

Maximum Entropy Inference for Mixed Continuous-Discrete Variables

Hermann Singer
FernUniversität in Hagen *

February 10, 2010

Abstract

We represent knowledge by probability distributions of mixed continuous and discrete variables. From the joint distribution of all items, one can compute arbitrary conditional distributions, which may be used for prediction. However, in many cases only some marginal distributions, inverse probabilities, or moments are known. Under these conditions, a principle is needed in order to determine the full joint distribution of all variables. The principle of maximum entropy (1; 2; 3; 4) ensures an unbiased estimation of the full multivariate relationships by using only known facts. For the case of discrete variables, the expert shell SPIRIT implements this approach (cf. 5; 6; 7). In this paper the approach is generalized to continuous and mixed continuous-discrete distributions and applied to the problem of credit scoring.

Key Words:

Expert Systems; Entropy; Boltzmann Distribution; Credit Scoring

*Lehrstuhl für angewandte Statistik und Methoden der empirischen Sozialforschung,
D-58084 Hagen, Germany, Hermann.Singer@FernUni-Hagen.de

1 Introduction

The principle of maximum entropy (maxent) (1; 2; 3; 4) guarantees the unbiased (prejudice free) estimation of unknown probability distributions when only some facts, such as certain moment constraints, are known. For example, knowledge of the mean and standard deviation leads to the Gaussian distribution, which is maximizing the entropy functional. More generally, moment constraints lead to distributions in the form of the exponential family. In the physics literature, one speaks of Boltzmann or Gibbs distributions. They maximize the entropy under mean energy and particle number constraints (canonical and grand canonical ensemble). In applications in economics, often discrete (e.g. binary) variables are utilized (cf. 5; 6; 7). On the other hand, many relevant variables are continuous, e.g. wages, asset returns, gross national product, etc. Of course one can discretize the continuous variables by using thresholds, but this leads to a loss of available information.

Therefore, a theory for mixed discrete/continuous variables is developed and implemented numerically. Moreover, conditional constraints such as conditional probabilities are allowed, in order to use the partial knowledge in certain subgroups of the full variable set.

The article is organized as follows: In sect. 2, the maximum entropy formalism is developed for unconditional restrictions, whereas conditional expectations are covered in sect. 3. The complete mixed variable theory with multidimensional variables and multiple restrictions is treated in sect. 4. Numerical considerations and conditionally gaussian models are topic of sects. 5 and 6. Finally, the method is applied to the problem of credit scoring.

2 Entropy

First we define the relative entropy

$$S[p, p_0] = - \int p(x) \log \frac{p(x)}{p_0(x)} dx \leq 0 \quad (1)$$

(cf. ref. 8) of a probability density $p(x)$ w.r.t. a reference density or prior $p_0(x)$. Using $\log x \leq x - 1$ the inequality in (1) is easily proved. The maximum is obtained when $p = p_0$, i.e. $S[p_0, p_0] = 0$. In the case of no constraints, the maximum entropy distribution is thus equal to the prior p_0 , e.g. the uniform distribution $p_0 = U(x; [a, b])$ on the interval $[a, b]$. On the

other hand, if knowledge is available in the form of moments

$$E[F(X)] = \int F(x)p(x)dx \stackrel{!}{=} f, \quad (2)$$

the entropy (1) must be maximized under this constraint. Introducing the Lagrangian

$$L[p, \lambda] = S[p, p_0] + \lambda(f - E[F]), \quad (3)$$

the extremal point is found by computing the functional derivative $\delta/\delta p(y)$

$$\frac{\delta}{\delta p(y)}L[p, \lambda] = -\log \frac{p(y)}{p_0(y)} - 1 - \lambda F(y) \stackrel{!}{=} 0 \quad (4)$$

using the differentiation rule

$$\frac{\delta p(x)}{\delta p(y)} = \delta(x - y), \quad (5)$$

where $\delta(x - y)$ is the Dirac delta function. Solving for $p(y)$ we obtain the normalized maximum entropy density (Boltzmann distribution)

$$p_B(y, \lambda) = Z^{-1}p_0(y) \exp[-\lambda F(y)] \quad (6)$$

where $Z(\lambda)$ is the partition function (Zustandssumme) ¹

$$Z(\lambda) = \int p_0(y) \exp[-\lambda F(y)]dy. \quad (7)$$

The unknown Lagrange parameter λ must be determined in order to fulfil the constraint (2). Inserting the Boltzmann distribution into the Lagrangian we obtain the concentrated Lagrangian ²

$$L[p_B, \lambda] = \log Z(\lambda) + \lambda f := L^*(\lambda). \quad (8)$$

Thus the derivative of $L^*(\lambda)$ w.r.t. λ leads to the constraint

$$\frac{\partial L^*(\lambda)}{\partial \lambda} = -Z^{-1} \int F(x)p_0(y) \exp[-\lambda F(y)]dy + f \quad (9)$$

$$= -E[F] + f \stackrel{!}{=} 0. \quad (10)$$

¹Alternatively, one can introduce a normalization constraint $\int p(x)dx = 1$ with Lagrange parameter λ_0 , i.e. $Z = \exp(\lambda_0)$.

²This terminology was borrowed from maximum likelihood theory, where the insertion of certain partial solutions $\sigma(\theta)$ leads to a concentrated likelihood $l(\theta, \sigma(\theta))$.

The second derivative is

$$\frac{\partial^2 L^*(\lambda)}{\partial \lambda^2} = -E[F]^2 + E[F^2] = \text{Cov}(F) \geq 0 \quad (11)$$

thus we seek a minimum of $L^*(\lambda)$. In general the solution must be found numerically, e.g. by using a quasi Newton method for minimization (cf. 9). In general one imposes $k = 1, \dots, K$ restrictions

$$E[F_k(X)] = \int F_k(x)p(x)dx \stackrel{!}{=} f_k \quad (12)$$

leading to a K -dimensional minimization problem. We obtain the second derivative

$$\frac{\partial^2 L^*(\lambda)}{\partial \lambda_k \lambda_{k'}} = -E[F_k]E[F_{k'}] + E[F_k F_{k'}] = \text{Cov}(F_k, F_{k'}) \geq 0, \quad (13)$$

which is the positive semidefinite covariance matrix of the constraints.

