Journal of Machine Learning Research 4 (2003) 257-291 Submitted 4/03; Published 7/03

Learning Probabilistic Models: An Expected Utility Maximization Approach

Craig Friedman CRAIG_FRIEDMAN@SANDP.COM
Sven Sandow SVEN_SANDOW@SANDP.COM
Standard & Poor's
Risk Solutions Group
55 Water Street
New York, NY 10041

Editors: David Maxwell Chickering

Abstract

We consider the problem of learning a probabilistic model from the viewpoint of an expected utility maximizing decision maker/investor who would use the model to make decisions (bets), which result in well defined payoffs. In our new approach, we seek good out-of-sample model performance by considering a one-parameter family of Pareto optimal models, which we define in terms of consistency with the training data and consistency with a prior (benchmark) model. We measure the former by means of the large-sample distribution of a vector of sample-averaged features, and the latter by means of a generalized relative entropy. We express each Pareto optimal model as the solution of a strictly convex optimization problem and its strictly concave (and tractable) dual. Each dual problem is a regularized maximization of expected utility over a well-defined family of functions. Each Pareto optimal model is robust: maximizing worst-case outperformance relative to the benchmark model. Finally, we select the Pareto optimal model with maximum (out-of-sample) expected utility. We show that our method reduces to the minimum relative entropy method if and only if the utility function is a member of a three-parameter logarithmic family.

Keywords: Learning Probabilistic Models, Expected Utility, Relative Entropy, Pareto Optimality, Robustness

1. Introduction

From the viewpoint of a rational decision maker in an uncertain world, the efficacy of a probabilistic model is directly related to the quality of the decisions that he makes, based on the model. If, for example, the decision maker is an investor in a financial market, a probabilistic market model can be used by the investor to design an optimal investment strategy; the efficacy of the model should be judged by the success of this (optimal with respect to the model) strategy. Should the decision maker build a model, he ought to take this into account. In principle, standard approaches to probabilistic modeling (see, for example, Vapnik, 1999, Berger, 1985, or Hastie, Tibshirani, and Friedman, 2001) allow for the model builder to incorporate the decision consequences of a model, usually in terms of a risk functional or a loss function. However, these approaches often give no guidance as to how to construct, from first principles, a risk functional or a loss function. In this paper, we propose a new approach to model building that takes *explicitly* into account the decision consequences of the model; we measure these decision consequences by the success of the strategy that a rational investor (who believes the model) would choose to place bets in a horse race (see, for example, Cover and Thomas, 1991 for a discussion of the horse race). In particular, we monetize the decision consequences by assuming that there is a market with payoffs associated with each state (the horse race). The assumption of a rational investor who bets on horses allows us to relate the model user's decisions and their consequences to the model itself; as we shall see, this assumption leads to tractable models. Our approach combines ideas from maximum entropy modeling and utility theory.

Maximum entropy inference, introduced by Jaynes (1957, 1982, 1984) in the context of statistical physics, has been successfully applied to image processing (see, for example, Wu, 1997, or Gull and Daniell, 1978) as well as to a wide range of problems in biology (see, for example, Burnham and Anderson, 2002), finance (see, for example, Avellaneda, 1998, Avellaneda et al., 1997, Samperi, 1997, Gulko, 2002 and Frittelli, 2000) and economics (see, for example, Golan et al., 1996). The basic idea of the maximum entropy approach is that one chooses a model that maximizes uncertainty, or, more generally, minimizes the information-theoretic (Kullback-Leibler) distance to a prior, while ensuring that important features of the data are reproduced. In many applications, in order to avoid overfitting, one has to allow for some error in the calibration of the model-expected feature values to the expected feature values under the empirical measure (see, for example, Daniell, 1991, Skilling, 1991, Wu, 1997, Chen and Rosenfeld, 1999, or Lebanon and Lafferty, 2001).

In order to evaluate a model in terms of the decision consequences of a rational decision maker who believes the model, we need to first relate the rational decisions to the model and then evaluate the consequences of these decisions. For both of these purposes, we employ a utility function, a well established concept in economics (see, for example, Neumann and Morgenstern, 1944, or Luenberger, 1998). One can show that, under some additional plausible assumptions, a decision maker has a well defined utility function if he has preferences between the possible states of the world and probability weighted combinations of these states. It follows from the axioms of utility theory that a rational decision maker acts to maximize his expected utility based on the model he believes; his decisions are uniquely (and explicitly) determined by his model. Utility theory also dictates that the consequences of these decisions should be evaluated by means of the expected utility they lead to.

Friedman and Sandow (2003a) consider the performance of probabilistic models from the point of view of an expected utility maximizing investor who bets on horses. In order to evaluate a particular model, we assume that there is an investor who believes the model. This investor places bets in a horse race so as to maximize his expected utility according to his beliefs, i.e., the investor bets so as to maximize the expectation of his utility under the model probability measure. We then measure the success of the investor's investment strategy in terms of the average utility the strategy provides on an out-of-sample data set. An investor who has a highly accurate model will be able to choose a sound investment strategy, while an investor with a less accurate model will sometimes overbet or underbet, and consequently, be less successful in the long run. Therefore, the success of the investor's strategy, as measured by the utility averaged over a test sample, is a measure of the quality of the model on which the investor bases his strategy. This measure was used to evaluate probabilistic models.

In that work, we assume that there exist a number of candidate models and that an investor seeks the best (in the expected utility sense) of these models. A related, but harder, question is: how can we learn a model (from data) so that an investor, who makes decisions according to the model, maximizes his expected utility? We address this question in this paper. Our modeling approach, which is different from existing approaches, is based on the idea that one can achieve good out-of-sample model performance (as measured by expected utility under the model-optimal strategy) by considering models on an efficient frontier (Pareto optimal models), which we define in terms of consistency with the training data and consistency with a prior (benchmark) model. We measure the former by means of the large-sample distribution of a vector of sample-averaged features, and the latter by means of the generalized relative entropy introduced by Friedman and Sandow (2003a). This generalized relative entropy is essentially the same as the one independently introduced by Grünwald and Dawid (2002). The models on the efficient frontier, each of which can be obtained by solving a convex optimization problem (see Problems 2 and 8) form a single-parameter family. We show that each Pareto optimal model is robust in the sense that, for its level of consistency with the data, the model maximizes the worst-case outperformance relative to the benchmark model (see Section 2.2.5 and Appendix A). For each level of consistency with the data, we derive the dual problem (see Problems 3, 4 and 9), which has a Pareto optimal measure as its solution; this dual problem, which is new for non-logarithmic utility functions, amounts to the maximization of expected utility with a regularization penalty over a well-defined family of functions. We rank the models on the efficient frontier by computing their expected utilities on a hold-out sample, and select the model with maximum estimated expected utility. For ease of exposition, we consider only one hold-out sample; our procedure can be modified for $k-$fold cross validation.

Our economic paradigm, in general, requires the specification of the payoff structure of the horse race. This requirement imposes an additional encumbrance on the model-builder. However, we show that the optimization problems that follow from our paradigm are independent of the payoffs if and only if the investor's utility function is in a three-parameter logarithmic family (see Theorem 3). This logarithmic family is rich enough to describe a wide range of risk aversions, and it can be used to well-approximate (under reasonable conditions) non-logarithmic utility functions (see Friedman and Sandow, 2003a); it is therefore applicable to many practical problems. In the case of a utility function from this logarithmic family, our method leads to a regularized relative entropy minimization similar to the ones by Daniell (1991), Skilling (1991), Wu (1997), or Lebanon and Lafferty (2001) (see Corollary 1). This means that our approach provides new motivation of this regularized relative entropy minimization. It is well known that the relative entropy can be related to the expected utility of an investor with a logarithmic utility (of the form $U(z) = \log(z)$) who bets optimally in a horse race (see, for example, Cover and Thomas, 1991). Our result, however, is more general since we allow the investor's utility to be any member of the three-parameter family of logarithmic functions given by $U(z) = \gamma_1 \log(z+\gamma) + \gamma_2$.

Some of our discussion, such as the definition and robustness of the Pareto optimal measures (See Appendix A) and the formulation of the primal problems (see Sections 2.2 and 3.2), can be developed in a more general setting than the horse race. However, our dual problem formulation (see Sections 2.3 and 3.3), depends on the horse race setting. To keep things as simple as possible, we have confined our discussion to the horse race setting.

In Section 2, we formulate our modeling approach in the simplest context: we seek a discrete probability model. In Section 3, we briefly discuss our approach in a more general context: we seek a model that describes the conditional distribution of a possibly vector-valued random variable with a continuous range and discrete point masses. Numerical experiments based on the methodology of this paper are reported by Friedman and Sandow (2003b), Friedman and Huang (2003) and Sandow et al. (2003).

2. Discrete Probability Models

In this section, we describe our modeling paradigm in the simplest context: discrete probabilities.

2.1 Preliminaries

This section sets the stage for our modeling approach for the case of discrete probabilities. The ideas outlined below are explained in more detail in Section 2 in Friedman and Sandow (2003a).

We seek a probabilistic model for the discrete random variable Y, which can take any of the m values $y_1, \ldots, y_m$. For later use we define the following three probability measures:

Definition 1

$$\begin{aligned} p_y &= \textit{true (unknown) probability that } Y = y \, , \\ \tilde{p}_y &= \textit{empirical probability that } Y = y \, , \\ q_y &= \textit{model probability that } Y = y \, . \end{aligned}$$

The true probabilities, $p = (p_1, \ldots, p_m)^T$, are unknown, but we assume their existence; the empirical probabilities, $\tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_m)^T$, are known from the data, and the model probabilities, $q = (q_1, \ldots, q_m)^T$, are the ones we are trying to find. We define the probability simplex

$$Q = \{q : q \geq 0, \sum_y q_y = 1\}.$$

We identify the probabilistic problem with the horse race (see , for example, Cover and Thomas, 1991, Chapter 6; Cover and Thomas discuss the horse race from the point of view of an investor with logarithmic utility).

Definition 2 *A horse race is a market characterized by the discrete random variable Y with m possible states, in which an investor can place a bet that $Y = y$, which pays $O_y > 0$ for each dollar wagered if $Y = y$, and 0, otherwise.*[1]

We consider a decision maker/investor[2] who places bets on horses; we assume that our investor allocates b_y to the event $Y = y$, where

$$\sum_y b_y = 1. \tag{1}$$

We note that Equation 1 corresponds to the assumption that the investor allocates all his wealth to bets on the horses, without cash borrowing or lending. This setting does not represent the most general financial market.

In order to quantify the benefits of a model q to an expected utility-maximizing investor, we consider the investor's utility function, U, and assume that it is

(i) strictly concave,

1. The investor does not get his dollar back in addition to the payoff, O_y. This horse race definition is general enough to allow for situations where the investor loses with certainty.
2. Throughout the rest of this paper we will use the term investor for any type of decision maker, while keeping in mind that the decision maker does not necessarily act in a financial market.

(ii) twice differentiable, and

(iii) strictly monotone increasing.

We note that many popular utility functions (for example, the logarithmic, exponential and power utilities (see Luenberger, 1998)) are consistent with these conditions. It is possible to develop the ideas in this paper under more relaxed assumptions.

Our investor allocates his assets so as to maximize his expected utility according to his beliefs, i.e., the investor allocates so as to maximize the expectation of his utility under the model probability measure. This means that our investor allocates according to

$$b^*(q) = \arg \max_{\{b:\sum_y b_y=1\}} \sum_y q_y U(b_y O_y). \tag{2}$$

It has been shown (see Theorem 1 from Friedman and Sandow, 2003a) that

$$b_y^*(q) = \frac{1}{O_y}(U')^{-1}\left(\frac{\lambda}{q_y O_y}\right), \tag{3}$$

where λ is the solution of the following equation:

$$\sum_y \frac{1}{O_y}(U')^{-1}\left(\frac{\lambda}{q_y O_y}\right) = 1 \ , \tag{4}$$

if the solution of Equation 4 exists, which we assume to be the case here:

Assumption 1 *There exists a solution to Equation 4.*

We note that there does not always exist a solution to Equation 4, however, there exists a solution for many common utilities, for example the logarithmic, power, exponential and quadratic utilities (see Corollary 1, Appendix B of Friedman and Sandow, 2003a).

Equipped with above tools, we can formulate our modeling objective:

Objective:

$$\textit{Find} \qquad \arg\max_{q\in Q} E_p[U(b^*(q), O)] \ , \tag{5}$$

where, slightly abusing notation,

$$E_p[U(b^*(q), O)] = \sum_y p_y U(b_y^*(q) O_y) \ .$$

Thus, it is our objective to find the model that maximizes the *true* expectation of the utility of an investor who bets according to the model.

