Comparing Default Probability Models

Rudi Zagst [a], Stephan Höcht [b]

[a] *Director, HVB-Institute for Mathematical Finance, Munich University of Technology, Boltzmannstrasse 3, 85748 Garching b. München, Germany. Email: zagst@ma.tum.de, tel.: ++49/89/289-17404 fax: ++49/89/289-17407.*

[b] *HVB-Institute for Mathematical Finance, Munich University of Technology, Boltzmannstrasse 3, 85748 Garching b. München, Germany. Email: hoecht@ma.tum.de, tel.: ++49/89/289-17417, fax: ++49/89/289-17407.*

Abstract

We compare different ways of modeling real-world probabilities of default over a fixed time horizon conditioned on a vector of explanatory variables. Besides a simple logistic regression we introduce an extended version of the logistic regression that allows for modeling further dependencies. Later, we discuss a maximum expected utility approach, which chooses the model measure from a one-parameter family of pareto-optimal measures defined in terms of consistency with the data and a prior measure. We apply this setting to a very general class of utility functions, namely the class of hyperbolic absolute risk aversion (HARA) utility functions. The numerical comparison is based on Fitch Risk's North American Loan Loss Database.

Key words: default probability model, expected utility, hyperbolic absolute risk aversion, logistic regression

JEL classification: C35, E51, G21

Introduction

Regulatory requirements as well as risk management considerations force holders of loan portfolios to face the default risk within their portfolio. According to that, a lot of progress has been made in the field of credit risk management in the last few years. Various rating methodologies and credit risk modeling approaches have been developed in this time. Very often, a large part of a banking institutions' portfolio often consists of loans to medium-size unrated firms. However, institutions' individual exposures to such firms are mostly very small and therefore it is uneconomical to put a great effort in analyzing single obligors. Hence, it is important to have a good model for measuring and controlling the credit risk of such loans.

In this paper various approaches to default probability models will be presented and empirically compared on a complete database of North American firms. We will restrict the description to real-world probability models, namely the linear logistic regression (LLR) and the maximum expected utility approach (MEU), which chooses a model according to consistency with the training data and consistency with a prior distribution. Therefore, the aim of Sections 1 and 2 is to find a model for the probability of default given side-information. In a more mathematical notation, we seek a model

$$q(1|x) = q_{1|x} = Q(Y = 1|X = x),$$

where the random vector X describes the given information, e.g. financial ratios or macroeconomic indicators, $Q(\cdot|X = x)$ is the conditional model probability measure given $X = x$ and the random variable Y indicates default, i.e. $Y = 1$ means default and $Y = 0$ survival. For the LLR modeling approach we will distinguish between a simple approach that only models linear dependencies in the data, like the LLR models in Friedmann and Huang (2003) and Friedmann *et al.* (2003), and an extended approach that allows for modeling further dependences. In contrast to the LLR models and the separable logit models in Friedmann and Huang (2003) and Friedmann *et al.* (2003) the extended logistic regression, which we will introduce in Section 1, enables us to model non-linear dependences and interactions among the variables. The MEU modeling approach in Section 2 follows the ideas of Friedmann and Sandow (2003a) and Friedmann and Sandow (2003b). For the sake of completeness we give all the statements and proves for a conditional model measure $q_{\cdot|x} = (q_{0|x}, q_{1|x})$. In contrast to the existing MEU models in the literature, we will not only apply this approach to an investor with a logarithmic utility function as in Friedmann and Huang (2003) and Friedmann *et al.* (2003) but to a more general class of utility functions, the hyperbolic absolute risk aversion (HARA) utility functions. Finally, in Section 3 we will compare the different modeling approaches empirically based on a dataset of North American firms.

1 Logistic Regression

We will start with a very simple modeling approach that only allows for linear dependencies in the data, the simple linear logistic regression (LLR). The basic equation of simple linear logistic regression models is given by

$$q_{1|x} = \frac{1}{1 + e^{-\beta^T x}}, \tag{1}$$

where the left-hand side is the probability of default conditioned on given explanatory variables and the right-hand side is a logistic function evaluated at the linear combination $\beta^T x$. The parameter vector β can be estimated via a maximum likelihood estimation which leads to a numerically robust procedure as we only have to solve an unconstrained strictly concave optimization problem. Another advantage of simple LLR models is that the distribution of a binary random variable Y is already completely determined by the probability $q_{1|x}$ and there are no distributional assumptions about the explanatory variables in the model. The only assumptions made by the model refer to the form of the regression equation.
Due to its very simple dependence structure the simple LLR models have some negative aspects. As one can see from (1) it is only a transformation of a linear function in x. In other words, if the data includes complex relationships, the models which were built with this simple LLR technique are to stiff to match the data and are hence unable to handle non-linear variables, non-monotonic dependences and interactions among the explanatory variables. According to the limitations of the simple LLR approach we allow for further dependences in the data. The extended LLR modeling approach, which is an extension of the separable logit model (see, e.g., Friedmann *et al.* (2003)), is based on the following equation:

$$q_{1|x} = \frac{1}{1 + e^{\beta^T g(x)}},$$

where g is a function of the variables x and β is the corresponding parameter vector. Again, the parameters are estimated by a maximum likelihood procedure and the extended logistic regression has the same positive aspects as the simple logistic regression. Furthermore, it is now possible to model more complex dependence structures in the data, e.g. non-linear and non-monotonic dependences or interactions between variables, by choosing different functions $g_j(x)$:

- Linear dependences: $g_j(x) = -x_j$
- Quadratic dependences: $g_{i,j}(x) = -x_i x_j$
- Cylindrical dependences: $g_{j,a}(x) = -e^{\frac{(x_j - a)^2}{\sigma^2}}$ for a given $\sigma^2 > 0$.

Here, x_j denotes the j-th component of the vector x.

2 Maximum Expected Utility Approach

We now want to give a class of default probability models which enables us to model the same dependence structures as in the extended LLR models, but which is based on a more general background. For this we describe the Maximum Expected Utility (MEU) approach as introduced in Friedmann and Huang (2003), Friedmann and Sandow (2003a) and Friedmann *et al.* (2003). The basic idea behind this modeling approach is to maximize the expected utility of an investor who bets on default events. The allocation of the investor's money on the different betting events and hence the utility of the investor will depend on his beliefs expressed by the model measure $q_{\cdot|x}$. As we do not know the true probabilities we approximate the expected utility by using the empirical probabilities from a given sample set. The aim is to find a model measure that is on the one hand able to handle different dependences in the data and on the other hand as close as possible to a given prior distribution. We will build such pareto-optimal measures using consistency measures with respect to the data and the prior distribution. We assume that the investor acts rational with a strictly concave, twice differentiable and strictly monotonic increasing utility function. Furthermore, according to Cover and Thomas (1991) and Friedmann and Sandow (2003a), we define a conditional horse race to be a market, characterized by the discrete random variable Y with m possible states, in which investors can place a bet that $Y = y$ after learning the value of $x \in \mathbb{R}^d$. The payout of the bet is $O_{x,y}$ for each dollar wagered if $Y = y$, and 0 otherwise. In our case, $Y \in \{0, 1\}$ where 1 indicates default and 0 non-default. We assume that the investor allocates, after learning the value of x, $b_{y|x}$ to the event $Y = y$, where $\sum_y b_{y|x} = 1$. If we denote the wealth of the investor after a conditional horse race with W, then $W = b_{y|x} O_{x,y}$ if horse y won the race.

2.1 Optimal Betting

Now, let p_x denote the true probability for $X = x$ and $p_{y|x}$ the true conditional probability for $Y = y$ given $X = x$. As mentioned above the true measures p_x and $p_{y|x}$ are unknown and we approximate them by taking the empirical measures of a test sample instead. In a first step, let us assume that the model $q_{\cdot|x} = \left(q_{0|x}, q_{1|x}\right)$ is known and that the investor wants to maximize the expected utility of his wealth. Therefore, he places his bets according to

$$b^*_{\cdot|x} = \arg \max_{\left\{b_{\cdot|x}: \sum_y b_{y|x}=1\right\}} \sum_x p_x \sum_y q_{y|x} U\left(b_{y|x} O_{x,y}\right). \tag{2}$$

It can be easily shown (see, e.g., Theorem 7 of Friedmann and Sandow (2003b)) that the optimal betting weights in (2) are given by

$$b^*_{y|x} = \frac{1}{O_{x,y}} \left(U'\right)^{-1} \left(\frac{\lambda_x}{p_x q_{y|x} O_{x,y}}\right), \tag{3}$$

where λ_x is the solution of

$$1 = \sum_y b^*_{y|x} = \sum_y \frac{1}{O_{x,y}} \left(U'\right)^{-1} \left(\frac{\lambda_x}{p_x q_{y|x} O_{x,y}}\right), \tag{4}$$

if there exists a solution to (4). In this case, the solution is unique and the optimal allocation $b^*_{\cdot|x}(q)$ continuously depends on q [1]. In a second step, we have to specify the model $q_{\cdot|x}$. We will do this by the examination of the efficient frontier of pareto-optimal measures with respect to consistency with the training data and a prior distribution. To obtain these consistency measures we divide the data set in a training data set T which is a (randomly generated) subset of the (full) data set containing N observations and a hold-out data set which consists of the remaining observations [2]. Finally, we will choose the measure from the efficient frontier which has the highest expected utility on the hold-out data set as our model measure.