2.1 Example

Using the indicator function of the interval A as constraint function, we obtain the probability restriction

$$P(A) = E[\chi_A(X)] = \int_A p(x)dx$$

$$\chi_A(X) = \begin{cases} 1 & X \in A \\ 0 & \text{otherwise} \end{cases}.$$

The Boltzmann distribution is

$$p_B(y, \lambda) = Z^{-1}p_0(y) \exp[-\lambda\chi_A(y)]$$

with normalization constant

$$Z(\lambda) = \int p_0(y) \exp[-\lambda\chi_A(y)]dy.$$

Using the interval $A = (-\infty, 25]$ and $P(A) = 0.4$ with prior $p_0(x) = \phi(x; 45, 15^2)$ the maximum entropy distribution has the shape of a piecewise distorted gaussian (cf. fig. 1). The example may be interpreted as representing the age of a sample of study subjects (a priori information: mean 45 years, standard deviation 15 years) and the additional information that 40% of the people are younger than 25 years.

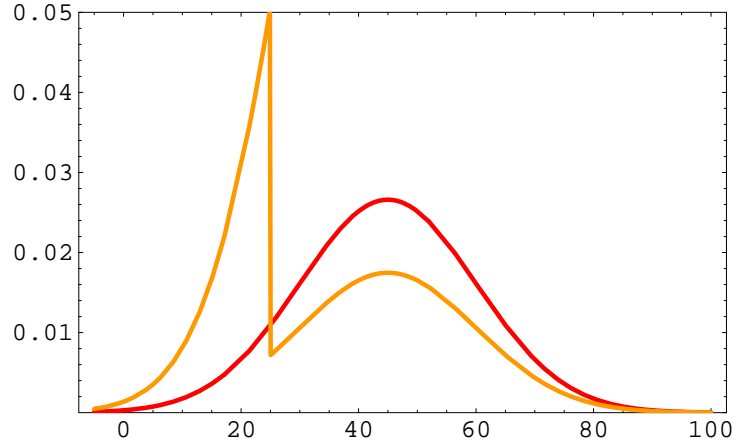


Figure 1: Boltzmann distribution for interval probability restriction $P(X \in (-\infty, 25]) = 0.4$ and prior $p_0(x) = \phi(x; 45, 15^2)$.

3 Conditional Restrictions

In many applications only conditional information is available. For example, we may know the age distribution in the group of good or bad customers. Similarly, the opinion of people may be known in the subgroups of preferences for certain political parties. From these conditional expectations one wants to estimate the full joint probability of the relevant variables.

We define the conditional constraint

$$\frac{E[F_1]}{E[F_2]} = f_{12} = \frac{\int F_1(x)p(x)dx}{\int F_2(x)p(x)dx}. \quad (14)$$

For example, the following indicator functions define the conditional probability

$$\begin{aligned} P(A|B) &= P(X \in A|X \in B) = f_{12} \\ F_1(x) &= \chi_A(x) \chi_B(x) = \chi_{A \cap B}(x) \\ F_2(x) &= \chi_B(x). \end{aligned}$$

These restrictions can be implemented with the Lagrange functional

$$L[p, \lambda] = S[p, p_0] + \lambda (f_{12}E[F_2] - E[F_1]). \quad (15)$$

Computing the functional derivatives as in (4) one obtains the maximum entropy distribution ³

$$p_B(x, \lambda) = Z^{-1} p_0(x) e^{\lambda[f_{12}F_2(x) - F_1(x)]} \quad (16)$$

$$Z(\lambda) = \int p_0(x) e^{\lambda[f_{12}F_2(x) - F_1(x)]} dx. \quad (17)$$

Inserting this into the Lagrangian (15) the unknown Lagrange parameters are found by solving the minimization problem

$$L^*[\lambda] = \log(Z)$$

$$\begin{aligned} L^{*'} &= Z^{-1} Z' \\ &= Z^{-1} \int [f_{12}F_2(x) - F_1(x)] e^{\lambda[f_{12}F_2(x) - F_1(x)]} dx \\ &= E[f_{12}F_2(X) - F_1(X)] = 0 \end{aligned}$$

$$\begin{aligned} L^{*''} &= Z^{-1} Z'' - Z^{-2} (Z')^2 \\ &= Z^{-1} \int [f_{12}F_2(x) - F_1(x)]^2 e^{\lambda[f_{12}F_2(x) - F_1(x)]} dx \\ &= E[f_{12}F_2(X) - F_1(X)]^2 \geq 0. \end{aligned}$$

The case of conditional restrictions includes the unconditional problem as special case, if we set $F_1 = F$, $F_2 = 1$ and $f_{12} = f$. Then, $L^*[\lambda] = \lambda f + \log \int p_0(x) \exp[-\lambda F(x)] dx$, thus recovering (8).

3.1 Example

We define the following conditional probability

$$\begin{aligned} P(A|B) &= P(X \in A | X \in B) = p_{AB} \\ F_1(x) &= \chi_A(x) \chi_B(x) = \chi_{A \cap B}(x) \\ F_2(x) &= \chi_B(x) \\ f_{12} &= p_{AB} = 0.3 \end{aligned}$$

and assume $A = [0, 3]$; $B = [2, 4]$. Under the prior $p_0(x) = \phi(x; 0, 10^2)$, the resulting Boltzmann distribution is shown in fig. (2). It fulfils the condition

³also called Boltzmann-, Gibbs- or exponential distribution.

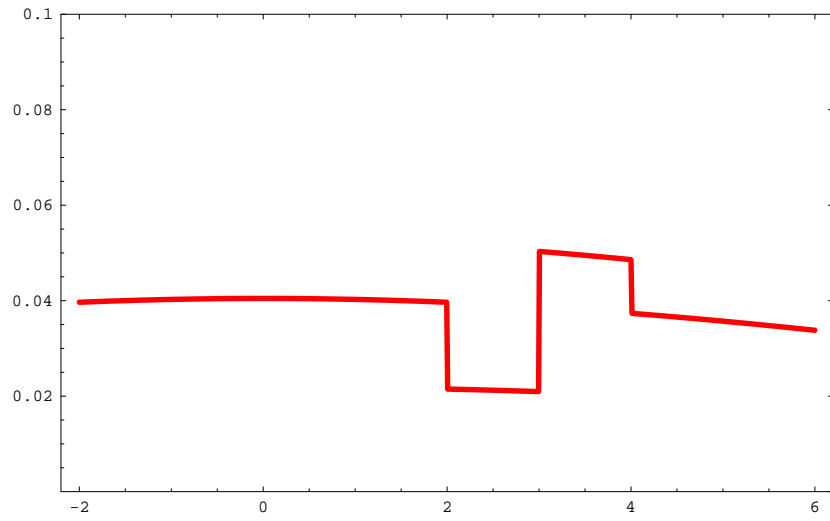


Figure 2: Boltzmann distribution for conditional probability restriction $P(A|B) = 0.3$ and prior $p_0(x) = \phi(x; 0, 10^2)$. $A = [0, 3]$; $B = [2, 4]$.