Since we don't know the true model, p, we cannot compute the $p-$expectation in Equation 5 exactly. Therefore, we approximate it by a sample average; in order to construct models that don't overfit, we approximate the $p-$expectation in Equation 5 by an average over a test sample:

$$E_p[U(b^*(q), O)] \approx E_{\tilde{p}}[U(b^*(q), O)] \ ,$$

where $\tilde{p}$ is the empirical measure of the test sample, which is different from the sample the model was trained on. It is obvious that one cannot maximize such an out-of-sample average analytically over an arbitrary family of models. However, one can easily (numerically) maximize $E_{\tilde{p}}[U(b^*(q), O)]$ over a one-parameter family of models. This is the approach we will take. In Sections 2.2 and 2.3 we will describe how one can construct a one-parameter family of candidate models based on the idea of an efficient frontier, which we define in terms of consistency with the training data and consistency with a prior distribution. We shall see that the construction of the candidate models involves a regularized in-sample expected utility maximization, and that each candidate model is robust in the sense that, for its level of consistency with the data, it maximizes the worst case outperformance relative to the benchmark model. This is another rationale for our choice of candidate models.

For our approach we will make use of the concept of generalized relative entropy, which was introduced in Friedman and Sandow (2003a) (Section 2.2). A very similar generalization of the relative entropy was independently introduced by Grünwald and Dawid (2002). The approach in Grünwald and Dawid (2002) is based on the idea of expected loss, and is therefore closely related to the approach in Friedman and Sandow (2003a), in which the utility function of an investor who bets (utility-optimally) on horses leads to an expected utility that can be viewed as the negative of an expected loss. Unlike Grünwald and Dawid (2002), however, Friedman and Sandow (2003a) explicitely link the decision-maker's/investor's action/investment-strategy to the probabilities he assigns to the states of the world. In Friedman and Sandow (2003a), the generalized relative entropy between the measures q and q^0 was defined as

$$D_{U,O}(q||q^0) = \sum_y q_y U(b^*_y(q) O_y) - \sum_y q_y U(b^*_y(q^0) O_y) \tag{6}$$

It can be interpreted as the loss in expected utility experienced by an investor who bets according to model q^0 when q is the true probability measure. It has been shown (see Friedman and Sandow, 2003a, Theorem 2) that $D_{U,O}(q||q^0)$ is a strictly convex function of q and that $D_{U,O}(q||q^0) \geq 0$ with equality if and only if $q = q^0$. We note that for $U(W) = \gamma_1 log(W+\gamma) + \gamma_2$, $D_{U,O}$ reduces to the Kullback-Leibler relative entropy, up to a constant factor (see Theorem 3 in Section 2.5, below).

2.2 Modeling Approach

We consider the tradeoff between consistency with the data and consistency with the investor's prior beliefs (as encoded in the prior measure, q^0). This approach is similar to others; see, for example, Lebanon and Lafferty (2001) or Wu (1997). However, we strive to formulate our problem(s) in economically meaningful ways. Given models equally consistent with the investor's prior beliefs, we assume that the investor prefers a model that is more consistent with the data; given models equally consistent with the data, we assume that the investor prefers a model that is more consistent with the investor's prior beliefs. We also show that these assumptions lead to measures which are robust, in the sense that they maximize a worst-case relative outperformance over the benchmark model. We make all of this precise below.

2.2.1 CONSISTENCY WITH THE DATA

For a model measure $q \in Q$, let $\mu^{data}(q)$ denote the investor's measure of the consistency of q with the data; this consistency is expressed in terms of expectations of the *feature vector*, $f(y) = (f_1(y), \ldots, f_J(y))^T \in \mathbf{R}^J$ where each feature, [3] f_j, is a mapping from $\mathbf{R}$ to $\mathbf{R}$. We make the following assumption:

3. For further discussion, see, for example, Vapnik (1999).

Assumption 2 *The investor measures the consistency,*[4] $\mu^{data}(q)$*, of the model* $q \in Q$ *with the data as a strictly monotone decreasing function of the large sample probability density of the sample feature means, evaluated at the model q feature expectations,* $E_q[f]$.

It is possible (for small sample sets, for example) to develop the theory under more general assumptions, by considering more general families of convex level sets.

To elaborate, for a fixed measure $q \in Q$, the model feature mean, $E_q[f_j]$, is a deterministic quantity depending on q. The sample mean of f_j, however, depends on the sample set and is therefore a random variable, ϕ_j. The quantity $E_{\tilde{p}}[f_j]$ is therefore an observation of the random variable ϕ_j. By the Central Limit Theorem, for a large number of observations, N, the random vector $\phi = (\phi_1, \ldots, \phi_J)^T$ is approximately normally distributed with mean $E_{\tilde{p}}[f]$ and covariance matrix $\frac{1}{N}\Sigma$, where Σ is the empirical feature covariance matrix. Therefore, for a given measure $q \in Q$, the probability density for the random variable ϕ, evaluated at $E_q[f]$ is (approximately) given by

$$p^c(c) \equiv pdf(\phi)|_{\phi=E_q[f_j]} = (2\pi)^{-\frac{1}{2}J} N^{\frac{1}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{N}{2}c^T\Sigma^{-1}c}, \tag{7}$$

where

$$c = (c_1, \ldots, c_J)^T$$

and

$$c_j = E_q[f_j] - E_{\tilde{p}}[f_j]. \tag{8}$$

We note that though we have used the Central Limit Theorem to *motivate* our assumption, we have not made any assumption on the probability distribution of the measures $q \in Q$. Our assumption allows us to parameterize the degree of consistency of a measure q with the data. Equally consistent measures, q, lie on the same level set of the function $\mu^{data}(q)$. We parameterize the nested family of sets, consisting of points $q \in Q$ that are equally consistent with the data.

Note that, by construction, $\mu^{data}(q)$ is invariant with respect to translations and rotations of the feature vectors. James Huang (2003) first pointed this out to the authors.

2.2.2 CONSISTENCY WITH THE PRIOR MEASURE

To quantify consistency of the model, $q \in Q$, with the investor's prior beliefs, we make use of the generalized entropy $D_{U,O}(q||q^0)$, where U is the investor's utility function and O is the set of odds ratios.

Assumption 3 *The investor measures the consistency,* $\mu^{prior}(q)$*, of the model* $q \in Q$ *with the prior,* q^0*, by using some strictly monotone increasing function of the generalized relative entropy* $D_{U,O}(q||q^0)$.

More precisely, low μ values are associated with highly consistent models and high μ values are associated with less consistent models. We shall see in Section 2.3 that generalized relative entropy is an appropriate measure of consistency, as it leads to models which asymptotically maximize expected utility.

4. More precisely, low μ values are associated with highly consistent models and high μ values are associated with less consistent models.

2.2.3 PARETO OPTIMAL MEASURES

To characterize the measures $q^* \in Q$ which are optimal (in a sense to be made precise), we define *dominance*, *Pareto optimal* probability measures, the set of *achievable measures*, and the *efficient frontier*. These notions are from vector optimization theory (see, for example, Boyd and Vandenberghe, 2001) and portfolio theory (see, for example, Luenberger, 1998).

Definition 3 *$q^1 \in Q$ dominates $q^2 \in Q$ with respect to μ^{data} and μ^{prior} if*

(i)

$$(\mu^{data}(q^1), \mu^{prior}(q^1)) \neq (\mu^{data}(q^2), \mu^{prior}(q^2))$$

and

(ii)

$$\mu^{data}(q^1) \leq \mu^{data}(q^2)$$

and

$$\mu^{prior}(q^1) \leq \mu^{prior}(q^2).$$

Observe that $q^1 \in Q$ dominates $q^2 \in Q$ with respect to μ^{data} and μ^{prior} if and only if $q^1 \in Q$ dominates $q^2 \in Q$ with respect to $t^{data}(\mu^{data})$ and $t^{prior}(\mu^{prior})$, where t^{data} and t^{prior} are strictly monotone increasing functions. Therefore, it follows from Equation 7 and Assumptions 2 and 3 that, without loss of generality, we may continue our discussion with[5]

$$\mu^{data}(q) = \alpha(q) \equiv N c^T \Sigma^{-1} c \geq 0 \tag{9}$$

and

$$\mu^{prior}(q) = D_{U,O}(q||q^0).$$

This choice is convenient as it leads to numerically tractable convex optimization problems (see Section 2.3, below). We note that $\alpha(q)$ is the Mahalanobis distance. For a definition of the Mahalanobis distance and its properties, see, for example, Kullback (1997), p. 190.

Definition 4 *A model, $q^* \in Q$, is Pareto optimal if and only if no measure $q \in Q$ dominates q^* with respect to $\mu^{data}(q) = \alpha(q)$ and $\mu^{prior}(q) = D_{U,O}(q||q^0)$. The efficient frontier is the set of Pareto optimal measures.*

We note that for any Pareto optimal measure $q^* \in Q$,

$$\alpha(q) \leq \alpha(q^*) \text{ implies that } D_{U,O}(q||q^0) \geq D_{U,O}(q^*||q^0) \tag{10}$$

for all $q \in Q$.

The Pareto optimal measures are contained in the *achievable set, $\mathcal{A}$*, which is defined as follows:

5. This form for μ^{data} leads to the regularization used, for example, in Wu (1997), and Gull and Daniell (1978).

Definition 5 *The achievable set, $\mathcal{A}$, is given by*

$$\mathcal{A} = \{(\alpha, D) | \alpha(q) \leq \alpha \text{ and } D_{U,O}(q||q^0) \leq D \text{ for some } q \in Q\} \subset \mathbf{R}^2.$$

We slightly abuse notation: we sometimes use α and D to denote functions, and at other times we use the same symbols to denote real values; our intentions should be clear from the context.

By Equations 8 and 9, measures q that are equally consistent with the data lie on the same level set of the function

$$\alpha(q) = N(E_q[f] - E_{\tilde{p}}[f])^T \Sigma^{-1} (E_q[f] - E_{\tilde{p}}[f]). \tag{11}$$

We parameterize the nested family of sets, consisting of points $q \in Q$ that are equally consistent with the data, by Equation 11. We note that Σ^{-1} is a nonnegative definite matrix, so $\alpha(q)$ is a convex function of q.

The achievable set, $\mathcal{A}$, is convex. To see this, recall that $\alpha(q)$ and $D_{U,O}(q||q^0)$ are convex functions of q (see the remark following Equation 11 for the convexity of $\alpha(q)$; see Friedman and Sandow (2003a), Theorem 2 for the strict convexity of $D_{U,O}(q||q^0)$). $\mathcal{A}$ is convex by the convexity of $\alpha(q)$ and $D_{U,O}(q||q^0)$ (see, for example, Boyd and Vandenberghe, 2001, Section 4.7). The convexity of the achievable set follows from the particular choice $\mu^{data}(q) = \alpha(q)$ and $\mu^{prior}(q) = D_{U,O}(q||q^0)$.

We may visualize the achievable set, $\mathcal{A}$, and the efficient frontier as displayed in Figure 1, which also incorporates the following lemma.

Lemma 1 *If q^* is a Pareto optimal measure, then*

(i) $\alpha(q^*) \leq \alpha_{max}$, *where*

$$\alpha_{max} = N(E_{q^0}[f] - E_{\tilde{p}}[f])^T \Sigma^{-1} (E_{q^0}[f] - E_{\tilde{p}}[f]). \tag{12}$$

(ii) $(\alpha(q^*), D_{U,O}(q^*||q^0))$ *lies on the lower $D-$boundary of $\mathcal{A}$.*

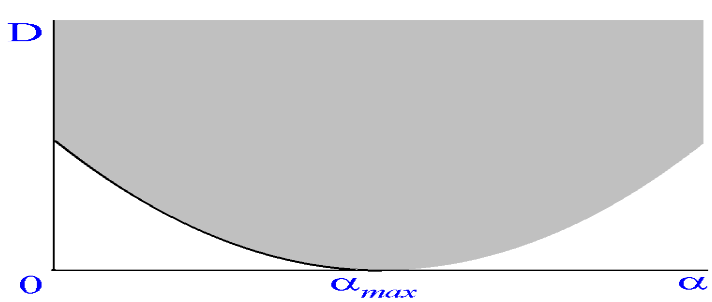

Figure 1: Achievable set, $\mathcal{A}$: shaded region above curve; Efficient Frontier: points on bold curve with $0 \leq \alpha \leq \alpha_{max}$.

