Measuring Model Risk*

Thomas Breuer  Imre Csicszár

Abstract

We propose to interpret model risk as sensitivity of expected loss to changes in the risk factor distribution and to measure the model risk of a portfolio by the maximum expected loss over a set of plausible models defined in terms of relative entropy with respect to an estimated model. We give explicit formulas for the calculation of the model risk measure both in the generic case and in the pathological cases. We explicitly determine the worst case model from the set of plausible models.

Keywords: multiple priors, model risk, ambiguity aversion, stress tests, relative entropy, maximum entropy principle, exponential family

JEL classification: C18, C44, C60, G01, G32, M48
AMS classification: 62C20, 90B50, 91B30, 94A17

1 Introduction

In many decision situations the distinction between Knightian ambiguity and risk is prominent. Ambiguity is uncertainty about the risk factor distribution, whereas risk is uncertainty about which risk factor values from a well-specified distribution are realised. The Ellsberg Paradox and related evidence have shown such a distinction to be behaviourally meaningful. The standard framework of subjective expected utility, however, excludes such a distinction.

In spite of the clear conceptual and empirical distinction between risk and ambiguity, ambiguity frequently appears in the finance literature under the name of ‘model risk’. The name implicitly suggests that ambiguity is simply another type of risk. The blurring of ambiguity and risk is perhaps encouraged by an isomorphy between risk measures and the representation of preferences under ambiguity aversion: A widely used class of preferences

*Thomas Breuer, PPE Research Centre, FH Vorarlberg, thomas.breuer@fhv.at. Imre Csizsár, Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, csiszar@renyi.hu. We gratefully acknowledge comments by Klaus Böcker, Freddy Delbaen, Georg Pflug, Chris Rogers, Walter Schachermayer, Martin Summer, as well as to seminar participants in London, Munich, and Vienna.
allowing for ambiguity aversion are the multiple priors preferences axiomatised by Gilboa and Schmeidler [1989]. According to this theory, ambiguity averse agents prefer acts with lower values of

$$\sup_{Q \in \Gamma} E_Q(u \circ L),$$

where \( L \) is the loss function associated to the act, \( \Gamma \) is some closed convex set of finitely additive probabilities and \( u \) is a continuous strictly increasing utility function of the payoffs. The set \( \Gamma \) is interpreted as a set of priors held by the agent, and ambiguity is reflected by the multiplicity of the priors. On the other side of the isomorphy there is a representation of risk measures. Risk measures assign to a portfolio a number interpreted as risk capital. Artzner et al. [1999] and Föllmer and Schied [2004] formulated requirements for risk measures and coined the terms ‘coherent’ resp. ‘convex’ for risk measures fulfilling them. Every coherent risk measure \( \rho \) can be represented as

$$\rho(-L) = \sup_{Q \in \Gamma} E_Q(L)$$

for some closed convex set \( \Gamma \) of distributions (Delbaen [2002, Thm. 3.2]). Interpreting the choice of a portfolio as an act, the preference representation (1) and the risk measure representation (2) agree apart from the utility term in (1), which disappears when utility is linear or when it is absorbed into \( L \) (so that \( L \) describes the disutility of a portfolio in dependence of the risk factors.) A decision maker who ranks portfolios by lower values of some coherent risk measure is ambiguity averse. And vice versa: An ambiguity averse decision maker acts as if she were minimising some coherent risk measure.

By virtue of Delbaen’s representation (2) any coherent risk measure can be interpreted as measuring how sensitively the expected portfolio loss reacts to changes of the risk factor distribution \( Q \) within some set \( \Gamma \). We propose this sensitivity as a measure of model risk. The distributions \( Q \) in \( \Gamma \) can be interpreted as priors, as plausible risk factor distributions, or as plausible alternative models. This interpretation is valid for all sets \( \Gamma \), but it is more natural if \( \Gamma \) is clearly related to possible modelling or estimation errors. We propose to choose

$$\Gamma = \{ Q : D(Q \| \nu) \leq k \},$$

where \( D \) is the relative entropy, defined in Section 4, of a distribution \( Q \) with respect to some fixed distribution \( \nu \). (Synonyms for relative entropy are Kullback-Leibler distance or \( I \)-divergence.) \( \Gamma \) contains all risk factor

---

\(^1\)While Gilboa and Schmeidler [1989] work in the setting of Anscombe and Aumann [1963] using lottery acts, Casadesus-Masanell et al. [2000] translated their approach to Savage acts. We use the framework of Savage acts to describe the isomorphy between risk measurement and the representation of ambiguity averse preferences.
distributions $Q$ whose relative entropy $D(Q||\nu)$ with respect to some reference distribution $\nu$ is smaller than some radius $k > 0$. The reference distribution $\nu$ typically results from the choice of a model class and some estimation procedure based on historical data. $\nu$ is the best guess of the risk factor distribution, but on account of model risk other distributions $Q \in \Gamma$ are considered as plausible alternatives. In Section 2 we discuss in more detail the reasons for taking relative entropy as a measure of plausibility underlying our choice of the priors set $\Gamma$.