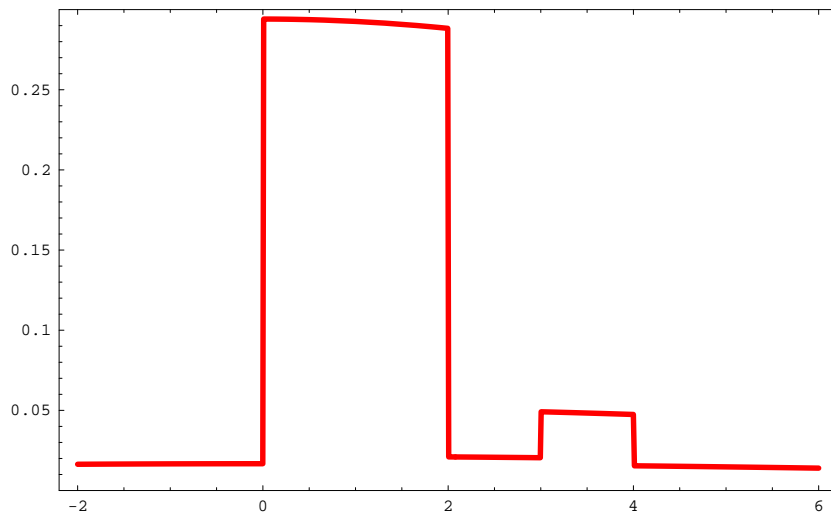


Figure 3: Boltzmann distribution for conditional probability restriction $P(A|B) = 0.3$. Additional constraint $P(A) = 0.6$.

$P([2, 3])/P([2, 4]) = 0.3$ and is as close as possible to the prior. Imposing the further condition $P(A) = 0.6$, fig. (3) is obtained. The function in the range $[0,3]$ is zoomed in order to get probability 0.6.

4 Mixed continuous-discrete problem

Introducing the discrete random variable I with distribution $p(i)$, we define the joint probability density $p(x, i)$ with normalization

$$\sum_i \int p(x, i) dx = 1. \quad (18)$$

In general we assume that $x = (x_1, \dots, x_p)$ is a p -vector, $dx = dx_1 \dots dx_p$ is a p -dimensional volume element and $i = (i_1, \dots, i_q)$ is a multi index. Using the rule

$$\frac{\delta p(x, i)}{\delta p(y, j)} = \delta(x - y) \delta_{ij} \quad (19)$$

with the Kronecker delta symbol $\delta_{ij} = 1, i = j; \delta_{ij} = 0$ otherwise, we obtain

$$p_B(x, i, \lambda) = Z^{-1} p_0(x, i) e^{\lambda [f_{12} F_2(x, i) - F_1(x, i)]} \quad (20)$$

$$Z(\lambda) = \sum_i \int p_0(x, i) e^{\lambda [f_{12} F_2(x, i) - F_1(x, i)]} dx. \quad (21)$$

In the case of $k = 1, \dots, K$ restrictions, we define the K -vectors $\lambda = (\lambda_1, \dots, \lambda_K)$, $f_{12} = (f_{12,1}, \dots, f_{12,K})$, $F_1(x, i) = (F_{11}(x, i), \dots, F_{1K}(x, i))$, $F_2(x, i) = (F_{21}(x, i), \dots, F_{2K}(x, i))$ and use the scalar product notation $x \cdot y = \sum_k x_k y_k$ and the Hadamard product $x y = (x_1 y_1, \dots, x_K y_K)$.

Thus, the general form of the Boltzmann distribution is

$$p_B(x, i, \lambda) = Z^{-1} p_0(x, i) e^{\lambda \cdot [f_{12} F_2(x, i) - F_1(x, i)]}. \quad (22)$$

4.1 Example

We assume 2 groups, $i = 1, 2$, with conditional restrictions $E(X^2|I = 1) = 1, P(0 < X < 1|I = 2) = 0.4$. This can be implemented by defining the vector functions

$$\begin{aligned} f_{12} &= (1, 0.4) \\ F_1(x, i) &= (x^2 \chi_1(i), \chi_{(0,1)}(x) \chi_2(i)) \\ F_2(x, i) &= (\chi_1(i), \chi_2(i)). \end{aligned}$$

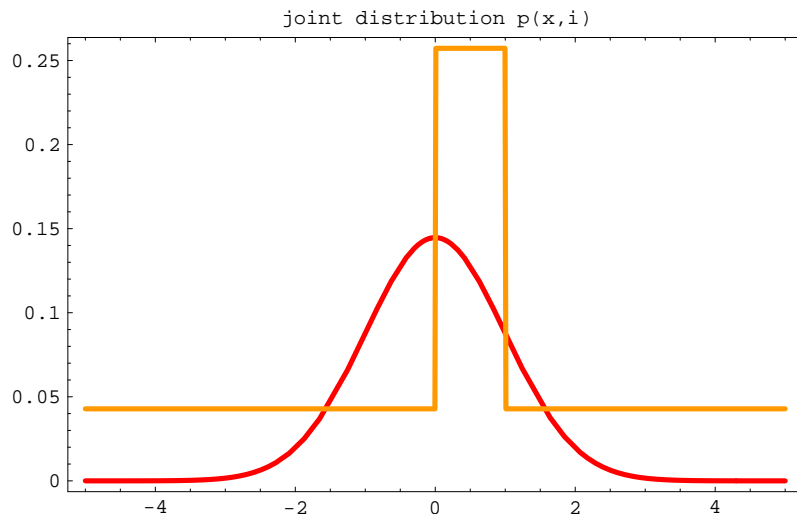


Figure 4: Boltzmann distribution $p(x, i)$ for the restrictions $E(X^2|I = 1) = 1, P(0 < X < 1|I = 2) = 0.4$.