2.2 Consistency with a prior distribution

To define the consistency with a prior distribution $q^0_{\cdot|x}$ we will need the concept of entropy and relative entropy known from information theory (see, e.g., Cover and Thomas (1991)), respectively a generalization of these information theoretic terms, the generalized conditional relative entropy (see, e.g., Friedmann and Sandow (2003b)). Given a utility function U and a system of outcomes O the generalized conditional relative entropy is given by

$$\begin{aligned}\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x}) &= \sum_x p_x \sum_y q_{y|x} \left[U\left(b^*_{y|x}(q_{y|x})O_{x,y}\right) - U\left(b^*_{y|x}(q^0_{y|x})O_{x,y}\right)\right] \\ &= E_{q_{\cdot|x}}\left[U\left(b^*_{y|x}(q_{y|x})O_{x,y}\right)\right] - E_{q_{\cdot|x}}\left[U\left(b^*_{y|x}(q^0_{y|x})O_{x,y}\right)\right]. \end{aligned} \tag{5}$$

The generalized conditional relative entropy is a measure for the difference in

[1] The condition $\frac{1}{(U')^{-1}(0)} < \sum_y \frac{1}{O_{x,y}} < \frac{1}{\max\{0,(U')^{-1}(\infty)\}}$ is necessary and sufficient for the existence of a solution to (4), which is fulfilled by the most common utility functions (a proof in the case of non-conditional probabilities can be found in Appendix B of Friedmann and Sandow (2003b)).

[2] We used 80% of the data as training data.

expected utility under the model measure $q_{\cdot|x}$ of an investor who acts optimal according to the model measure $q_{\cdot|x}$ and an investor who acts optimal according to the prior measure $q^0_{\cdot|x}$. Some very useful properties, which we will need later, are stated in the following theorem (see Theorem 8 of Friedmann and Sandow (2003b)):

Theorem 2.1 *For $\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x})$ from (5) we have:*

(1) $\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x}) \geq 0$ with equality if and only if $q_{\cdot|x} = q^0_{\cdot|x}$ (information inequality)
(2) $\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x})$ is a strictly convex function of $q_{\cdot|x}$.

For a proof see Appendix A.

Now, we can define the consistency of the model measure $q_{\cdot|x}$ with the investor's prior beliefs, expressed by the prior measure $q^0_{\cdot|x}$, as the Generalized Conditional Relative Entropy $\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x})$.

2.3 Consistency with the data

The consistency of the model measure $q_{\cdot|x}$ with the (training) data is given by the measure $\Delta_{data}(q_{\cdot|x}||p_{\cdot|x})$. This consistency measure will be expressed in terms of expectations of a feature vector $f(y,x) = (f_1(y,x), \ldots, f_J(y,x))^T \in \mathbb{R}^J$ where each feature f_j is a mapping from $\mathbb{R}^{d+1}$ to $\mathbb{R}$ and d denotes the length of the vector x. The use of features is a very popular concept in statistical learning theory. With the help of features we try to model the different possible relationships in the data. Later, we will introduce different kinds of features, one for linear relationships in the data, one for quadratic relationships and one with an exponential term. In our context the introduction of more features will bring more feature constraints and therefore more consistency with the data. We define

$$\Delta_{data}(q_{\cdot|x}||p_{\cdot|x}) := N c^T \Sigma^{-1} c \geq 0, \tag{6}$$

where Σ is the empirical covariance matrix of f, $c = (c_1, \ldots, c_J)$ with

$$c_j = E_{q_{\cdot|x}}[f_j] - E_{p_{\cdot|x}}[f_j],$$

$E_{q_{\cdot|x}}[f_j] = \sum_x p_x \sum_y q_{y|x} f_j(y,x)$, $E_{p_{\cdot|x}}[f_j] = \sum_x p_x \sum_y p_{y|x} f_j(y,x)$, and N the number of observations in the training data set [3]. Note that $\Delta_{data}(q_{\cdot|x}||p_{\cdot|x})$ from (6) is the so-called Mahalanobis distance (see, e.g., Fahrmeir *et al.* (1996)).

[3] For a derivation of (6) from a large-sample distribution of a vector of sample-averaged features see Friedmann and Sandow (2003a).

Our feature vector $f(y,x)$, that we will use for the MEU modeling approach, consists of three different types of features similar to the functions g_j for the extended LLR approach, i.e. $f(y,x) = (f_L(y,x), f_Q(y,x), f_C(y,x))^T$ where f_L, f_Q and f_C are vectors themselves, defined as follows:

(1) Linear features $f_L(y,x) = (f_j(y,x))_{j=1,\dots,m}$, where

$$f_j(y,x) = (y - \frac{1}{2})x_j = (\frac{1}{2} - y)g_j(x),$$

with $y \in \{0,1\}$ and x_j the j-th coordinate of x.

(2) Quadratic features $f_Q(y,x) = (f_{i,j}(y,x))_{1 \le i \le j \le m}$, where

$$f_{i,j}(y,x) = (y - \frac{1}{2})x_i x_j = (\frac{1}{2} - y)g_{i,j}(x).$$

(3) Cylindrical kernel features $f_C(y,x) = (f_{j,a}(y,x))_{\substack{a \in \{0,0.25,0.5,0.75,1\} \\ j=1,\dots,m}}$, where

$$f_{j,a}(y,x) = (y - \frac{1}{2})g_a(x_j) = (\frac{1}{2} - y)g_{j,a}(x),$$

with $g_a(x_j) = e^{-\frac{\left(x_j - a\right)^2}{\sigma^2}}$ and $\sigma = 0.35$ [4].

As mentioned above the linear features model linear relationships in the data, the quadratic features model quadratic relationships and the cylindrical kernel features model some kind of relationship with an exponential term.

We can now combine the concepts of consistency with the data and consistency with the prior distribution and call a measure $q^*_{\cdot|x}$ pareto-optimal if and only if no measure $q_{\cdot|x}$ dominates $q^*_{\cdot|x}$ with respect to $\Delta_{data}(q_{\cdot|x}||p_{\cdot|x})$ and $\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x})$ [5]. The set of all pareto-optimal measures is then called the efficient frontier and we assume that the rational investor selects his model measure on the efficient frontier. To say it in other words, the investor chooses, from a set of measures equally consistent with the prior distribution, the measure that has the highest degree of consistency with the data. The other way round the investor chooses, from a set of measures equally consistent with the data, the measure that has the highest degree of consistency with the prior distribution. The investor makes no assumption about the preferential treatment of one of these concepts.

In the following lemma, which is a generalisation of Lemma 1 in Friedmann and Sandow (2003a) on the case of conditional probabilities, we give an upper

[4] For a better comparability we have chosen a and σ as in Friedmann *et al.* (2003)

[5] For a definition of pareto optimality, dominance and efficient frontier see, e.g., Ingersoll Jr. (1987).

boundary for the consistency with the data of a pareto-optimal measure, which we will need for the formulation of the optimization problem:

Lemma 2.2 *For each pareto-optimal measure $q^*_{\cdot|x}$ we have:*

$$\Delta_{data}(q^*_{\cdot|x}||p_{\cdot|x}) \leq \alpha_0 := N \left(E_{q^0_{\cdot|x}}[f] - E_{p_{\cdot|x}}[f] \right)^T \Sigma^{-1} \left(E_{q^0_{\cdot|x}}[f] - E_{p_{\cdot|x}}[f] \right)$$

with $E_{q^0_{\cdot|x}}[f] = \sum_x p_x \sum_y q^0_{y|x} f(y,x)$ and $E_{p_{\cdot|x}}[f] = \sum_x p_x \sum_y p_{y|x} f(y,x)$.

Proof: For the prior measure $q^0_{\cdot|x}$ we have $\Delta_{data}(q^0_{\cdot|x}||p_{\cdot|x}) = \alpha_0$ and $\Delta_{prior}(q^0_{\cdot|x}||q^0_{\cdot|x}) = 0$. If $\Delta_{data}(q_{\cdot|x}||p_{\cdot|x}) > \alpha_0$, then $q_{\cdot|x} \neq q^0_{\cdot|x}$ and therefore $\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x}) > 0$ (see also Appendix A). Hence $q_{\cdot|x}$ is dominated by $q^0_{\cdot|x}$ and can not be efficient. □

2.4 Optimization Problem

In order to find all pareto-optimal measures we can, according to Lemma 2.2, express this problem for each level α as an optimization problem, i.e. as α ranges form 0 to α_0 we want to solve the following problem for all $q_{\cdot|x}$ with $\Delta_{data}(q_{\cdot|x}||p_{\cdot|x}) = \alpha$.

