The Most Dangerous Model:
A Natural Benchmark for Assessing Model Risk

John A. Major, ASA
Ruodu Wang, Ph.D.
Micah G. Woolstenhulme, FCAS[1]

We examine the problem of decision making using a probabilistic model when there is material uncertainty concerning the accuracy of the model coupled with limited information about it. Such conditions could hold, for example, for the user of a complex commercial model of natural catastrophe insurance risk. Working within an ambiguity-averse decision framework, we define bounds for a set of plausible alternative models, centered on the "baseline" model provided to the user. Three types of bounds are defined, reflecting the model user's assumptions about the unknown and inaccessible data to which the baseline model was fit. Given a utility function for a decision option and a bound, we first address the corresponding optimization problem of finding the "worst" (most adverse expected utility) model within the set of plausible models. Second, we construct posterior mean utilities among the unbounded set of alternatives and show the existence of a posterior utility-minimizing worst credible model, i.e. the "most dangerous model." Among all alternative models to the baseline, this model has the highest product of expected disutility times probability that it, and not the baseline, is the correct model. We present a case study of how the most dangerous model can be used as a naturally occurring benchmark when making decisions in the presence of model risk.

Keywords: ambiguity aversion, robust control, model risk, Gilboa-Schmeidler, model uncertainty

JEL classification: D81, C44, C61, G22, G32, Q54

1 Introduction

1.1 Motivation and approach

As the use of mathematical models in the financial services industry increases, the concept of "model risk" is gaining attention. While it

[1] john.a.major@guycarp.com (corresponding author), wang@uwaterloo.ca, micah.g.woolstenhulme@guycarp.com

encompasses multiple dimensions [Derman 1996], a key component is "model uncertainty" or "distribution model risk."

Uncertainty inherent in commercial natural catastrophe ("cat") model outputs is widely recognized within the insurance industry and has long been a prominent topic at modeling conferences. In a 1997 conference in Bermuda, model vendor representatives were pressed for a quantitative measure of uncertainty around 100-year return periods (99% Value at Risk quantiles); no modeler was willing or able to provide one. Moore [1998] and Miller [1999] derived rough estimates of uncertainty in some situations. Since that time there has been some advance in users' understanding and vendors' reporting of uncertainty, but the cat model user still cannot get the type of uncertainty assessment that is routinely provided in statistical analysis [Major, 2011].

Motivated by the concerns of users of commercial cat models, we examine the problem of assessing model risk under conditions of very limited information. Our methods and conclusions are not limited to cat models however; these concerns are shared by users of other types of complex probabilistic models as well.

In this paper, we identify models with probability distributions. The intent is to address stochastic models whose primary outputs are functionals of distributions (moments, quantiles, etc.), and, often, representations (simulated samples) of the distributions themselves. Regardless of the mathematical form or computational implementation of a stochastic model, its essence lies in the output distribution. For this reason, we feel free to ignore the particular structure and operation of a model and concern ourselves solely with its output distribution.

Gilboa and Schmeidler [1989] established that a set of preferences over "acts" that satisfied certain properties, including uncertainty aversion, could be represented by a maximin utility criterion.

We start with a simplified decision framework that resembles that of Gilboa and Schmeidler's and later use their framework as designed. In the simplified framework, the user specifies a set of alternative models Q surrounding a "baseline" model p, and chooses from a set of decision options A on the basis of a utility criterion:

$$\arg\max_{a \in A} \min_{q \in Q} E_q\left[U_a(\omega)\right] \qquad [1]$$

where $a \in A$ is the decision (action) to be taken, $\omega \in \Omega$ is the random state of nature realized after the decision is made, U_a is the utility function corresponding to the decision, and expectation is taken with respect to model q. This formulation assumes that actions do not affect the probabilities of states of nature. The set Q is a proper subset of all possible models. We will call Q the set of "plausible" models and define it in terms of one of three statistical detection criteria: the *likelihood ratio* between q and p, the Kullback-Leibler [1951] *information gain*, and the *significance level* of the log likelihood statistic.

Our first task is to locate q^*, the "worst plausible model" in Q. "Worst" means that the user's utility function as described by U is minimized. We fix an $a \in A$ and compute:

$$q^* = \arg\min_{q \in Q} E_q\left[U_a(\omega)\right]. \qquad [2]$$

Solution procedures are presented for all three detection criteria.

We then introduce priors (Gilboa-Schmeidler as designed) to develop Bayesian posterior mean models combining the baseline and alternative model, in a manner reminiscent of credibility theory. We show the existence of and construct "worst credible" models that minimize the posterior expected utility. Whereas identification of the worst plausible model requires selection of a detection threshold, identifying the worst credible model does not. It consolidates both the likelihood and impact of the true model differing from the baseline model and stands out as an important natural benchmark for assessing model risk and making robust decisions. We therefore suggest that the worst credible model is indeed the "most dangerous model" to which a user should pay close attention.

While the hypothetical user does not have access to the data nor fitting procedures underlying the baseline model, we nonetheless need to make some assumptions about what the user knows and believes about potential alternative models relative to the baseline. We argue that these are reasonable in light of the goals and general lack of knowledge on the part of the user.

The specific contributions of this paper include:

- providing actuaries with an accessible introduction to ambiguity averse model risk assessment,
- proposing different bounding criteria for "worst case" construction that have more natural and intuitive interpretations than the information gain metric popular in the literature, and
- presenting a natural benchmark "worst credible case" that does not require selection of a plausibility threshold nor penalty coefficient, as is common in the literature.

1.2 Background on robust control and ambiguity aversion

The theory of robust control, developing in the late 1970s and early 1980s, aims to design control systems that perform well despite their design assumptions not holding exactly or at all times [Zhou & Doyle 1997]. Robust control typically borrows the maximin principle from game theory [Osborne & Rubinstein 1994]. Decision theorists have been aware of problems that classical expected-utility decision theory has in dealing with ambiguity (Knightian uncertainty) since the Ellsberg [1961] Paradox. Actuaries are more familiar with the concept of probabilistic ambiguity under the name of parameter risk [Venter & Sahasrabuddhe 2012].

Robust control concepts migrated to economics in the 1980s and 1990s, with axiomatic characterizations of "ambiguity aversion" by Gilboa and Schmeidler [1989], Epstein and Wang [1994], Anderson et al. [2000], Hansen and Sargent [2008], and others. Ambiguity aversion causes the decision maker to doubt the currently accepted probability model, and to consider alternatives, making decisions that are robust with respect to model misspecification.

In the Gilboa-Schmeidler framework, the utility maximization problem considers a bounded set of alternative models. In the Hansen-Sargent framework, the set of alternative models is unbounded, but there is a "penalty" function that compensates utility as the alternative models get "farther away" from the baseline model. When Gilboa-Schmeidler bounds are defined in terms of Kullback-Leibler information gain, a Hansen-Sargent penalty function naturally arises from the Lagrange multiplier formulation of the bounded optimization problem. Thus the two frameworks are intimately connected.

Friedman [2002] addresses the problem of extracting the worst model given a bound on information gain. Breuer & Csiszár [2013] additionally solve for bounds defined by Bregman distance and *f*-divergences, which include information gain and likelihood as special cases.

Goldfarb & Iyengar [2003], Garlappi, Uppal, and Wang [2007], and other authors apply these frameworks to portfolio selection. Barillas, Hansen, and Sargent [2009] find that ambiguity aversion provides a promising explanation for the equity premium puzzle [Mehra & Prescott 1985].

More specific to insurance, Kunreuther et al. [1993] present evidence that insurers are ambiguity averse. Zhao and Zhu [2011] develop an insurance pricing framework that incorporates both risk aversion and ambiguity aversion. Venter et al.'s [2004] work in applying Moeller's [2004] maximum entropy martingale measure to reinsurance pricing has mathematical connections with this work as well. Föllmer & Schied [2002] articulate the connection between "worst model" methods and coherent risk measures.

Some of the ideas in this paper have been presented in other venues [Major and Woolstenhulme 2011; Woolstenhulme and Major 2011; Major 2012, 2014;].

1.3 Organization of the paper

The remainder of the paper is organized as follows. Section 2 presents a "model of the model user" and lays out three key assumptions about what the user knows and believes. It also introduces a simple example that will motivate and illustrate the theory in subsequent sections. Section 3 defines the three measures of plausibility and solves for the worst plausible model under each definition. Section 4 introduces priors and solves for the worst credible ("most dangerous") models. Section 5 with appendix G applies these constructs in a case study. Section 6 concludes. Appendix H is an index of symbols. Table 1 organizes the principal results of sections 3 and 4.

Table 1: Guide to principal results

		Worst Plausible Model	Worst Credible Model
Measure of Plausibility	Likelihood Ratio	§3.2 Existence and uniqueness of q^*, one-dimensional root search. Theorem 1, appendix A.	§4.3 Existence of $q^{\varnothing}$, one-dimensional minimum search. Theorem 6, appendix E.
	Information Gain	§3.3 Existence and uniqueness of q^*, one-dimensional root search. Theorem 2, appendix B.	§4.4 Unable to formulate a meaningful definition of $q^{\varnothing}$.
	Statistical Significance	§3.4 Existence of q^*, characterization, multi-dimensional minimum search. Theorems 4, 5; appendices C, D.	§4.5 Existence (if simplified) of $q^{\varnothing}$, multi-dimensional minimum search. Theorem 7, appendix F.

2 A Model of the Model User

We model the decision maker / model user as being in possession of a "baseline" model p consisting of probabilities for a finite set of states of the world at some future date. Specifically, p is an element in the simplex $\mathcal{S}$ = $\{(p_0, p_1,\ldots, p_{D-1})\in\Re^D$: all $p_i>0$ and $\Sigma p_i=1\}$. This represents a probability distribution for the categorical random variable ω taking values in $\Omega = \{0, 1, 2, \ldots, D\text{-}1\}$. While the theory could be developed for a more general probability space, this should suffice for practitioners. Also, the application of these ideas in a parametric setting is fairly straightforward.

The user also has utility functions $U_a(\omega)$ for various possible actions $a \in A$ to be taken (or decisions to be made). A utility function is simply a vector of length D representing the utility to the user of each outcome ω. For an insurer, this might be the estimated (negative) dollar value of losses and expenses resulting from a catastrophe of the indicated category. We will further assume that the values of this vector are unique (no repeats). This can be done without loss of generality by combining categories that have the same utility.

The user is concerned that the baseline model might be materially wrong; that utility-maximizing decisions being made, had they been based on the "correct" model, would be materially different. Yet, she has no meaningful access to the underlying historical data upon which the model is based nor the model-building methodology that was used. Not being privy to these, she is unable to apply the statistician's usual methods for assaying uncertainty, constructing confidence intervals or posterior distributions around results, etc. There is no possibility of examining the goodness of fit of the model to the data. She cannot see a likelihood function nor a Fisher information matrix. There is no possibility of "bootstrapping" [Efron 1982] the data to refit the model.

Understanding the extent of uncertainty around the baseline model is important because the user is *ambiguity averse*. She prefers to take actions that may sacrifice expected utility as implied by p, but perform well across a range of elements in $\mathcal{S}$.

However, the user is not completely in the dark. We assume she has the following knowledge and beliefs:

> **(1) SAMPLE SIZE:** The baseline model p was fitted (competently) to n independent and identically distributed observations that emerged from the same process that the future state of the world will follow;
>
> **(2) PREFERRED LIKELIHOOD:** Given a proposed alternative model q, the user believes that the likelihood of model p on the fitted data is greater than the likelihood of model q;

(3) PRIOR SYMMETRY: Given a proposed alternative model q, the user believes that prior to the availability of data, p and q are equally probable.

The sample size assumption is relatively uncontroversial, being a necessary abstraction of the complexity of real-world model building. Preferred likelihood and prior symmetry require a bit more explanation.

If the preferred likelihood assumption were simultaneously applied to all possible alternative models, then the logical conclusion would be that p is the maximum likelihood model among all possible models, that is, it is the empirical probability function of the underlying data. This assumption, however, may be too strong to represent the reality of professional probabilistic modelling which often synthesizes both statistical and structural elements.

Model-based decisions should recognize that an alternative model may in fact better represent the true process, and if so, the expected utility implied by decisions may be quite different. The user is concerned with alternative models that are both *plausible* and *material*. Models that are close to the baseline are not of concern because they have little impact on expected utility. There is a neighborhood around the baseline model consisting of alternatives which are therefore not useful in the decision-making process. The preferred likelihood assumption is not required to hold for those alternatives.

