MEAN-COVARIANCE ROBUST RISK MEASUREMENT

VIET ANH NGUYEN
Stanford University and VinAI Research

SOROOSH SHAFIEEZADEH-ABADEH
Tepper School of Business, CMU

DAMIR FILIPOVIĆ
EPFL and Swiss Finance Institute

DANIEL KUHN
Risk Analytics and Optimization Chair, EPFL

We introduce a universal framework for mean-covariance robust risk measurement and portfolio optimization. We model uncertainty in terms of the Gelbrich distance on the mean-covariance space, along with prior structural information about the population distribution. Our approach is related to the theory of optimal transport and exhibits superior statistical and computational properties than existing models. We find that, for a large class of risk measures, mean-covariance robust portfolio optimization boils down to the Markowitz model, subject to a regularization term given in closed form. This includes the finance standards, value-at-risk and conditional value-at-risk, and can be solved highly efficiently.

KEYWORDS: Robust optimization, risk measurement, optimal transport.

1. INTRODUCTION

Portfolio managers distribute their funds over multiple assets with the aim to optimize future returns. The workhorse in the finance industry is the Markowitz model [Markowitz, 1952], which maximizes the expected portfolio return adjusted for its standard deviation. Its optimal portfolio is given in closed form in terms of the mean and covariance matrix of the asset returns. The classical Markowitz model assumes symmetrically distributed asset returns, so that minimizing the standard deviation is equivalent to minimizing risk. However, asset returns are known to be skewed, and the standard deviation is unable to distinguish undesirable deviations below the mean from desirable deviations above the mean. Numerous propositions for a more appropriate risk assessment have made research on downside risk measures a vibrant field spanning economics and finance. These notably include the industry standards, value-at-risk (VaR) [Jorion, 1996, Longerstaey and Spencer, 1996, Duffie and Pan, 1997] and conditional value-at-risk (CVaR) [Artzner et al., 1999, Rockafellar and Uryasev, 2000], as well as a plethora of more general distribution-based risk measures that have emerged over the last two decades, see, *e.g.*, [Kusuoka, 2001, Acerbi, 2002].

A distribution-based risk measurement requires the precise knowledge of the joint distribution of the underlying asset returns, which is unobservable in practice. For assets traded in

Viet Anh Nguyen: viet-anh.nguyen@stanford.edu
Soroosh Shafieezadeh-Abadeh: sshafiee@andrew.cmu.edu
Damir Filipović: damir.filipovic@epfl.ch
Daniel Kuhn: daniel.kuhn@epfl.ch

We thank Erick Delage for valuable comments on an earlier version of this paper. This research was supported by the Swiss National Science Foundation under the NCCR Automation (grant agreement 51NF40_180545) and under an Early Postdoc.Mobility Fellowship awarded to the second author (grant agreement P2ELP2_195149).

public markets, one may attempt to estimate this distribution from historical data. However, Roy [1952] pointed out that the mean and the covariance matrix are the only quantities that can reasonably be distilled out of financial time series. This observation has spurred interest in robust risk measurement by maximizing a given distribution-based risk measure across all asset return distributions in a Chebyshev ambiguity set, which consists of all distributions with a fixed mean and covariance matrix [El Ghaoui et al., 2003, Yu et al., 2009, Zymler et al., 2013b, Rujeerapaiboon et al., 2016, Li, 2018, Cai et al., 2020].

While taking the worst case over a Chebyshev ambiguity set can immunize the risk measurement against uncertainty in the *shape* of the asset return distribution, it does not take account of uncertainty in its *mean* and *covariance matrix*. This is worrying, because estimators for the mean display a notoriously high variance irrespective of the sampling frequency of the historical return data [Luenberger, 1997, § 8.5]. In addition, estimating high-dimensional covariance matrices is a formidable challenge in statistics that necessitates structural information [Ledoit and Wolf, 2003]. Unfortunately, estimation errors in the mean and covariance matrix have a detrimental impact on the solution of a portfolio optimization problem in that the optimal portfolio's out-of-sample performance falls severely short of its in-sample performance [Michaud, 1989, Best and Grauer, 1991, Chopra and Ziemba, 1993].

In this paper we introduce a universal framework for robustifying any distribution-based risk measurement against mean-covariance uncertainty. To this end, we propose the Gelbrich distance [Gelbrich, 1990] as a metric on the space of mean-covariance pairs. We define the Gelbrich ambiguity set as the family of all asset return distributions with a given structure whose mean-covariance pairs reside in a Gelbrich ball around an empirical mean-covariance pair estimated from data. Here, the structure of a distribution refers to any of its properties that are complementary to location and scale. Examples include symmetry, unimodality, log-concavity or Gaussianity. We think of the structure as reflecting domain knowledge that is uninformed by data. Unlike the Chebyshev ambiguity set, the Gelbrich ambiguity set takes account of the uncertainty in the shape *as well as* the mean and covariance matrix of the asset return distribution. For any given distribution-based risk measure, we then define the Gelbrich risk as the worst-case risk over the Gelbrich ambiguity set.

We find that, if the underlying risk measure is law invariant, translation invariant and positive homogeneous, then the Gelbrich risk reduces to a regularized mean-standard deviation risk measure. The Gelbrich risk minimization problem is therefore equivalent to a regularized Markowitz portfolio selection problem, which can be solved highly efficiently. We thus revive and legitimate the Markowitz model, subject to a fully explicit and tractable regularization, irrespective its aforementioned shortcomings. The underlying risk measure and the structure of the asset return distribution impact the Gelbrich risk only through a scalar coefficient, which can be computed offline and is available in closed form for all coherent, spectral and distortion risk measures. In addition, the weight of the regularization term scales with the radius of the Gelbrich ambiguity set. As a corollary, we obtain that the equally weighted portfolio minimizes the Gelbrich risk under extreme mean-covariance uncertainty, that is, in the limit of an infinitely large Gelbrich ambiguity set.

For any fixed portfolio, we then analytically characterize the worst-case asset return distributions that maximize the underlying risk measure over the Gelbrich ambiguity set. These distributions reveal the portfolio's vulnerabilities and are straightforwardly applicable for stress tests.

The Gelbrich risk is intimately related to the theory of optimal transport [Villani, 2008]. Indeed, the Gelbrich distance between two mean-covariance pairs coincides with the 2-Wasserstein distance between the corresponding Gaussian distributions [Gelbrich, 1990]. Defining the Wasserstein risk as the worst-case risk over all distributions with a given structure that reside in a 2-Wasserstein ball around a nominal distribution estimated from data, we show that the Gelbrich risk upper bounds the Wasserstein risk if the underlying ambiguity sets have the same radius. This bound is sharp and collapses to an equality if the underlying ambiguity sets contain only Gaussian distributions.

We also show that the Gelbrich risk provides a finite-sample upper confidence bound on the true risk under the population distribution if the radius of the Gelbrich ambiguity set scales with the inverse square root of the sample size. This finite-sample bound is dimension-free in the sense that the rate does not depend the number of assets. This result contrasts sharply with existing out-of-sample guarantees for the Wasserstein risk [Mohajerin Esfahani and Kuhn, 2018], which rely on a measure concentration result by Fournier and Guillin [2015] and suffer from a curse of dimensionality. Our result is also orthogonal to the dimension-free finite-sample guarantees for the Wasserstein risk by Gao [2020], which rely on concepts of hypothesis complexity such as covering numbers or Rademacher complexities and which apply only to worst-case expectations but may not easily generalize to other risk measures.

The study of decision problems under distributional uncertainty has a long and distinguished history in economics dating back at least to Keynes [1921] and Knight [1921]. Ellsberg [1961] was the first to document that most individuals have a low tolerance for distributional uncertainty. This phenomenon is often used to motivate the maxim that different decision alternatives should be ranked in view of their worst-case performance with respect to all distributions in some ambiguity set. A rigorous axiomatic justification for this decision rule is due to Gilboa and Schmeidler [1989]. In operations research, decision problems under uncertainty are typically framed as distributionally robust optimization problems, and research focuses primarily on deriving tractable reformulations and efficient solution algorithms [Delage and Ye, 2010, Goh and Sim, 2010, Wiesemann et al., 2014].

Considerable efforts were directed to investigating various distributionally robust portfolio optimization problems. When modeling distributional uncertainty via Chebyshev ambiguity sets, the worst-case risk admits a tractable reformulations for the VaR [El Ghaoui et al., 2003, Zymler et al., 2013b, Rujeerapaiboon et al., 2016, 2018], the CVaR [Natarajan et al., 2010, Chen et al., 2011a, Zymler et al., 2013b] as well as for all spectral risk measures [Li, 2018] and all distortion risk measures [Cai et al., 2020, Pesenti et al., 2020]. These reformulations will emerge as special cases of our results because any Chebyshev ambiguity set constitutes a Gelbrich ambiguity set with a vanishing radius. We also stress that if the mean and covariance matrix of the asset returns are estimated from data, the corresponding Chebyshev ambiguity set will *not* contain the data-generating distribution with probability one. Thus, the Chebyshev risk fails to provide a safe estimate for the true risk. To obtain a safe estimate, one could inflate the Chebyshev ambiguity set by allowing the mean and covariance matrix of the asset return distributions to range over simple box-type or semidefinite-representable confidence sets [El Ghaoui et al., 2003, Delage and Ye, 2010, Zymler et al., 2013b, Rujeerapaiboon et al., 2016]. However, the design of these confidence sets is driven by computational rather than economic or statistical considerations. Moreover, tractability results are limited to special risk measures such as VaR or CVaR.

Other ambiguity sets used in distributionally robust portfolio optimization impose asymmetric moment bounds [Chen et al., 2010, Natarajan et al., 2008, 2018], marginal moment bounds [Doan et al., 2015] or structural properties such as symmetry, unimodality or tail convexity etc. [Popescu, 2005, Yu et al., 2009, Van Parys et al., 2016, Lam and Mottet, 2017], or they stipulate a factor model for the asset returns [Nemirovski and Shapiro, 2007].

Distributionally robust portfolio optimization problems with 1-Wasserstein ball ambiguity sets for discrete asset return distributions are addressed by Pflug and Wozabal [2007] via exhaustive search methods from global optimization. Using robust optimization techniques, Postek et al. [2016] show that these problems are actually equivalent to tractable convex programs and thus amenable to efficient iterative algorithms. When the ambiguity set may contain generic non-discrete distributions, the worst case of any convex distortion risk measure over a p-Wasserstein ball for $p \geq 1$ coincides with the nominal risk adjusted by a regualization term penalizing some dual norm of the portfolio weight vector [Wozabal, 2014]; see also [Pichler, 2013, § 4] for a related discussion. However, all of these results apply only to *convex* risk measures (thus excluding VaR) and fail to account for structural distributional information (meaning that the ambiguity set may contain unrealistic pathological distributions). Also, the only universal statistical guarantees on the true risk known to date suffer from a curse of dimensionality [Mohajerin Esfahani and Kuhn, 2018].

The regularization terms penalizing the norm of the portfolio weight vector, which emerge from our reformulation of the Gelbrich risk, can be viewed as *implicit* norm constraints. In the context of a Markowitz model, Jagannathan and Ma [2003] show that imposing *explicit* norm constraints is equivalent to shrinking the estimator of the covariance matrix. DeMiguel et al. [2009] offer a Bayesian interpretation for the resulting optimal portfolios and show empirically that they display an excellent out-of-sample performance. In contrast, our paper offers a new probabilistic interpretation for norm regularization terms and unveils how they depend on the risk measure and any available structural information.

The Gelbrich ambiguity set was first proposed by Nguyen et al. [2021b] to robustify minimum mean square error estimation problems against distributional uncertainty. Similar ideas were also used by Nguyen et al. [2021a] and Shafieezadeh-Abadeh et al. [2018] in the context of inverse covariance matrix estimation and Kalman filtering, respectively.

The remainder of the paper is structured as follows. Section 2 formally introduces the Gelbrich risk. Section 3 investigates its attractive conceptual and statistical properties. Section 4 demonstrates its efficient computability for law invariant, translation invariant and positive homogeneous base risk measures. Section 5 concludes. The proofs of our main results are relegated to the appendix. An online appendix contains all remaining proofs as well as additional tractability results for mean-variance risk measures, which fail to be positive homogeneous.

Notation. The 2-norm of a vector $x \in \mathbb{R}^n$ is denoted by $\|x\|$. Similarly, we use $\|A\|$ and $\|A\|_F$ to denote the spectral and Frobenius norms of a square matrix $A \in \mathbb{R}^{n\times n}$, respectively. The space of all symmetric matrices in $\mathbb{R}^{n\times n}$ is denoted as $\mathbb{S}^n$, while $\mathbb{S}^n_{++}$ ($\mathbb{S}^n_{+}$) stands for the cone of all positive (semi)definite matrices in $\mathbb{S}^n$. For any $A \in \mathbb{S}^n$, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of A, respectively. We denote the Borel σ-algebra on $\mathbb{R}^n$ by $\mathcal{B}(\mathbb{R}^n)$, the space of Borel-measurable functions from $\mathbb{R}^n$ to $\mathbb{R}$ by $\mathcal{L}_0$, and the set of all probability distributions on $\mathcal{B}(\mathbb{R}^n)$ by $\mathcal{M}$. The expectation of a random variable $\ell \in \mathcal{L}_0$ under $\mathbb{Q} \in \mathcal{M}$ is denoted by $\mathbb{E}_{\mathbb{Q}}[\ell]$. We denote by $\mathcal{M}_2$ the set of all $\mathbb{Q} \in \mathcal{M}$ with finite second moments, that is, with $\mathbb{E}_{\mathbb{Q}}[\|\xi\|^2] < \infty$.

2. PROBLEM STATEMENT

We study decision problems under distributional uncertainty that aim to minimize the risk of a loss affected by a vector $\xi \in \mathbb{R}^n$ of *risk factors*. Formally, a *loss function* $\ell \in \mathcal{L}_0$ assigns each realization $\xi \in \mathbb{R}^n$ of the risk factors a loss $\ell(\xi) \in \mathbb{R}$. Different loss functions correspond to different decision alternatives available to a risk-averse decision maker. If the risk factors are governed by a probability distribution $\mathbb{P} \in \mathcal{M}$, then the decision maker ranks the loss functions according to a risk measure $\mathcal{R}_{\mathbb{P}} : \mathcal{L}_0 \to \mathbb{R} \cup \{+\infty\}$, which usually depends on $\mathbb{P}$. We thus define the *risk* of any loss function $\ell \in \mathcal{L}_0$ under $\mathbb{P}$ as $\mathcal{R}_{\mathbb{P}}(\ell)$. In addition, we define the *optimal risk* corresponding to a set $\mathcal{L} \subseteq \mathcal{L}_0$ of *admissible* loss functions under $\mathbb{P}$ as $\inf_{\ell \in \mathcal{L}} \mathcal{R}_{\mathbb{P}}(\ell)$. Unfortunately, the true probability distribution $\mathbb{P}$ is almost never known in practice, and thus neither the risk of a fixed loss function nor the optimal risk can be evaluated reliably.

We henceforth assume that the decision maker has only access to limited statistical information, along with some prior structural information, about $\mathbb{P}$. Formally, she only knows that $\mathbb{P}$ lies in some ambiguity set $\mathcal{P} \subseteq \mathcal{M}$. For a given family of risk measures $\{\mathcal{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathcal{M}}$ this leads to the corresponding *worst-case risk* of any loss function $\ell \in \mathcal{L}_0$ given by

$$\mathcal{R}_{\mathcal{P}}(\ell) = \sup_{\mathbb{Q}\in\mathcal{P}} \mathcal{R}_{\mathbb{Q}}(\ell). \tag{1}$$

Ranking different loss functions by their worst-case risk, the decision maker thus solves a distributionally robust optimization problem that finds the *optimal worst-case risk*

$$\mathcal{R}_{\mathcal{P}}(\mathcal{L}) = \inf_{\ell\in\mathcal{L}} \mathcal{R}_{\mathcal{P}}(\ell) = \inf_{\ell\in\mathcal{L}} \sup_{\mathbb{Q}\in\mathcal{P}} \mathcal{R}_{\mathbb{Q}}(\ell). \tag{2}$$

In the following we discuss the choice of the ambiguity set $\mathcal{P}$. Specifically, in Section 2.1 we first formalize the modeling of structural information, and in Section 2.2 we address the modeling of statistical information. We thereby focus on mean-covariance uncertainty.

2.1. *Structural Information*

We encode structural information about the unknown probability distribution $\mathbb{P}$ by a structural ambiguity set defined as follows.

DEFINITION 1—Structural ambiguity set: *A structural ambiguity set $\mathcal{S}$ is a subset of $\mathcal{M}_2$ that is closed under positive semidefinite affine pushforwards. That is, for any $\mathbb{Q} \in \mathcal{S}$ and any transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ of the form $f(\xi) = A\xi + b$ for some $A \in \mathbb{S}^n_+$ and $b \in \mathbb{R}^n$, the pushforward distribution $\mathbb{Q} \circ f^{-1}$ belongs to $\mathcal{S}$.*

The entire set $\mathcal{M}_2$ trivially constitutes a structural ambiguity set. Below we discuss non-trivial examples of structural ambiguity sets that will be addressed in this paper.

DEFINITION 2—Symmetric distribution: *The probability distribution $\mathbb{Q} \in \mathcal{M}_2$ is symmetric if there exists $\mu \in \mathbb{R}^n$ with $\mathbb{Q}[\xi \leq \mu - \tau] = \mathbb{Q}[\xi \geq \mu + \tau]$ for all $\tau \in \mathbb{R}^n$.*

By definition, $\mathbb{Q}$ is symmetric about μ if and only if the random vectors $\xi - \mu$ and $\mu - \xi$ have the same cumulative distribution function under $\mathbb{Q}$, which is equivalent to the condition that $\mathbb{Q}[(\xi - \mu) \in B] = \mathbb{Q}[(\mu - \xi) \in B]$ for all Borel sets $B \subseteq \mathbb{R}^n$. Hence, the set of symmetric probability distributions is closed under positive semidefinite affine pushforwards and thus constitutes a structural ambiguity set; see also [Yu et al., 2009, Lemma 1].

While there is consensus about what it means for a *univariate* distribution to be unimodal, there are several non-equivalent notions of unimodality for *multivariate* distributions. In the following we will argue that two of these notions, namely linear unimodality and log-concavity, give rise to structural ambiguity sets.

DEFINITION 3—Linear unimodal distribution: *The probability distribution $\mathbb{Q} \in \mathcal{M}_2$ is linear unimodal if there exists $\mu \in \mathbb{R}^n$ such that the cumulative distribution function of $w^\top \xi$ under $\mathbb{Q}$ is convex on $(-\infty, w^\top \mu]$ and concave on $[w^\top \mu, +\infty)$ for all $w \in \mathbb{R}^n$.*

It is easy to show that the set of linear unimodal distributions is closed under positive semidefinite affine pushforwards and thus also constitutes a structural ambiguity set.

DEFINITION 4—Log-concave distribution: *The probability distribution $\mathbb{Q} \in \mathcal{M}_2$ is log-concave if for any Borel sets $B_1, B_2 \in \mathcal{B}(\mathbb{R}^n)$ and for any scalar weight $\theta \in [0, 1]$ we have*

$$\mathbb{Q}(\theta B_1 + (1-\theta)B_2) \geq \mathbb{Q}(B_1)^\theta \mathbb{Q}(B_2)^{1-\theta},$$

where the convex combination of B_1 and B_2 is understood in the sense of Minkowski.

Log-concave distributions play an important role in statistics and optimization. Many standard distributions such as the uniform distributions on convex sets as well as the Gaussian, Wishart or Dirichlet distributions are log-concave. All log-concave distributions have sub-exponential tails [Borell, 1983]. Moreover, by [Dharmadhikari and Joag-Dev, 1988, Lemma 2.1] the family of all log-concave distributions is closed under positive semidefinite affine pushforwards and thus constitutes a structural ambiguity set.

DEFINITION 5—Elliptical distribution: *The probability distribution $\mathbb{Q} \in \mathcal{M}_2$ is elliptical if its characteristic function $\mathbb{E}_\mathbb{Q}[\exp(i\tau^\top \xi)]$ can be written as $\exp(i\tau^\top \mu)\phi(\tau^\top \Sigma \tau)$ for some location parameter $\mu \in \mathbb{R}^n$, dispersion matrix $\Sigma \in \mathbb{S}_+^n$ and characteristic generator $\phi : \mathbb{R}_+ \to \mathbb{R}$, where $i = \sqrt{-1}$ denotes the imaginary unit.*

By [Cambanis et al., 1981, Theorem 1], the family of all elliptical distributions with the same characteristic generator is closed under positive semidefinite affine pushforwards and thus constitutes a structural ambiguity set. The same theorem implies that every elliptical distribution is symmetric. However, not every elliptical distribution is unimodal. Examples of elliptical distributions that fail to be unimodal include certain Kotz-type or multivariate Bessel distributions. The location parameter μ of an elliptical distribution $\mathbb{Q}$ always matches the mean of $\mathbb{Q}$. In addition, one can show that the generator ϕ of $\mathbb{Q}$ may not be chosen freely but is only admissible if $\phi(\|\tau\|^2)$ represents the characteristic function of some probability distribution $\mathbb{Q}'$ with finite second moments. Note that $\mathbb{Q}'$ differs from $\mathbb{Q}$ unless $\mu = 0$ and $\Sigma = I_n$. As probability distributions are normalized, this condition can only hold if $\phi(0) = 1$. And as any characteristic function is continuous thanks to the dominated convergence theorem, this condition further implies that ϕ must be continuous. One can also show that the covariance matrix of $\mathbb{Q}$ is given by $-2\phi'(0)\Sigma$, where $\phi'(0)$ stands for the right derivative of $\phi(u)$ at $u = 0$. Assuming that $\mathbb{Q}$ has finite second moments is thus equivalent to assuming that $\phi'(0)$ exists and is finite. In the remainder of the paper we will assume without loss of generality that $\phi'(0) = -\frac{1}{2}$, which means that the dispersion matrix Σ coincides with the covariance matrix of $\mathbb{Q}$. Indeed, changing the characteristic generator to $\phi(\frac{-u}{2\phi'(0)})$ and the dispersion matrix to $-2\phi'(0)\Sigma$ does not change $\mathbb{Q}$ but ensures that the dispersion matrix and the covariance matrix of $\mathbb{Q}$ coincide. Note that the families of all symmetric, linear unimodal or log-concave distributions as well as the family of all elliptical distributions with a given generator fail to be convex. For example, the set of Gaussian distributions, which is obtained by setting $\phi(u) = e^{-u/2}$, is non-convex because mixtures of Gaussian distributions are generically multimodal and thus not Gaussian.