Problem 1 (Initial Problem) *Let α be fixed with $0 \leq \alpha \leq \alpha_0$.*

$$\textit{Find} \quad \arg\min_{q_{\cdot|x} \in \mathbb{R}^2} \Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x}) \tag{7}$$

$$s.t.\ 1 = \sum_y q_{y|x} \tag{8}$$

$$0 \leq q_{y|x} \tag{9}$$

$$\Delta_{data}(q_{\cdot|x}||p_{\cdot|x}) = \alpha \tag{10}$$

Since (10) is a non-affine equality constraint, Problem 1 is not a standard convex optimization problem (for a definition see, e.g., Boyd and Vandenberghe (2004)). But we can easily derive a standard convex optimization problem which admits the same solutions as Problem 1 (for a proof in the case of non-cond. probabilities see also Appendix B of Friedmann and Sandow (2003a)):

Problem 2 (Initial Strictly Convex Problem) *Let* α *be fixed with* $0 \leq \alpha \leq \alpha_0$.

$$\begin{aligned} \textit{Find} \quad & \arg\min_{q_{\cdot|x} \in \mathbb{R}^2} \Delta_{prior}(q_{\cdot|x} || q^0_{\cdot|x}) \\ s.t. \; 1 = & \sum_y q_{y|x} \\ 0 \leq & \, q_{y|x} \\ \Delta_{data}(q_{\cdot|x} || p_{\cdot|x}) \leq & \, \alpha \end{aligned}$$

Lemma 2.3 *Problem 2 is a strictly convex optimization problem and Problems 2 and 1 have the same unique solution.*

For a proof see Appendix B.

When we solve Problem 2 for different levels of α we get a one-parameter family of pareto-optimal measures, $q^*_{\cdot|x}(\alpha)$, indexed by α, from which we choose the measure with the highest expected utility on the hold-out data set as our model measure (see Subsection 2.1). It is no problem to numerically solve such maximization problems for a one-parameter family of candidate measures. As was shown above, the search for a pareto-optimal measure for a given α leads to a strictly convex optimization problem, Problem 2. We now present and prove two different versions of the dual problem of Problem 2:

Problem 3 (Dual Problem 1) *Let* α *be fixed with* $0 \leq \alpha \leq \alpha_0$.

$$\begin{aligned} \textit{Find } \beta^* = & \arg\max_\beta h(\beta) \\ \textit{with } h(\beta) = & \sum_x p_x \sum_y p_{y|x} U(b^*_{y|x}(q^*_{y|x}) O_{x,y}) - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}} \qquad (11) \\ \textit{where } b^*_{y|x}(q^*_{y|x}) = & \frac{1}{O_{x,y}} (U')^{-1} \left(\frac{\lambda^*_x}{p_x q^*_{y|x} O_{x,y}} \right) \\ \textit{and } q^*_{y|x} = & \frac{\lambda^*_x}{p_x O_{x,y} U'(U^{-1}(G(x,y,q^0_{\cdot|x},\beta,\mu^*)))} \qquad (12) \\ \textit{with } G(x,y,q^0_{\cdot|x},\beta,\mu^*_x) = & \, U(b^*_{y|x}(q^0_{y|x}) O_{x,y}) + \beta^T f(y,x) - \mu^*_x \\ \textit{where } \mu^*_x \textit{ solves } 1 = & \sum_y \frac{1}{O_{x,y}} U^{-1}(G(x,y,q^0_{\cdot|x},\beta,\mu^*_x)) \\ \textit{and } \lambda^*_x = & \left\{ \sum_y \frac{1}{p_x O_{x,y} U'(U^{-1}(G(x,y,q^0_{\cdot|x},\beta,\mu^*_x)))} \right\}^{-1}. \end{aligned}$$

Theorem 2.4 *Problems 2 and 3 lead to the same unique solution $q^*_{\cdot|x}$ of Problem 2.*

For a proof see Appendix C. A proof in the case of non-conditional probabilities can be found in Appendix C of Friedmann and Sandow (2003a). As can be seen from the proof, $h(\beta)$ can be rewritten as

$$h(\beta) = \beta^T E_{p_{\cdot|x}}(f) - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}} - \sum_x p_x \mu^*_x.$$

The first term of the objective function in Problem 3 is the utility of an utility maximizing investor over the training sample. Therefore Problem 3 can be interpreted as a risk-adjusted maximization of the average utility over the training sample. The dual problem, Problem 3, is a J-dimensional unconstrained, concave maximization problem. From (11) we get the following theorem about asymptotic optimality (see also Theorem 2 of Friedmann and Sandow (2003a)):

Theorem 2.5 *For $N \to \infty$ the optimal solution of Problem 2 maximizes the investor's expected utility over the family of measures parameterized by (12).*

Proof:
Follows from (11) and Theorem 2.4. □

2.5 MEU for a HARA Utility Function

After we have run through the whole model building process for a general utility function U in the last subsection, we now want to restrict ourselves to a class of utility functions which is often used when dealing with a risk avers investor. The class of hyperbolic absolute risk aversion (HARA) or linear risk tolerance (LRT) utility functions is given by

$$U(x) = \frac{1-b}{b}\left[\left(\frac{ax}{1-b} + c\right)^b - d\right], \tag{13}$$

for $a, b > 0$ and $x \in \left\{x : \frac{ax}{1-b} + c > 0\right\}$. This class includes most of the common utility functions, e.g. logarithmic, exponential, power and quadratic utility functions. The approach with the class of HARA utility functions is a generalisation of the modeling approach introduced in Friedmann and Huang (2003) and Friedmann *et al.* (2003), who only use a logarithmic utility function. As one can easily show Problem 3 with a HARA utility function is equivalent to the following optimization problem.

Problem 4 (Dual Problem for HARA utility function) *Let α be fixed with $0 \leq \alpha \leq \alpha_0$.*

$$
\begin{aligned}
&\textit{Find } \beta^* = \arg\max_{\beta} h(\beta) \\
&\textit{with } h(\beta) = \sum_x p_x \sum_y p_{y|x} \frac{1-b}{b} \left[\left(\frac{\lambda_x}{a p_x q^*_{y|x} O_{x,y}} \right)^{\frac{b}{b-1}} - d \right] - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}} \\
&\textit{and } q^*_{y|x} = \frac{\lambda_x}{a p_x O_{x,y} \left[\left(\frac{\lambda_x}{a p_x q^0_{y|x} O_{x,y}} \right)^{\frac{b}{b-1}} + \frac{b}{1-b} \left(\beta^T f(y,x) - \mu_x \right) \right]^{\frac{b-1}{b}}}
\end{aligned} \tag{14}
$$

where μ_x solves

$$
\frac{a}{1-b} + c \sum_y \frac{1}{O_{x,y}} = \sum_y \frac{1}{O_{x,y}} \left[\left(\frac{\lambda_x}{a p_x q^0_{y|x} O_{x,y}} \right)^{\frac{b}{b-1}} + \frac{b}{1-b} \left(\beta^T f(y,x) - \mu_x \right) \right]^{\frac{1}{b}} \tag{15}
$$

$$
\textit{and } \frac{1}{\lambda_x} = \sum_y \frac{1}{a p_x O_{x,y} \left[\left(\frac{\lambda_x}{a p_x q^0_{y|x} O_{x,y}} \right)^{\frac{b}{b-1}} + \frac{b}{1-b} \left(\beta^T f(y,x) - \mu_x \right) \right]^{\frac{b-1}{b}}} \tag{16}
$$

.

Unfortunately the equality constraints can not be solved analytically for λ_x and μ_x. Hence, Problem 4 is a constrained optimization problem. There are a lot of optimization routines that solve these constrained optimization problems (see, e.g., Boyd and Vandenberghe (2004)), but in practice it is not possible to solve Problem 4 in an acceptable time as the dimension of the constraints can be very high. Alternatively we could put the equality constraints in the objective function and solve this unconstrained optimization problem with a penalty method. However, as the dimension of the optimization variables of this problem can become very large, this problem would also be very hard to solve in practice. In the following we will discuss two possibilities of solving this problem.

The first idea is to simplify Problem 4 to obtain an unconstrained optimization problem. We have done this with a first-order Taylor expansion. Using the simplifying assumptions $O_{x,y} = O_x$ and $q^0_{y|x} = q^0_x = \frac{1}{m}$ with m denoting the number of possible states of Y, Problem 4 becomes an unconstrained optimization problem:

Problem 5 (First-Order Simpl. Dual Problem for HARA) *Let α be fixed with $0 \leq \alpha \leq \alpha_0$.*

$$
\begin{aligned}
\textit{Find } \beta^* &= \arg\max_{\beta} h(\beta) \\
\textit{with } h(\beta) &= \sum_x p_x \sum_y p_{y|x} \frac{1-b}{b} \left[\left(\frac{\lambda_x}{a p_x q^*_{y|x} O_x} \right)^{\frac{b}{b-1}} - d \right] - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}} \\
\textit{and } q^*_{y|x} &= \frac{1}{2} \left[1 + \left(c + \frac{a}{2(1-b)} O_x \right)^{-b} \beta^T f(y,x) \right] \\
\lambda_x &= \frac{a p_x O_x}{m} \left[\frac{a}{2(1-b)} O_x + c \right]^{b-1} \\
-1 &\leq - \left(c + \frac{a}{2(1-b)} O_x \right)^{-b} \beta^T f(y,x) \\
-1 &\leq \left(c + \frac{a}{2(1-b)} O_x \right)^{-b} \beta^T f(y,x)
\end{aligned}
$$

