A toolkit for robust risk assessment using F-divergences

Kruse, Thomas; Schneider, Judith C.; Schweizer, Nikolaus

Published in:
Management Science

DOI:
10.1287/mnsc.2020.3822

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Toolkit for Robust Risk Assessment Using F-Divergences

Thomas Kruse , Judith C. Schneider , Nikolaus Schweizer

To cite this article:
Thomas Kruse , Judith C. Schneider , Nikolaus Schweizer (2021) A Toolkit for Robust Risk Assessment Using F-Divergences. Management Science 67(10):6529-6552. https://doi.org/10.1287/mnsc.2020.3822

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 67, No. 10, October 2021, pp. 6529–6552
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

A Toolkit for Robust Risk Assessment Using *F*-Divergences

Thomas Kruse,[a] Judith C. Schneider,[b,c] Nikolaus Schweizer[d]

[a] Institute of Mathematics, Justus Liebig University, 35392 Giessen, Germany; [b] Chair of Finance, Leuphana University, 21335 Lüneburg, Germany; [c] Finance Center Münster, University of Münster, 48143 Münster, Germany; [d] Department of Econometrics and Operations Research, Tilburg University, 5037AB Tilburg, Netherlands

Contact: thomas.kruse@math.uni-giessen.de, https://orcid.org/0000-0003-2388-3929 (TK); judith.schneider@leuphana.de, https://orcid.org/0000-0001-6955-021X (JCS); n.f.f.schweizer@uvt.nl, https://orcid.org/0000-0002-5807-7321 (NS)

Received: March 17, 2019
Revised: June 10, 2020
Accepted: August 17, 2020
Published Online in Articles in Advance: March 18, 2021

https://doi.org/10.1287/mnsc.2020.3822

Abstract. This paper assembles a toolkit for the assessment of model risk when model uncertainty sets are defined in terms of an *F*-divergence ball around a reference model. We propose a new family of *F*-divergences that are easy to implement and flexible enough to imply convincing uncertainty sets for broad classes of reference models. We use our theoretical results to construct concrete examples of divergences that allow for significant amounts of uncertainty about lognormal or heavy-tailed Weibull reference models without implying that the worst case is necessarily infinitely bad. We implement our tools in an open-source software package and apply them to three risk management problems from operations management, insurance, and finance.

History: Accepted by Baris Ata, stochastic models and simulation.

Supplemental Material: The data files are available at https://doi.org/10.1287/mnsc.2020.3822.

Keywords: *F*-divergence • model risk • risk management • robustness

1. Introduction

F-divergences, also known as φ-divergences or f-divergences, are a class of distance measures for probability models that were first proposed independently in the 1960s by Csiszár (1963), Morimoto (1963), and Ali and Silvey (1966). Since then, they have found applications in various fields such as probability theory, statistics, information theory, operations research, and most recently, machine learning. This paper is motivated by applications of *F*-divergences in the assessment of model risk in economics and business. The basic idea is that a decision maker who is uncertain about the accuracy of his reference model can augment calculations under this model with best and worst cases over all models that lie within a divergence ball of fixed radius from the reference model. This idea goes back to the operations research literature on robust optimization and robust control (Whittle 1990, Bertsimas and Sim 2004, Ben-Tal et al. 2009). Related ideas have been applied and developed in quantitative finance and economics (see Föllmer and Schied 2011 and Hansen and Sargent 2011 for seminal monographs).

The *F*-divergence quantifies how strongly an alternative model differs from a given reference model. To this end, one asks how strongly the two models differ in the probabilities they assign to possible outcomes. The function F that defines the *F*-divergence can be interpreted as a cost function that is applied to the relative differences in probabilities, that is, to the change of measure from the reference model to the alternative. The *F*-divergence between the models is then simply the expectation over potential outcomes of this cost function under the reference model. If the function F is relatively flat, the divergence makes little distinction between small and large local differences in assigned probabilities. In contrast, a function F with a steep growth behavior puts considerable penalties on large differences between the two models.

The most prominent *F*-divergence is Kullback-Leibler (KL) divergence, also known as relative entropy. KL-divergence corresponds to a particular choice of the cost function F and became the default choice in many applications (Hansen and Sargent, 2011). However, for many purposes, other *F*-divergences have similar theoretical properties and could be used just as well. For instance, the monograph by Pardo (2005) develops rich statistical theories that apply, essentially, to the whole class of *F*-divergences. For our application of assessing model risk, the possibility of a flexible choice of F is emphasized, for example, in Breuer and Csiszár (2016), Csiszár and Breuer (2018), and Ben-Tal et al. (2013). Going one step further, Kruse et al. (2019)

argues that carefully choosing the divergence is necessary for assessing model risk convincingly across different reference models, especially when the state space is unbounded. They show that there is a considerable gap between the most commonly used divergences, KL-divergence and polynomial divergence, with many important reference models falling into the range where neither divergence is suitable.

A consequence of a fixed cost function F is that, depending on the reference model, F may be growing too quickly or too slowly, thus taking too few or too many alternative models into account. This results in a worst-case analysis that is either overly optimistic or pessimistic. The function F thus has to be chosen with the reference model in mind.

The aim of this paper is to develop the modeling of uncertainty sets with flexible choices of F from a theoretical possibility into a viable option for handling model uncertainty in challenging practical problems using an open-source software package. This package makes it easy to incorporate the resulting worst- and best-case analysis into management processes.

As a foundational step, we provide a methodology for constructing F-divergences that fit a previously defined purpose while still being straightforward to implement in relevant practical problems. The theoretical toolkit we assemble consists of three components.

The first component is a new parametrization of the class of F-divergences. The idea is to specify the divergence not in terms of the cost function F itself but in terms of its derivative, the marginal cost function. In particular, we suggest to build divergences from marginal cost functions that have closed-form inverse functions. The advantage is that when using an F-divergence constraint in an optimization problem, first-order conditions can easily be resolved. The only price we pay is that when evaluating the divergence, the dimensionality of the integral we need to compute is increased by one. To make this tradeoff more tangible, in our numerical illustration of worst-case hedging errors, we trade in a 180-dimensional integral for a 181-dimensional integral by specifying the marginal cost function F' in place of the cost function F. What we gain is a worst-case change of measure that can be computed efficiently without numerically inverting a nonlinear function at each evaluation.

The second component of the toolkit is a characterization of worst-case analysis under this new class of divergences. Technically, our results build on the earlier literature, in particular Breuer and Csiszár (2016). However, our focus on a broad but well-behaved subclass of F-divergences yields explicit sufficient conditions and expressions for worst-case models that are easy to handle in practice. In addition, we provide new results that address situations where a quantity of interest depends on a possibly high-dimensional vector of input risk factors. We show how to carry over the properties needed in the worst-case analysis from the input risk factors, which tend to be easier to analyze, to the quantity of interest. Moreover, we prove that it is irrelevant whether uncertainty sets are defined on the level of the joint distribution of the risk factors or on the level of the quantity of interest. It is thus without loss of generality to conduct a worst-case analysis on the level of the univariate output quantity. This property is a major advantage of uncertainty sets that are defined in terms of F-divergences rather than, for example, Wasserstein distances. A valid worst-case analysis under F-divergences can be conducted in a postprocessing step, relying solely on observations of the quantity of interest without access to the underlying risk factors.

The third component is a series of results on the contents of divergence balls for given combinations of divergence and reference model. The purpose of these results is to guide the choice or design of F-divergences in applications to robust risk assessment. In particular, we characterize those marginal cost functions that are *critical* for a given reference model in the sense that models in the uncertainty set have finite moments up to a certain order but not necessarily higher. For the classical divergences, KL-divergence and polynomial divergence, criticality results of this type are shown in Kruse et al. (2019). For instance, KL-divergence is critical for Gaussian reference models with a cutoff at the second moment. Thus, KL-divergence balls around Gaussian models contain models with a diverse range of tail behaviors; however, there is the implicit moment constraint that expected value and variance must be finite, guaranteeing a well-defined worst-case analysis. Similarly, polynomial divergences are critical for power law reference models. Our new results characterize F-divergences with similar implicit moment constraints for general reference models. The order of finite moments is a parameter of choice in the construction.

After introducing this toolkit, we proceed by applying it to two broad classes of reference models. To this end, we provide explicit constructions of divergences that are critical for Weibull-type models and for generalized lognormal models. Although the names of both model classes refer to well-known parametric families of distributions, our divergences have broader applicability. For example, Gaussian tail behavior is Weibull type because only the tail behavior determines which worst-case changes of measure are critical. To be more precise, if the log-density behaves asymptotically like a (possibly fractional) polynomial, then the distribution is Weibull type. What heavy-tailed Weibull models and generalized lognormal models have in common is that the density decreases faster than polynomially but more slowly than exponentially in the tail. Functions that fall between polynomial and exponential growth behavior are not as well studied as

their polynomial and exponential counterparts. Closed-form inverses for derivatives are rare. Thus, the construction strongly relies on the first three parts of the toolbox: The third component determines the ideal growth behavior of F, the first component clarifies that it suffices to specify the marginal cost function, and the second component provides the technical results that provide the foundation for the worst-case analysis.

In the open-source Python package divbox, we provide an implementation of best- and worst-case analysis with a flexible choice of the divergence. divbox takes a sample from the nominal model as an input and returns worst- and best-case expected values together with suitable asymptotic confidence bounds. With this package, worst-case analysis is effectively as easy as computing a sample mean with associated standard errors.

Our paper concludes with three practical illustrations from the fields of operations management, insurance, and finance. In the first illustration, we study worst- and best-case scenarios for the gains from having a centralized inventory in a multilocation newsvendor problem with heavy-tailed demand. In the second illustration, we consider worst cases of an insurance loss that is modeled as a sum of correlated Weibull random variables. In the third example, we consider the absolute hedging error that arises from hedging a call option only over a sequence of discrete time points, trading off hedging quality against transaction costs. We impose a lognormal (Black-Scholes) reference model for the option's underlying stock.

In all three settings, the actual target quantity no longer has a Weibull or lognormal distribution, but the quantity of interest inherits the type of tail behavior from the input risk factors. Our tools enable us to conduct worst-case analyses that are easy to implement and provide an explicit control on the amount and type of model uncertainty. Moreover, we take into account the possibility of qualitatively heavier tails than the reference model while still obtaining finite worst cases. In future work, these worst-case computations could serve an input for a robust optimization of strategic decisions.

2. A Toolkit for Robust Risk Assessment

In this section, we introduce the three components of our toolkit for robust risk assessment with general F-divergences. Section 2.1 introduces the setting and the new parametrization of F-divergences. Section 2.2 characterizes worst-case distributions for this class of divergences. Section 2.3 studies the contents of uncertainty sets induced by F-divergences. In particular, we provide conditions under which an F-divergence is critical for a given reference model in the following sense: Finiteness of a certain moment determines whether an alternative model is a candidate for inclusion in the uncertainty set or not.[1]

2.1. A Class of F-Divergences

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable. We denote by ν the distribution of X under $\mathbb{P}$ and call ν the reference model. Throughout the paper, we denote by $\mathcal{S} \subseteq \mathbb{R}$ the (bounded or unbounded) support of ν. We assume that ν has a density[2] f such that $\int_{\mathcal{S}} |x| f(x) dx < \infty$. We are interested in uncertainty sets around this reference model that are defined in terms of an F-divergence that quantifies how alternative models differ from ν. To this end, let η be a second distribution on $\mathcal{S}$, which is absolutely continuous with respect to ν. Here and throughout the paper, we denote the density of η by g.[3] For a convex function F with $F(1) = 0$, the F-divergence between ν and η is defined as

$$D_F(\eta|\nu) = \int_{\mathcal{S}} F\left(\frac{g(x)}{f(x)}\right) f(x) dx, \tag{1}$$

with the convention that $F(\frac{0}{0}) \cdot 0 = 0$. If η is not absolutely continuous with respect to ν, then we set $D_F(\eta|\nu) = +\infty$. This is the classical definition of F-divergences from Csiszár (1963), Morimoto (1963), and Ali and Silvey (1966).

How differences between models are measured depends crucially on the choice of the function F. In this paper, we propose a new parametrization of the function F. This parametrization facilitates choosing F in a problem-specific way while keeping the worst-case analysis of the following sections practicable. The key idea is to specify F in terms of a transformed derivative function H.

Let $H : \mathbb{R} \to \mathbb{R}$ be a continuous, strictly increasing function such that $\lim_{y\to\infty} H(y) = \infty$ and $H(0) = 0$. We define the function $F : (0, \infty) \to \mathbb{R}$ by[4]

$$F(y) = \int_1^y H(\log(z)) dz \tag{2}$$

for all $y \in (0, \infty)$. We extend F to a (convex) function on $[0, \infty)$ by defining $F(0) = \lim_{y\downarrow 0} F(y) \in (0, \infty]$. The next lemma proves several technical properties of the function F. In particular, it verifies that we have a valid F-divergence; that is, F is convex with $F(1) = 0$. Like all our results, the lemma is proved in the Appendix.

Lemma 1. *F is strictly convex, continuously differentiable and satisfies $F(1) = 0$, $F'(1) = 0$, and $\lim_{y\to\infty} \frac{F(y)}{y} = \infty$ (i.e., F is cofinite). In particular, F is nonnegative and attains its minimum 0 at 1. If H is continuously differentiable with $H'(0) = 1$, then F is twice continuously differentiable with $F''(1) = 1$.*

Let us briefly discuss our assumptions and the implied properties of F.

Remark 1. (i) We thus restrict attention to divergences with $F'(1) = 0$. This is without loss of generality: We can always change the slope of $F(y)$ in a single point by adding $c(y-1)$, $c \in \mathbb{R}$, to F without changing the value of the divergence in (1). The assumption ensures that F is nonnegative and possesses a global minimum in 1. This is in line with the interpretation of F as a cost function that is minimal when the model is not changed. (ii) Any sufficiently smooth, convex function F with $F(1) = 0$ can be written in the form (2) for some increasing H. Thus, in principle, the class of divergences in (2) is almost the same as the usual class of F-divergences. The advantage of defining F-divergences in terms of $F'(\cdot) = H(\log(\cdot))$ rather than F comes from a modeling perspective. In applications to worst-case analysis and robust optimization, it is a considerable advantage to have a closed-form expression for the inverse of the derivative F' as this appears in the worst-case change of measure. See Section 2.2 for details. If the function H is chosen such that it has a closed-form inverse, then this property comes for free. (iii) The property $F''(1) = 1$ is important in statistical applications of F-divergences (see Pardo 2005 who calls F-divergences with this property *standard*). It can be understood as a normalization of the divergence for small perturbations. When we develop F into a Taylor approximation around 1, $F(y) \approx F'(1)(y-1) + \frac{1}{2}F''(1)(y-1)^2$, where the linear term is spurious as argued in (i), we see that two divergences with the same value of $F''(1)$ will tend to give similar values when $g/f \approx 1$. In terms of the function H, the property $F''(1) = 1$ becomes $H'(0) = 1$.

To illustrate how F and H are related, we consider the two most prominent F-divergences: KL-divergence and polynomial divergence.[5] In particular, we see that KL-divergence corresponds to the identity function $H(y) = y$. In this sense, divergences that arise from other choices of H can be viewed as distortions of KL-divergence.

Example 1.

(i) KL-divergence lies in our class of divergences: Choosing $H(y) = y$, $y \in \mathbb{R}$, in (2) leads to $F_{KL}(y) := F(y) = y\log(y) - y + 1$ and thus

$$D_{KL}(\eta|\nu) := D_F(\eta|\nu) = \int_{\mathcal{S}} \left[\frac{g(x)}{f(x)}\log\left(\frac{g(x)}{f(x)}\right) - \frac{g(x)}{f(x)} + 1\right] \times f(x)dx = \int_{\mathcal{S}} \log\left(\frac{g(x)}{f(x)}\right) g(x)dx,$$

for any distribution η that is absolutely continuous with respect to ν. This is not the *textbook* choice of F_{KL}, which would be $y\log(y)$, but rather a modification that satisfies $F'(1) = 0$ as discussed in Remark 1 (i).

(ii) Polynomial divergences also lie in our class of divergences: Let $p > 1$. Then the choice $H(y) = \frac{e^{(p-1)y}-1}{p-1}$, $y \in \mathbb{R}$, leads to $F_p(y) := F(y) = \frac{y^p - p(y-1) - 1}{p(p-1)}$ and thus

$$D_p(\eta|\nu) := D_F(\eta|\nu) = \int_{\mathcal{S}} \left[\frac{\left(\frac{g(x)}{f(x)}\right)^p - p\left(\frac{g(x)}{f(x)} - 1\right) - 1}{p(p-1)}\right] \times f(x)dx = \int_{\mathcal{S}} \left[\frac{\left(\frac{g(x)}{f(x)}\right)^p - 1}{p(p-1)}\right] f(x)dx \tag{3}$$

for any distribution η that is absolutely continuous with respect to ν.

