Applied Robust Performance Analysis for Actuarial Applications

Jose Blanchet* Henry Lam† Qihe Tang‡ Zhongyi Yuan§

February 7, 2017

Abstract

This paper investigates techniques for the assessment of model error in the context of insurance risk analysis. The methodology is based on finding bounds for quantities of interest, such as loss probabilities and conditional value-at-risk, which are obtained by solving optimization problems where the variable to optimize is the model itself in a non-parametric framework. The non-parametric aspect of the approach is crucial for model error quantification.

1 Introduction

We shall study the problem of quantifying the impact of model assumptions that are uncertain or might actually be incorrect in actuarial risk analysis. We present a general methodology that applies to a wide range of problems. However, to explain both the motivation and the proposed methodology, let us fix a canonical example throughout our discussion.

Let us say that an insurer is interested in evaluating the expected shortfall or the conditional value-at-risk (C-VaR) of a given portfolio of risk exposures. The expected shortfall is the conditional expected size of the deficit faced by the company, given the unlikely event of insufficiency of statutory solvency capital. The solvency capital is set at a high level, so that the likelihood of a loss in a given year is small. According to Solvency II, the capital should be sufficient to withstand losses with probability not lower than .995 during a one-year time horizon [1].

*Department of Industrial Engineering and Operations Research, Columbia University, New York, NY.
†Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI.
‡Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA.
§Department of Risk Management, Pennsylvania State University, University Park, PA.

There are two main reasons why the evaluation of performance risk measures (such as C-VaR) is challenging. First, in typical applications, there is simply not enough data to accurately calibrate models relative to such a low-probability event. Second, underlying risk factors often possess a dependence structure that might be difficult to estimate, once again, with the required degree of accuracy in the context of tail events.

To cope with these challenges actuaries and statisticians have developed models and calibration procedures that are designed to capture stylized features believed, based on experience or expert knowledge, to accurately describe common underlying risk factors. Nevertheless, the reality is that it is extremely difficult to be highly certain of the validity of a probabilistic model that is postulated to accurately assess events that are supposed to happen once every hundreds of years (as prescribed by Solvency II).

A natural approach taken in practice is to carry out scenario analysis and stress testing, but it is not clear how to generate the scenarios or which variables to stress and at which level. In addition, it is important to keep in mind that actuaries must balance model fidelity (i.e., descriptive power of the model) with tractability. This additional constraint compounds the difficulty in accurate risk quantification to an even higher level.

The methodology that we study in this paper has the following characteristics:

1) Its starting point is a baseline model that, for whatever reason (maybe a good balance of fidelity and tractability), is currently employed by the actuary.

2) The method employs optimization theory to find a bound for the underlying risk metric of interest (say, C-VaR), where the optimization is non-parametrically imposed over all probabilistic models that are in the neighborhood of the baseline model.

3) Typically the optimal solution (i.e., the model obtained from the optimization procedure) can be written in terms of the baseline model. So the proposed procedure can be understood as a correction to quantify for the possibility of model misspecification.

The approach that we discuss in this paper has its roots in areas such as economics, operations research and statistics. Here we start the discussion with some background behind the formulation we propose. In the context of economics, the work of two Nobel laureates L. P. Hansen and T. J. Sargent [14] studies optimal decision making when the decision maker lacks full information

about the underlying probabilistic model, focusing in particular on its macroeconomic implications. Similar ideas have been used in portfolio optimization [11] and in quantifying the so-called model risks [12] in finance. In stochastic control, best control policies are derived under ambiguous rather than fully specified transition distributions [24, 22, 15]. In the operations research literature, the general topic of distributionally robust optimization was developed in [9, 3, 16, 20], where reformulations and efficient algorithms were studied to handle uncertainty in the probabilistic assumptions in various stochastic optimization problems. For consistency, throughout this paper, we will adopt the terminology from the operations research literature and call our methodology *distributionally robust optimization*.

Our contribution in this paper is to bring the combination of the ideas in these areas to the attention of the actuarial community. In addition, given that risk assessment is of special importance for actuaries, we also discuss the implications of the distributionally robust methodology that we advocate in the setting of tail events. These types of discussions are not standard in the literature, and they provide, we believe, additional insights relevant to actuarial risk analyses.

Throughout the rest of the paper, we will explain our modeling approach and discuss its validity. We have chosen to present examples that are relatively simple because of pedagogical purposes, but our goal is to convince the reader that the proposed methodology is substantially general.

In Section 2 we will provide the elements of the proposed methodology. To keep technicalities at a minimum, we shall assume that underlying probability models are built on a finite outcome space. The purpose of Section 2 is to concentrate on a conceptual discussion; future sections are designed to illustrate the conceptual elements discussed in Section 2. Hence, the organization of the rest of the paper will be discussed at the end of Section 2. We shall return to the evaluation of C-VaR toward the end of the paper.

2 Basic Distributionally Robust Problem Formulation

In a typical application of stochastic modeling, one is interested in computing the expected value of some performance measure that is a function of underlying risk factors. In particular, let us say that we are interested in estimating $E_{true}\left(h\left(X\right)\right)$, where X is a random variable (or random vector)

taking values in R^d, and $h : R^d \to R$ is a performance measure of interest. The notation $E_{true}(\cdot)$ denotes the expectation operator associated with an underlying "true" or correct probabilistic model, which is unknown to the actuary.

To have a concrete example to guide our discussion, let us assume that X is the time-until-death of an individual seeking to acquire a whole life insurance contract that pays \$1 of benefit to the beneficiaries of the customer at time X. If we assume that the force of interest (continuously compounded) is given by $r > 0$, and we are interested in the expected net present value of the benefits paid, then $h(x) = \exp(-rX)$, and such expected value is given by $E_{true}(\exp(-rX)) = E_{true}(h(X))$. A more realistic example might consider stochastic interest rates or a portfolio of different types of contracts. But let us continue with the current standard example for simplicity.

The first step in estimating $E_{true}(h(X))$ is to approximate the true distribution of X using available information. In the context of our simple example, let us assume that the individual in consideration is 20 years old and that X takes values only on the set $\{1, ..., 100\}$, meaning that the individual could die at ages 21, 22,..., 120 and that the benefit is paid in integer years. So, once a baseline probability distribution $p_0(k) = P_0(X = k)$ has been calibrated using some procedure, the actuary uses the estimate

$$E(\exp(-rX)) \approx \sum_{k=1}^{100} p_0(k) \exp(-rk),$$

and depending on the procedure used to calibrate $p_0(\cdot)$, a confidence interval can be developed to obtain $E(h(X))$.

Now, imagine a situation in which we know that the individual has a medical condition for which not much is known, so there might be substantial variability in our estimation of the distribution of X. Or perhaps there is a significant likelihood that medical advances might be expected to occur in the next 10 years or so, and therefore, if $p_0(\cdot)$ were calibrated based on historical data, $p_0(\cdot)$ might simply be an incorrect model.

To quantify model error, we propose several alternative methods. The first involves solving the following pair of maximization and minimization problems, which we shall call the basic distribu-

tionally robust (BDR) problem formulation:

$$\min / \max \sum_{k=1}^{100} p(k)\, h(k) \tag{1}$$

s.t.

$$\sum_{k=1}^{100} p(k) \log\left(\frac{p(k)}{p_0(k)}\right) \leq \delta\,, \tag{2}$$

$$\sum_{k=1}^{100} p(k) = 1, \quad p(k) \geq 0 \text{ for } k \geq 1\,,$$

where δ should be suitably calibrated (chosen as small as possible) to guarantee that

$$\sum_{k=1}^{100} p_{true}(k) \log\left(\frac{p_{true}(k)}{p_0(k)}\right) \leq \delta\,. \tag{3}$$

The reason for such a selection of δ is the following. First, by choosing δ so that inequality (3) holds, we guarantee that $\{p_{true}(k) : 1 \leq k \leq 100\}$ is in the feasible region described by (2). Second, δ should be chosen as small as possible so that the interval obtained by computing the minimum and maximum in (1) is small. In the end, the actuary will provide a robustified interval (corresponding to the minimum and maximum solutions, respectively) that is guaranteed to contain $E_{true}(h(X))$, assuming that δ is properly chosen as in (3).

The BDR problem (1) can be postulated in great generality. The left side in (2) is known as the Kullback-Leibler (KL) divergence [18] and is denoted by $D(p(\cdot)\,\|p_0(\cdot))$. The KL divergence plays an important role in information theory and statistics (with connections to concepts such as entropy [7] and Fisher information [8]; it is also known as the relative entropy, a term we will use interchangeably in this paper), and it has been substantially studied. In general, given two probability distributions, $P(\cdot)$ and $P_0(\cdot)$, of the random element X, not necessarily supported on a discrete set of points, we have that

$$D(P\|P_0) = E_P\left(\log\left(\frac{dP}{dP_0}(X)\right)\right) = \int \log\left(\frac{dP}{dP_0}(x)\right) dP(x)\,,$$

where the integral is taken over the region in which X takes its values and dP/dP_0 is the likelihood ratio between the probability models P and P_0. Virtually all of what we will discuss about (1) gen-

eralizes to non-discrete, even infinite dimensional outputs (e.g., when x is, for instance, a stochastic process such as Brownian motion). However, since we want to avoid technicalities, we will continue our discussion in the setting of (1).

Let us briefly discuss two key properties of the KL divergence that make the formulation (1) appealing. First, by Jensen's inequality, we have that

$$D(p(\cdot)\,\|p_0(\cdot)) = -\sum_{k=1}^{100} p(k)\log\left(\frac{p_0(k)}{p(k)}\right) \geq -\log\left(\sum_{k=1}^{100} p(k)\,\frac{p_0(k)}{p(k)}\right) = 0,$$

and $D(p(\cdot)\,\|p_0(\cdot)) = 0$ if and only if $p(k) = p_0(k)$ for all k. In other words, the $D(\cdot)$ allows us to compare the discrepancies between any two models, and the models agree if and only if there is no discrepancy. That is, in principle, we can include any distribution $\{p(k) : 1 \leq k \leq 100\}$ in the feasible region, and this non-parametric feature is, we believe, crucial when trying to quantify model errors.

It should be noted that $D(\cdot)$ is not a distance in the mathematical sense, because it does not obey the triangle inequality. However, and this is the second key property of the KL divergence, $D(p(\cdot)\,\|p_0(\cdot))$ is a convex function of $p(\cdot)$. To see this, first note that the function $l(x) = x\log(x)$ is convex. Second, observe that if $\{p(k) : 0 \leq k \leq 100\}$ and $\{q(k) : 0 \leq k \leq 100\}$ are two probability distributions and $\alpha \in (0,1)$, then, because of Jensen's inequality, for any fixed k,

$$l\left(\frac{\alpha p(k) + (1-\alpha)\,q(k)}{p_0(k)}\right) \geq \alpha l\left(\frac{p(k)}{p_0(k)}\right) + (1-\alpha)\,l\left(\frac{q(k)}{p_0(k)}\right).$$

Thus, we conclude (multiplying and dividing by $p_0(k)$ inside the summation defining the KL divergence),

$$\begin{aligned} D(\alpha p(\cdot) + (1-\alpha)\,q(\cdot)\,\|p_0(\cdot)) &= \sum_{k=1}^{100} p_0(k)\,l\left(\frac{\alpha p(k) + (1-\alpha)\,q(k)}{p_0(k)}\right) \\ &\geq \sum_{k=1}^{100} \alpha l\left(\frac{p(k)}{p_0(k)}\right) + (1-\alpha)\,l\left(\frac{q(k)}{p_0(k)}\right) \\ &\geq \alpha D(p(\cdot)\,\|p_0(\cdot)) + (1-\alpha)\,D(p(\cdot)\,\|p_0(\cdot)). \end{aligned}$$

Consequently, BDR problem (1) is a convex optimization problem with a linear objective func-

tion. These types of problems have been studied substantially in the operations research literature, which makes the proposed formulation computationally tractable. And this is a key point for any optimization approach that attempts to quantify model error in a non-parametric way.