We now define the smallest structural ambiguity set that contains a given $\widehat{\mathbb{P}} \in \mathcal{M}_2$.

DEFINITION 6—Structural ambiguity set generated by $\widehat{\mathbb{P}}$: *The structural ambiguity set generated by $\widehat{\mathbb{P}} \in \mathcal{M}_2$ is the family of all positive semidefinite affine pushforwards of $\widehat{\mathbb{P}}$.*

By [Cambanis et al., 1981, Theorem 1], the set of all elliptical distributions with generator ϕ can be viewed as the structural ambiguity set generated by the standardized elliptical distribution $\widehat{\mathbb{P}}$ with generator ϕ, mean $\mu = 0$ and covariance matrix $\Sigma = I$. In contrast, the sets of all symmetric, linear unimodal or log-concave distributions are not generated by any single distribution. Also, not every distribution in the structural ambiguity set $\mathcal{S}$ generated by $\widehat{\mathbb{P}}$ generates all of $\mathcal{S}$. For example, the Dirac distribution δ_0 that concentrates unit mass at 0 constitutes a (degenerate) Gaussian distribution with mean $\mu = 0$ and covariance matrix $\Sigma = 0$. However, δ_0 fails to generate the family of all Gaussian distributions because any positive semidefinite affine pushforward of δ_0 is also Dirac distribution.

2.2. *Statistical Information*

In addition to structural information captured by the ambiguity set $\mathcal{S}$, the decision maker may have access to a finite training sample that provides statistical information about the unknown probability distribution $\mathbb{P}$. In a financial context, such training samples are routinely used to construct estimators $\widehat{\mu}$ and $\widehat{\Sigma}$ for the mean and covariance matrix of $\mathbb{P}$, respectively. Indeed, Roy [1952] asserts that the first and second moments of $\mathbb{P}$ '*are the only quantities that can be distilled out of our knowledge of the past.*' Moreover, he adds that '*the slightest acquaintance with problems of analysing economic time series will suggest that this assumption is optimistic rather than unnecessarily restrictive.*' Roy's warning alerts us that the true mean μ and the true covariance matrix Σ of $\mathbb{P}$ typically differ from their noisy estimators $\widehat{\mu}$ and $\widehat{\Sigma}$. In the following, we quantify the corresponding estimation errors via the Gelbrich distance in the space of mean-covariance pairs.

DEFINITION 7—Gelbrich distance: *The Gelbrich distance between two mean-covariance pairs (μ_1, Σ_1) and (μ_2, Σ_2) in $\mathbb{R}^n \times \mathbb{S}^n_+$ is given by*

$$\mathbb{G}\big((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)\big) = \sqrt{\|\mu_1 - \mu_2\|^2 + \operatorname{Tr}\left[\Sigma_1 + \Sigma_2 - 2\big(\Sigma_2^{\frac{1}{2}} \Sigma_1 \Sigma_2^{\frac{1}{2}}\big)^{\frac{1}{2}}\right]}.$$

One can show that the Gelbrich distance is non-negative, symmetric and subadditive and that it vanishes if and only if $(\mu_1, \Sigma_1) = (\mu_2, \Sigma_2)$, which implies that it represents a metric on $\mathbb{R}^n \times \mathbb{S}^n_+$ [Givens and Shortt, 1984, pp. 239].

If $\mu_1 = \mu_2$, then the Gelbrich distance reduces to the Bures distance that measures the dissimilarity between density matrix operators in quantum information theory [Bhatia et al., 2018, 2019]. In this case, the Gelbrich distance induces a Riemannian metric on the space of positive semidefinite matrices. If, in addition, the two covariance matrices are diagonal, then the Gelbrich distance simplifies to the Hellinger distance, which is closely related to the Fisher-Rao metric ubiquitous in information theory [Liese and Vajda, 1987].

We can thus define the mean-covariance uncertainty set as the ball of radius $\rho \geq 0$ centered at the estimators $\widehat{\mu}$ and $\widehat{\Sigma}$ of the mean and covariance matrix of $\mathbb{P}$,

$$\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma}) = \left\{(\mu, \Sigma) \in \mathbb{R}^n \times \mathbb{S}^n_+ : \mathbb{G}\big((\mu, \Sigma), (\widehat{\mu}, \widehat{\Sigma})\big) \leq \rho\right\}.$$

We can then introduce the *Gelbrich ambiguity set* as the family of all probability distributions of ξ that are consistent with the available structural and statistical information. Formally, the Gelbrich ambiguity set is defined as the pre-image of $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ under the transformation that maps $\mathbb{Q} \in \mathcal{S}$ to its first and second moments.

DEFINITION 8—Gelbrich ambiguity set: *The Gelbrich ambiguity set is given by*

$$\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma}) = \left\{\mathbb{Q} \in \mathcal{S} : \big(\mathbb{E}_{\mathbb{Q}}[\xi], \mathbb{E}_{\mathbb{Q}}[(\xi - \mathbb{E}_{\mathbb{Q}}[\xi])(\xi - \mathbb{E}_{\mathbb{Q}}[\xi])^\top]\big) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})\right\}.$$

Note that the Gelbrich ambiguity set contains all distributions in the structural ambiguity set $\mathcal{S}$ whose mean-covariance pairs have a Gelbrich distance of at most ρ from the estimators $(\widehat{\mu}, \widehat{\Sigma})$. If the estimation error of $(\widehat{\mu}, \widehat{\Sigma})$ is at most ρ, then the unknown true distribution $\mathbb{P}$ is thus guaranteed to belong to the Gelbrich ambiguity set. In this case the true risk of a Borel-measurable loss function $\ell \in \mathcal{L}_0$ cannot be evaluated because $\mathbb{P}$ is unknown (except in trivial cases, *e.g.*, when ℓ is constant). However, we can evaluate the worst-case risk of ℓ with respect to all probability distributions in the Gelbrich ambiguity set $\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$. We thus define the *Gelbrich risk* $\mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})}(\ell)$ as the worst-case risk (1) with respect to the ambiguity set $\mathcal{P} = \mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$. In addition, we define the *optimal Gelbrich risk* $\mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})}(\mathcal{L})$ corresponding to a set $\mathcal{L} \subseteq \mathcal{L}_0$ of admissible loss functions as the infimum of the Gelbrich risk over all loss functions $\ell \in \mathcal{L}$, defined as in (2) with $\mathcal{P} = \mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$. Note that by solving the optimal Gelbrich risk problem, the decision maker anticipates the worst possible probability distribution in the Gelbrich ambiguity set and seeks a decision alternative that results in the least possible risk under this worst-case distribution. Computing $\mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})}(\mathcal{L})$ thus amounts to solving a distributionally robust optimization problem.

3. PROPERTIES OF THE GELBRICH AMBIGUITY SET

We now provide conceptual, statistical and computational justification for modeling distributional uncertainty via Gelbrich ambiguity sets. In Section 3.1 we first show that the Gelbrich ambiguity set is closely related to the Wasserstein ambiguity set, which is widely used in distributionally robust optimization. In Section 3.2 we then investigate the statistical and in Section 3.3 the computational properties of the Gelbrich ambiguity set.

3.1. *Relation between the Gelbrich and Wasserstein Ambiguity Sets*

Instead of directly estimating the mean and covariance matrix from a training sample, one could construct a *nominal* probability distribution $\widehat{\mathbb{P}}$ representing a best guess of the unknown true distribution $\mathbb{P}$. Throughout the rest of the paper we assume that $\widehat{\mathbb{P}}$ belongs to the structural ambiguity set $\mathcal{S}$ and that $\widehat{\mu}$ and $\widehat{\Sigma}$ coincide with the mean and covariance matrix of $\widehat{\mathbb{P}}$, respectively. Note that the first and second moments of $\widehat{\mathbb{P}}$ exist because $\mathcal{S} \subseteq \mathcal{M}_2$. Of course, the decision maker should not put full trust into the nominal distribution $\widehat{\mathbb{P}}$, which can be viewed as a noisy estimator for $\mathbb{P}$, and the corresponding estimation errors are conveniently measured by the 2-Wasserstein distance on the space $\mathcal{M}_2$.

DEFINITION 9—Wasserstein distance: *The 2-Wasserstein distance between two distributions $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{M}_2$ is*

$$\mathbb{W}(\mathbb{Q}_1, \mathbb{Q}_2) = \min_{\pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2)} \left(\int_{\mathbb{R}^n \times \mathbb{R}^n} \|\xi_1 - \xi_2\|^2 \, \pi(\mathrm{d}\xi_1, \mathrm{d}\xi_2) \right)^{\frac{1}{2}},$$

where $\Pi(\mathbb{Q}_1, \mathbb{Q}_2)$ denotes the set of all couplings of $\mathbb{Q}_1$ and $\mathbb{Q}_2$, that is, the set of all joint distributions of $\xi_1 \in \mathbb{R}^n$ and $\xi_2 \in \mathbb{R}^n$ with marginal distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$, respectively.

The 2-Wasserstein distance is non-negative, symmetric and subadditive, and it vanishes only if $\mathbb{Q}_1 = \mathbb{Q}_2$, which implies that it represents a metric on $\mathcal{M}_2$ [Villani, 2008, p. 94]. In addition, the minimization problem over π is always solvable [Villani, 2008, Theorem 5.9], and $\mathbb{W}(\mathbb{Q}_1, \mathbb{Q}_2)^2$ can be viewed as the minimum cost of transporting the distribution $\mathbb{Q}_1$ to $\mathbb{Q}_2$, assuming that the cost of moving a unit probability mass from ξ_1 to ξ_2 amounts to $\|\xi_1 - \xi_2\|^2$. The variable π thus encodes a (probability) mass transportation plan.

We can now define the Wasserstein ambiguity set with structural information

$$\mathcal{W}_\rho(\widehat{\mathbb{P}}) = \left\{ \mathbb{Q} \in \mathcal{S} : \mathbb{W}(\widehat{\mathbb{P}}, \mathbb{Q}) \leq \rho \right\}$$

as the ball of radius $\rho \geq 0$ in the structural ambiguity set $\mathcal{S}$ centered at the nominal distribution $\widehat{\mathbb{P}}$ with respect to the 2-Wasserstein distance. Intuitively, the radius ρ of this ambiguity set quantifies the decision maker's distrust in the nominal distribution $\widehat{\mathbb{P}}$. Note that since $\widehat{\mathbb{P}} \in \mathcal{S}$ and $\mathbb{W}(\widehat{\mathbb{P}}, \widehat{\mathbb{P}}) = 0$, the Wasserstein ambiguity set is non-empty for every $\rho \geq 0$.

If the true probability distribution $\mathbb{P}$ of the risk factors is unknown, then the decision maker could rank different loss functions in view of their worst-case risk over the Wasserstein ambiguity set. We thus define the *Wasserstein risk* $\mathcal{R}_{\mathcal{W}_\rho(\widehat{\mathbb{P}})}(\ell)$ as the worst-case risk (1) with $\mathcal{P} = \mathcal{W}_\rho(\widehat{\mathbb{P}})$. In addition, we define the *optimal Wasserstein risk* $\mathcal{R}_{\mathcal{W}_\rho(\widehat{\mathbb{P}})}(\mathcal{L})$ corresponding to a set $\mathcal{L} \subseteq \mathcal{L}_0$ of admissible loss functions as the infimum of the Wasserstein risk over all loss functions $\ell \in \mathcal{L}$, defined as in (2) with $\mathcal{P} = \mathcal{W}_\rho(\widehat{\mathbb{P}})$.

Distributionally robust optimization with Wasserstein ambiguity sets enjoys increasing popularity in economics and operations research because it offers attractive out-of-sample performance and asymptotic consistency guarantees while being computationally tractable [Kuhn et al., 2019]. For example, if the risk measure $\mathcal{R}_\mathbb{Q}$ coincides with the expected value under $\mathbb{Q}$, the loss function ℓ is representable as a pointwise maximum of finitely many concave functions, $\widehat{\mathbb{P}}$ is discrete and $\mathcal{S} = \mathcal{M}_2$, then, under mild technical conditions, the optimal Wasserstein risk $\mathcal{R}_{\mathcal{W}_\rho(\widehat{\mathbb{P}})}(\mathcal{L})$ can be computed by solving a tractable convex optimization problem [Zhen et al., 2021, § 6]. However, tractability results for more general risk measures, nominal distributions and structural ambiguity sets are scarce. And even if a tractable reformulation exists, its size typically scales with the cardinality of the support of $\widehat{\mathbb{P}}$. In this section we will show that the Gelbrich ambiguity set provides an outer approximation of the Wasserstein ambiguity set and that this approximation becomes exact if the structural ambiguity set $\mathcal{S}$ is generated by the nominal distribution $\widehat{\mathbb{P}}$. This result is significant for several reasons. First, it implies that if the Wasserstein ambiguity set is designed as a confidence region for $\mathbb{P}$, then the corresponding Gelbrich ambiguity set also constitutes a confidence region for $\mathbb{P}$ with the same coverage probability. The Gelbrich risk thus inherits any known statistical guarantees for the Wasserstein risk. Moreover, this result will later allow us to construct tractable conservative approximations or exact tractable reformulations of the Wasserstein risk that do not grow with the cardinality of the support of the nominal distribution. We emphasize that these approximations and reformulations are available for a broad range of risk measures, nominal distributions and structural ambiguity sets for which there currently exist no tractability results.

To show that the Gelbrich risk upper bounds the Wasserstein risk, we recall that the Gelbrich distance provides a lower bound on the 2-Wasserstein distance between two probability distributions that depends exclusively on their mean vectors and covariance matrices.

THEOREM 1—Gelbrich bound [Gelbrich, 1990, Theorem 2.1]: *For any distributions $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{M}_2$ with mean vectors μ_1, $\mu_2 \in \mathbb{R}^n$ and covariance matrices Σ_1, $\Sigma_2 \in \mathbb{S}^n_+$, respectively, we have $\mathbb{W}(\mathbb{Q}_1, \mathbb{Q}_2) \geq \mathbb{G}((\mu_1, \Sigma_1), (\mu_2, \Sigma_2))$.*

The Gelbrich bound of Theorem 1 may be useful when the exact Wasserstein distance between two probability distributions is inaccessible. Indeed, computing Wasserstein distances is generically #P-hard [Taşkesen et al., 2021, Theorem 2.2]. Even though the Gelbrich distance is non-convex, the squared Gelbrich distance is jointly convex in both of its arguments. This is evident from the proof of Theorem 1, which shows that $\mathbb{G}^2((\mu_1,\Sigma_1),(\mu_2,\Sigma_2))$ equals the optimal value of a semidefinite program, and because convexity is preserved under partial minimization [Bertsekas, 2009, Proposition 3.3.1]. Maybe surprisingly, the Gelbrich bound is tight in several cases of practical interest.

THEOREM 2—Tightness of the Gelbrich bound: *Suppose that the distributions $\mathbb{Q}_1,\mathbb{Q}_2 \in \mathcal{M}_2$ have mean vectors $\mu_1,\mu_2 \in \mathbb{R}^n$ and covariance matrices $\Sigma_1,\Sigma_2 \in \mathbb{S}^n_+$, respectively. If $\Sigma_1 \succ 0$ and $\mathbb{Q}_2$ is a positive semidefinite affine pushforward of $\mathbb{Q}_1$, then we have*

$$\mathbb{W}(\mathbb{Q}_1,\mathbb{Q}_2) = \mathbb{G}\big((\mu_1,\Sigma_1),(\mu_2,\Sigma_2)\big).$$

Recall that $\mathbb{Q}_2$ is a positive semidefinite affine pushforward of $\mathbb{Q}_1$ if there exists an affine function $f(\xi) = A\xi + b$ with $A \in \mathbb{S}^n_+$ and $b \in \mathbb{R}^n$ such that $\mathbb{Q}_2 = \mathbb{Q}_1 \circ f^{-1}$. The proof of Theorem 2 reveals that f is uniquely determined by μ_1, μ_2, Σ_1 and Σ_2 via the relations

$$A = \Sigma_1^{-\frac{1}{2}}\big(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\big)^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}} \qquad \text{and} \qquad b = \mu_2 - A\mu_1. \tag{3}$$

Note that the inverse of $\Sigma_1^{-\frac{1}{2}}$ exists because of our assumption that Σ_1 is positive definite. We emphasize, however, that Theorem 2 remains valid if Σ_1 is only positive *semi*definite and if $\mathbb{Q}_2 \circ P_{\Sigma_1}^{-1} = \mathbb{Q}_1 \circ f^{-1}$, where f is parametrized as in (3) with Σ_1^{-1} representing the Moore-Penrose inverse of Σ_1, while P_{Σ_1} denotes the orthogonal projection onto the column space of Σ_1 [Gelbrich, 1990, Theorem 2.1]. To keep this paper self-contained, we prove Theorem 2, which is weaker but sufficient for our purposes, in the online appendix.

If $\mathbb{Q}$ belongs to the structural ambiguity set generated by $\widehat{\mathbb{P}}$, then it constitutes a positive semidefinite affine pushforward of $\widehat{\mathbb{P}}$. If in addition $\widehat{\Sigma} \succ 0$, then Theorem 2 implies that the 2-Wasserstein distance between $\mathbb{Q}$ and $\widehat{\mathbb{P}}$ coincides with the Gelbrich distance between their mean-covariance pairs.

We now show that $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ covers the projection of the 2-Wasserstein ball $\mathcal{W}_\rho(\widehat{\mathbb{P}})$ onto the space of mean-covariance pairs. A sharper results is available if $\widehat{\Sigma} \succ 0$, in which case $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ coincides exactly with the aforementioned projection.

PROPOSITION 1—Projection of $\mathcal{W}_\rho(\widehat{\mathbb{P}})$ onto the mean-covariance space: *If the nominal distribution $\widehat{\mathbb{P}}$ has mean $\widehat{\mu} \in \mathbb{R}^n$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}^n_+$, then we have*

$$\Big\{\big(\mathbb{E}_{\mathbb{Q}}[\xi], \mathbb{E}_{\mathbb{Q}}[(\xi - \mathbb{E}_{\mathbb{Q}}[\xi])(\xi - \mathbb{E}_{\mathbb{Q}}[\xi])^\top]\big) : \mathbb{Q} \in \mathcal{W}_\rho(\widehat{\mathbb{P}})\Big\} \subseteq \mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma}). \tag{4}$$

If in addition $\widehat{\Sigma} \succ 0$, then the inclusion becomes an equality.

The above results culminate in the following main theorem.

THEOREM 3—Relation between the Gelbrich and Wasserstein ambiguity sets: *If the nominal distribution $\widehat{\mathbb{P}}$ has mean $\widehat{\mu} \in \mathbb{R}^n$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}^n_+$, then we have $\mathcal{W}_\rho(\widehat{\mathbb{P}}) \subseteq \mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})$. In addition, if $\mathcal{S}$ is the structural ambiguity set generated by $\widehat{\mathbb{P}}$ and if $\widehat{\Sigma} \succ 0$, then the inclusion becomes an equality.*

Theorem 3 implies that the Gelbrich ambiguity set $\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$ constitutes an outer approximation for the Wasserstein ambiguity set $\mathcal{W}_\rho(\widehat{\mathbb{P}})$ that ignores any information about the nominal distribution except for its structure, mean and covariance matrix. Discarding all higher-order moments can be interpreted as a compression of the available information. Below we will argue that this compression can be leveraged to construct tractable reformulations of the (optimal) Wasserstein risk. An immediate consequence of Theorem 3 is that the (optimal) Gelbrich risk provides an upper bound on the (optimal) Wasserstein risk. We formalize this insight in the following corollary, which we state without proof.

COROLLARY 1— Gelbrich risk versus Wasserstein risk: *If the nominal distribution $\widehat{\mathbb{P}}$ has mean $\widehat{\mu} \in \mathbb{R}^n$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}_+^n$, then we have $\mathcal{R}_{\mathcal{W}_\rho(\widehat{\mathbb{P}})}(\ell) \leq \mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})}(\ell)$ for all $\ell \in \mathcal{L}$ and $\mathcal{R}_{\mathcal{W}_\rho(\widehat{\mathbb{P}})}(\mathcal{L}) \leq \mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})}(\mathcal{L})$. In addition, if $\mathcal{S}$ is the structural ambiguity set generated by $\widehat{\mathbb{P}}$ and if $\widehat{\Sigma} \succ 0$, then these inequalities become equalities.*

Corollary 1 implies that any finite-sample guarantee for the (optimal) Wasserstein risk immediately leads to a finite-sample guarantee for the (optimal) Gelbrich risk. For example, if $\widehat{\mathbb{P}}$ is set to the discrete empirical distribution of the training sample, then the measure concentration results by [Fournier and Guillin, 2015] can be used to calibrate the radius ρ to the sample size so that the unknown true distribution $\mathbb{P}$ belongs to the Wasserstein ambiguity set with probability $1-\eta$ for any given $\eta \in (0,1)$ [Mohajerin Esfahani and Kuhn, 2018, Theorem 3.4]. For this choice of ρ, the Wasserstein and the Gelbrich risk exceed the true risk with probability at least $1-\eta$. However, the underlying measure concentration results suffer from a fundamental curse of dimensionality, which implies that ρ must decay extremely slowly with the sample size even if ξ has moderate dimension. In the next section we will demonstrate that this curse of dimensionality can be circumvented by working with the Gelbrich instead of the Wasserstein ambiguity set.

3.2. *Statistical Properties of the Gelbrich Ambiguity Set*

We now show that the Gelbrich distance enjoys attractive measure concentration properties, which make it ideal to construct ambiguity sets for sub-Gaussian distributions. These measure concentration properties thus lend themselves for calibrating the radius ρ of the Gelbrich ambiguity set $\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$. Throughout this section we denote by $\mathbb{P}$ the unknown true distribution of the vector $\xi \in \mathbb{R}^n$ of risk factors, and we assume that $\mathbb{P}$ has a finite mean $\mu = \mathbb{E}_\mathbb{P}[\xi]$, second moment matrix $M = \mathbb{E}_\mathbb{P}[\xi\xi^\top]$ and covariance matrix $\Sigma = M - \mu\mu^\top$. Similarly, we define the sample mean, the sample second moment matrix and the sample covariance matrix corresponding to N independent points $\xi_1, \ldots, \xi_N$ sampled from $\mathbb{P}$ as

$$\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^N \xi_i, \quad \widehat{M}_N = \frac{1}{N}\sum_{i=1}^N \xi_i\xi_i^\top \quad \text{and} \quad \widehat{\Sigma}_N = \widehat{M}_N - \widehat{\mu}_N\widehat{\mu}_N^\top, \tag{5}$$

respectively. We stress that $\widehat{\Sigma}_N$ is defined without the usual Bessel correction, which will simplify our analysis. Throughout this section we will assume that $\mathbb{P}$ is sub-Gaussian.