It is well known that $D_p(\eta|\nu)$ converges to $D_{KL}(\eta|\nu)$ in the limit $p \downarrow 1$ (if all these divergences are finite). The convergence is based on the fact that $x \approx e^x - 1$ if x is close to zero. Hence, it follows that $\frac{e^{(p-1)y}-1}{p-1} \to y$ as $p \downarrow 1$. Example 1 shows that, for fixed $p > 1$, there is nevertheless a considerable gap between the functions H in the two cases: Although H grows linearly in the case of KL-divergence, it grows exponentially for polynomial divergences.[6] In Section 3, we discuss examples of divergences that lie between these extremes. In these examples, H has, respectively, polynomial growth behavior, or superpolynomial but subexponential growth behavior.

2.2. Worst-Case Analysis

In this section, we study the worst-case expected value of the random variable X under all distributions within a radius $\kappa > 0$ from the reference model ν,

$$W(\nu, F, \kappa) = \sup_{\eta: D_F(\eta|\nu) \le \kappa} E_\eta[X] = \sup_{\eta: D_F(\eta|\nu) \le \kappa} \int_{\mathcal{S}} xg(x)dx, \tag{4}$$

where distances between models are measured by an F-divergence that satisfies the assumptions of the previous section. As we define the supremum as the worst case, X should be thought of as a loss. Equation (4) thus spells out the worst-case expected loss for a given triple (ν, F, κ). From a risk management perspective, the quantity $W(\nu, F, \kappa)$ coincides with the F-entropic value at risk, a coherent risk measure introduced in Ahmadi-Javid (2012). In Section 4, we extend the discussion to best-case expected values.

In the following, we derive a condition that ensures that the worst-case problem for our F-divergence has a finite solution and provide an expression for the worst case. To achieve this, we build on results of Breuer and Csiszár (2016). The major difference to their more general approach is that our sufficient conditions

and worst-case densities can be formulated explicitly in terms of the function H without invoking the machinery of convex analysis.

Proposition 1. *Let $L = \lim_{y\to-\infty} H(y) \in [-\infty, 0)$. Assume that for all $\alpha \in [0,\infty)$ it holds that*

$$\int_{\mathcal{S}} |x| \exp\big(H^{-1}(\alpha|x|)\big) f(x)dx < \infty. \tag{5}$$

Then the worst case in (4) is finite and there exists $\kappa_{\max} \in (0,\infty]$ such that for all $\kappa \in (0,\kappa_{\max})$ there exists $(\alpha_1^{wc}, \alpha_2^{wc}) \in \mathbb{R}\times(0,\infty)$ such that the worst-case model η^{wc} in (4) has the density

$$g^{wc}(x) = \mathbf{1}_{\{\alpha_1^{wc}+\alpha_2^{wc}x>L\}} \exp\big(H^{-1}\big(\alpha_1^{wc}+\alpha_2^{wc}x\big)\big) f(x), \quad x \in \mathcal{S}. \tag{6}$$

Moreover, for all $\kappa \in (0,\kappa_{\max})$ the parameters $(\alpha_1^{wc}, \alpha_2^{wc}) \in \mathbb{R}\times(0,\infty)$ are uniquely characterized by the conditions that g^{wc} integrates to 1 and that $D_F(\eta^{wc}|\nu) = \kappa$. Furthermore, if the support of ν is unbounded from above, that is, $\sup\mathcal{S} = \infty$, then $\kappa_{\max} = \infty$ and the previous statements hold for all $\kappa > 0$.

For KL-divergence and most other concrete examples of F-divergences considered in this paper, we have $L = -\infty$ so that the indicator function in the worst-case density can be ignored. For polynomial divergence with $p > 1$, we obtain $L = -\frac{1}{p-1}$. Intuitively, the positive parameter α_2^{wc} can be understood as a generalized *exponential tilting* parameter that increases the weight on large realizations of X compared with the reference model. The parameter α_1^{wc} is then chosen in such a way that g^{wc} is a probability density, balancing the effect of α_2^{wc}.

To understand the role of $\kappa_{\max}$, consider for a moment a discrete example where X takes values 1 and 0 with equal probability under ν. X can be thought of as the indicator for some adverse event. In this case, the worst possible alternative model η^* puts all probability mass on the adverse outcome 1 so that $E_{\eta^*}[X] = 1$. Thus, the worst possible change of measure doubles the probability of one outcome while setting the probability of the other to zero. A quick calculation shows that $D_{KL}(\eta^*|\nu) = \log(2)$.[7] The worst possible case can thus be achieved within a finite KL-radius of the nominal model, and we have $\kappa_{\max} = \log(2)$. For $\kappa < \kappa_{\max}$, the worst case can be characterized using a discrete analogue of (6). For $\kappa \geq \kappa_{\max}$, we simply have $W(\nu, F, \kappa) = 1$ and worst-case probabilities of 1 and 0.

In principle, the same situation could arise for continuous distributions. If the support of ν is bounded from above then $\kappa_{\max}$ may be finite and thus the worst-case model is given by (6) only for sufficiently small radii κ. In Appendix B, we investigate this issue further, showing that $\kappa_{\max}$ is infinite even with bounded support if we use KL-divergence, polynomial divergence or one of the new divergences from Section 3 and if X is continuously distributed near the top of its support under ν.

The integral condition (5) of Proposition 1 ensures that the worst case is finite and the worst-case model is of the form (6) for sufficiently small radii κ. Because by assumption we have that $\int_{\mathcal{S}} |x| f(x)dx < \infty$ and that H^{-1} is bounded on every compact subset of $[0,\infty)$, it follows that condition (5) is always satisfied if $\mathcal{S}$ is bounded. The following lemma gives a simple sufficient condition for (5) for cases where $\mathcal{S}$ is unbounded, postulating that the integrand in (5) vanishes faster than $|x|^{-1}$. Moreover, we show that if a reference model satisfies (7), then this property also holds for all more light-tailed reference models.

Lemma 2.

(i) *A sufficient condition for (5) is that for all $\alpha \in (0,\infty)$ there exists $\epsilon > 0$ such that*[8]

$$\limsup_{|x|\to\infty} |x|^{2+\epsilon} \exp\big(H^{-1}(\alpha|x|)\big) f(x) < \infty. \tag{7}$$

(ii) *Consider two reference models ν and $\hat{\nu}$ with respective densities f and $\hat{f}$. Assume that (7) is satisfied for ν and that*

$$\limsup_{|x|\to\infty} \frac{\hat{f}(x)}{f(x)} < \infty. \tag{8}$$

Then $\hat{\nu}$ satisfies (7) as well.

For the case of KL-divergence, proposition 3.1 in Glasserman and Xu (2014) provides an extension of (6) to the case where additional constraints of the form $E_\eta[\Phi(X)] \leq \lambda$ are imposed on the worst-case distribution. Although a thorough analysis of this case is beyond the scope of this paper, we note that formally going through the Lagrangian analysis (similar to Kruse et al. 2019, p.432) shows that the worst-case density with additional constraints is given by

$$g^{wc}(x) = \mathbf{1}_{\{\alpha_1^{wc}+\alpha_2^{wc}x+\alpha_3^{wc}\Phi(x)>L\}} \exp\big(H^{-1}\big(\alpha_1^{wc}+\alpha_2^{wc}x + \alpha_3^{wc}\Phi(x)\big)\big) f(x), \quad x \in \mathcal{S}, \tag{9}$$

analogously to the statement in Glasserman and Xu (2014), where the additional parameter α_3^{wc} is determined by the additional constraint.

2.3. Contents of Divergence Balls

The third component of our toolkit consists of two propositions that investigate the contents of divergence balls around a given reference model implied by a choice of H. Understanding what a divergence ball of a given radius contains is crucial for interpreting

the results of a worst-case analysis from an applied point of view.

For the baseline case of a KL-divergence ball around a Gaussian nominal model, the situation is clear: KL-divergence balls around Gaussian reference models contain models with finite second (and lower) moments but no models for which the second moment is infinite. Put differently, we can observe some densities that behave like $|x|^{-t-1}$ in the tail (or tails) when studying a KL-divergence ball around a Gaussian model if $t > 2$. No such densities exist in the ball for $t < 2$. Gaussian models thus admit some power-law models in the divergence ball but not all. We are interested in extending this type of result to general H and ν.

In Kruse et al. (2019), it is shown that it crucially depends on the interplay between the reference model and the F-divergence to decided which alternative models are taken into consideration. For a broad class of reference models that includes (heavy-tailed) Weibull and lognormal models, the traditional choices of KL-divergence and polynomial divergence have their drawbacks: KL-divergence balls contain power laws with arbitrarily heavy tails so that the worst-case expected value is infinite even in arbitrarily small divergence balls. In contrast, polynomial divergence balls only contain models for which all moments are finite and for which the heaviness of tails is similar to the reference model.

In the following, we demonstrate that by creating a suitable match between the function H and the reference model we can control the number of moments that alternative models in the divergence ball possess.[9] The motivation for this analysis is twofold. First, it suggests a method for constructing divergences that come with an explicit control on the heaviness of tails of alternative models. We apply this method to concrete classes of distributions in the next section. Second, on a more abstract level, the analysis here and, especially, the examples in the next section demonstrate that defining divergences in terms of the function H enables us to tailor divergences to specific needs without giving up closed-form worst-case densities.

Up to some technical conditions, the two results can be summarized as follows: Suppose that there exists $\theta > 1$ such that the functions $H(\varphi(x))$ and x^θ have a similar growth behavior as $x \to \infty$ where we define $\varphi = -\log(f)$. Then all models η that admit a finite moment of order θ have a finite divergence $D_F(\eta|\nu)$ (Proposition 2). Finiteness of higher moments is not guaranteed, however. Conversely, all sufficiently regular models whose tails are heavier than x^{-t-1} for some $t < \theta$ have infinite divergence from ν (Proposition 3).[10] In the initial example of KL-divergence, $H(y) = y$, and a Gaussian reference model, $\varphi \sim x^2$, we find that $H(\varphi(x))$ is a quadratic polynomial. Indeed, the second moment, $\theta = 2$, is the critical level for inclusion in the divergence ball. For the construction of new divergences in the remainder of the paper, the main implication of the two propositions is that they suggest to create a match between the nominal model and the divergence by choosing H such that $H(\varphi(x))$ grows like x^θ for some θ.

Proposition 2. *Let $\theta > 1$. Assume that $F(0) < \infty$ and that*[11] $\limsup_{|x|\to\infty} \frac{H(\varphi(x))\mathbf{1}_S(x)}{|x|^\theta} < \infty$, *where $\varphi = -\log(f)$. Let η be a distribution on $\mathcal{S}$ that is absolutely continuous with respect to ν and has density g. Suppose that $\int_{\mathcal{S}\cap[-k,k]} F(\frac{g(x)}{f(x)})f(x)dx < \infty$ for all $k \in (0,\infty)$, that $\limsup_{|x|\to\infty} g(x) < 1$, and that $\int_{\mathcal{S}} |x|^\theta g(x)dx < \infty$. Then $D_F(\eta|\nu) < \infty$.*

Remark 2. In the preceding proposition, we consider alternative models with finite θth moment $\int_0^\infty |x|^\theta g(x)dx < \infty$. As argued previously (see also Kruse et al. 2019), when one wants to include models with moments up to order θ in the uncertainty sets, the divergence needs to be tailored to the reference model (and the threshold θ). This observation is reflected by the condition $\limsup_{|x|\to\infty} \frac{H(\varphi(x))\mathbf{1}_S(x)}{|x|^\theta} < \infty$, which connects the divergence D to the reference model η and the threshold θ. The indicator function $\mathbf{1}_S$ ensures that the condition only imposes a growth assumption where the support $\mathcal{S}$ of ν is unbounded. The remaining conditions are also needed to ensure that $D_F(\eta|\nu) < \infty$ but they do not concern the tail behavior of η. For example, ν being not absolutely continuous with respect to η and $F(0) = \infty$ would imply $D_F(\eta|\nu) = \infty$. Therefore, we impose $F(0) < \infty$. This assumption holds true for KL-divergence, α-divergence, and the new divergences proposed in the next section. Likewise, the condition $\int_{\mathcal{S}\cap[-k,k]} F(\frac{g(x)}{f(x)})f(x)dx < \infty$ for all $k \in (0,\infty)$ rules out alternative distributions that already have infinite distance to ν on bounded intervals. It is satisfied, for example, if the density $\frac{g}{f}$ of η with respect to ν is bounded from above on every compact subset of $\mathcal{S}$. Finally, $\limsup_{|x|\to\infty} g(x) < 1$ is a weak regularity assumption on η.

We now turn to the converse direction and verify that all sufficiently regular models whose tails are heavier than a power law with some infinite moment of order $t < \theta$ are excluded from any divergence ball around the reference model. To this end, we assume that the support of ν is unbounded from above. A similar statement holds in the case where $\mathcal{S}$ is unbounded from below.

Proposition 3. *Let $\theta > 1$ and let $t \in (1,\theta)$. Assume that $F(0) < \infty$, that there exists $\hat{x} \in (0,\infty)$ such that $(\hat{x},\infty) \subset \mathcal{S}$ and that there exists $\tilde{y} \in (1,\infty)$ such that H is continuously differentiable on $(\tilde{y},\infty)$. Moreover, assume that $\limsup_{y\to\infty} \frac{H'(y)}{H(y)} < \infty$ and that for all $c \in (0,\infty)$ it holds that $\liminf_{x\to\infty} \frac{H(\varphi(x)) - (t+1)\log(x) - c)}{x^t} > 0$. Let η be a distribution on $\mathcal{S}$ that is*

absolutely continuous with respect to ν and has density g. Suppose that $\liminf_{x\to\infty} x^{t+1}g(x) > 0$ *and that* $\lim_{x\to\infty} \frac{g(x)}{f(x)} = \infty$. *Then* $D_F(\eta|\nu) = +\infty$.

3. Explicit Divergences for Weibull and Lognormal Models

In this section, we use the tools assembled in the previous section to propose two new explicit F-divergences. Following the discussion in Section 2.3, the intended reference models for these new divergences are, respectively, of Weibull type and of generalized lognormal type. In both cases, the support of the reference models is the nonnegative real line. From a technical perspective, a reference model is of Weibull type if the log-density φ is a (possibly fractional) polynomial, whereas it is of generalized lognormal type if the log-density is a polynomial in $\log(x)$. Just like polynomial divergence, both new divergences have an additional parameter that can be used to adjust the growth behavior of the cost function and thus the level of the implicit moment constraint in the sense of Section 2.3.

3.1. The Weibull Case

Throughout this section, we assume that $\mathcal{S} = [0,\infty)$ and that there exists a constant $k > 0$ such that the reference model satisfies $\varphi \in \Theta(x^k)$ where $\varphi = -\log(f)$.[12] We refer to this assumption on the tail asymptotics of the reference model as the Weibull case because varying $k > 0$ covers the entire family of Weibull distributions including the heavy-tailed case $k \in (0,1)$. For the heavy-tailed case, previous divergences have considerable limitations as argued in Kruse et al. (2019). However, the assumption on the reference model covers other reference distributions such as exponential tail behavior ($k = 1$) and Gaussian tail behavior ($k = 2$). If X is Gaussian, then $|X|^q$, $q > 0$, satisfies the assumption with $k = 2/q$. In particular, in order to apply the results of this section, it is not necessary that the reference density is available in closed form: Tail asymptotics in the sense of $\varphi \in \Theta(x^k)$ are sufficient.

The divergences we construct can be combined with any model from these classes to obtain divergence balls that respect finiteness of the θth moment as an additional constraint. To be more precise, for $\theta > \max\{k,1\}$, our divergences for the Weibull case are given by

$$H(y) = \begin{cases} \frac{k}{\theta}\left((y+1)^{\frac{\theta}{k}} - 1\right) & \text{for } y \geq 0 \\ y & \text{for } y < 0. \end{cases} \tag{10}$$

The inverse is given by

$$H^{-1}(x) = \begin{cases} \left(\frac{\theta}{k}x + 1\right)^{\frac{k}{\theta}} - 1 & \text{for } x \geq 0 \\ x & \text{for } x < 0. \end{cases} \tag{11}$$

In the limiting case $\theta = k$, we thus recover the KL-divergence $H(y) = y$. Generally, guided by the results of the previous section, the function H is chosen in such a way that $H(\varphi(x))$ behaves like x^θ for large x, as is verified here. The next lemma states some key properties of H.

Lemma 3. *The function H is strictly increasing with $\lim_{y\to\infty} H(y) = \infty$, $H(0) = 0$ and $L = \lim_{y\to-\infty} H(y) = -\infty$. Moreover, H is convex and continuously differentiable with $H'(0) = 1$. Finally, H is increasing in θ.*

The fact that $H(y) \geq y$ for all $y \geq 0$ and $H(y) = y$ for all $y \leq 0$ ensures that $F \geq F_{KL}$ (cf. Example 1) and consequently that the divergences defined through H are weakly larger than the KL-divergence. Similarly, the fact that H increases in θ on $[0,\infty)$ and does not depend on θ on $(-\infty,0)$, implies that the divergence $D_F(\eta|\nu)$ increases in θ. For an uncertainty set with fixed radius κ, this implies that the set of models that are taken into account as alternatives shrinks as we impose finiteness of more moments. The next lemma confirms that the tools of the previous section are applicable in this setting: Worst cases are finite, and the θth moment is the cutoff for inclusion into divergence balls.