At the other extreme, models that are *very* different from the baseline are implausible, and therefore can be ignored. That leaves the user with a "shell" of alternative models – not too close and not too distant – that are of concern. See figure 1. At this point in the exposition, such a shell is only a vague notion. Below, we will make it precise.

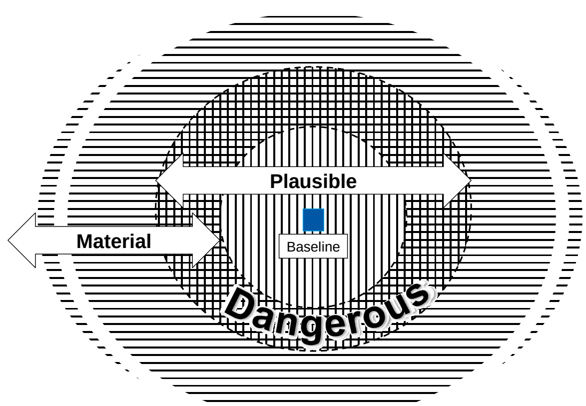

Figure 1: Dangerous models are materially different yet plausible.

Prior symmetry is another concept that should not be pushed too far. For example, one might argue that if p and q are equally probable, and p and q' are equally probable, then q and q' must be equally probable as well. Therefore all models (or at least all of those in the shell) must be equally probable and, if we can agree on a suitable measure for the space of models, the user must believe in a "flat" prior over all relevant models.

This is not the intent of prior symmetry. Prior symmetry is intended to model the user's beliefs only about pairwise comparisons and only when one of the pairs is the baseline model. We will show below that the use made of this assumption is equivalent to the user believing in a particular *class* of priors in the Gilboa-Schmeidler decision framework.

In order to illustrate the theory, we will work with a very simple example with D=3. Suppose natural catastrophe experience during a year is either "mild" ($\omega = 0$), "moderate" ($\omega = 1$), or "severe" ($\omega = 2$). Say the baseline model puts the probabilities at p = (0.88, 0.10, 0.02). Figure 2 presents this distribution as a histogram.

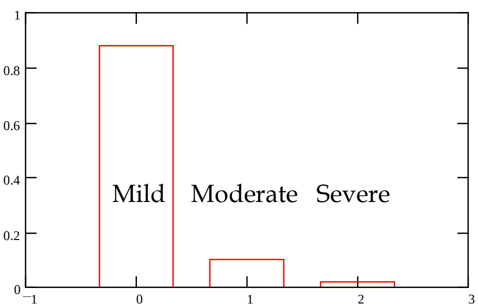

Figure 2: Histogram view of baseline model p.

Because there are only three components, we may represent the model as a point in the two-dimensional simplex, and we may choose to view the space of components 1 and 2 (moderate and severe), leaving component 0 (mild) unstated because it must carry all of the complementary probability. Figure 3 presents model p in this view.

The user is presented with a proposed alternative model q^0 = (0.81, 0.15, 0.04). This alternative implies that severe experience is twice as likely, and moderate experience is 50% more likely, than what model p supposes. Clearly this is a materially different model. The user wonders whether the alternative model is plausible. Might the vendor have "got it wrong" in developing model p? The user assumes that n = 100 observations were used to fit model p.

The user is concerned with other alternative models as well, and uses the utility function U = (0,-1,-10) that assigns \$0 loss to $\omega = 0$ "mild" experience, \$1 to $\omega = 1$ "moderate" experience, and \$10 to $\omega = 2$ "severe" experience to express risk preferences. Any model q under consideration implies a particular expected utility $E_q[U] = U \cdot q^{\mathsf{T}}$. The baseline model p has expected utility of -0.3; the alternative q^0 has expected utility of -0.55.

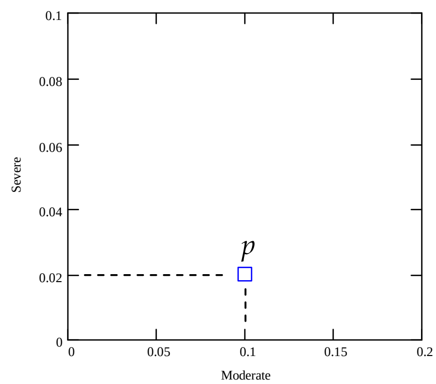

Figure 3: State space view of baseline model p.

3 Worst plausible models

In this section we address the problem of finding the most adverse model given a plausibility bound. In section 3.1 we review the applicable probability function and log likelihood ratio statistics. In sections 3.2 through 3.4 we solve the problem with plausibility defined by likelihood ratio, information gain, and statistical significance, respectively.

3.1 Likelihood ratio of alternative models

Samples drawn from a categorical distribution follow the multinomial distribution:

$$\Pr\left\{\begin{pmatrix} x_0 \\ x_1 \\ x_2 \end{pmatrix} \middle| \begin{pmatrix} p_0 \\ p_1 \\ p_2 \end{pmatrix}\right\} = \frac{(x_0 + x_1 + x_2)!}{(x_0!)\cdot(x_1!)\cdot(x_2!)} \cdot p_0^{\,x_0} \cdot p_1^{\,x_1} \cdot p_2^{\,x_2}. \qquad [3]$$

Here, the vector $X=(x_0, x_1, x_2)$ represents counts of how many of $n=x_0+x_1+x_2$ realizations of a single draw of ω fall into each category.

For the example, figure 4 shows the baseline model p (square), the alternative model q^0 (large dot) and 250 independent random samples, each of size n=100, (small crosses) drawn from model p. The random samples are represented by their empirical probability functions (epfs). Because the epfs can only take on values that are multiples of $1/n$ = 0.01, the plotted values have been "jittered" (displaced by a small random amount) to reduce overplotting and so permit better visualization.

It appears it is possible for p to generate data that would lead one to infer q^0. But does possible mean plausible?

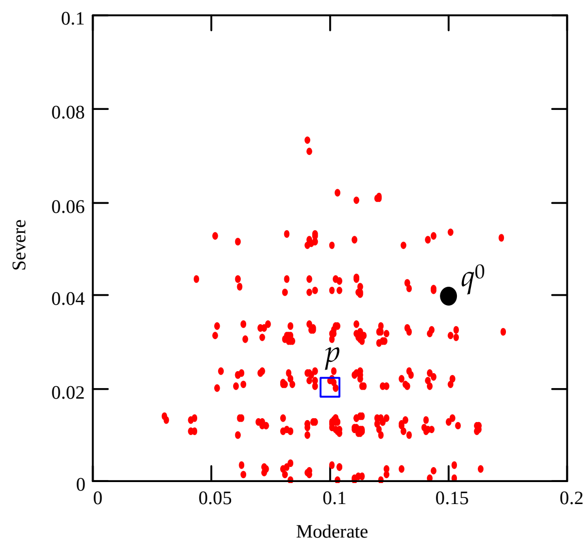

Figure 4: Models p and q^0 and 250 sample epfs (n=100) from p.

The log likelihood ratio $\lambda(\omega)$ between model p and a model q is a random variable defined over the D elements of Ω by $\lambda(\omega) = \ln(p_\omega/q_\omega)$. We can extend the concept across the sample to compute the sum of log likelihood ratios:

$$\lambda(X, p, q) = \sum_{i=0}^{D-1} x_i \cdot \ln\left(\frac{p_i}{q_i}\right). \qquad [4]$$

As a function of the data, this is the log likelihood ratio statistic.

3.2 Most adverse given a bound on the likelihood ratio

The user does not know the data X to which the model p was fit, only its size n. The simplest assumption is that the data conforms to the model: $X = n \cdot p$, or, equivalently, model p is the epf of the data.[2] Equation 4 becomes

[2] While technically this constrains the vector $n \cdot p$ to consist of integers, we will make use of the continuity of subsequent algebra to ignore this constraint. This could be formalized through a notion of "fractional" or weighted observations, but that is beyond the scope of this paper.

$$\lambda(n \cdot p, p, q) = n \cdot \sum_{i=0}^{D-1} p_i \cdot \ln\left(\frac{p_i}{q_i}\right). \quad [5]$$

Figure 5 shows contours of λ as a function of q.

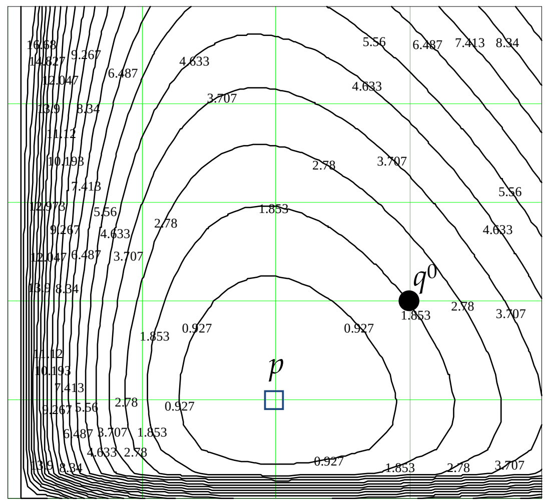

Figure 5 – Contours of log likelihood ratio (equation 5) as a function of q; same axes as figure 3.

Model q^0 has λ = 1.853, meaning that it is exp(-1.853) = 0.1567 times as likely to generate data $X=n{\cdot}p$ as p is.

Our first plausibility bound is thus defined in terms of the (log) likelihood: **An alternative model** q **will be deemed plausible** relative to the baseline model p and number of observations n if $\lambda(n{\cdot}p,p,q) \le \lambda_0$. Keep in mind that different models on a single λ contour line will in general have different expected utilities.

Say the user is concerned with models that are at least as plausible as the alternative q^0 with λ = 1.853, and within that set, she wants to know which alternative q^* produces the worst (lowest) expected utility $E_{q^*}[U]$. The solution is provided by the following:

Theorem 1: Given a baseline model p with all positive components, utility function U with distinct components, and a threshold $\lambda_0 > 0$ defining a set $Q=\{q: \lambda(n{\cdot}p,p,q) \le \lambda_0\}$ of alternative models. The problem $q^* = \arg\min_{q \in Q} E_q[U(\omega)]$ has a unique solution of the form

$$q_i^* = k \cdot (S + U_i)^{-1} \cdot p_i, \quad [6]$$

where k is a normalizing factor and S satisfies

$$\lambda_0/n = \sum_i p_i \cdot \ln(S + U_i) + \ln\left[\sum_i p_i \cdot \ln(S + U_i)^{-1}\right].$$

Proof: Proof is provided in appendix A. The result is a special case of equation 38 of [Breuer and Csiszár 2013], but we are not aware of an explicit discussion of this case in the literature. ▓

Remark: In order to provide positive components to q^*, it must be the case that $S > -\min(U_i)$. This is a one-dimensional search problem in S and is easily solved by iterative methods. The solution will always be on the boundary : $\lambda(n \cdot p, p, q^*) = \lambda_0$.

Example: Figure 6 illustrates the minimizing (and maximizing) constrained solutions where $\lambda_0 = 1.853$.

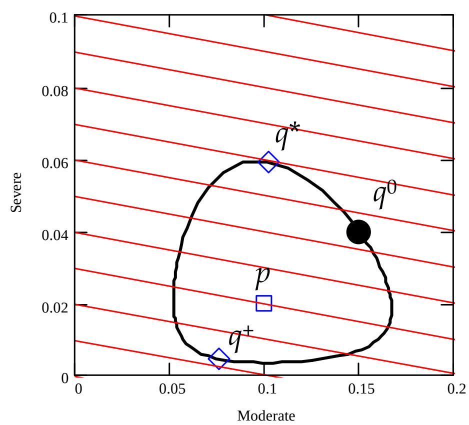

Figure 6: Constrained solutions for extreme $E_{q^*}[U]$.

The baseline model p is represented by the square inside the thick curve marking the boundary of models q satisfying $\lambda(n \cdot p, p, q) \leq \lambda_0$. Model q^0 is represented by the large dot in the upper right quadrant of that boundary. Contours of models with equal $E_q[U]$, spaced at units of ΔU=0.1 (higher utility on lower contours), are indicated by the thin diagonal lines. The most adverse model q^* = (0.8386, 0.1022, 0.0592) with EU = -0.6943 and least adverse model q^+ = (0.9191, 0.0764, 0.0045) with EU = -0.1213 are represented by diamonds.

The identification of q^* and q^+ is significant because either one (or, for that matter, q^0) is equally likely to produce data X resembling p, but expected utility is very different between them. They are models which the prudent decision maker should know about.