DEFINITION 10—Sub-Gaussian probability distribution: *The probability distribution $\mathbb{P}$ of $\xi \in \mathbb{R}^n$ is sub-Gaussian with variance proxy $\sigma^2 \in \mathbb{R}_+$ if*

$$\mathbb{E}_\mathbb{P}\left[\exp\left(z^\top(\xi - \mathbb{E}_\mathbb{P}[\xi])\right)\right] \leq \exp\left(\tfrac{1}{2}\|z\|^2\sigma^2\right) \quad \forall z \in \mathbb{R}^n.$$

THEOREM 4—Finite-sample guarantee: *Suppose that $\mathbb{P}$ is sub-Gaussian with variance proxy σ^2, mean μ and covariance matrix $\Sigma \succ 0$, and denote by $\widehat{\mu}_N$ the sample mean and by $\widehat{\Sigma}_N$ the sample covariance matrix corresponding to N independent points sampled from $\mathbb{P}$ as in (5). Then, there exist positive constants c_1 and c_2 that depend on $\mathbb{P}$ only through μ, Σ, σ^2 and n such that for any significance level $\eta \in (0, 1]$ we have*

$$\mathbb{P}^N \left[\mathbb{G}\big((\widehat{\mu}_N, \widehat{\Sigma}_N), (\mu, \Sigma)\big) \leq \rho(\eta) \right] \geq 1 - \eta, \quad \textit{where} \quad \rho(\eta) = (c_1 + c_2 \log(1/\eta)) \,/ \sqrt{N}.$$

Theorem 4 guarantees that if $\rho \geq (c_1 + c_2 \log(1/\eta))/\sqrt{N}$, then the Gelbrich ambiguity set $\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$ contains the unknown data-generating distribution $\mathbb{P}$ with probability at least $1 - \eta$ under $\mathbb{P}^N$. This in turn immediately implies that the Gelbrich risk $\mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})}(\ell)$ of any fixed loss function $\ell \in \mathcal{L}_0$ provides an upper bound on the true risk $\mathcal{R}_{\mathbb{P}}(\ell)$, and the optimal Gelbrich risk $\mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})}(\mathcal{L})$ provides an upper bound on the true optimal risk $\inf_{\ell \in \mathcal{L}} \mathcal{R}_{\mathbb{P}}(\ell)$, with probability at least $1 - \eta$ under $\mathbb{P}^N$.

3.3. *Computational Properties of the Gelbrich Ambiguity Set*

We now establish basic properties of the Gelbrich ambiguity set that are conducive to the solvability and computational tractability of distributionally robust optimization problems with Gelbrich ambiguity sets. To this end, we first study the ball $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ in the space of mean-covariance pairs. It is natural to expect that this set is compact and convex because $\mathbb{G}$ represents a metric on $\mathbb{R}^n \times \mathbb{S}^n_+$. The next proposition formalizes this intuition.

PROPOSITION 2—Properties of $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$: *For any $\widehat{\mu} \in \mathbb{R}^n$, $\widehat{\Sigma} \in \mathbb{S}^n_+$ and $\rho \in \mathbb{R}_+$, the set $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is compact and convex.*

Consider now the Gelbrich ambiguity set $\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$ of Definition 8. Note that the mean $\mu = \mathbb{E}_{\mathbb{Q}}[\xi]$ and the second moment matrix $M = \mathbb{E}_{\mathbb{Q}}[\xi\xi^\top]$ are linear in the underlying probability distribution $\mathbb{Q}$, whereas the covariance matrix $\Sigma = M - \mu\mu^\top$ is indefinite quadratic in $\mathbb{Q}$. The constraint requiring the mean and covariance matrix of $\mathbb{Q}$ to fall into the convex set $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ thus appears to be non-convex in $\mathbb{Q}$. One might therefore suspect that the Gelbrich ambiguity set is non-convex and that evaluting the Gelbrich risk $\mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})}(\ell)$ is hard. We will now show that the Gelbrich ambiguity set is nonetheless convex. To this end, we define the following transform of $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$,

$$\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma}) = \left\{ (\mu, M) \in \mathbb{R}^n \times \mathbb{S}^n_+ : (\mu, M - \mu\mu^\top) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma}) \right\}.$$

Even though it is constructed as the pre-image of a convex set under an indefinite quadratic transformation, one can show that $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is convex.

PROPOSITION 3—Properties of $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$: *For any $\widehat{\mu} \in \mathbb{R}^n$, $\widehat{\Sigma} \in \mathbb{S}^n_+$ and $\rho \in \mathbb{R}_+$, the set $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is compact and convex.*

Proposition 3 implies that the Gelbrich ambiguity set $\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is convex because it can be viewed as the pre-image of the convex set $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ under the linear transformation that maps any probability distribution to its mean and second moment matrix. We formalize this insight in the next corollary, which we state without proof.

COROLLARY 2—Convexity of $\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$: *The Gelbrich ambiguity set is convex.*

To close this section, we establish a decomposition of the Gelbrich ambiguity set that will prove useful for evaluating $\mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})}(\ell)$ and $\mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})}(\mathcal{L})$. By definition, the Gelbrich ambiguity set $\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$ encompasses all distributions in $\mathcal{S}$ whose mean vectors and covariance matrices belong to $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$. Denoting by $\mathcal{C}(\mu, \Sigma)$ the *structured Chebyshev ambiguity set* that contains all distributions in $\mathcal{S}$ with mean μ and covariance matrix Σ, the Gelbrich ambiguity set can be decomposed as

$$\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma}) = \bigcup_{(\mu,\Sigma)\in\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \mathcal{C}(\mu, \Sigma). \tag{6}$$

In particular, we have $\mathcal{G}_0(\widehat{\mu}, \widehat{\Sigma}) = \mathcal{C}(\widehat{\mu}, \widehat{\Sigma})$. The decomposition (6) indicates that if $\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$ contains a particular distribution $\mathbb{Q}$, then it contains *all* distributions in $\mathcal{S}$ with the same mean and covariance matrix as $\mathbb{Q}$. Moreover, it allows us to represent the Gelbrich risk of any loss function $\ell \in \mathcal{L}_0$ as

$$\begin{aligned} \mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})}(\ell) &= \sup_{(\mu,\Sigma)\in\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \; \sup_{\mathbb{Q}\in\mathcal{C}(\mu,\Sigma)} \mathcal{R}_\mathbb{Q}(\ell) && \text{(7a)} \\ &= \sup_{(\mu,M)\in\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})} \; \sup_{\mathbb{Q}\in\mathcal{C}(\mu,M-\mu\mu^\top)} \mathcal{R}_\mathbb{Q}(\ell), && \text{(7b)} \end{aligned}$$

where the second equality holds because $(\mu, M) \in \mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ if and only if $(\mu, M - \mu\mu^\top) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$. The reformulation (7b) suggests that the Gelbrich risk evaluation problem may be computationally tractable in situations of practical interest. To see this, note that the inner maximization problem in (7b) simply evaluates the worst-case risk over the Chebyshev ambiguity set of all probability distributions with mean μ and second moment matrix M. If the risk measure $\mathcal{R}_\mathbb{Q}(\ell)$ is linear in $\mathbb{Q}$ (*e.g.*, if it represents the expected loss) or concave in $\mathbb{Q}$ (*e.g.*, if it represents the variance, the VaR or the CVaR of the loss), then this inner problem constitutes a convex mazimization problem over all probability distributions $\mathbb{Q}$ that satisfy the linear equality constraints $\mathbb{E}_\mathbb{Q}[\xi] = \mu$ and $\mathbb{E}_\mathbb{Q}[\xi\xi^\top] = M$. As concavity is preserved under partial maximization, the optimal value of the inner problem in (7a) is jointly concave in the constraint right hand sides μ and M. The outer problem in (7b) thus maximizes a concave function (*i.e.*, the worst-case Chebyshev risk) over all mean vectors and second moment matrices in the convex set $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$; see Proposition 3. Consequently, both the inner and the outer maximization problems in (7b) are convex, and thus there is hope that both of them are computationally tractable. We will further investigate the decomposition (7) in Section 4.

Conceptually, the inner problems in (7a) and (7b) hedge against uncertainty in the shape and the outer problems hedge against uncertainty in the location and dispersion of the distribution of the risk factors. We are not the first to study two-layer distributionally robust optimization problems with an outer layer that hedges against mean-covariance uncertainty. As the worst-case risk over a Chebyshev ambiguity set is concave in (μ, M) but non-concave in (μ, Σ) for most common risk measures, moment uncertainty has mostly been modeled through convex uncertainty sets for (μ, M). This choice leads to convex outer-layer problems. For example, uncertainty sets that restrict μ to an ellipsoid and M to the intersection of two positive semidefinite cones were proposed by Delage and Ye [2010], whereas rectangular uncertainty sets for (μ, M) were studied by Zymler et al. [2013b] and Hanasusanto et al. [2015]. Generic convex uncertainty sets for (μ, Σ) render the outer-layer problems convex only in special situations, *e.g.*, when the loss function is quadratic and the VaR is used as a risk measure; see [El Ghaoui et al., 2003, Rujeerapaiboon et al., 2016]. Our new convex uncertainty set $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ for (μ,Σ) is not only remarkable due to its connection to the Wasserstein ambiguity set but also because it leads to a convex outer-layer problem in (7a) irrespective of the loss function (as long as the risk measure is concave in $\mathbb{Q}$).

4. GELBRICH RISK OF PORTFOLIO LOSS FUNCTIONS

We now derive explicit formulas for the Gelbrich risk of portfolio loss functions of the form $\ell(\xi) = -w^\top \xi$, where ξ stands for the vector of asset returns, and w collects the portfolio weights. Thus, $\ell(\xi)$ represents the negative portfolio return. As a preparation, Section 4.1 reviews several basic properties of risk measures. Section 4.2 then shows that if the risk measure at hand is law-invariant, translation invariant and positive homogeneous (but not necessarily convex) and the structural ambiguity set satisfies a stability condition, then the Gelbrich risk simplifies to a regularized mean-standard deviation risk measure, which is convex and can be minimized efficiently. In addition, we analytically characterize the extremal distributions that attain the supremum in evaluating $\mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})}(\ell)$. Remarkably, the Gelbrich risk, its optimal portfolios as well as the corresponding extremal distributions depend on the underlying risk measure only through a scalar, which we term the *standard risk coefficient*, and which can be calculated offline. In Section 4.3 we thus provide closed-form expressions for the standard risk coefficients of the VaR, the CVaR and the mean-standard deviation risk measure. We also derive the standard risk coefficients of all spectral risk measures, all risk measures that admit a Kusuoka representation and all distortion risk measures. The online appendix shows that most results of Sections 4.2–4.3 extend to the mean-variance risk measures, even though they fail to be positive homogeneous.

4.1. *Basic Properties of Risk Measures*

Virtually all risk measures used in economics and finance are law-invariant [Föllmer and Schied, 2008b, § 4.5]. Such risk measures are usually defined in view of the probability distribution $\mathbb{P}$ of the relevant risk factors. As we study situations in which $\mathbb{P}$ is ambiguous, we now extend the notion of law-invariance to *families* of risk measures $\{\mathcal{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathcal{M}}$.

DEFINITION 11—Law-invariant family of risk measures: *The family of risk measures $\{\mathcal{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathcal{M}}$ is law-invariant if $\mathcal{R}_{\mathbb{Q}_1}(\ell_1) = \mathcal{R}_{\mathbb{Q}_2}(\ell_2)$ for any loss functions $\ell_1,\ell_2 \in \mathcal{L}_0$ and probability distributions $\mathbb{Q}_1,\mathbb{Q}_2 \in \mathcal{M}$ such that the distribution of $\ell_1(\xi)$ under $\mathbb{Q}_1$ matches the distribution of $\ell_2(\xi)$ under $\mathbb{Q}_2$.*

Note that if the family of risk measures $\{\mathcal{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathcal{M}}$ is law-invariant in the sense of Definition 11, then the risk measure $\mathcal{R}_\mathbb{Q}$ is law-invariant in the sense of [Föllmer and Schied, 2008b, § 4.5] for any fixed $\mathbb{Q}\in\mathcal{M}$. Conversely, the following remark shows that any law-invariant risk measure $\mathcal{R}_\mathbb{P}$ associated with a continuous probability distribution $\mathbb{P}\in\mathcal{M}$ naturally induces a law-invariant family of risk measures $\{\mathcal{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathcal{M}}$.

REMARK 1—Constructing a law-invariant family of risk measures: *Assume that $\mathcal{R}_\mathbb{P}$ is a law-invariant risk measure associated with a continuous probability distribution $\mathbb{P}\in\mathcal{M}$, and let $\mathbb{Q}\in\mathcal{M}$ be any other probability distribution. In particular, $\mathbb{Q}$ does not have to be continuous. Denote by $\varphi_\mathbb{P} : \mathbb{R}^n \to [0,1]^n$ the Rosenblatt transformation corresponding to $\mathbb{P}$ [Rosenblatt, 1952] and by $\psi_\mathbb{Q} : [0,1]^n \to \mathbb{R}^n$ the inverse Rosenblatt transformation corresponding to $\mathbb{Q}$ [Chen et al., 2011b, § 2.5]. As $\mathbb{P}$ is continuous, one can show that $\varphi_\mathbb{P}(\xi)$ is uniformly* *distributed on $[0,1]^n$ under $\mathbb{P}$. In addition, $\psi_{\mathbb{Q}}(\varphi_{\mathbb{P}}(\xi))$ follows the distribution $\mathbb{Q}$ under $\mathbb{P}$. We can now define a risk measure $\mathcal{R}_{\mathbb{Q}}$ corresponding to $\mathbb{Q}$ by setting $\mathcal{R}_{\mathbb{Q}}(\ell) = \mathcal{R}_{\mathbb{P}}(\ell(\psi_{\mathbb{Q}}(\varphi_{\mathbb{P}}(\xi))))$ for all $\ell \in \mathcal{L}_0$. The family $\{\mathcal{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathcal{M}}$ constructed in this way is law-invariant in the sense of Definition 11 as $\mathcal{R}_{\mathbb{P}}$ is law-invariant in the usual sense.*

Note that ‘risk measures’ in colloquial English (*e.g.*, the ‘variance,’ the ‘VaR’ or the ‘CVaR’ etc.) make no reference to a specific probability distribution and are therefore naturally interpreted as families of risk measures of the form $\{\mathcal{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathcal{M}}$. All of these standard families of risk measures are in fact law-invariant. We now recall some basic properties displayed by many popular risk measures.

DEFINITION 12—Properties of risk measures: *A risk measure $\mathcal{R}_{\mathbb{Q}}$ associated with a probability distribution $\mathbb{Q} \in \mathcal{M}$ is*

◇ *translation invariant if $\mathcal{R}_{\mathbb{Q}}(\ell + \lambda) = \mathcal{R}_{\mathbb{Q}}(\ell) + \lambda$ for all $\ell \in \mathcal{L}_0$, $\lambda \in \mathbb{R}$;*
◇ *positive homogeneous if $\mathcal{R}_{\mathbb{Q}}(\lambda\ell) = \lambda\mathcal{R}_{\mathbb{Q}}(\ell)$ for all $\ell \in \mathcal{L}_0$, $\lambda \in \mathbb{R}_+$;*
◇ *monotonic if $\mathcal{R}_{\mathbb{Q}}(\ell_1) \leq \mathcal{R}_{\mathbb{Q}}(\ell_2)$ for all $\ell_1, \ell_2 \in \mathcal{L}_0$ such that $\ell_1 \leq \ell_2$ $\mathbb{Q}$-almost surely;*
◇ *convex if $\mathcal{R}_{\mathbb{Q}}(\lambda\ell_1 + (1-\lambda)\ell_2) \leq \lambda\mathcal{R}_{\mathbb{Q}}(\ell_1) + (1-\lambda)\mathcal{R}_{\mathbb{Q}}(\ell_2)$ for all $\ell_1, \ell_2 \in \mathcal{L}_0$, $\lambda \in [0,1]$.*

A risk measure is called coherent if it satisfies all of the above properties.

4.2. *Law Invariant, Translation Invariant and Positive Homogeneous Risk Measures*

We will now demonstrate that many commonly used families of risk measures impact the Gelbrich risk of a linear loss function only through a scalar, which we define as follows.

DEFINITION 13—Standard risk coefficient: *The standard risk coefficient of a family $\{\mathcal{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathcal{M}}$ of risk measures corresponding to a structural ambiguity set $\mathcal{S}$, mean $\mu \in \mathbb{R}^n$, covariance matrix $\Sigma \in \mathbb{S}^n_+$ and portfolio vector $w \in \mathbb{R}^n$ with $w^\top \Sigma w \neq 0$ is*

$$\alpha(\mu, \Sigma, w) = \sup_{\mathbb{Q}\in\mathcal{C}(\mu,\Sigma)} \mathcal{R}_{\mathbb{Q}}\left(-\frac{w^\top(\xi-\mu)}{\sqrt{w^\top \Sigma w}}\right). \tag{8}$$

Recall that $\mathcal{C}(\mu,\Sigma)$ denotes the structured Chebyshev ambiguity set of all distributions in $\mathcal{S}$ with mean μ and covariance matrix Σ. For generic families of risk measures, the standard risk coefficient of Definition 13 depends on μ, Σ and w. However, if the family of risk measures is law-invariant and the structural ambiguity set is stable in the sense of the following definition, then the standard risk coefficient is constant in these parameters and depends solely on the family of risk measures and the structural ambiguity set at hand.

DEFINITION 14—Stable structural ambiguity set: *A structural ambiguity set $\mathcal{S}$ is stable if it is closed under arbitrary affine pushforwards and convolutions. Thus, if $\mathbb{Q} \in \mathcal{S}$ and $f : \mathbb{R}^n \to \mathbb{R}^n$ is of the form $f(\xi) = A\xi + b$ for some $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, then $\mathbb{Q} \circ f^{-1} \in \mathcal{S}$. Similarly, if $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{S}$, then $\mathbb{Q}_1 * \mathbb{Q}_2 \in \mathcal{S}$, where the convolution is defined through $\mathbb{Q}_1 * \mathbb{Q}_2(B) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \mathbb{1}_{\xi_1+\xi_2\in B}\, \mathrm{d}\mathbb{Q}_1(\xi_1)\, \mathrm{d}\mathbb{Q}_2(\xi_2)$ for all Borel sets $B \in \mathcal{B}(\mathbb{R}^n)$.*

To motivate our terminology, recall that a distribution is called stable if any linear combination of two independent random variables with this distribution has the same distribution up to location and scaling. For example, the Gaussian, Cauchy and Lévy distributions are stable. Definition 14 generalizes this notion to ambiguity sets. Indeed, it implies that if two independent random vectors ξ_1 and ξ_2 have distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$, respectively, both of which belong to the same stable structural ambiguity set $\mathcal{S}$, and if $A_1, A_2 \in \mathbb{R}^{n \times n}$, then the probability distribution of the linear combination $\xi = A_1\xi_1 + A_2\xi_2$ also belongs to $\mathcal{S}$. The structural ambiguity set $\mathcal{M}_2$ of all distributions with finite second moment is trivially stable. By [Yu et al., 2009, § 2], the sets of all symmetric, all symmetric linear unimodal and all log-concave distributions with finite second moments also constitute stable ambiguity sets. In addition, the ambiguity set of all Gaussian distributions is stable. However, some structural ambiguity sets fail to be stable. The set of all elliptical distributions with the same characteristic generator, for example, is not necessarily stable. Indeed, the convolution of two Laplace distributions is not a Laplace distribution, for instance. One can also show that the structural ambiguity set of all linear unimodal (but not necessarily symmetric) distributions also fails to be stable even in the univariate case. One can further show that the structural ambiguity set generated by a distribution $\widehat{\mathbb{P}}$ is stable only if $\widehat{\mathbb{P}}$ is Gaussian.

We can now state the announced result.

PROPOSITION 4—Standard risk coefficient: *If $\{\mathcal{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathcal{M}}$ is a law-invariant family of risk measures and the structural ambiguity set $\mathcal{S}$ is stable, then the corresponding standard risk coefficient α is independent of μ, Σ and w.*

Proposition 4 is a key ingredient to prove our following main result.

THEOREM 5—Gelbrich risk of linear loss functions: *If $\{\mathcal{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathcal{M}}$ is a law-invariant family of translation invariant and positive homogeneous risk measures, the structural ambiguity set $\mathcal{S}$ is stable and the corresponding standard risk coefficient satisfies $0 \leq \alpha < +\infty$, then the Gelbrich risk of the portfolio loss function $\ell(\xi) = -w^\top\xi$ is given by*

$$\sup_{\mathbb{Q}\in\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})} \mathcal{R}_\mathbb{Q}\left(-w^\top\xi\right) = -\widehat{\mu}^\top w + \alpha\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+\alpha^2}\,\|w\|. \tag{9}$$

In addition, if $\mathcal{S}$ is the structural ambiguity set generated by a Gaussian nominal distribution $\widehat{\mathbb{P}}$ with $\widehat{\Sigma} \succ 0$, then the Wasserstein risk coincides with the Gelbrich risk.

We emphasize that the standard risk coefficient α can in general be negative. In this case, evaluating the Gelbrich risk of $-w^\top\xi$ requires the solution of a non-convex optimization problem, and Theorem 5 no longer holds (see problem (13) in the proof of Theorem 5). A sufficient condition for the non-negativity of α is described in the following proposition.