The second idea is to choose a special utility function from the class of HARA utility functions to simplify Problem 4. This can be done by choosing $b = 0$, which leads to a logarithmic family of utility functions. The resulting optimization problem is an unconstrained, concave optimization problem, which is independent of $O_{x,y}$:

Problem 6 (Dual Problem for U in Logarithmic Family) *Let α be fixed with $0 \leq \alpha \leq \alpha_0$.*

$$
\begin{aligned}
\textit{Find } \beta^* &= \arg\max_{\beta} h(\beta) \\
\textit{with } h(\beta) &= \sum_x p_x \sum_y p_{y|x} \log(q^*_{y|x}) - \sqrt{\frac{\alpha}{\gamma_1^2} \frac{\beta^T \Sigma \beta}{N}} \\
\textit{with } q^*_{y|x} &= \frac{q^0_{y|x} e^{\frac{1}{\gamma_1} \beta^T f(y,x)}}{\sum_y q^0_{y|x} e^{\frac{1}{\gamma_1} \beta^T f(y,x)}}.
\end{aligned}
$$

Lemma 2.6 *Problem 4 with $a = \gamma_1^{\frac{1}{b}}$, $b = 0$, $c = \gamma_1^{\frac{1}{b}} \gamma$ and $d = \gamma_1 - \gamma_2 b$ is equivalent to Problem 6.*

For a proof see Appendix D.

In the practical considerations for modeling default probabilities we will mostly restrict ourselves to a logarithmic utility function of the form $U(W) = \log(W)$; $\forall W > 0$. Moreover we set

$$p_x = \frac{1}{N} \sum_{\{i \in T : x_i = x\}} 1,$$

where N is the number of observations in the training data set T [6], and $p_{y|x}$ is the empirical conditional default probability in the training data set, i.e.

$$p_{1|x} = \frac{\sum_{\{i \in T : y_i = 1, x_i = x\}} 1}{\sum_{\{i \in T : x_i = x\}} 1}.$$

Furthermore, we choose the empirical conditional default probability in the (full) data set as prior measure $q^0_{\cdot|x}$ [7]. In practice, we also may not search over the range $\alpha \in (0, \alpha_0)$, but over the range $\alpha \in (0, \alpha_{\text{search}})$, where $\alpha_{\text{search}} = \min(\alpha_0, \alpha_l)$ and α_l is the $100 \cdot l\%$-quantile of a χ^2-distribution with J degrees of freedom. This is possible, because we get for the true expectation μ, under the hypothesis $H_0 : \mu = E_{p_{\cdot|x}}(f)$ that

$$N c^T \Sigma^{-1} c \sim \chi^2_J$$

for a given Σ (see, e.g., Davidson and MacKinnon (1993) or Fahrmeir *et al.* (1996)). So, we search either until we are $100 \cdot l\%$ confident that the true value of c is within the region $N c^T \Sigma^{-1} c \leq \alpha$ or until this region includes $q^0_{\cdot|x}$ and hence $q^*_{\cdot|x}$ is insensitive to further increasing the value of α (see Friedmann and Sandow (2003a)). Let $(x_k, y_k)_{k=1,\ldots,N}$ be the observed (x, y)-pairs. Then we can simplify Problem 6 to the following optimization problem:

Problem 7 (Optimization Problem for an Optimal PD Model) *Let α be fixed with $0 \leq \alpha \leq \alpha_{search}$.*

$$\begin{aligned}
\textit{Find } \beta^* &= \arg\max_{\beta} h(\beta) \\
\textit{with } h(\beta) &= \frac{1}{N} \sum_{k=1}^{N} \log q^{(\beta)}_{y_k|x_k} - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}} \\
\textit{and } q^{(\beta)}_{y|x} &= \frac{1}{Z_x(\beta)} e^{\beta^T f(y,x)} q^0_{y|x} \\
\textit{and } Z_x(\beta) &= \sum_{y=0,1} q^0_{y|x} e^{\beta^T f(y,x)}
\end{aligned}$$

[6] The training data set T is a randomly generated subset of the data set.

[7] Note that this prior measure $q^0_{\cdot|x}$ is not equivalent to the empirical conditional default probability in the training data set $p_{\cdot|x}$. Otherwise $q^0_{\cdot|x}$ would be a trivial solution of Problems 1 and 2.

Problem 7 is a risk-adjusted maximum likelihood estimation of the parameter vector β.

The connection between the LLR modeling approach and the MEU approach will be discussed in the following Lemma.

Lemma 2.7 *A MEU model with* $U(x) = \log(x)$, $\alpha = 0$, $q^0_{y|x} = 0.5$ *and feature vector*

$$\widetilde{f}(y,x) := \frac{y-1}{y-0.5} f(y,x)$$

is equivalent to an extended LLR model with $g(x) = \widetilde{f}(0,x)$.

Proof:
Inserting α, $q^0_{y|x}$ and $\widetilde{f}(y,x)$ in Problem 7 and dropping all constant terms out leads to the likelihood equation of the corresponding LLR model. □

This means that the LLR approach is equivalent to the MEU approach with a logarithmic utility function and a special choice of α, $q^0_{y|x}$ and $f(y,x)$, where $\alpha = 0$ means that there is no risk-adjustment according to the consistency with the data and $q^0_{y|x} = \frac{1}{2}$ says we don't have any a priory information.

3 Empirical Comparison

For the empirical comparison of the different models that were introduced in Section 2 we used Fitch Risk's North American Loan Loss Database. The data was trimmed so that it only includes firms for which a full set of financials is obtainable. We compared the models on a one- and a five-year time horizon, i.e. we calculated probabilities of default in year one after the financial statement and in year five after the financial statement. For the one-year time horizon the dataset contained 301 defaulters and 17646 non-defaulters, for the five-year time horizon 109 defaulters and 10922 non-defaulters. This data set contains data from North American firms between 1984 and 2001 from 13 different regions [8] and 31 different industries [9]. In Figure 1 one can see the asset sizes

[8] New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, Pacific, U.S. National, Canada, International, Not Known

[9] aerospace; automobile; beverage, food and tobacco; buildings and real estate; cargo transport; chemicals, plastics and rubber; containers, packaging and glass; div. natural resources, precious metals and minerals; diversified/conglomerate manufacturing; diversified/conglomerate service; ecological; electronics; farming and agriculture; government; grocery; healthcare, education and childcare; home/office furnishing, houseware; hotels, motels, inns and gaming; leisure, amusement, motion

of the firms used to build the models.

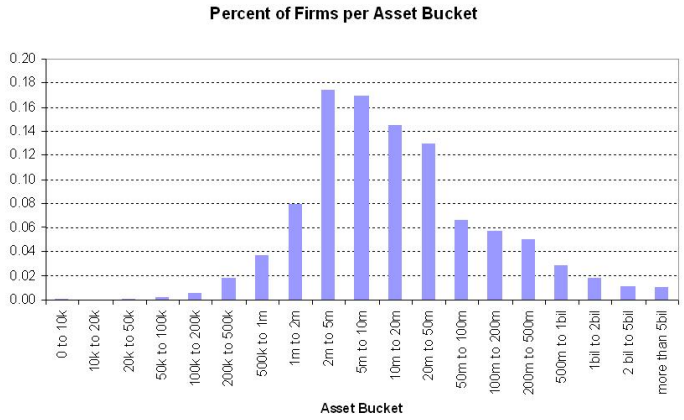

Fig. 1. Percent of Firms per Asset Bucket.

The next question that arises is how to choose the explanatory variables, the drivers on which the models are based. As the LLR and MEU models shall be compared, the same drivers are chosen for models with the same time horizon. The explanatory variables were chosen in a procedure that includes various methodologies starting with the examination of univariate relationships in the data. The financial ratios that showed the strongest dependence were chosen. Different combinations of those ratios were used to generate a series of logistic regression models. The best regression model was then selected based on three different criteria "credit sense" (Does the relationship between the financial drivers and the model output make sense?), performance (How well does the model distinguish between defaulters and non-defaulters?) and parsimony (a model with less input parameters is - similar performance assumed - in general preferred). This procedure led to the choice of financial ratios as explanatory variables shown in Table 1. It turned out that different factors are successful at distinguishing between defaulters and non-defaulters on different time horizons. For example liquidity with respect to long-term debt has more explorative power for longer periods of time than liquidity with respect to short-term debt, whereas today's unemployment rate is less important for the prediction of the default status in five years. The usage of unemployment data makes it possible to include information of the macroeconomic environment in the model building process.

pictures, entertainment; machinery; mining, steel, iron and non-precious metals; oil and gas; other; personal and non-durable consumer products; personal, food and miscellaneous services; printing, publishing and broadcasting; retail stores, telecommunications, textiles and leather; transportation; N/A

Table 1
Explanatory Variables.