Furthermore, the uniform prior $p_0(x, i) = \frac{1}{2}U(x; [-5, 5])$ was used. The result is displayed in fig. (4). In group $i = 1$, we obtain a gaussian distribution due to a quadratic moment restriction, whereas in group $i = 2$ a piecewise constant density results. In applications, one is interested in the posterior distribution $p(i|x)$, e.g. in categorical regression or discriminant analysis. Using the Bayes formula, one obtains the response functions as displayed in fig. 5. They can be interpreted as Bayesian discriminant functions for the groups numbered by $i = 1, 2$. In the ranges $A = [-1.56, 0]$ and $B = [1, 1.56]$, the posterior probability is larger for group 1, thus we obtain a quite complicated assignment rule (fig. 5).

5 Numerical considerations

The use of functional derivatives leads to the remarkable result, that the infinite dimensional probability density $p(x)$ can be computed explicitly without using a finite dimensional parametrization $p(x, \theta)$, where θ is a parameter vector. In the discrete variable case, one has only the finite or countable set of probabilities $p(i) = p_i$. This usually leads to the idea of representing the

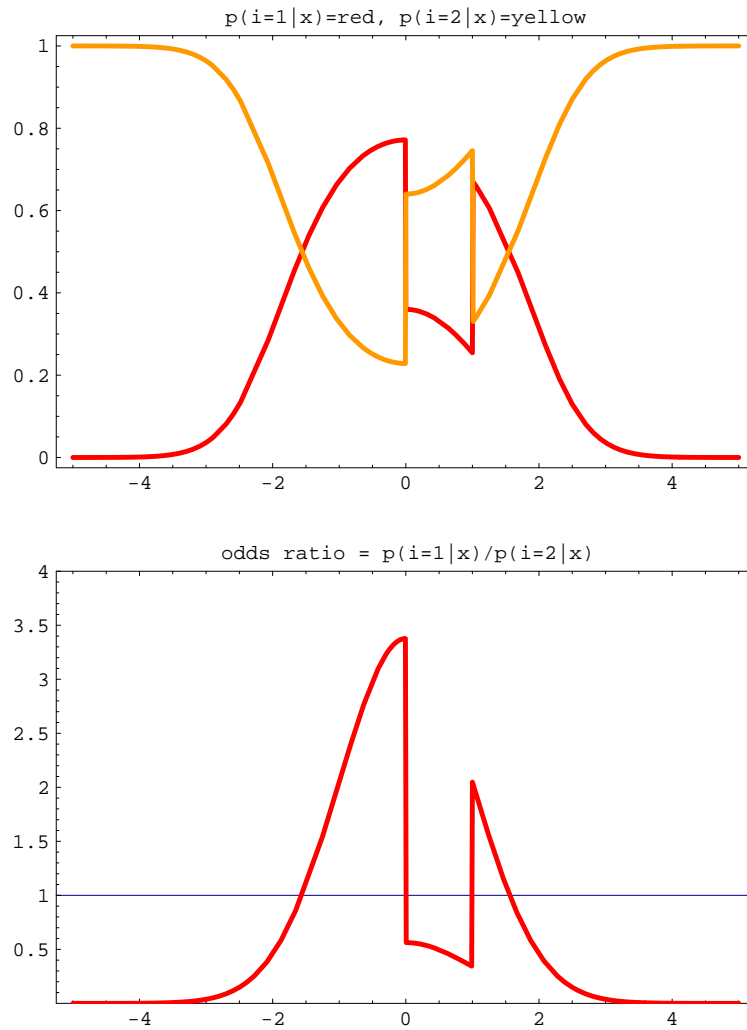


Figure 5: Posterior distribution $p(i|x)$ as function of x (above), odds ratio (below).

continuous density by some parametrization, such as the simple function

$$p(x; p_i) = \sum_i (p_i / \Delta x_i) \chi_{A_i}(x)$$

$$A_i = (x_i, x_i + \Delta x_i]$$

with a partition A_i of the real axis. Then, integrals such as $\int p \log p dx$ degenerate to sums $\sum p_i \log p_i$. This is only an approximation, however, and as shown in the preceding sections, not necessary at all.

Using the explicit form of the maximum entropy distribution (16) we just have to compute the minimum of the concentrated Lagrangian $L^*[\lambda] = \log(Z)$. Thus the key problem is the computation of the p -dimensional integral in the partition function

$$Z(\lambda) = \int p_0(x) e^{\lambda \cdot [f_{12} F_2(x) - F_1(x)]} dx.$$

If the prior is gaussian, one can easily use Gauss-Hermite integration. If not, one can insert a model gaussian $\phi(x; \mu, \Sigma)$ with a suitable choice of parameters. Using the transformed Gauss-Hermite sample points $\xi_l = \mu + \Sigma^{1/2} \zeta_l$, $l = (l_1, \dots, l_p)$, where $\zeta_l = (\zeta_{l_1}, \dots, \zeta_{l_p})$ are the standardized sample points, $\Sigma^{1/2}$ is a matrix square root, and weights $w_l = w_{l_1, \dots, l_p} = w_{l_1} \dots w_{l_p}$ we obtain the Gauss-Hermite approximation

$$Z(\lambda) \approx \sum_l w_l \frac{p_0(\xi_l)}{\phi(\xi_l; \mu, \Sigma)} \exp(\lambda \cdot [f_{12} F_2(\xi_l) - F_1(\xi_l)]). \quad (23)$$

If p_0 is gaussian, the choice $\mu = \mu_0$, $\Sigma = \Sigma_0$ leads to an important simplification. If not, one can choose μ and Σ in order to make $p_0(x)/\phi(x; \mu, \Sigma)$ as flat as possible. Generally, it is important to shift the sample points ξ_l to regions with a high contribution of the integrand. Alternatively, one may use Gauss quadrature or other numerical integration formulas.

The minimization of $L^*(\lambda)$ was achieved using quasi-Newton methods such as the BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm (cf. 9). This method avoids the computation of the Hessian $L^{*''}$ by using a secant update J_m . The optimal estimate of lambda is found by iterating

$$\lambda_{m+1} = \lambda_m + J_m^{-1} L^{*'}(\lambda_m) \quad (24)$$

with the initial choice $J_0 = I_K$ (K -dimensional unit matrix), $\lambda_0 = 0_k$ or some other starting values. The algorithm is stopped if $|\lambda_{m+1} - \lambda_m| < \epsilon_1$ and

$\|L^{*'}(\lambda_m)\| < \epsilon_2$ for suitable choices of ϵ_i , e.g. 10^{-4} . One can use numerical derivatives of $L^*(\lambda)$ or analytical gradients

$$L_k^{*'} = E[f_{12,k}F_{2k}(X) - F_{1k}(X)], k = 1, \dots, K.$$

Using the expectation value, one must again calculate the integrals by numerical quadrature.