PROPOSITION 5—Non-negative standard risk coefficient: *If $\{\mathcal{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathcal{M}}$ is a law-invariant family of coherent risk measures and the structural ambiguity set $\mathcal{S}$ contains a symmetric distribution, then the corresponding standard risk coefficient α is non-negative.*

From now on we assume that $\{\mathcal{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathcal{M}}$ is a law-invariant family of translation invariant and positive homogeneous risk measures and that $\mathcal{S}$ is a stable ambiguity set with standard risk coefficient $\alpha \geq 0$. Denoting by $\mathcal{L}$ the set of portfolio loss functions $\xi \mapsto \ell(\xi) = -w^\top\xi$ with portfolio weights w belonging to a set $\Omega \subseteq \mathbb{R}^n$, Theorem 5 allows us to reformulate the optimal Gelbrich risk $\mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})}(\mathcal{L})$ as

$$\min_{w\in\Omega}\sup_{\mathbb{Q}\in\mathcal{G}_\rho(\widehat{\mathbb{P}})} \mathcal{R}_\mathbb{Q}(-w^\top\xi) = \min_{w\in\Omega}\; -\widehat{\mu}^\top w + \alpha\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+\alpha^2}\|w\|. \tag{10}$$

Note that if Ω is convex, then (10) constitutes a finite convex program. In particular, if Ω is representable via second-order cone constraints, then problem (10) reduces to a tractable second-order cone program that can be solved highly efficiently with off-the-shelf solvers. As the Gelbrich risk upper bounds the Wasserstein risk by virtue of Corollary 1, the convex program (10) provides a conservative and efficiently computable proxy for the optimal Wasserstein risk $\mathcal{R}_{\mathcal{W}_\rho(\widehat{\mathbb{P}})}(\mathcal{L})$, which may be hard to compute exactly. Note also that (10) can be interpreted as a regularized Markowitz portfolio selection problem with an ℓ_2-regularization term that scales with the size parameter ρ of the Gelbrich ambiguity set.

Under mild conditions on Ω, one can show that the optimizer of the Gelbrich risk portfolio selection problem (10) converges to the equally weighted portfolio as ρ tends to infinity (*i.e.*, in the limit of extreme uncertainty). A similar result was proved by Pflug et al. [2012] for a specific class of *convex* risk measures and for a Wasserstein ambiguity set.

COROLLARY 3—The equally weighted portfolio is optimal under high uncertainty: *If the assumptions of Theorem 5 hold, $\Omega \subseteq \{w \in \mathbb{R}^n : e^\top w = 1\}$ is a closed set of portfolio weights with $e \in \mathbb{R}^n$ being the vector of all ones, and if Ω contains the equally weighted portfolio $\frac{1}{n}e$, then the unique minimizer of* (10) *converges to $\frac{1}{n}e$ as ρ tends to ∞.*

Worst-case distributions that maximize the risk of a fixed portfolio over the Gelbrich ambiguity set can expose potential threats to the portfolio or may be useful for stress test experiments. Therefore, we now aim to characterize the worst-case distributions $\mathbb{Q}^\star$ that attain the supremum in evaluating the Gelbrich risk $\mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})}(\ell)$ for $\ell(\xi) = -w^\top \xi$.

PROPOSITION 6—Worst-case moments: *If $\{\mathcal{R}_\mathbb{Q}\}_{\mathbb{Q}\in\mathcal{M}}$ is a law-invariant family of translation invariant and positive homogeneous risk measures, the structural ambiguity set $\mathcal{S}$ is stable, the standard risk coefficient satisfies $0 < \alpha < +\infty$ and $\widehat{\Sigma} \succ 0$, then any extremal distribution $\mathbb{Q}^\star$ that attains the Gelbrich risk of the loss function $\ell(\xi) = -w^\top\xi$ has the same mean $\mu^\star \in \mathbb{R}^n$ and covariance matrix $\Sigma^\star \in \mathbb{S}^n_+$, where*

$$\mu^\star = \widehat{\mu} - \frac{\rho}{\sqrt{1+\alpha^2}\|w\|} w \quad \textit{and}$$

$$\Sigma^\star = \left(I + \frac{\rho\alpha w w^\top}{\sqrt{1+\alpha^2}\|w\|\sqrt{w^\top\widehat{\Sigma}w}} \right) \widehat{\Sigma} \left(I + \frac{\rho\alpha w w^\top}{\sqrt{1+\alpha^2}\|w\|\sqrt{w^\top\widehat{\Sigma}w}} \right).$$

Proposition 6 characterizes only the first two moments of the extremal distributions that attain the Gelbrich risk. In general, $\mathcal{S}$ may contain multiple distributions with these moments. When $\mathcal{S}$ is the set of all Gaussian distributions, however, the Gelbrich risk is uniquely attained by the Gaussian distribution with mean $\mu^\star$ and covariance matrix $\Sigma^\star$.

4.3. *Calculation of the Standard Risk Coefficient*

We now show that the standard risk coefficient α is given closed form for a large class of risk measures. First, we focus on the value-at-risk (VaR). For any probability distribution $\mathbb{Q} \in \mathcal{M}$, the VaR at level $\beta \in (0,1)$ of any loss function $\ell \in \mathcal{L}_0$ is defined as

$$\mathbb{Q}\text{-}\operatorname{VaR}_\beta(\ell) = \inf\left\{\tau \in \mathbb{R} : \mathbb{Q}[\ell(\xi) \le \tau] \ge 1-\beta\right\}.$$

VaR fails to be convex, yet it is widely used by financial institutions and regulators [Jorion, 1996, Longerstaey and Spencer, 1996, Duffie and Pan, 1997]. In addition, VaR induces a law-invariant family of translation invariant and positive homogeneous risk measures, and thus

Proposition 4 and Theorem 5 apply whenever the structural ambiguity set $\mathcal{S}$ is stable. The following proposition describes situations in which α is available in closed form.

PROPOSITION 7—Standard risk coefficient for VaR: *If $\beta \in (0,1)$ and $\mathcal{R}_{\mathbb{Q}} = \mathbb{Q}\text{-VaR}_\beta$ for $\mathbb{Q} \in \mathcal{M}$, then α is available in closed form for several stable structural ambiguity sets.*

(i) If $\mathcal{S} = \mathcal{M}_2$, then $\alpha = \sqrt{(1-\beta)/\beta}$.

(ii) If $\mathcal{S}$ is the set of all symmetric distributions in $\mathcal{M}_2$, then $\alpha = \sqrt{1/(2\beta)}$ for $\beta < \frac{1}{2}$ and $\alpha = 0$ for $\beta \geq \frac{1}{2}$.

(iii) If $\mathcal{S}$ is the set of all symmetric linear unimodal distributions in $\mathcal{M}_2$, then $\alpha = 2/(3\sqrt{2\beta})$ for $\beta < \frac{1}{2}$ and $\alpha = 0$ for $\beta \geq \frac{1}{2}$.

(iv) If $\mathcal{S}$ is the set of all Gaussian distributions, then $\alpha = \Phi^{-1}(1-\beta)$, where Φ denotes the cumulative distribution function of the standard Gaussian distribution.

In assertions *(i)*, *(ii)* and *(iii)* of Proposition 7 the standard risk coefficient α is non-negative for all $\beta \in (0,1)$ even though Proposition 5 does not apply (as VaR fails to be convex). In assertion *(iv)*, on the other hand, α becomes negative for $\beta > \frac{1}{2}$. Hence, Theorem 5 does *not* apply to the VaR at level $\beta > \frac{1}{2}$ if $\mathcal{S}$ is the family of all Gaussian distributions.

We now address the conditional value-at-risk (CVaR). For any distribution $\mathbb{Q} \in \mathcal{M}$, the CVaR at level $\beta \in (0,1)$ of any loss function $\ell \in \mathcal{L}_0$ is defined as

$$\mathbb{Q}\text{-}\operatorname{CVaR}_\beta(\ell) = \inf_{\tau \in \mathbb{R}} \left\{ \tau + \frac{1}{\beta} \mathbb{E}_{\mathbb{Q}} \left[\max\{\ell(\xi) - \tau, 0\} \right] \right\}.$$

It is well known that CVaR induces a law-invariant family of coherent risk measures [Artzner et al., 1999, Rockafellar and Uryasev, 2000], and thus Proposition 4 and Theorem 5 apply whenever the structural ambiguity set $\mathcal{S}$ is stable. The following proposition shows that α is again available in closed form in several situations of practical interest.

PROPOSITION 8—Standard risk coefficient for CVaR: *If $\beta \in (0,1)$ and $\mathcal{R}_{\mathbb{Q}} = \mathbb{Q}\text{-CVaR}_\beta$ for $\mathbb{Q} \in \mathcal{M}$, then α is available in closed form for several stable structural ambiguity sets.*

(i) If $\mathcal{S} = \mathcal{M}_2$, then $\alpha = \sqrt{(1-\beta)/\beta}$.

(ii) If $\mathcal{S}$ is the set of all symmetric distributions in $\mathcal{M}_2$, then $\alpha = \sqrt{1/(2\beta)}$ for $\beta < \frac{1}{2}$ and $\alpha = \sqrt{1-\beta}/(\sqrt{2}\beta)$ for $\beta \geq \frac{1}{2}$.

(iii) If $\mathcal{S}$ is the set of all symmetric linear unimodal distributions in $\mathcal{M}_2$, then $\alpha = 2/(3\sqrt{\beta})$ for $\beta \leq \frac{1}{3}$, $\alpha = \sqrt{3}(1-\beta)$ for $\frac{1}{3} < \beta \leq \frac{2}{3}$ and $\alpha = 2\sqrt{1-\beta}/(3\beta)$ for $\beta > \frac{2}{3}$.

(iv) If $\mathcal{S}$ is the set of all Gaussian distributions, then $\alpha = (\sqrt{2\pi}\beta)^{-1}\exp(-(\Phi^{-1}(1-\beta))^2/2)$, where Φ denotes the cumulative distribution function of the standard Gaussian distribution.

Propositions 7 *(i)* and 8 *(i)* imply that the standard risk coefficients for VaR and CVaR coincide if $\mathcal{S}$ represents the family of all distributions with finite second moments. This result is reminiscent of the observation that distributionally robust chance constraints are equivalent to distributionally robust CVaR constraints when the distributional uncertainty is modeled by a Chebyshev ambiguity set [Zymler et al., 2013a, Theorem 2.2]. We emphasize that the standard risk coefficients for VaR and CVaR differ under the structural ambiguity sets of assertions *(ii)*, *(iii)* and *(iv)* of Propositions 7 and 8, respectively.

Next, we study mean-standard deviation risk measures that are ubiquitous in classical portfolio theory [Rockafellar et al., 2002]. For any distribution $\mathbb{Q} \in \mathcal{M}$, the mean-standard deviation risk measure with risk-aversion coefficient $\beta \geq 0$ of any loss function $\ell \in \mathcal{L}_0$ is defined as $\mathcal{R}_{\mathbb{Q}}(\ell) = \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] + \beta(\mathbb{V}\mathrm{ar}_{\mathbb{Q}}(\ell(\xi)))^{1/2}$, where $\mathbb{V}\mathrm{ar}_{\mathbb{Q}}(\ell(\xi))$ denotes the variance of $\ell(\xi)$ under $\mathbb{Q}$. For any fixed β, the mean-standard deviation risk measure induces a law-invariant family of translation invariant and positive homogeneous risk measures, and thus Proposition 4 and Theorem 5 apply if $\mathcal{S}$ is stable. Note that this includes the expected loss, for $\beta = 0$. Proposition 9 below derives α again in closed form.

PROPOSITION 9—Standard risk coefficient for mean-standard deviation risk measures: *If $\mathcal{R}_{\mathbb{Q}}$ is the mean-standard deviation risk measure with risk-aversion coefficient $\beta \geq 0$ for every $\mathbb{Q} \in \mathcal{M}$ and if the structural ambiguity set $\mathcal{S}$ is stable, then $\alpha = \beta$.*

Consider now the family of spectral risk measures introduced by Acerbi [2002]. In the following discussion, for any $\ell \in \mathcal{L}_0$ and $\mathbb{Q} \in \mathcal{M}$ we use $F^{\mathbb{Q}}_{\ell(\xi)}$ to denote the cumulative distribution function of $\ell(\xi)$ under $\mathbb{Q}$. In addition, we define the quantile function $(F^{\mathbb{Q}}_{\ell(\xi)})^{-1}$ through $(F^{\mathbb{Q}}_{\ell(\xi)})^{-1}(\tau) = \inf\{q \in \mathbb{R} : F^{\mathbb{Q}}_{\ell(\xi)}(q) \geq \tau\}$ for all $\tau \in (0,1)$.

DEFINITION 15—Spectral risk measures: *An admissible spectrum is a right-continuous and nondecreasing function $\psi : [0,1) \to \mathbb{R}_+$ with $\int_0^1 \psi(\tau)\mathrm{d}\tau = 1$. The spectral risk measure $\mathcal{R}_{\mathbb{Q}}$ induced by ψ under a given distribution $\mathbb{Q} \in \mathcal{M}$ of the risk factors is defined through*

$$\mathcal{R}_{\mathbb{Q}}(\ell) = \int_0^1 \psi(\tau)(F^{\mathbb{Q}}_{\ell(\xi)})^{-1}(\tau)\mathrm{d}\tau \quad \forall \ell \in \mathcal{L}_0.$$

For any fixed $\mathbb{Q}$, the set of all spectral risk measures coincides with the family of all coherent, law-invariant and comonotonic risk measures that satisfy a nonrestrictive Fatou property [Kusuoka, 2001, Theorem 7]. On the other hand, any fixed admissible spectrum ψ induces a family of spectral risk measures parametrized by the distributions $\mathbb{Q} \in \mathcal{M}$. This family is law-invariant by construction, and thus Proposition 4 and Theorem 5 apply whenever $\mathcal{S}$ is stable. The following proposition evaluates α in closed form for $\mathcal{S} = \mathcal{M}_2$.

PROPOSITION 10—Standard risk coefficient for spectral risk measures: *If there exists a square-integrable admissible spectrum ψ such that $\mathcal{R}_{\mathbb{Q}}$ is the spectral risk measure induced by ψ for every $\mathbb{Q} \in \mathcal{M}$ and if $\mathcal{S} = \mathcal{M}_2$, then $\alpha = (\int_0^1 \psi(\tau)^2\mathrm{d}\tau - 1)^{\frac{1}{2}}$.*

Note that the CVaR at level $\beta \in (0,1)$ is a spectral risk measure with spectrum $\psi(\tau) = \beta^{-1}\mathbb{1}_{[1-\beta,1)}(\tau)$; see, *e.g.*, [Föllmer and Schied, 2008b, Definition 4.43 and Lemma 4.46]. By Proposition 10, the standard risk coefficient of the CVaR is thus given by $(\int_0^1 \psi(\tau)^2\mathrm{d}\tau - 1)^{\frac{1}{2}} = \sqrt{(1-\tau)/\tau}$, which confirms the formula derived in Proposition 8 *(i)*.

Next, we address risk measures that admit a Kusuoka representation [Kusuoka, 2001, Shapiro, 2013] and can be expressed as suprema over families of spectral risk measures.

DEFINITION 16—Kusuoka representation: *A risk measure $\mathcal{R}_{\mathbb{Q}}$ admits a Kusuoka representation under the distribution $\mathbb{Q} \in \mathcal{M}$ of the risk factors if there exists a set Ψ of admissible spectra in the sense of Definition 15 such that*

$$\mathcal{R}_{\mathbb{Q}}(\ell) = \sup_{\psi \in \Psi} \int_0^1 \psi(\tau)(F^{\mathbb{Q}}_{\ell(\xi)})^{-1}(\tau)\mathrm{d}\tau \quad \forall \ell \in \mathcal{L}_0.$$

For any fixed $\mathbb{Q}$, the set of all risk measures that admit a Kusuoka representation coincides with the family of all coherent law-invariant risk measures satisfying the Fatou property [Kusuoka, 2001, Theorem 10]. On the other hand, any fixed set Ψ of admissible spectra induces a law-invariant family of coherent risk measures parametrized by the distributions $\mathbb{Q} \in \mathcal{M}$, and thus Proposition 4 and Theorem 5 apply whenever $\mathcal{S}$ is stable. The following proposition presents a closed-form expression for α if $\mathcal{S} = \mathcal{M}_2$.

PROPOSITION 11—Standard risk coefficient for risk measures with a Kusuoka representation: *If there exists a set Ψ of square-integrable admissible spectra such that $\mathcal{R}_{\mathbb{Q}}$ admits a Kusuoka representation induced by Ψ for every $\mathbb{Q} \in \mathcal{M}$ and if $\mathcal{S} = \mathcal{M}_2$, then $\alpha = \sup_{\psi \in \Psi}(\int_0^1 \psi(\tau)^2 \mathrm{d}\tau - 1)^{\frac{1}{2}}$.*

Lastly, we study the family of distortion risk measures, which measure the risk of an uncertain loss function by its expected value under a distorted distribution [Yaari, 1987].

DEFINITION 17—Distortion risk measures: *An admissible distortion is a nondecreasing function $h : [0,1] \to [0,1]$ with $\lim_{\tau \downarrow 0} h(\tau) = h(0) = 0$ and $\lim_{\tau \uparrow 1} h(\tau) = h(1) = 1$. The distortion risk measure $\mathcal{R}_{\mathbb{Q}}$ induced by h under the distribution $\mathbb{Q} \in \mathcal{M}$ is defined through*

$$\mathcal{R}_{\mathbb{Q}}(\ell) = \int_0^{\infty} \left(1 - h\left(F^{\mathbb{Q}}_{\ell(\xi)}(\tau)\right)\right) \mathrm{d}\tau - \int_{-\infty}^{0} h\left(F^{\mathbb{Q}}_{\ell(\xi)}(\tau)\right) \mathrm{d}\tau \quad \forall \ell \in \mathcal{L}_0.$$

If the distortion h is right-continuous, then we have

$$\mathcal{R}_{\mathbb{Q}}(\ell) = \int_0^1 \left(F^{\mathbb{Q}}_{\ell(\xi)}\right)^{-1}(\tau)\, \mathrm{d}h(\tau) = \int_{\mathbb{R}} \tau\, \mathrm{d}h\left(F^{\mathbb{Q}}_{\ell(\xi)}(\tau)\right), \tag{11}$$

where the first equality follows from [Cai et al., 2020, Lemma 1], whereas the second equality follows from the definition of the Lebesgue-Stieltjes integral [Riesz and Sz.-Nagy, 1990, § 3]. The resulting reformulation reveals that $\mathcal{R}_{\mathbb{Q}}(\ell)$ can be viewed as the expected value of the distorted cumulative distribution function $h \circ F^{\mathbb{Q}}_{\ell(\xi)}$. The simplest distortion risk measure is the ordinary expectation, which is induced by the trivial distortion $h(\tau) = \tau$. Other examples include the VaR and the CVaR at level $\beta \in (0,1)$, which are induced by the distortions $h(\tau) = \mathbb{1}_{[\beta,1]}(\tau)$ and $h(\tau) = \max\{\tau - 1 + \beta, 0\}/\beta$, respectively; see, *e.g.*, [Föllmer and Schied, 2008b, Definition 4.43 and Lemma 4.46]. Moreover, one readily verifies that the spectral risk measure with admissible spectrum ψ coincides with the distortion risk measure induced by the convex continuous distortion $h(\tau) = \int_0^{\tau} \psi(t)\mathrm{d}t$. A comprehensive list of distortion risk measures is provided by Cai et al. [2020].

Note that $(F^{\mathbb{Q}}_{\ell(\xi)})^{-1}(\tau) = \mathbb{Q}\text{-}\mathrm{VaR}_{1-\tau}(\ell)$ for all $\tau \in (0,1)$, and recall that VaR induces a law-invariant family of translation invariant and positive homogeneous risk measures. The second expression in (11) thus implies that the distortion risk measure corresponding to a right-continuous distortion h represents an average of VaRs with different levels τ. Hence, any such distortion risk measure induces a law-invariant family of translation invariant and positive homogeneous risk measures, which implies that Proposition 4 and Theorem 5 apply if $\mathcal{S}$ is stable. Next, we provide a closed-form expression for α when $\mathcal{S} = \mathcal{M}_2$.

PROPOSITION 12—Standard risk coefficient for distortion risk measures: *If there exists an admissible right-continuous distortion h such that $\mathcal{R}_{\mathbb{Q}}$ is the distortion risk measure induced by h for every $\mathbb{Q} \in \mathcal{M}$ and if $\mathcal{S} = \mathcal{M}_2$, then $\alpha = (\int_0^1 h'_{\mathrm{cvx}}(\tau)^2 \mathrm{d}\tau - 1)^{\frac{1}{2}}$, where h'_{cvx} denotes the derivative of the convex envelope of h, which exists almost everywhere.*

5. CONCLUSION

Our novel notion of structural Gelbrich ambiguity provides a universal framework for mean-covariance robust risk measurement and portfolio optimization. Our approach is related to Wasserstein distributionally robust optimization and exhibits superior statistical and computational properties than existing approaches. A large class of law and translation invariant, positive homogeneous risk measures boil down to the Markowitz model, subject to an additional term penalizing the norm of the portfolio weight vector. The risk-aversion coefficient is implied and given explicitly in terms of the mean-covariance uncertainty parameter. Our approach unifies and nests much of the standing and more recent literature on distributionally robust portfolio optimization of the past two decades. It thus paves the way for further empirical research.

APPENDIX: PROOFS OF SECTION 4

PROOF OF PROPOSITION 4: Define $\mathcal{F}$ as the set of all cumulative distribution functions on $\mathbb{R}$. Thus, $\mathcal{F}$ contains all nondecreasing and right-continuous functions $F : \mathbb{R} \to [0,1]$ with $\lim_{t\downarrow\infty} F(t) = 0$ and $\lim_{t\uparrow\infty} F(t) = 1$. For any loss function $\ell \in \mathcal{L}_0$ and probability distribution $\mathbb{Q} \in \mathcal{M}$, we use $F^{\mathbb{Q}}_{\ell(\xi)} \in \mathcal{F}$ to denote the distribution function of the random variable $\ell(\xi)$ under $\mathbb{Q}$. As the family of risk measures $\{\mathcal{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathcal{M}}$ is law-invariant, there exists a distribution functional $\varrho : \mathcal{F} \to \mathbb{R}$ with $\mathcal{R}_{\mathbb{Q}}(\ell) = \varrho(F^{\mathbb{Q}}_{\ell(\xi)})$ for all $\ell \in \mathcal{L}_0$ and $\mathbb{Q} \in \mathcal{M}$.