Proof: (i) For the measure q^0, we have $\alpha(q^0) = \alpha_{max}$ and $D = D_{U,O}(q^0||q^0) = 0$. If $\alpha(q) > \alpha_{max}$, then q cannot be identical to q^0, so $D_{U,O}(q||q^0) > 0$, so q is dominated by q^0 and cannot be efficient. (ii) follows directly from Equation 10. □

We shall make use of the preceding lemma when we formulate our optimization problem.

We make the following assumption, which serves as one of our guiding principles.

Assumption 4 *The investor selects a measure on the efficient frontier.*

Thus, given a set of measures equally consistent with the prior, our investor prefers measures that are more consistent with the data, and, given a set of measures equally consistent with the data, he prefers measures that are more consistent with the prior. He makes no assumptions about the precedence of these two preferences. We shall see (in Section 2.2.5) that every Pareto optimal measure is robust in the sense that it maximizes, over all measures, the worst-case (over measures equally consistent with the data) relative outperformance of the model over the benchmark model.

2.2.4 CONVEX OPTIMIZATION PROBLEM

We seek the set of Pareto optimal measures. That is, motivated by Lemma 1, for all $q \in Q$ with $\alpha(q) = \alpha$, we seek all solutions of the following problem, as α ranges from 0 to α_{max}, where α_{max} is defined in Equation 12.

Problem 1 *(Initial Problem, Given* $\alpha, 0 \leq \alpha \leq \alpha_{max}$*)*

$$
\begin{aligned}
\textit{Find} \quad & \arg \inf_{q \in (R^+)^m, c \in R^J} D_{U,O}(q||q^0) && (13)\\
\textit{under the constraints } 1 \quad &= \sum_y q_y && (14)\\
\textit{and} \quad Nc^T \Sigma^{-1} c \quad &= \alpha && (15)\\
\textit{where} \quad c_j \quad &= E_q[f_j] - E_{\tilde{p}}[f_j] \ . && (16)
\end{aligned}
$$

Problem 1 is not a standard convex optimization problem (see, for example, Berkovitz, 2002), since Equation 15 is a non-affine equality constraint. However, we formulate a different (strictly convex optimization) problem, which, as we shall show, has the same solutions:

Problem 2 *(Initial Strictly Convex Problem, Given* $\alpha, 0 \leq \alpha \leq \alpha_{max}$*)*

$$
\begin{aligned}
\textit{Find} \quad & \arg \min_{q \in (R^+)^m, c \in R^J} D_{U,O}(q||q^0) && (17)\\
\textit{under the constraints } 1 \quad &= \sum_y q_y && (18)\\
\textit{and} \quad Nc^T \Sigma^{-1} c \quad &\leq \alpha && (19)\\
\textit{where} \quad c_j \quad &= E_q[f_j] - E_{\tilde{p}}[f_j] \ . && (20)
\end{aligned}
$$

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

Proof: See Appendix B.

In order to visualize the solutions to Problem 2, we define

$$S_\alpha = \{q : Nc^T\Sigma^{-1}c = \alpha, q \in Q_c\}, \tag{21}$$

where

$$Q_c = \{q : q \geq 0, \sum_y q_y = 1, \sum_y q_y f_j(y) = c_j + E_{\tilde{p}}[f_j], j = 1, \ldots, J\}.$$

By solving Problem 2, for each α, we generate a one-parameter family of candidate models, $q^*(\alpha)$, indexed by α. We can visualize these models as the points of tangency (on the probability simplex, Q) of the nested surfaces of the families S_α and the level sets of $D_{U,O}(q||q^0)$ (see Figure 2). Each candidate model, $q^*(\alpha)$, is a solution of Problem 2; accordingly, each point $(\alpha, D_{U,O}(q^*(\alpha)||q^0))$ is a point on the efficient frontier (see Figure 1), and the efficient frontier consists of all points of the form $(\alpha, D_{U,O}(q^*(\alpha)||q^0))$, as α ranges from $(0, \alpha_{max})$.

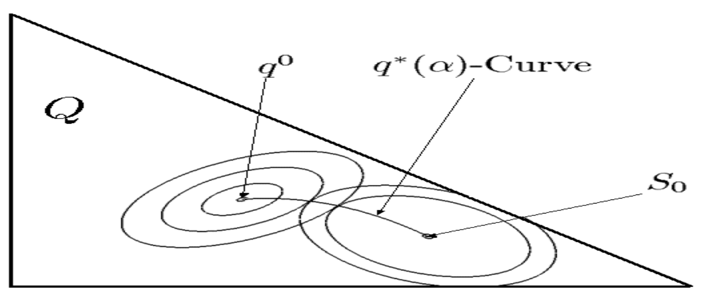

Figure 2: The sets S_α (see Equation 21), centered at S_0, the $q^*(\alpha)$-curve, and the level sets of $D_{U,O}(q||q^0)$, centered at q^0, on the probability simplex Q

2.2.5 ROBUSTNESS OF THE PARETO OPTIMAL MEASURES

We can measure the quality of a model q by the relative outperformance of the model over the benchmark model,[6] q^0: the gain in expected (under the true measure) utility experienced by an investor who invests optimally according to the model relative to an investor who invests optimally according to the benchmark model q^0 (see Friedman and Sandow, 2003a). That is, the relative outperformance is given by $E_{p'}\left[U(b^*(q), O) - U(b^*(q^0), O)\right]$, where p' is a potential *true* probability measure. It follows from the fact that the Pareto optimal measures are solutions of Problem 2 and from Theorem 5 (in Appendix A) that every Pareto optimal measure, $q^*(\alpha)$, is robust in the following sense: it maximizes, over all measures, the worst-case (with respect to potential true measures

6. Here, we look at the prior, q^0, from a slightly different perspective; in this context, we view q^0 as a model against which we benchmark performance.

equally consistent with the data) relative outperformance of the model over the benchmark model, i.e.,

$$q^*(\alpha) = \arg\max_{q\in Q} \min_{p'\in S_\alpha} \left\{ E_{p'}\left[U(b^*(q),O)\right] - E_{p'}\left[U(b^*(q^0),O)\right]\right\} .$$

2.2.6 Choosing a Measure on the Efficient Frontier

According to our paradigm, the best candidate model lies on a one-parameter efficient frontier. In order to choose the best candidate model from this one-parameter family, we make the following assumption.

Assumption 5 *The investor chooses* α *so as to maximize his expected utility on an out-of-sample data set.*

Thus, given a utility function, U, odds ratios, O, and a prior belief, q^0, Assumptions 1 to 5 lead to a method for finding a probability measure

$$\begin{aligned} q^{**} &= q^*(\alpha^*)\ , \\ \text{with } \alpha^* &= \arg\max_{\alpha} E_{\tilde{p}}[U(b^*(q^*(\alpha)),O)]\ , \end{aligned}$$

where $\tilde{p}$ is the empirical measure of the test set. In our method, the relative importance of the data and the prior is determined by the out-of-sample performance (expected utility) of the model.

2.2.7 Informal Comment: Practical Bound for α

In practice, given a confidence level, l, under Assumption 2, we can search over the range,

$$\alpha \in (0, \alpha_{search}),$$

where

$$\alpha_{search} = min(\alpha_l, \alpha_{max})$$

and

$$\alpha_l = (cdf_{\chi^2_J})^{-1}(l)$$

(see, for example, Davidson and MacKinnon, 1993 or Wu, 1997). That is, we search until

(*i*) we are $100 \cdot l\%$ confident that the true value of c is within the region $Nc^T\Sigma^{-1}c \leq \alpha$,

(*ii*) the region $Nc^T\Sigma^{-1}c \leq \alpha$ includes q^0 and q^* is insensitive to further increasing the value of α.

In practice, the covariance matrix Σ may be nearly singular, so we may need to regularize it to insure that our search spans $c-$space.

2.3 Dual Problem

We have shown in Section 2.2 that, in order to find the Pareto optimal model, q^*, for a given α, we have to solve Problem 2. As we have seen, this problem is strictly convex. Convex problems are known to have so called dual problems.

We show in Appendix C that the dual of Problem 2 can be formulated as:

Problem 3 *(Easily Interpreted Version of Dual Problem, Given* α*)*

$$\begin{aligned}
\textit{Find } \beta^* &= \arg\max_{\beta} h(\beta) && (22)\\
\textit{with} \quad h(\beta) &= \sum_y \tilde{p}_y U(b_y^*(q^*) O_y) - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}}, && (23)\\
\textit{where } b_y^*(q^*) &= \frac{1}{O_y} (U')^{-1} \left(\frac{\lambda^*}{q_y^* O_y} \right) && (24)\\
\textit{and} \quad q_y^* &= \frac{\lambda^*}{O_y U'(U^{-1}(G_y(q^0, \beta, \mu^*)))}, && (25)\\
\textit{with} \quad G_y(q^0, \beta, \mu^*) &= U(b_y^*(q^0) O_y) + \beta^T f(y) - \mu^*, && (26)\\
\textit{where} \quad \mu^* \textit{ solves} \quad 1 &= \sum_y \frac{1}{O_y} U^{-1} \left(G_y(q^0, \beta, \mu^*) \right), && (27)\\
\textit{and} \quad \lambda^* &= \left\{ \sum_y \frac{1}{O_y U'(U^{-1}(G_y(q^0, \beta, \mu^*)))} \right\}^{-1}. && (28)
\end{aligned}$$

Equation 25 is often referred to as the connecting equation (see, for example, Lebanon and Lafferty, 2001).

We also show in Appendix C that an alternative formulation of the dual problem is the following:

Problem 4 *(Easily Implemented Version of Dual Problem, Given* α*)*

$$\begin{aligned}
\textit{Find } \beta^* &= \arg\max_{\beta} h(\beta) && (29)\\
\textit{with} \quad h(\beta) &= \beta^T E_{\tilde{p}}[f] - \mu^* - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}}, && (30)\\
\textit{where} \quad \mu^* \textit{ solves} \quad 1 &= \sum_y \frac{1}{O_y} U^{-1} \left(G_y(q^0, \beta, \mu^*) \right) && (31)\\
\textit{with} \quad G_y(q^0, \beta, \mu^*) &= U(b_y^*(q^0) O_y) + \beta^T f(y) - \mu^*. && (32)
\end{aligned}$$

The optimal probability distribution is then

$$\begin{aligned}
q_y^* &= \frac{\lambda^*}{O_y U'(U^{-1}(G_y(q^0, \beta^*, \mu^*)))}, && (33)\\
\textit{with} \quad \lambda^* &= \left\{ \sum_y \frac{1}{O_y U'(U^{-1}(G_y(q^0, \beta^*, \mu^*)))} \right\}^{-1}. && (34)
\end{aligned}$$

We state the following theorem:

Theorem 1 *Problems 2, 3 and 4 have the same unique solution, q^*_y.*

Proof: see Appendix C.

Problems 3 and 4 are equivalent. Problem 4 is easier to implement and Problem 3 is easier to interpret. The first term in the objective function of Problem 3 is the utility (of the utility maximizing investor) averaged over the training sample. Thus, our dual problem is a regularized maximization of the training-sample averaged utility, where the utility function, U, is the utility function on which the generalized relative entropy $D_{U,O}(q||q^0)$ depends.

The dual problems, Problems 3 and 4, are J-dimensional (J is the number of features), unconstrained, concave maximization problems (see Boyd and Vandenberghe, 2001, p. 159 for the concavity). The primal problem, Problem 2, on the other hand, is an m-dimensional (m is the number of states) convex minimization with convex quadratic constraints. The dual problem, Problem 4, may be easier to solve than the primal problem, Problem 2, if $m > J$. In the more general context discussed in section 3, the dual problem will always be easier to solve than the primal problem.

We note that we can obtain the same $\alpha-$parameterized family of solutions to Problems 3 and 4, if we allow α to vary over $[0,\infty)$, by dropping the square roots in Equations 23 and 30; we show that this is so in Appendix E.

2.3.1 Asymptotic Optimality

It follows from Equation 23 that

Theorem 2 *As $N \to \infty$, to leading order, the optimal solution to Problem 2 maximizes (over the parametric family prescribed by the connecting equation, Equation 25) the expected utility for the investor.*

2.3.2 Example: A Logarithmic Family of Utilities

We consider a utility of the form

$$U(z) = \gamma_1 log(z+\gamma) + \gamma_2 \ , \ \gamma > -\frac{1}{\sum_y \frac{1}{O_y}} \ , \ \gamma_1 > 0 \ , \tag{35}$$

(see Theorems 3 and 4 in Friedman and Sandow, 2003a). This logarithmic family is rich enough to describe a wide range of risk aversions; and it can be used to approximate non-logarithmic utility functions (see Friedman and Sandow, 2003a).