Lemma 4. *It holds that $F(0) = 1$, that* $\limsup_{x\to\infty} \frac{H(\varphi(x))}{x^\theta} < \infty$, *that* $\limsup_{y\to\infty} \frac{H'(y)}{H(y)} < \infty$, *and that* $\liminf_{x\to\infty} \frac{H(\varphi(x)) - (t+1)\log(x) - c)}{x^t} > 0$ *for all $t \in (1,\theta)$ and $c \in (0,\infty)$. Moreover, Condition (7) is satisfied. Thus, the assertions of Propositions 1–3 hold true.*

The main implication of Proposition 1 is an explicit formula for worst-case densities, $g^{wc}(x) = h^{wc}(x)f(x)$, where the worst-case change of measure h^{wc} is given by

$$h^{wc}(x) = \begin{cases} \exp\left(\left(\frac{\theta}{k}\left(\alpha_1^{wc} + \alpha_2^{wc}x\right) + 1\right)^{\frac{k}{\theta}} - 1\right) \\ \quad \text{for } \alpha_1^{wc} + \alpha_2^{wc}x \geq 0 \\ \exp\left(\alpha_1^{wc} + \alpha_2^{wc}x\right) \\ \quad \text{for } \alpha_1^{wc} + \alpha_2^{wc}x < 0, \end{cases} \tag{12}$$

and where the two parameters $(\alpha_1^{wc}, \alpha_2^{wc}) \in \mathbb{R} \times (0,\infty)$ are such that g^{wc} is a probability density that has a specified divergence from f.[13] Moreover, Propositions 2 and 3 confirm that the θth moment is indeed the cutoff for inclusion in the divergence ball.

Remark 3. In the previous exposition, we presented the function H in terms of k and θ and thus in dependence on the reference model and a target constraint on moments.

This facilitates an application of Propositions 2 and 3. Alternatively, one can also introduce the same function independently of the reference model with a single parameter $\beta \geq 1$:

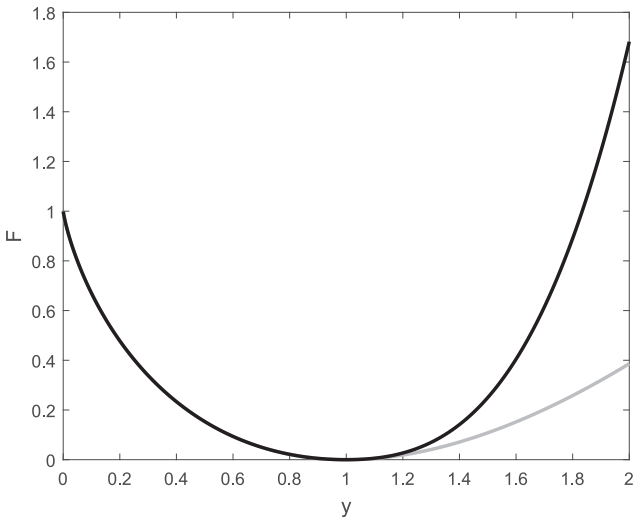
$$H(y) = \begin{cases} \frac{1}{\beta}\left((y+1)^{\beta}-1\right) & \text{for } y \geq 0 \\ y & \text{for } y < 0. \end{cases} \tag{13}$$

Similarly to polynomial divergence, this gives a one-parameter family of divergences that converge to KL-divergence in the limit $\beta \downarrow 1$.[14] The case distinction in the definition of H implies that we treat $y \leq 0$ as in KL-divergence. As y corresponds to the logarithm of the change of measure between the reference and the alternative model, these are exactly the possible outcomes that are more likely under the reference model than under the alternative. Our function differs from KL-divergence for those outcomes that are more likely under the alternative than under the reference model. A piecewise definition like this is convenient to avoid issues such as taking powers of negative numbers. Figure 1 displays the function F implied by (13) for the case $\beta = 1$ that corresponds to KL-divergence (in gray) and for a case with $\beta > 1$ (in black). The functions coincide for $y \leq 1$, whereas for $y > 1$, we see indeed a markedly higher penalty for $\beta > 1$ than in the KL case.

3.2. The Generalized Lognormal Case

In this section, we provide an analogous analysis to the previous section for reference models that are generalized lognormal distributions. Although the assumptions on the reference model are more rigid than in the previous section,[15] the results can nevertheless be transferred to more general settings involving lognormal distributions using Proposition 4. This is demonstrated in the setting of worst-case hedging errors in Section 6.

Figure 1. Function F Induced by (13) for $\beta = 1$ (Gray Line) and for $\beta = \theta/k$ with $\theta = 2$ and $k = 0.3$ (Black Line)



Throughout this section, we assume that the reference model is a generalized lognormal distribution. Thus, the support is given by $\mathcal{S} = [0, \infty)$, and there exist $r \geq 2$, $\sigma > 0$, and $\mu \in \mathbb{R}$ such that

$$f(x) = \frac{1}{Z \cdot x} \exp\left(-\frac{1}{r\sigma^r}|\log(x) - \mu|^r\right), \tag{14}$$

with $Z = 2r^{1/r}\sigma\Gamma(1 + 1/r)$. Similarly, as in the Weibull case, our divergences depend on a parameter $\theta > 1$ that can be interpreted in terms of a moment restriction. The divergence is implicitly defined through

$$H(y) = \begin{cases} \frac{1}{(\theta\sigma)^r}\left(e^{\left(r(\theta\sigma)^r y+1\right)^{\frac{1}{r}}-1} - 1\right) & \text{for } y \geq 0 \\ y & \text{for } y < 0. \end{cases} \tag{15}$$

The inverse is thus given by

$$H^{-1}(x) = \begin{cases} \frac{1}{r(\theta\sigma)^r}\left(\left(\log((\theta\sigma)^r x + 1) + 1\right)^r - 1\right) & \text{for } x \geq 0 \\ x & \text{for } x < 0. \end{cases} \tag{16}$$

The next lemma verifies that this choice of H satisfies the basic properties we need.

Lemma 5. *The function H is strictly increasing with $\lim_{y\to\infty} H(y) = \infty$, $H(0) = 0$ and $L = \lim_{y\to-\infty} H(y) = -\infty$. Moreover, H is continuously differentiable with $H'(0) = 1$. In the standard lognormal case $r = 2$, H is also convex and increasing in θ.*

Notice that we only have convexity of H in the special case $r = 2$. This case corresponds to the standard lognormal distribution found, for example, in financial applications. In that case, we also have monotonicity in θ, implying that with fixed radius a larger θ yields smaller worst-case expectations because of heavier restrictions on the divergence balls. The next lemma confirms that the tools for worst-case analysis from Section 2 are again applicable.

Lemma 6. *It holds that $F(0) = 1$, that $\limsup_{x\to\infty} \frac{H(\varphi(x))}{x^{\theta}} < \infty$, that $\limsup_{y\to\infty} \frac{H'(y)}{H(y)} < \infty$, and that $\liminf_{x\to\infty} \frac{H(\varphi(x)-(t+1)\log(x)-c)}{x^t} > 0$ for all $t \in (1, \theta)$ and $c \in (0, \infty)$. Moreover, Condition (7) is satisfied. Thus, the assertions of Propositions 1–3 hold true.*

In particular, the worst-case densities are given by $g^{wc}(x) = h^{wc}(x)f(x)$ with

$$h^{wc}(x) = \begin{cases} \exp\left(\frac{1}{r(\theta\sigma)^r}\left(\left\{\log\left((\theta\sigma)^r\left(\alpha_1^{wc} + \alpha_2^{wc}x\right)+1\right)+1\right\}^r - 1\right)\right) & \text{for } \alpha_1^{wc} + \alpha_2^{wc}x \geq 0 \\ \exp\left(\alpha_1^{wc} + \alpha_2^{wc}x\right) & \text{for } \alpha_1^{wc} + \alpha_2^{wc}x < 0. \end{cases} \tag{17}$$

Finally, although the divergence is defined in terms of the reference model via σ and r, the dependence on σ

can formally be removed by defining the divergence via

$$H(y) = \begin{cases} \frac{1}{\beta^r}\left(e^{(r\beta^r y+1)^{\frac{1}{r}-1}} - 1\right) & \text{for } y \geq 0 \\ y & \text{for } y < 0, \end{cases}$$

for $\beta \geq 0$. For a generalized lognormal reference model with volatility parameter σ, finiteness of the $\frac{\beta}{\sigma}$th moment is thus the threshold for inclusion of alternative models into divergence balls as discussed in Propositions 2 and 3.

4. Risk Factors, Transformations, and the Wasserstein Distance

In many applications, the univariate output quantity X that is the object of our worst-case analysis depends in a complex way on many driving risk factors $Z = (Z_1, \ldots, Z_n)$. For example, in Section 6, we study a financial setting where X is the absolute hedging error of a discrete hedging strategy, whereas the Z_i represent stock prices at different time points. In a situation like this, it seems plausible that the true model uncertainty concerns the joint distribution of the stock prices Z_i while the mapping $X = G(Z)$ from prices Z to terminal portfolio values X is a known deterministic function. The two main results of this section, Propositions 4 and 5, reconcile this multivariate viewpoint with the univariate worst-case analysis of Section 2.2. Proposition 4 shows how to transfer the sufficient condition (5) from Proposition 1 from the risk factors to the output quantity X. Proposition 5 shows that the univariate, output-based worst-case analysis of Section 2.2 is equivalent to a multivariate worst-case analysis at the level of the risk factors. In the final part of this section, we argue that this equivalence is an important advantage of the F-divergence approach to model uncertainty compared with the Wasserstein approach, which has recently attracted considerable attention (Pflug and Pichler 2014, Esfahani and Kuhn 2018).

4.1. Sufficient Conditions for Multivariate Problems

In typical applications where X is a function of many risk factors, $X = G(Z)$, the marginal distributions of the Z_i are chosen from well-known parametric families, whereas the nominal distribution of X is more complex. The nominal density f of X may thus not be available explicitly. This makes a direct application of Proposition 1 to X difficult because Condition (5) cannot be verified directly. In Proposition 4, we provide sufficient conditions under which the nominal distribution of X inherits Condition (5) from the marginal distributions of the risk factors.

For instance, in the hedging example of Section 6, the stock prices Z_i at different time points are lognormally distributed under the nominal model, whereas the absolute hedging error X is a complicated nonlinear function of the Z_i whose density is not known in closed form. Using Proposition 4, we can verify Condition (5) individually for the risk factors Z_i and then conclude that the condition is inherited by X.

Proposition 4. *Fix a function H that satisfies the standing assumptions from the beginning of Section 2.1 and $n+1$ reference models ν and $\nu_1, \ldots, \nu_n$. Assume* (i) *that there are random variables X and $Z_1, \ldots, Z_n$ with marginal distributions $X \sim \nu$ and $Z_i \sim \nu_i$ and a constant C such that*

$$|X| \leq C\left(1 + \sum_{i=1}^{n} |Z_i|\right),$$

almost surely (a.s.), (ii) *that the distributions $\nu_1, \ldots, \nu_n$ all satisfy (5), and* (iii) *that there exists $\bar{x} \in [0, \infty)$ such that $x \mapsto x\exp(H^{-1}(x))$ is convex on $(\bar{x}, \infty)$. Then ν also satisfies (5).*

The proposition is based only on the marginal distributions and does not make any assumptions on the dependence structure of the Z_i. The convexity condition (iii) on $x \mapsto x\exp(H^{-1}(x))$ is satisfied by all specific examples of divergences considered in this paper.

4.2. Equivalence of Univariate and Multivariate Worst-Case Analysis

When X is a function of many underlying risk factors Z_i, it may seem more natural to formulate uncertainty sets at the level of the multivariate joint distribution of the Z_i than at the level of the univariate distribution of X like we did thus far. We next show that this does not affect the outcome of a worst-case analysis. If we define the uncertainty set with the same F and κ on the joint distribution of the risk factors rather than on the marginal distribution of X, we obtain the same worst-case distribution and worst-case expectation for X. Moreover, the worst-case joint distribution of the Z_i can be recovered from the univariate worst-case analysis. This equivalence between the univariate and multivariate worst-case problems under F-divergence is established in Proposition 5.

Proposition 5. *Suppose the real-valued random variable X satisfies our standing assumptions, and there exists a random variable $Z = (Z_1, \ldots, Z_n)$ taking values in $U \subseteq \mathbb{R}^n$ such that $X = G(Z)$ a.s. for some measurable function $G : \mathbb{R}^n \to \mathbb{R}^+$. Denote by ν_Z the distribution of Z under the reference model and assume that it has Lebesgue density $f_Z : U \to \mathbb{R}$. In particular, the implied distribution of $G(Z)$ under ν_Z is ν. Consider the multivariate analogue of (4) with the same F and κ,*

$$\sup_{\eta_Z : D_F(\eta_Z|\nu_Z) \leq \kappa} E_{\eta_Z}[G(Z)], \tag{18}$$

where the supremum now runs over n-variate distributions η_Z and the F-divergence is the n-variate generalization of (1), that is,

$$D_F(\eta_Z|\nu_Z) = \int_U F\left(\frac{g_Z(z)}{f_Z(z)}\right) f_Z(z)dz,$$

with g_Z denoting the Lebesgue density of η_Z. Then the two suprema in (4) and (18) coincide. Moreover, if the supremum in (4) is finite and attained by a distribution with density $g^{wc}(x) = h^{wc}(x)f(x)$, then the supremum in (18) is attained by

$$\hat{g}_Z^{wc}(z) = h^{wc}(G(z))f_Z(z). \tag{19}$$

Likewise, if the supremum in (18) is finite and attained by a distribution with density $\hat{g}^{wc}(z) = h_Z^{wc}(z)f_Z(z)$, then the supremum in (4) is attained by

$$\hat{g}^{wc}(x) = E_{\nu_Z}\left[h_Z^{wc}(Z)|G(Z) = x\right]f(x).$$

Remark 4. (i) Already the univariate version of this result has some interesting implications. Suppose that Z is a real-valued random variable and $X = G(Z)$ is a monotonic transformation of Z. Then, for any fixed radius κ, it does not make a difference whether we maximize the expectation of X over an F-divergence ball around the nominal distribution of X or the expectation of $G(Z)$ over an F-divergence ball around the nominal distribution of Z. Transforming the data does not change the worst case. (ii) As a special case of (i), Proposition 5 enables us to translate the worst-case analysis of Proposition 1 into a best-case analysis. Denote by $\bar{\nu}$ the nominal distribution of $Z = -X$. We can write the best-case problem for X as

$$\begin{aligned}\inf_{\eta:D_F(\eta|\nu)\le\kappa} E_\eta[X] &= -\left(\sup_{\eta:D_F(\eta|\nu)\le\kappa} E_\eta[-X]\right)\\ &= -\left(\sup_{\bar{\eta}:D_F(\bar{\eta}|\bar{\nu})\le\kappa} E_{\bar{\eta}}[Z]\right),\end{aligned}$$

where the second step applies Proposition 5 with $G(z) = -z$. We thus arrive at a worst-case problem for $-X$, which can be solved using Proposition 1.[16] Denoting by $\bar{g}^{wc}$ the resulting worst-case density for $-X$, we find that the best-case density for X is of the form

$$\begin{aligned}g^{bc}(x) = \bar{g}^{wc}(-x) &= \mathbf{1}_{\{\alpha_1^{bc}-\alpha_2^{bc}x>L\}}\\ &\times \exp\left(H^{-1}\left(\alpha_1^{bc}-\alpha_2^{bc}x\right)\right)f(x), \quad x\in\mathcal{S},\end{aligned} \tag{20}$$

with $(\alpha_1^{bc}, \alpha_2^{bc}) \in \mathbb{R}\times(0,\infty)$. The only difference to the worst-case density is thus the minus sign in front of α_2^{bc}, corresponding to an exponential tilting toward small realizations of X.

A major consequence of Proposition 5 is that a worst-case analysis that is based on the univariate worst case problem (4) leads to the same result as a multivariate worst-case analysis that is based on perturbing the distributions of the underlying risk factors. Moreover, by (19), the solution to the univariate problem pins down an explicit worst case for the joint distribution of the risk factors.

4.3. Comparison with the Wasserstein Approach

Our observed equivalence between the univariate and multivariate worst-case problems stands in marked contrast to what happens under the main competitor of the F-divergence approach to model uncertainty: the Wasserstein approach. For two distributions μ_1 and μ_2 on $\mathbb{R}^n$, the Wasserstein distance d_W is given by

$$d_W(\mu_1,\mu_2) = \inf_{(X_1,X_2)\sim\mu\in M(\mu_1,\mu_2)} E_\mu[|X_1 - X_2|],$$

where $|\cdot|$ denotes the Euclidean norm. The infimum runs over the set $M(\mu_1,\mu_2)$ of all couplings of μ_1 and μ_2, that is, over the set of all possible joint distributions of X_1 and X_2 on $\mathbb{R}^n\times\mathbb{R}^n$ such that $X_1\sim\mu_1$ and $X_2\sim\mu_2$.