We shall later in the paper expand on the implications of choosing the KL divergence as a notion of discrepancy between a chosen baseline model (p_0) and feasible models from which to choose to obtain bounds on the performance measure of interest. In particular, we will see that other discrepancy measures also can be used, but the KL divergence, in the context of tail risk modeling, tends to be a safe and conservative choice.

There are direct variations of the BDR problem formulation that can be easily accommodated within the same convex optimization framework. For instance, suppose that the actuary is less confident about the estimate for $p_{true}(k)$ for small values of k; that is, assume that $p_0(k) \approx p_{true}(k)$ for large values of k, but the actuary is uncertain about how similar $p_0(k)$ is to $p_{true}(k)$ for small values of k. Then we can replace the inequality constraint (2) by introducing a weighting function $\{w(k) : 1 \leq k \leq 100\}$ with $w(k) > 0$ which is *increasing,* thus obtaining

$$\sum_{k=1}^{100} w(k)\, p(k) \log\left(\frac{p(k)}{p_0(k)}\right) \leq \delta\ .$$

To understand why $w(\cdot)$ should be increasing in a setting in which we wish to quantify the impact of model errors arising due to misspecification of $p_0(k)$ for small values of k, think of the case in which $w(k) = \varepsilon > 0$ for $k \leq k_0$ for some $k_0 > 0$ and ε small. Also assume that $w(k) = 1$ for $k > k_0$. Then observe that the constraint (2) is relatively insensitive to the value of $p(k)$ for $k \leq k_0$; therefore, the convex optimization program will have more freedom to optimize the objective function without having a significant impact on feasibility.

Another variation of the BDR formulation includes moment constraints. For instance, let us assume that substantial additional information is known for the expected time-until-death for individuals who have a particular underlying medical condition and are at least 30 years old. For example, one may gather such information from a series of medical studies. Using such information one might impose a constraint of the form $E(X|X \geq 30) \in [a_-, a_+]$ for a specified range $[a_-, a_+]$; equivalently, $E(XI(X \geq 30)) \in [P(X \geq 30)a_-, P(X \geq 30)a_+]$. (Throughout the paper, we use

$I(A)$ to denote the indicator of A, that is, $I(A) = 1$ if A occurs, and $I(A) = 0$ if A does not occur.) Using this information, one can add the constraints

$$a_- \sum_{k=30}^{100} p(k) \leq \sum_{k=30}^{100} p(k)\, k \leq a_+ \sum_{k=30}^{100} p(k), \tag{4}$$

which are linear inequalities, and therefore, the optimization problem is still of convex form and also admits a closed form solution. In Section 9 we will discuss additional constraints that can inform the BDR formulation with other forms of expert knowledge.

At this point, several questions might be in order: How do we solve the optimization problem BDR and its variations? How can we understand such solution intuitively? What is the role of the constraint (2)? How do we choose δ? How do we extend the methodology to deal with continuous, possibly multidimensional distributions? Our goal is to address these questions throughout the rest of this paper.

3 Solving the BDR Formulation

Our goal in this section is to describe how to solve (1) and use this explanation to transition toward the most general version of (1) in an intuitive way. We concentrate on the problem of maximization; the minimization counterpart is analogous, and we will summarize the differences at the end of our discussion.

3.1 The Maximization Form

Let us now introduce Lagrange multipliers to solve the convex optimization problem (1). The Lagrangian takes the form

$$g(p(1), ..., p(100), \lambda_1, \lambda_2) = -\sum_{k=1}^{100} p(k)\, h(k) + \lambda_1 \left(\sum_{k=1}^{100} p(k) \log \left(\frac{p(k)}{p_0(k)} \right) - \delta \right) + \lambda_2 \left(\sum_{k=1}^{100} p_0(k) - 1 \right).$$

Note the negative sign multiplying the objective function, $h(\cdot)$. It has been introduced to transform the problem in standard minimization form over a convex domain. The KKT (Karush-Kuhn-Tucker) [5] conditions in our (convex optimization) setting characterize the optimal solution. The

KKT conditions for the optimal selection, $\{p_+(k) : 1 \leq k \leq 100\}$, λ_1^+ and λ_2^+ in (1), are as follows (we use "+" to denote the optimal solution for the maximization formulation, as opposed to "−" for the minimization counterpart, which we shall discuss momentarily as well):

I) Stationarity.

$$0 = \frac{\partial g}{\partial p(k)}\left(p_+(1), ..., p_+(100), \lambda_1^+, \lambda_2^+\right) = -h(k) + \lambda_1^+\left(1 + \log\left(\frac{p_+(k)}{p_0(k)}\right)\right) + \lambda_2^+. \tag{5}$$

II) Primal feasibility. We must have that

$$\sum_{k=1}^{100} p_+(k) \log\left(\frac{p_+(k)}{p_0(k)}\right) \leq \delta, \quad \sum_{k=1}^{100} p_+(k) = 1, \quad p_+(k) \geq 0 \text{ for } 1 \leq k \leq 100.$$

III) Dual feasibility. We must have that $\lambda_1^+ \geq 0$ (due to the ≤ 0 form in the constraint associated with λ_1^+). The sign of λ_2^+ is free.

IV) Complementary slackness condition. At the optimum, we must have the complementary slackness condition

$$\lambda_1^+\left(\sum_{k=1}^{100} p_+(k) \log\left(\frac{p_+(k)}{p_0(k)}\right) - \delta\right) = 0.$$

Using (5), and writing $1/\lambda_1^+ = \theta_+ > 0$ and $\lambda_2^+/\lambda_1^+ + 1 = \psi_+$ to make a connection to natural exponential families more evident, we conclude that

$$p_+(k) = p_0(k) \exp\left(\theta_+ h(k) - \psi_+\right). \tag{6}$$

Note that $p_+(\cdot)$ is a member of a "natural exponential family,"also known as "exponential tilting" distribution. These types of distributions arise often when one is guessing the form of optimal solutions in BDR problem formulations based on the KL divergence.

The case $\lambda_1^+ = 0$ (or equivalently $\theta_+ = \infty$) is understood as a limit, and we will show in Case 1 below that such limit corresponds to the conditional distribution X, given that X lies on the set of maximizers of $h(\cdot)$. Enforcement of the constraint that $p_+(\cdot)$ must be a probability mass function

requires that for $\theta_+ > 0$ we must pick $\psi_+ := \psi(\theta_+)$ so that

$$\sum_{k=1}^{100} p_0(k) \exp(\theta_+ h(k)) = \exp(\psi(\theta_+)). \tag{7}$$

Note that this satisfies the second and third conditions in primal feasibility. Moreover, to enforce complementary slackness, we must have that

$$\lambda_1^+ \left(\theta_+ \frac{\sum_{k=1}^{100} p_0(k) \exp(\theta_+ h(k)) h(k)}{\sum_{j=1}^{100} p_0(j) \exp(\theta_+ h(j))} - \log \left(\sum_{k=1}^{100} p_0(j) \exp(\theta_+ h(j)) \right) - \delta \right) = 0.$$

Therefore, we have two cases to consider, namely, $\lambda_1^+ = 0$ and $\lambda_1^+ > 0$.

3.2 Case 1: $\lambda_1^+ = 0$

Let us first consider the case $\lambda_1^+ = 0$, which implies that $\theta_+ = \infty$. Let us write $p_{+,\infty}(\cdot)$ to denote the degenerate distribution corresponding to $\theta_+ = \infty$. To characterize this distribution, define

$$\mathcal{M}_+ = \{k : h(k) = \max(h(j) : 1 \leq j \leq 100)\}.$$

In simple words, $\mathcal{M}_+$ is the set on which $h(\cdot)$ achieves its maximum value. So if $k \in \mathcal{M}_+$, we have that

$$\lim_{\theta \to \infty} \exp(\theta(h(j) - h(k))) = \begin{cases} 0 & \text{if} \quad j \notin \mathcal{M}_+ \\ 1 & \text{if} \quad j \in \mathcal{M}_+ \end{cases}.$$

Consequently, from (6) and (7) we obtain that

$$p_{+,\infty}(k) = \lim_{\theta \to \infty} p_0(k) \frac{\exp(\theta h(k))}{\sum_{j=1}^{100} p_0(j) \exp(\theta h(j))} = \frac{p_0(k) I(k \in \mathcal{M}_+)}{\sum_{j \in \mathcal{M}_+} p_0(j)} = P_0(X = k \mid X \in \mathcal{M}_+).$$

In simple words, $p_{+,\infty}(\cdot)$ is the conditional distribution of X, given $X \in \mathcal{M}_+$ under the model $p_0(\cdot)$, which clearly (if feasible!) would be optimal. Now, to verify primal feasibility, we must have that

$$\sum_{k=1}^{100} p_{+,\infty}(k) \log \left(\frac{p_{+,\infty}(k)}{p_0(k)} \right) = \log \left(\frac{1}{P_0(X \in \mathcal{M}_+)} \right) \leq \delta. \tag{8}$$

So if inequality (8) holds, then we have that $p_+(k) = p_{+,\infty}(k)$ because the KKT conditions I to IV hold. Inequality (8), however, is a degenerate case and will rarely be satisfied in practice, so let us consider now the second case.

3.3 Case 2: $\lambda_1^+ > 0$

If inequality (8) does not hold, then $p_{+,\infty}(\cdot)$ is not feasible and therefore $\theta_+ \in (0, \infty)$ must be chosen to enforce dual feasibility. In particular, to enforce complementary slackness, we most compute θ_+ so that

$$\theta_+ \frac{\sum_{k=1}^{100} p_0(k) \exp(\theta_+ h(k)) h(k)}{\sum_{j=1}^{100} p_0(j) \exp(\theta_+ h(j))} - \log\left(\sum_{k=1}^{100} p_0(j) \exp(\theta_+ h(j))\right) = \delta. \tag{9}$$

There exists a unique solution $\theta_+ > 0$ satisfying the previous equation (9). The reason for the existence is that here, in the convex optimization problem (1), the Slater condition (see [5]) is satisfied and therefore the KKT conditions are necessary and sufficient for optimality. Moreover, θ_+ satisfying equation (9) is unique because the left side of (9) is increasing and continuous as a function of $\theta_+ > 0$ (continuity is immediate, and monotonicity can be verified by somewhat tedious but elementary differentiation). Thus, using Newton's method or a line search to solve for θ_+, we obtain the solution given by (6) and (7).

3.4 The Minimization Form

The minimization form of the BDR problem is analogous to the maximization case. In particular, in this case, the Lagrangian takes the form

$$g(p(1), ..., p(100), \lambda_1, \lambda_2) = \sum_{k=1}^{100} p(k) h(k) + \lambda_1 \left(\sum_{k=1}^{100} p(k) \log\left(\frac{p(k)}{p_0(k)}\right) - \delta\right) + \lambda_2 \left(\sum_{k=1}^{100} p_0(k) - 1\right).$$

Note that the sign in the objective function is now positive. The KKT conditions are the same. A careful reading of our previous discussion yields two cases analogous to those described in the previous subsection.