Driver	1yr	5yrs
Log(Asset Values)	×	×
Cash and Equivalents / Long Term Debt		×
Cash and Equivalents / Short Term Debt	×	
EBITDA / Assets	×	×
Revenue / Total Assets	×	
Total Liabilities / Total Equity	×	
US Employment Rate Quarterly	×	
US Employment Rate Quarterly Change	×	×

3.1 Model Performance

The examination of model performance is an integral aspect in credit risk modeling. In our case, where we look at models for default probabilities, we will mostly concentrate on two different kinds of model performance measures to determine the goodness of a model: model power on the one hand and model calibration on the other hand. The power of a model describes how good a model is able to differentiate between defaulting and non-defaulting firms. If we look for example at two models that only differentiate between "good" and "bad" obligors, then the model that has a higher percentage of "defaults" in his "bad" category and a higher percentage of "non-defaults" in his "good" category is the more powerful model. Some examples for these power measures are contingency tables and power curves like cumulative accuracy profile (CAP) plots (see Figure 2) as well as the corresponding Accuracy Ratio ($AR = \frac{A_I}{A_{II}}$) (see, e.g., Keenan *et al.* (2000) or Stein (2002)). A quite similar concept is given by the receiver operating characteristic (ROC) curves and the Area under the ROC curve A, which can be directly calculated from the Accuracy Ratio via $A = \frac{1}{2}AR + \frac{1}{2}$ (see, e.g., Keenan *et al.* (2000) or Stein (2002)).
In contrast to this, the model calibration describes how good the predicted probabilities of default match the actual realizations, i.e. we measure the distance between the conditional default probabilities that our model predicts and the actual outcomes. These calibration measures can be summarized by the key word likelihood measures. In particular, we will look at the so-called *expected utility* model performance measure

$$\sum_x p_x \sum_y p_{y|x} U\left(b^*_{y|x}(q_{y|x}) O_{x,y}\right)$$

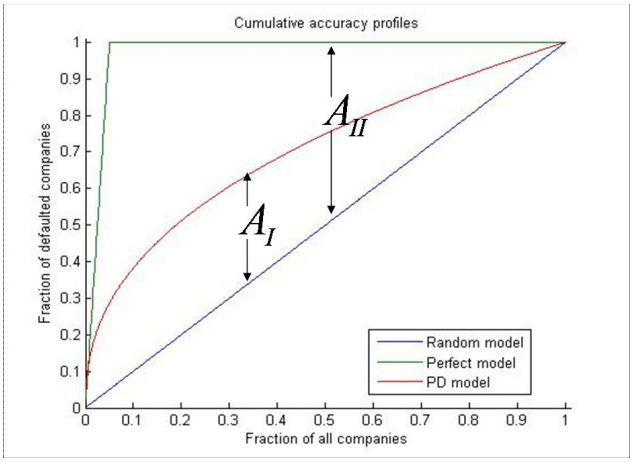

Fig. 2. CAP Plot.

and the *Kullback-Leibler relative entropy* (see, e.g., Cover and Thomas (1991))

$$\sum_x p_x \sum_y p_{y|x} \log\left(\frac{q_{y|x}}{p_{y|x}}\right).$$

The relative entropy can be interpreted as an estimate for the difference in the expected utility of an investor with a logarithmic utility function who works under the model $q_{\cdot|x}$ and uses the empirical conditional default probability $p_{\cdot|x}$ as benchmark model.

3.2 Comparison of models with logarithmic utility functions

First, we compared the models from Sections 1 and 2 that are based on a logarithmic utility function of the form $U(x) = \gamma_1 \ln(x + \gamma) + \gamma_2$, namely the linear logistic regression and the MEU models from Problem 6 and 7. As one can see from Problem 6, the only parameter of the three parameter logarithmic utility function that influences the MEU models is γ_1. It turned out that this parameter has no great influence on the model performance. Hence, we will use $U(x) = \ln(x)$ in the following. We started with an in-sample comparison of five different models, a simple linear logistic regression (LLR1), a LLR model with linear and quadratic dependencies (LLR2), a LLR model with linear, quadratic and cylindrical dependencies (LLR3), a MEU model with logarithmic utility function and linear and quadratic features (MEU2), and a MEU model with a logarithmic utility function and linear, quadratic and cylindrical features (MEU3).

Table 2
1yr Performance Measures In-Sample Comparison.

1yr	LLR1	MEU2	LLR2	MEU3	LLR3
AR	0.7538	0.7792	0.7817	0.7912	0.8082
A	0.8769	0.8896	0.8909	0.8956	0.9041
% of right predictions	79.82	81.50	81.53	81.90	82.42
Log-Likelihood	-504.65	-498.48	-487.28	-485.61	-466.51
Expected Utility	-0.0647	-0.0640	-0.0625	-0.0623	-0.0599
Relative Entropy	0.0204	0.0212	0.0227	0.0229	0.0253

Table 3
5yr Performance Measures In-Sample Comparison.

5yr	LLR1	MEU2	LLR2	MEU3	LLR3
AR	0.3836	0.4019	0.4346	0.4507	0.5074
A	0.6918	0.7010	0.7173	0.7253	0.7537
% of right predictions	64.62	66.06	66.27	67.92	69.22
Log-Likelihood	-253.19	-250.25	-246.55	-245.87	-240.85
Expected Utility	-0.0529	-0.0522	-0.0515	-0.0513	-0.0503
Relative Entropy	0.0026	0.0032	0.0040	0.0041	0.0052

In Tables 2 and 3 one can see that the extended LLR model with linear, quadratic and cylindrical dependencies clearly outperforms all the other models with respect to all performance measures. One can also see that the LLR models outperform the corresponding MEU models that model the same dependences in the data.
We also tested the MEU approach for a general HARA utility function. As mentioned above this is only possible with some simplifying assumptions. We have run in-sample comparisons for a MEU approach simplified with a first- and a second-order Taylor series expansion for different parameters in the HARA utility function. It turned out that those models were clearly outperformed by the models examined previously. E.g. on the five-year time horizon the accuracy ratio of the first-order simplified model for $U(x) = exp(-x)$ was 0.3904 and for $U(x) = 2\sqrt{x}$ we had 0.1922.
We also performed an out-of-sample comparison between the LLR and the MEU modeling approach. For this, the data set was randomly divided into two parts: the estimation data (70 % of the data) and the testing data (30 % of the data). The model parameters were estimated on the estimation data and the performance of those fitted models was tested on the testing data. This procedure was repeated 30 times and the averages of the performance

measures were calculated (see Tables 4 and 5).

Table 4
1yr Performance Measures Out-of-Sample Comparison.

1yr	LLR1		LLR2		MEU3	
	Mean	Stdev	Mean	Stdev	Mean	Stdev
AR	0.7556	0.0368	0.7558	0.0414	0.7597	0.0347
A	0.8778	0.0184	0.8779	0.0207	0.8799	0.0173
% of right pred.	80.89	1.99	81.37	2.29	81.12	1.68

Table 5
5yr Performance Measures Out-of-Sample Comparison.

5yr	LLR1		LLR2		MEU3	
	Mean	Stdev	Mean	Stdev	Mean	Stdev
AR	0.3812	0.0755	0.3568	0.0985	0.3520	0.1022
A	0.6906	0.0377	0.6784	0.0492	0.6760	0.0511
% of right pred.	65.86	2.64	66.76	3.95	65.65	3.77

As the dataset we used is quite small, this procedure didn't work for the logistic regression with linear, quadratic and cylindrical dependencies. Hence, we performed the out-of-sample comparison for the best possible models of simple and extended LLR and MEU. The results in this out-of-sample comparison are no longer that clear as for the in-sample comparison. One possible reason for this is the fact that the dataset we used is not big enough.

4 Summary and Conclusion

In this paper we described different approaches to modeling (real-world) probabilities of default and compared them empirically based on a database of North American firms. We started with a simple linear logistic regression as in Friedmann and Huang (2003) and Friedmann *et al.* (2003) that only allowed to model linear dependencies in the data. As this approach has, due to its simple form, some negative aspects, we introduced an extension of the linear logistic regression that allows for more complex dependence structures. This extended linear logistic regression model has the same positive aspects than the simple LLR but enables us to model different dependencies in the data and interactions among the explanatory variables. In addition to that, we examined a maximum expected utility modeling approach as in Friedmann and Sandow (2003a) which chooses a model on an efficient frontier defined in

terms of consistency with the training data and consistency with a prior distribution. The first is measured by a large-sample distribution of a vector of sample-averaged features, the latter by means of relative entropy. This leads to an optimization problem and its dual with an objective function which is asymptotically the expected utility of an investor who uses the model to bet on default events. We extended this MEU modeling approach by applying it not only to a logarithmic utility function but to a very general class of utility functions, the class of HARA utility functions.
In the empirical part it was shown that the more complex the dependence structure is modeled, the better the performance. The LLR models and the MEU models with a logarithmic utility function outperform the approximated MEU models. Also, the extended LLR models outperform the MEU models with respect to the in-sample tests, whereas no clear outperformance turned up in the out-of-sample comparison. Another disadvantage of the MEU approach that turned up in the empirical comparison is the larger computing time in comparison to the (extended) LLR models.

Acknowledgement

The authors would like to thank Algorithmics Inc. for the support and the provision of the Fitch Risk North American Loan Loss Database. Special thanks are to Colin Farquhar and Bernd Schmid for the valuable discussions and the insightful comments.