The Wasserstein analogues of the two equivalent worst-case problems in Proposition 5 are given by

$$\sup_{\eta:d_W(\eta,\nu)\le\kappa} E_\eta[X] \tag{21}$$

and

$$\sup_{\eta_Z:d_W(\eta_Z,\nu_Z)\le\kappa} E_{\eta_Z}[G(Z)]. \tag{22}$$

These two problems are far from equivalent. The univariate problem (21) is almost degenerate. The worst-case distribution is simply an upward shift of ν by κ.[17] The worst-case expected value is thus $E_\nu[X] + \kappa$, simply adding the radius as a safety margin.[18] In particular, this univariate Wasserstein approach does not respect the structure of the problem; that is, there does not necessarily exist a distribution of the underlying risk factors that can generate the worst-case distribution for $G(Z)$. For instance, if $X = G(Z)$ is a probability that takes values in $[0,1]$ depending on the realization of the risk factors, the worst-case model may shift X outside the unit interval. The resulting worst case cannot be rationalized with realizations of Z or meaningfully interpreted as a probability. The multivariate problem (22) does not face these drawbacks because it directly perturbs the distribution of the risk factors. However, despite impressive progress in the recent literature (Esfahani and Kuhn 2018 and the references therein), solving this problem may be quite challenging depending on the complexity of G and the dimensionality of Z.

The F-divergence approach provides an attractive middle ground between the extremes of the two Wasserstein problems (21) and (22). Solving the univariate worst-case problem using Proposition 1 may not be quite as easy as adding a safety margin, but depending on the setting, it may be much easier than solving a complex, high-dimensional optimization problem. Proposition 5 guarantees that the solution of the univariate problem coincides with that of its multivariate version. This guarantee does not even require explicit knowledge of the function G and the risk factors Z. It does not matter which model of the form $X = G(Z)$ generates the nominal distribution of X. The univariate F-divergence worst case can be rationalized within any such model (i.e., by Proposition 5, there always exists a distribution of Z that generates the right distribution of $G(Z)$).

The equivalence between the univariate and the multivariate worst-case problems under F-divergence arises because the worst-case change of measure in the multivariate problem is a deterministic function of $X = G(Z)$. Thus, even though the uncertainty set in the multivariate problem is much richer than in the univariate one, the worst case is achieved by a change of measure that lies in both uncertainty sets and thus only depends on X. From a practical perspective, this implies that a worst-case analysis based on Proposition 1 can be conducted in a postprocessing step given only a sample of X without precise knowledge of the underlying model, exactly in the same way in which one might compute a sample mean or standard deviation.[19] This direction is explored further in Section 5.

5. Numerical Approach

In this section, we outline our computational strategy and introduce divbox, a Python package for worst- and best-case analysis with F-divergences.

5.1. Computational Strategy

5.1.1. Nonlinear Importance Sampling. Our numerical approach is based on Monte Carlo simulation with importance sampling as proposed in Glasserman and Xu (2014). To this end, we generate a sample of m independent copies of X, denoted by x_1 to x_m, under the reference model and approximate the worst-case expected value by importance sampling,

$$E_{\eta^{wc}}[X] = E_\nu\left[XT\left(\alpha_1^{wc} + \alpha_2^{wc}X\right)\right] \approx \frac{1}{m}\sum_{j=1}^{m} x_j T\left(\hat{\alpha}_1^m + \hat{\alpha}_2^m x_j\right), \tag{23}$$

where the function T is the tilting function in the change of measure in (6), that is,

$$g^{wc}(x) = T\left(\alpha_1^{wc} + \alpha_2^{wc}x\right)f(x).$$

As we do not know the exact values of α_1^{wc} and α_2^{wc}, we determine their approximations $\hat{\alpha}_1^m$ and $\hat{\alpha}_2^m$ numerically in such a way that the sample equivalents of the constraints $D_F(\eta^{wc}|\nu) = \kappa$ and $E_\nu[h^{wc}(X)] = 1$ are satisfied:

$$\frac{1}{m}\sum_{j=1}^{m} F\left(T\left(\hat{\alpha}_1^m + \hat{\alpha}_2^m x_j\right)\right) = \kappa \tag{24}$$

and

$$\frac{1}{m}\sum_{j=1}^{m} T\left(\hat{\alpha}_1^m + \hat{\alpha}_2^m x_j\right) = 1. \tag{25}$$

5.1.2. Confidence Bounds. We thus approximate the worst-case expected value by the exact worst-case expectation for the empirical distribution of the sample. This is analogous to an approximation of the expected value by a sample mean. However, because of the nonlinear dependence of the parameters $\hat{\alpha}^m$ on the sample, we cannot apply the usual Monte Carlo confidence intervals to quantify the sampling uncertainty in (23). Thus, we resort to classical tools from the econometrics literature. The estimator (23) can be interpreted as a method of moments estimator. In particular, following the reasoning around theorem 2 in chapter 8 of Manski (1988), under certain technical conditions the approximation error

$$\mathcal{E}(m) = E_\nu\left[XT\left(\alpha_1^{wc} + \alpha_2^{wc}X\right)\right] - \frac{1}{m}\sum_{j=1}^{m} x_j T\left(\hat{\alpha}_1^m + \hat{\alpha}_2^m x_j\right)$$

vanishes as m goes to infinity and the distribution of the scaled error $\sqrt{m}\mathcal{E}(m)$ converges to a normal distribution with mean zero and variance $v^2 = w^\top (B^{-1})^\top \Sigma B^{-1} w$. Defining, $U = \alpha_1^{wc} + \alpha_1^{wc}X$, Σ is the covariance matrix of the random vector

$$\begin{pmatrix} T(U) \\ F(T(U)) \\ XT(U) \end{pmatrix}$$

under the nominal model ν, the 3×3 matrix B is defined as

$$B = \begin{pmatrix} E_\nu[T'(U)] & E_\nu[F'(T(U))T'(U)] & 0 \\ E_\nu[T'(U)X] & E_\nu[F'(T(U))T'(U)X] & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and the vector w is

$$w = \begin{pmatrix} E_\nu[XT'(U)] \\ E_\nu[X^2T'(U)] \\ -1 \end{pmatrix}.$$

The basic idea is to linearize the estimator (23) and the constraints around the true parameter values α_1^{wc} and α_2^{wc} to obtain expressions that behave like sample averages and thus satisfy classical stochastic limit theorems. This leads to consistency and asymptotic

normality first for the pair of parameters $(\hat{\alpha}_1^m, \hat{\alpha}_2^m)$ and then also for the estimated worst-case expected value. Finally, to practically compute asymptotic standard errors based on $v^2 = w^\top (B^{-1})^\top \Sigma B^{-1} w$, we estimate B, Σ and w from their sample analogues.

5.1.3. Implementation. To achieve an efficient implementation for the new divergences from Section 3, we need to exploit that evaluating F involves many very similar numerical integrations, especially for large sample sizes. To this end, in every call, we first evaluate F on a grid that covers the entire range of sample values and then use spline interpolation to fill in the gaps. This works very well as the functions F are constructed to be smooth (see Appendix C.2 for more details and an example).

Moreover, we need appropriate starting values when solving the nonlinear system of Equations (24) and (25) for $\hat{\alpha}_1^m$ and $\hat{\alpha}_2^m$.[20] To this end, we exploit the observation from Remark 1 (iii) that, up to first order, all divergences with $F''(1) = 1$ have similar magnitudes in small balls. Thus, starting values that work well for one such divergence can be expected to also work well for the others. The polynomial divergence with $p = 2$ is particularly tractable. For this divergence, we have $T(y) = 1 + y$ and $F(y) = \frac{1}{2}(y^2 - 1)$, where we neglect the indicator function involving L to obtain simpler results. Denoting by a_1 and a_2 the resulting values of α_1^{wc} and α_2^{wc}, (25) becomes $a_1 = -a_2 E_\nu[X]$ and thus (24) turns into $a_2^2 \mathrm{Var}_\nu(X) = 2\kappa$. The positive solution of this equation[21] is given by $a_2 = \sqrt{2\kappa / \mathrm{Var}_\nu(X)}$. Replacing the mean and variance by their empirical counterparts, we obtain starting values for searching $\hat{\alpha}_1^m$ and $\hat{\alpha}_2^m$, which work very well in practice.[22]

5.2. The divbox Package

For KL-divergence, polynomial divergence and the two new divergences from Section 3, we implemented this approach for computing worst- and best-case expected values and asymptotic standard errors in the Python package divbox. As discussed in Section 4, a major advantage of the F-divergence approach to model uncertainty is that we obtain worst-case models that respect the structure of the underlying problem while requiring minimal knowledge about the nominal model in the worst-case analysis. The worst-case analysis happens in a postprocessing step, similar to the computation of a sample mean or sample quantile. A typical application of divbox is as follows:

```
from divbox import supmean
X=SimulateLosses()
WCMean=supmean(X,div='KL',kappa=1)
```

Here, the first line loads the supmean command from divbox. In the second line, SimulateLosses stands for a generic, problem-dependent function that simulates a vector X of realized loss scenarios under the nominal model. This may well be a complex simulation based on a black-box scenario generator. The third line computes the worst-case mean of X over a KL-divergence ball with radius 1 using the methodology outlined previously. A comprehensive introduction to divbox is provided in Appendix C.

In the previous description, we focused on applications of our toolkit within a well-specified nominal model coupled with a Monte Carlo approach. However, the tools from the divbox package can also directly be applied to empirical rather than simulated data. In fact, our reasoning about convergence of the estimator as the sample size increases is based on ideas from the econometrics literature that were originally developed for that type of setting.

It is a well-known advantage of the Wasserstein approach to model risk that, even in a continuous setting, the true data-generating process will lie within a sufficiently large ball around the nominal model. In contrast, under the F-divergence paradigm, a ball around a distribution with discrete support only includes distributions with (at most) the same discrete support. The reason is that the F-divergence approach reweights probability mass, whereas the Wasserstein approach shifts it around. Yet, for the same reason, as argued in Section 4, worst-case postprocessing based on F-divergences respects the structure of an underlying multivariate model, whereas the Wasserstein approach does not. Together with our asymptotic standard errors and the resulting confidence bounds, worst-case expectations that are estimated from a sample are informative about the underlying data-generating process while respecting the structure of the problem.

6. Illustrations

This section illustrates our toolkit in the context of three practical problems from the fields of operations management, insurance, and finance. We focus on situations where the quantity of interest has Weibull or lognormal tail behavior so that the newly constructed F-divergences from Section 3 come in naturally. Of course, these are just some examples for the many applications of heavy-tailed reference models. In actuarial and financial regulation, the lognormal distribution appears in many standard models for losses under Solvency II and Basel III (Frachot et al. 2004, Hürlimann 2009). For applications beyond finance, see Kleiber and Kotz (2003) and Clauset et al. (2009). The latter paper also emphasizes that if one heavy-tailed model is plausible, then it is often hard to rule out others as alternatives, leading to model misspecification risk

regarding tail behavior. Moreover, in all three examples, a worst-case analysis based on KL-divergence would lead to infinite worst cases.

6.1. Inventory Pooling Under Heavy-Tailed Demand

Bimpikis and Markakis (2016) study the value of having a pooled inventory in a static multilocation newsvendor problem where demand at individual locations follows a heavy-tailed distribution. Intuitively, having a pooled inventory for several locations enables the firm to balance demand fluctuations across locations, thus reducing costs from holding a too large or too small inventory compared with a situation where each location has its own inventory. Bimpikis and Markakis (2016) argue that the gains from inventory pooling are smaller under heavy-tailed demand than an intuition based on light-tailed demand models might suggest. The reason is that a sum of heavy-tailed random variables tends to be dominated by very few large summands. In the following, we study how sensitive the gains from inventory pooling are when it comes to model misspecification.

In the model of Bimpikis and Markakis (2016), a firm sells an identical good at n locations. Under the lognormal nominal model, demand is independent across locations and demand D_i at location i is given by $D_i = a\exp(bZ_i)$, where Z_i is standard normal and $(a, b) = (100, 2^{\frac{1}{2}})$. The firm faces backorder costs p for each unit by which demands exceeds inventory and holding costs h for each unit by which inventory exceeds demand where $(h, p) = (1, 1)$. The firm's realized costs $C_d(q)$ under a fully decentralized inventory scheme where each location has inventory q is given by[23]

$$C_d(q) = h\sum_{i=1}^{n}\max(q - D_i, 0) + p\sum_{i=1}^{n}\max(D_i - q, 0).$$

Under a centralized inventory scheme where all locations share the same inventory of size Q, realized costs are given by

$$C_c(Q) = h\max\left(Q - \sum_{i=1}^{n} D_i, 0\right) + p\max\left(\sum_{i=1}^{n} D_i - Q, 0\right).$$

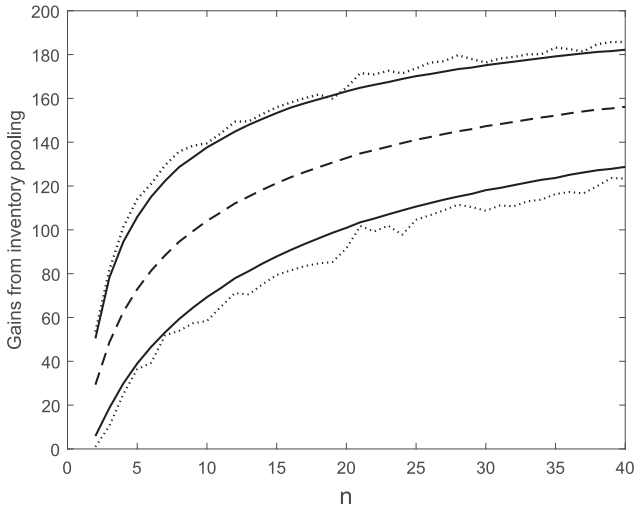
We denote by q_ν^* and Q_ν^* the nominally optimal inventory levels, that is, the respective minimizers of $E_\nu(C_d(q))$ and $E_\nu(C_c(Q))$.[24] The nominal gains per location from inventory pooling are thus given by $E_\nu[X]$ where $X = (C_d(q_\nu^*) - C_c(Q_\nu^*))/n$. In the following, we study how the expectation of X reacts to worst-case perturbations of the underlying demand model for varying numbers of locations n.

Our analysis of worst- and best-case gains from inventory pooling is based on the divergence for the lognormal case $r = 2$ defined in (15) with $\theta = 2$. To determine the volatility parameter σ in the divergence, we apply Proposition 4: Whereas X itself is not lognormally distributed, it is a piecewise linear function of the location-specific demands D_i that are lognormally distributed with volatility parameter b. Thus, we can choose $\sigma = b$.

For each n, we follow a Monte Carlo approach with 500,000 simulations of X. To compute worst and best cases, we use the supmean and infmean commands from the divbox package. To illustrate the asymptotic standard errors from divbox, we also present confidence bounds based on a smaller sample of 1,000 simulations. Here, we add two standard errors to the supmean for an upper bound and subtract two standard errors from the infmean for a lower bound. All computations are based on a radius of $\kappa = 0.1$.

The dashed line in Figure 2 shows the expected gains from centralization under the nominal model as the number of locations increases from 2 to 40. These gains increase from 29.3 to 156.2 as the number of locations goes up. Compared with the expected costs of $E[C_d(q_\nu^*)] = 229.1$ without centralization, this corresponds to cost reductions by, respectively, 12.8% and 68.2%. Thus, there are sizeable economies of scale in inventory pooling. The solid lines show the corresponding best- and worst-case expected values. Here, we observe similar economies of scale although the level of the cost reduction has changed substantially. The dotted lines provide a test of our asymptotic standard errors. Although these more conservative bounds are based on much noisier estimates,

Figure 2. Worst- and Best-Case Expected Gains from Inventory Pooling as Functions of n (Solid Lines)



Note. The dashed line is the nominal mean, and the dotted lines are confidence bounds based on a smaller sample size.

they provide meaningful confidence bounds for our more precise estimates represented by the solid lines.[25]

6.2. Proportional Reinsurance Under Different Claim Dependences

As a second example, we consider proportional reinsurance. The reinsurer provides protection for a fixed percentage of the total claim amount. The total claim amount is equal to the sum of individual claims or losses L_i, $i = 1, \dots, n$. Our quantity of interest is thus

$$R = \rho \sum_{i=1}^{n} L_i, \tag{26}$$

where $\rho \in (0,1)$ is the fraction the reinsurer covers. Under the reference model ν, we assume that loss L_i has a Weibull distribution with parameters (k_i, λ_i). The marginal density of L_i is thus given by

$$f_i(x) = \frac{k_i}{\lambda_i}\left(\frac{x}{\lambda_i}\right)^{k_i-1} e^{-(x/\lambda_i)^{k_i}}.$$

This information about marginal distributions is sufficient to compute the reference mean $E_\nu[R]$. In the following, we study the impact of different assumptions about the dependence structure of the L_i on worst-case expected values. In particular, we compare the cases where the L_i are independent or dependent with a dependence structure modeled through a Gaussian copula or a t-copula. The t-copula is particularly well suited for modeling dependence in the tails. Although the density of R is not available in closed form even in the independent case, Proposition 4 implies that the divergences defined in (10) are applicable with $k = \min_i k_i$. The parameter k in the divergence is thus chosen such that it matches the most heavy-tailed of the risk factors L_i. For the remaining parameter in the divergence, we set $\theta = 2$ throughout the analysis. For the parameters, we consider $n = 5$ risk factors L_i with

$$(k_1, \dots, k_5) = (0.3, 0.4 \dots, 0.7) \quad \text{and} \quad \lambda_i = 1.1^{-\frac{1}{k_i}}$$

and $\rho = 0.3$. The function F with $(\theta, k) = (2, 0.3)$ thus corresponds to the black curve in Figure 1.