Case 1: Define

$$\mathcal{M}_- = \{k : h(k) = \min(h(k) : 1 \le k \le 100)\}.$$

Then if $\log\left(1/P_0\left(X \in \mathcal{M}_-\right)\right) \leq \delta$, we have that the optimal solution $p_{-,\infty}\left(\cdot\right)$ is given by

$$p_{-,\infty}\left(k\right) = P_0\left(X = k | X \in \mathcal{M}_-\right).$$

Case 2: If $\log\left(1/P_0\left(X \in \mathcal{M}_-\right)\right) > \delta$, then the optimal solution to the minimization BDR formulation is given by $p_-\left(\cdot\right)$, defined via

$$\begin{aligned} p_-\left(k\right) &= p_0\left(k\right)\exp\left(-\theta_- h\left(k\right) - \psi\left(\theta_-\right)\right), \\ \exp\left(\psi\left(\theta_-\right)\right) &= \sum_{k=1}^{100} p_0\left(k\right)\exp\left(-\theta_- h\left(k\right)\right), \end{aligned}$$

with $\theta_- > 0$ satisfying

$$-\theta_- \frac{\sum_{k=1}^{100} p_0\left(k\right)\exp\left(-\theta_- h\left(k\right)\right) h\left(k\right)}{\sum_{k=1}^{100} p_0\left(j\right)\exp\left(-\theta_- h\left(j\right)\right)} - \log\left(\sum_{k=1}^{100} p_0\left(j\right)\exp\left(-\theta_- h\left(j\right)\right)\right) = \delta.$$

3.5 Numerical Example of BDR Formulation

To illustrate numerically the BDR formulation (1), we consider the estimation of the expected payoff $h(X) = \exp(-rX)$, where X is the time-until-death of a policyholder at age 20 years. We set the baseline distribution $p_0(k)$ from the recent static mortality table under the Internal Revenue Code ([2], Appendix, "unisex" column). We are interested in investigating the estimates of the expected payoff if the true mortality distribution deviates from $p_0(k)$.

Figure 1 shows the maximum and minimum values of (1) under different δ, when the force of interest is set at $r = .015$. The maximum value increases with δ, while the minimum value decreases with δ, both in a concave manner. The effects on the maximum values seem to be stronger than the minimum values when the distribution changes from the baseline, as shown by a larger magnitude of the upward movement as δ increases.

Figure 2 shows the shapes of the distributions giving rise to the maximum and minimum expected payoffs, at $\delta = 2$, compared with the baseline distribution. We can see a sharp concentration of mass close to 0 for the maximal distribution, as higher chance of sooner death leads to a larger expected present value of payoff. In contrast, more mass is located toward older ages for the minimal

distribution, leading to a reduction in the expected present value of payoff.

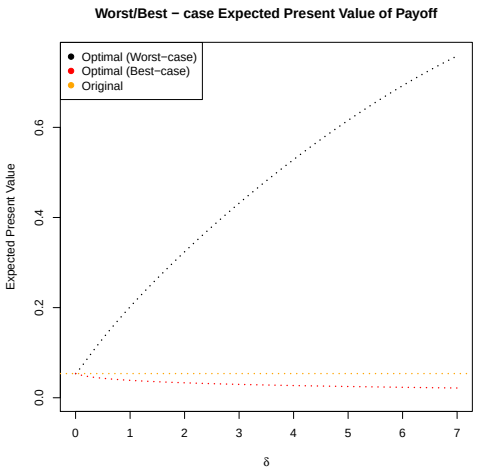

Figure 1: Robust estimates against δ

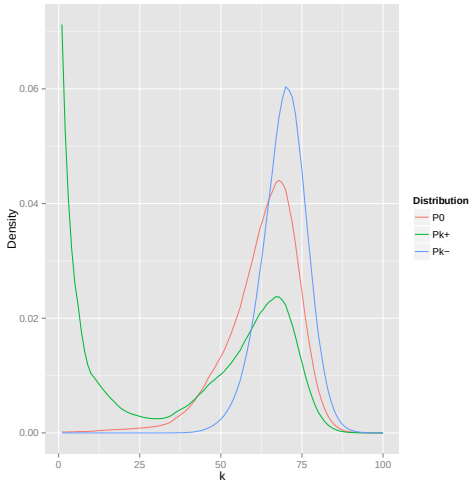

Figure 2: Optimal probability mass function at $\delta = 2$

4 Distributionally Robust Analysis of Dependence

Consider now the following variation of the BDR formulation which is suitable for quantifying the impact of dependence. Again, we revisit the case of whole life insurance, but this time we assume that a couple is interested in a contract that pays \$1 of benefit at the time of the first death among the couple. In this case, the payoff equals $h(X,Y) = \exp(-r\min(X,Y))$, where X and Y are the time-until-death of the individual and his or her spouse, respectively. Let us assume that both spouses are 20 years old at the time of signing the risk premium valuation, so we will assume that the pair (X,Y) is supported on the set $\{1,...,100\} \times \{1,...,100\}$.

We are interested in the potential impact on the estimated actuarial net present value of the benefits due to unaccounted dependence structures. Suppose that, this time, enough information is available to estimate the distributions of X and Y marginally with relatively high accuracy, but the joint distribution is difficult to estimate. A joint baseline distribution might be calibrated, leading to the model $p_0(i,j) = P_0(X = i, Y = j)$. We assume, for the purpose of illustrating the analysis

of dependence structures, that the marginals are known—that is,

$$\begin{aligned} P_{true}\left(Y=j\right) &= \sum_{i=1}^{100} P_0\left(X=i,Y=j\right), \\ P_{true}\left(X=i\right) &= \sum_{j=1}^{100} P_0\left(X=i,Y=j\right). \end{aligned}$$

Then the distributionally robust (maximization) formulation is given by

$$\begin{aligned} &\max \sum_{i=1}^{100}\sum_{j=1}^{100} p\left(i,j\right) h\left(i,j\right) \\ &\text{s.t.} \\ &\sum_{i=1}^{100}\sum_{j=1}^{100} p\left(i,j\right) \log\left(\frac{p\left(i,j\right)}{p_0\left(i,j\right)}\right) \leq \delta\ , \\ &\sum_{j=1}^{100} p\left(i,j\right) = P_0\left(X=i\right) \text{ for all } i\ , \quad \sum_{i=1}^{100} p\left(i,j\right) = P_0\left(Y=j\right) \text{ for all } j\ , \\ &p\left(i,j\right) \geq 0 \text{ for all } i,j\ . \end{aligned} \tag{10}$$

We concentrate on the case analogous to Case 2 analyzed in the BDR formulation. This is the most interesting case in practice, because this is the one in which the relative entropy constraint, involving δ, is satisfied with equality at the optimum.

It is worth mentioning that $p = p_0$ is a feasible solution, and therefore we can verify that the Slater condition (see [5] and recall Section 3.3) is satisfied. Such condition guarantees that strong duality holds and that the KKT conditions are necessary and sufficient for optimality. Commercial packages can be used to solve a convex optimization problem such as (10) very quickly. Here, we will describe an iterative procedure that is based on using an exponential tilting form similar to that in problem (1). We guess, using the stationarity KKT condition, that the optimal solution must take the form

$$p_+\left(i,j\right) = P_0\left(X=i,Y=j\right)\exp\left(\theta_+ h\left(i,j\right) - \alpha_{\theta_+}\left(i\right) - \beta_{\theta_+}\left(j\right)\right),$$

where

$$\sum_j \exp\left(\theta_+ h\left(i,j\right) - \beta_{\theta_+}\left(j\right)\right) P_0\left(X=i, Y=j\right) = \exp\left(\alpha_{\theta_+}\left(i\right)\right) P_0\left(X=i\right), \tag{11}$$

$$\sum_i \exp\left(\theta_+ h\left(i,j\right) - \alpha_{\theta_+}\left(i\right)\right) P_0\left(X=i, Y=j\right) = \exp\left(\beta_{\theta_+}\left(j\right)\right) P_0\left(Y=j\right). \tag{12}$$

These equations can be solved efficiently using the following iterative scheme (see, e.g., [10, 13]): First, pick $\alpha^0_{\theta_+}$. At the kth iteration, for $k \geq 0$, compute

$$\exp\left(\bar{\alpha}^k_{\theta_+}\left(i\right)\right) = \sum_j \frac{\exp\left(\theta_+ h\left(i,j\right)\right) p_0\left(i,j\right) P_0\left(Y=j\right)}{\sum_l \exp\left(\theta_+ h\left(l,j\right) - \alpha^k_{\theta_+}\left(l\right)\right) p_0\left(l,j\right) P_0\left(X=i\right)},$$

and renormalize to obtain $\alpha^{k+1}_{\theta_+}\left(i\right)$ via the equation

$$\exp\left(-\alpha^{k+1}_{\theta_+}\left(i\right)\right) = \frac{\exp\left(-\bar{\alpha}^k_{\theta_+}\left(i\right)\right)}{\sum_i \sum_j p_0\left(i,j\right) \exp\left(\theta_+ h\left(i,j\right) - \bar{\alpha}^k_{\theta_+}\left(i\right)\right) / \left(\sum_l \exp\left(\theta_+ h\left(l,j\right) - \alpha^k_{\theta_+}\left(l\right)\right) p_0\left(l,j\right) / P_0\left(Y=j\right)\right)}.$$

Similar iterations are available for β_{θ_+}, namely, given $\beta^0_{\theta_+}$, evaluate at the kth iteration,

$$\exp\left(\bar{\beta}^k_{\theta_+}\left(j\right)\right) = \sum_i \frac{\exp\left(\theta_+ h\left(i,j\right)\right) p_0\left(i,j\right) P_0\left(X=i\right)}{\sum_l \exp\left(\theta_+ h\left(i,l\right) - \beta^k_{\theta_+}\left(i\right)\right) p_0\left(i,l\right) P_0\left(Y=j\right)},$$

and renormalize to obtain

$$\exp\left(-\beta^{k+1}_{\theta_+}\left(i\right)\right) = \frac{\exp\left(-\bar{\beta}^k_{\theta_+}\left(i\right)\right)}{\sum_i \sum_j p_0\left(i,j\right) \exp\left(\theta_+ h\left(i,j\right) - \bar{\beta}^k_{\theta_+}\left(i\right)\right) / \left(\sum_l \exp\left(\theta_+ h\left(l,j\right) - \beta^k_{\theta_+}\left(l\right)\right) p_0\left(l,j\right) / P_0\left(Y=j\right)\right)}.$$

Then θ_+ can ultimately be chosen so that

$$\begin{aligned}\delta &= \theta_+ E\left(\exp\left(\theta_+ h\left(X,Y\right) - \alpha_{\theta_+}\left(X\right) - \beta_{\theta_+}\left(X\right)\right) h\left(X,Y\right)\right) \\ &\quad - E\left(\exp\left(\theta_+ h\left(X,Y\right) - \alpha_{\theta_+}\left(X\right) - \beta_{\theta_+}\left(Y\right)\right)\left(\alpha_{\theta_+}\left(X\right) + \beta_{\theta_+}\left(Y\right)\right)\right),\end{aligned}$$

by performing a line search procedure.

4.1 Numerical Example on Joint Mortality

We illustrate numerically the solution to optimization (10). Consider, as discussed above, the estimation of the expected payoff $h(X, Y) = \exp(-r \min(X, Y))$, where X and Y are the times-until-death of a couple, both at 20 years old. We set the baseline joint distribution $p_0(i, j)$ from the same mortality table used in Section 3.5, but now assuming independence of the times-until-death of the couple, i.e. $p_0(i, j) = P_0(X = i)P_0(Y = j)$. This independent baseline model could describe the simplest assumption made by insurers when pricing life products, and we are interested in a robust estimate of the expected payoff when this assumption is violated.

Figure 3 shows the optimal values of (10) under different values of δ, setting $r = .015$. As depicted, there is a concavely increasing trend for the worst-case expected payoff as the level of dependency deviates from independence. While the estimate is 0.0729 according to the independent model, it can potentially rise to 0.0734 when the true model is misspecified from being independence to a level of $\delta = 0.002$.

Moreover, Figures 4 and 5 further show the approximate surfaces of the optimal joint probability mass function at $\delta = 0.013$ and the original independent joint probability mass function respectively. Their differences are more clearly visualized in Figure 6.