If one uses the second derivative (Hessian), either numerically or analytically

$$L_{kk'}^{*''} = E[f_{12,k}F_{2k}(X) - F_{1k}(X)][f_{12,k'}F_{2k'}(X) - F_{1k'}(X)], \\ k, k' = 1, \dots, K,$$

the classical Newton algorithm is obtained.

6 Conditionally Gaussian models

It is interesting to explore the models, when the information is given in form of conditional moments, e.g. the conditional mean and variance. For example, in discriminant analysis it is assumed that the groups $i = 1, \dots, G$ are normally distributed, i.e.

$$E[X|I = i] = \mu_i \\ \text{Var}[X|I = i] = \Sigma_i.$$

Thus we set the restrictions ($i' = 1, \dots, G$)

$$F_{1i'}(x, i) = (x, x^2 - \mu_{i'}^2)\chi_{i'}(i) \quad (25)$$

$$F_{2i'}(x, i) = (1, 1)\chi_{i'}(i) \quad (26)$$

$$E[F_{1i'}]/E[F_{2i'}] = (\mu_{i'}, \Sigma_{i'}) := f_{12i'}. \quad (27)$$

The resulting joint probability is given explicitly

$$p(x, i, \lambda) = Z^{-1} \exp\left[\sum_{i'} \lambda_{i'}(f_{12i'}F_{2i'}(x, i) - F_{1i'}(x, i))\right] \\ = p(x|i)p(i) = \phi(x; \mu_i, \Sigma_i) p(i) := \phi_i p(i).$$

Thus imposing two moment restrictions leads to the conditionally gaussian model. For the purpose of prediction, the posterior probabilities are computed by the Bayes formula

$$\begin{aligned} p(i|x) &= \frac{p(x|i)p(i)}{p(x)} = \frac{p(x|i)p(i)}{\sum_j p(x|j)p(j)} \\ &= \frac{1}{\sum_j \frac{p(x|j)p(j)}{p(x|i)p(i)}} = \frac{1}{\sum_j \frac{\phi_j p(j)}{\phi_i p(i)}}. \end{aligned}$$

Especially in the case $G = 2$ (dichotomous variable $i = 1, 2$) we obtain

$$\begin{aligned} p(1|x) &= \frac{p(x|1)p(1)}{p(x)} = \frac{p(x|1)p(1)}{p(x|1)p(1) + p(x|2)p(2)} \\ &= \frac{1}{1 + \frac{p(x|2)p(2)}{p(x|1)p(1)}} = \frac{1}{1 + \frac{\phi_2 p(2)}{\phi_1 p(1)}}. \end{aligned}$$

Explicitly, the exponent of the quotient $\frac{\phi_2}{\phi_1}$ of the two gaussians is

$$\frac{1}{2}[(x - \mu_1)' \Sigma_1^{-1} (x - \mu_1)' - (x - \mu_2)' \Sigma_2^{-1} (x - \mu_2)'],$$

leading to the linear and quadratic discriminant functions

$$d_1(x) = (\mu_2 - \mu_1)' \Sigma^{-1} x + \dots \quad (28)$$

(equal variances)

$$d_2(x) = \frac{1}{2} x' (\Sigma_1^{-1} - \Sigma_2^{-1}) x + (\mu_2' \Sigma_2^{-1} - \mu_1' \Sigma_1^{-1}) x + \dots \quad (29)$$

(unequal variances).

Comparing with

$$\begin{aligned} l(x) &= \frac{1}{1 + e^{-x}} \\ p(i = 1|x) &= [1 + \alpha e^{-\beta' x}]^{-1} \end{aligned}$$

it is seen that the logistic function naturally occurs in the context of conditionally gaussian models. In the realistic case of unequal variances, a quadratic generalized logistic regression of the form

$$p(i = 1|x) = [1 + \alpha e^{x' \Gamma x - \beta' x}]^{-1} \quad (30)$$

is appropriate.

6.1 Example

$N = 200$ data points (x, i) were simulated from two populations with probabilities $p(1) = p(2) = 0.5$. The continuous random variable X was drawn from the uniform distribution $U[0, 1]$ if $i = 1$ and $U[1, 2]$ if $i = 2$. For the density estimation problem, only the sample means and variances $\bar{x}_1 = 0.512, s_1^2 = 0.0767, \bar{x}_2 = 1.489, s_2^2 = 0.0864$ were substituted for the population moments (the true values are 0.5, 1/12=0.0833, 1.5, 1/12). The restrictions were implemented as in eqn. (25). The solution is displayed in fig. 6. Since the variances are nearly equal, the result is very similar to a logistic regression. Using data with unequal variances, one obtains fig. 7. The shape of the nonlinear response function strongly deviates from the logistic form. In the range of about $x = -0.5, \dots, 1$, population 1 dominates, whereas for higher and lower x values, the posterior $p(i = 2|x)$ is higher. This stems from the fact that a normal distribution with higher variance dominates in the outer tails of the distribution. From the data set one can compute higher order moments and use more restrictions. Adding the values of conditional skewness and kurtosis to the restrictions, a Boltzmann distribution with exponents up to 4th order is obtained (fig. 8). The data fit is better and the response function more sharply discriminates between the groups $i = 1, 2$.

7 Application: credit scoring

In this section the maximum entropy algorithm is applied to the problem of credit scoring. We seek for variables which serve to predict the ability of credit customers to repay a credit. The data set was taken from a south german bank ⁴ and contains 21 variables of continuous, ordinal and nominal scales. In this context I concentrate on the continuous variables $x = \text{creditsum}$ (in Deutsche Mark) and $y = \text{age}$ (in years). The credit variable is $i = 1$ (if the credit was paid back) and $i = 0$ otherwise. The data set contains $n = 1000$ subjects and $n_1 = 700$ loans were repayed correctly. The means and standard deviations in the groups $i = 0, 1$ are given by $\bar{x}_0 = 3938.13$ DM, $s_{x0} = 3535.82$ DM, $\bar{x}_1 = 2985.44$ DM, $s_{x1} = 2401.5$ DM, $\bar{y}_0 = 33.96$ years, $s_{y0} = 11.2252$ years, $\bar{y}_1 = 36.22$ years, $s_{y1} = 11.3474$ years. An in-

⁴The data set was discussed in (10), and is available at <http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit.html>.