We follow a Monte Carlo approach with $m = 500,000$ simulations and apply the divbox package for the worst-case analysis. When the L_i are independent, we simulate independent uniform random variables and transform them into Weibull distributed variables by applying the quantile functions of the marginal Weibull distributions. When assuming a copula for the dependence structure, we similarly transform a sample of dependent uniform random variables into dependent Weibull. The Gaussian copula needs as an input a correlation matrix Σ, whereas the t-copula needs both a correlation matrix and a degrees of freedom parameter v. Throughout, our simulations we use $v = 2$ and

$$\Sigma = \begin{pmatrix} 1 & -0.25 & 0.25 & 0 & 0.25 \\ -0.25 & 1 & 0 & 0.25 & 0.25 \\ 0.25 & 0 & 1 & 0 & 0.25 \\ 0 & 0.25 & 0 & 1 & 0.5 \\ 0.25 & 0.25 & 0.25 & 0.5 & 1 \end{pmatrix}.$$

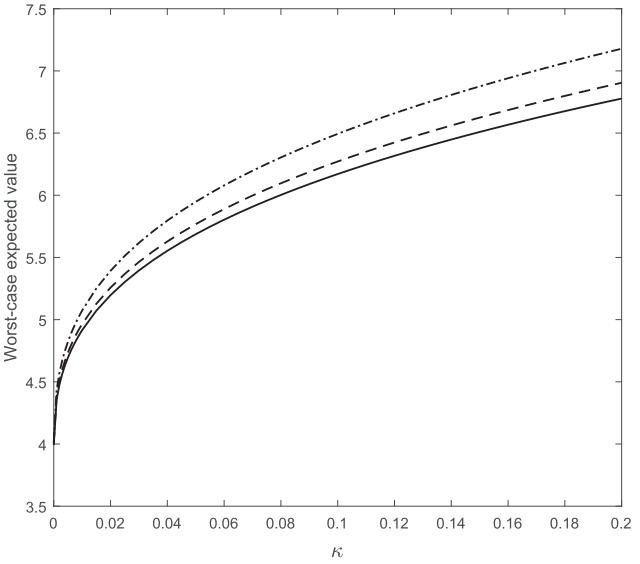
In Figure 3, we illustrate the impact of increasing the radius κ and thereby the model uncertainty on the worst-case expected value under independence, the Gaussian copula and the t-copula. Recall that the expected value under the reference model is identical for all three dependence assumptions and equals $E_\nu[R] \approx 4.02$. As expected, the worst-case expected values increase quickly with κ. Moreover, the worst cases are quite sensitive to the dependence between the risk factors. In particular, the t-copula, which accounts for tail dependence, yields markedly higher worst-case expected values for any given radius.

6.3. Discrete Hedging of a Call Option

In this section, we consider the problem of discrete hedging of a call option under a lognormal (i.e., Black-Scholes) reference model with transaction costs. Model risk for this problem has previously been studied (Glasserman and Xu 2014, Schneider and Schweizer 2015). The latter paper also shows that worst-cases based on KL-divergence are in general infinite for this problem.

The call option has maturity $T > 0$ and strike price K. Its underlying is the stock S, so that the terminal payoff is given by $h(S_T) = \max(S_T - K, 0)$. Under the

Figure 3. Worst-Case Expected Values as Functions of κ for Independent Losses (Solid Line), the Gaussian Copula (Dashed Line), and the t-Copula Case (Dashed-Dotted Line)



reference model, S is a geometric Brownian motion with drift μ_S, volatility σ_S and initial value S_0. Moreover, there is a risk-free bond B in the market with interest rate ρ. Over the time horizon $[0, T]$, a risk manager hedges the option through a trading strategy in the underlying and the bond with portfolio changes at a fixed sequence of n trading dates $t_i = iT/n$, $i = 0, \dots, n-1$. The hedging strategy in S and B is financed by an initial investment. The bond position is adjusted only to finance the changes in the position in S and to cover transaction costs. We denote by $\phi^S_{t_i}$ the number of stocks and by $\phi^B_{t_i}$ the number of bonds held over the interval $[t_i, t_{i+1})$. We assume that the hedge position $\phi^S_{t_i}$ in S is chosen as the position one would hold at t_i under continuous trading in the absence of transaction costs, that is, as a Black-Scholes call delta. This implies in particular that $0 < \phi^S_{t_i} < 1$. The initial investment corresponds to the price of the option in the Black-Scholes model. Thus, in the limit $n \to \infty$ and in the absence of model risk and transaction costs, the hedging strategy would perfectly replicate the payoff. Transaction costs consist of a fixed component k_0 and a component k that is proportional to trading volume. In particular, when the number of stocks held changes by v_i at time t_i, the agent incurs transaction costs of $k_0 + k|v_i|S_{t_i}$.

Our key quantity of interest is the absolute terminal hedging error resulting from such a discretized delta hedge:

$$X = |h(S_T) - \phi^S_{t_{n-1}} S_T - \phi^B_{t_{n-1}} B_T|.$$

Even if the reference model is correct, the claim is not perfectly replicated because of the discrete trading and the transaction costs. Although the distribution of X is not available in closed form, the facts that transaction costs are linear in S_{t_i} and that $\phi^S_{t_i}$ is bounded, $0 < \phi^S_{t_i} < 1$, imply that there exists a constant C such that

$$X \le C\left(1 + \sum_{i=1}^{n} S_{t_i}\right).$$

The individual stock prices S_{t_i} are lognormally distributed with volatility parameter $\sigma_i = \sigma_S\sqrt{t_i}$. We can thus conclude from Proposition 4 that the divergence for the lognormal case $r = 2$ defined in (15) is applicable if we choose $\sigma = \max_i \sigma_i = \sigma_S\sqrt{T}$. We fix $\theta = 2$ throughout the analysis.

We set the parameters of the reference model to $\mu = 0.05$, $\sigma = 0.3$, $\rho = 0.01$, $T = 1$, $S_0 = 1$, and $K = 1$. For the transaction costs, we assume $k_0 = 0.0002$ and $k = 0.005$. The number of portfolio rebalancings n is varied in the following illustration. For the radius, we choose $\kappa = 0.3$. Our illustrations use 500,000 simulated hedging errors and the worst-case computations are based on the `divbox` package.

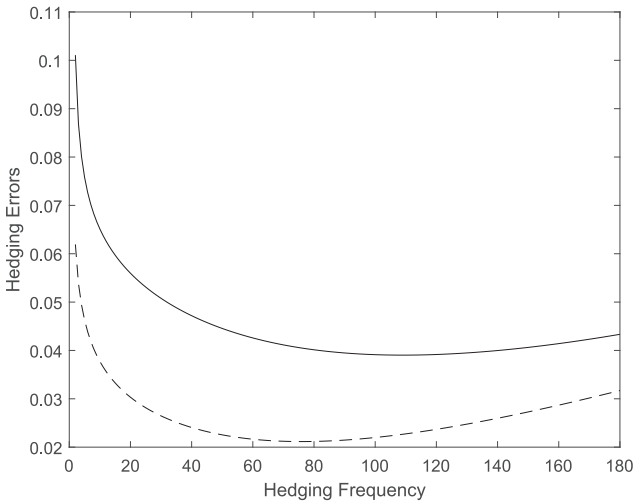
Figure 4 shows how varying the hedging frequency from 12 to 180 times per year affects the absolute hedging error. Increasing the number of portfolio rebalancings first decreases and then increases the hedging error, both under the reference model and under the worst-case model. This U-shape results from a tradeoff between transaction costs and risk reduction. As the hedging frequency increases, the stochastic component of the payoff is replicated more closely, but there are also higher and higher transaction costs. In the worst case, hedging errors are significantly larger. Importantly, the optimal hedging frequency is quite different under the nominal model and the worst case. Although the nominal model implies an optimal number of 76 rebalancings, the worst-case analysis suggests to rebalance the portfolio 109 times and thus 43.4% more frequently.

For all trading frequencies, the hedging errors can be written as functions of the stock prices under the finest trading frequency. In this sense, all the hedging errors in the figure can be interpreted as worst cases over the same uncertainty set by Proposition 5, making the numbers comparable.

7. Managerial Implications

In this paper, we provide a toolkit for making the divergence approach to model uncertainty applicable to a broad set of problems from finance, insurance, and beyond. Our toolkit consists of three components: a new parametrization of F-divergences, a collection of results that define a meaningful worst-case analysis for the new divergences in practical problems, and verifiable conditions on the divergences that enable risk managers to tailor them to specific purposes

Figure 4. Expected Absolute Hedging Errors as Functions of the Number of Portfolio Rebalancings n Under the Reference Model (Dashed Line) and in Worst-Case Expectation (Solid Line)



guided by economic considerations. In particular, unlike some previous approaches, our tools are well suited for lognormal and Weibull-type reference models that are popular modeling frameworks for moderately heavy-tailed phenomena. The Python implementation of our approach in the divbox package makes it easy to integrate into existing risk management processes.

Although the illustrations of the previous section give some indication of the potential practical implications of our approach, they are naturally just first steps in this direction. A brief look at the earlier literature on robustness and model risk based on traditional divergence measures confirms that a lot more can be done. For instance, there are many further potential applications. Any field where risk management calculations meet stochastic modeling is a potential field of application for our toolkit.

There are also further ways in which our tools can be used. In the hedging example, we study the robustness of a specific hedging strategy: delta hedging. In a natural next step, one might compare the robustness of different types of strategies using our approach. For instance, in the hedging example, one could study how a change from delta hedging to delta-gamma hedging affects hedging errors under model uncertainty. From these comparisons, it is only a small conceptual step to robust optimization: A risk manager could use the uncertainty sets implied by our new divergence measures to determine, for example, hedging strategies that respond to model uncertainty in an optimal way. Similarly, in the inventory management example, a next step might be to consider alternative inventory strategies that account for model risk, thus integrating our toolkit into strategic decision.

Finally, thus far, we implicitly assumed that a decision maker is interested in our toolkit because of an intrinsic preference for good risk management under uncertainty. Although this motivation is perfectly valid, other motivations are conceivable as well. For instance, our tools might be used to address model risk when constructing internal models within a regulatory framework such as Solvency II or Basel III. In fact, providing a better protection against model risk has been a common theme of various recent reforms in financial regulation.[26]

Acknowledgments

The authors thank Thomas Breuer, Pavel Cizek, and Volker Krätschmer for valuable comments.

Appendix A. Proofs

Proof of Lemma 1. Clearly, F is continuously differentiable with increasing derivative $F'(y) = H(\log(y))$. Cofiniteness then follows from $\lim_{y\to\infty} F'(y) = \infty$ and L'Hospital's rule. By definition, we have $F(1) = 0$. Convexity and $F'(1) = 0$ imply nonnegativity of F. Finally, if H is continuously differentiable with $H'(0) = 1$, we find $F''(y) = H'(\log(y))/y$ and thus $F''(1) = 1$. □

Proof of Proposition 1. In order to connect our claims to the results of Breuer and Csiszár (2016),[27] we need to introduce some concepts from convex analysis. We extend the definition of F *to* $(-\infty, 0)$ by setting $F(y) = +\infty$ for $y \in (-\infty, 0)$. Let $K : \mathbb{R}^2 \to (-\infty, \infty]$ satisfy for all $(\alpha_1, \alpha_2) \in \mathbb{R}^2$ that

$$K(\alpha_1, \alpha_2) = \int_{\mathcal{S}} F^*(\alpha_1 + \alpha_2 x) f(x) dx,$$

where $F^*:\mathbb{R} \to \mathbb{R}$ is the convex conjugate of F; that is, F^* satisfies for all $x \in \mathbb{R}$ that $F^*(x) = \sup_{y\in\mathbb{R}}(xy - F(y))$. In order to show that the worst case is given by (6), we need to show that $K(\alpha_1, \alpha_2)$ is finite. The fact that H is strictly increasing implies that $F^*(x) = x\exp(H^{-1}(x)) - F(\exp(H^{-1}(x)))$ for all $x > L$ and $F^*(x) = -F(0)$ for all $x \le L$. Moreover, it holds that $(F^*)'(x) = \mathbf{1}_{\{x>L\}} e^{H^{-1}(x)}$. Because F (and hence also $F(0)$) is nonnegative, it holds for all $\alpha_1, \alpha_2 \in \mathbb{R}$ that

$$\begin{aligned} K(\alpha_1, \alpha_2) &\le \int_{\mathcal{S}} \mathbf{1}_{\{\alpha_1+\alpha_2 x>L\}} \Big[(\alpha_1 + \alpha_2 x) e^{H^{-1}(\alpha_1+\alpha_2 x)} \\ &\qquad - F\Big(e^{H^{-1}(\alpha_1+\alpha_2 x)}\Big)\Big] f(x) dx \\ &\le \int_{\mathcal{S}} \mathbf{1}_{\{\alpha_1+\alpha_2 x>L\}} (\alpha_1 + \alpha_2 x) e^{H^{-1}(\alpha_1+\alpha_2 x)} f(x) dx \\ &\le \int_{\mathcal{S}} (|\alpha_1| + |\alpha_2||x|) e^{H^{-1}(|\alpha_1|+|\alpha_2||x|)} f(x) dx. \end{aligned} \tag{A.1}$$

Because $x \mapsto (|\alpha_1| + |\alpha_2||x|)e^{H^{-1}(|\alpha_1|+|\alpha_2||x|)}$ is bounded on any compact subinterval of $\mathbb{R}$ and increasing in $|\alpha_1| + |\alpha_2||x|$, Assumption (5) then ensures that $K(\alpha_1, \alpha_2) < \infty$ for all $(\alpha_1, \alpha_2) \in \mathbb{R}^2$ because $\alpha|x| > |\alpha_1| + |\alpha_2||x|$ for $\alpha > |\alpha_2|$ and $|x|$ is sufficiently large. It follows that K is differentiable on $\mathbb{R}^2$ (see the text preceding (3.20) in Breuer and Csiszár 2016 or Csiszár and Matúš 2012, corollary 3.8). In particular, K is essentially smooth.[28] It follows from Breuer and Csiszár (2016, corollary 4.6) that $\kappa_{\max} > 0$ (in the notation of Breuer and Csiszár 2016). If the support of ν is unbounded to the right, this implies that $\kappa_{\max} = \infty$ (see the last sentence in Breuer and Csiszár 2016, remark 3.2). Because F is convex and satisfies $F(1) = 0$, assumption (3.5) in Breuer and Csiszár (2016) is satisfied. Then, theorem 4.2 of Breuer and Csiszár (2016) ensures that for all $\kappa < \kappa_{\max}$ that there exists $(\alpha_1^{wc}, \alpha_2^{wc}) \in \mathbb{R} \times (0, \infty)$ with

$$\int_{\mathcal{S}} g^{wc}(x) dx = 1 \quad \text{and} \quad \alpha_1^{wc} + \alpha_2^{wc} \int_{\mathcal{S}} x g^{wc}(x) dx - K\big(\alpha_1^{wc}, \alpha_2^{wc}\big) = \kappa, \tag{A.2}$$

where

$$\begin{aligned} g^{wc}(x) &= \big(F^*\big)'\big(\alpha_1^{wc} + \alpha_2^{wc} x\big) f(x) \\ &= \mathbf{1}_{\{\alpha_1^{wc}+\alpha_2^{wc} x>L\}} \exp\big(H^{-1}\big(\alpha_1^{wc} + \alpha_2^{wc} x\big)\big) f(x). \end{aligned} \tag{A.3}$$