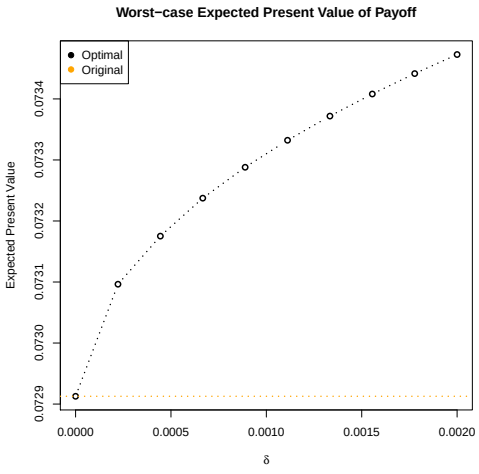

Figure 3: Robust estimate against δ

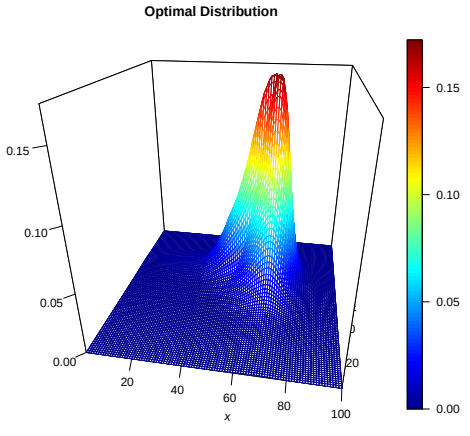

Figure 4: Optimal probability mass function at $\delta = 0.013$

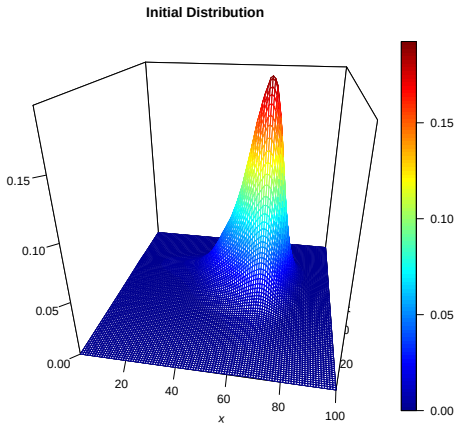

Figure 5: Baseline probability mass function

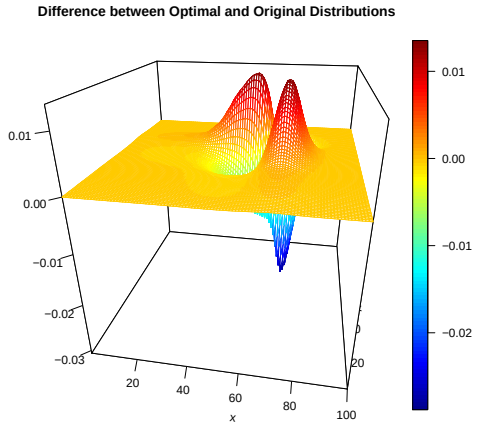

Figure 6: Difference between optimal and baseline probability mass functions

5 General Distributions and Interpretation of Optimal Solutions

In the most general (maximization) form, the BDR formulation is written in an equivalent form that optimizes over all random variables $Z(\omega)$, where ω is an element of the underlying outcome space Ω. Precisely, and using $E_0(\cdot)$ to denote the expectation under the baseline model $P_0(\cdot)$), the general BDR formulation takes the following form:

$$\begin{aligned} &\max E(h(X)) \\ &\text{s.t. } D(P||P_0) \leq \delta \ . \end{aligned} \tag{13}$$

This is equivalent, as we shall explain, to

$$\begin{aligned} &\max E_0(h(X)\, Z) \\ &\text{s.t.} \\ &E_0(Z \log(Z)) \leq \delta \\ &Z(\omega) \geq 0, \text{ for all } \omega, \text{ and } \ E_{P_0}(Z) = 1 \ . \end{aligned} \tag{14}$$

The solution to (14), denoted as $(Z_+(\omega) : \omega \in \Omega)$, can be used to obtain the solution, $P_+(\cdot)$, to (13). It turns out that $P_+(\cdot)$ is defined via

$$P_+(X \in A) = E_0(Z_+ I(X \in A)).$$

Conceptually, the most general problem formulation is not much different from the case in which X takes finitely many outcomes. It can be shown (e.g., [24, 6]) that exactly the same form of the optimal solution applies as in the finite case. For instance, if $f_0(\cdot)$ is the density of X in R^d, then

$$\begin{aligned} Z_+(x) &= \frac{f_+(x)}{f_0(x)} = \exp(\theta_+ h(x) - \psi(\theta_+)), \\ \psi(\theta_+) &= \log E(\exp(\theta_+ h(X))), \end{aligned}$$

with

$$\theta_+ \frac{E_0(\exp(\theta_+ h(X))\, h(X))}{E_0(\exp(\theta_+ h(X)))} - \log(E_0(\exp(\theta_+ h(X)))) = \delta$$

in case $\log(1/P_0(X \in \mathcal{M}_+)) > \delta$, where

$$\mathcal{M}_+ = \{x : h(x) = \sup(h(x) : x \in R^d)\}.$$

However, if $\log(1/P_0(X \in \mathcal{M}_+)) \leq \delta$, then the optimal solution to the general BDR problem is given by the probability distribution given by

$$P_{+,\infty}(X \in A) = P_{+,\infty}(X \in A | X \in \mathcal{M}_+).$$

To gain intuition concerning the meaning of the BDR solution, let us consider the important case of performance analysis of rare events.

5.1 Distributionally Robust Performance Analysis for Rare Events

Let us assume that $h(X) = I(X \in B)$ for some set B so that $P_{true}(X \in B)$ is known to be small. We wish to solve the BDR maximization problem

$$\begin{aligned} &\max P(X \in B) \\ &s.t. \;\; D(P||P_0) \leq \delta \,, \end{aligned}$$

where the distribution P_0 of X has been suitably calibrated. Since we are in a rare-event setting, it is reasonable to assume that $\log(1/P_0(X \in B)) > \delta$. Therefore, applying the general solution form of the BDR maximization problem, we know that the optimal solution is given by $P_+(\cdot)$, defined so that for each set A,

$$\begin{aligned} P_+(X \in A) &= \frac{E_0(\exp(\theta_+ I(X \in B))\, I(X \in A))}{E_0(\exp(\theta_+ I(X \in B)))} \\ &= \frac{\exp(\theta_+) P_0(X \in A, X \in B)}{\exp(\theta_+) P_0(X \in B) + P_0(X \notin B)} + \frac{P_0(X \in A, X \notin B)}{\exp(\theta_+) P_0(X \in B) + P_0(X \notin B)}, \end{aligned}$$

for some $\theta_+ \in (0, \infty)$. To obtain a clearer interpretation of the optimal solution, let us write

$$\begin{aligned} P_+(X \in A) &= \frac{(\exp(\theta_+) - 1) P_0(X \in A, X \in B)}{\exp(\theta_+) P_0(X \in B) + P_0(X \notin B)} + \frac{P_0(X \in A)}{\exp(\theta_+) P_0(X \in B) + P_0(X \notin B)} \\ &= \alpha(\theta_+) P_0(X \in A | X \in B) + (1 - \alpha(\theta_+)) P_0(X \in A), \end{aligned}$$

where

$$\alpha(\theta_+) = \frac{(\exp(\theta_+) - 1) P_0(X \in B)}{\exp(\theta_+) P_0(X \in B) + P_0(X \notin B)} > 0$$

and $\theta_+ > 0$ satisfies

$$\theta_+ \frac{\exp(\theta_+) P_0(X \in B)}{(\exp(\theta_+) - 1) P_0(X \in B) + 1} - \log((\exp(\theta_+) - 1) P_0(X \in B) + 1) = \delta. \tag{15}$$

Consequently, in simple words, in the context of distributionally robust performance analysis of rare-event probabilities, the worst-case measure is a (specific) mixture between the conditional distribution of X given that $X \in B$, under the baseline measure (regardless of how it was calibrated)

and the nominal distribution. This is a remarkable insight, because the BDR formulation allows the actuary to bound rare events of interest in terms of any baseline measure chosen. Moreover, note that the optimal solution is given by

$$P_+(X \in B) = \frac{\exp(\theta_+) P_0(X \in B)}{\exp(\theta_+) P_0(X \in B) + P_0(X \notin B)} = \frac{\exp(\theta_+) P_0(X \in B)}{1 + (\exp(\theta_+) - 1) P_0(X \in B)}.$$

But let us continue observing what occurs in (15) when we fix δ and consider an asymptotic analysis as $P_0(X \in B) \to 0$. This asymptotic, namely, fixing δ while sending the probability of the rare event of interest to 0, is actually very relevant from an applied standpoint. It models, in particular, a situation in which we have limited data, a specific parametric model (namely, P_0), and we want to estimate probabilities that are very hard (due to lack of data) to estimate non-parametrically with good relative accuracy.

To perform the asymptotic analysis describing the behavior of $P_+(X \in B)$ as $P_0(X \in B) \to 0$ while fixing $\delta > 0$, let us first assume that

$$\theta_+ \exp(\theta_+) P_0(X \in B) = \eta, \tag{16}$$

for some $\eta > 0$, which will be chosen to remain bounded away from zero as $P_0(X \in B) \to 0$. Ultimately, we will see that $\delta \approx \eta$ as $P_0(X \in B) \to 0$. From (16), we must (since $\eta > 0$ is fixed) have that $\theta_+ \to \infty$ as $P_0(X \in B) \to 0$, and therefore we have $\exp(\theta_+) P_0(X \in B) \to 0$ as $P_0(X \in B) \to 0$. Consequently, letting $x = (\exp(\theta_+) - 1) P_0(X \in B)$ and using (16) in (15), we conclude that

$$\frac{\eta}{1+x} - \log(1+x) = \delta,$$

which implies, after expanding as $x \to 0$, that $\eta = \delta + (1+\delta) x + ...$ The bottom line is that as $P_0(X \in B) \to 0$, we obtain that

$$P_+(X \in B) \approx \frac{\delta}{\log(1/P_0(X \in B))}.$$

To understand the decision-making implications of the BDR formulation, suppose that the potential losses of an insurance company are modeled as X, and under P_0, we have that $P_0(X > t) =$

$\exp(-\Lambda_0(t))$ for some function $\Lambda_0(\cdot)$. In simple words, under P_0, X has cumulative hazard function $\Lambda_0(\cdot)$. Now assume that we wish to compute b so that $P_{true}(X > b) \leq .005$. Suppose that $\delta = .1$, but the model P_{true} is not known. Then, based on the optimal solution of our robust analysis, we must choose b so that

$$P_+(X > b) \approx \frac{\delta}{\Lambda_0(b)} = \frac{.1}{\Lambda_0(b)} \leq .005, \tag{17}$$

which implies that $b \approx \Lambda_0^{-1}(20)$. For example, if X is exponentially distributed with unit mean under P_0, we have that $\Lambda_0(b) = b$; thus, we must have that $b \approx 20$. In contrast, if one trusts the exponential model fully, then one would choose $\exp(-b) = .005$, which yields $b \approx 5.3$.

Paraphrasing, if b is the value of the statutory solvency capital needed to withstand losses with probability at least .995, and if the actuary uses an exponential model with unit mean, then a deviation of .1 units measured in KL divergence between the assumed (exponential) model and the (unknown) reality might result in underestimating the statutory capital by a factor of about $20/5.3 \approx 3.7$.

But another important message to keep in mind is that the BDR formulation might provide very conservative estimates for rare-event probabilities when δ (i.e., the size of the uncertainty) is large relative to the rare-event probability of interest. Once again, if X is exponentially distributed with unit mean under P_0, then (17) indicates that $P_+(X > b) \approx .1/b$, and therefore X has Pareto-type tail behavior under P_+.