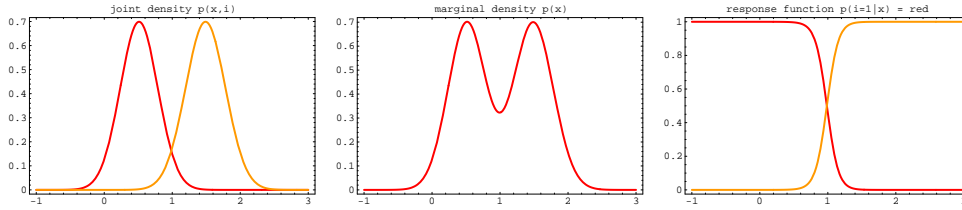


Figure 6: Joint $p(x, i)$, marginal $p(x)$ and posterior distribution $p(i|x)$ as function of x . Two conditional moments (Gaussian case).

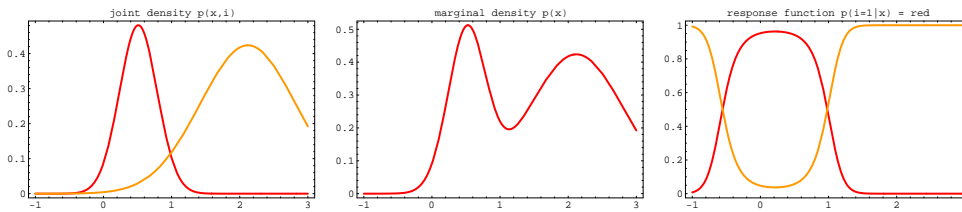


Figure 7: Joint $p(x, i)$, marginal $p(x)$ and posterior distribution $p(i|x)$ as function of x . Unequal variances.

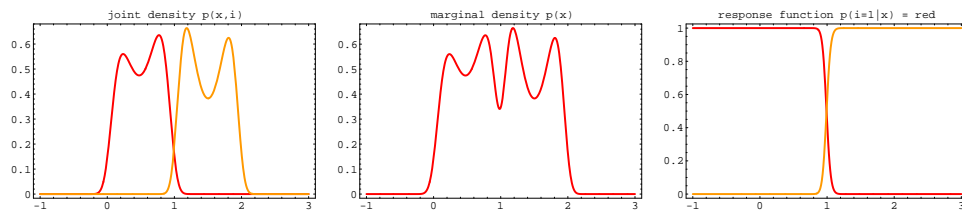


Figure 8: Joint $p(x, i)$, marginal $p(x)$ and posterior distribution $p(i|x)$ as function of x using 4 moments (mean, variance, skewness and kurtosis).

spection of fig. (9) shows that the data are strongly skewed. Nevertheless we start with a model using the conditional moments $E[X|I = i]$, $\text{Var}(X|I = i)$, $E[Y|I = i]$, $\text{Var}(Y|I = i)$; $i = 0, 1$ and prior probability $p_1 = E[\chi_1(I)]$ leading to a conditionally gaussian model. In order to use the information in the data, we substitute the sample moments for the expectation values, but continue to use the notation $p(x, y, i)$ for the estimated density function.

As discussed in sect. 6, the resulting posterior distribution is of logistic type, but with a quadratic exponent (discriminant function). It is displayed in fig. (12). The different standard deviations for the credit sum $s_{x0} = 3535.82$ DM, $s_{x1} = 2401.5$ DM (fig. 10) are mirrored in the nonlinear response function (12). In a range of credit sum up to about 7000 DM, the probability of successful repayment is higher, whereas for higher values of the credit sum, problems with repaying dominate. Note that the prior probabilities $p_0 = 0.3$ and $p_1 = 0.7$ must be considered as well.

Using 4 moments the skewness of the distributions $p(x|i)$, $p(y|i)$ is modeled much better (fig. 11). The response function for x now favors good credits up to about DM 12000, which is consistent with the prior $p_1 = 0.7$ and the slightly higher density $p(x|i = 0)$ (fig. 11, left). From this point on, the bad credits dominate.

The response functions of $p(i|y = age)$ are somewhat different. In the gaussian case, the variances are nearly equal (fig. 10, right), leading to approximate logistic behavior (fig. 12, right). Using higher moments, the skewness differences of $p(y|1)$ and $p(y|2)$ are modeled much better (fig. 11, right), leading to a strongly nonlinear response function (fig. 13, right).

For comparison, kernel density estimates of $p(x, y|i)$ were computed using gaussian kernels. Using the Bayes formula, posterior probabilities $p(i|x, y)$ are obtained. The result is shown in fig. 14. The curves for credit sum are similar to the maxent solution, but showing more detail (using a larger bandwidth, stronger smoothing can be obtained). It is remarkable, that the response function $p(i|y = age)$ is different in the tails of the distribution. Here, the bad credits dominate, as can also be seen from the marginal densities (fig. 9). Of course, there are only few data points in these regions.

The simultaneous influence of the independent variables x, y can be seen from the response surface $p(i|x, y)$ (fig. 15), which is quite complex. For comparison, the kernel density result is displayed showing even more details (fig. 16).

Finally, we show the gaussian case with equal variances in credit sum x (the variances $\sigma_{x0}^2 = \sigma_{x1}^2$ were set equal to the pooled sample variance $s_x^2 =$

$0.3s_{x_0}^2 + 0.7s_{x_1}^2$). The resulting response functions are the well known logistic regression curves.

For comparison, the percentages of correct and misclassification are displayed using the maximum a posteriori probability (MAP) prediction rule

$$\hat{i} = \arg \max_i p(i|x, y) \in \{0, 1\} \quad (31)$$

All data points were used for prediction and cross tabulated with the true class $i = 0, 1$. The error rates $p_{01} + p_{10}$ are displayed in table 1. The prediction rules perform only slightly better than the simple maximum a priori rule $\arg \max_i p(i) = 1$ (line 1). As expected, maxent with 2 moments is equivalent (up to rounding errors) to Bayesian discriminant analysis (DA; computed with SAS/JMP). Equal variances correspond to linear DA (lines 4, 7), whereas unrestricted variances are equivalent to quadratic DA (lines 2, 8; cf. equ. 28).

The solution with four multivariate moments (lines 5, 6) is better than maxent/2, but only equal to the linear discriminant analysis result (=maxent/2, equal variances). Using a uniform prior $p(x, y, i) \propto U[0, 20000] * U[0, 100]$ somewhat improves the error rate (line 3).