In Appendix D, we show that the dual problem is given by:

Problem 5 *(Dual Problem for our Logarithmic Family of Utilities)*

$$\begin{aligned}
\textit{Find } \beta^* &= \arg\max_{\beta} h(\beta) \\
\textit{with} \quad h(\beta) &= \sum_y \tilde{p}_y \log q^*_y - \sqrt{\frac{\alpha}{\gamma_1^2}\frac{\beta^T \Sigma \beta}{N}} \ , \\
\textit{where} \quad q^*_y &= \frac{1}{\sum_y q^0_y e^{\beta^T f(y)}} \, q^0_y \, e^{\beta^T f(y)} \ .
\end{aligned}$$

This problem is equivalent to a regularized maximum likelihood search, which is independent of the odds ratios, O; this is consistent with Section 2.5, where we show that the odds ratios drop out of the primal problem for this logarithmic family of utility functions.

2.3.3 EXAMPLE: POWER UTILITY

We consider a utility of the form

$$U(z) \;=\; \frac{z^{1-\kappa}-1}{1-\kappa}\;\;. \tag{36}$$

(see Section 2.1.1 in Friedman and Sandow, 2003a). In order to specify the dual problem for this utility, note that

$$\begin{aligned} U'(z) &= z^{-\kappa}, & (37)\\ U^{-1}(z) &= [1+(1-\kappa)z]^{\frac{1}{1-\kappa}} & (38)\\ \text{and } U'(U^{-1}(z)) &= [1+(1-\kappa)z]^{\frac{-\kappa}{1-\kappa}}\;\;. & (39) \end{aligned}$$

One can show that

$$\begin{aligned} b^*(q) &= \frac{(q_y O_y)^{\frac{1}{\kappa}}}{O_y B(q,O)} & (40)\\ \text{with } B(q,O) &= \sum_y \frac{1}{O_y}(q_y O_y)^{\frac{1}{\kappa}}\;\;, & (41) \end{aligned}$$

(see Section 2.1.1 in Friedman and Sandow, 2003a). Using this equation, we can write $G_y(q^0,\beta,\mu^*)$ from Equation 26 as

$$G_y(q^0,\beta,\mu^*) \;=\; \frac{1}{1-\kappa}\left[\left(\frac{(q^0_y O_y)^{\frac{1-\kappa}{\kappa}}}{(B(q^0,O))^{1-\kappa}}\right)-1\right] \;+\; \beta^T f(y) \;-\; \mu^*\;\;. \tag{42}$$

Inserting Equations 38 and 42 into Equation 27 gives

$$1 \;\;=\;\; \sum_y \frac{1}{O_y}\left[\frac{(q^0_y O_y)^{\frac{1-\kappa}{\kappa}}}{(B(q^0,O))^{1-\kappa}}+(1-\kappa)[\beta^T f(y)-\mu^*]\right]^{\frac{1}{1-\kappa}}\;\;, \tag{43}$$

which is our condition for μ^*. Next we specify the condition Equation 28 for λ^*. We use Equations 39 and 42 to write Equation 28 as

$$\lambda^* \;\;=\;\; \left\{\sum_y \frac{1}{O_y}\left[\frac{(q^0_y O_y)^{\frac{1-\kappa}{\kappa}}}{(B(q^0,O))^{1-\kappa}}+(1-\kappa)[\beta^T f(y)-\mu^*]\right]^{\frac{\kappa}{1-\kappa}}\right\}^{-1}\;\;. \tag{44}$$

By means of Equations 25, 39 and 42 we obtain for the optimal probability distribution

$$q^*_y \;\;=\;\; \frac{1}{O_y}\left[\frac{(q^0_y O_y)^{\frac{1-\kappa}{\kappa}}}{(B(q^0,O))^{1-\kappa}}+(1-\kappa)[\beta^T f(y)-\mu^*]\right]^{\frac{\kappa}{1-\kappa}}\;\;. \tag{45}$$

Collecting Equations 43, 41, 44 and 45, we obtain:

Problem 6 *(Dual problem for power utility)*

$$\begin{aligned}
\textit{Find } \beta^* &= \arg\max_{\beta} h(\beta) \\
\textit{with} \quad h(\beta) &= \beta^T E_{\tilde{p}}[f] - \mu^* - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}} \ , \\
\textit{where} \quad \mu^* \textit{ solves} \quad 1 &= \sum_y \frac{1}{O_y} \left[\frac{(q_y^0 O_y)^{\frac{1-\kappa}{\kappa}}}{(B(q^0, O))^{1-\kappa}} + (1-\kappa)[\beta^T f(y) - \mu^*] \right]^{\frac{1}{1-\kappa}} \\
\textit{with} \quad B(q^0, O) &= \sum_y \frac{1}{O_y} (q_y^0 O_y)^{\frac{1}{\kappa}} \ .
\end{aligned}$$

The optimal probability distribution is then

$$\begin{aligned}
q_y^* &= \frac{\lambda^*}{O_y} \left[\frac{(q_y^0 O_y)^{\frac{1-\kappa}{\kappa}}}{(B(q^0, O))^{1-\kappa}} + (1-\kappa)[\beta^{*T} f(y) - \mu^*] \right]^{\frac{\kappa}{1-\kappa}} \\
\textit{with} \quad \lambda^* &= \left\{ \sum_y \frac{1}{O_y} \left[\frac{(q_y^0 O_y)^{\frac{1-\kappa}{\kappa}}}{(B(q^0, O))^{1-\kappa}} + (1-\kappa)[\beta^{*T} f(y) - \mu^*] \right]^{\frac{\kappa}{1-\kappa}} \right\}^{-1}
\end{aligned}$$

2.4 Summary of Modeling Approach

The modeling approach described in Sections 2.2 and 2.3 is based on the idea that our investor selects a Pareto optimal model, i.e. a model on an efficient frontier, which we have defined in terms of consistency with the training data and consistency with a prior distribution. We measured the former by means of the large-sample distribution of a vector of sample-averaged features, and the latter by means of a generalized relative entropy. We have seen that the measures on the efficient frontier form a family which is parameterized by the single parameter $\alpha \in (0, \alpha_{max})$, and that, for a given α, the Pareto optimal measure is the unique solution of Problem 2, which is a strictly convex optimization problem. Moreover, the Pareto optimal measures are robust in the sense of Theorem 5. For a given α, the Pareto optimal measure can be found by solving the dual (concave maximization) problem in the form of Problem 3 or in the form of Problem 4. Solving this dual problem amounts to a regularized expected utility maximization (over the training sample) over a certain family of measures; for many practical examples, solving the dual problem can be easier than solving the primal problem. Having thus computed an α-parameterized family of Pareto optimal measures, we pick the measure with highest expected utility on a hold-out sample.

We note that the procedure to select α is, by virtue of the fact that α is one-dimensional, both tractable and barely susceptible to overfitting on the hold-out sample set.

Our approach boils down to the following procedure:

1. Break the data into a training set and a hold-out sample. (In their numerical experiments, Friedman and Sandow, 2003b, Friedman and Huang, 2003, and Sandow et al., 2003 used 75% or 80% of the data, selected randomly, to train the model.)
2. Choose a discrete set $A = \{\alpha_k \in (0, \alpha_{max}), k = 1, \ldots, K\}$.
3. For $k = 1, \ldots, K$,
   - Solve Problem 4 for $\beta^*(\alpha_k)$, based on the training set,
   - Compute the out-of-sample performance $P_k = E_{\tilde{p}}[U(b^*(q^*(\alpha_k)), O)]$ on the out-of-sample test set, where $\tilde{p}$ is the empirical measure on this test set, and q^*, is determined from Equation 33 with parameters $\beta^*(\alpha_k)$, and b^* is determined from Equation 3.
4. Put $k^* = \arg\max_k P_k$.
5. Our model, q^{**}, is determined from Equation 33 with parameters $\beta^*(\alpha_{k^*})$.

2.5 Utilities Admitting Odds Ratio Independent Problems: a Logarithmic Family

Model builders who use probabilistic models make decisions (bets) which result in well defined benefits or ill effects (payoffs) in the presence of risk. In principle, the payoffs associated with the various outcomes can be assigned precise values; in practice, it may be difficult to assign such values. Outside the financial modeling context, for example, there may be no "market makers" who set odds ratios. Even in the financial modeling context, the data for the payoffs (or equivalently, market prices or odds ratios) may not exist or be of poor quality. In this context, given market prices on instruments which have nonzero payoffs for more than one state, we would need a complete market in order to calculate the odds ratios (see, for example, Duffie, 1996, for a definition of complete markets). In the absence of high quality data, one might consider modeling the odds ratios, but that introduces additional complexity; moreover, the resulting model, under a general utility function, will be sensitive to the odds ratio model.

For these reasons, we seek the most general family of utility functions for which our problem formulation is independent of the odds ratios. This family is specified in the following theorem

Theorem 3 *The generalized relative entropy, $D_{U,O}(q||q^0)$, is independent of the odds ratios, O, for any candidate model q and prior measure, q^0, if and only if the utility function, U, is a member of the logarithmic family*

$$U(W) = \gamma_1 \log(W + \gamma) + \gamma_2 \ , \ \forall\, W > \max\{0, -\gamma\} \ , \tag{46}$$

where $\gamma_1 > 0$, γ_2 and $\gamma > -\frac{1}{\sum_y \frac{1}{O_y}}$ are constants. In this case,

$$D_{U,O}(q||q^0) = \gamma_1 E_q \left[\log \left(\frac{q}{q^0} \right) \right] \ , \tag{47}$$

which depends on γ_1 in a trivial way and is independent of γ and γ_2.

Proof: First we prove that if the utility function has the form Equation 46 then $D_{U,O}(q||q^0)$ is independent of O, for any measures q and q^0, and Equation 47 holds. Theorem 4 in Friedman and Sandow (2003a) states that, if the utility function has the form Equation 46 then the relative performance measure[7]

$$\Delta_U(q^1,q^2,O) = \sum_y \tilde{p}_y[U(b^*_y(q^2)O_y) - U(b^*_y(q^1)O_y)] \tag{48}$$

is independent of O, for any measures q^1,q^2, and $\tilde{p}$. Putting $\tilde{p}=q$, $q^1=q^0$ and $q^2=q$, we see from Equations 48 and 6 that

$$D_{U,O}(q||q^0) = \Delta_U(q^1,q^2,O) \ .$$

It follows from Theorem 4 in Friedman and Sandow (2003a) that

$$D_{U,O}(q||q^0) = \gamma_1 E_q\left[\log\left(\frac{q}{q^0}\right)\right] \ ,$$

which is independent of O, for any measures q and q^0.

Next we prove the reverse: If $D_{U,O}(q||q^0)$ is independent of O, for any measures q, and q^0, then the utility function has the form Equation 46. If $D_{U,O}(q||q^0)$ is independent of O, for any measures q and q^0, then both $D_{U,O}(\tilde{p}||q^1)$ and $D_{U,O}(\tilde{p}||q^2)$ are independent of O, for any measures $\tilde{p},q^1$ and q^2. Consequently, the performance measure

$$\Delta_U(q^1,q^2,O) = D_{U,O}(\tilde{p}||q^1) - D_{U,O}(\tilde{p}||q^2)$$

is independent of O, for any measures $\tilde{p},q^1$ and q^2. It follows then from Theorem 3 in Friedman and Sandow (2003a) that the utility function has the form Equation 46. $\square$

From this theorem and Problem 2, we obtain

Corollary 1 *For utility functions of the form Equation 35, Problem 2 reduces to Problem 7.*

Problem 7 *(Initial Strictly Convex Problem for U in our Logarithmic Family, Given $\alpha, 0 \le \alpha \le \alpha_{max}$)*

$$\begin{aligned} \textit{Find} \qquad & \arg\min_{q\in(R^+)^m, c\in R^J} \gamma_1 E_q\left[\log\left(\frac{q}{q^0}\right)\right] \\ \textit{under the constraints } 1 \ &= \ \sum_y q_y \\ \textit{and} \ \ Nc^T\Sigma^{-1}c \ &\le \ \alpha \\ \textit{where} \ \ c_j \ &= \ E_q[f_j] - E_{\tilde{p}}[f_j] \ . \end{aligned}$$

We have already explicitly derived the dual problem for utility functions of the form Equation 46 in Section 2.3.2.