Moreover, any $(\alpha_1, \alpha_2) \in \mathbb{R} \times (0, \infty)$ satisfying (A.2) leads to a probability measure η (via (A.3)) that attains the sup in (4). If the support of ν is bounded from above, then the worst case in (4) is clearly finite for every $\kappa > 0$. If the support is unbounded from above, then the worst case in (A.2) is for every $\kappa > 0$ given by $\int_{\mathcal{S}} x g^{wc}(x) dx$ and is finite by (A.2). By strict

convexity of F, the maximizer in (4) is unique (see also the text following display (3.7) in Breuer and Csiszár 2016). It follows that $(\alpha_1^{wc}, \alpha_2^{wc})$ is the unique solution of (A.2) in $\mathbb{R} \times (0, \infty)$. Finally, observe that

$$\begin{aligned}\kappa &= \alpha_1^{wc} + \alpha_2^{wc} \int_{\mathcal{S}} x g^{wc}(x)dx - K\big(\alpha_1^{wc}, \alpha_2^{wc}\big)\\ &= \int_{\mathcal{S}} \big[\big(\alpha_1^{wc} + \alpha_2^{wc} x\big)\big(F^*\big)'\big(\alpha_1^{wc} + \alpha_2^{wc} x\big) - F^*\big(\alpha_1^{wc} + \alpha_2^{wc} x\big)\big] f(x)dx\\ &= \int_{\mathcal{S}} F\left(\frac{g^{wc}(x)}{f(x)}\right) f(x)dx = D_F\big(\eta^{wc}|\nu\big).\end{aligned}$$

Hence, $(\alpha_1^{wc}, \alpha_2^{wc}) \in \mathbb{R} \times (0, \infty)$ is the unique solution of $\int_{\mathcal{S}} g^{wc}(x)dx = 1$ and $D_F(\eta^{wc}|\nu) = \kappa$ in $\mathbb{R} \times (0, \infty)$. □

Proof of Lemma 2. We begin with (i) and fix $\alpha > 0$. Because H^{-1} is bounded from above on every compact subset of $[0, \infty)$, it follows that $\int_{\mathcal{S} \cap [-r,r]} |x| \exp(H^{-1}(\alpha|x|)) f(x)dx < \infty$ for every $r > 0$. Condition (7) ensures that there exists $C > 0$ such that $|x| \exp(H^{-1}(\alpha|x|)) f(x) \le \frac{C}{|x|^{1+\varepsilon}}$ for all $x \in \mathcal{S}$ with $|x|$ large enough. This proves (5). It remains to prove (ii). This claim follows immediately from

$$\begin{aligned}&\limsup_{|x|\to\infty} |x|^{2+\epsilon} \exp\big(H^{-1}(\alpha|x|)\big)\hat{f}(x)\\ &\quad\le \left[\limsup_{|x|\to\infty} |x|^{2+\epsilon} \exp\big(H^{-1}(\alpha|x|)\big) f(x)\right]\left[\limsup_{|x|\to\infty} \frac{\hat{f}(x)}{f(x)}\right] < \infty\end{aligned}$$

where we used that all factors are nonnegative. □

Proof of Proposition 2. Let $\gamma = -\log(g)$. First, observe that the fact that H is increasing implies for all $y > 1$ that

$$F(y) \le (y-1)H\big(\log(y)\big) \le yH\big(\log(y)\big). \tag{A.4}$$

By assumption, there exist $\bar{x}, C \in (0, \infty)$ such that $g(x) \le 1$ and $H(\varphi(x)) \le C|x|^\theta$ for all $x \in \mathcal{S}$ with $|x| \ge \bar{x}$. This implies for all $x \in \mathcal{S}$ with $|x| \ge \bar{x}$ that

$$H\left(\log\left(\frac{g(x)}{f(x)}\right)\right) = H\big(\varphi(x) - \gamma(x)\big) \le H\big(\varphi(x)\big) \le C|x|^\theta.$$

This together with (A.4) implies that

$$\begin{aligned}D_F(\eta|\nu) &= \int_{\mathcal{S}} \mathbf{1}_{\{g(x)\le f(x)\}} F\left(\frac{g(x)}{f(x)}\right) f(x)dx\\ &\quad + \int_{\mathcal{S}} \mathbf{1}_{\{g(x)> f(x)\}} F\left(\frac{g(x)}{f(x)}\right) f(x)dx\\ &\le F(0) + \int_{\mathcal{S}\cap[-\bar{x},\bar{x}]} F\left(\frac{g(x)}{f(x)}\right) f(x)dx\\ &\quad + \int_{\mathcal{S}\cap[-\bar{x},\bar{x}]^c} \mathbf{1}_{\{g(x)> f(x)\}} F\left(\frac{g(x)}{f(x)}\right) f(x)dx\\ &\le F(0) + \int_{\mathcal{S}\cap[-\bar{x},\bar{x}]} F\left(\frac{g(x)}{f(x)}\right) f(x)dx\\ &\quad + \int_{\mathcal{S}\cap[-\bar{x},\bar{x}]^c} \mathbf{1}_{\{g(x)> f(x)\}} H\left(\log\left(\frac{g(x)}{f(x)}\right)\right)\\ &\qquad \times g(x)dx\\ &\le F(0) + \int_{\mathcal{S}\cap[-\bar{x},\bar{x}]} F\left(\frac{g(x)}{f(x)}\right) f(x)dx\\ &\quad + C\int_{\mathcal{S}\cap[-\bar{x},\bar{x}]^c} \mathbf{1}_{\{g(x)> f(x)\}} |x|^\theta g(x)dx < \infty.\end{aligned}$$ □

Proof of Proposition 3. First observe that because $\limsup_{y\to\infty} \frac{H'(y)}{H(y)} < \infty$, there exist $y_0 \in (\tilde{y}, \infty)$ and $C \in (0, \infty)$, such that $H'(y) \le CH(y)$ for all $y \in (y_0, \infty)$. This implies for all $d > 0$ that

$$\begin{aligned}&\frac{\int_{y_0}^{y_0+d} H\big(\log(z)\big)dz}{(y_0+d)H\big(\log(y_0+d)\big) - y_0 H\big(\log(y_0)\big)}\\ &\quad= \frac{\int_{y_0}^{y_0+d} H\big(\log(z)\big)dz}{\int_{y_0}^{y_0+d} H\big(\log(z)\big) + H'\big(\log(z)\big)dz} \ge \frac{1}{1+C}.\end{aligned}$$

This proves that

$$\begin{aligned}\frac{\int_1^{y_0+d} H\big(\log(z)\big)dz}{(y_0+d)H\big(\log(y_0+d)\big)} &\ge \frac{\int_1^{y_0} H\big(\log(z)\big)dz}{(y_0+d)H\big(\log(y_0+d)\big)}\\ &\quad + \frac{1}{1+C}\left(1 - \frac{y_0 H\big(\log(y_0)\big)}{(y_0+d)H\big(\log(y_0+d)\big)}\right).\end{aligned}$$

Taking the limit $d \to \infty$ shows that

$$\begin{aligned}\liminf_{y\to\infty} \frac{F(y)}{yH\big(\log(y)\big)} &= \liminf_{y\to\infty} \frac{\int_1^y H\big(\log(z)\big)dz}{yH\big(\log(y)\big)}\\ &\ge \frac{1}{1+C} > 0.\end{aligned}$$

This implies that there exist $\bar{y}, \epsilon_1 \in (0, \infty)$ such that $F(y) \ge \epsilon_1 yH(\log(y))$ for all $y \in (\bar{y}, \infty)$. By assumption there exist $\bar{x} \in (\hat{x}, \infty)$ and $\delta \in (0, 1)$ such that $g(x) \ge \frac{\delta}{x^{t+1}}$, that $H(\varphi(x) - (t+1)\log(x) + \log(\delta)) \ge \epsilon_2 x^t$, and that $g(x) \ge \bar{y} f(x)$ for all $x \in (\bar{x}, \infty)$. This implies that

$$\begin{aligned}D_F(\eta|\nu) &\ge \int_{\bar{x}}^\infty \mathbf{1}_{\{g(x)\ge \bar{y} f(x)\}} F\left(\frac{g(x)}{f(x)}\right) f(x)dx\\ &\ge \epsilon_1 \int_{\bar{x}}^\infty \mathbf{1}_{\{g(x)\ge \bar{y} f(x)\}} H\left(\log\left(\frac{g(x)}{f(x)}\right)\right) g(x)dx\\ &\ge \epsilon_1 \int_{\bar{x}}^\infty H\big(\varphi(x) - (t+1)\log(x) + \log(\delta)\big)\\ &\quad \times g(x)dx \ge \epsilon_1\epsilon_2 \int_{\bar{x}}^\infty x^t g(x)dx\\ &\ge \epsilon_1\epsilon_2\delta \int_{\bar{x}}^\infty \frac{1}{x}dx = \infty.\end{aligned}$$ □

Proof of Lemma 3. Monotonicity in y, convexity, and the limiting values are obvious. For continuous differentiability, it suffices to consider the left and right derivatives at $y = 0$. We have $\lim_{y\uparrow 0} H'(y) = 1$ and

$$\lim_{y\downarrow 0} H'(y) = \lim_{y\downarrow 0} (y+1)^{\frac{\theta}{k}-1} = 1.$$

To see monotonicity in θ, notice first that $H(y)$ is independent of θ for $y \le 1$. For $y > 1$, we can write

$$H(y) = \int_1^y H'(x)dx = \int_1^y (y+1)^{\frac{\theta}{k}-1}dx.$$

The integrand is increasing in θ and thus so is the integral. □

Proof of Lemma 4. It holds that

$$F(0) = \lim_{y\downarrow 0} F(y) = \lim_{y\downarrow 0} y\log(y) - y + 1 = 1.$$

Next, observe that

$$\limsup_{x\to\infty} \frac{H(\varphi(x))}{x^\theta} = \frac{k}{\theta} \limsup_{x\to\infty} \left[\left(\frac{\varphi(x)}{x^k} + \frac{1}{x^k} \right)^{\frac{\theta}{k}} - \frac{1}{x^\theta} \right] = \frac{k}{\theta} \left(\limsup_{x\to\infty} \frac{\varphi(x)}{x^k} \right)^{\frac{\theta}{k}} < \infty,$$

and that

$$\liminf_{x\to\infty} \frac{H(\varphi(x)-(t+1)\log(x)-c)}{x^t} = \frac{k}{\theta} \liminf_{x\to\infty} \left[\left(\frac{\varphi(x)-(t+1)\log(x)-c+1}{x^{tk/\theta}} \right)^{\frac{\theta}{k}} - \frac{1}{x^t} \right] \geq \frac{k}{\theta} \left(\liminf_{x\to\infty} \frac{\varphi(x)}{x^k} \right)^{\frac{\theta}{k}} > 0,$$

for all $t \in (1,\theta)$ and $c \in (0,\infty)$. Furthermore, it holds that $H'(y) = (y+1)^{\frac{\theta}{k}-1}$ for all $y \in (0,\infty)$, which implies that $\limsup_{y\to\infty} \frac{H'(y)}{H(y)} = 0$. Finally, let $\alpha > 0$ and $\epsilon > 0$. The assumption $\varphi \in \Theta(x^k)$ ensures that

$$\limsup_{x\to\infty} \left(\frac{\left(\frac{\theta}{k}(\alpha x)+1\right)^{\frac{k}{\theta}} - 1 + (2+\epsilon)\log(x)}{x^k} - \frac{\varphi(x)}{x^k} \right) < 0.$$

This implies that

$$\limsup_{x\to\infty} \left(H^{-1}(\alpha x) + (2+\epsilon)\log(x) - \varphi(x) \right) = \limsup_{x\to\infty} \left[x^k \left(\frac{\left(\frac{\theta}{k}(\alpha x)+1\right)^{\frac{k}{\theta}} - 1 + (2+\epsilon)\log(x)}{x^k} - \frac{\varphi(x)}{x^k} \right) \right] = -\infty.$$

Hence, Condition (7) is satisfied. □

Proof of Lemma 5. Monotonicity and the limiting values are obvious. For continuous differentiability, we consider the left and right derivatives at $y = 0$. We have $\lim_{y\uparrow 0} H'(y) = 1$ and

$$\lim_{y\downarrow 0} H'(y) = \lim_{y\downarrow 0} e^{\left(r(\theta\sigma)^r y+1\right)^{\frac{1}{r}-1} \left(r(\theta\sigma)^r y+1\right)^{\frac{1}{r}-1}} = 1.$$

Convexity for $r = 2$ follows from $H''(y) = 0$ for $y < 0$ and

$$H''(y) = e^{\left(2(\theta\sigma)^2 y+1\right)^{-\frac{1}{2}}-1} \times \frac{(\theta\sigma)^2 \left(\left(2(\theta\sigma)^2 y+1\right)^{\frac{1}{2}} - 1 \right)}{\left(2(\theta\sigma)^2 y+1\right)^{\frac{3}{2}}} \geq 0,$$

for $y \geq 0$. Monotonicity of $H(y)$ in θ for $r = 2$ and $y \geq 0$ follows from $\frac{d^2}{d\theta\, dy} H(y) \geq 0$ and $\frac{d}{d\theta} H(y)|_{y=0} = 0$. To see this, note that

$$\frac{d}{d\theta} H(y) = \frac{2y e^{\left(2(\theta\sigma)^2 y+1\right)^{-\frac{1}{2}}-1}}{\theta\sigma\left(2(\theta\sigma)^2 y+1\right)^{\frac{1}{2}}} - \frac{2\left(e^{\left(2(\theta\sigma)^2 y+1\right)^{-\frac{1}{2}}-1} - 1 \right)}{(\theta\sigma)^3},$$

and

$$\frac{d^2}{d\theta\, dy} H(y) = \frac{2\theta\sigma y e^{\left(2(\theta\sigma)^2 y+1\right)^{-\frac{1}{2}}-1}}{\left(2(\theta\sigma)^2 y+1\right)^{\frac{3}{2}}} \times \left(\left(2(\theta\sigma)^2 y+1\right)^{\frac{1}{2}} - 1 \right) \geq 0. \quad \square$$

Proof of Lemma 6. As in the proof of Lemma 4, it holds that $F(0) = 1$. Observe that $\varphi(x) = \frac{1}{r\sigma^r}|\log(x)-\mu|^r + \log(Zx)$. It follows that

$$\begin{aligned}
&\lim_{x\to\infty} \left(r(\theta\sigma)^r \varphi(x) + 1 \right)^{1/r} - \theta\log(x) \\
&= \lim_{x\to\infty} \left(\theta^r |\log(x)-\mu|^r + r(\theta\sigma)^r \log(Zx) + 1 \right)^{1/r} - \theta\log(x) \\
&= \theta \lim_{x\to\infty} \log(x) \left(\left(\left| 1 - \frac{\mu}{\log(x)} \right|^r + \frac{r\sigma^r \log(Zx) + \frac{1}{\theta}}{\log(x)^r} \right)^{1/r} - 1 \right) \\
&= \theta \lim_{h\to 0} \frac{\left(|1-\mu h|^r + r\sigma^r h^{r-1} + \left(r\sigma^r \log(Z) + \frac{1}{\theta} \right) h^r \right)^{1/r} - 1}{h} \\
&= \frac{\theta}{r} \lim_{h\to 0} \frac{-r\mu(1-\mu h)^{r-1} + r(r-1)\sigma^r h^{r-2} + r\left(r\sigma^r \log(Z) + \frac{1}{\theta} \right) h^{r-1}}{\left(|1-\mu h|^r + r\sigma^r h^{r-1} + \left(r\sigma^r \log(Z) + \frac{1}{\theta} \right) h^r \right)^{(r-1)/r}} \\
&= \theta\left(-\mu + (r-1)\sigma^r \mathbf{1}_{\{r=2\}} \right).
\end{aligned}$$

This implies that

$$\limsup_{x\to\infty} \frac{H(\varphi(x))}{x^\theta} = \limsup_{x\to\infty} \frac{1}{(\theta\sigma)^r} \left(e^{\left(r(\theta\sigma)^r \varphi(x)+1 \right)^{\frac{1}{r}} - \theta\log(x) - 1} - \frac{1}{x^\theta} \right) < \infty.$$

Next, observe that, for all $t \in (1,\theta)$ and $c \in (0,\infty)$, it holds that

$$\lim_{x\to\infty} \left(r(\theta\sigma)^r \left(\varphi(x) - (t+1)\log(x) - c \right) + 1 \right)^{1/r} - t\log(x) = \infty.$$

This implies for all $t \in (1,\theta)$ and $c \in (0,\infty)$ that

$$\liminf_{x\to\infty} \frac{H(\varphi(x) - (t+1)\log(x) - c)}{x^t} = \infty.$$

Furthermore, it holds that $H'(y) = (r(\theta\sigma)^r y+1)^{\frac{1}{r}-1} e^{\left(r(\theta\sigma)^r y+1\right)^{\frac{1}{r}}-1}$ for all $y \in (0,\infty)$, which implies that $\limsup_{y\to\infty} \frac{H'(y)}{H(y)} = 0$. Finally, let $\alpha > 0$ and $\epsilon > 0$. It holds that

$$\limsup_{x\to\infty} \left(\frac{\left(\log((\theta\sigma)^r(\alpha x)+1) + 1 \right)^r - 1 + r(\theta\sigma)^r (2+\epsilon)\log(x) - r(\theta\sigma)^r \varphi(x)}{\log(x)^r} \right) = 1 - \theta^r < 0.$$