We put that model user in the position of deciding that alternative models beyond a certain likelihood ratio (0.1567) to the baseline were deemed implausible. This is an intuitively reasonable conception that can be presented to and discussed with non-technical audiences.

3.3 Most adverse given bounds on information gain

Define:

$$\mu_q \equiv \mu(q,p) \equiv \sum_{i=0}^{D-1} q_i \cdot \ln\left(\frac{q_i}{p_i}\right) \tag{7}$$

This is the expected value of the log likelihood ratio (based on a single observation; multiply by n for larger samples) of q compared to p, given that model q generated the sample. It allows for the possibility that $X = n\cdot p$ was generated by q, but measures the average $\ln(q/p)$ across samples that are more probably closer to q than to p.

The quantity μ_q is variously known as *Kullback-Leibler* [1951] *divergence, cross-entropy, relative entropy,* and *information gain*. It has deep connections with information theory [Rényi 1961] and is itself a reasonable metric for the (directed) distance from q to p. Burnham and Anderson [2002] show how information gain underpins the Akaike [1981] Information Criterion in selecting models. Hansen and Sargent [2008] base their theory, robust decision making in the face of ambiguity, on information gain as the plausibility metric, and solve a problem related to Theorem 2 (below).

Remark: Reversing the roles of p and q gives us (up to the multiplier N) the log likelihood developed in section 3.2. This N=1 version of log likelihood is sometimes referred to as *reversed relative entropy* [Breuer and Csiszár 2013] or *Burg entropy* [Ben-Tal et al. 2013].

Information gain attains its minimum at $\mu = 0$ when $q = p$. The definition is extended to include some $q_j=0$ by continuity because $\lim_{q\to 0^+} q\cdot\ln(q) = 0$.

Our second plausibility bound is thus defined in terms of information gain: **An alternative model q will be deemed plausible** relative to the baseline model p and number of observations n if $n\cdot\mu(q,p) \le \mu_0$.

The following result is well-known in the literature.

Theorem 2: Given a baseline model p, utility function U with distinct component values, and an information gain threshold $\mu_0 < \max_i|\ln(p_i)|$ defining a set $Q=\{q{:}\mu_q \le \mu_0\}$ of alternative models. If Q is interior to the simplex $\mathcal{S}$, then there is a unique solution to the problem $q^* = \arg\min_{q\in Q} E_q[U(\omega)]$ given by

$$q_i^* = k \cdot \exp(c \cdot U(i)) \cdot p_i, \qquad [8]$$

where k is a normalizing factor and c satisfies

$$f(c) \equiv \frac{\sum_i \exp(c \cdot U_i) \cdot (c \cdot U_i) \cdot p_i}{\sum_i \exp(c \cdot U_i) \cdot p_i} - \ln\left(\sum_i \exp(c \cdot U_i) \cdot p_i\right) = \mu_0. \qquad [9]$$

Proof: Proofs appear in [Friedman 2002] and [Breuer & Csiszár 2013**].** A proof is included in appendix B for convenience. ▓

Remarks: Equation 8 is the well-known Esscher [1932] transform. In the context of distributions, it is typically developed from a dual problem, as the minimum information gain (maximum entropy) solution subject to a utility function constraint. The equation $f(c) = \mu_0$ is a one-dimensional search problem in c and is easily solved by iterative methods. The solution will always be on the boundary : $\mu_{q^*} = \mu_0$.

Example: Continue with the same p, U, and identified alternative q^0 as before. We take $\mu_0 = \mu_{q^0} = 0.0214$ as the information gain constraint. Figure 7 shows a trace of $f(c)$ against c. There is a positive as well as a negative solution. The positive solution is $c = 0.3371$; the negative is $c = -0.1052$. The corresponding models are $q^* = (0.9243, 0.0750, 0.0007)$ and $q^+ = (0.8394, 0.1060, 0.0546)$.

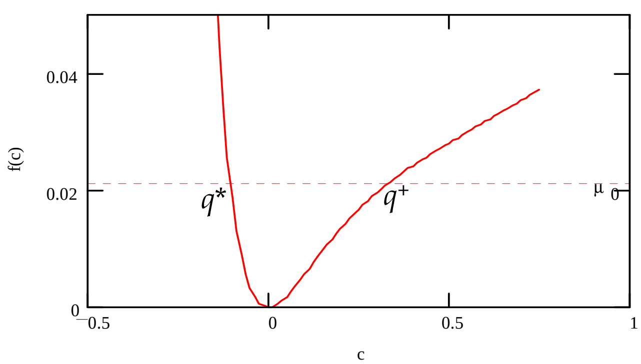

Figure 7: Root finding for the information gain constraint.

The information gain-bounded extreme models are close to, but not equal to, the solutions found in section 3.2 based on likelihood bounds, despite the fact that the same utility function U was used for minimization and the same alternative q^0 was used to define the bounds of plausibility.

The reason for this can be seen in figure 8; the constraints provide different bounds. This should come as no surprise; while ln(p/q) = -ln(q/p) for each scenario, the expectations for the log are driven by p for the λ bounds and by q for the μ bounds.

Interpreting information gain probabilistically is not quite as straightforward as interpreting likelihood. With likelihood, we were measuring the relative probabilities of assumed $X=n{\cdot}p$ data under p and its alternative q. With information gain, we no longer make use of assumptions about X and instead inquire about the typical behavior of the log likelihood if q were to be the data generating process.

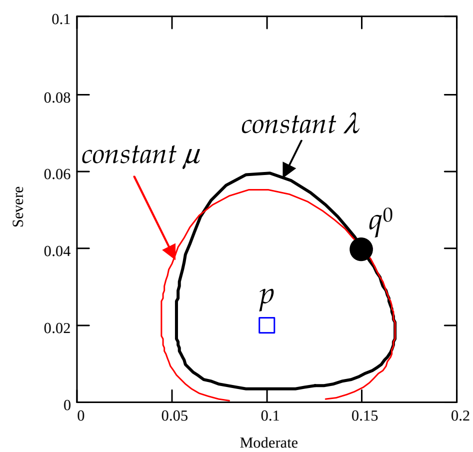

Figure 8: log likelihood λ and information gain μ plausibility bounds.

The closest we can come to a probabilistic interpretation is to exponentiate the information gain to become a likelihood ratio; specifically, the geometric mean likelihood ratio. But now it is p that is unlikely relative to q. So q is to be interpreted as implausible because it implies that p would be unlikely, if q were true.

Creating information gain bounds for Q is therefore somewhat problematic for our hypothetical user. What criteria can she use to select a threshold? How can she explain what this threshold means to a non-technical business audience?

3.4 Most adverse given significance of hypothesis test

The interpretation of the "preferred likelihood" assumption implemented in section 3.2 (likelihood ratio bound) is quite stringent; p is exactly the epf of the data. The interpretation of preferred likelihood in section 3.3 (information gain bound) is inscrutable; p is the target for likelihood comparisons, but specific assumptions about X are not used.

In this section, we want to interpret p as a very good, but not necessarily perfect, representative of the data. In order to implement this, we turn to formal hypothesis testing with the likelihood ratio test [Neyman & Pearson 1933].

Figure 9 shows the familiar p and q^0 models, along with 250 (jittered) epfs corresponding to size n = 100 random samples drawn from q^0. Clearly, samples are more likely to resemble q^0 than p.

Three diagonal lines are also shown. These indicate samples X (represented by epfs X/n) sharing the same log likelihood of data X when comparing p to q^0. These contours are straight lines because λ is linear in X. The top line has λ = 2.14. Samples along this line favor q by a factor of exp(2.14) = 8.51. The bottom line has λ = -1.85; samples favor p by a factor of exp(1.85) = 6.38 The center line has λ = 0. *Samples on this line favor neither model.*

The user imagines performing a likelihood ratio test, deciding to reject the hypothesis that alternative q produced the data upon which p was built, in favor of p, when $\lambda(X,q,p)<0$. The user will not be able to perform this test, because X is not available, but she is willing to assume that X lies below the middle line, i.e. $\lambda<0$. Her metric of plausibility for alternative model q is the significance of the test: **An alternative model q will be deemed plausible** relative to p and the number of observations n if $\Pr\{\lambda(X,q,p)<0 \mid X\sim q\} \equiv \alpha \geq \alpha_0$. We no longer interpret preferred likelihood as meaning that X/n lies exactly at p, rather, we take it to mean only that it lies below the middle line.

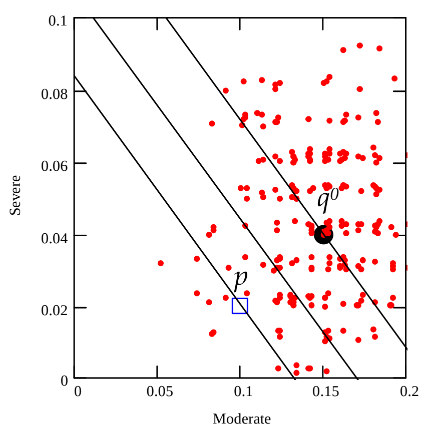

Figure 9: Contours of $\lambda(X,q,p)$.

An exact calculation of this probability is given by

$$\alpha = \sum_{X \in S} \Pr(X|q) \quad where \quad S = \{X | \lambda(X, q, p) < 0\} \qquad [10]$$

where $\Pr(X \mid q)$ is given by equation 3. That is, all possible samples X of size n satisfying $\lambda<0$ are identified and their occurrence probability summed. The total number of possible samples (which is the number of terms in a multinomial sum of D variables raised to the nth power) is given by $(n+D-1)!/(n! \cdot D!)$ [Feller 1950].

For $D=3$ and $n=100$, the number of distinct possible samples is 176,851. As D increases, this quantity grows exponentially. For $D=5$, it is over 96 million. For $D=10$ it is nearly 47 trillion. This is one reason we seek an approximation to equation 10.

Define μ_q as above in equation 7 and define

$$\sigma_q = \sqrt{\sum_{i=0}^{D-1} q_i \cdot \left[\ln\left(\frac{q_i}{p_i} \right) \right]^2 - \mu_q^2} \qquad [11]$$

Whereas μ_q is the sample mean of the log likelihood ratio, σ_q^2 is its variance.

Theorem 3: Let Φ be the standard normal cdf and define $\alpha = \Pr\{\lambda(X, q, p) < 0 | X \sim q\}$ and

$$\Xi(q, p, n) \equiv \Phi\left(-\sqrt{n} \cdot \frac{\mu_q}{\sigma_q} \right). \qquad [12]$$

Then α and Ξ are asymptotically equivalent, that is, as the sample size n increases without bound, α/Ξ converges to one.

Proof: This follows from the central limit theorem. ▓

Example: An exact calculation summing $\Pr(X \mid q^0)$ over all possible samples X where $\lambda(X,q^0,p)<0$ yields $\alpha = 0.1691$ whereas theorem 3's equation 12 yields 0.1667.

We will use this formula, rather than the exact calculation, in the sequel. One reason is its tractability, compared to tabulating $\Pr(X \mid q)$ and λ across all possible samples. A second, and more important, reason is that the exact calculation is not continuous in q.

Figure 10 illustrates this, showing precise and approximate α for a section of alternative models with fixed $q_2=0.03$, against the baseline model p and $n=100$. The crosses are calculated from equation 10 whereas the line is calculated from equation 12.

We saw previously that two models q and q' having the same λ may not have the same μ. Similarly, they may not have the same α, because to do so they would need to have the same ratio μ_q / σ_q. If they were to have the same information gain, μ, they must also have the same value for σ. But this is not true in general. If figure 8 were to be redrawn with all three definitions of plausibility, the new locus of models with constant significance α would nearly, but not quite, coincide with the locus of constant likelihood λ. The locus of constant information gain μ would remain quite distinct by comparison.

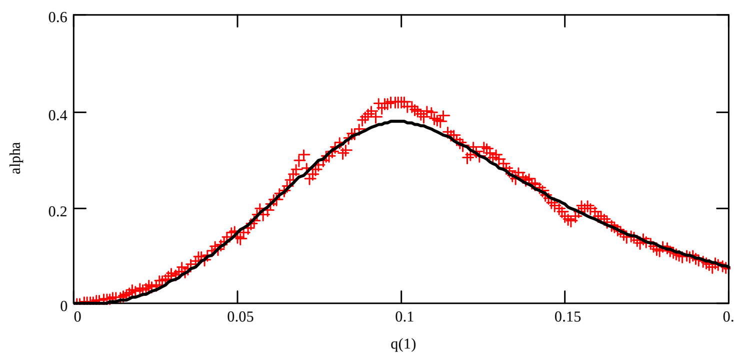

Figure 10: α for q=(0.97-q_1,q_1,0.03) is discontinuous.