Define now $\mathcal{F}(0,1) \subseteq \mathcal{F}$ as the family of all cumulative distribution functions of the random variables of the form $v^\top \xi$ for some $v \in \mathbb{R}^n$ under any probability distribution $\mathbb{D} \in \mathcal{S}$ under which $v^\top \xi$ has zero mean and unit variance, that is, we set

$$\mathcal{F}(0,1) = \left\{ F^{\mathbb{D}}_{v^\top \xi} : v \in \mathbb{R}^n,\ \mathbb{D} \in \mathcal{S} \text{ such that } \mathbb{E}_{\mathbb{D}}[v^\top \xi] = 0 \text{ and } \mathbb{E}_{\mathbb{D}}[(v^\top \xi)^2] = 1 \right\}.$$

Next, choose any $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{S}^n_+$ and $w \in \mathbb{R}^n$ with $w^\top \Sigma w > 0$, and set $m = w^\top \mu$ and $s = \sqrt{w^\top \Sigma w}$. Recalling that the structured Chebyshev ambiguity set $\mathcal{C}(\mu, \Sigma)$ contains all distributions $\mathbb{Q} \in \mathcal{S}$ with mean μ and covariance matrix Σ, we will first show that

$$\mathcal{F}(0,1) = \left\{ F^{\mathbb{Q}}_{-(w^\top \xi - m)/s} : \mathbb{Q} \in \mathcal{C}(\mu, \Sigma) \right\}. \tag{12}$$

The proof proceeds in two steps. In the first step, we prove that the left hand side of (12) is a subset of the right hand side. To this end, select any $F \in \mathcal{F}(0,1)$. Thus, there exists a random vector $\xi_1 \in \mathbb{R}^n$ with probability distribution $\mathbb{D}_1 \in \mathcal{S}$ and a deterministic vector $v \in \mathbb{R}^n$ such that $v^\top \xi_1$ has zero mean and unit variance under $\mathbb{D}_1$ and such that $F = F^{\mathbb{D}_1}_{v^\top \xi_1}$. Using similar ideas as in [Yu et al., 2009, Theorems 1 and 2], we construct another random vector $\xi_2 \in \mathbb{R}^n$ with distribution $\mathbb{D}_2 \in \mathcal{C}(0, I)$ that is independent of ξ_1. We can then construct a new random vector $\xi \in \mathbb{R}^n$ as the following linear combination of ξ_1 and ξ_2.

$$\xi = \mu - \tfrac{1}{s} \Sigma w v^\top \xi_1 + \left(I - \tfrac{1}{s^2} \Sigma w w^\top \right) \Sigma^{\frac{1}{2}} \xi_2$$

Next, we define $\mathbb{Q}$ as the probability distribution of ξ. As $\mathbb{D}_1$ and $\mathbb{D}_2$ belong to the structural ambiguity set $\mathcal{S}$ and as $\mathcal{S}$ is stable and therefore closed under (not necessarily positive semidefinite) affine pushforwards and under convolutions, we may conclude that $\mathbb{Q} \in \mathcal{S}$. In addition, an elementary calculation exploiting our knowledge that $v^\top \xi_1$ and ξ_2 are standardized under the distributions $\mathbb{D}_1$ and $\mathbb{D}_2$, respectively, reveals that ξ has mean μ and covariance matrix Σ under $\mathbb{Q}$. Hence, we have shown that $\mathbb{Q} \in \mathcal{C}(\mu, \Sigma)$. By the construction of ξ and the definitions of m and s, we finally have $-(w^\top \xi - m)/s = v^\top \xi_1$ and thus $F^{\mathbb{Q}}_{-(w^\top \xi - m)/s} = F$. This implies that F belongs to the set on the right hand side of (12).

In the second step, we prove that the right hand side of (12) is a subset of the left hand side. To this end, select any $\mathbb{Q} \in \mathcal{C}(\mu, \Sigma)$, and use F as a shorthand for $F^{\mathbb{Q}}_{-(w^\top \xi - m)/s}$. To show that $F \in \mathcal{F}(0,1)$, set $v = -w/s$, and define the pushforward distribution $\mathbb{D} = \mathbb{Q} \circ f^{-1}$ with respect to the transformation $f(\xi) = \xi + mv/(s\|v\|^2)$. Note first that $\mathbb{D} \in \mathcal{S}$ because the structural ambiguity set $\mathcal{S}$ is closed under affine pushforwards. In addition, one readily verifies that $-(w^\top \xi - m)/s = v^\top f(\xi)$ has zero mean and unit variance under $\mathbb{Q}$, which means that $v^\top \xi$ has zero mean and unit variance under $\mathbb{D}$. We thus have $F \in \mathcal{F}(0,1)$. In combination, the first and the second step of the proof establish equation (12).

We may now conclude that the standard risk coefficient satisfies

$$\alpha = \sup_{\mathbb{Q} \in \mathcal{C}(\mu,\Sigma)} \mathcal{R}_{\mathbb{Q}} \left(-\frac{1}{s}(w^\top \xi - m) \right) = \sup_{\mathbb{Q} \in \mathcal{C}(\mu,\Sigma)} \varrho \left(F^{\mathbb{Q}}_{-(w^\top \xi - m)/s} \right) = \sup_{F \in \mathcal{F}(0,1)} \varrho(F),$$

where the second equality re-expresses the risk measures $\mathcal{R}_{\mathbb{Q}}$ for $\mathbb{Q} \in \mathcal{M}$ in terms of the distribution functional ϱ, and the second equality exploits (12). The last expression is manifestly independent of μ, Σ and w. This observation completes the proof. *Q.E.D.*

PROOF OF THEOREM 5: Decomposing the Gelbrich ambiguity set into disjoint Chebyshev ambiguity sets allows us to rewrite the Gelbrich risk as in (7a). As the loss function $\ell(\xi) = -w^\top \xi$ is linear, the inner maximization problem in (7a) further simplifies to

$$\begin{aligned} \sup_{\mathbb{Q} \in \mathcal{C}(\mu,\Sigma)} \mathcal{R}_{\mathbb{Q}} \left(-w^\top \xi \right) &= -w^\top \mu + \sup_{\mathbb{Q} \in \mathcal{C}(\mu,\Sigma)} \mathcal{R}_{\mathbb{Q}} \left(-w^\top (\xi - \mu) \right) \\ &= -w^\top \mu + \sqrt{w^\top \Sigma w} \sup_{\mathbb{Q} \in \mathcal{C}(\mu,\Sigma)} \mathcal{R}_{\mathbb{Q}} \left(-\frac{w^\top (\xi - \mu)}{\sqrt{w^\top \Sigma w}} \right) = -w^\top \mu + \alpha \sqrt{w^\top \Sigma w} \end{aligned}$$

where the first two equalities exploit the translation invariance and the positive homogeneity of the risk measures $\mathcal{R}_{\mathbb{Q}}$, $\mathbb{Q} \in \mathcal{M}$, respectively, whereas the third equality follows from the definition of the standard risk coefficient α. Note that α is independent of μ, Σ and w thanks to Proposition 4, which applies because the structural ambiguity set is stable and the family of risk measures is law-invariant. Using the definition of the set $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$, the outer maximization problem in (7a) can then be reformulated as

$$\begin{aligned} \sup_{\mathbb{Q} \in \mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})} \mathcal{R}_{\mathbb{Q}}(-w^\top \xi) &= \sup_{(\mu,\Sigma) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})} -\mu^\top w + \alpha \sqrt{w^\top \Sigma w} \\ &= \left\{ \begin{array}{ll} \sup\limits_{\mu, \Sigma \succeq 0} & -\mu^\top w + \alpha \sqrt{w^\top \Sigma w} \\ \text{s. t.} & \|\mu - \widehat{\mu}\|^2 + \operatorname{Tr}\left[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}} \big)^{\frac{1}{2}} \right] \le \rho^2. \end{array} \right. \end{aligned} \tag{13}$$

If $\rho = 0$, then $(\mu, \Sigma) = (\widehat{\mu}, \widehat{\Sigma})$ is the only feasible solution of problem (13), in which case (9) trivially holds. Similarly, if $\alpha = 0$, then $\Sigma = \widehat{\Sigma}$ and $\mu = \widehat{\mu} - \rho w / \|w\|$ are optimal in (13), and (9) again trivially holds. From now on we may thus assume without loss of generality that $\rho > 0$ and $\alpha > 0$. In this case problem (13) is equivalent to

$$\sup_{\mu, \Sigma \succeq 0} \inf_{\gamma \ge 0} \quad -\mu^\top w + \alpha \sqrt{w^\top \Sigma w} + \gamma \left[\rho^2 - \|\mu - \widehat{\mu}\|^2 - \operatorname{Tr}\left[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}} \big)^{\frac{1}{2}} \right] \right]$$

$$= \inf_{\gamma \geq 0} \left\{ \gamma\big(\rho^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\big) + \sup_{\mu} \left\{ -\mu^\top w - \gamma \|\mu - \widehat{\mu}\|^2 \right\} \right.$$
$$\left. + \sup_{\Sigma \succeq 0} \left\{ \alpha \sqrt{w^\top \Sigma w} + \gamma \operatorname{Tr}\big[-\Sigma + 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}} \big] \right\} \right\},$$

where the first equality follows from strong duality, which holds because $(\widehat{\mu}, \widehat{\Sigma})$ constitutes a Slater point for the primal convex program (13), and from a simple rearrangement. As $\gamma \geq 0$, the embedded quadratic maximization problem over μ is convex and can be solved analytically. Indeed, for $\gamma > 0$ we have $\sup_\mu \{-\mu^\top w - \gamma\|\mu - \widehat{\mu}\|^2\} = -\widehat{\mu}^\top w + \frac{\|w\|^2}{4\gamma}$. For $\gamma = 0$, on the other hand, the supremum over μ evaluates to 0 if $w = 0$ and to $+\infty$ otherwise. Thus, the formula on the right hand side of the above expression remains valid for $\gamma = 0$ if we interpret it as the limit when γ tends to 0 from above. Similarly, as $\gamma \geq 0$ and $\alpha > 0$, one may introduce an auxiliarly epigraphical variable t to reformulate the embedded maximization problem over Σ as the convex program

$$\begin{array}{ll} \sup\limits_{t, \Sigma} & \alpha t + \gamma \operatorname{Tr}\big[-\Sigma + 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}} \big] \\ \text{s.t.} & t \geq 0,\ \Sigma \succeq 0,\ t^2 - w^\top \Sigma w \leq 0, \end{array}$$

which manifestly satisfies Slater's condition. Suppose now temporarily that $\widehat{\Sigma} \succ 0$. By invoking strong duality and using the variable transformation $B \leftarrow (\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}$, the above optimization problem can be recast as

$$\begin{aligned} &\inf_{\lambda \geq 0} \sup_{t \geq 0, \Sigma \succeq 0} \ \alpha t - \lambda t^2 + \operatorname{Tr}\big[\Sigma(\lambda w w^\top - \gamma I)\big] + 2\gamma \operatorname{Tr}\big[\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big] \\ &= \inf_{\lambda \geq 0} \sup_{t \geq 0, B \succeq 0} \ \alpha t - \lambda t^2 + \operatorname{Tr}\big[B^2 \Delta_\lambda\big] + 2\gamma \operatorname{Tr}\big[B\big], \end{aligned} \tag{14}$$

where $\Delta_\lambda = \widehat{\Sigma}^{-\frac{1}{2}}(\lambda w w^\top - \gamma I)\widehat{\Sigma}^{-\frac{1}{2}}$ for any $\lambda \geq 0$. Note that Δ_λ is well-defined because $\widehat{\Sigma}$ is invertible. Note also that the inner maximization problem in (14) is separable with respect to t and B. Consider first the maximization problem over t. As $\alpha > 0$, its supremum evaluates to $\alpha^2/(4\lambda)$ and is attained at $t^\star = \alpha/(2\lambda)$ whenever $\lambda > 0$. Otherwise, if $\lambda = 0$, then its supremum evaluates to $+\infty$ for. From now on we may thus assume without loss of generality that the outer minimization over λ in (14) is subject to the strict constraint $\lambda > 0$. Consider now the maximization problem over B. The proof of [Nguyen et al., 2021a, Proposition 2.8] implies that if $\Delta_\lambda \not\prec 0$, then the supremum over B evaluates to $+\infty$. From now on we may thus assume without loss of generality that the outer minimization over λ in (14) is subject to the strict constraint $\gamma I - \lambda w w^\top \succ 0$, which is equivalent to $\lambda < \gamma\|w\|^{-2}$ and guarantees that $\Delta_\lambda \prec 0$. As $\lambda > 0$, this in turn implies that $B^\star = -\gamma \Delta_\lambda^{-1}$ is strictly positive definite and satisfies the first-order optimality condition $B\Delta_\lambda + \Delta_\lambda B + 2\gamma I = 0$. This condition can be interpreted as a continuous Lyapunov equation, and therefore its solution $B^\star$ is in fact unique; see, *e.g.*, [Hespanha, 2009, Theorem 12.5]. By making the implicit constraints on λ explicit and by eliminating the supremum operator by evaluating the objective function at $t^\star$ and $B^\star$, problem (14) can now be reformulated as

$$\inf_{0 < \lambda < \gamma\|w\|^{-2}} \frac{\alpha^2}{4\lambda} + \gamma^2 \operatorname{Tr}\big[\widehat{\Sigma}^{\frac{1}{2}}(\gamma I - \lambda w w^\top)^{-1} \widehat{\Sigma}^{\frac{1}{2}}\big]$$

$$= \inf_{0<\lambda<\gamma\|w\|^{-2}} \frac{\alpha^2}{4\lambda} + \gamma \operatorname{Tr}\left[\widehat{\Sigma}\right] + \frac{w^\top \widehat{\Sigma} w}{\lambda^{-1} - \|w\|^2/\gamma} = \gamma \operatorname{Tr}\left[\widehat{\Sigma}\right] + \frac{\alpha^2}{4}\frac{\|w\|^2}{\gamma} + \alpha\sqrt{w^\top \widehat{\Sigma} w}.$$

Here, the first equality exploits the Sherman-Morrison formula [Bernstein, 2009, Corollary 2.8.8] to rewrite the inverse matrix, and the second equality solves the resulting minimization problem analytically. Indeed, the infimum is attained in the interior of the feasible set at the unique solution $\lambda^\star$ of the first-order condition $\frac{1}{\lambda} = \frac{\|w\|^2}{\gamma} + \frac{2}{\alpha}\sqrt{w^\top \widehat{\Sigma} w}$. In summary, we have derived closed-form solutions for both subproblems over μ and Σ in the objective function of the problem dual to (13). Hence, the Gelbrich risk satisfies

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})} \mathcal{R}_{\mathbb{Q}}(-w^\top\xi) &= \inf_{\gamma\geq 0}\ -\widehat{\mu}^\top w + \alpha\sqrt{w^\top\widehat{\Sigma}w} + \gamma\rho^2 + \frac{1+\alpha^2}{4}\frac{\|w\|^2}{\gamma}\\ &= -\widehat{\mu}^\top w + \alpha\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+\alpha^2}\,\|w\|,\end{aligned}$$

where the second equality holds because the unique solution of the minimization problem in the first line is given by $\gamma^\star = (2\rho)^{-1}\sqrt{1+\alpha^2}\,\|w\|$. We have thus established (9) for $\widehat{\Sigma} \succ 0$.

To demonstrate that (9) remains valid for all $\widehat{\Sigma} \succeq 0$, we denote by $J(\widehat{\Sigma})$ the optimal value of problem (13) as a function of $\widehat{\Sigma}$. Thus, $J(\widehat{\Sigma})$ coincides with the Gelbrich risk. Applying Berge's maximum theorem [Berge, 1963, pp. 115–116] to problem (13), it is easy to show that $J(\widehat{\Sigma})$ is continuous on $\mathbb{S}^n_+$. Next, define $\bar{J}(\widehat{\Sigma}) = -\widehat{\mu}^\top w + \alpha\sqrt{w^\top\widehat{\Sigma}w} + \rho\sqrt{1+\alpha^2}\,\|w\|$ as the right hand side of (9), which is manifestly continuous on $\mathbb{S}^n_+$. From the first part of the proof we know that $J(\widehat{\Sigma}) = \bar{J}(\widehat{\Sigma})$ for all $\widehat{\Sigma} \succ 0$. As both $J(\widehat{\Sigma})$ and $\bar{J}(\widehat{\Sigma})$ are continuous on $\mathbb{S}^n_+$ and as any positive semidefinite matrix can be expressed as a limit of positive definite matrices, we thus have $J(\widehat{\Sigma}) = \bar{J}(\widehat{\Sigma})$ for all $\widehat{\Sigma} \in \mathbb{S}^n_+$. This proves (9) for all $\widehat{\Sigma} \succeq 0$.

Finally, if $\mathcal{S}$ is the structural ambiguity set generated by a Gaussian nominal distribution $\widehat{\mathbb{P}}$ with $\widehat{\Sigma} \succ 0$, then the Wasserstein risk equals the Gelbrich risk by Corollary 1. *Q.E.D.*

PROOF OF PROPOSITION 5: Fix any $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{S}^n_+$ and $w \in \mathbb{R}^n$, and select any symmetric probability distribution $\mathbb{Q} \in \mathcal{S}$, which exists by assumption. As $\mathcal{S}$ is closed under positive semidefinite affine pushforwards, we may assume without loss of generality that ξ has mean μ and covariance matrix Σ under $\mathbb{Q}$. We then find

$$\begin{aligned}0 = \mathcal{R}_{\mathbb{Q}}(0) &= \mathcal{R}_{\mathbb{Q}}\left(\tfrac{1}{2}\tfrac{w^\top(\xi-\mu)}{\sqrt{w^\top\Sigma w}} - \tfrac{1}{2}\tfrac{w^\top(\xi-\mu)}{\sqrt{w^\top\Sigma w}}\right)\\ &\leq \frac{1}{2}\mathcal{R}_{\mathbb{Q}}\left(\frac{w^\top(\xi-\mu)}{\sqrt{w^\top\Sigma w}}\right) + \frac{1}{2}\mathcal{R}_{\mathbb{Q}}\left(-\frac{w^\top(\xi-\mu)}{\sqrt{w^\top\Sigma w}}\right) = \mathcal{R}_{\mathbb{Q}}\left(-\frac{w^\top(\xi-\mu)}{\sqrt{w^\top\Sigma w}}\right) \leq \alpha,\end{aligned}$$

where the first equality and the first inequality follow from the positive homogeneity and convexity of the coherent risk measure $\mathcal{R}_{\mathbb{Q}}$, respectively. The last equality exploits the law-invariance of $\mathcal{R}_{\mathbb{Q}}$ and the symmetry of $\mathbb{Q}$, which implies that $w^\top(\xi-\mu)/\sqrt{w^\top\Sigma w}$ and $-w^\top(\xi-\mu)/\sqrt{w^\top\Sigma w}$ have the same distribution under $\mathbb{Q}$. Finally, the last inequality follows from the definition of α and the observation that $\mathbb{Q} \in \mathcal{C}(\mu,\Sigma)$. *Q.E.D.*

PROOF OF COROLLARY 3: Note that problem (10) has a unique minimizer for every $\rho > 0$ because its feasible set Ω is closed and its objective function is strictly convex and coercive on Ω. This minimizer coincides with the unique optimal solution $w^\star(\lambda)$ of

$$\min_{w\in\Omega} \; -\lambda\widehat{\mu}^\top w + \lambda\alpha\sqrt{w^\top\widehat{\Sigma}w} + \sqrt{1+\alpha^2}\|w\|,$$

where $\lambda = 1/\rho$. By Berge's maximum theorem [Berge, 1963, pp. 115–116], the function $w^\star(\lambda)$ is continuous on $[0,1]$, and therefore $w^\star(\lambda)$ converges to $w^\star(0)$ as λ tends to 0. As $\frac{1}{n}e \in \Omega$, however, one readily verifies that $w^\star(0) = \frac{1}{n}e$. As $\lambda = 1/\rho$, this reasoning shows that the unique minimizer of (10) converges to $\frac{1}{n}e$ as ρ tends to ∞. *Q.E.D.*

PROOF OF PROPOSITION 6: If $\rho = 0$, then all distributions in the Gelbrich ambiguity set, and in particular all maximizers that attain the Gelbrich risk, have mean $\widehat{\mu}$ and covariance matrix $\widehat{\Sigma}$. The claim then follows because, for $\rho = 0$, the formulas for $\mu^\star$ and $\Sigma^\star$ reduce to $\widehat{\mu}$ and $\widehat{\Sigma}$, respectively. Assume from now on that $\rho > 0$. As all assumptions of Theorem 5 are satisfied, we may proceed as in the proof of Theorem 5 to show that

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})} \mathcal{R}_{\mathbb{Q}}\left(-w^\top\xi\right) &= \begin{cases} \max \; -\mu^\top w + \alpha t \\ \text{s.t.} \;\; (\mu,\Sigma) \in \mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma}),\; t \geq 0,\; t^2 \leq w^\top\Sigma w \end{cases} \\ &= \min_{\gamma\geq 0,\lambda\geq 0} \; \sup_{\mu,\Sigma\succeq 0,t\geq 0} L(\mu,\Sigma,t,\gamma,\lambda),\end{aligned}$$

where the last equality follows from strong duality, which applies because the primal problem has a compact feasible set. The Lagrangian in the last expression is defined through

$$\begin{aligned}L(\mu,\Sigma,t,\gamma,\lambda) = {}& \gamma\big(\rho^2 - \|\mu-\widehat{\mu}\|^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\big) - \mu^\top w + \alpha t - \lambda t^2 \\ & + \operatorname{Tr}\big[\Sigma(\lambda w w^\top - \gamma I)\big] + 2\gamma \operatorname{Tr}\big[\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big].\end{aligned}$$

From the proof of Theorem 5 we know that the dual problem is uniquely solved by

$$\gamma^\star = (2\rho)^{-1}\sqrt{1+\alpha^2}\,\|w\| \quad \text{and} \quad \frac{1}{\lambda^\star} = \frac{\|w\|^2}{\gamma^\star} + \frac{2}{\alpha}\sqrt{w^\top\widehat{\Sigma}w}.$$

In addition, [Ben-Tal and Nemirovski, 2001, Theorem D.4.1] implies that any primal maximizer $(\mu^\star, \Sigma^\star, t^\star)$ must also be a maximizer of

$$\max_{\mu,\Sigma\succeq 0,t\geq 0} \; L(\mu,\Sigma,t,\gamma^\star,\lambda^\star). \tag{15}$$