The nonparametric regression is best, because it uses all available information. It should be noted, that no general conclusions can be drawn from one data set.

Discussion The example shows, that real data may require modifications from the usual logistic scenario in two respects:

1. The variances in the groups are unequal, leading to quadratic exponents in the logistic functions.
2. Skewness and kurtosis of the conditional distribution leads to cubic and quartic effects.

Of course, one may prefer kernel density estimates of the posterior distributions $p(i|x, y)$, but the maximum entropy distribution attains a parametric form (exponential density) which summarizes the information contained in simple moment information (means, variances, skewness, kurtosis, etc.).

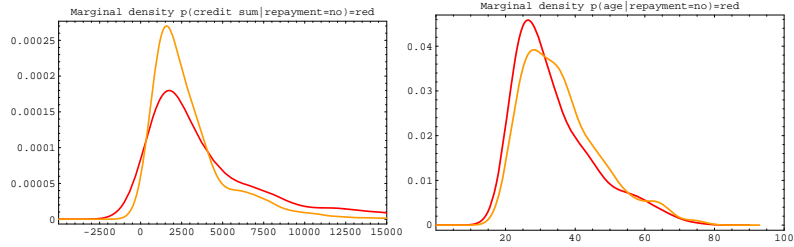


Figure 9: Credit scoring: Kernel density estimates of the conditional densities $p(x|i)$ (x =credit sum, left), $p(y|i)$ (y =age, right), $i = 0, 1$ (repayment no/yes).

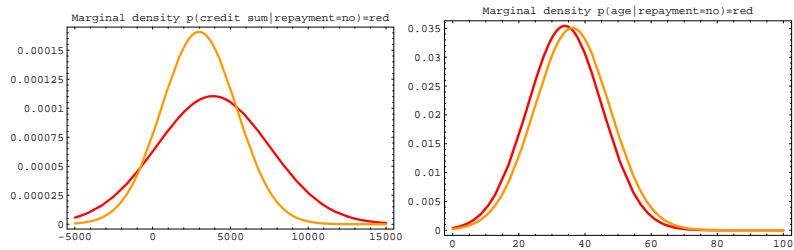


Figure 10: Credit scoring: Maximum entropy estimates of the conditional densities $p(x|i), p(y|i), i = 0, 1$ using 2 moments (gaussian distribution).

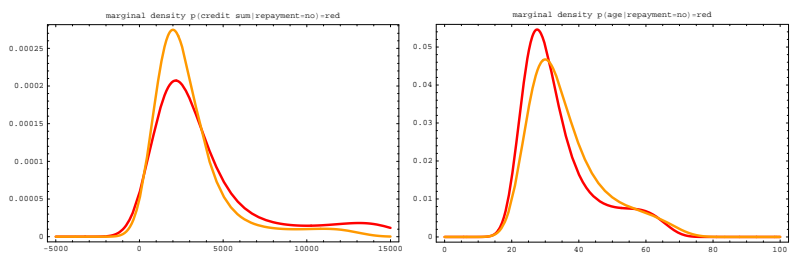


Figure 11: Credit scoring: Maximum entropy estimates of the conditional densities $p(x|i), p(y|i), i = 0, 1$ using 4 moments (exponential distribution).

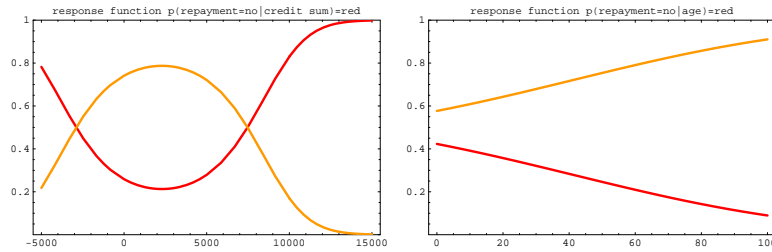


Figure 12: Credit scoring: Response function $p(i|x), p(i|y), i = 0, 1$ using 2 moments (gaussian distribution).

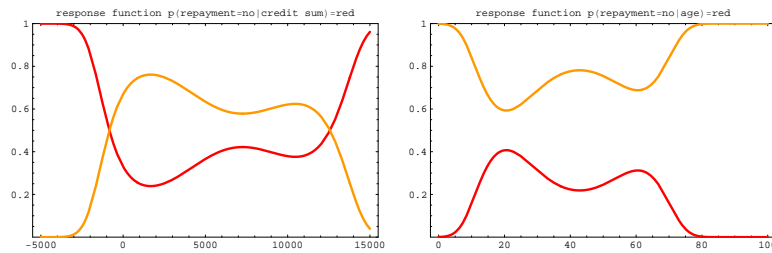


Figure 13: Credit scoring: Response function $p(i|x), p(i|y), i = 0, 1$ using 4 moments (exponential distribution).

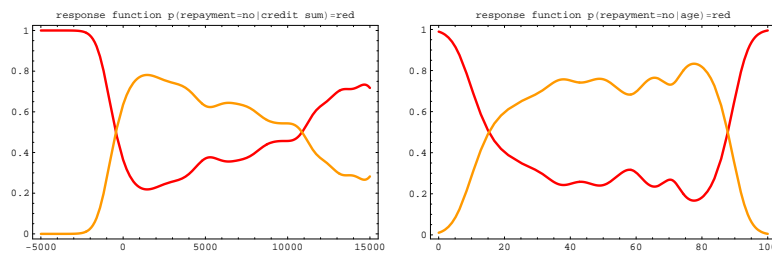


Figure 14: Credit scoring: Response function $p(i|x), p(i|y), i = 0, 1$ using kernel density estimates.

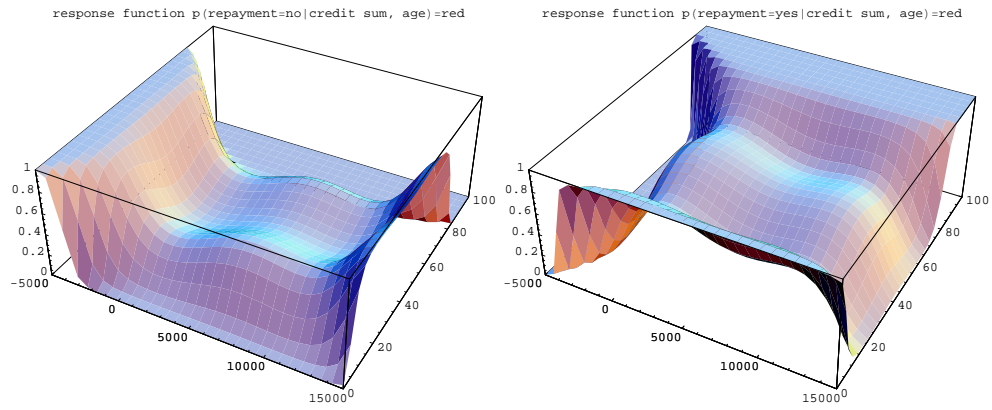


Figure 15: Response function $p(i|x, y), i = 0, 1$ using 4 moments (exponential distribution).