Some properties of $\alpha = \Xi(q,p,n)$ are given in the following:

Theorem 4: Given a baseline model p, with all positive components, and an alternative model q,

a. $0 < \alpha < ½$ if $q \neq p$.
b. As q approaches p, $\alpha \to ½$.
c. As q approaches a vertex of $\mathcal{S}$ (one component equal to one, all others zero), $\alpha \to 0$.
d. As q approaches a non-vertex boundary of $\mathcal{S}$ (at least one but fewer than D-1 components equal to zero), α approaches a number strictly between 0 and ½.

Proof: The proof is presented in appendix C. ▓

The next result, to our knowledge, is new.

Theorem 5: Given a baseline model p, utility function U, and a significance threshold α defining a set $Q = \{ q : \Xi(q,p,n) \geq \alpha \}$ of alternative models, then $q^* \equiv \arg\min_{q \in Q} E_q[U(\omega)]$ exists and takes the form

$$q_i^* = k \cdot \exp\left(\pm \sqrt{A - \frac{U_i}{\theta}} \right) \cdot p_i , \qquad [13]$$

where k is the normalizing factor, and A and θ are scalars such that q^* satisfies the constraints

$$\alpha = \Xi(q^*,p,n) \quad \text{and} \quad k = \exp(R_\alpha \cdot \mu_{q^*} - 1)$$

where $R_\alpha \equiv \dfrac{\sigma_{q^*}{}^2 + \mu_{q^*}{}^2}{\mu_{q^*}{}^2}$.

Proof: The proof is presented in appendix D. ▓

Remark: If the "±" of equation 13 were "+" or "-" consistently across all components i, q^* could be found through a two-dimensional (A,θ) constraint satisfaction search. Unfortunately that is not always the case, making the search more difficult. Therefore, we recommend robust numerical search procedures that directly target the expected utility minimizing q^*. The theorem can then be used to verify that the putative solution is of the right form. We conjecture that the solution is unique as long as the components of U are distinct.

Example: The Cross-Entropy ("CE") method [Rubenstein & Kroese 2004] was applied to find the most adverse q^* with the same α-level as q^0 (α = 0.1667). The solution is q^* = (0.8382, 0.1033, 0.0586). Regressing $[\ln(q^*/p)\text{-}R_\alpha \cdot \mu+1]^2$ on U, we find A = 3.202, θ = -3.627 and that "±" resolved into "-" for all components. This solution is not far from that of section 3.2.

4 Worst credible (most dangerous) models

In this section we address finding credibility-weighted adverse models, or, equivalently, posterior expected utility given priors of a certain form. The weighting or priors reflect an interpretation of the prior symmetry assumption. In section 4.1 we introduce the weighting scheme. In section 4.2 we show it is equivalent to a particular version of the Gilboa-Schmeidler framework as they designed it to operate on priors. In sections 4.3 to 4.5 we address the implementation using likelihood, information gain, and statistical significance, respectively, as the basis for conditional probability.

4.1 Credibility and posterior expectation

If one had the conditional probability of the data given a model, call it $\chi(X \mid q)$, and a prior probability of each model, call it $\pi(q)$, then one could use Bayes' Rule to compute the posterior probability of a model given the data. In the case of only two models p and q, with $\pi(p)+\pi(q)=1$, the formula is

$$\pi(q|X) = \frac{\pi(q)\cdot\chi(X|q)}{\pi(p)\cdot\chi(X|p)+\pi(q)\cdot\chi(X|q)}. \qquad [14]$$

If we interpret the prior symmetry assumption to mean $\pi(p)=\pi(q)=1/2$, then the formula reduces to

$$\pi(q|X) = \frac{\chi(X|q)}{\chi(X|p)+\chi(X|q)}. \qquad [15]$$

This carries over to posterior expectations, as well. Conditional on the data X, we can compute the posterior expected utility as

$$E[U|X,q] = \frac{\chi(X|p)\cdot p+\chi(X|q)\cdot q}{\chi(X|p)+\chi(X|q)}\cdot U^{\top} \equiv q^{w}\cdot U^{\top}. \qquad [16]$$

Actuaries are familiar with this sort of calculation as a credibility-weighted average between $E_p[U]$ and $E_q[U]$, with weights proportional to the conditional probability of the data given a model.

We may then pose a problem similar to that of equation 1 of section 1.1:

$$\arg\max_{a\in A}\min_{q\in\mathcal{B}} E_q[U_a|X,q] \qquad [17]$$

This seeks the best choice where expected outcomes are calculated not on the worst plausible (bounded) models, but on the worst credibility weighted (and unbounded) model.

4.2 Gilboa-Schmeidler

In the Gilboa-Schmeidler framework, one solves the optimization problem

$$\arg\max_{a\in A}\min_{\pi\in\Psi} E^{\pi}[E_{\theta}[U_a]] \qquad [18]$$

where now $\pi \in \Psi$ is a *prior distribution* on the space of models $\mathcal{S}$, and Ψ is a specified set of such priors. Here, $\theta \in \mathcal{S}$ is a random model distributed as π. Specifically,

$$E^{\pi}[\varphi] = \frac{\int \varphi \cdot \chi(X|\theta) \cdot d\pi(\theta)}{\int \chi(X|\theta) \cdot d\pi(\theta)} \qquad [19]$$

(with a suitable interpretation of the integrals). Thus, rather than seeking out the most adverse model q, one seeks out the most adverse prior, π.

We now interpret the operations of section 4.1 in terms of this framework.

Let $\pi_p(\cdot)$ represent a "typical" (proper or improper) prior distribution, but assume that $E^{\pi_p}[\theta] = p$. That is, assume the model developers used this prior to fit the data and that p was the posterior mean distribution. This may be interpreted as a version of the preferred likelihood assumption. Because $E_\theta[U]$ is linear in θ, this also means that $E^{\pi_p}[E_\theta[U]] = E_p[U]$. Let $\pi_q(\cdot)$ represent a point mass prior distribution giving probability one for any model subset A that includes q and zero for all others. Thus $E^{\pi_q}[E_\theta[U]] = E_q[U]$.

Finally, let $\psi_q(\cdot) = (\pi_p(\cdot)+\pi_q(\cdot))/2$ and let $\Psi = \{\psi_\theta \,|\, \theta \in \mathcal{S}\}$. Clearly all elements of Ψ are priors on $\mathcal{S}$. It is evident that now the problem of finding the utility minimizing prior $\psi_{q^\varnothing}$ in equation 18 is equivalent to finding the utility minimizing model $q^\varnothing$ in equation 17.

We are not quite done, however, because Gilboa & Schmeidler's representation theorem requires that Ψ be closed and convex. We first extend Ψ to Ψ' by including all finite convex combinations of elements $\psi \in \Psi$. Because posterior utility E^ψ is linear in ψ, extremal values of E^ψ where $\psi = \alpha \cdot \psi_1 + (1-\alpha) \cdot \psi_2$ will be attained at $\alpha = 0$ or $\alpha = 1$, so the minimizing solutions in equation 18 over Ψ' in fact occur within Ψ. Finally, extending Ψ' to a closed $\underline{\Psi}'$ by including all limits (in the sense of weak convergence) does not add new extremes.

This establishes our procedure of section 4.1 as being a version of the Gilboa-Schmeidler equation 18. Note, however, that most implementations of the theory use information gain with respect to a baseline π_p to define the set Ψ; information gain between π_p and ψ_q here is undefined due to the existence of singularities (non-coinciding mass points).

Remark: The set $\underline{\Psi}'$ constructed above is known as the ε-contamination of $\pi_p(\cdot)$ with all other priors, and with ε set to one-half, constantly. Such sets appear often in the robust Bayesian literature.

4.3 Most adverse credible using likelihood

Using likelihood as our plausibility metric, we effectively assume that $X = n \cdot p$, so that $\chi(X \mid q)$ = Pr{ $n \cdot p \mid q$} from equation 3. Rewriting in terms of log likelihood ratios, equation 16 becomes

$$E[U|X,q] = \frac{E_q[U] + \exp(\lambda(n \cdot p, p, q)) \cdot E_p[U]}{1 + \exp(\lambda(n \cdot p, p, q))}, \qquad [20]$$

from whence it follows that

$$E[U|X,q] - E_p[U] = \frac{E_q[U] - E_p[U]}{1 + \exp(\lambda(n \cdot p, p, q))}. \qquad [21]$$

Thus, the amount by which the posterior expected utility differs from the baseline model utility is a ratio: the numerator is the difference in the two models' utility; the denominator is a deflator that increases with size as the plausibility of model q decreases.

Figure 11 illustrates. Recall from figure 6 that expected utility contours are shallow straight lines sloping down to the right. Such a line through the baseline model p would neatly divide the two sets of concentric posterior utility contours. Models near the baseline differ little in posterior utility from p's own expected utility; neither do the implausible models that are far away from p. Two global extrema are evident. A global minimum exists at a model exhibiting slightly higher probability for moderate events (horizontal) and substantially higher probability for extreme events (vertical). A global maximum exists in the opposite direction.

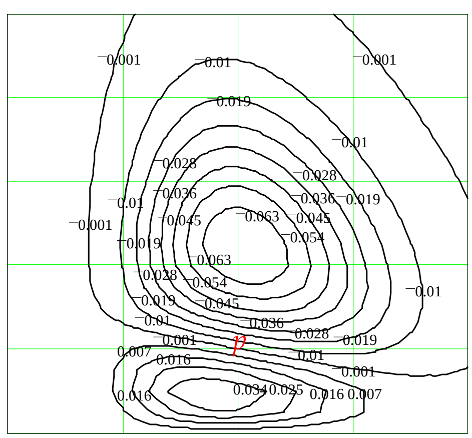

Figure 11: Posterior expected utility difference for alternative models (same axes as figure 3)

We seek to minimize this difference, that is to say, to maximize it in the negative, disutility, direction. The next theorem tells us how:

Theorem 6: Given a baseline model p and utility function U, the solution to the problem of minimizing E[U|X,q] (defined in equations 20 and 21), by choosing $q \in \mathcal{S}$ unrestricted, takes the form of:

$$q_i^{\varnothing} = k \cdot (B + U_i)^{-1} \cdot p_i \qquad [22]$$

where k is the normalizing factor, and B is a scalar constant.

Proof: The proof is presented in appendix E. ▓

Remark: Note that $q^{\varnothing}$ takes the same form as q^* from equation 6. The derivative dE[U|X, $q^{\varnothing}$]/dB is sufficiently complicated that a direct search for the minimizing B in E[U|X, $q^{\varnothing}$] is preferred. We do not know if the solution is unique. Fortunately, in any particular application, a trace like figure 12 will reveal the nature of solutions.

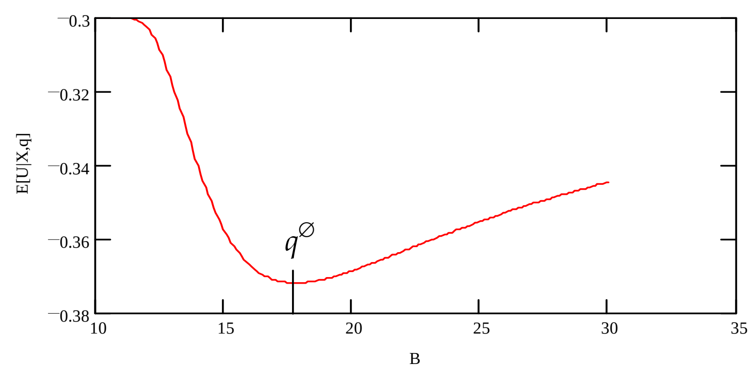

Figure 12: Posterior utility as a function of the parameter B

Example: Figure 12 shows the trace of E[U|X,q] as a function of B for positive B. Clearly, the desired solution will use B>0 as that will increase the probability on components with more negative U. The solution is $q^{\varnothing}$ = (0.8531, 0.1027, 0.0442) with E[U|X,$q^{\varnothing}$] = -0.3718, some 24% worse than $E_p[U]$ = -0.3. This solution corresponds to q^* at exp(-λ) = 0.415.