The main points of this paper are the following. (1) We propose to interpret model risk as sensitivity of expected loss to changes in the risk factor distribution and (2) to measure the model risk of a portfolio with loss function $L$ by

$$MR(L,k) := \sup_{Q : D(Q||\nu) \leq k} \mathbb{E}_Q(L).$$

(3) We give explicit formulas for the calculation of the model risk measure $MR$ both in the generic case (Theorem 1) and in the pathological cases (Theorem 2). (4) We explicitly determine the worst case model from the set $\Gamma$ of plausible models.

These results also have implications in other contexts. In the context of the Gilboa and Schmeidler [1989] theory in the Casadesus-Masanell et al. [2000] formulation, Theorems 1 and 2 provide explicit expressions for objective function of decision makers under ambiguity, in the special case that the priors set is given by (3). This is the basis for portfolio optimisation problems under ambiguity aversion.

In the context of stress testing, the alternative models (risk factor distributions) in $\Gamma$ can be interpreted as mixed scenarios satisfying the plausibility constraint $D(Q||\nu) \leq k$. The idea of taking stress scenarios to be distributions was pioneered by Kupiec [1998] and Berkowitz [2000]. In this context, the loss maximisation in (4) can be seen as a systematic stress testing procedure searching for a worst case scenario. Such a systematic procedure can overcome two criticisms raised against traditional stress tests with handpicked scenarios. First, scenarios picked by hand may be too implausible. In traditional stress tests the plausibility of scenarios is rarely quantified. Second, it remains unclear whether there are more severe scenarios of similar plausibility. If the scenarios considered are harmless, either because stress testers lack proficiency or wish to hide risks, stress tests convey a feeling of safety which might be false, see Berkowitz [2000].

A first step towards more objective stress tests was made by Studer [1997, 1999]. He proposed to perform stress tests systematically. Instead of considering just a few hand-picked scenarios Studer searches for the worst

---

2 The term ‘mixed’ scenarios is in analogy with game theory, which uses mixed strategies along with pure strategies, or with physics, which uses mixed states along with pure states. Sometimes the term generalised scenario is used for mixed scenarios, see Delbaen [2002].
scenario among a set of plausible pure scenarios. In this way one ensures that no plausible scenario is missed and that only scenarios of sufficient plausibility are considered. Studer searches for the worst case over the ellipsoid of pure scenarios whose Mahalanobis distance is smaller than some threshold. He quantifies the harm done in a pure scenario \( r \) by the loss \( L(r) \). MR as defined in (4) builds on a similar intuition but differs in important aspects: it uses mixed scenarios instead of pure scenarios, measures their plausibility by the relative entropy instead of the Mahalanobis distance, and quantifies their harm by the expected loss instead of the loss. This circumvents the drawbacks of Studer’s method pinpointed in Breuer [2008]. In this context, the contribution of the paper is an explicit solution to the systematic stress test procedure specifying the worst case scenario and the expected loss in the worst case scenario (Theorems 1 and 2).

In the context of risk measurement, the model risk measure MR defined in (4) is yet another coherent risk measure. (Coherency follows from the representation theorem of Delbaen [2002, Thm. 3.2].) It has important advantages over traditional risk measures, like Value at Risk or Expected Shortfall. Those assign risk numbers on the basis of the profit loss distribution, which arises from the portfolio when a risk factor distribution is given. For a fixed portfolio, a different risk factor distribution gives rise to a different profit loss distribution, and therefore often to a different risk capital requirement. Thus, traditional risk measures rely on a specific model, which perhaps uses inappropriate risk factors, or which works with a misspecified type of risk factor distribution, or with misestimated distribution parameters. MR measures exactly this model dependence. In addition, the risk measure MR is law-invariant (Proposition 1).

The rest of the paper is structured in the following way: In Section 2 we discuss the choice of relative entropy as a measure of model distance. In Section 3 we describe the intuition of the solution of the expected loss maximisation problem (4). This problem will turn out to be ‘dual’ to a maximum entropy problem. In Section 4 we formulate Theorems 1 and 2. The proofs are in Section 5.

2 Relative Entropy as a Measure of Model Plausibility

Model risk stems from the use of inappropriate risk factor distributions. A risk factor distribution may be inappropriate for two main reasons: parameter estimation errors and misspecification of the model.\(^3\) The model

---

\(^3\) Another way a model can be flawed is the choice of a portfolio loss function \( L \) which does not take into account all risk factors relevant for the portfolio value, or which specifies a wrong portfolio loss from the risk factor values. We do not deal with this type of model failure.
Figure 1: **Wrong risk factor distributions arising from estimation errors and/or model specification errors.** The model assumptions specify a family of distributions. The historical data determine by a parameter estimation procedure an estimated risk factor distribution $\nu$ from that family.

assumptions and the historical data determine by a parameter estimation procedure an estimated risk factor distribution $\nu$. But the true distribution differs from $\nu$ if the model assumptions are violated or if a parameter estimation error occurred. In case of parameter estimation errors the true distribution is not $\nu$ but some other distribution from the model family. In case of model misspecification the true distribution is not even in the family specified by the model. The situation is illustrated in Fig. 1.