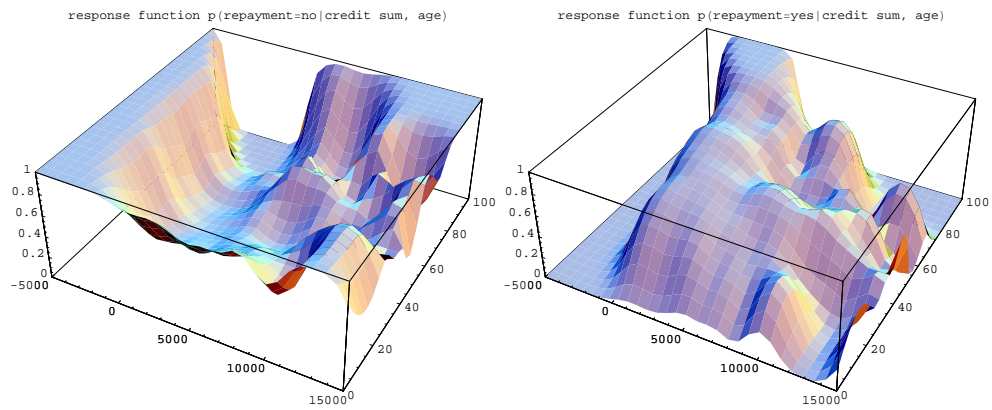


Figure 16: Response function $p(i|x, y), i = 0, 1$ using kernel density estimates.

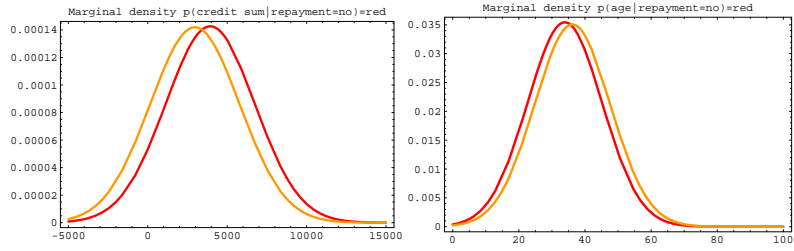


Figure 17: Equal variances $\sigma_{x1} = \sigma_{x2}$: Maximum entropy estimates of the conditional densities $p(x|i), p(y|i), i = 0, 1$ using 2 moments (gaussian distribution).

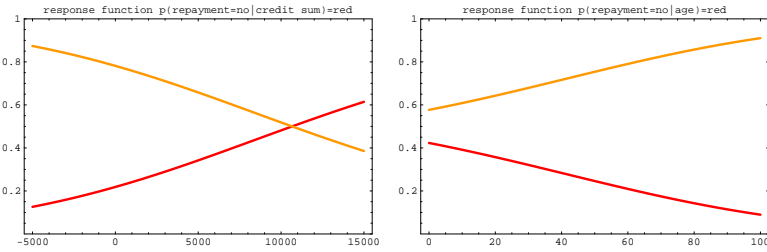


Figure 18: Equal variances $\sigma_{x1} = \sigma_{x2}$: Logistic response functions $p(i|x), p(i|y), i = 0, 1$.

8 Conclusion

We have demonstrated how moment restrictions for mixed continuous-discrete random variables can be used to compute the joint maximum entropy distribution of the variables. The explicit form of the Boltzmann distribution was found without parametrization or discretization. Only the Lagrange parameters (intensities) must be found using numerical procedures. Conditional restrictions using two moments lead to the well known logistic response functions, but we advocate the use of quadratic regressions due to unequal variances in empirical data. Higher order moment information such as skewness and kurtosis can be considered, leading to exponential densities with cubic and quartic terms. In contrast to kernel density approaches, the method also works if only summary statistics, and not a complete data set, are available.

number	method	error $p_{01} + p_{10}$
1	base rate $p(i = 1) = 0.7$: predict $i = 1$	0.3
maximum entropy		
2	2 moments	0.298
3	2 moments, prior $U[0, 20000] * U[0, 100]$	0.289
4	2 moments, equal variances	0.286
5	4 moments	0.291
6	4 moments, multivariate	0.288
discriminant analysis (SAS/JMP)		
7	linear	0.288
8	quadratic	0.299
9	nonparametric regression	0.275

Table 1: Error rates of several algorithms.

Acknowledgement

I would like to thank Wilhelm Rödter for many discussions on the topic.

References

- [1] Jaynes ET. Information Theory and Statistical Mechanics. *Physical Review*. 1957;106, 108:620–630, 171–190.
- [2] Jaynes ET. Probability Theory – The Logic of Science. Cambridge, UK: Cambridge University Press; 2003.
- [3] Haken H. Synergetics. Berlin: Springer; 1977.
- [4] Guiasu S, Shenitzer A. The Principle of Maximum Entropy. *Mathematical Intelligencer*. 1985;7, 1:42–48.
- [5] Rödder W. Conditional logic and the Principle of Entropy. *Artificial Intelligence*. 2000;117:83–106.
- [6] Rödder W, Meyer CH. Coherent knowledge processing at maximum entropy by SPIRIT. In: *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*. San Francisco, Cal.: Morgan Kaufmann; 2006.
- [7] Rödder W, Reucher E, Kulmann F. Features of the Expert-System-Shell SPIRIT. *Logical Journal of the IGPL*. 2006;14, 3:483–500.
- [8] Kullback S. Information Theory and Statistics. Wiley; 1959.
- [9] Dennis Jr JE, Schnabel RB. Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Englewood Cliffs: Prentice Hall; 1983.
- [10] Fahrmeir L, Hamerle A, Tutz G, editors. *Multivariate Statistische Methoden*. 2nd ed. Berlin, New York: de Gruyter; 1996.