This implies that

$$\begin{aligned}&\limsup_{x\to\infty}\left(H^{-1}(\alpha x)+(2+\epsilon)\log(x)-\varphi(x)\right)\\&=\limsup_{x\to\infty}\left[\frac{\log(x)^r}{r(\theta\sigma)^r}\left(\frac{\left(\log((\theta\sigma)^r(\alpha x)+1)+1\right)^r-1+r(\theta\sigma)^r(2+\epsilon)\log(x)-r(\theta\sigma)^r\varphi(x)}{\log(x)^r}\right)\right]\\&=-\infty.\end{aligned}$$

Hence, Condition (7) is satisfied. □

Proof of Proposition 4. Let $\alpha>0$. We need to show that

$$E\left[|X|\exp\left(H^{-1}(\alpha|X|)\right)\right]<\infty.$$

Let $\psi:[0,\infty)\to[0,\infty)$ be the function that satisfies for all $x\in[0,\infty)$ that $x\mapsto x\exp(H^{-1}(\alpha x))$. It follows from monotonicity of ψ and the assumption that $|X|\le C(1+\sum_{i=1}^n|Z_i|)$ that

$$\begin{aligned}E\left[|X|\exp\left(H^{-1}(\alpha|X|)\right)\right]&=E\left[\psi(|X|)\right]\\&\le E\left[\psi\left(C\left(1+\sum_{i=1}^n|Z_i|\right)\right)\right].\end{aligned}$$

Moreover, by assumption, ψ is convex for sufficiently large x for $\alpha=1$. This implies that for any fixed α there exists $\hat{x}$ such that ψ is convex on $(\hat{x},\infty)$. Without loss of generality, assume that $C\ge\hat{x}$. It follows from Jensen's inequality that

$$\begin{aligned}&E\left[\psi\left(C\left(1+\sum_{i=1}^n|Z_i|\right)\right)\right]\\&\quad=E\left[\psi\left(\frac{1}{n}\sum_{i=1}^nC(1+n|Z_i|)\right)\right]\le\frac{1}{n}\sum_{i=1}^nE\left[\psi(C(1+n|Z_i|))\right].\end{aligned}$$

For each summand on the right-hand side, we obtain finiteness from (5), arguing as in the proof of Proposition 1 (see the text succeeding (A.1)). This concludes the proof. □

Proof of Proposition 5. For the proof, we first redefine our two optimization problems (4) and (18) as optimization problems over changes of measures. Define sets of measurable functions

$$\begin{aligned}S_X=\{&h:\mathcal{S}\to\mathbb{R}|h\ge0,\\&E_\nu[h(X)]=1,E_\nu[F(h(X))]\le\kappa\}\end{aligned}$$

and similarly

$$\begin{aligned}S_Z=\{&h:U\to\mathbb{R}|h\ge0,\\&E_{\nu_Z}[h(Z)]=1,E_{\nu_Z}[F(h(Z))]\le\kappa\}.\end{aligned}$$

Problems (4) and (18) then become

$$\sup_{h\in S_X}E_\nu[Xh(X)]\quad\text{and}\quad\sup_{h\in S_Z}E_{\nu_Z}[G(Z)h(Z)].$$

We wish to show that both suprema coincide. For $h\in S_X$, define $h_G:U\to[0,\infty)$ by $h_G(z)=h(G(z))$. Then we can write

$$\begin{aligned}\sup_{h\in S_X}E_\nu[Xh(X)]&=\sup_{h\in S_X}E_{\nu_Z}[G(Z)h_G(Z)]\\&\le\sup_{h\in S_Z}E_{\nu_Z}[G(Z)h(Z)].\end{aligned}$$

To see this, it suffices to note that for all $h\in S_X$, we have $h_G\in S_Z$, that is,

$$E_{\nu_Z}[h_G(Z)]=E_{\nu_Z}[h(G(Z))]=E_\nu[h(X)]=1$$

and

$$\begin{aligned}E_{\nu_Z}[F(h_G(Z))]&=E_{\nu_Z}[F(h(G(Z)))]\\&=E_\nu[F(h(X))]\le\kappa.\end{aligned}$$

For the converse inequality, define for $h\in S_Z$ the univariate function

$$\bar{h}(x)=E_{\nu_Z}[h(Z)|G(Z)=x].$$

Indeed, we have the chain of inequalities

$$\begin{aligned}\sup_{h\in S_Z}E_{\nu_Z}[G(Z)h(Z)]&=\sup_{h\in S_Z}E_{\nu_Z}\left[G(Z)E_{\nu_Z}[h(Z)|G(Z)]\right]\\&=\sup_{h\in S_Z}E_\nu\left[X\bar{h}(X)\right]\le\sup_{h\in S_X}E_\nu[Xh(X)],\end{aligned}$$

where the last inequality uses that $h\in S_Z$ implies $\bar{h}\in S_X$. To see this, notice that

$$E_\nu\left[\bar{h}(X)\right]=E_{\nu_Z}\left[E_{\nu_Z}[h(Z)|G(Z)]\right]=E_{\nu_Z}[h(Z)]=1,$$

and, by Jensen's inequality and the convexity of F,

$$\begin{aligned}E_\nu\left[F(\bar{h}(X))\right]&=E_{\nu_Z}\left[F\left(E_{\nu_Z}[h(Z)|G(Z)]\right)\right]\\&\le E_{\nu_Z}[F(h(Z))]\le\kappa.\end{aligned}$$

Thus, the two suprema coincide. Moreover, it follows that if $h^{wc}\in S_X$ attains the supremum in (4), then $h_G^{wc}\in S_Z$ attains the supremum in (18). Similarly, if $h^{wc}\in S_Z$ attains the supremum in (18), then $\bar{h}^{wc}\in S_X$ attains the supremum in (4). □

Appendix B. Behavior with Bounded Support

When the support of the nominal model is bounded from above, then the worst-case expected value converges to the upper bound of the support as the radius κ increases under very mild conditions on the underlying F-divergence. The results of this section clarify whether this convergence to the upper bound, denoted by M, happens gradually as the radius κ converges to infinity or whether an expected value of M can already be achieved for balls with a finite radius. In the latter case, there exists a threshold $\kappa_{\max}<\infty$ such that for all $\kappa\ge\kappa_{\max}$, the worst case expected value equals M. In this section, we show that for the divergences we are interested in the picture is very simple. If the nominal model has an atom at M, $\nu(\{M\})>0$, then the model that puts all mass on M lies within a finite radius of the nominal model. If the nominal model does not possess such an atom on M, then the worst-case expected value is strictly less than M for any finite radius κ. It only converges to M in the limit $\kappa\to\infty$.[29]

Our analysis is based on the study of a function $G(b)$ that corresponds to the minimal radius that is necessary to include an alternative model with mean b in the divergence ball around a given nominal model. The crucial question then becomes whether $G(b)$ converges to infinity as b approaches M or not. We proceed in three steps. In the first

step, we show the result we need for KL-divergence: Up to technical conditions, G_{KL} diverges and $\kappa_{\max}$ is infinite if and only if ν does not possess an atom in M.

Proposition B.1. *Let ν be a distribution with finite mean and support $\mathcal{S} \subset \mathbb{R}$. Suppose that $M = \sup\mathcal{S} < \infty$ and define $m = \inf\mathcal{S} \in \{-\infty\} \cup \mathbb{R}$. For a distribution η with support in $\mathcal{S}$ let $\bar{\eta} = \int_{\mathcal{S}} x\eta(dx) \in [-\infty, M]$ denote its mean. Let $G_{KL}:(m, M] \to [0, \infty]$ be given by $G_{KL}(b) = \inf_{\eta|\bar{\eta}=b} D_{KL}(\eta|\nu)$.*

(i) *If $\nu(\{M\}) = 0$, then $\lim_{b\nearrow M} G_{KL}(b) = \infty$.*

(ii) *If $\nu(\{M\}) > 0$, then $\lim_{b\nearrow M} G_{KL}(b) = -\log(\nu(\{M\}))$.*

Proof of Proposition B.1. Note that G_{KL} is a convex function with $G_{KL}(b) \geq G_{KL}(\bar{\eta}) = 0$.

(i) Kullback's inequality ensures that for all distributions η with $\bar{\eta} \in \mathbb{R}$ and for all $t \in \mathbb{R}$ we have that

$$D_{KL}(\eta|\nu) \geq t\bar{\eta} - \log\left(\int_{\mathcal{S}} e^{tx}\nu(dx)\right).$$

This implies for all $b \in (m, M]$ and $t \in \mathbb{R}$ that

$$\begin{aligned} G_{KL}(b) &\geq tb - \log\left(\int_{\mathcal{S}} e^{tx}\nu(dx)\right) \\ &= -t(M-b) - \log\left(\int_{\mathcal{S}} e^{-t(M-x)}\nu(dx)\right). \end{aligned}$$

By choosing $t = \frac{1}{M-b}$ we obtain that

$$G_{KL}(b) \geq -1 - \log\left(\int_{\mathcal{S}} e^{-\frac{M-x}{M-b}}\nu(dx)\right).$$

Because $e^{-\frac{M-x}{M-b}} \to 0$ for ν-almost all $x \in \mathcal{S}$ as $b \nearrow M$, we obtain with dominated convergence that

$$\begin{aligned} \lim_{b\nearrow M} G_{KL}(b) &\geq -1 - \log\left(\lim_{b\nearrow M}\int_{\mathcal{S}} e^{-\frac{M-x}{M-b}}\nu(dx)\right) \\ &= -1 - \log(0) = \infty. \end{aligned}$$

(ii) For all $n \in \mathbb{N}$, let $\eta_n = (1 - \frac{1}{n})\delta_M + \frac{1}{n}\nu$. Then it holds that $\bar{\eta}_n \to M$ as $n \to \infty$. Moreover, we have that

$$\begin{aligned} G_{KL}(\bar{\eta}_n) &\leq D_{KL}(\eta_n|\nu) \\ &= \int_{\mathcal{S}\setminus\{M\}} F_{KL}\left(\frac{1}{n}\right)d\nu + F_{KL}\left(\frac{1 - \frac{1}{n} + \frac{1}{n}\nu(\{M\})}{\nu(\{M\})}\right)\nu(\{M\}) \\ &= F_{KL}\left(\frac{1}{n}\right)(1 - \nu(\{M\})) + F_{KL}\left(\frac{1 - \frac{1}{n} + \frac{1}{n}\nu(\{M\})}{\nu(\{M\})}\right)\nu(\{M\}) \\ &\to -\log(\nu(\{M\})). \end{aligned}$$

This implies that $\lim_{b\nearrow M} G_{KL}(b) \geq -\log(\nu(\{M\}))$. Convexity of F_{KL} ensures that $\lim_{b\nearrow M} G_{KL}(b) \leq G_{KL}(M) = -\log(\nu(\{M\}))$, which completes the proof. □

In the second step, we show that under mild conditions on the function H the associated F-divergence can be bounded from below by an increasing linear function of KL-divergence.

Lemma B.1. *Consider a function H with associated F-divergence D_F as introduced in Section 2.1. Assume that $\liminf_{y\to\infty} \frac{H(y)}{y} > 0$, that H is differentiable at 0 with $H'(0) > 0$, and that $F(0) < \infty$. Then there exists $\delta_1 > 0$ and $\delta_2 \geq 0$ such that for all nominal models ν and alternative distributions η, we have $D_F(\eta|\nu) \geq \delta_1 D_{KL}(\eta|\nu) - \delta_2$.*

Proof of Lemma B.1. Let $R:\mathbb{R} \to [0, \infty)$ satisfy for all $y \in \mathbb{R} \setminus \{0\}$ that $R(y) = \frac{|H(y)|}{|y|}$ and that $R(0) = |H'(0)|$. Continuity of H, differentiability of H at 0, and the fact that $H(0) = 0$ ensure that R is a continuous function. Moreover, the fact that $H'(0) > 0$ and the fact that H is strictly increasing show that $\inf_{y\in K} R(y) = \min_{y\in K} R(y) > 0$ for all compact subsets $K \subset \mathbb{R}$. This together with the assumption $\liminf_{y\to\infty} R(y) > 0$ implies that there exists $\delta_1 > 0$ such that for all $y \geq 0$, we have $R(y) \geq \delta_1$. Therefore, we obtain for all $x \in [1, \infty)$ that $F(x) = \int_1^x H(\log(z))dz \geq \delta_1 \int_1^x \log(z)dz = \delta_1 F_{KL}(x)$. Furthermore, as $F(0) < \infty$, both F_{KL} and F are bounded and continuous functions over the interval $[0, 1]$. Thus, there exists $\delta_2 \geq 0$ such that $F(x) \geq \delta_1 F_{KL}(x) - \delta_2$ for all $x \geq 0$. This yields the claim. □

Under the conditions of the previous lemma, the F-divergence ball of radius κ is contained in the KL-divergence ball of radius $(\kappa + \delta_2)/\delta_1$. Thus, if an infinite KL-radius is needed to attain an expected value of M, the same is true under the F-divergence. Together with a matching result about the case where ν has an atom in M, this is the content of the next corollary.

Corollary B.1. *Consider a function H with associated F-divergence D_F as introduced in Section 2.1 and assume that $F(0) < \infty$. Let ν be a distribution with finite mean and support $\mathcal{S} \subset \mathbb{R}$. Suppose that $M = \sup\mathcal{S} < \infty$ and define $m = \inf\mathcal{S} \in \{-\infty\} \cup \mathbb{R}$. For a distribution η with support in $\mathcal{S}$, let $\bar{\eta} = \int_{\mathcal{S}} x\eta(dx) \in [-\infty, M]$ denote its mean. Let $G_F:(m, M] \to [0, \infty]$ be given by $G_F(b) = \inf_{\bar{\eta}=b} D_F(\eta|\nu)$, $b \in \mathcal{S}$.*

(i) *Suppose that $\nu(\{M\}) = 0$, that $\liminf_{y\to\infty} \frac{H(y)}{y} > 0$ and that H is differentiable at 0 with $H'(0) > 0$. Then $\lim_{b\nearrow M} G_F(b) = \infty$.*

(ii) *Suppose that $\nu(\{M\}) > 0$, then*

$$\begin{aligned} \lim_{b\nearrow M} G_F(b) &= F(0)(1 - \nu(\{M\})) \\ &\quad + F\left(\frac{1}{\nu(\{M\})}\right)\nu(\{M\}) < \infty. \end{aligned}$$

Proof of Corollary B.1. Claim (i) follows from Proposition B.1 (i) and the observation that by Lemma B.1 there exists $\delta_1 > 0$ and $\delta_2 \geq 0$ such that $G_F(b) \geq \delta_1 G_{KL}(b) - \delta_2$. Thus, when G_{KL} goes to infinity, G_F goes to infinity as well. Claim (ii) follows by the same argument as in part (ii) of Proposition B.1. □

Remark B.1. The divergences for the Weibull and generalized lognormal cases from Section 3 satisfy the conditions in the corollary. To see this, notice first that the associated functions H are differentiable with $H'(0) = 1$ and that $H(y)$ grows faster than linearly for $y > 0$. Moreover, for both divergences, we have $F(0) = F_{KL}(0) = 1$. Similarly, for polynomial divergence with $p > 1$ as defined in Remark 1, we have $F_p(0) = 1/p < \infty$, and the function H grows exponentially and is smooth. Thus, both parts of the corollary apply. In particular, we have $\kappa_{\max} = \infty$ under all three divergences unless ν has an atom in M.