A partial analysis of model risk often addresses parameter estimation errors but assumes the model to be well-specified. Parameter estimation errors may lead to a distribution differing from the true one in mean, correlations, volatilities, etc. Which range of distribution parameters is plausible enough to be considered in a model risk analysis? If the model class is an exponential family, the confidence regions are specified in terms of relative entropy. Many of the common distributions of statistical interest are of exponential type: normal, $\chi^2$, Poission, binomial, multinomial, negative binomial, etc. The maximum likelihood estimate of the parameter vector of the exponential family amounts to a minimisation of the relative entropy with respect to the sample. If the sample size $m$ is big enough for the central limit theorem to apply, the $(1 - \alpha)$-confidence regions of the parameter vector are asymptotically determined by the equation

$$D(O_m||\nu) \leq \chi^2(\alpha, h)/2, \tag{5}$$

where $\chi^2(\alpha, h)$ is the value for which the $\chi^2$-distribution with $h$ degrees of freedom yields $\text{Prob}(\chi^2 \geq \chi^2(\alpha, h)) = \alpha$ and $h$ is the number of dimensions of the parameter space (Kullback [1959, p. 102, eq. (5.11)]). The right hand side of (5) determines the value of $k$. $D(O_m||\nu)$ is the minimum dis-
crimination information between a population with a distribution from the exponential family, with parameter vector the same as in the sample, and a population with density $\nu$. The larger the value, the worse the resemblance between the sample and $\nu$. (For details and examples about the use of relative entropy in the definition of confidence regions for parameter estimates we refer to Kullback [1959, Chap. 5].)

Eq. (5) is a good reason to choose the set $\Gamma$ of plausible models in terms of relative entropy $D$. In Figure 1 the confidence region defined by (5) is represented by the intersection of the plane representing the model class and the ball $\{Q : D(Q||\nu) \leq k\}$.

A more comprehensive analysis of model risk allows not only for estimation errors but also for misspecifications of the model class. Our choice of $\Gamma$ includes distributions which are not from the specified model family, but still close enough to the estimated distribution $\nu$ (see Figure 1). We take closeness in the sense of relative entropy. In the literature, various distances of probability distributions are used. One family of such distances, the $f$-divergences of Csiszár [1963], Ali and Silvey [1966], and Csiszár [1967], correspond to convex functions $f$ on the positive numbers. Relative entropy corresponds to $f(t) = t \log t$, several other choices of $f$ also give distances often used in statistics. For details about $f$-divergences see Liese and Vajda [1987].

From the range of possible distances we have chosen relative entropy, which appears the most versatile one with many applications in statistics, information theory, statistical physics, see e.g. Kullback [1959], Csiszár and Körner [1981], Cover and Thomas [2006], Jaynes [1968, 1982]. Relative entropy has already been used in econometrics, see Golan et al. [1996] or Grechuk et al. [2009], or robust portfolio selection, see Calafiore [2007]. Using it also in the context of risk measurement looks certainly reasonable, though we do not claim that among the various distances of distributions this one is necessarily the best for this purpose.

In the context of inference the method of maximum entropy is distinguished by axiomatic considerations. Shore and Johnson [1980], Paris and Vencovská [1990], Jones and Byrne [1990] and Csiszár [1991] showed that it is the only method that satisfies certain intuitively desirable postulates. Still, as Uffink [1995, 1996] argued, relative entropy cannot be singled out as providing the only reasonable method of inference. Csiszár [1991] determined what alternatives come into account if some postulates are relaxed. Grunwald and Dawid [2004] argue that distances between distributions might be chosen in a utility dependent way. Relative entropy is natural only for decision makers with logarithmic utility. Picking up this idea, for decision

---

$^4$Distance is meant in a broad sense, requiring neither symmetry nor the triangle inequality; those properties of distances in the narrow sense do not hold even for relative entropy.
makers with non-logarithmic utility one might define the radius of the scenario set in terms of some utility dependent distance. But this is not the approach of this paper. In our framework utility may enter into the function $L$ (making it a disutility instead of a loss) but not into the scenario set.