4.4 Can information gain be used for credibility?

For reasons explained in the latter part of section 3.3, it seems impractical to use information gain *between models* to define posterior utility. Information gain *between priors*, used as a mechanism to bound the set of priors, would be mathematically practical. However, the problem of articulating an information gain threshold in a business context, alluded to in section 3.3, is now compounded by the fact that one would be operating at a whole new level of abstraction: Bayesian priors over models, not just models.

Therefore we will not attempt to use information gain to define credibility.

4.5 Most adverse credible using significance

Whereas plausibility–as-likelihood assumes X=$n \cdot p$, plausibility-as-statistical significance assumes merely that a log-likelihood ratio test (with threshold zero) of the data X would prefer p to q. Specifically, $\alpha = \Pr\{\lambda(X,q,p) < 0 | X \sim q\}$ is the probability that a sample X (here a random variable) generated by q would indicate p as being more likely in a likelihood ratio test. Conversely, $\beta = 1 - \Pr\{\lambda(X,p,q) < 0 | X \sim p\}$ is the probability that a sample generated by p would prefer p to q. We will use the approximation from equation 12 and write $\alpha = \chi(X\,|\,q) = \Xi(q,p,n)$ and $\beta = \chi(X\,|\,p) = 1 - \Xi(p,q,n)$. Substituting into equation 16 we get

$$E[U|X,q] = \frac{\alpha \cdot E_q[U] + \beta \cdot E_p[U]}{\alpha + \beta}. \qquad [23]$$

Equivalently, we have

$$E[U|X,q] - E_p[U] = \frac{\alpha}{\alpha + \beta} \cdot \left(E_q[U] - E_p[U]\right). \qquad [24]$$

Similar to the situation in section 4.3, the amount by which the posterior expected utility differs from the baseline model utility is a product: the second factor is the difference in the two models' utility; the first factor is a weight $w = \alpha/(\alpha+\beta)$ that decreases with size as the plausibility of model q decreases.

We seek to minimize this difference, that is to say, to maximize it in the negative, disutility, direction. Unfortunately, we do not have a theorem that tells us how to do this; we cannot even be assured that a solution exists for the problem as posed.

We note some properties of β and $w = \alpha/(\alpha+\beta)$:

Theorem 7: Given a baseline model p, with all positive components, and an alternative model q,

a) If $q \neq p$ then $\frac{1}{2} < \beta < 1$, and $\alpha < \alpha/(\alpha+1) < w < \alpha/(\alpha+\frac{1}{2}) < 2 \cdot \alpha$.
b) As q approaches p, $\beta \rightarrow \frac{1}{2}$, and so does w.
c) As q approaches a vertex of $\mathcal{S}$ (one component equal to one, all others zero), $w \rightarrow 0$.
d) As q approaches a non-vertex boundary of $\mathcal{S}$ (at least one but fewer than D-1 components equal to zero), neither β nor w necessarily converge.

Proof: The proof is provided in appendix F. ▓

Remark: Despite not converging to the same value from all directions at a boundary of the simplex, β and w are nonetheless constrained by the bounds given in part (a), where α has computable values on the boundary; β and w do not explode.

Despite not having theoretical backing for a solution, we may nonetheless apply numerical methods in search of one. If numerical methods indicate the worst credible model is away from the boundary, we may feel comfortable that a solution has been found. Still, with no proof of uniqueness, all we can say is the solution provides a local minimum.

Example: The CE method was applied to find the worst credible alternative model using the significance as weight function. The result is $q^{\varnothing} = (0.8427, 0.1038, 0.0535)$ with $E[U \mid X, q^{\varnothing}] = -0.3629$. This solution applies 21% more probability to the "severe" component than does the solution in section 4.3, but it produces 2.4% more favorable posterior expectation.

In our numerical work, we have discovered that worst credible $q^{\varnothing}$ are not generally equal to worst plausible q^* at any α level.

For these reasons, we may want to modify our model to use an approximation: $\beta = 1-\alpha$. This is not entirely nonsensical; both are computed from the coefficient of variation of the random variable $\ln(q/p)$. The difference is that one uses q as the probability distribution and the other uses p. As q approaches p, the approximation converges to an equality. Figure 13 shows $\alpha+\beta$ for our example.

With this modification, we have $w = \alpha$, and theorem 4 applies to w. The posterior mean function becomes continuous on the entire simplex, and thus attains global extrema somewhere. Moreover, the problem of finding the minimal $q^{\varnothing}$ becomes, in principle, a one-dimensional search over α, because $q^{\varnothing} = q^*$. We conjecture the solution is unique.

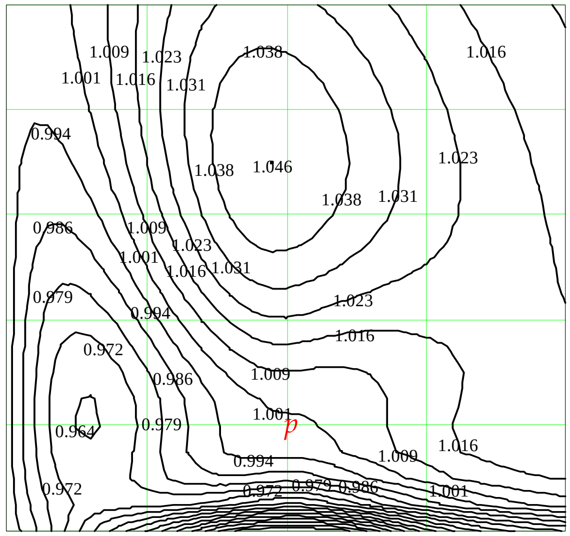

Figure 13: α+β is approximately one.

Example: Using CE with this modified weighting, the most adverse credible model is $q^{\varnothing}$ = (0.8424, 0.1031, 0.0544), which is within 0.0012 (Euclidean distance) of the unmodified solution. The posterior expectation is -0.3628, only 0.0001 away from the previous solution. This solution corresponds to the most adverse model bounded by α = 0.1878.

5 Case study: decision making with model risk

In this section we address a realistic reinsurance decision problem. The probability distribution is taken from the cat model results of a real company, but rescaled to make it anonymous. Pricing assumptions are based on an arbitrage-free methodology consistent with observed cat reinsurance prices at a particular point in time.

5.1 Statement of the problem

The company provided a modeled distribution of annual catastrophe losses, presented as a D=29-dimensional categorical distribution. The probabilities and corresponding direct (before reinsurance) losses are set out in the first two columns of table 4 in appendix G. We visualize these losses in figure 14 on the vertical axis against return period (inverse of exceedance probability) on the horizontal axis. The expected direct loss under this baseline model is 1.527.

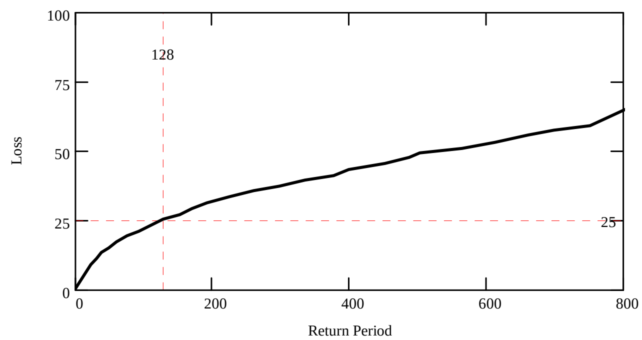

Figure 14: Baseline model, direct loss

The company writes multiple lines of business and allocates 25 units of capital to this particular business segment. The model predicts a probability of 0.78%, or once every 128 years, that 25 units or more will be lost due to this risk. Losses in excess of 25 are of particular concern for management.

Table 2 shows the reinsurance options being considered. They are all of the aggregate excess-of-loss (XOL) type, without reinstatement provisions.

Table 2: Reinsurance program choices

Program	Attach	Payout	Limit	Exhaust	Premium
0 – No reins (bare)	0	0	0	0	0
1 – Full/low	25	100%	10	35	0.315
2 – Stretch/low	25	60%	15	50	0.357
3 – Stretch/high	30	66.7%	20	60	0.365
4 – Two layers	25 35	50% 33.3%	5 10	35 65	0.310

A program will reimburse direct losses during the year that exceed the attachment point, at the payout rate, until the limit has been paid:

$$\text{Reimbursement} = \max\{0,\min[(\text{DirectLoss-Attach})\cdot\text{Payout},\text{Limit}]\}. \qquad [25]$$

The reimbursement limit will have been reached when the direct loss equals the exhaustion point. The annual premium is listed in the last column. Program "0" corresponds to no reinsurance.

The company is concerned with capital deficit, i.e., net loss amounts in excess of 25, where net loss is defined as direct loss plus reinsurance premium minus reimbursement. The criterion for ranking the programs will be the expected capital deficit, corresponding to utility functions that are the negative of capital deficit, as depicted in figure 15. For example, if direct loss realized equals 60, program 3 provides the best coverage because utility is the highest (least negative).

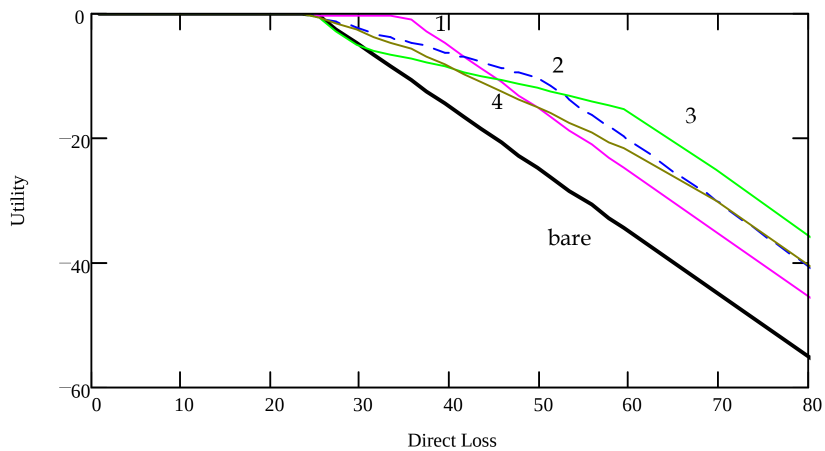

Figure 15: Utility functions of reinsurance programs

The expected utility (using the baseline cat model) for each program is as follows:

Program	0	1	2	3	4
$E_p[U]$	-0.130	-0.079	-0.077	-0.082	-0.086

Using expected utility as our conventional decision criterion, we reach the conclusion that all reinsurance options are superior to going without (program 0). Program 2 is the best option and program 1 is next best.

We can now apply the results of this paper to seek a program whose performance is robust across an array of models, one of which may describe the true loss process. The company may be willing to sacrifice some expected utility as measured by the baseline model p if a program performs adequately across a larger range of uncertainty.

5.2 Worst plausible models

We now find, for each program, the worst plausible model using the likelihood method of section 3.2. We opt for likelihood here because it is the simplest computationally. We assume the baseline cat model p is founded on n=250 observations. A threshold of 3 is used, that is, models with likelihood less than 1/3 of the baseline model are considered implausible.

The alternative models are very similar to each other and are shown in figure 16.

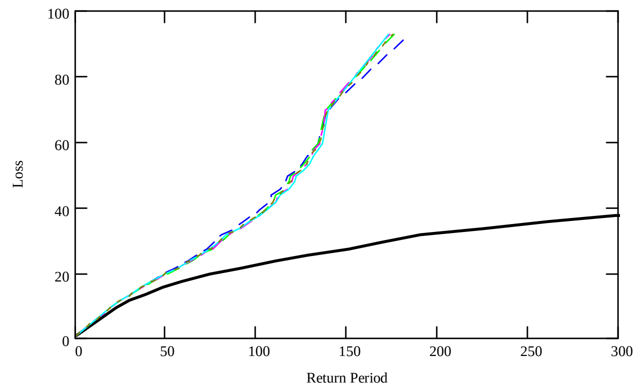

Figure 16: Worst plausible alternative models by program

The worst plausible alternative model for program 0 is listed in appendix G, column 3. The expected direct loss associated with this model is 2.104. The expected utilities associated with each program are as follows:

Program	0	1	2	3	4
$E_{q^*}[U]$	-0.537	-0.419	-0.387	-0.363	-0.397

These results suggest that if we prefer a reinsurance program which will perform well in the event that model p is wrong, program 3 is the best option; program 2 is now second best. This relative comparison of expected utility under worst plausible conditions is useful if we have significant doubts about the accuracy of model p.

5.3 Worst credible models

We now find, for each program, the worst credible model using the likelihood method of section 4.3. That n=250 is still assumed.