As the Lagrangian is additively separable in μ, Σ and t, the maximizers $\mu^\star$, $\Sigma^\star$ and $t^\star$ can be determined separately. Indeed, one readily verifies that $\mu^\star$ must be a maximizer of the problem $\max_\mu\{-\mu^\top w - \gamma^\star\|\mu-\widehat{\mu}\|^2\}$, which is uniquely solved by $\mu^\star = \widehat{\mu} - w/(2\gamma^\star)$. Similarly, $\Sigma^\star$ must be a maximizer of the problem

$$\max_{\Sigma\succeq 0} \operatorname{Tr}\big[\Sigma(\lambda w w^\top - \gamma I)\big] + 2\gamma\operatorname{Tr}\big[\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\big] = \max_{B\succeq 0} \operatorname{Tr}\big[B^2\Delta\big] + 2\gamma\operatorname{Tr}\big[B\big], \tag{16}$$

where the equality exploits the substitution $B \leftarrow (\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}}$ and the definition $\Delta = \widehat{\Sigma}^{-\frac{1}{2}}(\lambda^\star w w^\top - \gamma^\star I)\widehat{\Sigma}^{-\frac{1}{2}}$. As in the proof of Theorem 5, one can show that the second maximization problem in (16) is uniquely solved by $B^\star = \widehat{\Sigma}^{\frac{1}{2}}(I - \frac{\lambda^\star}{\gamma^\star}ww^\top)^{-1}\widehat{\Sigma}^{\frac{1}{2}}$, which implies that the first maximization problem in (16) is uniquely solved by

$$\begin{aligned}\Sigma^\star &= \left(I - \frac{\lambda^\star}{\gamma^\star} w w^\top\right)^{-1} \widehat{\Sigma} \left(I - \frac{\lambda^\star}{\gamma^\star} w w^\top\right)^{-1} \\ &= \left(I + \frac{\lambda^\star}{\gamma^\star - \lambda^\star \|w\|^2} w w^\top\right) \widehat{\Sigma} \left(I + \frac{\lambda^\star}{\gamma^\star - \lambda^\star \|w\|^2} w w^\top\right).\end{aligned}$$

Here, the last equality exploits the Sherman-Morrison-Woodbury identity [Bernstein, 2009, Corollary 2.8.8]. Finally, $t^\star$ must be a maximizer of $\max_{t \geq 0} \alpha t - \lambda^\star t^2$, which is uniquely solved by $t^\star = \alpha/(2\lambda^\star)$. The claim now follows from the formulas for the dual variables $\gamma^\star$ and $\lambda^\star$ substituted into the formulas for $\mu^\star$, $\Sigma^\star$ and $t^\star$. *Q.E.D.*

PROOF OF PROPOSITION 7: Proposition 1 by Yu et al. [2009] provides an analytical formula for the Chebyshev risk of a portfolio loss function with respect to VaR. As the Chebyshev risk coincides with the Gelbrich risk for $\rho = 0$, the standard risk coefficient is readily found by comparison with (9). *Q.E.D.*

PROOF OF PROPOSITION 8: This follows from [Yu et al., 2009, Proposition 2], similar to Proposition 7. *Q.E.D.*

PROOF OF PROPOSITION 9: Equation (9) for $\rho = 0$ reveals that $\alpha = \beta$. *Q.E.D.*

PROOF OF PROPOSITION 10: This follows from [Li, 2018, Theorem 2], similar to Proposition 7. *Q.E.D.*

PROOF OF PROPOSITION 11: This follows from [Li, 2018, Theorem 3], similar to Proposition 7. *Q.E.D.*

PROOF OF PROPOSITION 12: This follows from [Cai et al., 2020, Theorem 3.10], similar to Proposition 7. The same theorem reveals that the Chebyshev risk does not change if h is replaced with its right-continuous modification. In light of (7), this invariance remains valid if uncertainty is modeled by a Gelbrich ambiguity set. Therefore, we may always assume without loss of generality that h is right-continuous. *Q.E.D.*

REFERENCES

C. Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM, 2001.

A. Ben-Tal, D. Den Hertog, and J. P. Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming*, 149(1-2):265–299, 2015.

C. Berge. *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*. Courier Corporation, 1963.

D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, 2009.

D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

M. J. Best and R. R. Grauer. On the sensitivity of mean-variance-efficient portfolios to changes in asset means: Some analytical and computational results. *The Review of Financial Studies*, 4(2):315–342, 1991.

R. Bhatia, T. Jain, and Y. Lim. Strong convexity of sandwiched entropies and related optimization problems. *Reviews in Mathematical Physics*, 30(9):1850014, 2018.

R. Bhatia, T. Jain, and Y. Lim. On the Bures-Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

J. Blanchet, L. Chen, and X. Y. Zhou. Distributionally robust mean-variance portfolio selection with Wasserstein distances. *Management Science*, 2020. Forthcoming.

C. Borell. Convexity of measures in certain convex cones in vector space σ-algebras. *Mathematica Scandinavica*, 53 (1):125–144, 1983.

J. Cai, J. Li, and T. Mao. Distributionally robust optimization under distorted expectations. *SSRN preprint 3566708*, 2020.

S. Cambanis, S. Huang, and G. Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981.

L. Chen, S. He, and S. Zhang. Tight bounds for some risk measures, with applications to robust portfolio selection. *Operations Research*, 59(4):847–865, 2011a.

S. Chen, J. Dick, and A. B. Owen. Consistency of Markov chain quasi-Monte Carlo on continuous state spaces. *The Annals of Statistics*, 39(2):673–701, 2011b.

W. Chen, M. Sim, J. Sun, and C. Teo. From CVaR to uncertainty set: Implications in joint chance-constrained optimization. *Operations Research*, 58(2):470–485, 2010.

V. K. Chopra and W. T. Ziemba. The effect of errors in means, variances and covariances on optimal portfolio choice. *Journal of Portfolio Management*, 19(2):6–11, 1993.

J. A. Cuesta-Albertos, L. Rüschendorf, and A. Tuero-Diaz. Optimal coupling of multivariate distributions and stochastic processes. *Journal of Multivariate Analysis*, 46(2):335–361, 1993.

J. A. Cuesta-Albertos, C. Matrán-Bea, and A. Tuero-Diaz. On lower bounds for the L^2-Wasserstein metric in a Hilbert space. *Journal of Theoretical Probability*, 9(2):263–283, 1996.

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

V. DeMiguel, L. Garlappi, F. J. Nogales, and R. Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812, 2009.

S. Dharmadhikari and K. Joag-Dev. *Unimodality, Convexity, and Applications*. Elsevier, 1988.

X. V. Doan, X. Li, and K. Natarajan. Robustness to dependency in portfolio optimization using overlapping marginals. *Operations Research*, 63(6):1468–1488, 2015.

D. Dowson and B. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

D. Duffie and J. Pan. An overview of value at risk. *The Journal of Derivatives*, 4(3):7–49, 1997.

L. El Ghaoui, M. Oks, and F. Oustry. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, 51(4):543–556, 2003.

D. Ellsberg. Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics*, 75(4):643–669, 1961.

H. Föllmer and A. Schied. Convex and coherent risk measures. In R. Cont, editor, *Encyclopedia of Quantitative Finance*, pages 355–363. Wiley, 2008a.

H. Föllmer and A. Schied. *Stochastic Finance. An Introduction in Discrete Time*. de Gruyter, 2008b.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

R. Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *arXiv preprint arXiv:2009.04382*, 2020.

M. Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.

C. Givens and R. Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58 (4):902–917, 2010.

G. Hanasusanto, D. Kuhn, S. W. Wallace, and S. Zymler. Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Mathematical Programming*, 152(1-2):1–32, 2015.

D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

J. P. Hespanha. *Linear Systems Theory*. Princeton University Press, 2009.

D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012.

R. Jagannathan and T. Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4):1651–1683, 2003.

P. Jorion. *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, 1996.

J. M. Keynes. *A Treatise on Probability*. Macmillan & Co., 1921.

F. H. Knight. *Risk, Uncertainty and Profit*. Houghton Mifflin, 1921.

M. Knott and C. S. Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

D. Kuhn, P. Mohajerin Esfahani, V. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *INFORMS TutORials in Operations Research*, pages 130–169, 2019.

S. Kusuoka. On law invariant coherent risk measures. In S. Kusuoka and T. Maruyama, editors, *Advances in Mathematical Economics*, pages 83–95. Springer, 2001.

H. Lam and C. Mottet. Tail analysis without parametric models: A worst-case perspective. *Operations Research*, 65 (6):1696–1711, 2017.

O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, 2003.

J. Y.-M. Li. Closed-form solutions for worst-case law invariant risk measures with application to robust portfolio optimization. *Operations Research*, 66(6):1533–1541, 2018.

F. Liese and I. Vajda. *Convex Statistical Distances*. Teubner, 1987.

M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284(1):193–228, 1998.

J. Longerstaey and M. Spencer. *Riskmetrics ^{TM}—-Technical Document*. Morgan Guaranty Trust Company of New York, 1996.

D. G. Luenberger. *Investment Science*. Oxford University Press, 1997.

L. Malagò, L. Montrucchio, and G. Pistone. Wasserstein Riemannian geometry of Gaussian densities. *Information Geometry*, 1(2):137–179, 2018.

H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

R. C. Merton. Lifetime portfolio selection under uncertainty: The continuous-time case. *The Review of Economics and Statistics*, 51(3):247–257, 1969.

R. O. Michaud. The Markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1): 31–42, 1989.

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

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

K. Natarajan, M. Sim, and J. Uichanco. Tractable robust expected utility and risk models for portfolio optimization. *Mathematical Finance*, 20(4):695–731, 2010.

K. Natarajan, M. Sim, and J. Uichanco. Asymmetry and ambiguity in newsvendor models. *Management Science*, 64 (7):3146–3167, 2018.

A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.

V. A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *Operations Research*, 2021a. Forthcoming.

V. A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization. *Mathematics of Operations Research*, 2021b. Forthcoming.

I. Olkin and F. Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

S. Pesenti, Q. Wang, and R. Wang. Optimizing distortion riskmetrics with distributional uncertainty. *arXiv preprint arXiv:2011.04889*, 2020.

G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

G. C. Pflug, A. Pichler, and D. Wozabal. The $1/N$ investment strategy is optimal under high model ambiguity. *Journal of Banking & Finance*, 36(2):410–417, 2012.

A. Pichler. Evaluations of risk measures for different probability measures. *SIAM Journal on Optimization*, 23(1): 530–551, 2013.

I. Popescu. A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operations Research*, 30(3):632–657, 2005.

K. Postek, D. den Hertog, and B. Melenberg. Computationally tractable counterparts of distributionally robust constraints on risk measures. *SIAM Review*, 58(4):603–650, 2016.

F. Riesz and B. Sz.-Nagy. *Functional Analysis*. Dover Publications, 1990.

R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–41, 2000.

R. T. Rockafellar, S. P. Uryasev, and M. Zabarankin. Deviation measures in risk analysis and optimization. *SSRN preprint 365640*, 2002.

M. Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

A. Roy. Safety first and the holding of assets. *Econometrica*, 20(3):431–449, 1952.

N. Rujeerapaiboon, D. Kuhn, and W. Wiesemann. Robust growth-optimal portfolios. *Management Science*, 62(7): 2090–2109, 2016.

N. Rujeerapaiboon, D. Kuhn, and W. Wiesemann. Chebyshev inequalities for products of random variables. *Mathematics of Operations Research*, 43(3):887–918, 2018.

B. A. Schmitt. Perturbation bounds for matrix square roots and Pythagorean sums. *Linear Algebra and its Applications*, 174:215–227, 1992.

S. Shafieezadeh-Abadeh, V. A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani. Wasserstein distributionally robust Kalman filtering. In *Advances in Neural Information Processing Systems*, pages 8483–8492, 2018.

A. Shapiro. On Kusuoka representation of law invariant risk measures. *Mathematics of Operations Research*, 38(1): 142–152, 2013.

B. Taşkesen, S. Shafieezadeh-Abadeh, and D. Kuhn. Semi-discrete optimal transport: Hardness, regularization and numerical solution. *arXiv preprint arXiv:2103.06263*, 2021.

B. P. Van Parys, P. J. Goulart, and D. Kuhn. Generalized Gauss inequalities via semidefinite programming. *Mathematical Programming*, 156(1-2):271–302, 2016.

C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

M. J. Wainwright. *High-dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6): 1358–1376, 2014.

D. Wozabal. Robustifying convex risk measures for linear portfolios: A nonparametric approach. *Operations Research*, 62(6):1302–1315, 2014.

M. E. Yaari. The dual theory of choice under risk. *Econometrica*, 55(1):95–115, 1987.

Y.-L. Yu, Y. Li, D. Schuurmans, and C. Szepesvári. A general projection property for distribution families. In *Advances in Neural Information Processing Systems*, pages 2232–2240, 2009.

J. Zhen, D. Kuhn, and W. Wiesemann. Mathematical foundations of robust and distributionally robust optimization. *arXiv preprint arXiv:2105.00760*, 2021.

S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198, 2013a.

S. Zymler, D. Kuhn, and B. Rustem. Worst-case value at risk of nonlinear portfolios. *Management Science*, 59(1): 172–188, 2013b.

APPENDIX A: PROOFS OF SECTION 3

We give a short proof of Theorem 1 in order to keep this paper self-contained.

PROOF OF THEOREM 1: By the definition of the 2-Wasserstein distance, we have

$$\mathbb{W}^2(\mathbb{Q}_1,\mathbb{Q}_2) = \min_{\pi\in\Pi(\mathbb{Q}_1,\mathbb{Q}_2)} \int_{\mathbb{R}^n\times\mathbb{R}^n} \|\xi_1-\xi_2\|^2\pi(\mathrm{d}\xi_1,\mathrm{d}\xi_2)$$

$$= \begin{cases} \min \|\mu_1-\mu_2\|^2 + \mathrm{Tr}\left[\Sigma_1+\Sigma_2-2C\right] \\ \mathrm{s.t.}\ \pi\in\Pi(\mathbb{Q}_1,\mathbb{Q}_2),\ C\in\mathbb{R}^{n\times n} \\ \quad \displaystyle\int_{\mathbb{R}^n\times\mathbb{R}^n} \begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix}\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix}^\top \pi(\mathrm{d}\xi_1,\mathrm{d}\xi_2) = \begin{bmatrix}\Sigma_1 & C\\ C^\top & \Sigma_2\end{bmatrix},\quad \begin{bmatrix}\Sigma_1 & C\\ C^\top & \Sigma_2\end{bmatrix}\succeq 0, \end{cases}$$

where the second equality follows from the observations that C is uniquely determined by π (thanks to the equality constraint in the last line) and that the conic constraint is redundant (because the second-order momement matrix of π is always positive semidefinite). Relaxing the last optimization problem by removing all constraints that involve the original decision variable π, which in turn allows us to remove π itself, we find

$$\mathbb{W}^2(\mathbb{Q}_1,\mathbb{Q}_2) \geq \begin{cases} \displaystyle\min_{C\in\mathbb{R}^{n\times n}} & \|\mu_1-\mu_2\|^2 + \mathrm{Tr}\left[\Sigma_1+\Sigma_2-2C\right] \\ \mathrm{s.t.} & \begin{bmatrix}\Sigma_1 & C\\ C^\top & \Sigma_2\end{bmatrix}\succeq 0. \end{cases} \tag{17}$$

By [Malagò et al., 2018, Proposition 2], the above semidefinite program can be solved analytically, and its optimal value is given by $\mathbb{G}^2((\mu_1,\Sigma_1),(\mu_2,\Sigma_2))$. We emphasize that the same analytical formula has also been reported in [Cuesta-Albertos et al., 1996, Dowson and Landau, 1982, Givens and Shortt, 1984, Knott and Smith, 1984, Olkin and Pukelsheim, 1982]. The claim then follows by taking square roots on both sides of (17). *Q.E.D.*

Our proof of Theorem 2 relies on the following preparatory lemma.

LEMMA 1—2-Wasserstein distances of perfectly correlated distributions [Cuesta-Albertos et al., 1993, Theorem 2.13]: *If* $\mathbb{Q}_1\in\mathcal{M}_2$ *has mean 0 and covariance matrix* $\Sigma_1\in\mathbb{S}^n_+$, *and if* $\mathbb{Q}_2=\mathbb{Q}_1\circ g^{-1}$ *is the pushforward distribution of* $\mathbb{Q}_1$ *under the positive semidefinite linear transformation* $g(\xi)=A\xi$ *with* $A\in\mathbb{S}^n_+$, *then*

$$\mathbb{W}(\mathbb{Q}_1,\mathbb{Q}_2) = \sqrt{\mathbb{E}_{\mathbb{Q}_1}\left[\|\xi-A\xi\|^2\right]} = \sqrt{\mathrm{Tr}\left[\Sigma_1+A\Sigma_1A-2\Sigma_1^{\frac{1}{2}}A\Sigma_1^{\frac{1}{2}}\right]}. \tag{18}$$

Lemma 1 implies that if ξ follows a distribution $\mathbb{Q}_1$ with vanishing mean and if $\mathbb{Q}_2$ is defined as the distribution of $A\xi$ for some $A\in\mathbb{S}^n_+$, then the 2-Wasserstein distance between the perfectly correlated distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ is given by (18). We can now use this (nontrivial) result to prove Theorem 2.

PROOF OF THEOREM 2: By assumption, we have $\mathbb{Q}_2 = \mathbb{Q}_1 \circ f^{-1}$, where $f(\xi) = A\xi + b$ is an affine transformation with parameters $A \in \mathbb{S}_+^n$ and $b \in \mathbb{R}^n$. In the following, it will be more convenient to re-express this affine transformation as $f(\xi) = A(\xi - \mu_1) + b'$, which involves the auxiliary parameter $b' = A\mu_1 + b$. Thus, we have

$$\mu_2 = \mathbb{E}_{\mathbb{Q}_2}[\xi] = \mathbb{E}_{\mathbb{Q}_1}[f(\xi)] = b',$$

where the three equalities follow from the definition of μ_2, the integration formula for pushforward distributions and the definitions of μ_1 and f, respectively. Similarly, we find

$$\Sigma_2 = \mathbb{E}_{\mathbb{Q}_2}[(\xi - \mu_2)(\xi - \mu_2)^\top] = \mathbb{E}_{\mathbb{Q}_1}[(f(\xi) - b')(f(\xi) - b')^\top] = A\Sigma_1 A,$$

where the second equality exploits our earlier insight that $\mu_2 = b'$. Multiplying the above expression from both sides with $\Sigma_1^{\frac{1}{2}}$ yields the quadratic equation $\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}} = (\Sigma_1^{\frac{1}{2}} A \Sigma_1^{\frac{1}{2}})^2$. As $\Sigma_1 \succ 0$ by assumption, this equation is uniquely solved by $A = \Sigma_1^{-\frac{1}{2}}(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}}$. This confirms that the affine function f is uniquely determined by the first- and second-order moments of $\mathbb{Q}_1$ and $\mathbb{Q}_2$ and that its parameters are given by (3).

Next, we define two distributions $\overline{\mathbb{Q}}_1, \overline{\mathbb{Q}}_2 \in \mathcal{M}_2$ through the relations $\overline{\mathbb{Q}}_1[\xi \in B] = \mathbb{Q}_1[(\xi + \mu_1) \in B]$ and $\overline{\mathbb{Q}}_2[\xi \in B] = \mathbb{Q}_2[(\xi + \mu_2) \in B]$ for all Borel sets $B \in \mathcal{B}(\mathbb{R}^n)$. Thus, $\overline{\mathbb{Q}}_1$ and $\overline{\mathbb{Q}}_2$ are obtained by shifting $\mathbb{Q}_1$ and $\mathbb{Q}_2$ so that their mean vectors vanish. By construction, we then have $\overline{\mathbb{Q}}_2 = \overline{\mathbb{Q}}_1 \circ g^{-1}$ where $g(\xi) = A\xi$. By the definition of the 2-Wasserstein distance and the shifted distributions $\overline{\mathbb{Q}}_1$ and $\overline{\mathbb{Q}}_2$, we further have

$$\mathbb{W}^2(\mathbb{Q}_1, \mathbb{Q}_2) = \|\mu_1 - \mu_2\|^2 + \mathbb{W}^2(\overline{\mathbb{Q}}_1, \overline{\mathbb{Q}}_2). \tag{19}$$

Finally, as $\overline{\mathbb{Q}}_2 = \overline{\mathbb{Q}}_1 \circ g^{-1}$, we may use Lemma 1 to conclude that

$$\begin{aligned}
\mathbb{W}^2(\overline{\mathbb{Q}}_1, \overline{\mathbb{Q}}_2) &= \mathrm{Tr}\left[\Sigma_1 + A\Sigma_1 A - 2\Sigma_1^{\frac{1}{2}} A \Sigma_1^{\frac{1}{2}}\right] \\
&= \mathrm{Tr}\left[\Sigma_1 + \left(\Sigma_1^{-\frac{1}{2}}(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}}\right)\Sigma_1\left(\Sigma_1^{-\frac{1}{2}}(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}}\right)\right] \\
&\quad - 2\,\mathrm{Tr}\left[\Sigma_1^{\frac{1}{2}}\left(\Sigma_1^{-\frac{1}{2}}(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}}\right)\Sigma_1^{\frac{1}{2}}\right] \\
&= \mathrm{Tr}\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right],
\end{aligned}$$

where the second equality uses the expression for A in (3). The claim then follows by substituting the above expression into (19) and taking square roots on both sides. *Q.E.D.*

The proof of Theorem 4 requires two preparatory lemmas.

LEMMA 2—Concentration inequality for the sample mean: *Suppose that $\mathbb{P}$ is sub-Gaussian with variance proxy σ^2, mean μ and covariance matrix $\Sigma \succ 0$, and denote by $\widehat{\mu}_N$ the sample mean corresponding to N independent points sampled from $\mathbb{P}$ as in (5). Then, there exists a positive constant $C \leq \sigma/\sqrt{\|\Sigma\|}$ such that $\mathbb{P}^N[\|\widehat{\mu}_N - \mu\| \leq \rho_\mu(\eta)] \geq 1 - \eta$ for any significance level $\eta \in (0, 1]$, where*

$$\rho_\mu(\eta) = C\left(\sqrt{\frac{\mathrm{Tr}\left[\Sigma\right]}{N}} + \sqrt{\frac{2\left\|\Sigma\right\|\log\left(1/\eta\right)}{N}}\right).$$

Lemma 2 establishes a variant of the Hanson-Wright inequality [Hanson and Wright, 1971] for sub-Gaussian distributions.