The next example is given to develop intuition concerning the shape of the optimal solution density of the BDR formulation.

5.2 The Shape of the Optimal Density: A Simple Example

We are interested in bounding the loss probability $P_{true}(L > b)$ for a given reserve b. The actuary considers a simple loss model of the form $L = X_1 + \cdots + X_n$, where the X_i's follow a multivariate Gaussian distribution under the baseline model P_0—that is, $(X_1, \ldots, X_d) \sim N(\mu, \Sigma)$ for some mean vector μ and covariance matrix Σ—with specific distributional parameters given in the sequel. Consequently,

$$P_0(L > b) = \overline{\Phi}\left(\frac{b - \mathbf{1}'\mu}{\sqrt{\mathbf{1}'\Sigma\mathbf{1}}}\right),$$

where $\mathbf{1}$ is a vector of entries equal to one, and $\overline{\Phi}(\cdot)$ is the tail distribution function of a standard Gaussian random variable.

Now suppose that the modeler is uncertain whether $(X_1, ..., X_d)$ follows a Gaussian model. So we formulate the corresponding BDR problem, namely,

$$\max P(L > b) \text{ subject to } D(P\|P_0) \leq \delta. \tag{18}$$

We first check whether $\log(1/P_0(L > b)) \leq \delta$. This is precisely the analogue of Case 1 in Section 3.1. If indeed $\log(1/P_0(L > b)) \leq \delta$, then the worst-case probability density for L is

$$f_+(x) = \frac{\phi\left((x - \mathbf{1}'\mu)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}\right)}{P_0(L > b)\sqrt{\mathbf{1}'\Sigma\mathbf{1}}} I(\mathbf{1}'x > b),$$

where $\phi(\cdot)$ is the density of a standard Gaussian random variable, and $P_+(L > b) = 1$. To see this, note that $f_+(\cdot)$ is the exact analogue of the solution discussed in Case 1 of Section 3.1; to draw the analogy, note that in our current setting, we have that $h(l) = I(l > b)$, so the optimal solution of the optimization problem then must be given by the conditional distribution of L given that $L > b$ under the model P_0, which is precisely obtained by the density $f_+(\cdot)$.

Otherwise, if $\log(1/P_0(L > b)) < \delta$, then the worst-case probability distribution is given by

$$f_+(x) = \begin{cases} \frac{\exp(\theta_+)\phi\left((x-\mathbf{1}'\mu)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}\right)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}}{\exp(\theta_+)P_0(L>b)+P_0(L\leq b)} & \text{for } \mathbf{1}'x > b \\ \frac{\phi\left((x-\mathbf{1}'\mu)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}\right)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}}{\exp(\theta_+)P_0(L>b)+P_0(L\leq b)} & \text{for } \mathbf{1}'x \leq b \end{cases}, \tag{19}$$

where $\theta_+ > 0$ is the root of the equation

$$\frac{\theta_+ \exp(\theta_+) P_0(L > b)}{\exp(\theta_+) P_0(L > b) + P_0(L \leq b)} - \log(\exp(\theta_+) P_0(L > b) + P_0(L \leq b)) = \delta.$$

Once again, this solution is obtained by applying the same reasoning as in Case 2 in Section 3.1. As indicated earlier, here $h(l) = I(l > b) = I(\mathbf{1}'x > b)$. Therefore, the associated exponential tilting

is given by

$$\begin{aligned} f_+(x) &\propto \phi\left(\left(x-\mathbf{1}'\mu\right)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}\right)\exp\left(\theta_+ h\left(l\right)\right) \\ &\propto \phi\left(\left(x-\mathbf{1}'\mu\right)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}\right)\exp\left(\theta_+\right) I\left(\mathbf{1}'x > b\right) \\ &\quad +\phi\left(\left(x-\mathbf{1}'\mu\right)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}\right) I\left(\mathbf{1}'x \leq b\right), \end{aligned}$$

which coincides with (19). Observe that the worst-case distribution weights the original distribution, thus preserving the shape, and it assigns a uniform (constant) weight to the region where $\mathbf{1}'x > b$ as large as possible but preserves the relative entropy constraint. In turn, the weight to the region $\mathbf{1}'x \leq b$ is enforced by the constraint that $f_+(\cdot)$ ultimately must be a probability density and therefore its integral must be equal to one.

We illustrate numerically for $n = 5$, with means $\mu_1 = 1.92, \mu_2 = 1.42, \mu_3 = 1.13, \mu_4 = 1.80, \mu_5 = 1.54$ (five numbers generated uniformly between 1 and 2) and covariance matrix

$$\Sigma = \begin{bmatrix} 1.11 & -0.21 & -0.42 & -0.72 & 1.30 \\ -0.21 & 4.82 & 0.89 & -0.31 & -1.50 \\ -0.42 & 0.89 & 1.05 & 0.61 & -0.52 \\ -0.72 & -0.31 & 0.61 & 2.58 & -0.45 \\ 1.30 & -1.50 & -0.52 & -0.45 & 1.91 \end{bmatrix}$$

which is randomly generated from a Wishart distribution with 5 degrees of freedom and an identity scale matrix. Setting $b = 10$, the probability $P_0(L > b)$ is now given by 0.23. Figure 7 shows the worst-case density under $\delta = 0.1$ and the baseline density. Note that the worst-case density puts more mass on the right side of b, and less on its left side, in order to boost the likelihood of a big loss.

We also plot the worst-case probability against δ in Figure 8, which shows a concave growth pattern.

To apply distributionally robust analysis, we have to calibrate δ. We discuss this next.

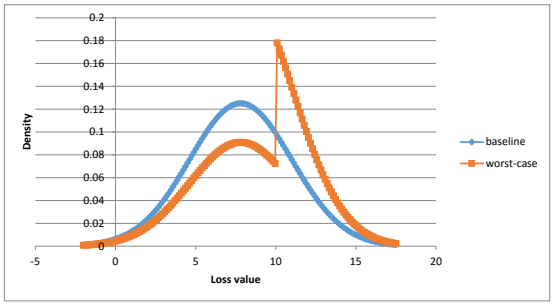

Figure 7: Approximate optimal density under $\delta = 0.1$

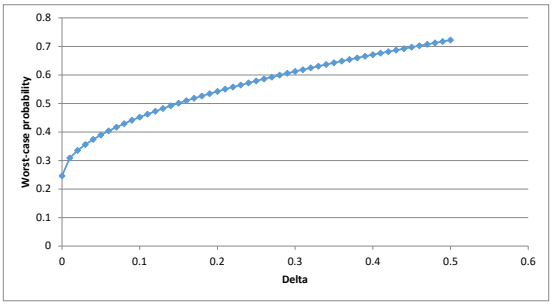

Figure 8: Worst-case probability against δ

6 The Feasible Region in the BDR Formulation

In this subsection, we discuss the selection of the parameter $\delta > 0$. We will discuss two main approaches: One is estimation using historical data. Another is to understand the choice of δ in terms of systematic stress testing.

6.1 Data-Driven Estimation

There are several ways in which one might approach the estimation of δ. For simplicity, we will assume that the true model has a density, $f_{true}(\cdot)$. Suppose that the baseline model also has a density, $f_0(\cdot)$. The most direct approach is to estimate

$$D(P_{true}||P_0) = \int \log\left(\frac{f_{true}(x)}{f_0(x)}\right) f_{true}(x)\, dx.$$

Some proposed procedures and guarantees on convergence analysis can be found in, e.g., [21, 25]. The problem with this approach, however, is that confidence intervals are difficult to obtain, because the rate of convergence of the estimator will depend on the dimension of the underlying density. This is because, indirectly, such a direct approach will attempt to estimate non-parametrically the underlying true density, and this leads to dimension-dependent rates of convergence. Therefore, instead, we use a different approach based on empirical likelihood.

Our starting point is an optimization problem that bears some similarities to the BDR formula-

tion. First, suppose that we observe $X_1, ...X_n$ as an independent and identically distributed (i.i.d.) sample from X and form the empirical probability mass function

$$\mu_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = x),$$

which is uniform on the set $\{X_1, ..., X_n\}$. Now, for any given set of weights $w = (w_1, ..., w_n)$ such that $w_i \geq 0$ and $\sum_{i=1}^{n} w_i = 1$, define

$$v_n(x, w) = \sum_{i=1}^{n} w_i I(X_i = x).$$

In simple words, $v_n(\cdot, w)$ is a probability mass function that assigns probability w_i to the value X_i. Then consider the function

$$\mathcal{R}_n(\gamma) = \min\left\{ D(v_n(\cdot, w)\,||\mu_n(\cdot)) : E_{v_n(\cdot,w)}(h(X)) = \sum_{x\in\{X_1,...,X_n\}} v_n(x)\,h(x) = \sum_{i=1}^{n} w_i h(X_i) = \gamma \right\}.$$

It turns out, under the null hypothesis that $E_{true}(h(X)) = \gamma$ and other mild conditions (including $Var_{true}(h(X)) < \infty$), we have

$$2n\mathcal{R}_n(\gamma) \Rightarrow \chi^2_1.$$

That is, under the null hypothesis, $2n\mathcal{R}_n(\gamma)$ follows approximately a chi-squared distribution with one degree of freedom. This result is classical in the theory of empirical likelihood [23]. To make our discussion as self-contained as possible, we provide formal derivation of the result in the appendix at the end of this section.

The connection with the BDR problem formulation can be established as follows. Assume P_0 is built from the empirical distribution function of the observed data; that is, we use $\mu_n(\cdot)$ as the distribution of X under the model P_0. If we knew the specific value $\gamma = E_{true}(h(X))$, it would make sense to choose $\delta \geq \mathcal{R}_n(\gamma)$ to solve the BDR problem, because the optimal solution, P^n_+, of

the corresponding empirical BDR problem, namely

$$\begin{aligned} \max \quad & \sum_{x \in \{X_1,\ldots,X_n\}} v_n(x)\, h(x) \\ \text{s.t.} \quad & D\left(v_n(\cdot, w) \,||\mu_n(\cdot)\right) \leq \delta \,, \end{aligned} \tag{20}$$

will yield the optimal value, $E_+^n\left(h\left(X\right)\right) = \sum_{k=1}^n w_k^+ h\left(X_k\right) \geq \gamma$. So, as $n \to \infty$, the BDR formulation will result in a correct upper bound for $E_{true}\left(h\left(X\right)\right)$.

Consequently, if one chooses P_0, based on the empirical distribution, it is reasonable to select δ so that

$$P\left(\delta > \mathcal{R}_n\left(\gamma\right)\right) \approx P\left(2n\delta > \chi_1^2\right) = .95, \tag{21}$$

thus providing a 95% confidence upper bound for γ.

6.2 A Systematic Approach to Stress Testing

In the absence of enough data to perform a non-parametric calibration of P_0 (and thus of δ) as suggested in the previous section, we suggest using the BDR formulation as an approach to perform systematic stress testing. The idea is to understand how the solution of the BDR optimization problem performs as we vary δ.

In practice—for instance, when computing C-VaR—it is customary to incorporate corrections to C-VaR estimates (i.e., robustify the C-VaR evaluation) by applying arbitrary shocks into the system (i.e., stressing the system by assuming that extreme events occur). However, such shocks are typically selected in an arbitrary way. Moreover, in the presence of multiple shocks affecting different risk factors, it might be difficult to argue that a particular combination of shocks is more reasonable than an alternative combination of shocks.

In contrast, the approach that we study here can be used to robustify more systematically. Suppose that an insurance company has a given baseline model, P_0, which has been calibrated using a combination of past observations and expert knowledge. The model P_0 is used to compute a given risk measure (or performance measure), say, C-VaR. Periodically, either by internal procedures or due to regulatory constraints, the company is requested to perform stress testing. The result of

these stress tests is the determination of capital requirement, which might typically be higher than the C-VaR obtained under P_0.

