ is

$$OR_{f'}(u, v \mid u', v') \;=\; OR_f\big(g^{-1}(u), h^{-1}(v) \mid g^{-1}(u'), h^{-1}(v')\big)\,.$$

***Random vectors.*** A typical example for transformations of random vectors $X$ and $Y$ are affine mappings $g(x) = A(x - b)$ and $h(y) = C(y - d)$ with appropriate quadratic non-singular matrices $A$, $C$ and vectors $b$, $d$. If the covariance matrices $\Sigma_X$ and $\Sigma_Y$ of $X$ and $Y$ are non-singular, then the choice of $A = \Sigma_X^{-1/2}$, $b = \mathrm{E}(X)$, $C = \Sigma_Y^{-1/2}$, $d = \mathrm{E}(Y)$ yield *standardized* random vectors $U$ and $V$ (i.e. with zero expectation and identity covariance matrix). Hence the odds ratio function of $(X, Y)$ is easily obtained from those of the *standardized pair* $(U, V)$.

## 2.5 Vector Spaces of Random Variables

Let $\mathscr{L}^0$ resp. $\mathscr{L}^0_X$, $\mathscr{L}^0_Y$ denote the vector space of all random variables from $\Omega$ resp. $\Omega_X$, $\Omega_Y$ into $\mathbb{R}$. For any $\varphi \in \Phi \subset \mathscr{L}^0$ the log odds ratio function $\psi^\circ_\varphi \in \mathscr{L}^0$ for the reference pair $(x^\circ, y^\circ)$ will now be characterized as a projection of $\varphi$ onto the linear subspace

$$\mathscr{A}^0 \;=\; \{\, \xi \in \mathscr{L}^0 \mid \xi(x^\circ, y) = \xi(x, y^\circ) = 0 \text{ for all } x,\, y \,\}.$$

For any pair $\eta_X \in \mathscr{L}^0_X$, $\eta_Y \in \mathscr{L}^0_Y$ the sum $\eta_X(x) + \eta_Y(y)$ defines a random variable on $\Omega$, denoted by $\eta_X + \eta_Y$, and the *marginal subspace* $\mathscr{M}^0$ is the space of all such sums

$$\mathscr{M}^0 \;=\; \{\, \eta_X + \eta_Y \mid \eta_X \in \mathscr{L}^0_X,\, \eta_Y \in \mathscr{L}^0_Y \,\}.$$

The space $\mathscr{L}^0$ is the direct sum $\mathscr{M}^0 \oplus \mathscr{A}^0$, i.e. any $\xi \in \mathscr{L}^0$ may uniquely be written as $\xi = \eta + \psi$ with $\eta \in \mathscr{M}^0$ and $\psi \in \mathscr{A}^0$. The unique *projection* $\eta$ resp. $\psi$ of $\xi$ onto $\mathscr{M}^0$ resp. $\mathscr{A}^0$ - is given by

$$\eta(x,y) \;=\; \xi(x,y^\circ) + \xi(x^\circ,y) - \xi(x^\circ,y^\circ)\;,$$
$$\psi(x,y) \;=\; \xi(x,y) + \xi(x^\circ,y^\circ) - \xi(x,y^\circ) - \xi(x^\circ,y)\;.$$

Hence the log odds ratio function may be written as a projection $\psi^\circ_\varphi = \Pi(\varphi\,|\,\mathscr{A}^0)$, where $\Pi(\text{-}\,|\,\mathscr{H})$ denotes the projection operator for a subspace $\mathscr{H}$.

## 2.6 The Space of Integrable Log-Densities

For a further representation of the log odds ratio function we consider the class $\mathscr{P}^1$ of probability measures $P$, for which the log-density $\varphi = \log dP/d\pi$ is *$\pi$-integrable*. Let $\mathscr{L}^1 \subset \mathscr{L}^0$ be the subspace of $\pi$-integrable random variables, $\mathscr{F}^1 = \{f \in \mathscr{F}\,|\,\log f \in \mathscr{L}^1\}$ the class of densities for $\mathscr{P}^1$, and $\Phi^1 = \Phi \cap \mathscr{L}^1$ the subspace of integrable log-densities. We show that for $\varphi \in \Phi^1$ its log odds ratio function $\psi_\varphi \in \mathscr{L}^1$ is uniquely determined by the projection of $\varphi$ onto a suitable linear subspace $\mathscr{A}^1 \subset \mathscr{L}^1$. Strictly speaking, this results holds only $\pi$-almost surely, i.e. in the quotient space $\mathscr{L}^1/\underset{\pi}{\sim}$ of equivalence classes with respect to $\underset{\pi}{\sim}$. Using the *marginal functions*

$$\xi^X(x) = \int \xi(x,y)\; d\pi_Y(y), \qquad \xi^Y(y) = \int \xi(x,y)\; d\pi_X(x) \qquad \text{for } \xi \in \mathscr{L}^1$$

the subspace is defined by $\mathscr{A}^1 = \{\, \xi \in \mathscr{L}^1 \,|\, \xi^X = 0,\, \xi^Y = 0 \,\}$. Considering the corresponding subspaces $\mathscr{L}^1_X \subset \mathscr{L}^0_X$ and $\mathscr{L}^1_Y \subset \mathscr{L}^0_Y$ of integrable functions, the *marginal subspace* of integrable functions is

$$\mathscr{M}^1 \;=\; \{\, \eta_X + \eta_Y \,|\, \eta_X \in \mathscr{L}^1_X,\, \eta_Y \in \mathscr{L}^1_Y \,\}.$$

Since the sum $\eta_X + \eta_Y$ does not uniquely determine its components we define further subspaces

$$\mathscr{M}^1_X \;=\; \{\, \beta \in \mathscr{L}^1_X \,|\, \int \beta\, d\pi_X = 0 \,\}\;, \qquad \mathscr{M}^1_Y \;=\; \{\, \gamma \in \mathscr{M}^1_Y \,|\, \int \gamma\, d\pi_Y = 0 \,\}.$$

Now for any $\eta \in \mathscr{M}^1$ there exists *unique* respresentation

$$\eta = \alpha + \beta + \gamma\;, \qquad \text{i.e.} \qquad \eta(x,y) = \alpha + \beta(x) + \gamma(y) \quad \text{for all } x, y$$

with $\alpha \in \mathbb{R}$, $\beta \in \mathscr{M}^1_X$ and $\gamma \in \mathscr{M}^1_Y$, namely

$$\alpha = \int \eta\, d\pi, \qquad \beta = \eta^X - \alpha, \qquad \gamma = \eta^Y - \alpha.$$

Now any $\xi \in \mathscr{L}^1$ with *finite* $\xi^X$ and $\xi^Y$ may uniquely be written as $\xi = \eta + \psi$ with $\eta \in \mathscr{M}^1$ and $\psi \in \mathscr{A}^1$ given by

$$\eta(x,y) \;=\; \xi^X(x) + \xi^Y(y) - \int \xi\, d\pi\;,$$
$$\psi(x,y) \;=\; \xi(x,y) - \xi^X(x) - \xi^Y(y) + \int \xi\, d\pi\;. \tag{3}$$

The uniqueness of the representation is obtained as follows. From $\xi = \eta + \psi$ and $\psi^X = 0$ we get $\xi^X = \eta^X + \psi^X = \eta^X$, and similiar $\xi^Y = \eta^Y$ and $\int \xi \, d\pi = \int \eta \, d\pi$. Hence, by the above decomposition $\eta = \alpha + \beta + \gamma = \eta^X + \eta^Y - \int \eta \, d\pi$ we get the first equation of (3) which in turn implies the second in view of $\xi = \eta + \psi$.

The unique $\eta$ resp. $\psi$ in the decomposition $\xi = \eta + \psi$ is the *projection* of $\xi$ onto the respective subspace. The *finiteness* of $\xi^X$ and $\xi^Y$ is not crucial, since both are finite $\pi$-almost surely. Hence a corresponding decomposition holds in the quotient space $\mathscr{L}^1/\underset{\pi}{\sim}$.

For any $P \in \mathscr{P}^1$ we can choose a log-density $\varphi \in \Phi^1$ with finite $\varphi^X$ and $\varphi^Y$. From the easily verified identities

$$\Pi(\varphi \,|\, \mathscr{A}^1) \;=\; \Pi\big(\Pi(\varphi \,|\, \mathscr{A}^0) \,|\, \mathscr{A}^1\big)\,, \qquad \Pi(\varphi \,|\, \mathscr{A}^0) \;=\; \Pi\big(\Pi(\varphi \,|\, \mathscr{A}^1) \,|\, \mathscr{A}^0\big)$$

we conclude that the log odds ratio function $\psi^\circ_\varphi = \Pi(\varphi \,|\, \mathscr{A}^0)$ determines and is determined by the projection $\Pi(\varphi \,|\, \mathscr{A}^1)$ of the log-density $\varphi$ onto the space $\mathscr{A}^1$. Note that this projection does not refer to a reference pair $(x^\circ, y^\circ)$.

Finally, we show that for any $f \in \mathscr{F}^1$ the *marginal* log-densities $\log f^X$ and $\log f^Y$ are integrable, which in turn implies the integrability of the *conditional* log-densities $\log f^{|X}$ and $\log f^{|Y}$. Indeed for any $\varphi \in \mathscr{L}^1$ and $\xi = \exp(\varphi)$ Jensen's inequality gives

$$\log \xi^X(x) \;=\; \log\big[\textstyle\int \xi(x,y)\, d\pi(y)\big] \;\geq\; \int \log \xi(x,y)\, d\pi(y) \;=\; \varphi^X(x)$$

and from $\log(a) \leq a - 1$ we get the fundamental inequality

$$\varphi^X \;\leq\; \log \xi^X \;\leq\; \xi^X - 1\,. \tag{4}$$

Hence $\log \xi^X$ is $\pi_X$-integrable, provided $\xi$ is $\pi$-integrable (e.g. for $\xi \in \mathscr{F}^1$), which by duality also implies that $\log \xi^Y$ is $\pi_Y$-integrable.

## 3. A Characterization of Association

This section contains our main result, the characterization of association in terms of the odds ratio function under mild conditions. First we will show that any joint distribution $P \in \mathscr{P}$ of $(X, Y)$ with given marginal distributions $\pi_X$ and $\pi_Y$ is uniquely determined by its odds ratio function $OR(P)$. In order to evaluate Kullback-Leibler information in the proof requires mild integrability conditions, which in particular hold if the log-density of $P$ is $\pi$-integrable (i.e. $P \in \mathscr{P}^1$) or if either $\Omega_X$ or $\Omega_Y$ is *finite*.