Appendix C. Documentation of the divbox Package

The divbox package implements worst-case and best-case expected values using the methods described in Section 5. The package is written in Python 3 and requires the numpy and scipy packages. divbox provides three commands: supmean, infmean, and supinfsummary. The easiest way to install the package is to copy the file divbox.py into the working directory and to load the commands within Python using, for example:

```
from divbox import supmean, infmean, supinfsummary
```

C.1. Use

All three commands get as first and mandatory input X, a sample of observations from the nominal model. supmean computes the maximal expected value over all alternative models within a given radius kappa of the empirical distribution using a specified divergence. For instance, for 100 draws from a uniform distribution on [0, 1], we can compute the largest and smallest possible expected values over a KL-divergence ball of radius 0.1 via

```
import numpy.random as rand
X=rand.uniform(0,1,100)
sm=supmean(X,div='KL',kappa=0.1)
im=infmean(X,div='KL',kappa=0.1)
```

and get as results, for example, sm = 0.627 and im = 0.378. The command supinfsummary provides those two numbers and additional summary statistics. The printed output of

```
supinfsummary(X,div='KL',kappa=0.1)
```

for this simulation is

```
Nominal Mean: 0.502886434127069 (s.e.: 0.0281273
12339710164)
Sample Range: [ 0.012850703030013455, 0.97788516
91135755 ]
Sample Size: 100
Sup-Mean: 0.6273753261429902 (s.e.: 0.02766167
0689737333)
Effective Sample Size: 83.06296783691072
Maximal Radius kappamax: 4.605170185988092
Inf-Mean: 0.3782897481811803 (s.e.: 0.0277872266
40836125)
Effective Sample Size: 82.99539693100944
Maximal Radius kappamax: 4.605170185988092
```

The first three output lines are descriptive of the data X. The command provides the nominal mean together with its standard error, that is, the sample standard deviation of X divided by the square root of the sample size. It also reports the range from the smallest to the largest sample and the sample size. The next three lines concern the computation of the sup-mean. This mean comes together with its asymptotic standard error as derived in Section 5. We also provide the value of the maximal radius kappamax for the empirical distribution, that is, the minimum necessary radius for shifting all probability mass to the sample maximum.[30] As an additional measure for the reliability of our approximation, we report the effective sample size, a standard diagnostic for weight degeneration in Importance Sampling estimators that can be interpreted as an equivalent number of independent, identically distributed samples that corresponds to our (weighted) sample (see chapter 9.3 of Owen 2013). For small kappa, this is only slightly below the original sample size, whereas for kappa close to kappamax, it declines to 1, reflecting the fact that all probability mass is then concentrated in a single sample. The final three output lines give the corresponding information for the inf-mean.[31]

In addition to KL-divergence, divbox contains polynomial divergences (with parameter p), the Weibull divergences of Section 3.1 (with parameters k and θ), and the generalized lognormal divergences of Section 3.2 (with parameters r, σ, and θ). These can be applied using

```
supinfsummary(X,div='Polynomial', kappa=0.1, p=2)
supinfsummary(X,div='Weibull', kappa=0.1, k=0.5,
  theta=2)
supinfsummary(X,div='Lognormal', kappa=0.1, r=2,
  sigma=1, theta=2)
```

In this example with bounded support, there is little difference between the four divergences regarding whether worst cases are well-defined and regarding the contents and interpretation of the uncertainty sets. These matters are completely different for the heavy-tailed illustrations in Section 6. The file illustrations.py contains functions that simulate scenarios for these settings. In run_illustrations.py, these simulations are evaluated using divbox with suitable divergence choices derived in Section 6.

C.2. Syntax

In the following, we describe in detail the possible input and return arguments of the function supmean. The function infmean takes and returns exactly the same arguments. The function supinfsummary takes the same arguments except the parameter monitor. It does not return any output besides a printed summary.

The only mandatory argument is the vector of data X. Calling supmean(X) simply returns the sample mean. X has to be the first argument. The remaining arguments can be entered in arbitrary order. If a certain divergence misses a given input parameter, that parameter is set to a default value. If a parameter is assigned that is not used by the chosen divergence, then this input is simply ignored. Thus, for example, the following three lines give the same output—the supremum over a KL-divergence ball of radius 0.1:

```
supmean(X,div='KL', kappa=0.1)
supmean(X, kappa=0.1)
supmean(X, theta=3, kappa=0.1, p=2, div='KL')
```

where the second line uses the default value of div, which is 'KL'. The parameters of the remaining divergences are as follows:

- Polynomial divergence is chosen with `div='Polynomial'`. This divergence, defined in Remark 1, has one parameter p > 0 with default value p=2. For p=1, the divergences collapse to KL-divergence.[32]
- Weibull divergence is chosen with `div='Weibull'`. This divergence, defined in (10), has two parameters, a parameter k, which can be interpreted as the maximal heaviness of Weibull tails, and a parameter theta, which can be interpreted as a

bound for existence of moments. The default values are (k,theta)=(0.5,2).[33]

- (Generalized) lognormal divergence is chosen with `div='Lognormal'`. This divergence, defined in (15), has three parameters: a parameter sigma>0, which can be interpreted as a bound on (generalized) volatility; a parameter theta>1, which is similar to Weibull divergence; and a parameter r>1, which parametrizes the generalized lognormal family of distributions. The default values are r=2, which corresponds to the usual lognormal distribution and (sigma,theta)=(1,2).

The remaining arguments monitor, prec, and nbin have a more technical role.

- The argument monitor takes values 0 and 1. Under the default, monitor=0, supmean computes and returns only the sup-mean. With the choice monitor=1, it returns (in this order) the sup-mean, its standard error, the effective sample size, and the value of kappamax. Thus, monitor=1 is, for example, used when calling supmean within supinfsummary, whereas monitor=0 may be preferable when using supmean within the objective of a minimization.
- The argument prec takes values 0 and 1 and determines whether splines are used in the evaluation of the function F for the Weibull and lognormal divergences. Under the default, prec=0, the function F is evaluated using quadrature for input vectors of length less than nbin. For longer input vectors, we compute the $1/(1+\text{nbin})$, $2/(1+\text{nbin})$,..., $\text{nbin}/(1+\text{nbin})$ sample quantiles of the input vector and evaluate the function F on these values and on the extremal sample values using quadrature. Afterward, we use cubic splines to extend the function to the entire sample. The default value of nbin is nbin=100. Under the slower but more precise alternative prec=1, the function F is evaluated using quadrature regardless of sample size. For example, the following three lines compute a sup-mean under the Weibull divergence with default parameters and radius 0.1 with increasing precision:

```
supmean(X, div='Weibull', kappa=0.1)
supmean(X, div='Weibull', kappa=0.1, nbin=1000)
supmean(X, div='Weibull', kappa=0.1, prec=1)
```

For a sample of 10,000 draws from a uniform distribution on $[0,1]$, the three commands have execution times of 0.1, 0.7, and 3.1 seconds, with results of 0.6120488907480726, 0.6120488907488051, and 0.6120488907488054, respectively. When we increase the sample size to a million, computation times become 9.5, 10.3, and 387.8 seconds, with results of 0.6148296693502018, 0.6148296693506876, and 0.6148296693506875, respectively. Thus, in this example, the precision of the computationally cheaper default version seems to be completely sufficient for practical purposes, especially when accounting for additional simulation error.

Endnotes

[1] This notion of criticality is introduced in Kruse et al. (2019) for the special cases of KL-divergence and polynomial divergence.

[2] By density, we always mean density with respect to the Lebesgue measure on $\mathbb{R}$. In particular, $f(x) = 0$ for $x \in \mathbb{R} \setminus \mathcal{S}$.

[3] Absolute continuity with respect to ν implies absolute continuity with respect to the Lebesgue measure.

[4] For $y < 1$, we follow the standard convention that $\int_1^y = -\int_y^1$.

[5] Polynomial divergence is variously known as α-divergence, power-divergence, or Cressie-Read divergence. Moreover, up to monotonic transformations, it coincides with the Rényi divergence and the Tsallis divergence.

[6] Kruse et al. (2019) show that there is a wide class of models for which neither KL-divergence nor polynomial divergence has certain desirable properties for robustness analysis. This is reflected in the vastly different growth behaviors of the functions H we see here.

[7] For more general F-divergences, we have $D_F(\eta^*|\nu) = \frac{1}{2}(F(0) + F(2))$, which is also finite as long as $F(0) < \infty$.

[8] Here and in the following, we denote by $|x| \to \infty$ the pair of limits $x \to \infty$ and $x \to -\infty$.

[9] Of course, this does not mean that such a tailored divergence only suits a single reference model. For instance, KL-divergence is meaningful also for non-Gaussian reference models. It is only the threshold at which existence of moments is (not) guaranteed that changes if we change the reference model or the divergence.

[10] These implicit moment constraints are much weaker than explicit moment constraints as considered in Glasserman and Xu (2014) (see also (9)). By imposing explicit upper bounds on moments of the worst-case distribution, finiteness of KL worst cases can easily be guaranteed. However, the resulting worst-case analysis is then conditional on these moment bounds being justified. In contrast, a worst-case analysis based on our results just requires finiteness of certain moments. It seems much more innocent to assume that, for example, a quantity has a finite variance than to assume that the variance is below an explicit upper bound.

[11] We define the expression in the limit as 0 for $x \notin \mathcal{S}$ such that $\mathbf{1}_{\mathcal{S}}(x) = 0$.

[12] Here $\varphi_1 \in \Theta(\varphi_2)$ means that φ_1 and φ_2 are asymptotically equivalent: $0 < \liminf_{x\to\infty} \frac{\varphi_1(x)}{\varphi_2(x)} \le \limsup_{x\to\infty} \frac{\varphi_1(x)}{\varphi_2(x)} < \infty$.

[13] This function h^{wc} is continuous in x as the terms in the exponent coincide at the cutoff points where $\alpha_1^{wc} + \alpha_2^{wc} x = 0$.

[14] The function H from (13) has a formal resemblance to the function F in the definition of polynomial divergence. However, as H enters the definition of F together with a logarithmic transformation, the function H associated with polynomial divergence looks rather different from the function here (see Remark 1 (ii)). In particular, the polynomial growth behavior of H lies between the linear growth of H in KL-divergence and the exponential growth of H in polynomial divergence. This leads to uncertainty sets that are more restrictive than those under KL-divergence but richer than those under polynomial divergence.

[15] In the Weibull case, it is sufficient to specify tail asymptotics rather than a specific class of reference densities. Intuitively, the lognormal case is harder to handle than the Weibull case because lognormal distributions are more heavy-tailed and closer to power laws than Weibull-type distributions. How well behaved the perturbed densities are thus becomes harder to guarantee.

[16] The technical conditions in Proposition 1 are such that if X satisfies them then $-X$ satisfies them as well.

[17] To see this, note that the upper bound $E_\eta[X] \le E_\nu[X] + d_W(\nu,\eta) \le E_\nu[X] + \kappa$ is attained by $\eta = \nu(\cdot - \kappa)$.

[18] The implications of this can be seen in the illustrations of Section 6. Under a univariate Wasserstein approach, the three curves in Figure 3 would all collapse to the same straight line, $E_\nu[X] + \kappa$, whereas the worst cases and best cases in Figures 2 and 4 would simply become upward and downward shifts of the nominal curves. The Wasserstein robustness analysis depends on the nominal model only through its mean, whereas the F-divergence approach uses the entire distribution of X.

[19] This argument does not stand in conflict with the discussion of Section 2.3. In *heavy-tailed* models, the *choice* of the divergence should be an informed one. When perturbing heavy-tailed models, potential

integrability problems should not be neglected—just like they should not be neglected when computing sample means and standard deviations in heavy-tailed models. Having some knowledge about the tail behavior of the underlying model is always useful.

[20] Basically, existence and uniqueness of a solution are guaranteed because these two equations describe the worst-case problem for the empirical distribution of the sample. This problem has similar properties as the continuous version studied in Proposition 1 but without potential integrability problems because of the finite support. We only need to verify that the radius κ is not too large as discussed in Appendix B. When it is admissible to put all mass on the largest realization in the sample, the worst-case mean has reached its largest possible value, and we have reached $\kappa_{\max}$.

[21] The negative solution corresponds to the best-case change of measure.

[22] For the case of KL-divergence, the approximation $\alpha_2^{wc} \approx \sqrt{2\kappa/\mathrm{Var}_\nu(X)}$ can also be found as the leading term of a rigorous asymptotic expansion in Lam (2016).

[23] For notational simplicity, we assume that inventory under decentralization is identical across locations. Given the symmetry and convexity of the objective, there can be no gain from choosing asymmetric inventory schemes.

[24] While q_ν^* is available in closed form as the $(p/(p+h))$-quantile of the lognormal variable D_i, we rely on Monte Carlo simulations with an independent sample of $m = 500,000$ copies of $\sum_{i=1}^n D_i$ to estimate Q_ν^*.

[25] In particular, as expected from a plot containing many realizations of an asymptotic 95% confidence bound, the dotted lines lie outside the solid lines most of the time but not always. We have not included confidence bounds based on the larger sample size of $m = 500,000$ because these bounds are so tight that they are visually indistinguishable from the solid lines.

[26] See, for example, the European Banking Authority's Guide for the Targeted Review of Internal Models (TRIM) of February 2017, which states that "An institution should have a model risk management framework in place that allows it to identify, understand and manage its model risk…." (European Banking Authority 2017)

[27] Observe that as opposed to (4), Breuer and Csiszár (2016) formulate the worst-case problem as a minimization problem (see equation (3.6) in Breuer and Csiszár 2016). When referring to Breuer and Csiszár (2016) in the arguments that follows, we tacitly make the appropriate adjustments.

[28] In particular, the conditions on $K(\alpha_1, \alpha_2)$ of proposition A2 in Breuer and Csiszár (2016) imply that K fulfills the property called essential smoothness and that its effective domain, that is, the subset of $\mathbb{R}^2$ where K is finite, is the whole space. If necessary, one can easily extend the result to the case where the effective domain is an open subset of $\mathbb{R}^2$ that contains some points with $\alpha_2 > 0$. See footnote 7 in Breuer and Csiszár (2016).

[29] By allowing ν to possess atoms, we make the framework of this appendix slightly more general than that of our main analysis in Section 2.

[30] As argued in Appendix B, when X is continuously distributed, then the theoretical $\kappa_{\max}$ is infinite, which implies that the empirical kappamax increases indefinitely with the sample size. For instance, with a sample size of 10,000, we obtain a kappamax of about 9.2 in this example.

[31] With continuously distributed data, the values of kappamax in the two computations will typically be the same. Differences appear with discrete data when the sample maximum and minimum are attained multiple times and in different multiplicities.

[32] In our theoretical results, we restrict attention to the case p>1. The reason is that polynomial divergences with 0<p<1 are less restrictive than KL-divergence regarding tail behavior. Thus, the theoretical worst-case mean tends to be infinite even for light-tailed nominal models with unbounded support (Kruse et al. 2019). Our implementation covers the case p<1 but the expected stability issues may appear.

[33] This divergence only depends on the ratio k/theta and collapses to KL-divergence for theta=k. The theoretical results of Section 3.1 require theta > max(1,k), but this restriction is not implemented in the code.

References

Ahmadi-Javid A (2012) Entropic value-at-risk: A new coherent risk measure. *J. Optim. Theory Appl.* 155(3):1105–1123.

Ali SM, Silvey SD (1966) A general class of coefficients of divergence of one distribution from another. *J. Royal Statist. Soc. B* 28(1):131–142.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton Univ. Press, NJ).

Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Sci.* 59(2): 341–357.

Bertsimas D, Sim M (2004) The price of robustness. *Oper. Res.* 52(1):35–53.

Bimpikis K, Markakis MG (2016) Inventory pooling under heavy-tailed demand. *Management Sci.* 62(6):1800–1813.

Breuer T, Csiszár I (2016) Measuring distribution model risk. *Math. Finance* 26(2):395–411.

Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. *SIAM Rev.* 51(4):661–703.

Csiszár I (1963) Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei.* 8(1-2):85–108.

Csiszár I, Breuer T (2018) Expected value minimization in information theoretic multiple priors models. *IEEE Trans. Inform. Theory* 64(6):3957–3974.

Csiszár I, Matúš F (2012) Generalized minimizers of convex integral functionals, Bregman distance, Pythagorean identities. *Kybernetika (Prague)* 48(4):637–689.

Esfahani PM, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Programming* 171(1-2): 115–166.

European Banking Authority (2017) *Guide for the Targeted Review of Internal Models (TRIM)*, version 1.0 (ECB Publication, Frankfurt, Germany).

Föllmer H, Schied A (2011) *Stochastic Finance: An Introduction in Discrete Time*, 3rd ed. (Walter de Gruyter, Berlin, Boston).

Frachot A, Moudoulaud O, Roncalli T (2004) Loss distribution approach in practice. Ong M, ed. *The Basel Handbook: A Guide for Financial Practitioners* (Risk Books, London), 527–554.

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

Hansen LP, Sargent TJ (2011) *Robustness* (Princeton Univ. Press, NJ).

Hürlimann W (2009) On the non-life solvency II model. Cruz M, ed. *The Solvency II Handbook* (Risk Books, London), 349–370.

Kleiber C, Kotz S (2003) *Statistical Size Distributions in Economics and Actuarial Sciences* (John Wiley & Sons, Hoboken, NJ).

Kruse T, Schneider JC, Schweizer N (2019) The joint impact of *F*-divergences and reference models on the contents of uncertainty sets. *Oper. Res.* 67(2):428–435.

Lam H (2016) Robust sensitivity analysis for stochastic systems. *Math. Oper. Res.* 41(4):1248–1275.

Manski CF (1988) *Analog Estimation Methods in Econometrics* (Chapman and Hall, New York).

Morimoto T (1963) Markov processes and the *H*-theorem. *J. Physical. Soc. Japan* 18(3):328–331.

Owen AB (2013) Monte Carlo Theory, Methods and Examples. Working paper, Stanford University, Stanford, CA, https://statweb.stanford.edu/~owen/mc/.

Pardo L (2005) *Statistical Inference Based on Divergence Measures* (Chapman and Hall/CRC, Boca Raton, FL).

Pflug GC, Pichler A (2014) *Multistage Stochastic Optimization* (Springer, New York).

Schneider JC, Schweizer N (2015) Robust measurement of (heavy-tailed) risks: Theory and implementation. *J. Econom. Dynamic Control* 61:183–203.

Whittle P (1990) *Risk-Sensitive Optimal Control* (John Wiley & Sons, Hoboken, NJ).