Let us assume that such a stress-testing procedure has occurred multiple times in the company—say, n times—using the customary approach described before. So the company has built certain experience of the increase in the capital requirement relative to the C-VaR determined by the stress-testing procedure. Based on this experience, one could calibrate values $\delta_1, ..., \delta_n$ so that the solution to the BDR formulation matches the increased capital requirement determined during the n sessions of stress testing. This experience eventually will allow not only the selection of δ, but also the adoption of a sequence of optimal solutions $P_+^1, ..., P_+^n$ that could provide insights in validating the shocks.

Ultimately, by simply choosing δ appropriately, the BDR formulation automatically takes care of inducing the shocks that can potentially cause the highest damage, subject to the constraint of being consistent with the baseline model formulation, P_0, to a certain extent (quantified by the relative entropy and the parameter δ).

7 Simulation-Based Solution Procedure

Many of the optimization problems that we consider might be difficult to solve in the sense that the optimal model P_+ might not be tractable in closed form, especially in the context of general distributions. Therefore, one might need to resort to stochastic simulation in order to solve the optimization problem at hand. The formulation is quite simple and takes the following form. Assume that P_0 is not analytically tractable or that we simply do not have access to a closed-form expression of the density of X under P_0, but we can simulate i.i.d. samples $X_1, ..., X_n$ of X from P_0.

We are interested in estimating $E_{true}\left(h\left(X\right)\right)$, but because P_0 might be incorrect, we use the BDR formulation (14). Then the corresponding empirical version of (14) takes the form given in

(20), which we write explicitly here as

$$\max \sum_{i=1}^{n} h\left(X_i\right) w_i$$
$$\text{s.t.} \quad \sum_{i=1}^{n} w_i \log\left(n w_i\right) \leq \delta \ , \quad \sum_{i=1}^{n} w_i = 1, \ w_i \geq 0 \text{ for all } 1 \leq i \leq n.$$

To illustrate the application of this approach, we include the following example.

7.1 Simulation-Based BDR Formulation: *t*-Copula Baseline Model

In connection with the example given in Section 5.2, suppose that the loss is given by $L = X_1 + \cdots + X_d$ where $(X_1, \ldots, X_d)$ have Gaussian marginal distributions $X_i \sim N(\mu_i, \sigma_i^2)$ and the dependency is modeled by a t-copula. A t-copula is a multivariate distribution denoted by

$$C^t_{\nu,\varrho}(u_1, \ldots, u_n) = \mathbf{t}_{\nu,\varrho}\left(t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_d)\right), \qquad (u_1, \ldots, u_d) \in (0,1)^d,$$

where ν is the degrees of freedom; $\varrho = (\rho_{ij})$ is a positive definite dispersion (or scale) matrix; $\mathbf{t}_{\nu,\varrho}$ is the joint distribution function of a d-dimensional t distribution with degrees of freedom ν, mean $\mathbf{0}$, and dispersion matrix ϱ; and t_ν^{-1} is the quantile function of a standard univariate t distribution with degrees of freedom ν.

Then the distribution function of $(X_1, \ldots, X_n)$ is

$$f(x_1, \ldots, x_n) = C^t_{\nu,\varrho}(\Phi_{\mu_1,\sigma_1^2}(x_1), \ldots, \Phi_{\mu_d,\sigma_d^2}(x_d)),$$

where $\Phi_{\mu,\sigma^2}(\cdot)$ denotes the distribution function of $N(\mu, \sigma^2)$.

It can be difficult to evaluate $P(L > b)$ in closed form, and one can resort to stochastic simulation. An unbiased estimate of $P_0(L > b)$ can be obtained by outputting $L = \Phi^{-1}_{\mu_1,\sigma_1^2}(t_\nu(Z_1)) + \cdots + \Phi^{-1}_{\mu_d,\sigma_d^2}(t_\nu(Z_d))$, where $(Z_1, \ldots, Z_d)$ is drawn from the multivariate t distribution $\mathbf{t}_{\nu,\varrho}$. Repeat this n times; say we get $L_1, \ldots, L_n$.

To solve the empirical BDR formulation, we proceed as follows. After sampling $L_1, \ldots, L_m$ as above, we check whether $\log(m/|\{j : L_j > b\}|) \leq \delta$, where $|\{j : L_j > b\}|$ is the cardinality of the

set $\{j : L_j > b\}$. In this case, we let

$$w_i^+ = \begin{cases} \frac{1}{|\{j:L_j>b\}|} & \text{for } i \text{ such that } L_i > b \\ 0 & \text{for } i \text{ such that } L_i \leq b \end{cases}.$$

This is the approximate worst-case probability distribution for L, and thus the approximate optimal value function of the BDR problem is 1.

Otherwise, if $\log(m/|\{j : L_j > b\}|) > \delta$, then output

$$w_i = \begin{cases} \frac{\exp(\theta_+)}{\exp(\theta_+)|\{i:L_i>b\}|+|\{i:L_i\leq b\}|} & \text{for } i \text{ such that } L_i > b \\ \frac{1}{\exp(\theta_+)|\{i:L_i>b\}|+|\{i:L_i\leq b\}|} & \text{for } i \text{ such that } L_i \leq b \end{cases}$$

where $\theta_+ > 0$ satisfies

$$\frac{\theta_+ \exp(\theta_+)}{\exp(\theta_+)|\{i : L_i > b\}| + |\{i : L_i \leq b\}|} - \log\left(\frac{1}{n}\exp(\theta_+)|\{i : L_i > b\}| + |\{i : L_i \leq b\}|\right) = \eta.$$

The probability weights $(w_i)_{i=1,\ldots,n}$ on $(L_i)_{i=1,\ldots,n}$ form an approximation for the worst-case probability distribution for L, and

$$\frac{\exp(\theta_+)|\{i : L_i > b\}|}{\exp(\theta_+)|\{i : L_i > b\}| + |\{i : L_i \leq b\}|}$$

is an approximation of the worst-case value of $P_+(L > b)$.

To illustrate numerically, we consider $d = 5$ and $(X_1, \cdots, X_5)$ each following a Gaussian marginal distribution with $\mu_1 = 2.20, \mu_2 = 2.73, \mu_3 = 2.73, \mu_4 = 2.42, \mu_5 = 2.27$ (five numbers generated uniformly between 2 and 3) and $\sigma_1 = 0.92, \sigma_2 = 0.39, \sigma_3 = 0.11, \sigma_4 = 0.56, \sigma_5 = 0.33$ (five numbers generated uniformly between 0 and 1), respectively. Moreover, the t-copula has degrees of freedom

10 and the following dispersion matrix:

$$\Sigma = \begin{bmatrix} 8.19 & -0.92 & -3.54 & 3.35 & -3.96 \\ -0.92 & 6.09 & 1.07 & 1.37 & -0.08 \\ -3.54 & 1.07 & 7.10 & 1.35 & -1.70 \\ 3.35 & 1.37 & 1.35 & 3.14 & -3.73 \\ -3.96 & -0.08 & -1.70 & -3.73 & 8.73 \end{bmatrix}$$

which is generated from a Wishart distribution with 5 degrees of freedom and an identity scale matrix. We use $n = 1,000$ to generate the baseline sample from the t-copula model. Setting $b = 10$, the estimated loss probability is 0.232 with 95% confidence interval $[0.206, 0.258]$. The histograms of the optimally weighted sample at $\delta = 0.1$ and the baseline sample are plotted in Figures 9 and 10. As can be seen, more weights are put on the right side of b, in a uniform manner, to boost the large loss probability.

We also plot the worst-case probability against δ in Figure 11, which, similarly to the example in Section 5.2, shows a concavely increasing pattern.

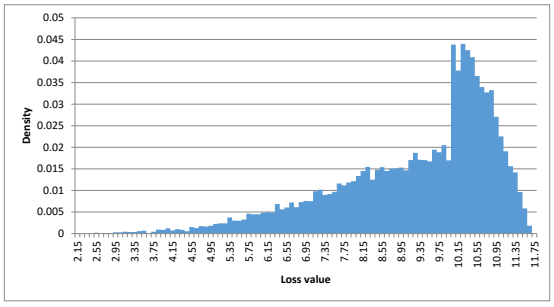

Figure 9: Approximate optimal density under $\delta = 0.1$

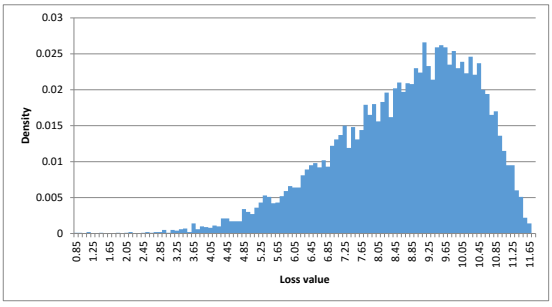

Figure 10: Approximate original density

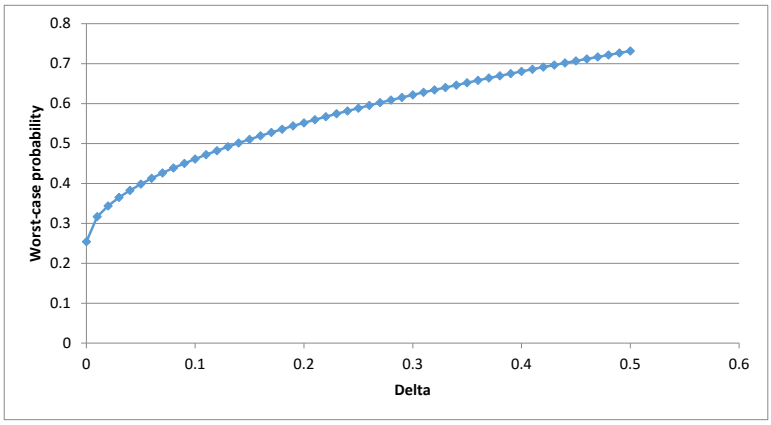

Figure 11: Worst-case probability against δ

8 Robust Conditional Value at Risk

This section demonstrate how our distributionally robust analysis can be further developed to handle risk-analytic problems involving optimization. For this discussion, we focus on the problem of computing, say, the 95% C-VaR of a random variable L which represents potential losses in a given year. In particular, we are interested in computing

$$\text{C-VaR}_{true}\left(\alpha\right) = E_{true}\left(X - b | X > b\right)$$

where $P_{true}\left(X > b\right) = 1 - \alpha$ and we take, for instance, $\alpha = .95$.

It is well known [26] that if L has a continuous distribution under the P_{true} model, then

$$\text{C-VaR}_{true}\left(\alpha\right) = \min_{\theta}\left(\theta + \frac{E_{true}\left(\max(L-\theta, 0)\right)}{1-\alpha}\right).$$

Therefore, we can proceed using a BDR-type formulation. Using the same principles described

earlier, the natural BDR-type extension for a robust upper bound for C-VaR$_{true}(\alpha)$ takes the form

$$\max_{D(P||P_0)\leq\delta} \min_{\theta} \left(\theta + \frac{E\left(\max(L-\theta,0)\right)}{1-\alpha} \right), \tag{22}$$

while the lower bound is given by

$$\min_{D(P||P_0)\leq\delta} \min_{\theta} \left(\theta + \frac{E\left(\max(L-\theta,0)\right)}{1-\alpha} \right). \tag{23}$$

The outer optimization problems are taken over probability models P.