Conversely, for a given $\psi \in \mathscr{A}^1$ we provide sufficient conditions for the existence of a $\pi$-integrable log-density $\varphi = \log(dP/d\pi) \in \Phi^1$ such that the joint distribution

$P \in \mathscr{P}^1$ has the pre-specified marginal distributions $\pi_X$ and $\pi_Y$ and the odds ratio function of $P$ corresponds to $\psi$, i.e. $\psi$ is the projection of $\varphi$ onto $\mathscr{A}^1$. The density $\varphi$ will be obtained as a limit of an iterative marginal fitting procedure.

The results of uniqueness and existence imply that the odds ratio function completely characterizes the association, i.e. the information in the joint distribution which is not contained in the marginal distributions. For this reason the odds ratio function will also be referred to as the *association function.*

### 3.1 Uniqueness

Suppose that $P_1$, $P_2 \in \mathscr{P}$ have the same marginal distributions $\pi_X$, $\pi_Y$ and a common odds ratio function $OR^\circ(P_1) = OR^\circ(P_2)$. We want to show that $P_1$ equals $P_2$. For the densities $f_i = dP_i/d\pi \in \mathscr{F}$ $(i = 1, 2)$ we have $f_i^X = dP_i^X/d\pi_X$ and $f_i^Y = dP_i^Y/d\pi_Y$ so that $P_i$ has the marginal distributions $\pi_X$ and $\pi_Y$ if and only if

$$f_i^X = 1\,, \qquad f_i^Y = 1 \qquad P_i\text{-almost surely.} \tag{5}$$

The projections of the log-densities $\varphi_i = \log f_i \in \Phi$ onto $\mathscr{A}^0$ are the log odds ratio functions and thus coincide for $i = 1,2$. Hence $\varphi_1 - \varphi_2$ equals its projection onto the marginal space $\mathscr{M}^0$ and may be written as

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

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

To establish $P_1 = P_2$ we show that the *Kullback-Leibler information* (Kullback 1959)

$$I(f_1, f_2) := \int f_1 \log(f_1/f_2)\, d\pi = \int (\varphi_1 - \varphi_2)\, dP_1 = \int (\beta_1 + \gamma_1)\, dP_1$$

is zero, which implies $f_1 = f_2$ $\pi$-almost surely, and hence $P_1 = P_2$ (cf. **A.1**). Now $I(f_1, f_2) = 0$ can be shown using the decomposition

$$I(f_1, f_2) = \int (\beta_1 + \gamma_1)\, dP_1 = \int \beta_1\, dP_1 + \int \gamma_1\, dP_1\,, \tag{6}$$

which holds under integrability conditions given in the following theorem.

***Uniqueness-Theorem:*** *Suppose $P_1$, $P_2 \in \mathscr{P}$ are joint distributions both with common marginal distributions $\pi_X$ resp. $\pi_Y$ and with a common odds ratio function $OR^\circ(P_1) = OR^\circ(P_2)$ with respect to an arbitrary reference pair $(x^\circ, y^\circ)$. Let $\varphi_i = \log(dP_i/d\pi)$ denote the log-densities for $i = 1,2$. Then any of the two integrability conditions implies $P_1 = P_2$:*

*(i)* $(\varphi_1 - \varphi_2)(\text{-}\,, y^\circ)$ *is $\pi_X$-integrable,*

*(ii)* $(\varphi_1 - \varphi_2)(x^\circ, \text{-})$ *is $\pi_Y$-integrable.*

*Note that (i) resp. (ii) always hold if $\Omega_X$ resp. $\Omega_Y$ is finite.*

***Corollary:*** *If the log-densities $\varphi_1$, $\varphi_2$ are $\pi$-integrable, then there exists $(x^\circ, y^\circ)$ such that (i) and (ii) hold.*

### 3.2 Existence

For a given function $\psi^\circ \in \mathscr{A}^0$ we now wish to construct a joint distribution $P \in \mathscr{P}^1$ with given marginal distributions $\pi_X$ and $\pi_Y$ such that the odds ratio function of $P$ corresponds to $\psi^\circ$, i.e. $\psi^\circ$ is the projection of the log-density $\varphi = \log(dP/d\pi)$ onto $\mathscr{A}^0$. As one might expect the construction of $P$ requires certain restrictions on $\psi^\circ$, the first being that $\psi^\circ$ is *$\pi$-integrable*, which will in turn produce a $\pi$-integrable log-density $\varphi$. Since $\psi^\circ \in \mathscr{L}^1$ is uniquely determined by its projection $\psi = \Pi(\psi^\circ \,|\, \mathscr{A}^1)$ via $\psi^\circ = \Pi(\psi \,|\, \mathscr{A}^0)$ we may alternatively start with a given $\pi$-integrable function $\psi \in \mathscr{A}^1$ and look for a log-density $\varphi$ with $\psi = \Pi(\varphi \,|\, \mathscr{A}^1)$, i.e. $\varphi$ is of the form $\varphi = \eta + \psi$ with $\eta \in \mathscr{M}^1$. We prove that such an $\eta \in \mathscr{M}^1$ exists if the following *existence conditions* hold:

(EC1) *There exists $\bar\beta \in \mathscr{L}^1_X$ such that $\exp(\bar\beta + \psi)$ is $\pi$-integrable.*

(EC2) *There exists $\bar\gamma \in \mathscr{L}^1_Y$ such that $\exp(\bar\gamma + \psi)$ is $\pi$-integrable.*

Passing from $\bar\beta$ to $\bar\beta - \int \bar\beta \, d\pi_X \in \mathscr{M}^1_X$ we may assume $\bar\beta \in \mathscr{M}^1_X$ in (EC1) and, by duality, $\bar\gamma \in \mathscr{M}^1_Y$ in (EC2). Equivalent formulations in terms of the marginal functions of $q = \exp\psi$ are the following:

(EC1)$'$ $\log q^X = \log\big[(\exp\psi)^X\big]$ *is $\pi_X$-integrable.*

(EC2)$'$ $\log q^Y = \log\big[(\exp\psi)^Y\big]$ *is $\pi_Y$-integrable.*

Suppose first that (EC1) holds and set $\xi = \exp(\bar\beta + \psi)$. Then

$$\xi^X(x) \;=\; \int \exp[\bar\beta(x) + \psi(x,y)]\, d\pi_Y(y) \;=\; \exp[\bar\beta(x)] \cdot q^X(x)$$

and $\xi^X$ is finite $\pi_X$-almost surely. Hence $\log q^X = \log \xi^X - \bar\beta$ is $\pi_X$-integrable by (4) which yields (EC1)$'$. Conversely, if (EC1)$'$ holds, then $q^X$ is finite $\pi_X$-almost surely and there exists a version $\bar\beta \in \mathscr{L}^1_X$ such that $\bar\beta = -\log q^X$ $\pi_X$-almost surely. Hence

$$\int \exp[\bar\beta(x) + \psi(x,y)]\, d\pi(x,y) \;=\; \int \exp[-\log q^X(x)] \cdot q^X(x)\, d\pi_X(x) \;=\; 1$$

which proves (EC1). By duality we get the equivalence of (EC2) and (EC2)$'$.

A stronger version of both (EC1) and (EC2) is obtained by replacing $\bar\beta$ or $\bar\gamma$ with the constant zero function , i.e.

(EC3) $\exp(\psi)$ *is $\pi$-integrable.*

This condition trivially holds for *bounded* $\psi$, which covers an important range of applications in practice where the sample space $\Omega$ typically is a *compact* subset of $\mathbb{R}^k$ and $\psi$ is a *continuous* function. Note that for *finite* $\Omega_X$ resp. $\Omega_Y$ the condition (EC3) is even equivalent to (EC1) resp. (EC2). However, for $\Omega = \mathbb{R}^k$ we will provide an example for $\pi$ (cf. **4.4**) , where (EC) fails but (EC1) and (EC2) hold.

The existence conditions (EC1) and (EC2) are *sufficient* for the existence of the wanted joint distribution:

***Existence-Theorem:*** *For any $\psi \in \mathscr{A}^1$ the existence conditions (EC1) and (EC2) imply the existence of a joint distribution $P \in \mathscr{P}^1$ with given marginal distributions $\pi_X$ and $\pi_Y$ and $\pi$-integrable log-density $\varphi = \log(dP/d\pi)$ such that $\psi = \Pi(\varphi \,|\, \mathscr{A}^1)$.*

Although the existence conditions will turn out (cf. **4.3**) to be weak enough for important applications, they are not *necessary* for the existence of $P$ - at least not for *binary* $Y$ (cf. **1**).

The desired log-density $\varphi$ will be constructed as a limit of an iterative procedure of so-called *marginal fittings*. Since $q^X$ is finite $\pi_X$-almost surely by (EC1)$'$, we will take a version of $\psi$ modulo $\underset{\pi}{\sim}$ such that $q^X$ is finite, and by duality from (EC2)$'$ we assume $q^Y$ to be finite, too. Furthermore, the functions given by (EC1) and (EC2), which will be chosen such that $\bar{\beta} \in \mathscr{M}_X^1$ and $\bar{\gamma} \in \mathscr{M}_Y^1$ hold. Denote

$$\bar{\eta} \;=\; \bar{\beta} + \bar{\gamma} \;\in \mathscr{M}^1, \qquad \bar{\varphi} \;=\; \bar{\eta} + \psi \;=\; \bar{\beta} + \bar{\gamma} + \psi \;\in \mathscr{L}^1 .$$

with marginal functions $\bar{\varphi}^X = \bar{\eta}^X = \bar{\beta}$ and $\bar{\varphi}^Y = \bar{\eta}^Y = \bar{\gamma}$. Since $\bar{\varphi}$ and the desired $\varphi$ both have the same projection $\psi$ in $\mathscr{A}^1$, $\varphi$ is necessarily of the form $\varphi = \eta + \bar{\varphi}$ with $\eta \in \mathscr{M}^1$. Hence we are looking for an element $\eta \in \mathscr{M}^1$ such that

$$\bar{f}(\eta) \;=\; \exp(\eta + \bar{\varphi}) \;=\; \exp(\eta + \bar{\beta} + \bar{\gamma} + \psi).$$

is the $\pi$-density of a probability measure $P$ with marginals $\pi_X$ and $\pi_Y$, i.e. $\bar{f}^X(\eta) = 1$ and $\bar{f}^Y(\eta) = 1$ almost surely. This element $\eta$ will be constructed as a limit of a sequence $\eta_n \in \mathscr{M}^1$. The definition of this sequence and the proof of its convergence is divided into several parts.