3 Duality to the Maximum Entropy Problem

Observe that Problem (4) is ‘dual’ to a problem of maximum entropy inference. If an unknown distribution $Q$ had to be inferred when the available information specified only a feasible set of distributions, and a distribution $\nu$ were given as a prior guess of $Q$, the maximum entropy\(^5\) principle would suggest to infer the feasible distribution $Q$ which minimizes $D(Q||\nu)$. In particular, if the feasible distributions were those with $\mathbb{E}_Q(L) = b$, for a constant $b$, we would arrive at the problem

$$\sup_{Q: \mathbb{E}_Q(L) = b} D(Q||\nu).$$

(6)

Note that the objective function of the worst case problem (1) is the constraint in the maximum entropy problem (2), and vice versa (Fig. 2). It is therefore intuitively expected that (taking $k$ and $b$ suitably related) both problems are solved by the same distribution $\overline{Q}$,

$$\arg \sup_{Q: D(Q||\nu) \leq k} \mathbb{E}_Q(L) = \arg \sup_{Q: \mathbb{E}_Q(L) = b} D(Q||\nu) =: \overline{Q}. \quad (7)$$

The literature on the maximum entropy problem establishes that (under some regularity conditions) the solution $\overline{Q}$ is a member of the exponential family of distributions with statistic $L$, which have a $\nu$-density of the form $\exp(\theta L(r))$ times a normalisation factor. Call a typical distribution in the exponential family $Q(\theta)$.

For members of the exponential family the two relevant quantities $D(Q(\theta)||\nu)$ and $\mathbb{E}_{Q(\theta)}(L)$ can be expressed with the help of a function

$$\Lambda(\theta, L) := \log \left( \int e^{\theta L(r)} d\nu(r) \right), \quad (8)$$

where $\theta$ is a positive real number. If the loss function $L$ is clear from the context, we will simply write $\Lambda(\theta)$. The expected loss can be written as

$$\mathbb{E}_{Q(\theta)}(L) = \int L(r) \exp(\theta L(r) - \Lambda(\theta)) d\nu(r) = \Lambda'(\theta), \quad (9)$$

\(^5\)This name refers to the special case when the prior guess $\nu$ is the uniform distribution; then minimising $D(Q||\nu)$ is equivalent to maximising the entropy of $Q$. 

7
Figure 2: ‘Duality’ of Worst Case and Maximum Entropy

What is the objective function in the worst case problem (4) is the constraint in the maximum relative entropy problem (6), and vice versa.

(where Λ′(θ) is the derivative of Λ(θ, L) with respect to θ) and the relative entropy as

\[
D(Q(\theta)||\nu) = \int \log \frac{dQ(\theta)}{d\nu}(r)dQ(\theta)(r) = \int (\theta L(r) - \Lambda(\theta))dQ(\theta)(r)
\]

\[
= \theta E_Q(L) - \Lambda(\theta) = \theta \Lambda'(\theta) - \Lambda(\theta).
\]

(10)

If the identity (7) holds one can directly calculate MR(L, k):

\[
\sup_{Q: D(Q||\nu) \leq k} E_Q(L) = \sup_{\theta: \theta \Lambda'(\theta) - \Lambda(\theta) \leq k} \Lambda'(\theta) =: \Lambda'(\overline{\theta}),
\]

where \(\overline{\theta}\) is defined to be the solution of \(\overline{\theta} \Lambda'(\overline{\theta}) - \Lambda(\overline{\theta}) = k\). (The last equality follows from the convexity of \(\Lambda\).) This solution is illustrated in Fig. 3. MR is the slope of the tangent to the curve \(\Lambda(\theta)\) passing through \((0, -k)\). \(\overline{\theta}\) is the \(\theta\)-coordinate of the tangent point. From the figure it is obvious that \(\overline{\theta} \Lambda'(\overline{\theta}) - \Lambda(\overline{\theta}) = k\).

So far the intuition about the solution. It requires two important assumptions: Identity (7) should hold and the equation \(\overline{\theta} \Lambda'(\overline{\theta}) - \Lambda(\overline{\theta}) = k\) should have a (unique) solution \(\overline{\theta}\). The above argument makes it intuitively clear that MR can be calculated as \(\Lambda'(\overline{\theta})\). The mathematical contribution of this paper is to give precise conditions under which the two assumptions hold and the solution is indeed of the generic form above (Theorem 1). Furthermore we give the solution also for ‘pathological’ cases where the conditions for the generic solution do not hold (Theorem 2).
Figure 3: Calculation of MR from $\Lambda$. MR is the slope of the tangent to the curve $\Lambda(\theta)$ passing through $(0, -k)$. $\bar{\theta}$ is the $\theta$-coordinate of the tangent point.

4 Main Results

We use the following notation: The sample space often but not always is some subset of a Euclidean space. Its elements are simultaneous realisations of all risk factors. In the context of stress testing they would be interpreted as pure scenarios. We denoted them by $r$. The consequences of a pure scenario $r$ for the portfolio is described by a loss function $L(r)$. The loss function characterises the portfolio.\(^6\)

We now explicitly calculate the model risk measure (4) as well as the worst case model. The solution relies on techniques familiar in the theory of exponential families, see Barndorff-Nielsen [1978], and large deviations theory, see Dembo and Zeitouni [1998]. Still, a self-contained development is provided, full proofs are given in Section 5.