Figure 17 shows the credibility-weighted results. That is, the alternatives $q^{\varnothing}$ are computed according to Theorem 6 (equation 22), and then the *weighted* models

$$q^w = p + (q^{\varnothing} - p)/(1+\exp(\lambda(n \cdot p,p,q^{\varnothing})) \qquad [26]$$

are shown in figure 17.

Again, the alternative models are very similar. The worst credible model for program 0 is listed in appendix G, column 4. The expected direct loss associated with this model is 1.671.

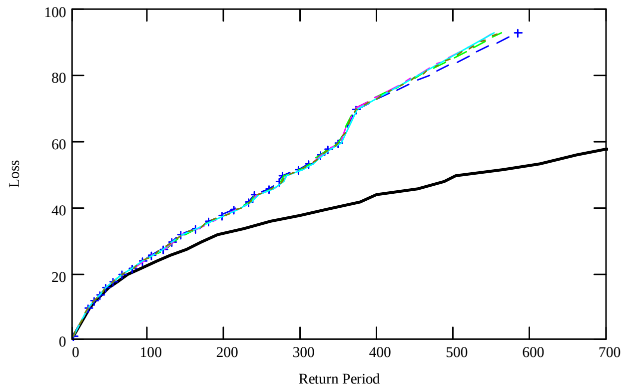

Figure 17: Worst credible models by program

Calculations of worst credible models using significance (full $\alpha+\beta$ method, CE algorithm) instead of likelihood produced very similar results.

The expected utilities across all models are given in table 3.

Table 3: Expected utilities under baseline and alternative models

Program	Model	0	1	2	3	4
$E_p[U]$	Baseline	-0.130	-0.079	-0.077	-0.082	-0.086
$E_{q^*}[U]$	Worst Plausible	-0.537	-0.419	-0.387	-0.363	-0.397
$E_{q^w}[U]$	Worst Credible	-0.232	-0.164	-0.155	-0.152	-0.163
$E_{q^w}[U]$	Worst Credible (significance)	-0.218	-0.151	-0.143	-0.142	-0.152

Figure 18 displays the difference in expected utility between each reinsurance program and program 0 using the four models for comparison.

5.4 Commentary

Program comparisons under adverse alternative models are useful additional criteria for the company's decisions.

Reinsurance program 2 is superior in meeting the company's objective according to the baseline model results. However, uncertainty inherent in the model ("model risk") casts doubt on that. Programs 1 and 4, inferior according to the baseline, proved *fragile* in the alternative analyses, showing even worse relative performance. They are to be rejected outright on those grounds. Program 3 was seen as a superior performer under the alternative models. It is *robust* in the face of uncertainty around the baseline model and deserves consideration alongside program 2.

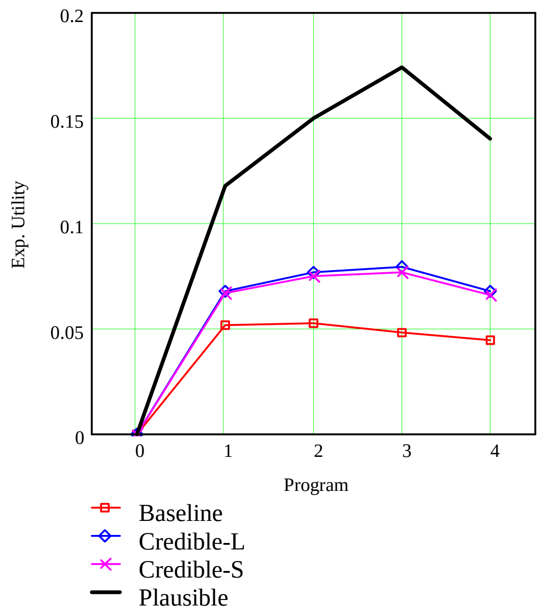

Figure18: Comparison of expected utilities

A more thorough analysis would expand the criteria further.

For example, the same calculations were performed with a 95-component discretization to test the sensitivity of the method. While the expected utility values were somewhat different, the overall preference orderings and conclusions were the same.

It might also be advisable to probe the sensitivity of the results to various utility functions. For example, a range of convex combinations of expected net loss and capital deficit could be explored. Determining the worst plausible and worst credible models using the likelihood definition of plausibility do not require lengthy simulations or arduous numerical calculations, so such exploration is quite feasible.

6 Conclusion

We have articulated an approach to assessing model risk (specifically model uncertainty or distribution model risk) by identifying certain adverse alternative models. These models should be of concern to an ambiguity-averse decision maker because they are (1) materially different from the baseline (accepted) model, yet (2) could plausibly represent reality better than does the baseline.

Our approach is generic and does not require deep insight into the specification of the baseline model nor access to the data upon which it was fit. This is particularly appealing for users of large-scale commercial models, for example, natural catastrophe models.

Whereas the literature has tended to focus on limiting alternatives by an information gain threshold or penalty, we introduced two other plausibility criteria that are arguably superior in terms of their interpretation for a business audience. The likelihood ratio (reverse entropy) criterion is equivalent to assuming that the baseline model coincides with the data upon which it was fit. The significance criterion is equivalent to assuming that the baseline model is a better fit than the alternative, although not necessarily the best fit.

Additionally, we introduced posterior or credibility-weighted plausibility, which identifies an important alternative model without requiring specification of the plausibility threshold. This procedure was shown to be an instance of the Gilboa-Schmeidler multiple-priors framework, using ε-contamination of the baseline prior.

The likelihood ratio criterion provides for efficient computation of the adverse alternative in either the plausibility or credibility context. The information gain criterion is efficient for limited plausibility, but suffers from conceptual difficulties for credibility. The significance criterion appears to be best approached by general-purpose optimization techniques.

Finally, we applied the likelihood plausibility and credibility methods to a reinsurance purchase case study. Two decision options could be dismissed as inferior. However, a third deserved consideration, despite not being the best according to the baseline model.

References

Akaike, Hirotugu (1981) "Likelihood of a Model and Information Criteria," *Journal of Econometrics* 16.1: 3-14.

Anderson, Evan, Lars Peter Hansen, and Thomas Sargent (2000) "Robustness, Detection and the Price of Risk," manuscript, Stanford University.

Barillas, Francisco, Lars Peter Hansen, and Thomas J. Sargent (2009) "Doubts or Variability?" *Journal of Economic Theory*, 144.6: 2388-2418.

Ben-Tal, A., D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen (2013) "Robust Solutions of Optimization Problems Affected by Uncertain Probabilities," *Management Science*, **59.2:** 341-357.

Breuer, Thomas, and Imre Csiszár (2013) "Measuring distribution model risk," *Mathematical Finance*.

Burnham, Kenneth P., and David R. Anderson (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach,* New York: Springer.

Crouhy, Michel, Dan Galai, and Robert Mark (2000) *Risk Management,* New York: McGraw-Hill.

Derman, Emanuel (1996) "Model Risk: What are the assumptions made in using models to value securities and what are the consequent risks?" *RISK-LONDON-RISK MAGAZINE LIMITED-* 9: 34-38.

Efron, Bradley (1982) *The Jackknife, the Bootstrap and Other Resampling Plans,* CBMS-NSF Regional Conference Series in Applied Mathematics Vol. 38, Philadelphia: Society for Industrial and Applied Mathematics.

Ellsberg, Daniel (1961) "Risk, Ambiguity, and the Savage Axioms," *The Quarterly Journal of Economics:* 643-669.

Epstein, Larry G., and Tan Wang (1994) "Intertemporal Asset Pricing Under Knightian Uncertainty," *Econometrica: Journal of the Econometric Society,* 283-322.

Esscher, F. (1932) "On the Probability Function in the Collective Theory of Risk," *Skand. Aktuarie Tidskr.* 15: 175-195.

Feller, William (1950) *An Introduction to Probability Theory and Its Applications: Volume One,* New York: John Wiley & Sons.

Föllmer, Hans, and Alexander Schied (2002) "Robust preferences and convex measures of risk," *Advances in Finance and Stochastics,* Springer Berlin Heidelberg, 39-56.

Friedman, Craig (2002) "Confronting model misspecification in finance: Tractable collections of scenario probability measures for robust financial optimization problems," *International Journal of Theoretical and Applied Finance*, 5:33

Garlappi, Lorenzo, Raman Uppal, and Tan Wang (2007) "Portfolio Selection with Parameter and Model Uncertainty: A Multi-Prior Approach," *Review of Financial Studies*, 20.1: 41-81.

Gilboa, Itzhak, and David Schmeidler (1989) "Maxmin Expected Utility with Non-Unique Prior," *Journal of Mathematical Economics* **18.2**: 141-153.

Goldfarb, Donald, and Garud Iyengar (2003) "Robust Portfolio Selection Problems," *Mathematics of Operations Research*, 28.1: 1-38.

Hansen, Lars Peter, and Thomas J. Sargent (2008) *Robustness,* Princeton, NJ: Princeton University Press.

Kocay, William, and Donald L. Kreher (2004) *Graphs, Algorithms, and Optimization,* Boca Raton, FL: CRC Press.

Kullback, Solomon, and Richard A. Leibler (1951) "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 79-86.

Kunreuther, Howard, Robin Hogarth, and Jacqueline Meszaros (1993) "Insurer Ambiguity and Market Failure," *Journal of Risk and Uncertainty*, 7:71-87.

Major, John A. (2011) "Managing Catastrophe Model Uncertainty: Issues and Challenges." Guy Carpenter & Company, London and New York.

Major, John A. (2012) "Using Models with Uncertainty: Applying Post-Modern 'Robust' Decision Theory," presented at the Casualty Actuarial Society Seminar on Reinsurance, Cambridge, MA. Available at https://cas.confex.com/cas/rein12/webprogramucas/Handout/Session5140/Major_Model_Uncertainty.pdf

Major, John A. (2014) "Model Risk: No More Pretending," presented at the Casualty Actuarial Society Seminar on Reinsurance, New York, NY. Available at https://cas.confex.com/cas/rein14/webprogram/Handout/Paper3234/CARe2014_C16_JAM_v7.pdf

Major, John A. and Micah Woolstenhulme (2011) "The Need For Robustness in Catastrophe Modeling," ASHK Newsletter volume 2, July. Available at http://www.actuaries.org.hk/upload/File/ Newsletter-volume2-2011%28FINAL%29.pdf

Mehra, Rajnish, and Edward C. Prescott (1985) "The Equity Premium: A Puzzle," *Journal of Monetary Economics,* 15.2: 145-161.

Miller, David (1999) "Uncertainty in Hurricane Risk Modeling and Implications for Securitization," *Discussion Paper Program*, Casualty Actuarial Society.

Møller, Thomas (2004) "Stochastic Orders in Dynamic Reinsurance Markets," *Finance and Stochastics* 8.4: 479-499.

Moore, James (1998) "Tail estimation and catastrophe security pricing: can we tell what target we hit if we are shooting in the dark?" Unpublished Working Paper, Wharton Financial Institutions Center.

Neyman, Jerzy, and Egon S. Pearson (1933) "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London, Series A,* vol. 231:289-337.

Osborne, Martin J., and Ariel Rubinstein (1994) *A Course in Game Theory,* Cambridge, MA: MIT Press.

Rényi, Alfred (1961) "On Measures of Entropy and Information," *Fourth Berkeley Symposium on Mathematical Statistics and Probability.*

Rubinstein, Reuven Y., and Dirk P. Kroese (2004) *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. New York: Springer Science + Business Media.

Venter, Gary, Jack Barnett, and Michael Owen (2004) "Market Value of Risk Transfer: Catastrophe Reinsurance Case," *International AFIR Colloquium*.

Venter, Gary and Rajesh Sahasrabuddhe (2012) "A Note on Parameter Risk," *Casualty Actuarial Society E-Forum*, Summer 2012.

Woolstenhulme, Micah and John A. Major (2011) "Model Risk in Financial Systems: The Need for Robust Decision Control," presented at the Casualty Actuaries of Greater New York Fall Meeting, New York, NY. Available at http://www.casact.org/community/affiliates/cagny/1211/Woolstenhulme.pdf

Zhao, Lin, and Wei Zhu (2011) "Ambiguity Aversion: A New Perspective on Insurance Pricing," *Astin Bulletin* 41.01: 157-189.

Zhou, Kemin, and John C. Doyle (1997) *Essentials of Robust Control*, Upper Saddle River, NJ: Prentice Hall.