### 3.3 Marginal Fitting

As a first step we define an operator $M_X$ for joint distributions $P \in \mathscr{P}^1$ given by a $\pi$-density of the form $\bar{f}(\eta) \in \mathscr{F}^1$ with $\eta \in \mathscr{M}^1$. The conditional density of $P$ given $X$ is of the same form, namely

$$\bar{f}(M_X(\eta))(x,y) \;=\; \bar{f}(\eta)(x,y) \,/\, \bar{f}^X(\eta)(x) \qquad \text{with}$$

$$M_X(\eta) \;=\; \eta - \log \bar{f}^X(\eta) \;\in\; \mathscr{M}^1$$

where $\log \bar{f}^X(\eta) \in \mathscr{L}_X^1$ follows from (4). The corresponding joint distribution, denoted by $M_X(P)$, has the same odds ratio function as $P$. And the marginal distribution of $X$ under $M_X(P)$ is $\pi_X$ since the marginal density of $\bar{f}(M_X(\eta))^X$ of $X$ is identically 1. Hence the *marginal fitting operator* $M_X$ adjusts the distribution of $X$ under $P$ to the given marginal $\pi_X$. Some basic properties of the operator $M_X$ are as follows.

For any representation $\eta = \eta_X + \eta_Y$ with $\eta_X \in \mathscr{L}_X^1$ and $\eta_Y \in \mathscr{L}_Y^1$ we obtain $\bar{f}^X(\eta) = \exp(\eta_X)\, \bar{f}^X(\eta_Y)$ and hence $M_X(\eta) = M_X(\eta_Y)$. In particular, the decompositions

$$\eta \;=\; \eta^X + \gamma\,, \qquad \gamma \;=\; \eta^Y - \int \eta\, d\pi \;=\; \Pi(\eta \,|\, \mathscr{M}_Y^1)$$

$$\eta \;=\; \beta + \eta^Y\,, \qquad \beta \;=\; \eta^X - \int \eta\, d\pi \;=\; \Pi(\eta \,|\, \mathscr{M}_X^1)$$

yield $$M_X(\eta^Y) \;=\; M_X(\eta) \;=\; M_X(\gamma) \;=\; \gamma - \log \bar{f}^X(\gamma). \tag{7}$$

Hence the projection of $M_X(\eta)$ onto $\mathscr{L}_X^1$ is

$$\big[M_X(\eta)\big]^X \;=\; -\log \bar{f}^X(\gamma)\,, \tag{8}$$

and (4) applied to $\xi = \exp\{\gamma + \bar{\varphi}\}$ yields the important inequality

$$1 - \bar{f}^X(\gamma) \;\le\; \big[M_X(\eta)\big]^X \;\le\; -\bar{\varphi}^X \;=\; -\bar{\beta}\,. \tag{9}$$

Furthermore, $M_X$ is idempotent, i.e. $M_X M_X = M_X$.

Now take an arbitrary $\eta_0 \in \mathscr{M}^1$ such that $\bar{f}(\eta_0) \in \mathscr{F}^1$ is the $\pi$-densitiy of a joint distribution $P_0 \in \mathscr{P}^1$. Iterating the operator $M_X$ and the dual operator $M_Y(\eta) = \eta - \log \bar{f}^Y(\eta)$ we obtain a sequence $\eta_n \in \mathscr{M}^1$, recursively given by

$$\eta_{n+1} = M_X\big(M_Y(\eta_n)\big)\,, \qquad n \ge 0,$$

and an accompanying sequence $\tilde{\eta}_n \in \mathscr{M}^1$, namely

$$\tilde{\eta}_n = M_Y(\eta_n) \qquad \text{with} \qquad \eta_{n+1} = M_X(\tilde{\eta}_n).$$

The sequence $(\eta_n)_{n \ge 0}$ will be called the *marginal fitting sequence* with starting value $\eta_0$ and is a straightforward generalization of the one used by Sinkhorn (1967) for simple random variables $X$ and $Y$. It will be shown to converge to a limit $\eta$ which provides the density $\bar{f}(\eta)$ of the wanted distribution $P$.

### 3.4 Maximizing Functional and Convergence Theorem

As we will see later, the desired joint distribution $P$ may also be obtained by maximizing the real-valued functional $\ell$ defined on $\mathscr{M}^1$ by

$$\ell(\eta) \;=\; \int [\eta - \bar{f}(\eta)]\, d\pi \;=\; \int l\big(\eta(x,y) \,|\, \bar{\varphi}(x,y)\big)\, d\pi(x,y) \qquad \text{with}$$

$$l(u \,|\, v) \;=\; u - \exp(u+v) \qquad \text{for } u, v \in \mathbb{R}.$$

The functional $\ell$ generalizes the function considered by Haberman (1974, Theorem 2.6) for (conditional) Poisson distributed $Y$. Note that $\ell(\eta)$ is finite for $\pi$-integrable $\bar{f}(\eta)$ and $-\infty$ otherwise. If $\bar{f}(\eta)$ is a probability density, we get

$$\ell(\eta) \;=\; \int \eta \, d\pi - 1 \qquad\qquad \text{for} \qquad \bar{f}(\eta) \in \mathcal{F}.$$

Maximizing $\ell(\eta)$ is equivalent to minimizing the Kullback-Leibler information $I\big(1,\bar{f}(\eta)\big)$ - the constant density 1 represents the product measure $\pi_X \times \pi_Y$ - since

$$\ell(\eta) \;=\; -\, I\big(1, \bar{f}(\eta)\big) - \int \psi \, d\pi - 1 \qquad \text{for} \qquad \bar{f}(\eta) \in \mathcal{F}.$$

For fixed $v$ the function $l(u \,|\, v)$ is strictly concave in $u$ and attains its unique maximum for $u = -\,v$. Hence $\ell$ is bounded from above

$$\ell(\eta) \;\le\; \ell(-\bar{\varphi}) \;=\; 1 - \int \bar{\varphi} \, d\pi \tag{10}$$

and strictly concave modulo $\underset{\pi}{\sim}$ on the convex set

$$\mathcal{M}^{1\pi} := \{\, \eta \in \mathcal{M}^1 \mid \bar{f}(\eta) \text{ is } \pi\text{-integrable} \,\} \;=\; \{\, \eta \in \mathcal{M}^1 \mid \ell(\eta) < \infty \,\}\,.$$

For any $\eta \in \mathcal{M}^{1\pi}$ we now establish an important connection between the functional $\ell$ and the marginal operator $M_X$ :

$$\ell(\eta) \;\le\; \ell\big(M_X(\eta)\big)\,, \tag{11}$$

$$\ell(\eta) \;=\; \ell\big(M_X(\eta)\big) \qquad \Leftrightarrow \qquad \bar{f}^X(\eta) = 1 \;\text{ resp. }\; \eta = M_X(\eta) \;\text{ almost surely.} \tag{12}$$

From the definitions we get

$$\begin{aligned} \ell\big(M_X(\eta)\big) \;&=\; \int \big(\eta - \log \bar{f}^X(\eta) - 1\big)\, d\pi \qquad\qquad \text{and} \\ \ell\big(M_X(\eta)\big) - \ell(\eta) \;&=\; \int \big[\bar{f}(\eta) - \log \bar{f}^X(\eta) - 1\big]\, d\pi \\ &=\; \int \big[\bar{f}^X(\eta) - \log \bar{f}^X(\eta) - 1\big]\, d\pi_X\,. \end{aligned}$$

Hence the results follow from the inequality $\log(a) \le a - 1$, where equality holds if and only if $a = 1$.

Now we can establish (cf. **A.3**) a convergence result for the iterated marginal fitting procedure from which the existence theorem will be derived (in **A.4**).

***Convergence-Theorem:*** *Let* $(\eta_n \in \mathcal{M}^1)_{n \ge 0}$ *be any marginal fitting sequence with starting value satisfying* $\int \bar{f}(\eta_0)\, d\pi = 1$. *If the existence conditions (EC1) and (EC2) hold, then any subsequence* $(\eta_{m(n)})$ *contains a further subsequence* $(\eta_{m'(n)})$ *which converges pointwise as well as in the mean to an element* $\eta \in \mathcal{M}^1$ *having the properties*

*(i)* $\quad M_X(\eta) = \eta\,, \qquad M_Y(\eta) = \eta \qquad$ *$\pi$-almost surely*

*(ii)* $\quad \int \bar{f}(\eta)\, d\pi = 1\,.$

*The limit $\eta$ is $\pi$-almost surely independent of the chosen subsequence $(\eta_{m(n)})$ and independent of $\eta_0$, i.e. for two starting values $\eta_{10}$, $\eta_{20}$ and any subsequences $(\eta_{1m(n)})$, $(\eta_{2m(n)})$ the corresponding limits $\eta_1$ and $\eta_2$ coincide $\pi$-almost surely.*

***Corollary**: For any starting value $\eta_0 \in \mathscr{M}^1$ the marginal fitting sequence $(\eta_n)_{n\geq 0}$ converges in the mean to an element $\eta \in \mathscr{M}^1$ and*

*(iii)* $\quad (\ell(\eta_n))_{n\geq 0}$ *is a non-decreasing sequence with limit $\ell(\eta)$.*

*Furthermore $\eta$ is the $\pi$-almost surely unique argument maximizing the functional $\ell$.*

## 4. Applications

The separation of the association from the marginals has an impact on statistical modelling which will now be illustrated. First, semi-parametric *association models* are introduced, which specify a model for the association function of $X$ and $Y$ but leave the marginal distributions completely arbitrary. Parametrizations of the joint and conditonal densities in terms of odds ratios are given which serve as a framework for further statistical analysis. Motivated by standard (regression) models we specialize to *log-bilinear associations* for which the existence conditions (EC1) and (EC2) may easily be checked using cumulant-generating functions. Second, we look at random *vectors* $X$ and $Y$ and characterize the *joint* normal distributions for $(X,Y)$ as those with normal marginals and a *log-bilinear association.* This emphasizes the importance of log-bilinear association models as a semi-parametric generalization of joint multivariate normality.