The relative entropy of a probability distributions $Q$ with respect to a reference distribution $\nu$ is defined as

$$D(Q\|\nu) := \begin{cases} \int \log \frac{dQ}{d\nu}(r)dQ(r) & \text{if } Q \ll \nu \\ +\infty & \text{if } Q \not\ll \nu \end{cases}$$

where $Q \ll \nu$ denotes absolute continuity of the distribution $Q$ with respect to the distribution $\nu$.

**Theorem 1.** (i) If $\text{ess sup}(L)$ is finite, assume $k$ is smaller than $k_{\text{max}} := -\log(\nu(\{r : L(r) = \text{ess sup}(L)\})$)

\(^6\)For stress testers it might seem a bold assumption to know explicitly the loss function $L(r)$. After all, stress testers often need days to evaluate a complex portfolio in a given scenario. On the other hand, all standard quantitative risk management frameworks do work with a loss function, see e.g. McNeil et al. [2005, Chpt 2.1].
(ii) Assume \( \theta_{\text{max}} := \sup\{\theta : \Lambda(\theta) < +\infty\} > 0 \),

(iii) If \( \theta_{\text{max}}, \Lambda(\theta_{\text{max}}), \) and \( \Lambda'(\theta_{\text{max}}) \) are all finite, assume \( k \) does not exceed
\[
k_{\text{max}} := \theta_{\text{max}} \Lambda'(\theta_{\text{max}}) - \Lambda(\theta_{\text{max}}).
\]
Under these assumptions the equation
\[
\theta \Lambda'(\theta) - \Lambda(\theta) = k
\]
has a unique positive solution \( \bar{\theta} \). The worst case scenario \( \bar{Q} \) is the distribution with \( \nu \)-density
\[
\frac{d\bar{Q}}{d\nu}(r) := \frac{e^{\bar{\theta}L(r)}}{\int e^{\bar{\theta}L(r)}d\nu(r)} = e^{\theta L(r) - \Lambda(\bar{\theta})}.
\]
The model risk of portfolio \( L \) equals
\[
\text{MR}(L, k) = \mathbb{E}_{\bar{Q}}(L) = \Lambda'(\bar{\theta}).
\]
This theorem gives conditions for the generic solution to apply and provides a practical procedure for calculating MR in the generic case (see Fig. 3).

1. Calculate \( \Lambda(\theta) \) from (8). This involves the evaluation of an \( n \)-dimensional integral.

2. Starting from the point \( (0, -k) \), lay a tangent to the curve \( \Lambda(\theta) \).

3. MR is given by the slope of the tangent.

How should one choose the radius \( k \)? \( k \) is a parameter in Problem (4), in the same way as the confidence level is a parameter for Value at Risk or Expected Shortfall. Which choice of \( k \) is sensible? MR(\( k \)) dominates Tail-VaR at the level \( \exp(-k) \):
\[
\sup_{A: \nu(A) \geq e^{-k}} \int A L(r)d\nu(r)/\nu(A) \leq \sup_{Q: D(Q||\nu) \leq k} \mathbb{E}_Q(L).
\]
(This is true because the distribution \( Q_A \) with density \( dQ_A/d\nu := 1_A/\nu(A) \) satisfies \( D(Q_A||\nu) \leq k \) if \( \nu(A) \geq \exp(-k) \).) This inequality suggests reasonable orders of magnitude for \( k \). For a 1%-tail the corresponding \( k \) is \(- \log(0.01) = 4.6 \). An alternative way to choose \( k \) would be to take \( k \)-values realised in historical crisis. In yet another way one could determine \( k \) from the right hand side of (5).

The second theorem deals with the pathological cases.

Theorem 2. (i) If ess sup(\( L \)) is finite, and \( k \geq k_{\text{max}} \), then MR(\( L, k \)) = ess sup(\( L \)).

(ii) If \( \theta_{\text{max}} = 0 \) then MR(\( L, k \)) = \( \infty \) for all \( k > 0 \).
(iii) If $0 < \theta_{\max} < +\infty$, and both $\Lambda(\theta_{\max})$ and $\Lambda'(\theta_{\max})$ are finite, and additionally $k > k_{\max}$, then

$$\text{MR}(L, k) = \frac{(k + \Lambda(\theta_{\max}))}{\theta_{\max}},$$

but there is no model achieving $\text{MR}(k)$: the supremum in eq. (4) is not a maximum.

The situation in the pathological cases (i) and (iii) is illustrated in Fig. 4.

![Figure 4: The pathological cases (i) and (iii).](image)

**Corollary 1.** MR is a law-invariant risk measure: If two portfolios $L_1, L_2$ have the same profit-loss distributions, $\nu \circ L_1^{-1} = \nu \circ L_2^{-1}$, then $\text{MR}(L_1, k) = \text{MR}(L_2, k)$.

## 5 Proofs

In the proof of Theorems 1 and 2 we will use the following properties of the function $\Lambda(\theta)$ defined by eq. (8), which are standard and easy to check.