We note that the family of utility functions Equation 46 admits a wide range of risk aversions (see the discussion in Friedman and Sandow, 2003a, Section 2.3). Moreover, for utilities not of this form and horse races with sufficiently homogeneous expected returns, we can approximate well $D_{U,O}(q||q^0)$ by $D_{log,O}(q||q^0)$; see Friedman and Sandow (2003a), Theorem 5. For utilities in this logarithmic family, the primal problem (Problem 2) and equivalent dual problems (Problems 3 and 4) are independent of the odds ratios.

7. Model q^2 outperforms model q^1 if and only if $\Delta_U(q^1,q^2,O) > 0$.

3. Conditional Density Models

In this section we briefly discuss our approach in the context of a conditional density model which may include point masses, i.e. for the case where the random variable Y has the continuous conditional probability density $q(y|x)$ on the finite set $\mathcal{Y} \subset \mathbf{R}^n$ and the finite conditional point probabilities $q_{\rho|x}$ on the set of points $\{y_\rho \in R^n, \rho = 1, 2, ..., m\}$, where x denotes a value of the vector X of explanatory variables which can take any of the values $x_1, ..., x_M, x_i \in \mathbf{R}^d$. This setting has interesting applications such as the modeling of recovery values of defaulted debt (see Friedman and Sandow, 2003b).

3.1 Preliminaries

We generalize the results and definitions from Section 2.1. Let us denote by $\tilde{p}_x$ the empirical probability of the vector, X, of explanatory variables, and define the following conditional probability measures:

Definition 6

$$\begin{aligned} p &= \{(p(y|x), p_{\rho|x}), y \in \mathcal{Y}, \rho = 1, 2, ..., m, x = x_1, ..., x_M\} \\ &= \textit{true (unknown) conditional probability measure} \\ \tilde{p} &= \{(\tilde{p}(y|x), \tilde{p}_{\rho|x}), y \in \mathcal{Y}, \rho = 1, 2, ..., m, x = x_1, ..., x_M\} \\ &= \textit{empirical conditional probability measure} \\ q &= \{(q(y|x), q_{\rho|x}), y \in \mathcal{Y}, \rho = 1, 2, ..., m, x = x_1, ..., x_M\} \\ &= \textit{model conditional probability measure} \end{aligned}$$

Following Lebanon and Lafferty (2001), we assume that the following relations between conditional and joint probabilities hold:

$$\begin{aligned} p(y,x) &= \tilde{p}_x p(y|x), \\ p_{\rho,x} &= \tilde{p}_x p_{\rho|x}, \\ q(y,x) &= \tilde{p}_x q(y|x), \text{ and} \\ q_{\rho,x} &= \tilde{p}_x q_{\rho|x}. \end{aligned}$$

Next we identify the probabilistic problem with the conditional horse race (Friedman and Sandow, 2003a, Definition 8), and consider an investor who places bets on horses. We assume that our investor allocates $b(y|x)$ to the event[8] $Y = y$ and $b_{\rho|x}$ to the event $Y = y_\rho$, if $X = x$ was observed, where

$$1 = \int_{\mathcal{Y}} b(y|x) dy + \sum_{\rho=1}^{m} b_{\rho|x} \,. \tag{49}$$

Our investor allocates his assets so as to maximize his utility function, U, which is strictly concave, twice differentiable, and strictly monotone increasing. This means that an investor who believes the model q allocates according to

$$b^*[q] = \arg\max_{\{b \in \mathcal{B}\}} \left[\int_{\mathcal{Y}} q(y|x) U(b(y|x) O(x,y)) dy + \sum_y q_{\rho|x} U(b_{\rho|x} O_{x,\rho}) \right] ,$$

8. To be precise, we have to bet on finite partitions of the interval $\mathcal{Y}$ as described by Friedman and Sandow (2003a), Section 3.2.

where

$$\mathcal{B} = \{(b(y|x), b_{\rho|x}) : \int_{\mathcal{Y}} b(y|x)dy + \sum_{\rho=1}^{m} b_{\rho|x} = 1\}'$$

denotes the set of betting weights consistent with Equation 49, and the odds ratios $O(x,y)$ and $O_{x,\rho}$ are defined by Friedman and Sandow (2003a). It is straightforward to generalize the results from Friedman and Sandow (2003a) for the optimal betting weights to:

$$b^*[q](y|x) = \frac{1}{O(x,y)}(U')^{-1}\left(\frac{\lambda_x^*}{\tilde{p}_x q(y|x) O(x,y)}\right), \tag{50}$$

$$b^*_{\rho|x}[q] = \frac{1}{O_{x,\rho}}(U')^{-1}\left(\frac{\lambda_x^*}{\tilde{p}_x q_{\rho|x} O_{x,\rho}}\right), \tag{51}$$

where λ_x^* is the solution of the following equation:

$$1 = \int_{\mathcal{Y}} \frac{1}{O(x,y)}(U')^{-1}\left(\frac{\lambda_x^*}{\tilde{p}_x q(y|x) O(x,y)}\right) dy + \sum_{\rho} \frac{1}{O_{x,\rho}}(U')^{-1}\left(\frac{\lambda_x^*}{\tilde{p}_x q_{\rho|x} O_{x,\rho}}\right) \tag{52}$$

In analogy with Assumption 1, we make the following assumption:

Assumption 6 *For each x, there exists a solution to Equation 52.*

Equipped with above tools, we can formulate our modeling objective:

Objective:

$$\textit{Find} \qquad \arg\max_{q \in Q} E_p[U(b^*[q], O)] \ ,$$

i.e., find the model that maximizes the true expectation,

$$E_p[U(b^*[q], O)] = \sum_x \tilde{p}_x \left\{ \int_{\mathcal{Y}} p(y|x) U(b^*[q](y|x) O(x,y)) dy + \sum_y p_{\rho|x} U(b^*_{\rho|x}[q] O_{\rho|x}) \right\} \ ,$$

of the utility of an investor who bets according to the model.

As in the context of discrete probabilities, we don't know the true measure, p, so that we cannot solve above optimization problem exactly. We will use the same ideas as in Section 2 to find an approximate solution (see the discussion after Equation 5. To this end, we need to define the generalized relative entropy for conditional probability densities with point masses. We notice that the generalized relative entropy was defined by Friedman and Sandow (2003b), Section 4.2 based on the notion of expected utility under q^1 for an investor who invests q^2-optimal, which in our case is

$$\begin{aligned} E_{q^1}[U(b^*[q^2], O)] &= \sum_x \tilde{p}_x \int_{\mathcal{Y}} q^1(y|x) U(b^*[q^2](y|x) O(x,y)) dy \\ &\quad + \sum_{x,\rho} \tilde{p}_x q^1_{\rho|x} U(b^*_{\rho|x}[q^2] O_{\rho|x}) \ . \end{aligned} \tag{53}$$

This suggests the following definition of the generalized relative entropy for conditional probability densities with point masses

$$D_{U,O}(q||q^0) \quad = \quad E_q[U(b^*[q],O)] \; - \; E_q[U(b^*[q^0],O)] \;, \tag{54}$$

where the expectation of a function $g_x(y)$ is defined as

$$\begin{aligned} E_q[g] \quad &= \quad \sum_x \tilde{p}_x E_q[g|x] && (55) \\ \text{with } E_q[g|x] \quad &= \quad \left\{ \int_{\mathcal{Y}} q(y|x) g_x(y) dy \; + \; \sum_{\rho} q_{\rho|x} g_x(y) \right\} \; . \end{aligned}$$

3.2 Modeling Approach

In this section, we generalize the modeling paradigm from Section 2.2 to the case of a conditional probability density with point masses. To this end, let us define the spaces

$$\begin{aligned} Q \quad &= \quad \{(q(y|x), q_{\rho|x}) : q(y|x) \in \mathbf{L}^M_l[\mathcal{Y}]\} \;, && (56) \\ \text{and } \; Q^+ \quad &= \quad \{q : q \in Q, q(y|x) \geq 0, q_{\rho|x} \geq 0\} \end{aligned}$$

where the integer index $l > 1$ (the power l up to which we can perform the integral $\int_{\mathcal{Y}} q^l(y|x)dy$) is chosen such that the integrals Equation 54 exits. We assume that $q \in Q^+$.

We further assume that Assumptions 2-5 hold. This leads to the following optimization problem or a given value of α, which is analogous to Problem 2:

Problem 8 *(Convex Problem: Conditional Probability Density, Given α)*

$$\begin{aligned} \textit{Find} \qquad & \arg \min_{q \in Q^+, c \in \mathbf{R}^J} D_{U,O}(q||q^0) && (57) \\ \textit{under the constraints } 1 \quad &= \quad E_q[1|x] && (58) \\ \textit{and } \; N c^T \Sigma^{-1} c \quad &\leq \quad \alpha && (59) \\ \textit{where } \; c_j \quad &= \quad E_q[f_j] - E_{\tilde{p}}[f_j] \; . && (60) \end{aligned}$$

Here, as in the context of discrete probabilities, f_j denotes a feature; there are J features, each of which is a real-valued function of x and y.

According to Assumption 5, our investor will choose the measure that maximizes his expected utility among the measures that are the family (parameterized by α) of solutions to Problem 8.

3.3 Dual Problem

Like Problem 2, Problem 8 has a dual. In order to derive this dual problem, we note that $Q \times \mathbf{R}^J$ is a convex subset of a vector space, the constraints expressed by Equations 58-60 can be rewritten in terms of convex mappings into a normed space, and the equality constraints expressed by Equations 58 and 60 are linear. By Theorem 1 of Section 8.6 in Luenberger (1969), the dual problem is the maximization over $\xi \geq 0$, $\beta = (\beta_1, ..., \beta_J)^T$, $\mu = \{\mu_x, x = x_1, ..., x_M\}$, and $\nu = \{(\nu(y|x) \geq 0, \nu_{\rho|x} \geq 0), y \in \mathcal{Y}, \rho = 1, 2, ..., m, x = x_1, ..., x_M\}$ of $\inf_{q \in Q, c \in \mathbf{R}^J} \mathcal{L}(q, c, \beta, \xi, \mu, \nu)$, where

$$\begin{aligned} \mathcal{L}(q,c,\beta,\xi,\mu,\nu) \quad &= \quad D_{U,O}(q||q^0) + \beta^T \left\{c - E_q[f] + E_{\tilde{p}}[f]\right\} \; + \; \xi \frac{1}{2} \left\{N c^T \Sigma^{-1} c - \alpha\right\} \\ & \quad + \; \sum_x \mu_x \tilde{p}_x \left\{E_q[1|x] - 1\right\} \; - \; E_q[\nu] \;\; , \end{aligned}$$

is a generalization of the Lagrangian Equation 84 for the case of discrete probabilities. One can find $\inf_{q\in Q, c\in \mathbf{R}^J} \mathcal{L}(q,c,\beta,\xi,\mu,\nu)$ the same way as we have done in Appendix C for discrete probabilities; the only difference is that we have to use Fréchet derivatives instead of ordinary ones. As a result, we obtain the analog of the connecting equation described in Appendix C. We can then continue along the lines from Appendix C, showing that $\nu = 0$ and finding ξ^* and μ^*. This leads to the dual of Problem 8:

Problem 9 *(Dual Problem: Conditional Probability Density, Given* α*)*

$$
\begin{aligned}
\textit{Find} \quad \beta^* &= \arg\max_{\beta} h(\beta) && (61)\\
\textit{with} \quad h(\beta) &= E_{\tilde{p}}[U(b^*[q^*],O)] - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}}\,, && (62)\\
\textit{where } b^*[q^*](y|x) &= \frac{1}{O(x,y)} (U')^{-1}\left(\frac{\lambda_x^*}{q^*(y|x)O(x,y)}\right), && (63)\\
b^*_{\rho|x}[q^*] &= \frac{1}{O_{x,\rho}} (U')^{-1}\left(\frac{\lambda_x^*}{q^*_{\rho|x} O_{x,\rho}}\right), && (64)\\
\textit{and} \quad q^*(y|x) &= \frac{\lambda_x^*}{O(x,y)U'\left(U^{-1}(G(x,y,q^0,\beta,\mu_x^*))\right)}\,, && (65)\\
q^*_{\rho|x} &= \frac{\lambda_x^*}{O_{\rho|x}U'\left(U^{-1}(G_{\rho|x}(q^0,\beta,\mu_x^*))\right)}\,, && (66)\\
\textit{with} \quad G(x,y,q^0,\beta,\mu_x^*) &= U(b^*[q^0](y|x)O(x,y)) + \beta^T f(x,y) - \mu_x^*, && (67)\\
G_{\rho|x}(q^0,\beta,\mu_x^*) &= U(b^*_{\rho|x}[q^0]O_{\rho|x}) + \beta^T f(x,y_\rho) - \mu_x^*\,, && (68)\\
\textit{where} \quad \mu_x^* \textit{ solves} \quad 1 &= \int_{\mathcal{Y}} \frac{1}{O(x,y)} U^{-1}\left(G(x,y,q^0,\beta,\mu_x^*)\right) dy && (69)\\
&\quad + \sum_{\rho} \frac{1}{O_{x,\rho}} U^{-1}\left(G_{\rho|x}(q^0,\beta,\mu_x^*)\right)\,, && (70)\\
\textit{and} \quad (\lambda_x^*)^{-1} &= \int_{\mathcal{Y}} \frac{1}{O(x,y)U'\left(U^{-1}(G(x,y,q^0,\beta,\mu_x^*))\right)} dy\\
&\quad + \sum_{\rho} \frac{1}{O_{\rho|x}U'\left(U^{-1}(G_{\rho|x}(q^0,\beta,\mu_x^*))\right)}\,. && (71)
\end{aligned}
$$

This dual problem is a straightforward generalization of the dual problem for discrete probabilities, Problem 3. In general, it is easier to solve the dual problem than the primal problem.