### 4.1 Association Models

From the uniqueness theorem we conclude that the joint distribution $P$ of $(X,Y)$ is determined by their marginal distributions $\pi_X$ and $\pi_Y$ and their association, i.e. the odds ratio function *OR*. If the focus of an investigation is on the association between $X$ and $Y$ rather than on the marginal distributions, then the appropriate models are semi-parametric *accociation models* or *odds ratio models*, which only specify the odds ratio function and leave the marginals completely arbitrary. The corresponding model for the density $f$ with respect to the product $\pi = \pi_X \times \pi_Y$ of the marginals (or any other product measure, cf. **2.3**) may be written in terms of the log-density $\varphi = \log f$ and the log odds ratio $\psi^\circ = \log OR$ as (cf. **2.5**)

$$\varphi(x,y) \;=\; \alpha^\circ + \beta^\circ(x) + \gamma^\circ(y) + \psi^\circ(x,y)\,.$$

Here $\psi^\circ$ is restricted to a subspace $\Psi^\circ \subset \mathscr{A}^0$ specifying the model and $\alpha^\circ \in \mathbb{R}$ as well as the functions $\beta^\circ \in \mathscr{L}_X^0$ and $\gamma^\circ \in \mathscr{L}_Y^0$ are completely arbitrary. Identifiability may be achieved through the constraints $\beta^\circ(x^\circ) = 0$ and $\gamma^\circ(y^\circ) = 0$, which will be assumed here. Note that the definition of $\mathscr{A}^0$ already imposes the constraints $\psi^\circ(x, y^\circ) = 0 = \psi^\circ(x^\circ, y)$ for all $x, y$. The model space $\Psi^\circ$ is typically parametrized by means of a parameter $\theta \in \Theta$, i.e. $\Psi^\circ = \{ \psi_\theta^\circ \mid \theta \in \Theta \}$. Assuming the log-density $\psi$ to be $\pi$-integrable, the model can be rewritten as (cf. **2.6**)

$$\varphi(x,y) \;=\; \alpha + \beta(x) + \gamma(y) + \psi(x,y)$$

with $\psi$ restricted to a subspace $\Psi \subset \mathscr{A}^1$, and arbitrary $\alpha \in \mathbb{R}$, $\beta \in \mathscr{M}_X^1$ and $\gamma \in \mathscr{M}_Y^1$. An important point, however, is that allowing arbitrary $\alpha$, $\beta$ and $\gamma$ is no guarantee that the model is semi-parametric, i.e. does not restrict the marginal distributions $\pi_X$ and $\pi_Y$. This, in fact, requires the existence theorem which explicitly states (under the existence conditions) that the any given marginal distributions may be obtained for suitable values of the parameters $\alpha$, $\beta$ and $\gamma$. Note that $X$ and $Y$ are *independent* under $P$, i.e. $P = \pi_X \times \mu_Y$, if and only if $\psi^\circ = 0$ resp. $\psi = 0$.

In statistical applications the model is often equivalently specified using the *conditional* density $f^{|X}(y \mid x)$ of $Y$ given $X$ by

$$\log f^{|X}(y \mid x) \;=\; \beta_X^\circ(x) + \gamma^\circ(y) + \psi^\circ(x,y)$$

with $\beta_X^\circ(x) = -\log \int \exp\left[\gamma^\circ(y) + \psi^\circ(x,y)\right] d\pi_Y(y)$, or by the log-density *ratio*

$$\log\left(f^{|X}(y \mid x) \,/\, f^{|X}(y^\circ \mid x)\right) \;=\; \gamma^\circ(y) + \psi^\circ(x,y)\ . \tag{13}$$

The dual formulation in terms of the conditional density of $X$ given $Y$ is

$$\log\left(f^{|Y}(x \mid y) \,/\, f^{|Y}(x^\circ \mid y)\right) \;=\; \beta^\circ(x) + \psi^\circ(x,y)\ .$$

The major advantage of association models is that statistical inference concerning the odds ratio function (or its parameter $\theta$) may be drawn from a sample of independent observations $(x_1, y_1), \ldots, (x_n, y_n)$ where each $(x_i, y_i)$ may be taken from any of the two conditional distributions $\mathscr{L}(Y \mid X = x_i)$ and $\mathscr{L}(X \mid Y = y_i)$ or from the joint distribution $\mathscr{L}(X, Y)$.

Returning to the discussion in the introduction, we first specialize to a *finite* sample space $\Omega_Y$ and then briefly look at *log-bilinear association models* for arbitrary $\Omega_Y$ which generalize several standard statistical models. However, statistical inference for these models is outside the scope of this paper and the reader is referred to van der Linde (2002ab) or van der Linde and Osius (2003) for detailed applications.

### 4.2 Output with Finite Range

Let $\Omega_Y$ be finite, say $\Omega_Y = \{0, 1, ..., K\}$. Then $\mathscr{L}(Y \mid X = x)$ has a multinomial distribution $M_{K+1}(1, \pi(x))$ with $K+1$ classes and probabilities $\pi_k(x) = P(Y = k \mid X = x) > 0$. Using the multivariate logistic transformation $\operatorname{logit} \pi_k(x) = \log\big(\pi_k(x) / \pi_0(x)\big)$ of $\pi(x)$, the association model (13) with $y^\circ = 0$ is equivalent to a *logistic regression model*,

$$\operatorname{logit} \pi_k(x) \;=\; \gamma_k^\circ + \psi_k^\circ(x) \,, \qquad\qquad k = 1, ..., K,$$

where the argument $y$ is replaced by an index $k$. Although this regression model is widely used, it has not yet been emphasized (or proved) that the model does *not* restrict the marginal distribution of $Y$ and thus represents an *association* model.

In a *linear* logistic regression model the log odds ratio functions $\psi_k^\circ$ are taken as linear functions of the form $\psi_k^\circ(x) = g(x)^T \theta_k$, where $g(x)$ is an $S$-dimensional vector of so called covariables and $\theta_k \in \mathbb{R}^S$ is an unknown parameter. Although typically the observation $x = (x_1, ..., x_t)$ is itself a finite-dimensional vector, the use of a transformation $g(x)$ instead of $x$ provides more flexible models. For example, $g(x)$ may contain powers $x_1$, $x_1^2$, ... of a continuous component $x_1$ as well as indicator variables $I\{x_2 = l\}$ for levels $l = 1, ..., L$ of a discrete component $x_2$. Introducing indicator variables $h_k(y) = I\{y = k\}$ for all values $k$ of $Y$, the model $\psi_k^\circ(x) = g(x)^T \theta_k$ may equivalently be written as

$$\psi^\circ(x,y) \;=\; \sum_k g(x)^T \theta_k h_k(y) \;=\; g(x)^T \theta\, h(y)$$

where $\theta$ is the corresponding $S \times K$ matrix with columns $\theta_1, ..., \theta_K$. This representation serves as a motivation for the association models considered next.

### 4.3 Log-Bilinear Association Models

Returning from finite to arbitrary $\Omega_Y$, let $U = g(X)$ and $V = h(Y)$ be random vectors given by measurable maps $g : \Omega_X \longrightarrow \mathbb{R}^{k_x}$ and $h : \Omega_Y \longrightarrow \mathbb{R}^{k_y}$. In a semi-parametric *log-bilinear association model* for $(X, Y)$ with respect to $(g, h)$ the log odds ratio function for the joint distribution of $(X, Y)$ is of the form

$$\text{(LBA)} \qquad \psi^\circ(x,y) \;=\; [g(x) - g(x^\circ)]^T A\, [h(y) - h(y^\circ)] \qquad\qquad \text{for all } x,\, y,$$

with a matrix $A$ of parameters, which may additionally be restricted to a suitable subspace of $k_x \times k_y$ matrices. The reference values and the functions $g$, $h$ will typically be chosen such that $g(x^\circ) = 0$ and $h(x^\circ) = 0$ thus making $\psi^\circ$ a bilinear function in the transformed variables $g(x)$ and $h(y)$.

The semi-parametric model (LBA) includes important regressions models, namely *Generalized Linear Models* with canonical links, i.e. linear resp. logistic or log-linear regression where the *conditional* distribution $\mathscr{L}(Y|X)$ is univariate normal resp. binomial or Poisson. Furthermore, we will show in **4.4** that (LBA) always holds if the *joint* distribution of $U$ and $V$ is multivariate *normal*, which serves as a basic model in multivariate analysis. Hence (LBA) generalizes several important statistical models - by removing restrictions of the marginal distributions - and may be considered a standard model for association. Further exploration of these ideas are given in van der Linde (2002ab) or van der Linde and Osius (2003).

The existence conditions for $\psi^\circ$ can be stated in terms of the cumulant generating function of $V$ resp. $U$. First we express the projection $\psi = \Pi(\psi^\circ \,|\, \mathscr{A}^1)$ in terms of the expectations $\mu_U = \mathrm{E}(U)$ and $\mu_V = \mathrm{E}(V)$ (which are assumed to exist) as

$$\begin{aligned} \psi(x,y) &= \psi^\circ(x,y) - \psi^{\circ X}(x) - \psi^{\circ Y}(y) + \int \psi^\circ \, d\pi \\ &= [g(x) - \mu_U]^T A \, [h(y) - \mu_V] \, . \end{aligned}$$

Next the marginal function $q^X$ of $q = \exp(\psi)$ can be computed via the moment generating function $m_V$ of $V$ by

$$q^X(x) \;=\; m_V(A^T [g(x) - \mu_U]) \times \exp\{-[g(x) - \mu_U]^T A \, \mu_V\} \, .$$

Finally, we get $\log(q^X)$ in terms of the cumulant generating function $\kappa_V = \log m_V$ of $V$ as

$$\log q^X(x) \;=\; \kappa_V(A^T [g(x) - \mu_U]) - [g(x) - \mu_U]^T A \, \mu_V .$$

Hence the existence condition (EC1)$'$ for $\psi^\circ$ in (LBA) is equivalent to

(EC1)$_A$ *The expectation of* $\kappa_V(A^T [U - \mu_U])$ *exists (i.e. is finite).*

By duality, (EC2)$'$ can be stated in terms of the cumulant generating function $\kappa_U$ of $U$. These conditions are easily checked for standard marginal distributions of $U$ and $V$. As a first example, (EC1)$_A$ holds if $V$ has a multivariate normal distribution $N(\mu_V, \Sigma_V)$ with $\kappa_V(t) = t^T \mu_V + \frac{1}{2} t^T \Sigma_V t$, provided the covariance matrix of $U$ exists. And second, suppose $V$ has a (univariate) Poisson distribution with $\kappa_V(t) = \mu_V[\exp(t) - 1]$. Then (EC1)$_A$ holds, if the value $m_U(A)$ of the moment generating function of $U$ is finite.