Problems (22) and (23) are challenging to solve in closed form, so it is natural to use the so-called "sample average approximation" (SAA) version of the problem; see [27, 17]. This can be viewed as a generalization of the simulation approach described in Section 7. For example, if we wish to compute

$$\text{C-VaR}_0(\alpha) = \min_{\theta} \left(\theta + \frac{E_0\left(\max(L-\theta,0)\right)}{1-\alpha} \right)$$

for a given baseline model P_0, we first simulate $L_1, ..., L_n$ i.i.d. copies under P_0 and then solve

$$\widehat{\text{C-VaR}}_0(\alpha, n) = \min_{\theta} \left(\theta + \frac{1}{n}\sum_{i=1}^{n} \frac{\max(L_i-\theta,0)}{1-\alpha}. \right) \tag{24}$$

It is known, from the theory of SAA, that under mild assumptions (for instance, if L has a continuous density under P_0),

$$\widehat{\text{C-VaR}}_0(\alpha, n) \approx \text{C-VaR}_0(\alpha) + \frac{\widehat{\sigma}\left(\theta_0^*(n)\right)}{n^{1/2}} Z, \tag{25}$$

where Z is a standard Gaussian random variable,

$$\widehat{\sigma}^2\left(\theta_0^*(n)\right) = \frac{1}{n-1}\sum_{j=1}^{n} \left(\theta_0^*(n) + \frac{\max(L_i-\theta_0^*(n),0)}{1-\alpha} - \widehat{\text{C-VaR}}_0(\alpha, n) \right)^2,$$

and $\theta_0^*(n)$ is the solution obtained by solving (24). In simple words, $\widehat{\sigma}^2\left(\theta_0^*(n)\right)$ is the empirical estimator of the variance of the optimal value of the objective function in (24).

A robust version of an upper bound for (24) is given by adding an outer maximization problem.

That is, we still keep $L_1, ..., L_n$ simulated under P_0, but we consider the problem

$$\begin{aligned} &\max_{w_1,...w_n} \min_{\theta} \left(\theta + \sum_{i=1}^{n} w_i \frac{\max(L_i - \theta, 0)}{1-\alpha} \right) \\ &s.t. \\ &\sum_{i=1}^{n} w_i \log(nw_i) \leq \delta \\ &\sum_{i=1}^{n} w_i = 1, \text{ and } w_i \geq 0 \text{ for all } 1 \leq i \leq n. \end{aligned} \tag{26}$$

We can select δ guided by the discussion in Section 6.

In the C-VaR setting, however, if we are building P_0 based on a non-parametric, purely data-driven approach, it turns out that δ should be selected so that $P\left(2n\delta > \chi_2^2\right) = 1 - \beta$for some $\beta > 0$, $(1-\beta)$ being the confidence level at which the obtained upper bound will hold for (22). Note that the degree of freedom has increased from one to two compared with the discussion in Section 6. The reason for changing the degrees of freedom is the appearance of the inner minimization in problem (26), which introduces another equality constraint in the empirical likelihood derivation outlined in the Appendix. Precisely, the appearance of the inner minimization problem yields the additional optimality constraint which forces the derivative of the objective function with respect to θ to vanish. The derivation is similar to that given in Section 6, and further details can be found in [19].

Similar to (26), the lower bound of (24) is given by

$$\begin{aligned} &\min_{w_1,...,w_n} \min_{\theta} \left(\theta + \sum_{i=1}^{n} w_i \frac{\max(L_i - \theta, 0)}{1-\alpha} \right) \\ &s.t. \\ &\sum_{i=1}^{n} w_i \log(nw_i) \leq \delta \\ &\sum_{i=1}^{n} w_i = 1, \text{ and } w_i \geq 0 \text{ for all } 1 \leq i \leq n. \end{aligned} \tag{27}$$

Once again, if P_0 is chosen non-parametrically as the empirical distribution of the observed losses, then it makes sense to choose δ so that $P\left(2n\delta > \chi_2^2\right) = 1 - \beta$. Note that the confidence interval obtained using distributionally robust analysis—namely, from solving (26) as upper bound and (27),

as lower bound—tends to have better small-sample properties than the one derived using (25). The next example illustrates the construction of robust C-VaR estimates in a simple setting.

8.1 Example Illustrating Robust C-VaR Estimation

To illustrate the applicability of the method, we consider the problem of estimating C-VaR in which we assume that the loss, L, follows the standard normal distribution under P_0. We set $\alpha = 0.9$ and generate $n = 1,000$ observations. Figure 12 shows the upper and lower robust bounds computed from (26) and (27) against different values of δ. As we can see, the width of the robust interval widens at a decreasing rate as δ increases.

Moreover, we also test the performance of the 95% (i.e., $\beta = 0.05$) confidence bounds using (26) and (27), by selecting δ such that $P(2n\delta > \chi^2_2) = 1-\beta$. We carry out the cases $n = 50$ and $n = 100$. Table 1 reports the point estimate of the coverage probability, mean lower and upper bounds, and the mean and standard deviation of the interval width for empirical likelihood, while Table 2 shows the results using the classical SAA theory via (25). We could see that empirical likelihood gives a higher and more accurate coverage, though the SAA counterpart gives tighter and less varied interval width.

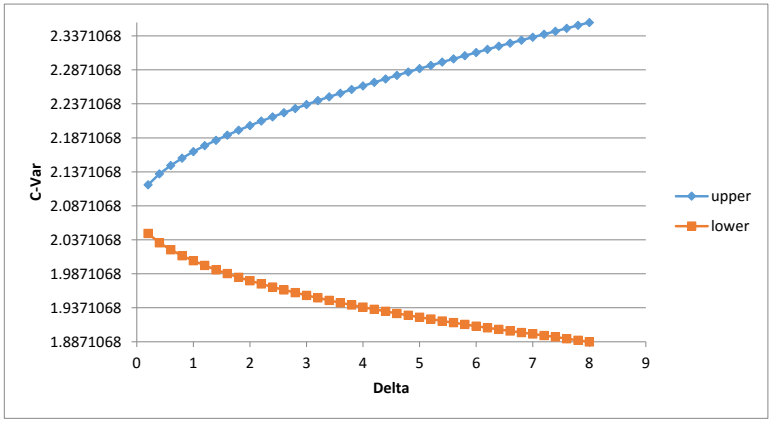

Figure 12: Worst-case C-VaR against δ

n	Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
50	0.90	1.22	2.33	1.11	0.43
100	0.94	1.32	2.26	0.94	0.26

Table 1: Statistical performances of empirical likelihood for different sample sizes

n	Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
50	0.86	1.21	2.26	1.05	0.47
100	0.84	1.34	2.05	0.71	0.21

Table 2: Statistical performances of standard confidence interval of SAA for different sample sizes

9 Additional Considerations

We shall discuss two alternatives that can be used for robust performance analysis. The first one involves the use of moment constraints, and the second one discusses different notions of discrepancy.

9.1 Robust Performance Analysis via Moment Constraints

In some cases, there might not be a direct baseline distribution P_0 that can be constructed. Alternatively, if information on moment constraints is available, we might consider worst-case optimizations under such information, in the form

$$\begin{array}{ll} \max & E\left(h(X)\right) \\ \text{s.t.} & E\left(v_i(X)\right) \leq \alpha_i,\ i = 1, \ldots, s \\ & E(v_i(X)) = \alpha_i,\ i = s+1, \ldots, m \end{array} \tag{28}$$

where the maximization is over all probability models P that satisfy the constraints. Again we focus on the maximization problem here. This is a general formulation that has m moment constraints, and $v_i(\cdot)$ can represent any function. For instance, for moment constraints involving means and variances, we can select $v_1(x) = x$ and $v_2(x) = -x$, $v_3(x) = x^2$, $v_4(x) = -x^2$, and $\alpha_1 = \overline{\mu}$, $\alpha_2 = -\underline{\mu}$, $\alpha_3 = \overline{\sigma}$, $\alpha_4 = -\underline{\sigma}$, and all constraints could be inequalities. There is a general procedure for solving these problems which builds on linear programming. The most tricky part involves finding the

support of the distribution. Observe that, if the support of the optimal distribution is known, then problem (28) is just a problem with a linear objective function and linear constraints, and is solvable using standard routines.

Finding the support involves a sequential search. More precisely, the procedure for solving (28) is shown in Algorithm 1 (which borrows from, e.g., [4]).

Algorithm 1 Generalized linear programming procedure for solving (28)

Initialization: An arbitrary probability distribution on the support $\{x_1, \ldots, x_L\}$, where $L \leq M+1$, that lies in the feasible region in (28). Set $\tau = L$.

Procedure: For each iteration $k = 1, 2, \ldots$, given $\{x_1, \ldots, x_\tau\}$:

1. Master problem solution. Solve

$$\begin{array}{ll} \max & \sum_{j=1}^{\tau} h(x_j) p_j \\ \text{s.t.} & \sum_{j=1}^{\tau} v_i(x_j) p_j \leq \alpha_i, \ i = 1, \ldots, s \\ & \sum_{j=1}^{\tau} v_i(x_j) p_j = \alpha_i, \ i = s+1, \ldots, M \\ & \sum_{j=1}^{\tau} p_j = 1 \\ & p_j \geq 0, \ j = 1, \ldots, \tau \end{array}$$

Let $\{p_1^k, \ldots, p_\tau^k\}$ be the optimal solution. Find the dual multipliers $\{\theta^k, \pi_1^k, \ldots, \pi_M^k\}$ that satisfy

$$\begin{array}{l} \theta^k + \sum_{i=1}^{M} \pi_i^k v_i(x_j) = v_0(x_j), \text{ if } p_j > 0, j = 1, \ldots, \tau \\ \theta^k + \sum_{i=1}^{M} \pi_i^k v_i(x_j) \geq v_0(x_j), \text{ if } p_j = 0, j = 1, \ldots, \tau \\ \pi_i^k \geq 0, i = 1, \ldots, s \end{array}$$

2. Subproblem solution. Find $x_{\tau+1}$ that maximizes

$$\rho(x; \theta^k, \pi_1^k, \ldots, \pi_M^k) = h(x) - \theta^k - \sum_{i=1}^{M} \pi_i^k v_i(x)$$

If $\rho(x_{\tau+1}; \theta^k, \pi_1^k, \ldots, \pi_M^k) > 0$, then let $\tau = \tau + 1$; otherwise, stop the procedure, and $\{x_1, \ldots, x_\tau\}$ are the optimal support points, with $\{p_1^k, \ldots, p_\tau^k\}$ the associated weights.

After the last iteration, output

$$\sum_{j=1}^{\tau} h(x_j) p_j^k$$

We discuss Algorithm 1 in the following several aspects:

1. *Interpretation:* The output of the procedure is an exact optimal value of (28). The worst-case probability distribution is a finite-support discrete distribution on $\{x_1, \ldots, x_\tau\}$ with weights $\{p_1^k, \ldots, p_\tau^k\}$ obtained in the last iteration.

2. *Comparison with standard and BDR formulation:* Unlike the BDR formulation, (28) does not have a baseline input distribution to begin with.

3. *Computational efficiency:* Step 1 in each iteration of Algorithm 1 can be carried out by standard linear programming solver, which can output both the optimal $\{p_j\}$ and the dual multipliers $\{\theta^k, \pi_1^k, \ldots, \pi_M^k\}$. Step 2 is a one-dimensional line search if X is one-dimensional.

4. *Minimization counterpart:* For a minimization problem, simply replace h with $-h$ in the whole procedure of Algorithm 1, except in the last, output $\sum_{j=1}^{\tau} h(x_j)p_j^k$.