Proof of Lemma 2: Set $\tilde{\xi}_i = \Sigma^{-\frac{1}{2}}(\xi_i - \mu)$, which is well-defined because $\Sigma \succ 0$, and note that $\tilde{\xi}_i$ is isotropic in the sense that $\mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i] = 0$ and $\mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i \tilde{\xi}_i^\top] = I$ for all $i = 1, \ldots, N$. Then, the probability distribution of $\tilde{\xi}_i$ is also sub-Gaussian with the variance proxy C^2 satisfying $C^2 \leq \sigma^2 / \|\Sigma\|$ because we have

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[\exp\left(z^\top(\tilde{\xi}_i - \mathbb{E}_{\mathbb{P}}[\tilde{\xi}_i])\right)\right] &= \mathbb{E}_{\mathbb{P}}\left[\exp\left(z^\top \Sigma^{-\frac{1}{2}}(\xi_i - \mathbb{E}_{\mathbb{P}}[\xi_i])\right)\right] \\ &\leq \exp\left(\tfrac{1}{2}\|\Sigma^{-\frac{1}{2}} z\|^2 \sigma^2\right) \leq \exp\left(\tfrac{1}{2}\|z\|^2 \sigma^2 / \|\Sigma\|\right)\end{aligned}$$

for all $z \in \mathbb{R}^n$. By the definition of the sample mean, we then find

$$\widehat{\mu}_N - \mu = \frac{1}{N}\sum_{i=1}^N \xi_i - \mu = \Sigma^{\frac{1}{2}}\left(\frac{1}{N}\sum_{i=1}^N \tilde{\xi}_i\right).$$

Note that the random vector $N^{-1}\sum_{i=1}^N \tilde{\xi}_i$ has zero mean and covariance matrix $N^{-1}I$, and thus it concentrates around 0 for large N. In addition, as $\tilde{\xi}_1, \ldots, \tilde{\xi}_N$ are mutually independent, one easily verifies that $N^{-1}\sum_{i=1}^N \tilde{\xi}_i$ is also sub-Gaussian with variance proxy C^2/N. Therefore, [Hsu et al., 2012, Theorem 2.1] guarantees that

$$\mathbb{P}^N\left[\left\|\Sigma^{\frac{1}{2}}\left(\frac{1}{N}\sum_{i=1}^N \tilde{\xi}_i\right)\right\|^2 \leq \frac{C^2}{N}\left(\mathrm{Tr}\left[\Sigma\right] + 2\sqrt{\mathrm{Tr}\left[\Sigma^2\right]\log(1/\eta)} + 2\|\Sigma\|\log(1/\eta)\right)\right] \\ \geq 1 - \eta.$$

The elementary inequality $\mathrm{Tr}\left[\Sigma^2\right] \leq \|\Sigma\|\,\mathrm{Tr}\left[\Sigma\right]$ for any $\Sigma \succeq 0$ further implies that

$$\mathrm{Tr}\left[\Sigma\right] + 2\sqrt{\mathrm{Tr}\left[\Sigma^2\right]\log(1/\eta)} + 2\|\Sigma\|\log(1/\eta) \leq \left(\sqrt{\mathrm{Tr}\left[\Sigma\right]} + \sqrt{2\|\Sigma\|\log(1/\eta)}\right)^2.$$

Combining this inequality with the above concentration bound yields

$$\begin{aligned}&\mathbb{P}^N\left[\|\widehat{\mu}_N - \mu\| \leq \rho_\mu(\eta)\right] \\ &= \mathbb{P}^N\left[\left\|\Sigma^{\frac{1}{2}}\left(\frac{1}{N}\sum_{i=1}^N \tilde{\xi}_i\right)\right\| \leq C\left(\sqrt{\frac{\mathrm{Tr}\left[\Sigma\right]}{N}} + \sqrt{\frac{2\|\Sigma\|\log(1/\eta)}{N}}\right)\right] \\ &\geq \mathbb{P}^N\left[\left\|\Sigma^{\frac{1}{2}}\left(\frac{1}{N}\sum_{i=1}^N \tilde{\xi}_i\right)\right\|^2 \leq \frac{C^2}{N}\left(\mathrm{Tr}\left[\Sigma\right] + 2\sqrt{\mathrm{Tr}\left[\Sigma^2\right]\log(1/\eta)} + 2\|\Sigma\|\log(1/\eta)\right)\right] \\ &\geq 1 - \eta,\end{aligned}$$

and thus the claim follows. *Q.E.D.*

LEMMA 3—Concentration inequality for the sample second moment matrix: *Suppose that $\mathbb{P}$ is sub-Gaussian with variance proxy σ^2 and second moment matrix M, and denote by $\widehat{M}_N$ the sample second moment matrix corresponding to N independent points sampled from $\mathbb{P}$ as in* (5)*. Then, there are universal constants $C_1 > 0$, $C_2 \geq 1$, $C_3 > 0$ such that $\mathbb{P}^N[\|\widehat{M}_N - M\| \leq \rho_M(\eta)] \geq 1-\eta$ for any significance level $\eta \in (0,1]$, where*

$$\rho_M(\eta) = \sigma^2 C_1 \left(\sqrt{\frac{n}{N}} + \frac{n}{N}\right) + \sigma^2 \left(\sqrt{\frac{\log(C_2/\eta)}{C_3 N}} + \frac{\log(C_2/\eta)}{C_3 N}\right).$$

PROOF OF LEMMA 3: By [Wainwright, 2019, Theorem 6.5], for any $\delta > 0$ there exist universal constants $C_1 > 0$, $C_2 \geq 1$, $C_3 > 0$ with

$$\mathbb{P}^N\left[\frac{\|\widehat{M}_N - M\|}{\sigma^2} > C_1 \left(\sqrt{\frac{n}{N}} + \frac{n}{N}\right) + \delta\right] \leq C_2 \exp\left(-C_3 N \min\{\delta, \delta^2\}\right),$$

which implies that

$$\mathbb{P}^N\left[\|\widehat{M}_N - M\| \leq \sigma^2 C_1 \left(\sqrt{\frac{n}{N}} + \frac{n}{N}\right) + \sigma^2 \max\left\{\sqrt{\frac{\log(C_2/\eta)}{C_3 N}}, \frac{\log(C_2/\eta)}{C_3 N}\right\}\right] \geq 1-\eta.$$

The claim then follows from the inequality $\max\{a,b\} \leq a+b$ for all $a,b \geq 0$. *Q.E.D.*

We are now armed to prove Theorem 4.

PROOF OF THEOREM 4: The Gelbrich distance satisfies

$$\begin{aligned}
\mathbb{G}\big((\widehat{\mu}_N, \widehat{\Sigma}_N), (\mu, \Sigma)\big) &\leq \sqrt{\|\widehat{\mu}_N - \mu\|^2 + \|\widehat{\Sigma}_N^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}\|_F^2} \\
&\leq \|\widehat{\mu}_N - \mu\| + \|\widehat{\Sigma}_N^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}\|_F \\
&\leq \|\widehat{\mu}_N - \mu\| + \frac{1}{\lambda_{\min}(\widehat{\Sigma}_N) + \lambda_{\min}(\Sigma)} \|\widehat{\Sigma}_N - \Sigma\|_F \\
&\leq \|\widehat{\mu}_N - \mu\| + \frac{1}{\lambda_{\min}(\Sigma)} \|\widehat{M}_N - M + \mu\mu^\top - \widehat{\mu}_N\widehat{\mu}_N^\top\|_F, \\
&\leq \|\widehat{\mu}_N - \mu\| + \frac{1}{\lambda_{\min}(\Sigma)} \|\widehat{\mu}_N\widehat{\mu}_N^\top - \mu\mu^\top\|_F + \frac{\sqrt{n}}{\lambda_{\min}(\Sigma)} \|\widehat{M}_N - M\|,
\end{aligned}$$

where the first inequality follows from [Bhatia et al., 2019, Theorem 1], the second inequality holds because $\sqrt{a^2+b^2} \leq a+b$ for all $a,b \geq 0$, and the third inequality uses [Schmitt, 1992, Equation (1.2)]. The last inequality exploits the triangle inequality and the observation that $\|A\|_F \leq \sqrt{n}\|A\|$ for all $A \in \mathbb{S}^n$. Note that all divisions by $\lambda_{\min}(\Sigma)$ are well-defined because $\Sigma \succ 0$ by assumption. An elementary calculation further shows that

$$\begin{aligned}
\|\widehat{\mu}_N\widehat{\mu}_N^\top - \mu\mu^\top\|_F &= \|\widehat{\mu}_N(\widehat{\mu}_N - \mu)^\top + (\widehat{\mu}_N - \mu)\mu^\top\|_F \\
&\leq \|\widehat{\mu}_N(\widehat{\mu}_N - \mu)^\top\|_F + \|(\widehat{\mu}_N - \mu)\mu^\top\|_F
\end{aligned}$$

$$\begin{aligned}
&\leq \|\widehat{\mu}_N\| \cdot \|\widehat{\mu}_N - \mu\| + \|\widehat{\mu}_N - \mu\| \cdot \|\mu\| \\
&= \|\widehat{\mu}_N - \mu\| \left(\|\widehat{\mu}_N\| + \|\mu\|\right) \\
&\leq 2\|\mu\| \cdot \|\widehat{\mu}_N - \mu\| + \|\widehat{\mu}_N - \mu\|^2.
\end{aligned}$$

Next, set $K = 1 + 2\|\mu\|/\lambda_{\min}(\Sigma)$ and introduce an auxiliary quadratic function $f_\mu : \mathbb{R}_+ \to \mathbb{R}_+$ defined through $f_\mu(x) = Kx + x^2/\lambda_{\min}(\Sigma)$ for all $x \in \mathbb{R}_+$. Similarly, introduce an auxiliary linear function $f_M : \mathbb{R}_+ \to \mathbb{R}_+$ defined through $f_M(x) = \sqrt{n}x/\lambda_{\min}(\Sigma)$ for all $x \in \mathbb{R}_+$. Using this notation, the above estimates imply that

$$\mathbb{G}\big((\widehat{\mu}_N, \widehat{\Sigma}_N), (\mu, \Sigma)\big) \leq f_\mu\big(\|\widehat{\mu}_N - \mu\|\big) + f_M\big(\|\widehat{M}_N - M\|\big).$$

For any $\eta_\mu, \eta_M \in (0, 1]$ and for any $\rho \geq f_\mu(\rho_\mu(\eta_\mu)) + f_M(\rho_M(\eta_M))$ we thus have

$$\begin{aligned}
&\mathbb{P}^N \left[\mathbb{G}\big((\widehat{\mu}, \widehat{\Sigma}), (\mu, \Sigma)\big) \leq \rho\right] \\
&\geq \mathbb{P}^N \left[f_\mu\big(\|\widehat{\mu}_N - \mu\|\big) + f_M\big(\|\widehat{M}_N - M\|\big) \leq \rho\right] \\
&\geq \mathbb{P}^N \left[f_\mu\big(\|\widehat{\mu}_N - \mu\|\big) \leq f_\mu(\rho_\mu(\eta_\mu)) \ \wedge \ f_M\big(\|\widehat{M}_N - M\|\big) \leq f_M(\rho_M(\eta_M))\right] \qquad (20) \\
&= \mathbb{P}^N \left[\|\widehat{\mu}_N - \mu\| \leq \rho_\mu(\eta_\mu) \ \wedge \ \|\widehat{M}_N - M\| \leq \rho_M(\eta_M)\right] \geq 1 - \eta_\mu - \eta_M,
\end{aligned}$$

where the equality holds because f_μ and f_M are strictly increasing on their domains, and the last inequality follows from Lemmas 2 and 3 and the reverse union bound. Next, the definition of $\rho_\mu(\eta_\mu)$ implies that

$$\begin{aligned}
f_\mu(\rho_\mu(\eta_\mu)) &= CK \left(\sqrt{\frac{\mathrm{Tr}\left[\Sigma\right]}{N}} + \sqrt{\frac{2\|\Sigma\| \log\left(1/\eta_\mu\right)}{N}}\right) \\
&\qquad + C^2 \left(\sqrt{\frac{\mathrm{Tr}\left[\Sigma\right]}{\lambda_{\min}(\Sigma)N}} + \sqrt{\frac{2\|\Sigma\| \log\left(1/\eta_\mu\right)}{\lambda_{\min}(\Sigma)N}}\right)^2 \\
&\leq \frac{CK\sqrt{\mathrm{Tr}\left[\Sigma\right]}\lambda_{\min}(\Sigma) + CK\lambda_{\min}(\Sigma) + 2C^2\,\mathrm{Tr}\left[\Sigma\right]}{\lambda_{\min}(\Sigma)\sqrt{N}} \\
&\qquad + \frac{\left(2CK\|\Sigma\|\lambda_{\min}(\Sigma) + 4C^2\|\Sigma\|\right)\log\left(1/\eta_\mu\right)}{\lambda_{\min}(\Sigma)\sqrt{N}},
\end{aligned}$$

where the inequality holds because $(a + b)^2 \leq 2a^2 + 2b^2$ for all $a, b \geq 0$, while $1/N \leq 1/\sqrt{N}$ for all $N \in \mathbb{N}$ and $\sqrt{x} \leq 1 + x$ for all $x \geq 0$. Similarly, from the definition of $\rho_M(\eta_M)$ we may conclude that

$$\begin{aligned}
f_M(\rho_M(\eta_M)) &= \frac{C_1\sigma^2\sqrt{n}}{\lambda_{\min}(\Sigma)} \left(\sqrt{\frac{n}{N}} + \frac{n}{N}\right) + \frac{\sigma^2\sqrt{n}}{\lambda_{\min}} \left(\sqrt{\frac{\log(C_2/\eta_M)}{C_3N}} + \frac{\log(C_2/\eta_M)}{C_3N}\right) \\
&\leq \frac{n\sigma^2C_1C_3(1 + \sqrt{n}) + \sigma^2\sqrt{C_3n}}{\lambda_{\min}(\Sigma)C_3\sqrt{N}} + \frac{\sigma^2\sqrt{n}(\sqrt{C_3} + 1)\log(C_2/\eta_M)}{\lambda_{\min}(\Sigma)C_3\sqrt{N}},
\end{aligned}$$

where the inequality holds again because $1/N \leq 1/\sqrt{N}$ for all $N \in \mathbb{N}$ and $\sqrt{x} \leq 1+x$ for all $x \geq 0$. Setting $\eta_\mu = \eta_M = \eta/2$, the sum of the above upper bounds on $f_\mu(\rho_\mu(\eta_\mu))$ and $f_M(\rho_M(\eta_M))$ equals $\rho(\eta) = (c_1 + c_2 \log(1/\eta))/\sqrt{N}$ for some positive constants c_1 and c_2 that depend only on μ, Σ, σ^2 and n. The claim then follows from (20) and our choice of η_μ and η_M. This observation completes the proof. *Q.E.D.*

PROOF OF PROPOSITION 1: The inclusion (4) follows immediately from the Gelbrich bound of Theorem 1. It thus suffices to prove the reverse inclusion for $\widehat{\Sigma} \succ 0$. To this end, select any $(\mu, \Sigma) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$, and construct the pushforward distribution $\mathbb{Q} = \widehat{\mathbb{P}} \circ f^{-1}$ using the affine function $f(\xi) = \widehat{\Sigma}^{-\frac{1}{2}}(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}})^{\frac{1}{2}} \widehat{\Sigma}^{-\frac{1}{2}}(\xi - \widehat{\mu}) + \mu$. As the structural ambiguity set $\mathcal{S}$ is closed under positive semidefinite affine pushforwards, we have $\mathbb{Q} \in \mathcal{S}$. In addition, Theorem 2 implies that $\mathbb{W}(\mathbb{Q}, \widehat{\mathbb{P}}) = \mathbb{G}\big((\mu, \Sigma), (\widehat{\mu}, \widehat{\Sigma})\big) \leq \rho$, where the inequality holds because $(\mu, \Sigma) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$. We may thus conclude that $\mathbb{Q} \in \mathcal{W}_\rho(\widehat{\mathbb{P}})$. Finally, an elementary calculation reveals that $\mathbb{Q}$ has mean μ and covariance matrix Σ. *Q.E.D.*

PROOF OF THEOREM 3: The claim trivially holds if $\mathcal{W}_\rho(\widehat{\mathbb{P}})$ is empty. From now on we thus assume that $\mathcal{W}_\rho(\widehat{\mathbb{P}})$ is non-empty. For any distribution $\mathbb{Q} \in \mathcal{W}_\rho(\widehat{\mathbb{P}})$ with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{S}^n_+$, Theorem 1 then implies that

$$\mathbb{G}\big((\mu, \Sigma), (\widehat{\mu}, \widehat{\Sigma})\big) \leq \mathbb{W}(\mathbb{Q}, \widehat{\mathbb{P}}) \leq \rho.$$

As $\mathbb{Q} \in \mathcal{S}$, we thus have $\mathbb{Q} \in \mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$. Hence, we may conclude that $\mathcal{W}_\rho(\widehat{\mathbb{P}}) \subseteq \mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$.

Assume now that the structural ambiguity set $\mathcal{S}$ is generated by $\widehat{\mathbb{P}}$ and that $\widehat{\Sigma} \succ 0$. Next, select any distribution $\mathbb{Q} \in \mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$ with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{S}^n_+$. As $\mathbb{Q} \in \mathcal{S}$ constitutes a positive semidefinite affine pushforward of $\widehat{\mathbb{P}}$, we thus have

$$\mathbb{W}(\mathbb{Q}, \widehat{\mathbb{P}}) = \mathbb{G}\big((\mu, \Sigma), (\widehat{\mu}, \widehat{\Sigma})\big) \leq \rho,$$

where the equality follows from Theorem 2. This implies that $\mathbb{Q} \in \mathcal{W}_\rho(\widehat{\mathbb{P}})$ and, as $\mathbb{Q} \in \mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$ was chosen arbitrarily, that $\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma}) \subseteq \mathcal{W}_\rho(\widehat{\mathbb{P}})$. Recalling the first part of the proof, we then obtain $\mathcal{W}_\rho(\widehat{\mathbb{P}}) = \mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$. *Q.E.D.*

PROOF OF PROPOSITION 2: The non-negativity of the Gelbrich distance implies that

$$\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma}) = \left\{ (\mu, \Sigma) \in \mathbb{R}^n \times \mathbb{S}^n_+ : \mathbb{G}^2\big((\mu, \Sigma), (\widehat{\mu}, \widehat{\Sigma})\big) \leq \rho^2 \right\}.$$

Recall that the squared Gelbrich distance is convex in $(\mu, \Sigma) \in \mathbb{R}^n \times \mathbb{S}^n_+$. The Hölder continuity of the matrix square root established in [Nguyen et al., 2021b, Lemma A.2] ensures that the squared Gelbrich distance is also continuous. Therefore, $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is convex and closed. Finally, one can show that for any $(\mu, \Sigma) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ we have $\|\mu - \widehat{\mu}\| \leq \rho$ and $0 \preceq \Sigma \preceq \big(\rho + \operatorname{Tr}\big[\widehat{\Sigma}\big]^{\frac{1}{2}}\big)^2 I$, see [Shafieezadeh-Abadeh et al., 2018, Lemma A.6]. This implies that $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is also compact. *Q.E.D.*

PROOF OF PROPOSITION 3: Recall from Proposition 2 that $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is compact, and note that $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ can be viewed as the image of $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ under the continuous transformation $f(\mu, \Sigma) = (\mu, \Sigma + \mu\mu^\top)$ defined on $\mathbb{R}^n \times \mathbb{S}^n_+$. Thus, the transformed set $\mathcal{V}_\rho(\widehat{\mu}, \widehat{\Sigma})$ inherits compactness from $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$. To show that $\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma})$ is convex, recall from the proof of Theorem 1 that the squared Gelbrich distance satisfies

$$\mathbb{G}^2\big((\mu,\Sigma),(\widehat{\mu},\widehat{\Sigma})\big) = \begin{cases} \inf \ \|\mu-\widehat{\mu}\|^2 + \mathrm{Tr}\left[\Sigma+\widehat{\Sigma}-2C\right] \\ \mathrm{s.t.}\ C\in\mathbb{R}^{n\times n},\ \begin{bmatrix}\Sigma & C\\ C^\top & \widehat{\Sigma}\end{bmatrix}\succeq 0 \end{cases}$$

$$= \begin{cases} \sup \ \|\mu-\widehat{\mu}\|^2 + \mathrm{Tr}\left[\Sigma(I-A_{11})+\widehat{\Sigma}(I-A_{22})\right] \\ \mathrm{s.t.}\ A_{11}\in\mathbb{S}^n_+,\ A_{22}\in\mathbb{S}^n_+,\ \begin{bmatrix}A_{11} & -I\\ -I & A_{22}\end{bmatrix}\succeq 0. \end{cases}$$

Here, the second equality follows from strong semidefinite duality [Ben-Tal and Nemirovski, 2001, Theorem 1.4.2], which holds because $A_{11}=A_{22}=2I$ constitutes a Slater point for the dual problem. Thus, we have

$$\mathbb{G}^2\big((\mu,M-\mu\mu^\top),(\widehat{\mu},\widehat{\Sigma})\big)$$
$$= \begin{cases} \sup\ \mathrm{Tr}\left[M(I-A_{22})\right] + \mu^\top A_{22}\mu - 2\mu^\top\widehat{\mu} + \mathrm{Tr}\left[\widehat{\Sigma}(I-A_{22})\right] + \|\widehat{\mu}\|^2 \\ \mathrm{s.t.}\ A_{11}\in\mathbb{S}^n_+,\ A_{22}\in\mathbb{S}^n_+,\ \begin{bmatrix}A_{11} & -I\\ -I & A_{22}\end{bmatrix}\succeq 0 \end{cases}$$

for any $\mu\in\mathbb{R}^n$ and $M\in\mathbb{S}^n_+$ with $M\succeq\mu\mu^\top$. Note that the objective function of the above maximization problem is jointly convex in μ and M for any feasible A_{11} and A_{22}. As convexity is preserved under maximization, we may thus conclude that $\mathbb{G}^2((\mu,M-\mu\mu^\top),(\widehat{\mu},\widehat{\Sigma}))$ is also jointly convex in μ and M. Hence, the set

$$\mathcal{V}_\rho(\widehat{\mu},\widehat{\Sigma}) = \left\{(\mu,M)\in\mathbb{R}^n\times\mathbb{S}^n_+ : M\succeq\mu\mu^\top,\ \mathbb{G}^2\big((\mu,M-\mu\mu^\top),(\widehat{\mu},\widehat{\Sigma})\big)\le\rho^2\right\}$$

is convex because it is representable as the feasible set of convex constraints. *Q.E.D.*

APPENDIX B: SUPPORT FUNCTION OF $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$

We now derive several technical results that will be needed in Appendix C to examine the Gelbrich mean-variance risk. These results are also relevant for robust optimization. Indeed, the mean-covariance uncertainty set $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ can conveniently be used in the context of classical robust optimization because a robust constraint that requires a concave function $h(\mu,\Sigma)$ to be nonpositive for all $(\mu,\Sigma)\in\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ can be reformulated as a convex constraint that involves the convex conjugate of $-h(\mu,\Sigma)$ as well as the support function of $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ [Ben-Tal et al., 2015, Theorem 2]. Formally, we have

$$\begin{aligned} &h(\mu,\Sigma)\le 0 \quad \forall(\mu,\Sigma)\in\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma}) \\ &\qquad\qquad \iff \quad \exists\, q\in\mathbb{R}^n, Q\in\mathbb{S}^n \text{ such that } (-h)^*(q,Q)+\delta^*_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q)\le 0. \end{aligned}$$

This constraint is computationally tractable for many commonly used constraint functions because the support function of $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ can be represented as the optimal value of a tractable conic minimization problem.