A Proof of Theorem 2.1

Since

$$b^*_{\cdot|x}(q) = \arg \max_{\left\{b_{\cdot|x}:\sum_y b_{y|x}=1\right\}} \sum_x p_x \sum_y q_{y|x} U\left(b_{y|x} O_{x,y}\right)$$

we have

$$\sum_x p_x \sum_y q_{y|x} U\left(b_{y|x} O_{x,y}\right) \leq \sum_x p_x \sum_y q_{y|x} U\left(b^*_{y|x}(q_{y|x}) O_{x,y}\right)$$

for a general $b_{\cdot|x}$. So, in particular

$$\sum_x p_x \sum_y q_{y|x} U\left(b^*_{y|x}(q^0_{y|x}) O_{x,y}\right) \leq \sum_x p_x \sum_y q_{y|x} U\left(b^*_{y|x}(q_{y|x}) O_{x,y}\right)$$

with equality for $q_{\cdot|x} = q^0_{\cdot|x}$. To see the last statement note that for two different measures $q_{\cdot|x}$ and $q^0_{\cdot|x}$ with $b^*_{y|x}(q_{y|x}) = b^*_{y|x}(q^0_{y|x})$ for all y we get from (3)

$$\frac{\lambda_x}{q_{y|x}} = \frac{\lambda^0_x}{q^0_{y|x}} \text{ for all } y.$$

Hence,

$$1 = \sum_y q^0_{y|x} = \frac{\lambda^0_x}{\lambda_x} \sum_y q_{y|x} = \frac{\lambda^0_x}{\lambda_x}$$

and thus $q_{\cdot|x} = q^0_{\cdot|x}$ for all y. Using the uniqueness of $b^*_{y|x}$ we can summarize

$$\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x}) \geq 0 \text{ with equality if and only if } q_{\cdot|x} = q^0_{\cdot|x}. \tag{A.1}$$

From (5) we get

$$\begin{aligned}
\frac{\partial \Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x})}{\partial q_{y|x}} =& p_x U(b^*_{y|x}(q_{y|x})O_{x,y}) - p_x U(b^*_{y|x}(q^0_{y|x})O_{x,y}) \\
&+ \sum_{x'} p_{x'} \sum_{y'} q_{y'|x'} \frac{\partial U(b^*_{y'|x'}(q_{y|x})O_{x',y'})}{\partial q_{y|x}} \\
=& p_x U(b^*_{y|x}(q_{y|x})O_{x,y}) - p_x U(b^*_{y|x}(q^0_{y|x})O_{x,y}) \\
&+ \sum_{x'} p_{x'} \sum_{y'} q_{y'|x'} O_{x',y'} \frac{\partial b^*_{y'|x'}(q_{y|x})}{\partial q_{y|x}} U'(b^*_{y'|x'}(q_{y|x})O_{x',y'}).
\end{aligned} \tag{A.2}$$

From (3) we know that

$$\lambda_x = p_x q_{y|x} O_{x,y} U'(b^*_{y|x}(q_{y|x})O_{x,y}).$$

With this we get

$$\begin{aligned}
\sum_{x'} p_{x'} \sum_{y'} q_{y'|x'} O_{x',y'} \frac{\partial b^*_{y'|x'}(q_{y|x})}{\partial q_{y|x}} U'(b^*_{y'|x'}(q_{y|x})O_{x',y'}) &= \sum_{x'} \lambda_{x'} \sum_{y'} \frac{\partial b^*_{y'|x'}(q_{y|x})}{\partial q_{y|x}} \\
&= \sum_{x'} \lambda_{x'} \frac{\partial}{\partial q_{y|x}} \sum_{y'} b^*_{y'|x'}(q_{y|x}) = \sum_{x'} \lambda_{x'} \frac{\partial}{\partial q_{y|x}} 1 = 0
\end{aligned}$$

Inserting this in (A.2) provides

$$\frac{\partial \Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x})}{\partial q_{y|x}} = p_x U(b^*_{y|x}(q_{y|x})O_{x,y}) - p_x U(b^*_{y|x}(q^0_{y|x})O_{x,y}) \tag{A.3}$$

To show that $\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x})$ is strictly convex, we consider the tangent hyperplane of $\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x})$ at the point $(\widetilde{p}_{\cdot|x}, \Delta_{prior}(\widetilde{p}_{\cdot|x}||q^0_{\cdot|x}))$, which is defined

by

$$\Delta_{prior}^{tangent}(q_{\cdot|x}||q_{\cdot|x}^0) = \Delta_{prior}(\widetilde{p}_{\cdot|x}||q_{\cdot|x}^0) + \sum_x \sum_y \left(q_{y|x} - \widetilde{p}_{y|x}\right) \frac{\partial \Delta_{prior}(q_{\cdot|x}||q_{\cdot|x}^0)}{\partial q_{y|x}}|_{q_{y|x}=\widetilde{p}_{y|x}}$$

With (A.3) we get

$$\begin{aligned}
&\Delta_{prior}^{tangent}(q_{\cdot|x}||q_{\cdot|x}^0) = \\
&= \Delta_{prior}(\widetilde{p}_{\cdot|x}||q_{\cdot|x}^0) + \sum_x p_x \sum_y (q_{y|x} - \widetilde{p}_{y|x}) \left(U(b_{y|x}^*(\widetilde{p}_{y|x})O_{x,y}) - U(b_{y|x}^*(q_{y|x}^0)O_{x,y})\right) \\
&= \sum_x p_x \sum_y q_{y|x} \left(U(b_{y|x}^*(\widetilde{p}_{y|x})O_{x,y}) - U(b_{y|x}^*(q_{y|x}^0)O_{x,y})\right)
\end{aligned}$$

Combining this with (5) we get

$$\begin{aligned}
\Delta_{prior}(q_{\cdot|x}||q_{\cdot|x}^0) - \Delta_{prior}^{tangent}(q_{\cdot|x}||q_{\cdot|x}^0) &= \sum_x p_x \sum_y q_{y|x} U(b_{y|x}^*(q_{y|x})O_{x,y}) \\
&\quad - \sum_x p_x \sum_y q_{y|x} U(b_{y|x}^*(\widetilde{p}_{y|x})O_{x,y}) \\
&= \Delta_{prior}(q_{\cdot|x}||\widetilde{p}_{\cdot|x}) \geq 0.
\end{aligned}$$

Using (A.1), we conclude that $\Delta_{prior}(q_{\cdot|x}||q_{\cdot|x}^0)$ as a function of $q_{\cdot|x}$ always lies above its tangent hyperplanes, i.e. $\Delta_{prior}(q_{\cdot|x}||q_{\cdot|x}^0)$ is a strictly convex function of $q_{\cdot|x}$. □

B Proof of Lemma 2.3

We know from Theorem 2.1 that the objective function $\Delta_{prior}(q_{\cdot|x}||q_{\cdot|x}^0)$ is strictly convex. Since Σ is a covariance matrix and hence positive semi-definite, the inequality constraint $\Delta_{data}(q_{\cdot|x}||p_{\cdot|x}) \leq \alpha$ is convex. The two equality constraints are affine linear and thus Problem 2 is a strictly convex optimization problem. In the case $\alpha = \alpha_0$, it is obvious that both problems have the same solution $q_{\cdot|x}^* = q_{\cdot|x}^0$. For $\alpha < \alpha_0$ we show that the solution of Problem 2 fulfills $\Delta_{data}(q_{\cdot|x}||p_{\cdot|x}) = \alpha$: From Theorem 2.1 we know that $\Delta_{prior}(q_{\cdot|x}||q_{\cdot|x}^0)$ is a strictly convex function of $q_{\cdot|x}$ and $\Delta_{prior}(q_{\cdot|x}||q_{\cdot|x}^0)$ is minimal if and only if $q_{\cdot|x} = q_{\cdot|x}^0$ which occurs if and only if $\alpha = \alpha_0$. Hence,

$$\nabla_{q_{\cdot|x}} \Delta_{prior}(q_{\cdot|x}||q_{\cdot|x}^0) \neq 0 \tag{B.1}$$

for $q_{\cdot|x} \neq q_{\cdot|x}^0$. Suppose that for the solution of Problem 2, $q_{\cdot|x}^*$, $\Delta_{data}(q_{\cdot|x}^*||p_{\cdot|x}) < \alpha$ holds. Then there exists a neighborhood of $q_{\cdot|x}^*$ in

$Q_x := \left\{q_{\cdot|x} = (q_{0|x}, q_{1|x}) : q_{y|x} \geq 0, \sum_y q_{y|x} = 1\right\}$ such that for all $q_{\cdot|x}$ in this neighborhood $\Delta_{data}(q_{\cdot|x}||p_{\cdot|x}) \leq \alpha$. As we know from (B.1) there is a direction of decrease of $\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x})$ on Q_x and so the solution $q^*_{\cdot|x}$ can't be optimal. Hence the assumption $\Delta_{data}(q^*_{\cdot|x}||p_{\cdot|x}) < \alpha$ can not be true and therefore $\Delta_{data}(q^*_{\cdot|x}||p_{\cdot|x}) = \alpha$. So the solution of Problem 2 is the solution of Problem 1. As the objective function $\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x})$ is strictly convex, the optimal solution of Problem 2 is unique and therefore the solution of Problem 1 is also unique. □

C Proof of Theorem 2.4

We will prove this theorem in 3 steps. As we want to derive the dual of Problem 2, we look at the Lagrangian of this problem (see for example Boyd and Vandenberghe (2004)) which is given by

$$\begin{aligned} L(q_{\cdot|x}, c, \beta, \xi, \mu, \nu) = &\Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x}) + \beta^T(c - E_{q_{\cdot|x}}(f) + E_{p_{\cdot|x}}(f)) \\ &+\frac{1}{2}\xi\left(Nc^T\Sigma^{-1}c - \alpha\right) + \sum_x p_x\mu_x\left(\sum_y q_{y|x} - 1\right) \\ &- \sum_x p_x \sum_y q_{y|x}\nu_{y|x}, \end{aligned} \tag{C.1}$$

where $\beta = (\beta_1, \ldots, \beta_J)^T$, $\xi \geq 0$, $\mu = (\mu_{x_1}, \ldots, \mu_{x_M})^T$ and $\nu = \{\nu_{y|x} \geq 0 : y = y_1, \ldots, y_m, x = x_1, \ldots, x_M\}$ [10].