The following theorem, which follows from Equation 62, is the analog of Theorem 2:

Theorem 4 *As $N \to \infty$, to leading order, the optimal solution to Problem 8 maximizes (over the parametric family prescribed by the connecting equation) the expected utility for the investor.*

As for the discrete probability models, discussed in Section 2.2.5, under mild regularity conditions, every Pareto optimal measure is robust in the sense that it maximizes, over all measures, the worst-case (over measures equally consistent with the data) relative outperformance of the model over the benchmark model (see Appendix A).

We note that we can obtain the same $\alpha-$parameterized family of solutions to Problem 10, if we allow α to vary over $[0,\infty)$, by dropping the square root in Equation 75; we show that this is so in Appendix E.

3.3.1 EXAMPLE: UTILITIES FROM OUR LOGARITHMIC FAMILY

Because of its practical relevance, we state the above dual problem for the case of a utility from the logarithmic family Equation 35. It is easy to see that, in this case, the Equations 50-52 for the optimal betting weights give:

$$b^*_{\rho|x}[q] = q_{\rho|x}\left[1+\gamma\sum_{\rho'}\frac{1}{O_{x,\rho'}}\right]-\frac{\gamma}{O_{x,\rho}} \tag{72}$$

$$b^*[q](y|x) = q(y|x)\left[1+\gamma\sum_{y'}\frac{1}{O(x,y')}\right]-\frac{\gamma}{O(x,y)}\;. \tag{73}$$

The generalized relative entropy, which enters Problem 8, is then

$$D_{U,O}(q||q^0) = \gamma_1 E_q\left[\log\left(\frac{q}{q^0}\right)\right]\;. \tag{74}$$

Inserting Equations 35, 72 and 73 into Problem 9, we derive the dual problem as:

Problem 10 *(Dual Problem for Probability Densities and our Logarithmic Family of Utilities)*

$$\begin{aligned}
\textit{Find}\ \beta^* &= \arg\max_{\beta} h(\beta) \\
\textit{with}\quad h(\beta) &= \frac{1}{N}\sum_i \log q^{(\beta)}(y_i|x_i) - \sqrt{\frac{\alpha}{\gamma_1^2}\frac{\beta^T\Sigma\beta}{N}}\,, && (75)\\
\textit{where}\quad q^{(\beta)}(y|x) &= Z_x^{-1}\, e^{\beta^T f(x,y)} \times \begin{cases} q^0_{\rho|x} & \textit{if } y=y_\rho \textit{ for some } \rho \\ q^0(y|x) & \textit{otherwise}\,, \end{cases} && (76)\\
\textit{and}\ Z_x &= \int_{\mathcal{Y}} q^0(y|x)e^{\beta^T f(x,y)}dy + \sum_\rho q^0_{\rho|x}e^{\beta^T f(x,y_\rho)}\;, && (77)
\end{aligned}$$

where the (x_i, y_i) are the observed values and N is the number of observations. The measure on the efficient frontier is then

$$\begin{aligned}
q^* &= \{(q^*(y|x), q^*_{\rho|x}), y\in\mathcal{Y}, \rho=1,2,...,m, x=x_1,...,x_M\} \\
\text{with } q^*(y|x) &= q^{(\beta^*)}(y|x) \\
\text{and } q^*_{\rho|x} &= q^{(\beta^*)}(y_\rho|x)\;.
\end{aligned}$$

We note that we can obtain the same $\alpha-$parameterized family of solutions to Problem 10, if we allow α to vary over $[0,\infty)$, by dropping the square root in Equation 75; we show that this is so in Appendix E.

3.3.2 EXAMPLE: LOGISTIC REGRESSION

We note that in the special case where $U(W)$ is in our logarithmic family Equation 35, $\mathcal{Y}=\emptyset$, $m=2$, the prior is flat, and $\alpha=0$, the dual problem, Problem 10, is the logistic regression problem.

3.4 Summary of Modeling Approach

The logic of our modeling approach in this section's more general context is similar to the logic described in Section 2.4. We have the following procedure:

1. Break the data into a training set and a hold-out sample. (In their numerical experiments, Friedman and Sandow, 2003b, Friedman and Huang, 2003, and Sandow et al., 2003) used 75% or 80% of the data, selected randomly, to train the model.)
2. Choose a discrete set $A = \{\alpha_k \in (0, \alpha_{max}), k = 1, \ldots, K\}$.
3. For $k = 1, \ldots, K$,
   - Solve Problem 9 for $\beta^*(\alpha_k)$,
   - Compute the out-of-sample performance $P_k = E_{\tilde{p}}[U(b^*(q^*(\alpha_k)), O)]$ on the out-of-sample test set, where $\tilde{p}$ is the empirical measure on this test set, and q^*, is determined from Equations 65 and 66 with parameters $\beta^*(\alpha_k)$, and b^* is determined from Equations 63 and 64.
4. Put $k^* = \arg\max_k P_k$.
5. Our model, q^{**}, is determined from Equations 65 and 66 with parameters $\beta^*(\alpha_{k^*})$.

Acknowledgments

We thank Max Chickering, James Huang and an anonymous referee for their insightful comments.

Appendix A. Robustness of the Minimum Generalized Relative Entropy Measure

In this appendix, we state and prove the following theorem, which is only a slight modification of a result from Grünwald and Dawid (2002) and is based on the logic from Topsøe (1979).

Theorem 5

$$\arg\min_{q \in K} D_{U,O}(q||q^0) = \arg\max_{q \in Q} \min_{p' \in K} \left\{ E_{p'}\left[U(b^*(q), O)\right] - E_{p'}\left[U(b^*(q^0), O)\right] \right\} ,$$

where Q is the compact convex set of all possible probability measures, and $K \subset Q$ is compact and convex.

Interpretation: We can measure the quality of a model q by the gain in expected (under the true measure) utility experienced by an investor who invests optimally according to the model relative to an investor who invests optimally according to the benchmark model q^0 (see Friedman and Sandow, 2003a), i.e. by $E_{p'}\left[U(b^*(q), O) - U(b^*(q^0), O)\right]$, where p' is the *true* probability measure. If we use this performance measure, Theorem 5 states the following: minimizing $D_{U,O}(q||q^0)$ with respect to $q \in K$ is equivalent to searching for the measure $q^* \in Q$ that maximizes the worst-case (with respect to the potential true measures, $p' \in K$) relative model performance. The optimal model, q^*, is robust in the sense that for any other model, q, the worst (over potential true measures $p' \in K$) relative performance is even worse than the worst-case relative performance under q^*. We do not know the true measure; an investor who makes allocation decisions based on q^* is prepared for the worst that nature can offer.

Proof: We start with the definition, Equation 6, of the generalized relative entropy:

$$D_{U,O}(q||q^0) = E_q\left[U(b^*(q), O)\right] - E_q\left[U(b^*(q^0), O)\right] \ .$$

By the definition (Equation 2) of the optimal betting weights, b^*,

$$E_q\left[U(b^*(q), O)\right] \geq E_q\left[U(b^*(\pi), O)\right] \ , \tag{78}$$

for any measure $\pi \in Q$. Therefore, we have

$$D_{U,O}(q||q^0) = \max_{\pi\in Q}\left\{E_q\left[U(b^*(\pi), O)\right] - E_q\left[U(b^*(q^0), O)\right]\right\} \ ,$$

and

$$\min_{q\in K} D_{U,O}(q||q^0) = \min_{q\in K}\max_{\pi\in Q}\left\{E_q\left[U(b^*(\pi), O)\right] - E_q\left[U(b^*(q^0), O)\right]\right\} \ .$$

Since both K and Q are compact and convex, and $E_q\left[U(b^*(\pi), O) - U(b^*(q^0), O)\right]$ is a continuous concave-convex function on $K \times Q$, we can apply a minimax theorem (see, for example, Frenk et al., 2002, Theorem 3); we obtain

$$\min_{q\in K} D_{U,O}(q||q^0) = \max_{\pi\in Q}\min_{q\in K}\left\{E_q\left[U(b^*(\pi), O)\right] - E_q\left[U(b^*(q^0), O)\right]\right\} \ .$$

The maximum is attained for some pair, (π^*, q^*). From Equation 78, it follows that $\pi^* = q^*$. To see this, suppose that $\pi^* \neq q^*$; then, with q^* fixed, we can increase the value of the term

$$\left\{E_{q^*}\left[U(b^*(\pi^*), O)\right] - E_{q^*}\left[U(b^*(q^0), O)\right]\right\}$$

by setting $\pi^* = q^*$, which contradicts our assumption that the maximum is attained for the pair (π^*, q^*) with $\pi^* \neq q^*$. So

$$\arg\min_{q\in K} D_{U,O}(q||q^0) = \arg\max_{\pi\in Q}\min_{q\in K}\left\{E_q\left[U(b^*(\pi), O)\right] - E_q\left[U(b^*(q^0), O)\right]\right\} \ .$$

After renaming the optimization variables of the right-hand-side: q as p' and π as q, we obtain Theorem 5. $\square$.

We note that this result can be applied directly to the discrete probability case (see Section 2.2.5). For conditional density models, we restrict the admissible probability measures so that the maximum value of the probability density is less than some finite number, q_{max}, to insure the compactness of K in the preceding theorem. We can always do so without changing the Pareto optimal measure by choosing

$$q_{max} = a \cdot \max_{x,y} q^*(y|x),$$

where $a > 1$ is some constant and $q^*(y|x)$ is given in Equation 65. Note that any measure in Q^+ can be approximated by a measure in K for a sufficiently large, and that $q^*(y|x)$ is finite for all utilities.

Appendix B. Proof of Lemma 2

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

Proof: We note that the objective function, $D_{U,O}(q||q^0)$, is strictly convex (see Theorem 2 in Friedman and Sandow, 2003a). The inequality constraint, Inequality 19, of Problem 2 is also convex; this follows from the fact that Σ is a covariance matrix and therefore nonnegative definite. The equality constraints, Equations 18 and 20, are both affine. Therefore, Problem 2 is a strictly convex programming problem (see, for example, the Convex Programming Problem II Berkovitz, 2002).

We now show that Problems 1 and 2 have the same unique solution.

We first *assume that* $\alpha < \alpha_{max}$, and show that in this case, the solution to Problem 2 satisfies

$$Nc^T\Sigma^{-1}c = \alpha.$$

To this end, we note that $D_{U,O}(q||q^0)$ is strictly convex in q, for q in the simplex Q, and that the global minimum of the function $D_{U,O}(q||q^0)$ occurs at $q = q^0$ (see Friedman and Sandow, 2003a, Theorem 2 for these $D_{U,O}(q||q^0)$ properties), which occurs only if $\alpha = \alpha_{max}$; therefore,

$$\nabla_q D_{U,O}(q||q^0) \neq 0 \tag{79}$$

for $q \neq q^0$. Suppose that q^* is such that $Nc(q^*)^T\Sigma^{-1}c(q^*) < \alpha$ where

$$c(q) = E_q[f] - E_{\tilde{p}}[f].$$

Then there exists a neighborhood of q^* on the simplex Q, such that for all q in the neighborhood, $Nc(q)^T\Sigma^{-1}c(q) \leq \alpha$. From Equation 79, we see that there is a direction of decrease of the objective function $D_{U,O}(q||q^0)$ on the simplex Q, so q^* cannot be the optimal solution. Therefore, we cannot have $Nc(q^*)^T\Sigma^{-1}c(q^*) < \alpha$. It follows that $Nc^T\Sigma^{-1}c = \alpha$, so the solution to Problem 2 is the solution to 1 for the case $\alpha < \alpha_{max}$.