## 4.4 Multivariate Normal Distributions

Let us now investigate the situation with multivariate *normal* marginal distributions $\mathscr{L}(X) = \pi_X = N_{k_x}(\mu_x, \Sigma_x)$ and $\mathscr{L}(Y) = \pi_Y = N_{k_y}(\mu_y, \Sigma_y)$ in more detail.

For simplicity we first assume the covariance matrices $\Sigma_x$ and $\Sigma_y$ to be *non-singular* (the singular case is considered later). Then $X$ has a positive density with respect to Lebesgue's measure $\lambda$

$$\log f_\lambda^x(x) \;=\; -\tfrac{1}{2}\,(x-\mu_x)^T \Sigma_x^{-1}(x-\mu_x) + c_x\,, \qquad c_x = -\tfrac{1}{2}\log\big[\,(2\pi)^{k_x}\det(\Sigma_x)\big]\,,$$

and $Y$ has a similiar density $f_\lambda^y$. Our aim is to characterize the *joint multivariate normal* distributions in $\mathscr{P}^1$ in terms of their association functions. The product $\pi = \pi_X \times \pi_Y$ is a multivariate normal distribution of dimension $k = k_x + k_y$ with expectation $\mu = (\mu_x, \mu_y)$ and nonsingular covariance matrix $\operatorname{diag}(\Sigma_x, \Sigma_y)$. Since $\pi$ has a positive density with respect to Lebesgue's measure $\lambda$ we have $\pi \ll \lambda \ll \pi$ and hence the class $\mathscr{P}$ consists of all joint distributions $P$ having a positive density with respect to $\lambda$.

First we take a multivariate normal distribution $P = N_k(\mu, \Sigma) \in \mathscr{P}^1$ and derive its association function. Since $P \in \mathscr{P}^1$ has a $\lambda$-density $f_\lambda$, the covariance matrix $\Sigma$ is nonsingular. The matrices $\Sigma$ and $Q = \Sigma^{-1}$ may be written as

$$\begin{aligned}
\Sigma \;&=\; \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_y \end{bmatrix}, & Q \;&=\; \begin{bmatrix} Q_x & Q_{xy} \\ Q_{xy}^T & Q_y \end{bmatrix}, \\
Q_x \;&=\; \big(\Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{xy}^T\big)^{-1}, & Q_{xy} &= -\,Q_x \Sigma_{xy}\Sigma_y^{-1}, \\
Q_y \;&=\; \Sigma_y^{-1} - \Sigma_y^{-1}\Sigma_{xy}^T Q_{xy}\,.
\end{aligned} \tag{14}$$

From

$$\log f_\lambda(x,y) \;=\;$$
$$-\tfrac{1}{2}\,\big[(x-\mu_x)^T Q_x (x-\mu_x) + 2(x-\mu_x)^T Q_{xy}(y-\mu_y) + (y-\mu_y)^T Q_y (y-\mu_y)\big] + \mathit{const}$$

we get the log odds ratio function of $P$ as

$$\psi_P^\circ(x,y) \;=\; -\,(x - x^\circ)^T Q_{xy}(y - y^\circ) \tag{15}$$

which is *bilinear* for the choices $x^\circ = 0$ and $y^\circ = 0$. The projection of $\psi_P^\circ$ onto the association space $\mathscr{A}^1$ is given by $\psi_P(x,y) = -\,(x-\mu_x)^T Q_{xy}(y-\mu_y)$ and coincides with $\psi_P^\circ$ if the expectations are taken as reference values. The log odds ratio function may be rewritten as

$$\psi_P^\circ(x,y) \;=\; (x - x^\circ)^T C^{-1} B (y - y^\circ) \tag{16}$$

with $B = \Sigma_{xy}\Sigma_y^{-1}$ and $C = Q_x^{-1}$ taken from the *conditional* distributions

$$\mathscr{L}(X \,|\, Y = y) \;=\; N_{k_x}(\mu_x + By\,, C) \qquad \text{for all } y. \tag{17}$$

More generally, the log-affine association model (16) already holds if the *conditional* distributions are given by (17) (i.e. without $Y$ being normal).

We now show (cf. **A.5**) that any joint distribution $P \in \mathscr{P}^1$ is *multivariate normal* if and only if its log odds ratio function is *bilinear*.

***Characterization Theorem:*** *Any joint distribution $P \in \mathscr{P}^1$ with marginal normal distributions $\pi_X = N_{k_x}(\mu_x, \Sigma_x)$ and $\pi_Y = N_{k_y}(\mu_y, \Sigma_y)$ is itself multivariate normal if and only if its log odds ratio function $\psi^\circ$ with respect to the reference values $x^\circ = 0$ and $y^\circ = 0$ is bilinear, i.e. there is a $k_x \times k_y$ matrix $A$ such that $\psi^\circ(x,y) = x^T A y$ for all $x$ and $y$.*
***Note:*** *The existence conditions $(EC1)_A$ and $(EC2)_A$ hold if both $\pi_X$ and $\pi_Y$ are normal.*

The odds ratio function $\exp[\psi^\circ(x,y)] = \exp(x^T A y)$ is in general *not* $\pi$-integrable, i.e. condition (EC3) need not hold when (EC1) and (EC2) are satisfied. Indeed, take $X$ and $Y$ both with standard multivariate normal distribution $N_k(0,I)$ and, for fixed $t \in \mathbb{R}$, choose $A$ such that $A^T A = 2\,t I$ is a multiple of the unit-matrix $I$. Then

$$\int \exp(x^T A y)\, d\pi(x,y) \;=\; \int \Big[ \int \exp(x^T A y)\, d\pi_X(x) \Big]\, d\pi_Y(y) \;=\; \int m_X(Ay)\, d\pi_Y(y)\,,$$

where $m_X$ is the moment-generating function of $X$, and thus

$$m_X(Ay) \;=\; \exp\big(\tfrac{1}{2}\, y^T A^T A y\big) \;=\; \exp\big(t\, y^T y\big)\,.$$

Hence the expectation of $m_X(AY)$ is the moment-generating function of $Y^T Y$ (which has a $\chi^2_k$-distribution) evaluated at $t$. Thus for $t < \frac{1}{2}$

$$\int \exp(x^T A y)\, d\pi(x,y) \;=\; (1-2t)^{-k/2}\,,$$

but for $t \geq \frac{1}{2}$ the integral is infinite.

So far we have assumed that the covariance matrices $\Sigma_x$ und $\Sigma_y$ of $X$ and $Y$ are non-singular. We now extend the above results to the more general case when $\Sigma_x$ and $\Sigma_y$ have any rank $r_x > 0$ and $r_y > 0$. Then there exists a $k_x \times r_x$ matrix $C_x$ of rank $r_x$ such that $\Sigma_x = C_x C_x^T$ and for a standardized $U \sim N_{rx}(0, I)$ the linear transform $g(U) = C_x U + \mu_x$ is distributed as $N_{k_x}(\mu_x, \Sigma_x)$. Therefore we may assume $X = g(U)$ and, by duality, $Y = h(V) = C_y V + \mu_y$ for a $V \sim N_{ry}(0, I)$.
Note that $g$ and $h$ have inverses, i.e. $g^{-1}(x) = C_x^-(x - \mu_x)$ with $C_x^- = (C_x^T C_x)^{-1} C_x^T$.

Now any joint normal distribution $P = N_k(\mu, \Sigma)$ for $(X,Y)$ is the image measure of a joint distribution $P' = N_r(\mu', \Sigma')$ for $(U,V)$ under the transformation given by $g$ and $h$. Furthermore we have $\pi \ll P \ll \pi$ if and only if $\pi' := \pi_U \times \pi_V \ll P' \ll \pi'$. The latter condition is equivalent to $\operatorname{rank}(\Sigma') = r_x + r_y$. Since $\Sigma$ and $\Sigma'$ have the same rank, we conclude

$$N_k(\mu, \Sigma) \in \mathscr{P} \qquad \Leftrightarrow \qquad \operatorname{rank}(\Sigma) = r_x + r_y\,.$$

Note that $P \in \mathscr{P}$ will not have a density with respect to Lebesgue's measure $\lambda$ unless $\Sigma$ is non-singular. By (15) the log odds ratio function of $P'$ with respect to $u^\circ = g^{-1}(x^\circ)$ and $v^\circ = h^{-1}(y^\circ)$ is given by $\psi^\circ_{P'}(u,v) = -(u-u^\circ)^T Q_{uv}(v-v^\circ)$ which yields the log odds ratio function of $P$ (cf. **2.4**) as $\psi^\circ_P(x,y) = -(x-x^\circ)^T C_x^{-T} Q_{xy} C_y^-(y-y^\circ)$. Hence even for a *singular* covariance matrix $\Sigma$ the log odds ratio function $\psi^\circ_P$ is bilinear (assuming $x^\circ = 0$ and $y^\circ = 0$). The characterization theorem can now be generalized to this more general situation (cf. **A.5**).

## 5. Discussion

The joint distribution $\mathscr{L}(X,Y)$ of two random elements $X$ and $Y$ has been shown to be completely determined by their marginal distributions $\mathscr{L}(X)$, $\mathscr{L}(Y)$ and an odds ratio function $OR(X,Y)$ derivable from the joint or any of the two conditional densities. Specifying each of these three parts separately yields (under mild integrability conditions) a unique joint distribution, which can be obtained as a limit of iterative marginal fittings. Hence the odds ratio function characterizes the association between $X$ and $Y$ in the sense that it carries all information on the joint distribution which is not contained in the marginal distributions. Thus the odds ratio function may be taken as a formal definition of association for arbitrary random elements. Since association is a *symmetric* concept, any result within this framework entails its *dual* by interchanging $X$ with $Y$.

The decomposition of a joint distribution into the association and the marginals is of importance whenever the association but not the marginals are of primary interest. *Association models* and corresponding parametrizations of the (conditional) densities given here provide a unifying formal framework for statistical studies on the relationship between $X$ and $Y$ based on *densities.* These models are semi-parametric in the sense that they do not restrict the marginal distributions. Hence statistical inference may be based on samples from the conditional distribution of either variable given the other.