(1) In the first step we will derive equation (12) which is often referred to as the connecting equation (see, e.g., Lebanon and Lafferty (2001)). In order to achieve this we have to solve the following equations:

$$\frac{\partial L(q_{\cdot|x}, c, \beta, \xi, \mu, \nu)}{\partial c_j} = 0$$

and

$$\frac{\partial L(q_{\cdot|x}, c, \beta, \xi, \mu, \nu)}{\partial q_{y|x}} = 0 \tag{C.2}$$

The first equation has the solution

$$c^* = -\frac{1}{\xi N}\Sigma\beta. \tag{C.3}$$

[10] Here, μ_x and $\nu_{y|x}$ are scaled with p_x to make the following calculations easier. The original Lagrange multipliers are $\widetilde{\mu}_x := p_x\mu_x$ and $\widetilde{\nu}_{y|x} := p_x\nu_{y|x}$.

From the proof of Theorem 2.1, (A.3), we know

$$\frac{\partial \Delta_{prior}(q_{\cdot|x}||q^0_{\cdot|x})}{\partial q_{y|x}} = p_x U(b^*_{y|x}(q_{y|x})O_{x,y}) - p_x U(b^*_{y|x}(q^0_{y|x})O_{x,y}). \quad \text{(C.4)}$$

If we insert (C.1) and (C.4) in (C.2), we get

$$0 = U(b^*_{y|x}(q_{y|x})O_{x,y}) - U(b^*_{y|x}(q^0_{y|x})O_{x,y}) - \beta^T f(y,x) + \mu_x - \nu_{y|x}. \quad \text{(C.5)}$$

We can rewrite this equation as

$$U(b^*_{y|x}(q_{y|x})O_{x,y}) = G(x, y, q^0_{\cdot|x}, \beta, \mu_x, \nu) \quad \text{(C.6)}$$

with

$$G(x, y, q^0_{\cdot|x}, \beta, \mu_x, \nu) = U(b^*_{y|x}(q^0_{y|x})O_{x,y}) + \beta^T f(y,x) - \mu_x + \nu_{y|x}. \quad \text{(C.7)}$$

Inserting (3) in (C.6) leads to

$$U\left((U')^{-1}\left(\frac{\lambda_x}{p_x q_{y|x} O_{x,y}}\right)\right) = G(x, y, q^0_{\cdot|x}, \beta, \mu_x, \nu). \quad \text{(C.8)}$$

Solving this equation for $q_{y|x}$, we get

$$q^*_{y|x} = \frac{\lambda_x}{p_x O_{x,y} U'(U^{-1}(G(x, y, q^0_{\cdot|x}, \beta, \mu_x, \nu)))}. \quad \text{(C.9)}$$

From this, the positivity of p_x and $O_{x,y}$ and the fact that U is a strictly monotone increasing function it follows that λ_x and all $q^*_{y|x}$ have the same sign. From $\sum_y q_{y|x} = 1$ we get that λ_x and $q^*_{y|x}$ are strictly positive. The Karush-Kuhn-Tucker conditions (see, e.g., Boyd and Vandenberghe (2004)) provide $\nu_{y|x} q^*_{y|x} = 0$ and therefore $\nu^*_{y|x} = 0$ for all (x, y). Therefore the dependence of L and G on ν can be left out. So (C.9) depends on the dual parameters λ_x, μ_x and β only.
Now, we want to express the dependence of λ_x and μ_x on β. By solving (C.8) for $(U')^{-1}\left(\frac{\lambda_x}{p_x q_{y|x} O_{x,y}}\right)$ and inserting this in (4), we get a condition for μ^*_x:

$$\sum_y \frac{1}{O_{x,y}} U^{-1}(G(x, y, q^0_{\cdot|x}, \beta, \mu^*_x)) = 1. \quad \text{(C.10)}$$

U^{-1} is a strictly monotone increasing function and G is monotone decreasing in μ_x (see (C.7)). Hence, the left-hand side of (C.10) is strictly monotone decreasing in μ^*_x. Let $\overline{\mu}_x := \max_y \beta^T f(y,x)$. Then $\beta^T f(y,x) - \overline{\mu}_x \leq 0$ for all y. From equation (C.7) we get $G(x, y, q^0_{\cdot|x}, \beta, \overline{\mu}_x) \leq U(b^*_{y|x}(q^0_{y|x})O_{x,y})$ for all y and, because of the monotony of U^{-1},

$$\sum_y \frac{1}{O_{x,y}} U^{-1}(G(x,y,q^0_{\cdot|x},\beta,\overline{\mu}_x)) \leq \sum_y \frac{1}{O_{x,y}} U^{-1}(U(b^*_{y|x}(q^0_{y|x})O_{x,y})) = \sum_y b^*_{y|x}(q^0_{y|x}) = 1$$

From (C.6) we know that $G(x,y,q^0_{\cdot|x},\beta,\overline{\mu}_x) \in dom(U^{-1})$. Using $\underline{\mu}_x := \min_y \beta^T f(y,x)$, we derive

$$\sum_y \frac{1}{O_{x,y}} U^{-1}(G(x,y,q^0_{\cdot|x},\beta,\underline{\mu}_x)) \geq 1$$

and thus, using the intermediate value theorem and the strict monotony and continuity of the left hand-side of (C.10), we conclude that there exists a unique solution μ^*_x of (C.10) which can be easily found by numerical methods.

To express λ_x in dependence of β and μ^*_x, we insert (C.9) in $1 = \sum_y q_{y|x}$ and get

$$1 = \lambda_x \sum_y \frac{1}{p_x O_{x,y} U'(U^{-1}(G(x,y,q^0_{\cdot|x},\beta,\mu^*_x)))}.$$

Solving this for λ_x we get

$$\lambda^*_x = \left(\sum_y \frac{1}{p_x O_{x,y} U'(U^{-1}(G(x,y,q^0_{\cdot|x},\beta,\mu^*_x)))} \right)^{-1}. \tag{C.11}$$

To sum it up, the connecting equation, which describes $q^*_{y|x}$ in dependence of β, can be written as

$$q^*_{y|x} = \frac{\lambda^*_x}{p_x O_{x,y} U'(U^{-1}(G(x,y,q^0_{\cdot|x},\beta,\mu^*_x)))} \tag{C.12}$$

with μ^*_x from (C.10) and λ^*_x from (C.11).

(2) In the next step we will show that Problem 3 with utility function $h(\beta) = \beta^T E_{p_{\cdot|x}}(f) - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}} - \sum_x p_x \mu^*_x$ is the dual of Problem 2:
From (C.3), (C.9), (C.10) and (C.11) we get c^*, the probabilities $q^*_{y|x}$ and the Lagrange multipliers μ^*_x and λ^*_x for which the Lagrangian is minimal for given β and ξ. Therefore, we can express the dual problem as an optimization problem depending on β and ξ. For this we have to calculate $L(q^*_{\cdot|x}, c^*, \beta, \xi, \mu^*_x)$. If we insert equations (5) and (C.3) in equation (C.1) we get

$$
\begin{aligned}
L(q^*_{\cdot|x}, c^*, \beta, \xi, \mu^*_x) = &\sum_x p_x \sum_y q^*_{y|x} U\left(b^*_{y|x}(q^*_{y|x}) O_{x,y}\right) \\
&- \sum_x p_x \sum_y q^*_{y|x} U\left(b^*_{y|x}(q^0_{y|x}) O_{x,y}\right) \\
&+ \beta^T \left(-\frac{1}{\xi N} \Sigma \beta - \sum_x p_x \sum_y q^*_{y|x} f(y,x) + E_{p_{\cdot|x}}(f) \right) \\
&+ \frac{1}{2} \xi \left(N \frac{1}{\xi^2 N^2} \beta^T \Sigma \beta - \alpha \right) \\
&+ \sum_x p_x \mu^*_x \left(\sum_y q_{y|x} - 1 \right)
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
L(q^*_{\cdot|x}, c^*, \beta, \xi, \mu^*_x) = &\sum_x p_x \sum_y q^*_{y|x} [U\left(b^*_{y|x}(q^*_{y|x}) O_{x,y}\right) \\
&- U\left(b^*_{y|x}(q^0_{y|x}) O_{x,y}\right) - \beta^T f(y,x) + \mu^*_x] \\
&+ \beta^T E_{p_{\cdot|x}}(f) - \frac{1}{2\xi N} \beta^T \Sigma \beta - \frac{1}{2} \xi \alpha - \sum_x p_x \mu^*_x.
\end{aligned}
$$