In the case $\alpha = \alpha_{max}$, it is obvious that both problems have the unique solution $q^* = q^0$.

The objective function, $D_{U,O}(q||q^0)$ is strictly convex in q, so the solution of Problem 2 is unique (see, for example, Rockafellar, 1970, Section 27). It follows that the solution to Problem 1 is also unique. $\square$

Appendix C. Proof of Theorem 1

We will show that Problem 2, which we restate below for convenience, has the (equivalent) dual formulations Problems 3 and 4.

Problem 2 *(Initial Convex Problem, Given* $\alpha, 0 \leq \alpha \leq \alpha_{max}$*)*

$$\begin{aligned}
\textit{Find} \quad & \min_{q\in(R^+)^m, c\in R^J} D_{U,O}(q||q^0) && (80)\\
\textit{under the constraints } 1 \quad &= \sum_y q_y && (81)\\
\textit{and } \ Nc^T\Sigma^{-1}c \quad &\leq \alpha && (82)\\
\textit{where } \ c_j \quad &= E_q[f_j] - E_{\tilde{p}}[f_j] \ . && (83)
\end{aligned}$$

We will derive the dual of Problem 2 now. To this end, note that the Lagrangian is given by

$$\begin{aligned}\mathcal{L}(q,c,\beta,\xi,\mu,\nu) &= D_{U,O}(q||q^0)+\beta^T\left\{c-E_q[f]+E_{\tilde{p}}[f]\right\}\\ &\quad +\xi\frac{1}{2}\left\{Nc^T\Sigma^{-1}c-\alpha\right\}+\mu\left\{\sum_y q_y-1\right\}\\ &\quad -\nu^T q, \end{aligned} \tag{84}$$

where $\xi \geq 0$, $\beta = (\beta_1,...,\beta_J)^T$, μ, and $\nu^T = (\nu_1,\ldots,\nu_m) \geq 0$ are Lagrange multipliers and q varies over $\mathbf{R}^m$.

C.1 The Connecting Equation

In order to derive the connecting equation, we have to solve

$$0 = \frac{\partial \mathcal{L}(q,c,\beta,\xi,\mu,\nu)}{\partial c_j} \tag{85}$$

$$\text{and } 0 = \frac{\partial \mathcal{L}(q,c,\beta,\xi,\mu,\nu)}{\partial q_y} \ . \tag{86}$$

The first one of these equations has solution

$$c = c^* \equiv -\frac{1}{\xi N}\Sigma\beta \ . \tag{87}$$

In order to solve Equation 86, we insert Equation 84 and the equation (see Lemma 2 from Friedman and Sandow, 2003a)

$$\frac{\partial D_{U,O}(q||q^0)}{\partial q_y} = U(b_y^*(q)O_y)-U(b_y^*(q^0)O_y) \ , \tag{88}$$

into Equation 86, and obtain

$$0 = U(b_y^*(q)O_y)-U(b_y^*(q^0)O_y) - \beta^T f(y) + \mu - \nu_y \ . \tag{89}$$

We rewrite this equation as

$$U(b_y^*(q)O_y) = G_y(q^0,\beta,\mu,\nu) \tag{90}$$

$$\text{with } G_y(q^0,\beta,\mu,\nu) = U(b_y^*(q^0)O_y) + \beta^T f(y) - \mu + \nu_y \ , \tag{91}$$

where $G_y(q^0,\beta,\mu,\nu)$ does not depend on q. In order to solve for q, we substitute Equation 3 into Equation 90, to obtain

$$U\left(U'^{-1}\left(\frac{\lambda}{q_y O_y}\right)\right) = G_y(q^0,\beta,\mu,\nu) \ . \tag{92}$$

Solving for q_y, we obtain the connecting equation

$$q_y^* \equiv \frac{\lambda}{O_y U'\left(U^{-1}(G_y(q^0,\beta,\mu,\nu))\right)} \ . \tag{93}$$

From Equation 93, by the positivity of the O_y and the fact the U is a monotone increasing function, we conclude that all of the q_y^* and λ have the same sign. We note, from Equation 81, that the q_y^* and λ must be positive. From the Karush-Kuhn-Tucker conditions, we must have $\nu_y q_y^* = 0$; it follows that $\nu_y^* = 0$ for all y. Accordingly, we may suppress the dependence of G and $\mathcal{L}$ on ν.

The connecting equation, Equation 93, depends on β, λ, and μ. We now show how to calculate λ and μ in terms of β. Solving Equation 92 for $U'^{-1}\left(\frac{\lambda}{q_y O_y}\right)$ and substituting into Equation 4, we obtain a condition for μ^*:

$$\sum_y \frac{1}{O_y} U^{-1}\left(G_y(q^0, \beta, \mu^*)\right) = 1 \,. \tag{94}$$

This equation is easy to solve numerically for μ^*, by the following lemma.

Lemma 3 *There exists a unique solution, μ^*, to Equation 94. The left hand side of Equation 94 is a strictly monotone decreasing function of μ^*.*

Proof: First, we note that since U is a strictly increasing function,

$$(U^{-1})' = \frac{1}{\frac{dU}{dW}} > 0,$$

so U^{-1} is a strictly increasing function and the left hand side of Equation 94 is a strictly decreasing function of μ^*.

Letting

$$\overline{\mu} = \max_y \beta^T f(y),$$

we see that

$$\beta^T f(y) - \overline{\mu} \leq 0 \text{ for all } y.$$

In this case, it follows from Equation 91 that

$$G_y(q^0, \beta, \overline{\mu}) \leq U(b_y^*(q^0) O_y) \text{ for all } y,$$

so, by the monotonicity of U^{-1},

$$\begin{aligned} \sum_y \frac{1}{O_y} U^{-1}\left(G_y(q^0, \beta, \overline{\mu})\right) &\leq \sum_y \frac{1}{O_y} U^{-1}\left(U(b_y^*(q^0) O_y)\right) \\ &= \sum_y b_y^*(q^0) = 1. \end{aligned} \tag{95}$$

Note that $G_y(q^0, \beta, \overline{\mu}) \in dom(U^{-1})$ for all y, by Equation 90. Similarly, by letting

$$\underline{\mu} = \min_y \beta^T f(y),$$

we can guarantee that

$$\sum_y \frac{1}{O_y} U^{-1}\left(G_y(q^0, \beta, \underline{\mu})\right) \geq 1.$$

By the Intermediate Value Theorem and the monotonicity and continuity of the left hand side of Equation 94, there exists a unique solution to Equation 94. □

We now show how to calculate λ in terms of β and μ^*. We insert Equation 93 into Equation 81, and obtain:

$$1 = \lambda \sum_y \frac{1}{O_y U'(U^{-1}(G_y(q^0, \beta, \mu^*)))} \ ;$$

solving for λ, we obtain

$$\lambda^* \equiv \left\{ \sum_y \frac{1}{O_y U'(U^{-1}(G_y(q^0, \beta, \mu^*)))} \right\}^{-1} \ . \tag{96}$$

Summarizing the result of this subsection:

The connecting equation, which describes q^* as a member of a parametric family (in β), is given by

$$q_y^* = \frac{\lambda^*}{O_y U'(U^{-1}(G_y(q^0, \beta, \mu^*)))}, \tag{97}$$

where we determine μ^* from Equation 94 via Lemma 3 and λ^* from Equation 96.

C.2 Dual Problems

We now show that

Lemma 4 *Problem 4 is the dual of Problem 2.*

Proof: Equations 87 and 93, together with the Equations 94 and 96, give the vector c^*, the probabilities q_y^* and the Lagrange multipliers μ^*, ν^* for which the Lagrangian is at its minimum for given multipliers β, ξ. This allows us to formulate the dual problem as an optimization with respect to β and ξ. To this end, we have to compute $\mathcal{L}(q^*, c^*, \beta, \xi, \mu^*)$. We insert Equations 6 and 87 into Equation 84), and obtain:

$$\begin{aligned}
\mathcal{L}(q^*, c^*, \beta, \xi, \mu^*) \quad = \quad & \sum_y q_y^* U(b_y^*(q^*) O_y) - \sum_y q_y^* U(b_y^*(q^0) O_y) \\
& + \beta^T \left\{ -\frac{1}{\xi N} \Sigma \beta - \sum_y q_y^* f(y) + E_{\tilde{p}}[f] \right\} \\
& + \xi \frac{1}{2} \left\{ N \frac{1}{\xi^2 N^2} \beta^T \Sigma \Sigma^{-1} \Sigma \beta - \alpha \right\} \\
& + \mu^* \left\{ \sum_y q_y^* - 1 \right\},
\end{aligned}$$

so

$$\begin{aligned} \mathcal{L}(q^*, c^*, \beta, \xi, \mu^*) &= \sum_y q_y^* \left\{ U(b_y^*(q^*)O_y) - U(b_y^*(q^0)O_y) - \beta^T f(y) + \mu^* \right\} \\ &\quad + \beta^T E_{\tilde{p}}[f] - \frac{1}{2\xi N}\beta^T \Sigma \beta - \frac{1}{2}\xi\alpha - \mu^* . \end{aligned}$$

Because of Equation 89, the first line on the r.h.s. of above equation is zero, i.e., we obtain

$$\mathcal{L}(q^*, c^*, \beta, \xi, \mu^*) = \beta^T E_{\tilde{p}}[f] - \frac{1}{2\xi N}\beta^T \Sigma \beta - \frac{1}{2}\xi\alpha - \mu^* .$$

The dual problem is to maximize the function $h(\beta) = \mathcal{L}(q^*, c^*, \beta, \xi, \mu^*)$ with respect to β, ξ. We can analytically maximize with respect to ξ, by solving $0 = \frac{\partial \mathcal{L}(q^*, c^*, \beta, \xi, \mu^*)}{\partial \xi}$ for ξ. The maximum is attained when

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

the Lagrangian at $\xi = \xi^*$ is given by

$$\mathcal{L}(q^*, c^*, \beta, \xi^*, \mu^*) = \beta^T E_{\tilde{p}}[f] - \mu^* - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}} . \tag{98}$$

Now we are ready to formulate the dual problem: maximize $h(\beta) = \mathcal{L}(q^*, c^*, \beta, \xi^*, \mu^*)$ with respect to β. From Equations 98, 91, 97, 96 and 94 we obtain Problem 4, which completes the proof of the equivalence of the solutions to Problems 4 and 2. □

In the following lemma, we show that we can express the dual problem objective function in a more easily interpreted form.

Lemma 5 *Problem 4 can be restated as Problem 3.*

Proof: Using Equation 89 to replace $\beta^T E_{\tilde{p}}[f] - \mu^*$ in Equation 30, and noticing that $U(b_y^*(q^0)O_y)$ does not depend on β, we obtain:

$$h(\beta) = \sum_y \tilde{p}_y U(b_y^*(q^*)O_y) - \sqrt{\alpha \frac{\beta^T \Sigma \beta}{N}} , \tag{99}$$

up to an unimportant constant. This means that the dual problem can be restated as in Problem 3. □

The proof of Theorem 1 is a direct consequence of Lemmas 4 and 5 and the fact that the primal problem satisfies the Slater condition and therefore there is no duality gap (see, for example Section V, Theorem 4.2 in Berkovitz, 2002). The primal problem is strictly convex and therefore has a unique solution (see, for example, Rockafellar, 1970, Section 27).