We introduced the class of *log-bilinear* association models which generalizes important standard models (by removing restrictions on the marginals), e.g. *Generalized Linear Models* with canonical link. In particular the logistic regression model (without restriction on the intercept parameters) represents a log-bilinear association model and hence does not restrict the marginal distribution of $Y$. And for random vectors $X$ and $Y$ with normal *marginals* the joint distribution is multivariate normal if and only if their assosciation is log-bilinear. Statistical applications of log-bilinear association models to dimension reduction and linear discriminant functions in multivariate analysis are given in van der Linde (2002a).

Finally we briefly indicate how the separation of association from the marginals may provide new perspectives in related areas not covered here. For instance, *measures* of *dependence* or *association* may be defined by taking marginal or joint expectations of the odds ratio $OR(X,Y)$ - or some function of it. Van der Linde (2002b) provides such applications in Bayesian statistics (where a *random* parameter $\Theta$ replaces $X$). An extentensive coverage of multivariate dependencies based on *distribution functions* and related concepts (rather than odds ratios) is given by Joe (1997).

Furthermore our existence theorem allows *classifications* as well as *constructions* of joint distributions with given marginals by specifying or modelling the odds ratio function. Distributions with *fixed* marginals have been studied extensively (with different methods and perspectives) and the reader is referred to Rüschendorf et al (1996), Joe (1997) and the literature cited therein for details. In particular joint distributions $\mathscr{L}(X,Y)$ with specified marginals for *real-valued* $X$ and $Y$ have been constructed using the joint and marginal *distribution functions* $F_{XY}$ and $F_X$, $F_Y$ (rather than *densities* used here). Plackett (1965) introduced a family of bivariate distributions depending on a single parameter and the marginals. A more general approach is given by Sklar's theorem which provides a relation $F_{XY}(x,y) = C\big(F_X(x), F_Y(y)\big)$ by means of a *copula* $C$, which is a function with certain properties. For details and further references on copulas we refer to Nelson (1999) and Joe (1997). Sklar's theorem extends to random vectors $X = (X_1, X_2 ...)$ and $Y = (Y_1, Y_2 ...)$ and gives their joint distribution function $F_{XY}(x_1, x_2 ..., y_1, y_2 ...) = C\big(F_{X_1}(x_1), F_{X_2}(x_2) ..., \;\; F_{Y_1}(y_1), F_{Y_2}(y_2) ...\big)$ in terms of a multivariate copula $C$ and the marginal distribution functions $F_{X_i}$ resp. $F_{Y_i}$ of the components $X_i$ resp. $Y_i$. Now the copula $C$ does not only capture the association between the vectors $X$ and $Y$ but also dependencies among the *components* $X_1$, $X_2$... of $X$ resp. $Y_1$, $Y_2$... of $Y$. Hence in the multivariate case the copula $C$ contains additional information about the (marginal) distribution of $X$ resp. $Y$ and does not represent the association between $X$ and $Y$ in the sense discussed here.

***Acknowlegdment***. The author thanks both referees for their helpful comments which improved the first draft of the paper.

# Appendix: Proofs

## A.1 Proof of the Uniqueness-Theorem

We first show that $I(f_1,f_2)=0$ holds if and only if $P_1=P_2$. Note that

$$0 \leq I(f_1,f_2) \leq \infty \tag{A1}$$

even if $\varphi_1-\varphi_2$ is only quasi-integrable (i.e. the positive or the negative part is integrable) with respect to $P_1$. In fact, from $-\log a \leq a^{-1}-1$ we conclude for the negative part $(\varphi_1-\varphi_2)^- = \big(\log(f_1/f_2)\big)^- \leq f_2/f_1$. Since $\int (f_2/f_1)\, dP_1 = \int f_2\, d\pi = 1$ we get

$$(\varphi_1-\varphi_2)^- \text{ is } P_1\text{-integrable.} \tag{A2}$$

Hence $(\varphi_1-\varphi_2)^-$ is $P_1$-quasi-integrable and $I(f_1,f_2) > -\infty$. Now the first part of (A1) follows from Jensen's inequality,

$$I(f_1,f_2) = \int -\log(f_2/f_1)\, dP_1 \geq -\log \int (f_2/f_1)\, dP_1 = 0 \ ,$$

where equality holds if and only if $f_1 = f_2$ $P_1$-almost surely - which in view of $P_1 \ll \pi \ll P_1$ is equivalent to $\pi$-almost surely - and hence if $P_1 = P_2$.

Now we justify (6) - which will be used to show $I(f_1,f_2)=0$ - under integrability conditions. For any *non-negative* $\beta \in \mathscr{L}_X^{\,0}$ we have by Fubini's theorem and (5)

$$\int \beta\, dP_i = \int \beta \cdot f_i\, dP = \int \beta \cdot f_i^X\, d\pi_X = \int \beta\, d\pi_X \qquad \text{for } i = 1, 2.$$

Separate application to the positive and negative part of $\beta_1$ leads to

$$\begin{aligned} \beta_1 \text{ is } P_i\text{-integrable} \quad &\Leftrightarrow \quad \beta_1 \text{ is } \pi_X\text{-integrable}\ , \\ &\Rightarrow \quad \int \beta_1\, dP_i = \int \beta_1\, d\pi_X\,. \end{aligned} \tag{A3}$$

The dual argument provides the corresponding result for $\gamma_1$, which even holds if *integrability* is replaced by *quasi-integrability*,

$$\begin{aligned} \gamma_1 \text{ is } P_i\text{-quasi-integrable} \quad &\Leftrightarrow \quad \gamma_1 \text{ is } \pi_Y\text{-quasi-integrable}\ , \\ &\Rightarrow \quad \int \gamma_1\, dP_i = \int \gamma_1\, d\pi_Y\,. \end{aligned} \tag{A4}$$

Now, by duality it suffices to prove the theorem under the assumption (*i*) that $\beta_1 = (\varphi_1-\varphi_2)(\cdot\,, y^\circ)$ is $\pi_X$-integrable and thus $P_1$-integrable by (A3). Hence the negative part of $\gamma_1 = (\varphi_1-\varphi_2)-\beta_1$ is $P_1$-integrable by (A2) and $\pi_Y$-integrable by (A4). This establishes (6) which, in view of (A1), implies

$$-\int \beta_1\, d\pi_X \leq \int \gamma_1\, d\pi_Y\,. \tag{A5}$$

Switching the indices 1 and 2 we get

$$\varphi_2-\varphi_1 = \beta_2+\gamma_2 \qquad \text{with} \qquad \beta_2 = -\beta_1, \qquad \gamma_2 = -\gamma_1.$$

Since $\beta_2 = -\beta_1$ is $\pi_X$-integrable by $(i)$, we obtain a result corresponding to (A5)

$$-\int \beta_2 \, d\pi_X \;\le\; \int \gamma_2 \, d\pi_Y .$$

Combined with (A5) we get $-\int \beta_1 \, d\pi_X = \int \gamma_1 \, d\pi_Y$ and hence $I(f_1, f_2) = 0$, which completes the proof of the theorem. The corollary follows since integrability of $\varphi_i$ implies that $\int |\varphi_i(\text{-}, y)| \, d\pi_X$ is finite for all $y$ except on a set of $\pi_Y$-measure zero. Hence there exist $y^\circ \in \Omega_Y$ with $\varphi_i(\text{-}, y^\circ) \in \mathscr{L}^1_X$ and $(i)$ holds, and by duality we get $(ii)$.

## A.2 Properties of Marginal Fitting Sequences

In this section we establish inequalities for the marginal fitting sequence $(\eta_n)_{n \ge 0}$ with starting value $\eta_0$ and the accompanying sequence $(\tilde\eta_n)_{n \ge 0}$ which are needed in the proof of the convergence theorem. The existence conditions (EC1) and (EC2) will *not* be assumed in this section. In the following the index "$n$" always ranges over non-negative integers $n \ge 0$. An inequality $\xi_1 \le \xi_2$ resp. $\xi_1 < \xi_2$ for *functions* $\xi_1$ and $\xi_2$ here means that $\xi_1(z) \le \xi_2(z)$ resp. $\xi_1(z) < \xi_2(z)$ holds for all arguments $z$.

We first show that both sequences $\ell(\eta_n)$ and $\ell(\tilde\eta_n)$ are non-decreasing and have the same limit $s$. Indeed, from (10) and (11) we get

$$\ell(\eta_n) \;\le\; \ell(\tilde\eta_n) \;\le\; \ell(\eta_{n+1}) \;\le\; \ell(\tilde\eta_{n+1}) \;\le\; \ell(-\bar\varphi) \tag{A6}$$

which implies the convergences

$$\begin{aligned} \ell(\eta_n) &= \alpha_n - 1 \longrightarrow s , & \alpha_n &= \int \eta_n \, d\pi , \\ \ell(\tilde\eta_n) &= \tilde\alpha_n - 1 \longrightarrow s , & \tilde\alpha_n &= \int \tilde\eta_n \, d\pi . \end{aligned} \tag{A7}$$

Hence $(\alpha_n)$ and $(\tilde\alpha_n)$ are non-decreasing too with the same limit $\alpha = s + 1$. Next we provide upper and lower bounds for

$$\eta_{n+1} = M_X(\tilde\eta_n) = M_X(\gamma_n) \qquad \text{with} \qquad \gamma_n = \tilde\eta_n^Y - \tilde\alpha_n = \Pi(\tilde\eta_n \,|\, \mathscr{M}^1_Y).$$

From (8) we get $\eta^X_{n+1} = -\log \bar f^X(\gamma_n)$, and hence $\eta_{n+1} = \gamma_n + \eta^X_{n+1}$. This allows us to derive bounds for $\eta_{n+1}$ from those for $\gamma_n$ and $\eta^X_{n+1}$. By (9), applied to $\tilde\eta_n$, we get

$$1 - \bar f^X(\gamma_n) \;\le\; \eta^X_{n+1} \;\le\; -\bar\beta , \tag{A8}$$

which already provides the upper bound for $\eta^X_{n+1}$. For the accompanying sequence $(\tilde\eta_n)$ we obtain the dual properties

$$\begin{aligned}
&\tilde{\eta}_n = M_Y(\eta_n) = M_Y(\beta_n) && \text{with} \quad \beta_n = \eta_n^X - \alpha_n = \Pi(\eta_n \,|\, \mathscr{M}_X^1)\,, \\
&\tilde{\eta}_n^Y = -\log \bar{f}^Y(\beta_n)\,, && \qquad\quad \tilde{\eta}_n = \beta_n + \eta_n^Y\,, \\
&1 - \bar{f}^Y(\beta_n) \le \tilde{\eta}_n^Y \le -\bar{\gamma}\,. && \qquad\qquad\qquad\qquad\qquad\qquad \text{(A9)}
\end{aligned}$$