Because of (C.5) the first two lines sum up to 0 and hence

$$
L(q^*_{\cdot|x}, c^*, \beta, \xi, \mu^*_x) = \beta^T E_{p_{\cdot|x}}(f) - \frac{1}{2\xi N} \beta^T \Sigma \beta - \frac{1}{2} \xi \alpha - \sum_x p_x \mu^*_x.
$$

$L(q^*_{\cdot|x}, c^*, \beta, \xi, \mu^*_x)$ can be maximized analytically with respect to ξ by solving

$$
\frac{\partial L(q^*_{\cdot|x}, c^*, \beta, \xi, \mu^*_x)}{\partial \xi} = 0.
$$

The solution of this equation is given by

$$
\xi^* = \sqrt{\frac{\beta^T \Sigma \beta}{N \alpha}} \geq 0.
$$

The Lagrangian evaluated at $\xi = \xi^*$ is then given by

$$
L(q^*_{\cdot|x}, c^*, \beta, \xi^*, \mu^*_x) = \beta^T E_{p_{\cdot|x}}(f) - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}} - \sum_x p_x \mu^*_x. \tag{C.13}
$$

For the dual problem we have to maximize $h(\beta) := L(q^*_{\cdot|x}, c^*, \beta, \xi^*, \mu^*_x)$ with respect to β. Combining (C.5) with the fact that $U(b^*_{y|x}(q^0_{y|x}) O_{x,y})$ is independent of β, we can rewrite $h(\beta)$ up to a constant factor as

$$
h(\beta) = \sum_x p_x \sum_y p_{y|x} U(b^*_{y|x}(q^*_{y|x}) O_{x,y}) - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}}.
$$

From (3), (C.7), (C.10), (C.11), (C.12) and (C.13) we get Problem 3.

(3) Theorem 2.4 is now a consequence of the first three steps and the fact that the solution of Problem 2 is unique (see Lemma 2.3).

□

D Proof of Lemma 2.6

With the help of the fact that $\lim_{k\to\infty}\left(1+\frac{w_k}{k}\right)^k = e^w$ for every sequence (w_k) with $w_k \to w$ (see, e.g., Section 8.1 of Königsberger (2004)) we get

$$\begin{aligned}
&\lim_{b\to 0} \frac{1}{\gamma_1^{\frac{1}{b}} p_x O_{x,y}} \left[\left(\frac{\lambda_x}{\gamma_1^{\frac{1}{b}} p_x q^0_{y|x} O_{x,y}} \right)^{\frac{b}{b-1}} + \frac{b}{1-b} \left(\beta^T f(y,x) - \mu_x \right) \right]^{\frac{1-b}{b}} \\
&= q^0_{y|x} \lim_{b\to 0} \left[\lambda_x^{\frac{b}{b-1}} + \gamma_1^{\frac{1}{b-1}} \left(p_x q^0_{y|x} O_{x,y} \right)^{\frac{b}{b-1}} \frac{b}{1-b} \left(\beta^T f(y,x) - \mu_x \right) \right]^{\frac{1-b}{b}} \\
&= q^0_{y|x} \lim_{k\to\infty} \left[\lambda_x^{-\frac{1}{k}} + \gamma_1^{-1} \left(\gamma_1 p_x q^0_{y|x} O_{x,y} \right)^{-\frac{1}{k}} \frac{1}{k} \left(\beta^T f(y,x) - \mu_x \right) \right]^k \\
&= q^0_{y|x} \frac{1}{\lambda_x} \lim_{k\to\infty} \left[1 + \frac{1}{k} \underbrace{\gamma_1^{-1} \left(\frac{\lambda_x}{\gamma_1 p_x q^0_{y|x} O_{x,y}} \right)^{\frac{1}{k}} \left(\beta^T f(y,x) - \mu_x \right)}_{w_k} \right]^k \\
&= q^0_{y|x} \frac{1}{\lambda_x} e^{\overbrace{\frac{1}{\gamma_1} \left(\beta^T f(y,x) - \mu_x \right)}^{w}}
\end{aligned} \tag{D.1}$$

and hence we can rewrite (16) and obtain

$$1 = \sum_y q^0_{y|x} e^{\frac{1}{\gamma_1}\left(\beta^T f(y,x) - \mu_x\right)}. \tag{D.2}$$

With (D.1) we can rewrite (14) and get

$$\begin{aligned}
q^*_{y|x} &= \frac{\lambda_x}{\gamma_1^{1/b} p_x O_{x,y} \left[\left(\frac{\lambda_x}{\gamma_1^{1/b} p_x q^0_{y|x} O_{x,y}} \right)^{\frac{b}{b-1}} + \frac{b}{1-b} \left(\beta^T f(y,x) - \mu_x \right) \right]^{\frac{b-1}{b}}} \\
&= q^0_{y|x} e^{\frac{1}{\gamma_1}\left(\beta^T f(y,x) - \mu_x\right)}.
\end{aligned} \tag{D.3}$$

Hence, equation (D.2) is equivalent to $\sum_y q^*_{y|x} = 1$ and

$$q^*_{y|x} = \frac{q^0_{y|x} e^{\frac{1}{\gamma_1}\left(\beta^T f(y,x) - \mu_x\right)}}{\sum_y q^0_{y|x} e^{\frac{1}{\gamma_1}(\beta^T f(y,x) - \mu_x)}} = \frac{q^0_{y|x} e^{\frac{1}{\gamma_1}\beta^T f(y,x)}}{\sum_y q^0_{y|x} e^{\frac{1}{\gamma_1}\beta^T f(y,x)}}. \tag{D.4}$$

With $\lim_{k\to\infty} k\left(x^{\frac{1}{k}} - 1\right) = \log(x)$ (see, e.g., Section 8.3 of Königsberger (2004)) we can rewrite the objective function of Problem 4 and obtain

$$\begin{aligned}
&\lim_{b\to 0} \frac{1-b}{b}\left[\left(\frac{\lambda_x}{\gamma_1^{\frac{1}{b}} p_x q^*_{y|x} O_{x,y}}\right)^{\frac{b}{b-1}} - \gamma_1 + \gamma_2 b\right] \\
&= \gamma_2 + \gamma_1 \lim_{b\to 0} \frac{1-b}{b}\left[\left(\frac{\gamma_1 p_x q^*_{y|x} O_{x,y}}{\lambda_x}\right)^{\frac{b}{1-b}} - 1\right] \\
&= \gamma_2 + \gamma_1 \lim_{k\to\infty} k\left[\left(\frac{\gamma_1 p_x q^*_{y|x} O_{x,y}}{\lambda_x}\right)^{\frac{1}{k}} - 1\right] \\
&= \gamma_2 + \gamma_1 \log\left(\frac{\gamma_1 p_x q^*_{y|x} O_{x,y}}{\lambda_x}\right) \\
&= \gamma_2 + \gamma_1 \log\left(q^*_{y|x}\right) + \gamma_1 \log\left(\frac{\gamma_1 p_x O_{x,y}}{\lambda_x}\right) \\
&\propto \gamma_1 \log\left(q^*_{y|x}\right).
\end{aligned} \tag{D.5}$$

Combining (D.4) with (D.5) leads to Problem 6. □

References

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. Wiley, New York.

Davidson, R. and MacKinnon, J. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, New York.

Fahrmeir, L., Hamerle, A., and Tutz, G., editors (1996). *Multivariate Statistische Verfahren*. De Gruyter, Berlin.

Friedmann, C. and Huang, J. (2003). Default probability modeling: A maximum expected utility approach. *Working Paper*.

Friedmann, C. and Sandow, S. (2003a). Learning probabilistic models: An expected utility maximization approach. *Journal of Machine Learning Research*, **4**, 257–291.

Friedmann, C. and Sandow, S. (2003b). Model performance measures for expected utility maximizing investors. *International Journal of Theoretical and Applied Finance*, **6**(4), 355–401.

Friedmann, C., Huang, J., De Servigny, A., and Salinas, E. (2003). A utility-based private firm default probability model. *Working Paper*.

Ingersoll Jr., J. (1987). *Theory of Financial Decision Making*. Rowmann and Littlefield, New York.

Keenan, J. R., Sobehart, S. C., and Stein, R. M. (2000). Benchmarking quantitative default risk models: A vaildation methodology. *Moody's Rating Methodology*.

Königsberger, K. (2004). *Analysis I*. Springer, Berlin.

Lebanon, G. and Lafferty, J. (2001). Boosting and maximum likelihood for exponential models. *Technical Report CMU-CS-01-144 School of Computer Science Carnegie Mellon University*.

Stein, R. M. (2002). Benchmarking default prediction models: Pitfalls and remedies in model validation. *Moody's KMV, New York, Technical Report #020305*.