The function $\Lambda(\theta)$ is convex and lower semicontinuous on $\mathbb{R}$, its value is 0 at $\theta = 0$, and—excluding the trivial case when $\Lambda(\theta) = +\infty$ at all $\theta \neq 0$—its essential domain $D_\Lambda := \{\theta : \Lambda(\theta) < +\infty\}$ is a finite or infinite interval. In this interval, $\Lambda(\theta)$ is continuous and has derivative given by (9): when $\theta \in D_\Lambda$ is an endpoint of this interval, the derivative $\Lambda'(\theta)$ is understood as one-sided and is not necessarily finite.

The derivative $\Lambda'(\theta)$ equals $E_{\nu}(L)$ at $\theta = 0$, see (9), and is strictly increasing in $D_\Lambda$ unless $L(r)$ is constant $\nu$-almost everywhere. Moreover, as $\theta$ goes (increasingly) to $\theta_{\max} = \sup\{\theta : \Lambda(\theta) < +\infty\}$, the limit of $\Lambda'(\theta)$ equals $\text{ess sup}(L)$ if $\theta_{\max} = +\infty$, while otherwise the limit equals $\Lambda'(\theta_{\max})$ or $+\infty$ according as $\theta_{\max}$ is in $D_\Lambda$ or not.
The function
\[ \Lambda^*(x) := \sup_{\theta} (\theta x - \Lambda(x)), \tag{15} \]
called the convex conjugate of \( \Lambda(x) \), is also convex and lower semicontinuous on \( \mathbb{R} \). Clearly,
\[ \Lambda^*(x) = \theta x - \Lambda(x) \text{ if } x = \Lambda'(\theta). \tag{16} \]
However, for some \( x \) perhaps no \( \theta \) satisfies \( x = \Lambda'(\theta) \).

**Lemma 1.** Excluding the cases (i) and (ii) of Theorem 2, there exists a unique \( x \) satisfying
\[ \Lambda^*(x) = k \text{ and } x > \mathbb{E}_{\nu}(L), \tag{17} \]
Except for the case (iii) of Theorem 2, this \( x \) equals \( \Lambda'(\theta) \) for some (unique) \( \theta > 0 \).

**Proof.** If \( \theta_{\max} \) is finite then \( \Lambda^*(x) \) is finite for all \( x > \Lambda'(0) = \mathbb{E}_{\nu}(L) \).
If \( \theta_{\max} = \infty \) then \( x \leq \lim_{\theta \to \infty} \Lambda'(\theta) \) is a necessary condition and \( \Lambda'(0) < x < \lim_{\theta \to \infty} \Lambda'(\theta) \) is a sufficient condition for \( \Lambda^*(x) < \infty \). In any case, \( \sup \{ x : \Lambda^*(x) < \infty \} \) is equal to \( \text{ess sup}(L) \).
Moreover, eqs. (16), (10) imply
\[ \Lambda^*(\Lambda'(0)) = 0 < D(\gamma \| \nu) = \Lambda^*(\Lambda'(\theta)) \]
if \( 0 < \gamma < \theta_{\max} \). Hence, since the function \( \Lambda^* \) is convex, it is strictly increasing in the interval \( [\Lambda'(0), \text{ess sup}(L)] \).
If \( \text{ess sup}(L) =: b < \infty \) then
\[ \Lambda^*(b) = \sup_{\theta} (\theta b - \Lambda(\theta)) = \sup_{\theta} \left[ -\log \int e^{\theta(L(r)-b)} d\nu(r) \right] = -\log \nu(\{ r : L(r) = b \}), \]
and \( \Lambda^*(x) \to \Lambda^*(b) \) as \( x \uparrow b \). It follows, since the case \( k \geq -\log \nu(\{ r : L(r) = b \}) \) has been excluded, that there exists a unique \( x \in (\Lambda'(0), b) \) satisfying \( \Lambda^*(x) = k \). Moreover, as \( \Lambda'(\theta) \) approaches \( b \) for \( \theta \to +\infty \), there exists \( \theta \in (0, +\infty) \) with \( x = \Lambda'(\theta) \).
If \( \text{ess sup}(L) = \infty \) then \( \Lambda^*(x) \) is a strictly increasing convex function in the interval \( [\Lambda'(0), \infty] \), hence it goes to \( +\infty \) as \( \theta \to \infty \). Thus, again, there exists a unique \( x \) satisfying (17). On account of (16), this \( x \) is equal to \( \Lambda'(\theta) \) for some positive \( \theta \) except for the case (iii) of Theorem 2.