In the following we use some evident properties (and their duals) of the operator $\bar{f}$:

$$\begin{aligned}
&\bar{f}(\eta + \eta') = \bar{f}(\eta) \cdot \exp(\eta')\,, && \eta' \in \mathscr{M}^1, \\
&\eta \le \eta' \quad \Rightarrow \quad \bar{f}(\eta) \le \bar{f}(\eta')\,, \quad \bar{f}^Y(\eta) \le \bar{f}^Y(\eta')\,, && \\
&\bar{f}^Y(\eta + \gamma) = \bar{f}^Y(\eta) \cdot \exp(\gamma), && \gamma \in \mathscr{L}_Y^1.
\end{aligned}$$

From (A8) and $\alpha_0 \le \alpha_{n+1}$ we get

$$\beta_{n+1} = \eta_{n+1}^X - \alpha_{n+1} \le -\bar{\beta} - \alpha_{n+1} \le -\bar{\beta} - \alpha_0$$

which implies

$$\bar{f}^Y(\beta_{n+1}) \le \bar{f}^Y(-\bar{\beta} - \alpha_0) = \bar{f}^Y(-\bar{\beta}) \cdot \exp(-\alpha_0)$$

and

$$-\tilde{\eta}_{n+1}^Y = \log \bar{f}^Y(\beta_{n+1}) \le \log \bar{f}^Y(-\bar{\beta}) - \alpha_0\,. \tag{A10}$$

Putting $q = \exp(\psi)$ we get $\bar{f}(-\bar{\beta}) = \exp(\bar{\gamma} + \psi)$ and

$$\log\big(\bar{f}^Y(-\bar{\beta})\big) = \log\big[\textstyle\int \exp(\bar{\gamma} + \psi)\, d\pi_X\big] = \bar{\gamma} + \log q^Y$$

which, in combination with (A10) and (A9), yields

$$(\alpha_0 - \bar{\gamma} - \log q^Y) \le \tilde{\eta}_{n+1}^Y \le -\bar{\gamma}\,.$$

This provides the bounds for $\tilde{\eta}_n^Y$

$$\delta_1 := \min(\tilde{\eta}_0^Y, \alpha_0 - \bar{\gamma} - \log q^Y) \le \tilde{\eta}_n^Y \le \max(\tilde{\eta}_0^Y, -\bar{\gamma}) =: \delta_2\,. \tag{A11}$$

From $\tilde{\alpha}_n \uparrow \alpha$ we obtain the desired bounds for $\gamma_n = \tilde{\eta}_n^Y - \tilde{\alpha}_n$,

$$\delta_3 := \delta_1 - \alpha \le \gamma_n \le \delta_2 - \tilde{\alpha}_0 =: \delta_4\,. \tag{A12}$$

Dualizing the arguments leading to (A11) and (A12), we get bounds for $\eta_n^X$ and $\beta_n$,

$$\begin{aligned}
\delta_5 := \min(\eta_0^X, \tilde{\alpha}_0 - \bar{\beta} - \log q^X) \le \eta_n^X \le \max(\eta_0^X, -\bar{\beta}) =: \delta_6\,, \\
\delta_7 := \delta_5 - \alpha \le \beta_n \le \delta_6 - \alpha_0 =: \delta_8\,.
\end{aligned} \tag{A13}$$

Finally, we obtain the bounds for $\eta_n$ and $\tilde{\eta}_n$,

$$\begin{aligned}
\delta_3 + \delta_5 \le \quad \eta_{n+1} = \gamma_n + \eta_{n+1}^X \le \delta_4 + \delta_6\,, \\
\delta_1 + \delta_7 \le \quad \tilde{\eta}_n = \beta_n + \tilde{\eta}_n^Y \le \delta_2 + \delta_8\,.
\end{aligned} \tag{A14}$$

The above bounds will be used in two ways. First, the bounded functions lie in *compact* sets, e.g. for any $y \in \Omega^Y$ we have $\gamma_n(y) \in K_y := [\,\delta_3(y), \delta_4(y)\,] \subset \mathbb{R}$, and hence

$$\gamma_n \in K := \prod_{y \in \Omega_Y} K_y \subset \mathbb{R}^{\Omega_Y}.$$

Now $K$ is a product of compact sets $K_y$ and by Tychonov's theorem $K$ is compact with respect to the product topology in which convergence of functions is given by *pointwise* convergence. - Second, all bounding functions $\delta_1, ..., \delta_8$ are *integrable* if the existence condition (EC1)$'$ and (EC2)$'$ hold, which allows integrating to the limit for pointwise convergent subsequences. Note that the integrability of the *upper* bounds does not require the existence conditions.

### A.3 Proof of the Convergence Theorem

In the following proof of the convergence theorem the indices $m(n)$ of a subsequence will simply be written as $m$, i.e. $(\eta_m)$ is a subsequence of $(\eta_n)$. Since the sequence $(\gamma_n)$ lies in a compact set $K$ there exists a (pointwise) convergent subsequence $k = k(n) > 1$ of the given subsequence $(\eta_m)$, i.e.

$$\gamma_k \longrightarrow \gamma \in K. \tag{A15}$$

Now $\gamma$ is measurable and bounded by integrable functions and hence $\gamma \in \mathscr{L}^1_Y$. Furthermore the integrable bounds also provide convergence in the mean in (A15). The convergence of $(\gamma_k)$ entails several *pointwise* convergences, some of which - in view of the integrable bounds derived in **A.2** - also imply convergence *in the mean.* First, from (A15) and (A12) we get pointwise convergences

$$\eta^X_{k+1} = -\log \bar{f}^X(\gamma_k) \longrightarrow -\log \bar{f}^X(\gamma)\,, \tag{A16}$$

$$\eta_{k+1} = M_X(\gamma_k) = \gamma_k - \log \bar{f}^X(\gamma_k) \longrightarrow \gamma - \log \bar{f}^X(\gamma) = M_X(\gamma) =: \eta\,, \tag{A17}$$

where the limits $-\log \bar{f}^X(\gamma)$ and $\eta$ are finite and integrable by (A13) and (A14). The convergences in (A16) and (A17) also hold *in the mean,* and the latter implies

$$\ell(\eta_{k+1}) = \int \eta_{k+1}\, d\pi - 1 \longrightarrow \ell(\eta) = \int \eta\, d\pi - 1\,. \tag{A18}$$

Taking $(\eta_{k+1})$ as the wanted subsequence $(\eta_{m'(n)})$ we now prove (*i*) and (*ii*). Since $M_X(\eta) = M_X(M_X(\gamma)) = M_X(\gamma) = \eta$ it remains to show $M_Y(\eta) = \eta$ $\pi_X$*-almost surely* and this is by (12) equivalent to

$$\ell(\eta) = \ell(M_Y(\eta)). \tag{A19}$$

Now (A17) and (A14) provide pointwise convergence (as well as in the mean) of the projections onto $\mathscr{M}^1_X$

$$\beta_{k+1} = \eta^X_{k+1} - \int \eta_{k+1}\, d\pi \longrightarrow \beta := \eta^X - \int \eta\, d\pi\,.$$

Replacing the sequence $(\gamma_k)$ by $(\beta_{k+1})$ in the arguments leading to (A18), we get

$$\ell(\tilde{\eta}_{k+1}) = \int \tilde{\eta}_{k+1}\, d\pi - 1 \;\longrightarrow\; \ell(M_Y(\beta)) = \int M_Y(\beta)\, d\pi - 1 \,.$$

By the dual of (7) we have $M_Y(\eta) = M_Y(\beta)$ and hence

$$\ell(\tilde{\eta}_{k+1}) \longrightarrow \ell(M_Y(\eta)) \,. \tag{A20}$$

Finally the sequences $\ell(\eta_{k+1})$ and $\ell(\tilde{\eta}_{k+1})$ have the same limit by (A7), and thus (A18) and (A20) establish (A19) and hence (*i*). From (*i*) we obtain $\bar{f}^X(\eta) = 1$ $\pi_X$-almost surely and hence $\int \bar{f}(\eta)\, d\pi = \int \bar{f}^X(\eta)\, d\pi^X = 1$ which proves (*ii*).

It remains to show that limit $\eta$ is $\pi$-almost surely independent of the subsequence and of the starting value. Indeed (as shown in **A.4**) *any* limit $\eta \in \mathscr{M}^1$ with (*i*) and (*ii*) provides a $\pi$-density $\bar{f}(\eta)$ of a probability measure $P$ with marginals $\pi_X$, $\pi_Y$ and a log odds ratio function determined by $\psi = \Pi(\varphi \,|\, \mathscr{A}^1)$. For two such limits $\eta_1$ and $\eta_2$ the uniqueness theorem (and its corollary) gives $\bar{f}(\eta_1) = \bar{f}(\eta_2)$ and hence $\eta_1 = \eta_2$ $\pi$-almost surely. This concludes the proof of the theorem.

To establish the *corollary,* we take $\eta$ as the limit provided by the theorem for the original sequence $(\eta_n)$ viewed as its own subsequence. Now any other subsequence of $(\eta_n)$ contains a further subsequence converging in the mean to a limit which coincides with the above $\eta$ $\pi$-almost surely and hence also converges in the mean to $\eta$. This proves convergence in the mean of $(\eta_n)$ to the limit $\eta$ and hence

$$\ell(\eta_n) \;=\; \int \eta_n\, d\pi - 1 \;\longrightarrow\; \int \eta\, d\pi - 1 \;=\; \ell(\eta) \,.$$

The sequence $\ell(\eta_n)$ is non-decreasing by (A6) which proves (*iii*). It remains to show, that $\eta$ is the $\pi$-almost surely-unique argument which maximizes the functional $\ell$. From (*iii*) and (A6) we first conclude $-\infty < \ell(\eta_0) \le \ell(\eta)$. For any $\eta^* \in \mathscr{M}^1$ with $\ell(\eta) \le \ell(\eta^*)$ we establish $\ell(\eta) = \ell(\eta^*)$ as follows. First $\ell(\eta^*)$ is *finite* and hence $\bar{f}(\eta^*)$ is $\pi$-integrable. Taking $\eta'_0 = M_X(\eta^*)$ as a starting value for a new marginal fitting sequence $(\eta'_n)$ we already know that that $(\eta'_n)$ converges in the mean to a limit $\eta' \in \mathscr{M}^1$. From (11) and (*iii*) - applied to $\eta'_n$ - we get $\ell(\eta) \le \ell(\eta^*) \le \ell(\eta'_0) \le \ell(\eta')$. However, by the theorem $\eta$ and $\eta'$ coincide $\pi$-almost surely, and hence $\ell(\eta) = \ell(\eta')$ which proves $\ell(\eta) = \ell(\eta^*)$. The functional $\ell$ is strictly concave modulo $\underset{\pi}{\sim}$ thus making $\eta$ the $\pi$-*almost surely-unique* argument maximizing $\ell$.