9.2 Renyi Divergence as Discrepancy Notion

While we have presented the method based on the relative entropy discrepancy, other notions can be used. For example, we could also consider the BDR formulation using the so-called Renyi divergence of degree $\alpha > 1$, defined via

$$D_\alpha\left(P||P_0\right) = \frac{1}{\alpha - 1}\log E_0\left(\left(\frac{dP}{dP_0}\right)^\alpha\right).$$

As $\alpha \to 1$, we recover Kullback-Leibler divergence or relative entropy; that is, we have that

$$D_\alpha\left(P||P_0\right) \to D\left(P||P_0\right) = E\left(\log\left(\frac{dP}{dP_0}\right)\right).$$

The corresponding distributionally robust formulation takes the form

$$\max\ E\left(h\left(X\right)\right) \quad \text{s.t.} \quad D_\alpha\left(P||P_0\right) \leq \delta,$$

and the optimal solution is obtained as follows. First, given $\theta_1, \theta_2 > 0$, define

$$Z_+\left(\theta_1, \theta_2\right) = \max(\theta_1 + \theta_2 h\left(X\right), 0)^{1/(1-\alpha)},$$

with θ_1, θ_2 chosen so that

$$E_0\left(Z_+\left(\theta_1, \theta_2\right)\right) = 1, \quad E_0\left(Z_+\left(\theta_1, \theta_2\right)^\alpha\right) = \exp\left(\delta\left(\alpha - 1\right)\right).$$

Finally, for any set A, we have that

$$P_+(X \in A) = E_0(Z_+(\theta_1, \theta_2) I(X \in A)).$$

The use of Renyi divergence would lead to somewhat less conservative estimates, but the parameter δ might be more difficult to estimate. Once again, we consider a rare-event estimation problem to develop greater intuition and contrast the effect of using Renyi divergence instead of relative entropy.

9.2.1 Robust Rare-Event Analysis via Renyi Divergence

As in the setting of relative entropy, we consider the case in which $h(X) = I(X \in B)$, so we have that

$$P_+(X \in B) = (\theta_1 + \theta_2)^{1/(1-\alpha)} P_0(X \in B),$$

and

$$\begin{aligned} E_0(Z_+(\theta_1, \theta_2)^\alpha) &= \theta_1^{\alpha/(1-\alpha)} P_0(X \notin B) + (\theta_1 + \theta_2)^{\alpha/(1-\alpha)} P_0(X \in B) = \exp(\delta(\alpha - 1)), \\ E_0(Z_+(\theta_1, \theta_2)) &= \theta_1^{1/(1-\alpha)} P_0(X \notin B) + (\theta_1 + \theta_2)^{1/(1-\alpha)} P_0(X \in B) = 1. \end{aligned}$$

Letting $(\theta_1 + \theta_2)^{\alpha/(1-\alpha)} = \eta / P_0(X \in B)$, and substituting in the previous display, we have

$$\begin{aligned} \theta_1^{\alpha/(1-\alpha)}(1 - P_0(X \in B)) + \eta &= \exp(\delta(\alpha - 1)), \\ \theta_1^{1/(1-\alpha)}(1 - P_0(X \in B)) + \eta^{1/\alpha} P_0(X \in B)^{1-1/\alpha} &= 1. \end{aligned}$$

As $P_0(X \in B) \to 0$, since $\alpha > 1$, we can select $\theta_1 \approx 1$ and $(\theta_1 + \theta_2)^{1/(1-\alpha)} \approx \eta^{1/\alpha} / P_0(X \in B)^{1/\alpha}$, with $\eta \approx \exp(\delta(\alpha - 1)) - 1$, concluding that

$$P_+(X \in B) \approx (\exp(\delta(\alpha - 1)) - 1)^{1/\alpha} P_0(X \in B)^{1-1/\alpha}.$$

Let us revisit the example discussed at the end of Section 5.1. Assume that X is exponentially

distributed with mean one under P_0. Select $B = [b, \infty)$ and $\alpha = 2$. Then we have that

$$P_+ (X > b) \approx (\exp(\delta) - 1)^{1/2} \exp(-b/2).$$

In contrast, when using the relative entropy for the case of exponentially distributed X, we obtain a much lower rate of decay (of the form δ/b for b large; see (17)). In this sense, relative entropy induces a much more cautious robust methodology than Renyi divergence.

10 Conclusions

We have discussed a systematic approach for the quantification of potential model error based on the BDR formulation, which is a convex optimization problem in the space of probability distribution.

The approach can be used to study the impact of dependence in multivariate models when evaluating a performance measure of interest. We also illustrated the use of the BDR formulation in the context of computing conditional value-at-risk. Important advantages of the approach are that it is non-parametric, substantially general and computationally tractable.

We discussed several ways to define the model uncertainty region, based on relative entropy, Renyi divergence or simple imposition of moment constraints. When using the relative entropy and non-parametric specifications of the baseline model, we discussed the connection between the method we advocate and empirical likelihood. This connection allows further specification of the model uncertainty region.

11 Appendix: Approximating Distribution for the Size of the Feasible Region

Similar to the use of the KKT conditions for the solution of the BDR problem formulation, we obtain that the optimal solution satisfies

$$w_i(\theta) = \frac{\exp(\theta h(X_i))}{\sum_{j=1}^{n} \exp(\theta h(X_j))},$$

where θ satisfies $\sum_{i=1}^{n} w_i(\theta)(h(X_i)-\gamma)=0$, namely

$$\frac{\sum_{i=1}^{n}\exp(\theta h(X_i))(h(X_i)-\gamma)}{\sum_{j=1}^{n}\exp(\theta h(X_j))}=0. \tag{29}$$

Now suppose that indeed $E_{true}(h(X)-\gamma)=0$, and let us consider $\bar{h}(x)=h(x)-\gamma$. We can rewrite (29)—after multiplying by $\sum_{j=1}^{n}\exp(\theta h(X_j)-\gamma\theta)/n^{1/2})$ —as

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n}\exp\left(\theta\bar{h}(X_i)\right)\bar{h}(X_i)=0.$$

Then let $\theta=\eta/n^{1/2}$ and perform a Taylor expansion to conclude that

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n}\bar{h}(X_i)\left(1+\eta\frac{\bar{h}(X_i)}{n^{1/2}}+...\right)=0. \tag{30}$$

Recall that the CLT states the approximation in distribution

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n}\bar{h}(X_i)\approx\sigma Z,$$

where $\sigma^2=Var_{true}(h(X))$, and Z is standard Gaussian. Therefore, solving for η in (30) and ignoring lower-order error terms, we obtain that

$$\eta\approx-\frac{1}{n^{1/2}}\sum_{i=1}^{n}\bar{h}(X_i)/\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)^2\right)\approx\frac{Z}{\sigma}.$$

Now consider $\mathcal{R}_n(\gamma)$, which can be written as

$$
\begin{aligned}
& \mathcal{R}_n(\gamma) \\
= & \sum_{i=1}^{n} w_i(\theta) \log\left(\frac{\exp\left(\theta\bar{h}(X_i)\right)}{n^{-1}\sum_{j=1}^{n}\exp\left(\theta\bar{h}(X_j)\right)}\right) \\
= & \theta\sum_{i=1}^{n} w_i(\theta)\bar{h}(X_i) - \log\left(\frac{1}{n}\sum_{i=1}^{n}\exp(\theta\bar{h}(X_i))\right) \\
= & \frac{\theta\frac{1}{n}\sum_{i=1}^{n}\exp(\theta\bar{h}(X_i))\bar{h}(X_i)}{\frac{1}{n}\sum_{i=1}^{n}\exp(\theta\bar{h}(X_i))} - \log\left(\frac{1}{n}\sum_{i=1}^{n}\exp(\theta\bar{h}(X_i))\right) \\
= & \frac{\theta\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i) + \frac{\theta}{n}\sum_{i=1}^{n}\bar{h}(X_i)^2 + \cdots\right)}{1 + \frac{\theta}{n}\sum_{i=1}^{n}\bar{h}(X_i) + \frac{\theta^2}{2n}\sum_{i=1}^{n}\bar{h}(X_i)^2 + \cdots} \\
& - \left(\frac{\theta}{n}\sum_{i=1}^{n}\bar{h}(X_i) + \frac{\theta^2}{2n}\sum_{i=1}^{n}\bar{h}(X_i)^2 - \frac{\theta^2}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)\right)^2 + \cdots\right) \\
\approx & \frac{\theta}{n}\sum_{i=1}^{n}\bar{h}(X_i) + \frac{\theta^2}{n}\sum_{i=1}^{n}\bar{h}(X_i)^2 \\
& - \theta^2\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)\right)^2 - \frac{\theta}{n}\sum_{i=1}^{n}\bar{h}(X_i)^2 - \frac{\theta^2}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)\right)^2\right) \\
& \text{by collecting terms up to } \theta^2 \\
= & \frac{\theta^2}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)\right)^2\right) \\
\approx & \frac{\eta^2\sigma^2}{2n} \approx \frac{Z^2}{2n}
\end{aligned}
$$

hence concluding that $2n\mathcal{R}_n(\gamma) \Rightarrow \chi^2_1$, as indicated in our discussion leading to (21).

Acknowledgement: We are grateful for the support received from the Society of Actuaries to carry out this research. We thank the Project Oversight Group for their thoughtful comments and feedback, which were very helpful to improve our work.

References

[1] Directive 2009/138/ec of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II). *Available at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:335:0001:0155:en:PDF*.

[2] Updated static mortality tables for defined benefit pension plans for 2016. *Available at https://www.irs.gov/pub/irs-drop/n-15-53.pdf*.

[3] Bertsimas, D. and I. Popescu (2005). Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization 15*(3), 780–804.

[4] Birge, J. R. and J. H. Dulá (1991). Bounding separable recourse functions with limited distribution information. *Annals of Operations Research 30*(1), 277–298.

[5] Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge University Press.

[6] Breuer, T. and I. Csiszár (2013). Measuring distribution model risk. *Mathematical Finance*.

[7] Cover, T. M. and J. A. Thomas (2012). *Elements of information theory*. John Wiley & Sons.

[8] Cox, D. R. and D. V. Hinkley (1979). *Theoretical statistics*. CRC Press.

[9] Delage, E. and Y. Ye (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research 58*(3), 595–612.

[10] Deming, W. E. and F. F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics 11*(4), 427–444.

[11] Glasserman, P. and X. Xu (2013). Robust portfolio control with stochastic factor dynamics. *Operations Research 61*(4), 874–893.

[12] Glasserman, P. and X. Xu (2014). Robust risk measurement and model risk. *Quantitative Finance 14*(1), 29–58.

[13] Glasserman, P. and L. Yang (2015). Bounding wrong-way risk in CVA calculation. *Available at SSRN 2607649*.

[14] Hansen, L. P. and T. J. Sargent (2008). *Robustness*. Princeton University Press.

[15] Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research 30*(2), 257–280.

[16] Jiang, R. and Y. Guan (2012). Data-driven chance constrained stochastic program. *Mathematical Programming*, 1–37.

[17] Kleywegt, A. J., A. Shapiro, and T. Homem-de Mello (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization 12*(2), 479–502.

[18] Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

[19] Lam, H. and E. Zhou (2015). Quantifying uncertainty in sample average approximation. In *Proceedings of the 2015 Winter Simulation Conference*, pp. 3846–3857. IEEE Press.

[20] Natarajan, K., D. Pachamanova, and M. Sim (2008). Incorporating asymmetric distributional information in robust value-at-risk optimization. *Management Science 54*(3), 573–585.

[21] Nguyen, X., M. J. Wainwright, and M. I. Jordan (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *Information Theory, IEEE Transactions on 56*(11), 5847–5861.

[22] Nilim, A. and L. El Ghaoui (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research 53*(5), 780–798.

[23] Owen, A. B. (2001). *Empirical likelihood*. CRC press.

[24] Petersen, I. R., M. R. James, and P. Dupuis (2000). Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *Automatic Control, IEEE Transactions on 45*(3), 398–412.

[25] Póczos, B. and J. Schneider (2011). On the estimation of alpha-divergences.

[26] Rockafellar, R. T. and S. Uryasev (2000). Optimization of conditional value-at-risk. *Journal of Risk 2*, 21–42.

[27] Shapiro, A., D. Dentcheva, et al. (2014). *Lectures on stochastic programming: Modeling and Theory*, Volume 16. SIAM.