Appendix A: proof sketch of Theorem 1 – most adverse given likelihood threshold

Theorem 1: Given a baseline model p with all positive components, utility function U with distinct components, and a threshold $\lambda_0 > 0$ defining a set Q={q: $\lambda(n\cdot p,p,q) \leq \lambda_0$} of alternative models. The problem $q^* = \arg\min_{q\in Q} E_q[U(\omega)]$ has a unique solution of the form

$$q_i^* = k\cdot(S+U_i)^{-1}\cdot p_i\,, \tag{6}$$

where k is a normalizing factor and S satisfies

$$\lambda_0/n = \sum_i p_i\cdot\ln(S+U_i)+\ln\left[\sum_i p_i\cdot\ln(S+U_i)^{-1}\right].$$

Proof: Notice that if all U components were equal then all $E_q[U]$ would be equal and there would be no unique minimum. The assumption of distinct components is sufficient to prevent this. Notice that the objective function is linear in q. Furthermore, the set Q={q: $\lambda(n\cdot p,p,q) \leq n\cdot\mu_0$} is bounded, because the space of q is bounded, and it is closed because λ (equation 5) is continuous in q and Q is thus the inverse image of a closed set. Therefore Q is compact. Thus, the minimum must occur on the boundary of Q [Kocay & Kreher 2004, p. 391]. Thus, we may rewrite the constraint as $\lambda(n\cdot p,p,q) = n\cdot\mu_0$.

This is equivalent to minimizing the following:

$$L=\theta\cdot\left(-\mu_0+\sum_i p_i\cdot\ln\left(\frac{p_i}{q_i}\right)\right)+S\cdot\left(-1+\sum_i q_i\right)+\left(\sum_i q_i\cdot U_i\right) \tag{A1}$$

where θ and S are the Lagrange multipliers associated with the μ_0 and probability constraints, respectively. Note that as any component q_i approaches 0 (all $p_i>0$ is assumed), the corresponding term $p_i\cdot\ln(p_i/q_i)$ increases without bound. Therefore Q is interior to the simplex domain $\mathcal{S}$, and therefore the edge constraints do not need to be represented. Differentiating on q_j, we obtain the first-order condition for a maximum or minimum of L:

$$\theta\cdot\left(\sum_i p_i\cdot\frac{d}{dq_j}\ln\left(\frac{p_i}{q_i}\right)\right)+S\cdot\left(\frac{d}{dq_j}\sum_i q_i\right)+\left(\frac{d}{dq_j}\sum_i q_i\cdot U_i\right)=0$$

$$-\theta\cdot\frac{p_j}{q_j}+S+U_j=0$$

$$q_j^*=\frac{\theta}{S+U_j}\cdot p_j \quad where \quad \theta=\left[\sum_i\frac{p_i}{S+U_i}\right]^{-1} \tag{A2}$$

Substituting the expressions for q and θ into the λ constraint, we get

$$\mu_0 = f(S) \equiv \sum_i \left[p_i \cdot \ln(S + U_i)\right] + \ln\left[\sum_i \frac{p_i}{S + U_i}\right] \qquad \text{[A3]}$$

This is a one-dimensional search problem in S. Feasible solutions lie in the disconnected set $\mathcal{S} = (-\infty,-\max(U)) \cup (-\min(U),\infty)$. For solutions S in the left half, θ will be negative to compensate. For those solutions, equation A3 is better represented as

$$\mu_0 = g(S) \equiv \sum_i \left[p_i \cdot \ln(|S| - U_i)\right] + \ln\left[\sum_i \frac{p_i}{|S| - U_i}\right] \qquad \text{[A3']}$$

so as to obviate the need to take logs of negative numbers.

We need to show such a solution exists and is unique. First, we show there exists a solution in each of the two disconnected halves of $\mathcal{S}$. In the right half we have:

$$\lim_{S\to\infty} f(S) = 0 \quad \textit{and} \quad \lim_{S\to -\min(U)^+} f(S) = \infty$$

(the second holding as long as the component m corresponding to the minimum of U does not have probability p_m=1, which it will not, given the assumption that all p_i>0 and assuming we are not dealing with the trivial case D=1). Note that f is continuous. Therefore a solution exists in the right half.

In the left half, similarly, we have:

$$\lim_{S\to\infty} g(S) = 0 \quad \textit{and} \quad \lim_{S\to -\max(U)^+} g(S) = \infty$$

with the same comments regarding the component m corresponding to the maximum of U. Therefore a solution exists in the left half as well.

Now we show that each of the two solutions is unique in its half. In the right half,

$$\frac{d}{dS} f(S) = \sum_i \left[p_i \cdot \frac{d}{dS}\ln(S + U_i)\right] + \frac{d}{dS}\ln\left[\sum_i \frac{p_i}{S + U_i}\right]$$

$$= \left(\sum_i \frac{p_i}{S + U_i}\right) - \frac{\sum_i \frac{p_i}{(S + U_i)^2}}{\sum_i \frac{p_i}{S + U_i}} = -\frac{\operatorname{var}\left[(S + U)^{-1}\right]}{E\left[(S + U)^{-1}\right]} < 0$$

the latter inequality holding unless all components of U are the same, which is assumed not the case. Thus the right half solution is unique. The calculation of the derivative in the left half leads to the same conclusion.

Inspecting equation A2, we can see that the right half solution puts more weight on components where U is lower; that is the most adverse solution. The left half solution is the least adverse. ▓

Appendix B: proof sketch of Theorem 2 – most adverse given information gain threshold

Theorem 2: Given a baseline model p, utility function U with distinct component values, and an information gain threshold $\mu_0 < \max_i|\ln(p_i)|$ defining a set $Q=\{q:\mu_q \le \mu_0\}$ of alternative models. If Q is interior to the simplex $\mathcal{S}$, then there is a unique solution to the problem $q^* = \arg\min_{q\in Q} E_q[U(\omega)]$ given by

$$q_i^* = k \cdot \exp(c \cdot U(i)) \cdot p_i\,, \tag{8}$$

where k is a normalizing factor and c satisfies

$$f(c) \equiv \frac{\sum_i \exp(c \cdot U_i)\cdot(c \cdot U_i)\cdot p_i}{\sum_i \exp(c \cdot U_i)\cdot p_i} - \ln\left(\sum_i \exp(c\cdot U_i)\cdot p_i\right) = \mu_0\,. \tag{9}$$

Proof: Notice that if all U components were equal then all $E_q[U]$ would be equal and there would be no unique minimum. The assumption of distinct components is sufficient to prevent this. The condition that $\mu_0 < \max_i|\ln(p_i)|$ is necessary because μ_q attains a maximum of $-\ln(p_i)$ for the largest p_i when q_i=1. For larger μ_0, no solution is possible.

The same logic in appendix A applies: the minimum must occur on the boundary of Q, thus, we may rewrite the constraint as $\mu_q = \mu_0$.

This is equivalent to minimizing the following:

$$L = \theta\cdot\left(-\mu_0 + \sum_i q_i \cdot \ln\left(\frac{q_i}{p_i}\right)\right) + S\cdot\left(-1+\sum_i q_i\right) + \left(\sum_i q_i \cdot U_i\right) \tag{B1}$$

where θ and S are the Lagrange multipliers associated with the μ_0 and probability constraints, respectively. Because we assume that Q is interior to the simplex domain $\mathcal{S}$, the edge constraints do not need to be represented. Differentiating on q_i, we obtain the first-order condition for a maximum or minimum of L:

$$\theta\cdot\left(\frac{d}{dq_j}\sum_i q_i\cdot\ln\left(\frac{q_i}{p_i}\right)\right) + S\cdot\left(\frac{d}{dq_j}\sum_i q_i\right) + \left(\frac{d}{dq_j}\sum_i q_i\cdot U_i\right) = 0$$

$$\theta\cdot\left(\ln\left(\frac{q_j}{p_j}\right)+1\right) + S + U_j = 0$$

$$\ln\left(\frac{q_j}{p_j}\right) = \frac{-S-U_j}{\theta} - 1$$

$$q_j^* = K \cdot \exp\left(-\frac{U_j}{\theta}\right) \cdot p_j \qquad \text{[B2]}$$

$$\textit{where} \quad K = \exp\left(-\frac{S}{\theta} - 1\right) \quad \textit{satisifies} \quad \sum_i q_i = 1 .$$

Evidently, this is a solution. Is it unique? Rewrite equation B2 as

$$q_i = \exp(c_1 \cdot U_i + c_2) \cdot p_i . \qquad \text{[B3]}$$

The solution must satisfy the two constraints:

$$\sum_i q_i = \sum_i \exp(c_1 \cdot U_i + c_2) \cdot p_i = 1 \qquad \text{[B4]}$$

$$\sum_i q_i \cdot \ln\left(\frac{q_i}{p_i}\right) = \sum_i \exp(c_1 \cdot U_i + c_2) \cdot (c_1 \cdot U_i + c_2) \cdot p_i = \mu_0 \qquad \text{[B5]}$$

We can eliminate c_2 by solving equation B4,

$$c_2 = -\ln\left(\sum_i \exp(c_1 \cdot U_i) \cdot p_i\right), \qquad \text{[B6]}$$

so equation B5 becomes equation 9, defining $f(c)$.

Define $g(c) = \sum_i \exp(c \cdot U_i) \cdot p_i$; then $f(c) = \frac{c \cdot g'(c)}{g(c)} - \ln(g(c))$ and

$$f'(c) = \frac{c}{g(c)^2} \cdot \left(g''(c) \cdot g(c) - g'(c)^2\right)$$

$$= \frac{c}{g(c)^2} \cdot \left[\left(\sum_i \exp(c \cdot u_i) \cdot u_i^{\,2} \cdot p_i\right) \cdot \left(\sum_i \exp(c \cdot u_i) \cdot p_i\right) - \left(\sum_i \exp(c \cdot u_i) \cdot u_i \cdot p_i\right)^2\right]$$

$$= \frac{c}{g(c)^2} \cdot \left(\sum_{i>j} \exp\left(c \cdot (u_i + u_j)\right) \cdot (u_i - u_j)^2 \cdot p_i \cdot p_j\right)$$

and evidently $f'(c)>0$ for $c>0$ and $f'(c)<0$ for $c<0$ because no $u_i=u_j$ so all terms in parentheses are positive. This means there are at most two values for c that satisfy $f(c)=\mu_0$; one positive and one negative. Obviously, one of them corresponds to the maximum and one to the minimum of $E_q[U]$. ∎

Appendix C: proof sketch of Theorem 4 –properties of α

Theorem 4: Given a baseline model p, with all positive components, and an alternative model q,

a. $0 < \alpha < ½$ if $q \neq p$.
b. As q approaches p, $\alpha \to ½$.
c. As q approaches a vertex of $\mathcal{S}$ (one component equal to one, all others zero), $\alpha \to 0$.
d. As q approaches a non-vertex boundary of $\mathcal{S}$ (at least one but fewer than D-1 components equal to zero), α approaches a number strictly between 0 and ½.

Proof:

(a) It follows from Gibbs' inequality that $\mu_q>0$ and from Jensen's inequality that $\sigma_q>0$, therefore the argument to Φ is negative.

(b) let $q = p+\varepsilon$ and note that $\Sigma_i\varepsilon_i = 0$ always holds. Expand $\ln(q_i/p_i) = \ln(1+\varepsilon_i/p_i)$ terms in Taylor series to order 2, obtaining:

$$\frac{\mu}{\sigma} = \frac{\frac{1}{2}\cdot\sum_i \frac{\varepsilon_i^2}{p_i} + O(\varepsilon^3)}{\sqrt{\sum_i \frac{\varepsilon_i^2}{p_i} + O(\varepsilon^3) - \left(\frac{1}{2}\cdot\sum_i \frac{\varepsilon_i^2}{p_i} + O(\varepsilon^3)\right)^2}}. \qquad \text{[C1]}$$

Notice that the numerator is of order 2 but the denominator is of order 1, therefore $\mu/\sigma \to 0$.

(c) With one component, say i=0, converging to 1, μ converges to $-\ln(p_0)>0$. However, σ converges to 0, so the ratio diverges to $+\infty$.