### A.4 Proof of the Existence Theorem

The existence theorem will be obatined from the convergence theorem. First, $\bar{f}(-\bar{\gamma}) = \exp(\bar{\beta} + \psi)$ is $\pi$-integrable (EC1) and hence $\eta_0 = -\bar{\gamma} - \log\left[\int \bar{f}(-\bar{\gamma})\, d\pi\right] \in \mathscr{M}^1$ provides a starting value with $\int \bar{f}(\eta_0)\, d\pi = 1$. The convergence theorem yields a limit $\eta \in \mathscr{M}^1$ satisfying (*i*) and (*ii*). We show that $\varphi = \eta + \bar{\varphi} = \eta + \bar{\beta} + \bar{\gamma} + \psi \in \mathscr{L}^1$ is the wanted log-density. Since $M_X(\eta) = \eta$ resp. $M_Y(\eta) = \eta$ is equivalent to $\bar{f}^X(\eta) = 1$ resp. $\bar{f}^Y(\eta) = 1$, the function $\bar{f}(\eta) = \exp(\varphi)$ is a $\pi$-density of a probability measure $P$ with marginal distributions $\pi_X$ and $\pi_Y$. Furthermore the log-odds ratio function of $P$ is determined by $\psi = \Pi(\varphi \mid \mathscr{A}^1)$, which proves the theorem.

### A.5 Proof of the Characterization Theorem

***Non-singular case:*** We first provide a proof for the case that both $\Sigma_x$ and $\Sigma_y$ are non-singular from which the general case will be derived. It was already shown in **4.4** that the log odds ratio function for any such normal $P$ is bilinear.

Conversely, for a given $P \in \mathscr{P}^1$ with normal marginals $\pi_X$, $\pi_Y$ and a bilinear log odds ratio function $\psi^\circ(x,y) = x^T A y$ we have to show that $P$ is *normal.* One possibility is to find the correponding covariance matrix $\mathrm{cov}(X,Y) = \Sigma_{xy}$ as a solution of the equation $A = Q_x \Sigma_{xy} \Sigma_y^{-1}$, where $Q_x$ given above also depends on $\Sigma_{xy}$. However, we will apply the theorems of convergence and uniqueness to show that $P$ is normal. In view of **2.4** it is sufficient to consider the special case with $\mu_x = 0$ and $\mu_y = 0$, for which $\psi^\circ$ coincides with its projection: $\psi = \Pi(\psi^\circ \mid \mathscr{A}^1) = \psi^\circ$.

We first note that the existence condition $(\mathrm{EC1})_A$ and its dual $(\mathrm{EC2})_A$ hold (cf. **4.3**). Next we construct a normal distribution $P_0$ with log odds ratio function $\psi^\circ$ (but not with the wanted marginals). Taking $P_0$ as a starting value in the iterated marginal fitting procedure, we get in the limit a distribution $P$ which turns out to be normal.

To obtain $P_0$ take any $a > 0$ such that $C = a^{-1} A$ is a correlation matrix, i.e. $|C_{ij}| < 1$ for all $i, j$. Now for standard normally distributed vectors $U \sim N_{kx}(0, I)$ and $V \sim N_{ky}(0, I)$ ($I$ denotes a unit matrix) there exists a joint normal distribution $P_C$ such that $C$ is the covariance matrix of $(U, V)$ under $P_C$. By (38) and (37) the log-odds ratio function of $P_C$ is $\psi_C(u,v) = u^T (BA) v$ with $B = a^{-1} (I - CC^T)^{-1}$. The joint distribution $P_0$ of $(BU, V)$ under $P_C$ is normal $N_k(0, \Sigma_0)$, and its log-odds ratio function turns out to be $\psi^\circ$ (cf. **2.4**). Since $\psi^\circ = \psi$, the log-density $\varphi_0 = \log(dP_0 / d\pi)$ is of the form $\varphi_0 = \eta_0 + \bar{\eta} + \psi$ with $\eta_0 \in \mathscr{M}^1$ and $\bar{\eta} = \bar{\beta} + \bar{\gamma}$ taken from the conditions (EC1) and (EC2).

Consider next the iterated marginal fitting sequence $\eta_n = M_X M_Y(\eta_{n-1})$ and the corresponding distributions $P_n$ with log-densities $\varphi_n = \log(dP_n / d\pi) = \eta_n + \bar{\eta} + \psi$. For any centered normal distribution $P' = N(0, \Sigma)$ the (conditional) distributions $M_X(P')$ and $M_Y(P')$ are normal and centered, too. Hence all $P_n$ are normal $N(0, \Sigma_n)$ and from the convergence theorem we obtain a subsequence $m = m(n)$ such that $(\eta_m)$ converges (pointwise) to an $\eta \in \mathscr{M}^1$. This implies convergence of the log-densities $\varphi_m = \eta_m + \bar{\eta} + \psi \longrightarrow \eta + \bar{\eta} + \psi =: \varphi$. Thus the log-densities $\varphi_{\lambda m}$ with respect to Lebesgue's measure $\lambda$ converge

$$\varphi_{\lambda m} \;=\; \varphi_m + \log f_\lambda^x + \log f_\lambda^y \quad\longrightarrow\quad \varphi + \log f_\lambda^x + \log f_\lambda^y \;=: \varphi_\lambda \tag{A21}$$

Since $\varphi$ is a log-density of a distribution $P_\varphi$ with marginals $\pi_X$, $\pi_Y$ and log odds ratio function $\psi^\circ = \psi$, the uniqueness theorem yields $P = P_\varphi$. Hence, for $P$ to be normal, it remains to show that $\varphi_\lambda$ is a log-density of a *normal* distribution. Now for any $z = (x, y)$ we have

$$\varphi_{\lambda m}(z) \;=\; -\tfrac{1}{2}\, z^T Q_m z + c_m\,, \qquad Q_m = \Sigma_m^{-1}\,, \qquad c_m = -\tfrac{1}{2} \log[(2\pi)^k \det(\Sigma_m)]\,.$$

Then, by (A21)

$$\begin{aligned} c_m = \varphi_{\lambda m}(0) \quad&\longrightarrow\quad \varphi_\lambda(0) =: \varphi, \\ 2[c_m - \varphi_{\lambda m}(z)] = z^T Q_m z \quad&\longrightarrow\quad 2[c - \varphi_\lambda(z)] \qquad \text{for all } z. \end{aligned} \tag{A22}$$

Choosing unit vectors and sums of those for $z$ proves that $(Q_m)$ converges to a matrix $Q$ and thus

$$\varphi_{\lambda m}(z) = -\tfrac{1}{2}\, z^T Q_m z + c_m \quad\longrightarrow\quad \varphi_\lambda(z) = -\tfrac{1}{2}\, z^T Q z + c$$

for any $z$. Hence $\varphi_\lambda$ is a log-density of a normal distribution, provided $Q$ is non-singular and positive-definite. Now (A22) implies convergence of $\log[\det(Q_m)] = -\log[\det(\Sigma_m)]$ and hence $\det(Q_m) \longrightarrow \det(Q) > 0$ which proves that $Q$ is non-singular. Since all $Q_m$ are positive-definite, so is their limit $Q$, which completes the proof.

***General case:*** For normal $P$ we already showed in **4.4** that the log odds ratio function is bilinear. Conversely, suppose $P \in \mathscr{P}^1$ has normal marginals $\pi_X$, $\pi_Y$ and a bilinear log odds ratio function $\psi^\circ(x, y) = x^T A y$ with respect to $x^\circ = 0$, $y^\circ = 0$. We have to show that $P$ is normal and may again assume $\mu_x = 0$ and $\mu_y = 0$ in view of **2.4**. Consider the joint distribution $P'$ of $U = C_x^- X$ and $V = C_y^- Y$ which has normal marginals and a bilinear log odds ratio function given by $\psi^{\circ\prime}(u, v) = u^T (C_x^T A C_y) v$. From the non-singular case of the theorem we conclude that $P'$ is normal and hence the joint distribution $P$ of $(X, Y)$ is normal, too.

## References

Breslow NE and Day NE (1980) *Statistical Methods in Cancer Research, Volume I: The Analysis of Case-Control Studies.* International Agency for Research on Cancer, Lyon.

Haberman SJ (1974) *The Analysis of Frequency Data.* The University of Chicago Press, Chicago.

Joe H (1997) *Multivariate Models and Dependence Concepts.* Chapman & Hall, London.

Kullback S (1959) *Information Theory and Statistics.* Mineola, New York: Dover Publications (1968).

van der Linde A (2002a) *Dimension reduction and linear discriminant functions based on an odds ratio parametrization.* Mathematik-Arbeitspapiere 56, Universität Bremen (http://www.math.uni-bremen.de/~avdl).

van der Linde A (2002b) *On the association between a random parameter and an observable.* Mathematik-Arbeitspapiere 57, Universität Bremen (http://www.math.uni-bremen.de/~avdl/). Final version: Test 13 (2004), 85-111.

van der Linde A and Osius G (2003) *Discrimination based on an odds ratio parametrization.* Seventh Valencia International Meeting on Bayesian Statistics, June 2002. Bernando, JM et al (Eds.) *Bayesian Statistics* 7, 233-248, Oxford University Press.

Nelson, RB (1999) *An Introduction to Copulas.* Lecture Notes in Statistics 139, Springer, New York.

Plackett RL (1965) *A class of bivariate distributions.* Journal of the American Statistical Association 60: 516-522.

Plackett RL (1974) *The Analysis of Categorical Data.* Griffin, London.

Rüschendorf L, Schweizer B and Taylor MD (eds.) (1996) *Distributions with Fixeed Marginals and Related Topics.* IMS Lecture Notes-Monograph Series 28. Institute of Mathematical Statistics, Hayward, California

Sinkhorn R (1967) *Diagonal Equivalence to Matrices With Prescribed Row and Column Sums.* American Mathematical Monthly 74: 402 - 405.

*Date: 24-Sept-2004*