Appendix D. Dual Problem for our Logarithmic Family

In order to specify the dual problem for our utility (Equation 35), we first notice that

$$U'(z) = \frac{\gamma_1}{z + \gamma} , \tag{100}$$

$$U^{-1}(z) = e^{\frac{z-\gamma_2}{\gamma_1}} - \gamma \tag{101}$$

$$\text{and } U'(U^{-1}(z)) = \gamma_1 e^{-\frac{z-\gamma_2}{\gamma_1}} . \tag{102}$$

Using the relation

$$b_y^*(q) = q_y \left[1+\gamma\sum_{y'}\frac{1}{O_{y'}}\right] - \frac{\gamma}{O_y} \ , \tag{103}$$

(see Theorem 4 in Friedman and Sandow, 2003a), we can write $G_y(q^0,\beta,\mu^*)$ from Equation 26 as

$$G_y(q^0,\beta,\mu^*) \ = \gamma_1 \left[\log\left(q_y^0 O_y \left[1+\gamma\sum_{y'}\frac{1}{O_{y'}}\right]\right) \ + \ \beta^T f(y) \ - \ \mu^* \right] + \gamma_2 \ . \tag{104}$$

Inserting Equations 101 and 104 into Equation 27 gives

$$\begin{aligned} 1 \quad &= \quad \sum_y \frac{1}{O_y}\left\{ q_y^0 O_y \left[1+\gamma\sum_{y'}\frac{1}{O_{y'}}\right] e^{\beta^T f(y)-\mu^*} \ - \ \gamma \right\} &(105)\\ &= \quad e^{-\mu^*} \left[1+\gamma\sum_{y'}\frac{1}{O_{y'}}\right] \sum_y \left[q_y^0 e^{\beta^T f(y)} \right] \ - \ \gamma\sum_{y'}\frac{1}{O_{y'}} \ , &(106)\end{aligned}$$

which can be solved for μ^*:

$$\mu^* \ = \ \log\left(\sum_y q_y^0 e^{\beta^T f(y)}\right) \ . \tag{107}$$

Next we solve Equation 28 for λ^*. We use Equations 102 and 104 to write Equation 28 as

$$\begin{aligned} 1 \quad &= \quad \lambda^* \sum_y \left\{ \frac{1}{O_y}\left(q_y^0 O_y \left[1+\gamma\sum_{y'}\frac{1}{O_{y'}}\right] e^{\beta^T f(y)-\mu^*}\right)\right\} \\ &= \quad \lambda^* \left[1+\gamma\sum_{y'}\frac{1}{O_{y'}}\right] \sum_y \left\{ q_y^0 e^{\beta^T f(y)-\mu^*}\right\} \ .\end{aligned}$$

After inserting Equation 107 we can solve for λ^* and get:

$$\lambda^* \ = \ \frac{1}{1+\gamma\sum_{y'}\frac{1}{O_{y'}}} \ . \tag{108}$$

By means of Equations 25, 102, 108 and 104 we obtain for the optimal probability distribution

$$\begin{aligned} q_y^* \quad &= \quad \frac{1}{1+\gamma\sum_{y'}\frac{1}{O_{y'}}}\,\frac{1}{O_y}\left(q_y^0 O_y \left[1+\gamma\sum_{y'}\frac{1}{O_{y'}}\right] e^{\beta^T f(y)-\mu^*}\right) \\ &= \quad q_y^0 \ e^{\beta^T f(y)-\mu^*} &(109)\\ &= \quad \frac{1}{\sum_y q_y^0 e^{\beta^T f(y)}}\, q_y^0 \ e^{\beta^T f(y)} \quad \text{(by Equation 107)} \ . &(110)\end{aligned}$$

We can now compute the objective function $h(\beta)$. Based on Equations 23 and 109, we obtain

$$h(\beta) = \sum_y \tilde{p}_y \log q_y^* \ - \ \sqrt{\frac{\alpha}{\gamma_1^2}\frac{\beta^T\Sigma\beta}{N}} \ , \tag{111}$$

up to the constants $E_{\tilde{p}}[\log q_y^0]$ and γ_2 and the factor γ_1.

Collecting Equations 110 and 111, we obtain Problem 5.

Appendix E. Family-Equivalent Dual Problems

In this appendix, we show that we can construct problems that are family-equivalent to the dual optimization problem 4 (or 3, or 9), where we call two problems family-equivalent if they have the same family of solutions (as parameterized by α). We first state and prove the following lemma:

Lemma 6 *Let K and g be functions on $\mathbf{R}^J$; and let κ be a strictly monotone increasing function on the range of g. Furthermore, let K and $\kappa \circ g$ be concave. If, for some $\alpha \geq 0$, the function*

$$h(\beta) = K(\beta) + \alpha g(\beta)$$

has its maximum at $\beta = \beta^* \in \mathbf{R}^J$, then there exists an $\tilde{\alpha} \geq 0$ such that the function

$$\tilde{h}(\beta) = K(\beta) + \tilde{\alpha}\kappa(g(\beta)) \tag{112}$$

has its maximum at $\beta = \beta^*$ too. Moreover, if K, g and κ are differentiable, and $\nabla_\beta g(\beta)\big|_{\beta=\beta^*} \neq 0$, then

$$\tilde{\alpha} = \frac{\alpha}{\kappa'(g(\beta^*))} \; . \tag{113}$$

Proof: We first show the existence of $\tilde{\alpha}$. Since β^* maximizes the function $K + \alpha g$, it corresponds to a Pareto optimal value of the vector $-(K, g)$, i.e., if, for some β, $(K(\beta), g(\beta)) \neq (K(\beta^*), g(\beta^*))$, then either $K(\beta) < K(\beta^*)$ or $g(\beta) < g(\beta^*)$, or both (see, for example, the section on scalarization in Boyd and Vandenberghe, 2001). Because κ is a strictly monotone increasing, i.e. order-preserving, function, β^* also corresponds to a Pareto optimal value of the vector $-(K, \kappa \circ g)$. Since K and $\kappa \circ g$ are concave, the set $\mathcal{A}$ of achievable $-(K, \kappa \circ g)$ is convex. Therefore, there exists a nonnegative $\tilde{\alpha}$ such that β^* maximizes the function $\tilde{h} = K + \tilde{\alpha}\kappa \circ g$ from Equation 112 ($\tilde{\alpha}$ defines a tangent on $\mathcal{A}$, see, for example Section 2.6 in Boyd and Vandenberghe, 2001).

Next, we show that, if K, g and κ are differentiable and $\nabla_\beta g(\beta)\big|_{\beta=\beta^*} \neq 0$, then Equation 113 holds. To this end, we consider the first-order conditions for the maximization of h and $\tilde{h}$:

$$\begin{aligned} 0 &= \nabla_\beta K(\beta)\big|_{\beta=\beta^*} + \alpha \nabla_\beta g(\beta)\big|_{\beta=\beta^*} \\ \text{and} \;\; 0 &= \nabla_\beta K(\beta)\big|_{\beta=\beta^*} + \tilde{\alpha}\kappa'(g(\beta^*)) \nabla_\beta g(\beta)\big|_{\beta=\beta^*} \; . \end{aligned} \tag{114}$$

Comparing these two equations results in Equation 113. □

As a direct consequence of Lemma 6, we have the following lemma:

Lemma 7 *Let K and g be functions on $\mathbf{R}^J$; and let κ be a strictly monotone increasing function on the range of g. Furthermore, let K, g and $\kappa \circ g$ be concave. Then the α-parameterized (with $\alpha \geq 0$) family of maxima of*

$$h(\beta) = K(\beta) + \alpha g(\beta) \tag{115}$$

is the same as the $\tilde{\alpha}$-parameterized (with $\tilde{\alpha} \geq 0$) family of maxima of

$$\tilde{h}(\beta) = K(\beta) + \tilde{\alpha}\kappa(g(\beta)) \; . \tag{116}$$

Proof: It follows from Lemma 6 that for every member β^* of the family of solutions of Equation 115 (with $\alpha \geq 0$) there exists a value $\tilde{\alpha} \geq 0$ such that β^* is a solution of Equation 116, i.e. a member of the family of solutions of Equation 116. Next we define the function $\overline{g} = \kappa^{-1} \circ g$, which has the property $\kappa \circ \overline{g} = g$. Using again Lemma 6 with g replaced by $\overline{g}$, we can see that for every member β^* of the family of solutions of Equation 116 (with $\tilde{\alpha} \geq 0$) there exists a value $\alpha \geq 0$ such that β^* is a solution of Equation 115, i.e. a member of family of solutions of Equation 115. Therefore, Equations 115 and 116 have the same family of solutions. $\square$.

Lemma 7 allows us to construct family-equivalent problems to our dual optimization problems (such as Problems 3, 4 or 9) if we allow α to vary form zero to infinity. This follows from the fact that our dual optimization problems have the form given by Equation 115 with concave K and g. (The concavity of K and g is a consequence of the concavity of $h = K + \alpha g$ for all $\alpha \geq 0$, which, in its turn, follows from the fact that, for any $\alpha \geq 0$, h is the objective function of the dual of a convex optimization problem.) In order to construct a family-equivalent problem, we have to choose a strictly monotone increasing function κ such that $\kappa \circ g$ is concave. For example, by choosing $\kappa(x) = x^2$, we obtain a family-equivalent problem to our dual optimization problem that does not have the square root in the second term.

References

M. Avellaneda. Minimum-relative-entropy calibration of asset pricing models. *Int. J. Theor. and Appl. Fin.*, 1(4):447, 1998.

M. Avellaneda, C. Friedman, R. Holmes, and D. Samperi. Calibrating volatility surfaces via relative-entropy minimization. *Applied Math Finance*, 1997.

J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 1985.

J. Berkovitz. *Convexity and Optimization in R^n*. John Wiley & Sons, Inc., New York, 2002.

S. Boyd and L. Vandenberghe. *Convex Optimization*. `http://www.stanford.edu/~boyd/bv_cvxbook_12_02.pdf`, 2001.

K. Burnham and D. Anderson. *Model Selection and Multimodel Inference*. Springer, New York, 2002.

S. F. Chen and R. Rosenfeld. A gaussian prior for smoothing maximum entropy models. *Technical Report CMU-CS-99-108, School of Computer Science Carnegie Mellon University, Pittsburgh, PA*, 1999.

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

G. J. Daniell. Of maps and monkeys. In B. Buck and V. A. Macaulay, editors, *Maximum Entropy in Action*. Clarendon Press, Oxford, 1991.

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

D. Duffie. *Dynamic Asset Pricing Theory*. Princeton University Press, Princeton, 1996.

G. Frenk, G. Kassay, and J. Kolumbán. Equivalent results in minimax theory. *working paper*, 2002.

C. Friedman and J. Huang. Default probabilty modeling: A maximum expected utility approach. *working paper*, 2003.

C. Friedman and S. Sandow. Model performance measures for expected utility maximizing investors. *International Journal of Theoretical and Applied Finance*, 6(4):355, 2003a.

C. Friedman and S. Sandow. Recovery rates of defaulted debt: A maximum expected utility approach. forthcoming. *Risk*, 2003b.

M. Frittelli. The minimal entropy martingale measure and the valuation problem in incomplete markets. *Math. Finance*, 10:39, 2000.

A. Golan, G. Judge, and D. Miller. *Maximum Entropy Econometrics*. Wiley, New York, 1996.

P. Grünwald and A. Dawid. Game theory, maximum generalized entropy, minimum discrepancy, robust bayes and pythagoras. *Proceedings ITW*, 2002.

L. Gulko. The entropy theory of bond option pricing. *International Journal of Theoretical and Applied Finance*, 5:355, 2002.

S. F. Gull and G. J. Daniell. Image reconstruction with incomplete and noisy data. *Nature*, 272:686, 1978.

T Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

J. Huang. Personal communication. 2003.

E. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620, 1957.

E. Jaynes. On the rationale of maximum entropy methods. *Proc IEEE*, 70:939, 1982.

E. Jaynes. Monkeys, kangeroos, and *n*. *Maximum Entropy And Bayesian Methods in Applied Statistics: Proceedings of the Fourth Maximum Entropy Workshop, University of Calgery*, page 26, 1984.

S. Kullback. *Information Theory and Statistics*. Dover, New York, 1997.

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

D. Luenberger. *Optimization by Vector Space Methods*. Wiley, New York, 1969.

D. Luenberger. *Investment Science*. Oxford University Press, New York, 1998.

J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1944.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

D. Samperi. *Model Selection and Entropy in Derivative Security Pricing*. Ph.D. Thesis, New York University, New York, 1997.

S. Sandow, C. Friedman, M. Gold, and P. Chang. Economy-wide bond default rates: A maximum expected utility approach. *Working Paper*, 2003.

J. Skilling. Fundamentals of maxent in data analysis. In B. Buck and V. A. Macaulay, editors, *Maximum Entropy in Action*. Clarendon Press, Oxford, 1991.

F. Topsøe. Economy-wide bond default rates: A maximum expected utility approach. *Kybernetika*, 15:8, 1979.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1999.

N. Wu. *The Maximum Entropy Method*. Springer, New York, 1997.