(d) Both μ and σ can be seen to converge to positive numbers. ▓

Appendix D: proof sketch of Theorem 5 – most adverse given significance threshold

Theorem 5: Given a baseline model p, utility function U, and a significance threshold α defining a set $Q = \{ q : \Xi(q,p,n) \geq \alpha \}$ of alternative models, then $q^* \equiv \arg\min_{q \in Q} E_q[U(\omega)]$ exists and takes the form

$$q_i^* = k \cdot \exp\left(\pm \sqrt{A - \frac{U_i}{\theta}} \right) \cdot p_i, \qquad [13]$$

where k is the normalizing factor, and A and θ are scalars such that q^* satisfies the constraints

$$\alpha = \Xi(q^*,p,n) \quad \text{and} \quad k = \exp(R_\alpha \cdot \mu_{q^*} - 1)$$

where $R_\alpha \equiv \dfrac{{\sigma_{q^*}}^2 + {\mu_{q^*}}^2}{{\mu_{q^*}}^2}$.

Proof: Theorem 4 allows us to extend the definition of $\alpha = \Xi(q,p,n)$ to be a continuous function on the entire (compact) simplex $\mathcal{S}$. Therefore, by Weierstrass's extreme value theorem, Ξ will attain its extrema.

First, we note that again the solution will occur on the boundary of Q, and the constraint $\alpha = \Xi(q,p,n) \equiv \Phi\left(-\sqrt{n} \cdot \dfrac{\mu_q}{\sigma_q} \right)$ is equivalent to

$$\frac{\sigma_q^2 + \mu_q^2}{\mu_q^2} = 1 + \frac{n}{\Phi^{-1}(\alpha)^2} \equiv R_\alpha . \qquad [D1]$$

We may then write the problem in Lagrange multiplier form as minimizing

$$L = \theta \cdot \left(\sum_i q_i \cdot \ln\left(\frac{q_i}{p_i}\right)^2 - R_\alpha \cdot \left(\sum_i q_i \cdot \ln\left(\frac{q_i}{p_i}\right)\right)^2 \right) + S \cdot \left(-1 + \sum_i q_i \right) + \left(\sum_i q_i \cdot U_i \right) [D2]$$

Differentiating with respect to q_j, we get

$$0 = \theta \cdot T_1 - \theta \cdot R_\alpha \cdot T_2 + S + U_j \quad \textit{where}$$
$$T_1 = \frac{d}{q_j} \sum_i q_i \cdot \ln\left(\frac{q_i}{p_i}\right)^2 = \ln\left(\frac{q_j}{p_j}\right)^2 + 2 \cdot \ln\left(\frac{q_j}{p_j}\right)$$
$$T_2 = \frac{d}{q_j} \left(\sum_i q_i \cdot \ln\left(\frac{q_i}{p_i}\right)\right)^2 = 2 \cdot \left(\ln\left(\frac{q_j}{p_j}\right) + 1\right) \cdot \sum_i q_i \cdot \ln\left(\frac{q_i}{p_i}\right)$$

Writing $\lambda_j = \ln(q_j / p_j)$ and $\mu = \sum_i q_i \cdot \lambda_i$, we have the quadratic equation

$$\lambda_j^2 + 2 \cdot (1 - R_\alpha \cdot \mu) \cdot \lambda_j + \left(\frac{U_j + S}{\theta} - 2 \cdot R_\alpha \cdot \mu\right) = 0 \qquad \text{[D3]}$$

whose solution is

$$\lambda_j = (R_\alpha \cdot \mu - 1) \pm \sqrt{A - \frac{U_j}{\theta}} \quad \textit{where} \qquad \text{[D4]}$$
$$A = 1 + R_\alpha{}^2 \cdot \mu^2 - \frac{S}{\theta}$$

thus proving the assertion. ▓

Appendix E: proof sketch of Theorem 6 – most adverse credible based on likelihood

Theorem 6: Given a baseline model p and utility function U, the solution to the problem of minimizing E[U | X,q] (defined in equations 20 and 21), by choosing $q \in \mathcal{S}$ unrestricted, takes the form of:

$$q_i^{\varnothing} = k \cdot (B + U_i)^{-1} \cdot p_i \qquad [22]$$

where k is the normalizing factor, and B is a scalar constant.

Proof: We seek to minimize equation 21, which we rewrite here as $L = \left(\sum_i U_i \cdot (q_i - p_i) \right) \cdot (1+\Lambda)^{-1}$ where $\Lambda = \exp(\lambda(n \cdot p, p, q))$.

As q approaches a boundary, the denominator 1+Λ increases without bound but the numerator remains finite. Thus we may extend L to the entire simplex with a value of zero on the boundary. Existence of a solution is therefore assured; not only will the function attain its extremes, but it will do so in the interior of the simplex. Moreover, for q=p, L=0. Small perturbations to q will result in positive or negative values for the numerator and positive values for the denominator, so L is not identically zero in the simplex.

Writing in Lagrange multiplier form, we seek to minimize

$$L = S \cdot \left(-1 + \sum_i q_i \right) + \left(\sum_i U_i \cdot (q_i - p_i) \right) \cdot (1+\Lambda)^{-1} \qquad [E1]$$

Notice that $d\Lambda/dq_j = -n \cdot (p_j/q_j) \cdot \Lambda$. Differentiating L with respect to q_j and setting to zero gives us

$$0 = S + U_j \cdot (1+\Lambda)^{-1} + \left(\sum_i U_i \cdot (q_i - p_i) \right) \cdot (1+\Lambda)^{-2} \cdot n \cdot \frac{p_j}{q_j} \cdot \Lambda . \qquad [E2]$$

Multiplying both sides by $(1+\Lambda)^2$ then rearranging, we get

$$q_j = \frac{\sum_i U_i \cdot (p_i - q_i) \cdot n \cdot \Lambda}{S \cdot (1+\Lambda)^2 + (1+\Lambda) \cdot U_j} \cdot p_j \equiv \frac{A}{B + U_j} \cdot p_j \qquad [E3]$$

where, of necessity, $A = \left(\sum_i \frac{p_i}{B + U_i} \right)^{-1}$.

Feasible solutions for B lie within the disconnected set $\mathcal{S}$ = $(-\infty, -\max(U)) \cup (-\min(U), \infty)$. The left hand portion (yielding A<0) will produce smaller denominators, thus higher weight, for larger values of U. Thus, the left hand portion contains the maximizing solution and the right hand portion contains the minimizing solution. ▓

Appendix F: proof sketch of Theorem 7 –properties of β, w

Theorem 7: Given a baseline model p, with all positive components, and an alternative model q,

a) If $q \neq p$ then $\frac{1}{2} < \beta < 1$, and $\alpha < \alpha/(\alpha+1) < w < \alpha/(\alpha+\frac{1}{2}) < 2 \cdot \alpha$.
b) As q approaches p, $\beta \rightarrow \frac{1}{2}$, and so does w.
c) As q approaches a vertex of $\mathcal{S}$ (one component equal to one, all others zero), $w \rightarrow 0$.
d) As q approaches a non-vertex boundary of $\mathcal{S}$ (at least one but fewer than D-1 components equal to zero), neither β nor w necessarily converge.

Proof:

(a) By Theorem 4(a), $0 < \alpha < \frac{1}{2}$. Because $\beta = 1-\alpha$ (with the model arguments reversed), it follows that $\frac{1}{2} < \beta < 1$. The rest follows from elementary algebra.

(b) Similarly, this follows from Theorem 4(b)

(c) It doesn't matter what β does; because $\alpha \rightarrow 0$ (Theorem 4(c)), so does w.

(d) Consider the following example with $D \geq 4$; q_0 and q_1 converging to 0, other components converging to values strictly between 0 and 1. Finite terms in the numerator and denominator are dominated by the unbounded q_0 and q_1 terms, so the limit expression is

$$\frac{\mu}{\sigma} \approx \left[\frac{\sum_{i=0,1} p_i \cdot \ln\left(\frac{p_i}{q_i}\right)^2}{\left(\sum_{i=0,1} p_i \cdot \ln\left(\frac{p_i}{q_i}\right)\right)^2} - 1 \right]^{-\frac{1}{2}} = \left[\frac{\sum_{i=0,1} p_i \cdot \ln(q_i)^2}{\left(\sum_{i=0,1} p_i \cdot \ln(q_i)\right)^2} - 1 \right]^{-\frac{1}{2}} . \qquad \text{[F1]}$$

Let convergence operate as $q_0 = t$, $q_1 = t^k$, $t \rightarrow 0$. In the limit the ratio becomes $(p_0+k^2 \cdot p_1)/(p_0+k \cdot p_1)^2$. This is not constant in k, so the limit does not exist. ▓

Appendix G: Details of the section 5 example

Table 4: Baseline and alternative models

Monetary Units	Probabilities		
Direct Loss	**Baseline**	**Worst Plausible, Program 0**	**Worst Credible, Program 0**
0.7392	0.95571	0.94882	0.95398
9.4234	0.01110	0.01102	0.01108
11.4227	0.00724	0.00719	0.00723
13.4211	0.00532	0.00528	0.00531
15.4421	0.00380	0.00377	0.00379
17.4578	0.00332	0.00330	0.00331
19.4496	0.00258	0.00256	0.00258
21.4375	0.00192	0.00191	0.00192
23.5410	0.00122	0.00121	0.00122
25.3559	0.00118	0.00118	0.00118
27.3889	0.00072	0.00074	0.00072
29.5455	0.00066	0.00070	0.00067
31.6000	0.00080	0.00087	0.00082
33.4667	0.00060	0.00067	0.00062
35.6667	0.00048	0.00056	0.00005
37.3889	0.00036	0.00043	0.00038
39.4118	0.00034	0.00042	0.00036
41.5714	0.00014	0.00018	0.00015
43.5333	0.00030	0.00040	0.00032
45.7500	0.00016	0.00022	0.00018
47.6667	0.00006	0.00009	0.00007
49.7273	0.00022	0.00033	0.00025
51.1429	0.00014	0.00021	0.00016
53.3333	0.00012	0.00019	0.00014
55.7500	0.00008	0.00014	0.00009
57.8000	0.00010	0.00018	0.00012
59.4286	0.00014	0.00026	0.00017
69.7606	0.00071	0.00177	0.00098
92.9583	0.00048	0.00543	0.00171
E[Direct Loss]	1.527	2.104	1.671

Direct Loss is the monetary loss experienced in the absence of reinsurance. *Baseline* is the corresponding probability according to the accepted cat model. *Worst plausible* and *worst credible* are, respectively, the most adverse likelihood 3:1 limited and most adverse posterior expected alternative models for the direct loss.

Appendix H: Index of symbols

Below are symbols, each with a brief definition and where they first occur.

$a \in A$	– decision option a out of choice set A; §2
α	– significance p-level for likelihood ratio test; §3.4, eq 10
β	– power of the test; §4.5
$\chi(X \mid q)$	– generic likelihood function; §4.1
D	– dimension of the probability space; §2
$\mathrm{E}_q[U]$	– expected utility based on specified model; §2
$\mathrm{E}[U \mid X, q]$	– posterior expected utility; §4.1, eq 16
$\mathrm{E}^{\pi}[\cdot]$	– expectation with respect to prior π; §4.2, eq 19
λ	– log likelihood ratio; §3, eq 4
μ	– expected log likelihood or information gain; §3.3
n	– number of observations in baseline fitted data; §3
p	– baseline model; §2
$\pi(q)$	– prior probability of model; §4.1
$\pi(q \mid X)$	– posterior probability of model; §4.1, eq 14
$\pi_p(\cdot)$	– working prior resulted in p as the posterior; §4.2
$\pi_q(\cdot)$	– point prior always results in q as the posterior; §4.2
Ψ	– set of prior distributions on models; §4.2
$\psi_q(\cdot)$	$= (\pi_p(\cdot)+\pi_q(\cdot))/2$; §4.2
q	– generic alternative model; §1.1
q^0	– a specifically identified alternative model; §3
q^*	– most adverse model given plausibility bound; §1.1
q^+	– least adverse model given plausibility bound; §3.2
$q^{\varnothing}$	– most adverse model in posterior expectation; §4.3
q^w	– credibility weighted $p, q^{\varnothing}$ models; §5.3, eq 26
Q	– generic set of plausible models; §1.1
$\mathcal{S}$	– D-dimensional simplex, space of models; §2
σ	– standard deviation of log likelihood ratio; §3.4, eq 11
θ	– random model distributed according to a prior; §4.2
$U_a(\omega)$	– utility function for decision option a; §2
w	$= \alpha/(\alpha+\beta)$; §4.5
$\omega \in \Omega$	– categorical r.v. taking values in a set of size D; §2
X	– sample data observed or generated by a model; §3
Ξ	– normal approximation to α; §3.4, eq 12