5.1 Proof of Theorem 1

Equation (11) has a unique positive solution \( \overline{\theta} \) due to Lemma 1 and (16). For \( \overline{Q} \) as defined in (12) we have
\[ \mathbb{E}_{\overline{Q}}(L) = \Lambda'(\overline{\theta}) \tag{18} \]
because of (9) as well as \( D(\tilde{Q}\||\nu) = k \), see (10). We now show that this \( \tilde{Q} \) attains the maximum \( E_Q(L) \) among all mixed scenarios \( Q \) with \( D(Q\||\nu) \leq k \). Take an arbitrary such \( Q \). Then we have

\[
\bar{\theta}\Lambda'(\bar{\theta}) - \Lambda(\bar{\theta}) = k \\
\geq D(Q\||\nu) \\
= \int \log \left( \frac{dQ}{d\nu}(r) \right) dQ(r) \\
= \int \left( \log \frac{dQ}{d\nu}(r) + \log \frac{\tilde{Q}(r)}{\nu}(r) \right) dQ(r) \\
= D(Q\||\tilde{Q}) + \int (\bar{\theta}L(r) - \Lambda(\bar{\theta}))dQ(r) \\
= D(Q\||\tilde{Q}) + \bar{\theta}E_Q(L) - \Lambda(\bar{\theta}),
\]

implying \( \bar{\theta}E_Q(L) \leq \bar{\theta}\Lambda'(\bar{\theta}) = \bar{\theta}E_{\tilde{Q}}(L) \). Since \( \bar{\theta} \) is positive we conclude \( E_Q(L) \leq E_{\tilde{Q}}(L) \).

5.2 Proof of Theorem 2

Proof. (i) First consider the case where \( \text{ess sup}(L) =: b \) is finite, and \( 0 < \nu(\{r : L(r) = b\}) =: \beta \). Then the measure \( \nu_b \ll \nu \) with

\[
\frac{d\nu_b}{d\nu}(r) := \begin{cases} 
1/\beta & \text{if } L(r) = b \\
0 & \text{otherwise}
\end{cases}
\]

satisfies

\[
D(\nu_b\||\nu) = \int \log \left( \frac{d\nu_b}{d\nu} \right) d\nu_b = -\log \beta,
\]

hence \( D(\nu_b\||\nu) \leq k \) if \( k \geq -\log \beta \). Then \( \text{MR}(k) \geq E_{\nu_b}(L) = b \). Trivially \( \text{MR}(k) \leq \text{ess sup}(L) = b \). The claim \( \text{MR}(k) = b \) follows.

(ii) Next consider the case \( \theta_{\max} = 0 \). Let \( \beta_{m,n} := \nu(\{r : -m \leq L(r) \leq n\}) \) and consider the measures \( \nu_{m,n} \ll \nu \) with

\[
\frac{d\nu_{m,n}}{d\nu}(r) := \begin{cases} 
1/\beta_{m,n} & \text{if } -m \leq L(r) \leq n \\
0 & \text{otherwise}
\end{cases}
\]

For any \( Q \ll \nu_{m,n} \),

\[
D(Q\||\nu) = \int \log \left( \frac{dQ}{d\nu_{m,n}} \right) d\nu_{m,n} = D(Q\||\nu_{m,n}) - \log \beta_{m,n}
\]

is arbitrarily close to \( D(Q\||\nu_{m,n}) \) if \( m \) and \( n \) are sufficiently large. Hence to prove that \( \text{MR}(k) = +\infty \) for all \( k > 0 \), it suffices to find to any given \( m \) and sufficiently large \( n \) distributions \( Q \ll \nu_{m,n} \) with \( D(Q\||\nu_{m,n}) \) arbitrarily close to zero and \( E_Q(L) \) arbitrarily large.
In the rest of part (ii) of this proof, $m$ is fixed and $n$ will go to $+\infty$. Define $Q$ and $\Lambda_{m,n}$ by

$$\frac{dQ}{d\nu_{m,n}}(r) := \frac{e^{\theta L(r)}}{\int e^{\theta L(r)}d\nu_{m,n}(r)} =: e^{\theta L(r) - \Lambda_{m,n}(\theta)}$$

for any $\theta > 0$. $Q$ and $\Lambda_{m,n}$ depend on $\theta$. As in the proof of Theorem 1, $E_Q(L) = \Lambda'_{m,n}(\theta)$ and $D(Q||\nu_{m,n}) = \theta \Lambda'_{m,n}(\theta) - \Lambda_{m,n}(\theta)$ for any $\theta > 0$. For each $\theta$,

$$-\theta m \leq \Lambda_{m,n}(\theta) = \int_0^\theta \Lambda'_{m,n}(\xi)d\xi \leq \theta \Lambda'_{m,n}(\theta). \quad (19)$$

For fixed $\theta > 0$, $\Lambda_{m,n} \to \infty$ as $n \to \infty$ since $\Lambda(\theta) = \infty$ by assumption. By (19) it follows that $\Lambda'_{m,n}(\theta) \to \infty$ as $n \to \infty$, and hence there exists a sequence $\theta_n \downarrow 0$ such that $\Lambda'_{m,n}(\theta_n) \to \infty$ and $\theta_n \Lambda'_{m,n}(\theta_n) \to 0$ as $n \to \infty$. By inequality (19), this implies $|\Lambda_{m,n}(\theta_n)| \to 0$ and hence $D(Q||\nu_{m,n}) \to 0$ as $n \to \infty$. This completes the proof that, for $Q$ defined with $\theta = \theta_n$, $E_Q(L)$ will be arbitrarily large and $D(Q||\nu_{m,n})$ arbitrarily small.

(iii) In case (iii) $0 < \theta_{\text{max}} < +\infty$, and both $\Lambda(\theta_{\text{max}})$ and $\Lambda'(\theta_{\text{max}})$ are finite, and additionally $k > k_{\text{max}} = \theta_{\text{max}} \Lambda'(\theta_{\text{max}}) - \Lambda(\theta_{\text{max}})$. Define $\overline{Q}$ as in (12) but with $\theta_{\text{max}}$ in the place of $\theta$. Then for all $Q \ll \overline{Q}$ we have

$$D(Q||\nu) = \int \log\left(\frac{dQ}{d\overline{Q}}\right) dQ = \int \log\left(\exp(\theta_{\text{max}} L(r) - \Lambda(\theta_{\text{max}}))\right) dQ(r) = \int D(Q||\overline{Q}) + \theta_{\text{max}} E_Q(L) - \Lambda(\theta_{\text{max}}). \quad (20)$$

Hence, if $D(Q||\nu) \leq k$ then

$$E_Q(L) \leq \left(k + \Lambda(\theta_{\text{max}}) - D(Q||\overline{Q})\right)/\theta_{\text{max}}, \quad (21)$$

proving that $\text{MR}(k) \leq \left(k + \Lambda(\theta_{\text{max}})\right)/\theta_{\text{max}}$. To show that equality holds, apply the result of (ii) to $\overline{Q}$ in the role of $\nu$, then the role of $\Lambda(\theta)$ is played by

$$\overline{\Lambda}(\theta) := \log \int e^{\theta L(r)} d\overline{Q}(r) = \Lambda(\theta + \theta_{\text{max}}) - \Lambda(\theta_{\text{max}}).$$

Clearly, $\overline{\Lambda}(\theta) = \infty$ for all $\theta > 0$, hence by the result of (ii) there exist distributions $\overline{Q}'$ with $D(\overline{Q}'||\overline{Q})$ arbitrarily small and $E_{\overline{Q}'}(L)$ arbitrarily large. Then, for any small $\epsilon > 0$, a suitable linear combination $Q$ of $Q'$ and $\overline{Q}$ satisfies $E_Q(L) = \left(k + \Lambda(\theta_{\text{max}}) - \epsilon\right)/\theta_{\text{max}}$ and $D(Q||\overline{Q}) < \epsilon$. For this $Q$, eq. (20) implies that $D(Q||\nu) \leq k$ and the claim $\text{MR}(k) \leq \left(k + \Lambda(\theta_{\text{max}})\right)/\theta_{\text{max}}$ follows. It implies that $\text{MR}(k) = \left(k + \Lambda(\theta_{\text{max}})\right)/\theta_{\text{max}}$.

Eq. (21) implies that in eq. (4) the supremum is not attained because $D(Q||\overline{Q})$ is strictly positive when $E_Q(L) > \Lambda'(\theta_{\text{max}}) = E_{\overline{Q}}(L)$. \qed
Remark: By Theorem 1, under its hypotheses $\text{MR}(k)$ is equal to $x = \Lambda'(\theta)$ for $\theta$ satisfying (11), which $x$ is the unique solution of (17). The last proof implies that MR always equals the solution of (17) when it exists, even if (11) does not have a solution.

5.3 Proof of Corollary 1

Proof. Assume two portfolios $L_1, L_2$ have the same profit-loss distributions, $\nu \circ L_1^{-1} = \nu \circ L_2^{-1}$, which we denote by $\mu_1, \mu_2$. Then

$$
\Lambda(\theta, L_1) = \log \left( \int_{\Omega} e^{\theta L_1(r)} d\nu(r) \right) = \log \left( \int_{\mathbb{R}} e^{\theta s} d\mu_1(s) \right) = \log \left( \int_{\mathbb{R}} e^{\theta s} d\mu_2(s) \right) = \Lambda(\theta, L_2).
$$

This holds for all $\theta$ for which $\Lambda$ is defined. The equality of distributions also implies that $\text{ess sup} L_1 = \text{ess sup} L_2$. Since the $\Lambda$-function of $L_1, L_2$ agree, their derivatives also agree, and by Theorems 1 and 2 their MR is equal. □

6 Conclusion

Theorem 1 suggests a close analogy between our model risk analysis and statistical mechanics. First, as pointed out above, the optimisation problem of finding the worst case model is ‘dual’ to the method of maximum entropy. Second, almost all quantities in the model risk analysis have counterparts in statistical mechanics: The worst case model (12) is the counterpart of the canonical distribution. $\theta$ is the counterpart of the temperature parameter $\beta = 1/kT$. The profit function $-L$ is the counterpart of the energy function $E$. The risk factor vector $r$ is the counterpart of the phase space points $(p, q)$. $\Lambda$ is the counterpart of the logarithm of the partition function $Z$.

References