PROPOSITION 13: *For any $\rho \geq 0$, $q \in \mathbb{R}^n$ and $Q \in \mathbb{S}^n$, the support function of $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ satisfies*

$$\delta^*_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q) = \begin{cases} \inf \ q^\top \widehat{\mu} + \tau + \gamma\big(\rho^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + \mathrm{Tr}\left[Z\right] \\ \mathrm{s.t.}\ \gamma \in \mathbb{R}_+,\ \tau \in \mathbb{R}_+,\ Z \in \mathbb{S}^n_+ \\ \qquad \begin{bmatrix} \gamma I - Q & \gamma \widehat{\Sigma}^{\frac{1}{2}} \\ \gamma \widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0, \quad \left\| \begin{bmatrix} q \\ \tau - \gamma \end{bmatrix} \right\| \leq \tau + \gamma. \end{cases}$$

PROOF OF PROPOSITION 13: Evaluating the support function of $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ at a given point $(q, Q) \in \mathbb{R}^n \times \mathbb{S}^n$ amounts to solving the finite convex program

$$\delta^*_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q) = \begin{cases} \sup\limits_{\mu, \Sigma \succeq 0} \ q^\top \mu + \mathrm{Tr}\left[Q\Sigma\right] \\ \mathrm{s.t.}\ \ \|\mu - \widehat{\mu}\|^2 + \mathrm{Tr}\Big[\Sigma + \widehat{\Sigma} - 2\big(\widehat{\Sigma}^{\frac{1}{2}} \Sigma \widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\Big] \leq \rho^2. \end{cases}$$

Using the semidefinite programming reformulation for the squared Gelbrich distance on the right hand side of (17) and introducing an auxiliary variable M that equals $\mu\mu^\top$ at optimality, we then obtain

$$\begin{aligned} \delta^*_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q) &= \begin{cases} \sup\limits_{\mu, C, \Sigma \succeq 0} \ q^\top \mu + \mathrm{Tr}\left[Q\Sigma\right] \\ \mathrm{s.t.}\quad \mathrm{Tr}\left[\mu\mu^\top - 2\widehat{\mu}\mu^\top + \widehat{\mu}\widehat{\mu}^\top\right] + \mathrm{Tr}\big[\Sigma + \widehat{\Sigma} - 2C\big] \leq \rho^2 \\ \qquad\quad \begin{bmatrix} \Sigma & C \\ C^\top & \widehat{\Sigma} \end{bmatrix} \succeq 0 \end{cases} \\ &= \begin{cases} \sup q^\top \mu + \mathrm{Tr}\left[Q\Sigma\right] \\ \mathrm{s.t.}\ \mu \in \mathbb{R}^n,\ C \in \mathbb{R}^{n \times n},\ \Sigma \in \mathbb{S}^n_+,\ M \in \mathbb{S}^n_+ \\ \qquad \mathrm{Tr}\left[M - 2\widehat{\mu}\mu^\top\right] + \|\widehat{\mu}\|^2 + \mathrm{Tr}\big[\Sigma + \widehat{\Sigma} - 2C\big] \leq \rho^2 \\ \qquad \begin{bmatrix} \Sigma & C \\ C^\top & \widehat{\Sigma} \end{bmatrix} \succeq 0, \ \begin{bmatrix} M & \mu \\ \mu^\top & 1 \end{bmatrix} \succeq 0. \end{cases} \end{aligned}$$

By strong conic programming duality [Ben-Tal and Nemirovski, 2001, Theorem 1.4.2], the resulting semidefinite program is equivalent to

$$\delta^*_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q,Q) = \begin{cases} \inf \ \gamma\left(\rho^2 - \|\widehat{\mu}\|^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\right) + \mathrm{Tr}\big[A_{22}\widehat{\Sigma}\big] + \beta \\ \mathrm{s.t.}\ \gamma \in \mathbb{R}_+,\ \begin{bmatrix} A_{11} & \gamma I \\ \gamma I & A_{22} \end{bmatrix} \in \mathbb{S}^{2n}_+,\ \begin{bmatrix} B & \gamma\widehat{\mu} + \frac{q}{2} \\ \big(\gamma\widehat{\mu} + \frac{q}{2}\big)^\top & \beta \end{bmatrix} \in \mathbb{S}^{n+1}_+ \\ \qquad A_{11} \preceq \gamma I - Q,\ B \preceq \gamma I. \end{cases} \tag{21}$$

Strong duality holds because $\gamma = \max\{\lambda_{\max}(Q), 0\} + 2$, $A_{11} = I$, $A_{22} = 2\gamma^2 I$, $B = I$ and $\beta = \|\gamma\widehat{\mu} + \frac{q}{2}\|^2 + 1$ represents a Slater point for the dual problem and because the primal problem is trivially feasible. Next, we simplify the semidefinite program (21) by eliminating the decision variables A_{11} and B, each of which appears in two opposing matrix inequalities. If $\gamma > 0$, then A_{22} has full rank thanks to [Bernstein, 2009, Corollary 8.2.2]. In this case, we obtain the following equivalences by Schur complementing the two matrix inequalities involving A_{11}.

$$\begin{aligned}\begin{bmatrix} A_{11} & \gamma I \\ \gamma I & A_{22} \end{bmatrix} \succeq 0,\ A_{11} \preceq \gamma I - Q \iff & \gamma^2 A_{22}^{-1} \preceq A_{11} \preceq \gamma I - Q \\ \iff & \begin{bmatrix} \gamma I - Q & \gamma I \\ \gamma I & A_{22} \end{bmatrix} \succeq 0 \end{aligned} \tag{22}$$

If $\gamma = 0$, on the other hand, we trivially have

$$\begin{bmatrix} A_{11} & \gamma I \\ \gamma I & A_{22} \end{bmatrix} \succeq 0,\ A_{11} \preceq \gamma I - Q \iff 0 \preceq A_{11} \preceq -Q,\ A_{22} \succeq 0 \iff \begin{bmatrix} -Q & 0 \\ 0 & A_{22} \end{bmatrix} \succeq 0.$$

Therefore, all equivalences in (22) hold for any $\gamma \geq 0$. If $\beta > 0$, then Schur complementing the two matrix inequalities involving B yields

$$\begin{aligned}\begin{bmatrix} B & \gamma\widehat{\mu} + \frac{q}{2} \\ \left(\gamma\widehat{\mu} + \frac{q}{2}\right)^\top & \beta \end{bmatrix} \succeq 0,\ B \preceq \gamma I \iff & \frac{1}{\beta}\left(\gamma\widehat{\mu} + \frac{q}{2}\right)\left(\gamma\widehat{\mu} + \frac{q}{2}\right)^\top \preceq B \preceq \gamma I \\ \iff & \begin{bmatrix} \gamma I & \gamma\widehat{\mu} + \frac{q}{2} \\ (\gamma\widehat{\mu} + \frac{q}{2})^\top & \beta \end{bmatrix} \succeq 0. \end{aligned} \tag{23}$$

If $\beta = 0$, on the other hand, we have

$$\begin{bmatrix} B & \gamma\widehat{\mu} + \frac{q}{2} \\ \left(\gamma\widehat{\mu} + \frac{q}{2}\right)^\top & \beta \end{bmatrix} \succeq 0,\ B \preceq \gamma I \iff 0 \preceq B \preceq \gamma I \iff \begin{bmatrix} \gamma I & 0 \\ 0 & 0 \end{bmatrix} \succeq 0,$$

where the first equivalence holds because $\beta = 0$ implies via the first matrix inequality that $\gamma\widehat{\mu} + \frac{q}{2} = 0$; see [Bernstein, 2009, Corollary 8.2.2]. Therefore, all equivalences in (23) hold for any $\beta \geq 0$. In summary, we have thus shown that

$$\delta^*_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q) = \begin{cases} \inf\limits_{\gamma,\beta,A_{22}} & \gamma\left(\rho^2 - \|\widehat{\mu}\|^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\right) + \mathrm{Tr}\left[A_{22}\widehat{\Sigma}\right] + \beta \\ \text{s.t.} & \begin{bmatrix} \gamma I - Q & \gamma I \\ \gamma I & A_{22} \end{bmatrix} \succeq 0,\ \begin{bmatrix} \gamma I & \gamma\widehat{\mu} + \frac{q}{2} \\ (\gamma\widehat{\mu} + \frac{q}{2})^\top & \beta \end{bmatrix} \succeq 0. \end{cases} \tag{24}$$

For any fixed $\gamma \geq 0$, problem (24) decomposes into two separate minimization problems over β and A_{22}, respectively. If $\gamma > 0$, the partial minimization problem over β reduces to

$$\begin{aligned} & \inf\left\{\beta : \begin{bmatrix} \gamma I & \gamma\widehat{\mu} + \frac{q}{2} \\ (\gamma\widehat{\mu} + \frac{q}{2})^\top & \beta \end{bmatrix} \succeq 0\right\} \\ &= \inf\left\{\beta : \beta - \gamma\|\widehat{\mu}\|^2 - q^\top\widehat{\mu} - \frac{1}{4\gamma}\|q\|^2 \geq 0\right\} \\ &= \inf\left\{\eta + \gamma\|\widehat{\mu}\|^2 : \eta \geq q^\top\widehat{\mu} + \tau,\ \left\|\begin{bmatrix} q \\ \tau - \gamma \end{bmatrix}\right\| \leq \tau + \gamma,\ \tau \geq 0\right\}, \end{aligned}$$

where the first equality exploits a standard Schur complement argument, and the second equality follows from the substitution $\eta \leftarrow \beta - \gamma\|\widehat{\mu}\|^2$ and from introducing an auxiliary variable $\tau \geq 0$ subject to the hyperbolic constraint $\tau \geq \|q\|^2/(4\gamma)$. Note also that the first and the last minimization problems in the above expression remain equivalent when $\gamma = 0$. Similarly, if $\gamma > 0$, then the partial minimization problem over A_{22} reduces to

$$\begin{aligned}
&\inf\left\{\operatorname{Tr}\left[A_{22}\widehat{\Sigma}\right] : \begin{bmatrix}\gamma I - Q & \gamma I\\ \gamma I & A_{22}\end{bmatrix} \succeq 0\right\}\\
&= \inf\left\{\operatorname{Tr}\left[Z\right] : Z \succeq \widehat{\Sigma}^{\frac{1}{2}}A_{22}\widehat{\Sigma}^{\frac{1}{2}}, A_{22} \succeq \gamma^2(\gamma I - Q)^{-1}\right\}\\
&= \inf\left\{\operatorname{Tr}\left[Z\right] : Z \succeq \gamma\widehat{\Sigma}^{\frac{1}{2}}(\gamma I - Q)^{-1}\gamma\widehat{\Sigma}^{\frac{1}{2}}\right\}\\
&= \inf\left\{\operatorname{Tr}\left[Z\right] : \begin{bmatrix}\gamma I - Q & \gamma\widehat{\Sigma}^{\frac{1}{2}}\\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z\end{bmatrix} \succeq 0\right\},
\end{aligned}$$

where the first and the third equalities follow from Schur complement arguments. As $\widehat{\Sigma} \succ 0$ by assumption, the first and the last minimization problems in the above expression remain equivalent when $\gamma = 0$. The claim then follows by substituting the reformulated partial minimization problems into (24) and eliminating η. *Q.E.D.*

The next lemma shows that the unique maximizer of the optimization problem defining the support function of $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ can be computed in quasi-closed form.

LEMMA 4: *Suppose that $\widehat{\Sigma} \succ 0$, $\rho > 0$ and either $q \neq 0$ or $\lambda_{\max}(Q) > 0$. Then the optimization problem*

$$\delta^*_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q) = \sup_{(\mu,\Sigma)\in\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})} q^\top \mu + \operatorname{Tr}\left[Q\Sigma\right] \tag{25a}$$

is uniquely solved by

$$\mu^\star = \widehat{\mu} + \frac{q}{2\gamma^\star} \quad \textit{and} \quad \Sigma^\star = \left(I - \frac{Q}{\gamma^\star}\right)^{-1}\widehat{\Sigma}\left(I - \frac{Q}{\gamma^\star}\right)^{-1}, \tag{25b}$$

where $\gamma^\star > \max\{\lambda_{\max}(Q), 0\}$ is the unique solution of the nonlinear algebraic equation

$$\frac{\|q\|^2}{4\gamma^2} + \operatorname{Tr}\left[\widehat{\Sigma}\left(I - \gamma(\gamma I - Q)^{-1}\right)^2\right] = \rho^2. \tag{25c}$$

In addition, if $Q \succeq 0$ then we have $\Sigma^\star \succeq \lambda_{\min}(\widehat{\Sigma})I$.

PROOF OF LEMMA 4: From the proof of Proposition 13 we know that problem (25a) is equivalent to the semidefinite program (24). We first prove that $\gamma > 0$ for any (γ, β, A_{22}) feasible in (24). To see this, assume for the sake of argument that $\gamma = 0$, in which case the two matrix inequalities in (24) reduce to

$$\begin{bmatrix}-Q & 0\\ 0 & A_{22}\end{bmatrix} \succeq 0 \quad \text{and} \quad \begin{bmatrix}0 & q/2\\ (q/2)^\top & \beta\end{bmatrix} \succeq 0.$$

However, these constraints are not satisfiable by any β and A_{22} under our assumption that either $q \neq 0$ or $\lambda_{\max}(Q) > 0$. Similarly, one can show that $\gamma > \lambda_{\max}(Q)$ for any (γ, β, A_{22}) feasible in (24). Indeed, we have

$$\begin{bmatrix} \gamma I - Q & \gamma I \\ \gamma I & A_{22} \end{bmatrix} \succeq 0,\ \gamma > 0 \quad \Longrightarrow \quad \gamma I - Q \succ 0 \quad \Longrightarrow \quad \gamma > \lambda_{\max}(Q),$$

where the first implication follows from [Bernstein, 2009, Corollary 8.2.2], which requires $\gamma I - Q$ to have full rank whenever the off-diagonal block γI has full rank. Schur complementing the two matrix inequalities in (24) yields

$$A_{22} \succeq \gamma^2 (\gamma I - Q)^{-1} \quad \text{and} \quad \beta \geq \|\gamma \widehat{\mu} + \tfrac{q}{2}\|^2/\gamma,$$

which is possible because $\gamma > \max\{\lambda_{\max}(Q), 0\}$. Using these relations to eliminate β and A_{22} from (24) yields

$$\begin{aligned} &\delta^*_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q) \\ &= \inf_{\gamma>0\gamma>\lambda_{\max}(Q)} q^\top \widehat{\mu} + \frac{\|q\|^2}{4\gamma} + \gamma\big(\rho^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\big) + \gamma^2 \operatorname{Tr}\big[(\gamma I - Q)^{-1}\widehat{\Sigma}\big]. \end{aligned} \tag{26}$$

In the following we use $f(\gamma)$ to denote the objective function of problem (26). As $\widehat{\Sigma} \succ 0$, $f(\gamma)$ satisfies

$$f(\gamma) \geq q^\top \widehat{\mu} + \frac{\|q\|^2}{4\gamma} + \gamma\big(\rho^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\big) + \lambda_{\min}(\widehat{\Sigma})\gamma^2 \operatorname{Tr}\big[(\gamma I - Q)^{-1}\big]$$

for all $\gamma \geq \max\{\lambda_{\max}(Q), 0\}$. Thus, $f(\gamma)$ diverges as γ decreases to $\max\{\lambda_{\max}(Q), 0\}$. Similarly, since $\rho > 0$, one readily verifies that $f(\gamma)$ grows indefinitely as γ increases to infinity. Noting that $f(\gamma)$ is smooth and strictly convex on its domain, these insights reveal that problem (26) has a unique minimizer $\gamma^\star \in (\max\{0, \lambda_{\max}(Q)\}, \infty)$. This minimizer can be found by solving the problem's first-order optimality condition $f'(\gamma^\star) = 0$, where

$$\begin{aligned} f'(\gamma) &= -\frac{\|q\|^2}{4\gamma^2} + \rho^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big] + 2\gamma \operatorname{Tr}\big[(\gamma I - Q)^{-1}\widehat{\Sigma}\big] - \gamma^2 \operatorname{Tr}\big[(\gamma I - Q)^{-2}\widehat{\Sigma}\big] \\ &= \rho^2 - \frac{\|q\|^2}{4\gamma^2} - \operatorname{Tr}\big[\widehat{\Sigma}\big(I - \gamma(\gamma I - Q)^{-1}\big)^2\big] \end{aligned}$$

denotes the derivative of $f(\gamma)$. Thus, the algebraic equation (25c) represents the (necessary and sufficient) first-order optimality condition of problem (26). In the remainder of the proof, we demonstrate that $(\mu^\star, \Sigma^\star)$ as defined in (25b) constitutes a global maximizer of problem (25a). To see this, note first that $\Sigma^\star \succeq 0$ and

$$\mathbb{G}^2\big((\mu^\star, \Sigma^\star), (\widehat{\mu}, \widehat{\Sigma})\big)^2 = \frac{\|q\|^2}{4(\gamma^\star)^2} + \operatorname{Tr}\big[\widehat{\Sigma}(I - \gamma^\star(\gamma^\star I - Q)^{-1})^2\big] = \rho^2,$$

where the first equality follows from (25b) and the definition of the Gelbrich distance, whereas the second equality holds because $\gamma^\star$ solves (25c). Thus, $(\mu^\star, \Sigma^\star)$ is feasible in problem (25a). Furthermore, the objective function value of $(\mu^\star, \Sigma^\star)$ in (25a) evaluates to

$$q^\top \mu^\star + \operatorname{Tr}\big[Q\Sigma^\star\big]$$

$$
\begin{aligned}
&= q^\top \widehat{\mu} + \frac{\|q\|^2}{2\gamma^\star} + (\gamma^\star)^2 \operatorname{Tr}\left[Q(\gamma^\star I - Q)^{-1}\widehat{\Sigma}(\gamma^\star I - Q)^{-1}\right] \\
&= q^\top \widehat{\mu} + \frac{\|q\|^2}{2\gamma^\star} + (\gamma^\star)^2 \operatorname{Tr}\left[(Q - \gamma^\star I + \gamma^\star I)(\gamma^\star I - Q)^{-1}\widehat{\Sigma}(\gamma^\star I - Q)^{-1}\right] \\
&= q^\top \widehat{\mu} + \frac{\|q\|^2}{2\gamma^\star} - (\gamma^\star)^2 \operatorname{Tr}\left[\widehat{\Sigma}(\gamma^\star I - Q)^{-1}\right] + (\gamma^\star)^3 \operatorname{Tr}\left[\widehat{\Sigma}(\gamma I - Q)^{-2}\right] \\
&= q^\top \widehat{\mu} + \frac{\|q\|^2}{2\gamma^\star} + \gamma^\star \left(\rho^2 - \frac{\|q\|^2}{4(\gamma^\star)^2} - \operatorname{Tr}\left[\widehat{\Sigma}\right]\right) + (\gamma^\star)^2 \operatorname{Tr}\left[(\gamma^\star I - Q)^{-1}\widehat{\Sigma}\right] \\
&= q^\top \widehat{\mu} + \frac{\|q\|^2}{4\gamma^\star} + \gamma^\star\left(\rho^2 - \operatorname{Tr}\left[\widehat{\Sigma}\right]\right) + (\gamma^\star)^2 \operatorname{Tr}\left[(\gamma^\star I - Q)^{-1}\widehat{\Sigma}\right] = \delta^*_{\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})}(q, Q),
\end{aligned}
$$

where the fourth equality holds because $\gamma^\star > 0$ solves (25c), and the last equality follows from the optimality of $\gamma^\star$ in (26). Thus, $(\mu^\star, \Sigma^\star)$ is optimal in (25a). As problem (25a) has a linear objective function and a strictly convex feasible set, the maximizwer $(\mu^\star, \Sigma^\star)$ is unique. To complete the proof, note that if $Q \succeq 0$, then $(I - Q/\gamma^\star)^{-1} \succeq I$ because $\gamma^\star I \succ Q$, and thus it is easy to verify that $\Sigma^\star \succeq \lambda_{\min}(\widehat{\Sigma}) I$. *Q.E.D.*

APPENDIX C: MEAN-VARIANCE RISK MEASURES

Mean-variance risk measures appear in static portfolio selection [Markowitz, 1952, Föllmer and Schied, 2008a] and in myopic reformulations of dynamic portfolio optimization [Merton, 1969], due to their analytic tractability. For any probability distribution $\mathbb{Q} \in \mathcal{M}$, the mean-variance risk measure with risk-aversion coefficient $\beta \geq 0$ of any loss function $\ell \in \mathcal{L}_0$ is defined as

$$
\mathcal{R}_{\mathbb{Q}}(\ell(\xi)) = \mathbb{E}_{\mathbb{Q}}[\ell(\xi)] + \beta \mathbb{V}\mathrm{ar}_{\mathbb{Q}}(\ell(\xi)),
$$

where $\mathbb{V}\mathrm{ar}_{\mathbb{Q}}(\ell(\xi))$ denotes the variance of the loss $\ell(\xi)$ under the probability distribution $\mathbb{Q}$. Even though the mean-variance risk measure gives rise to a law-invariant family of translation invariant risk measures, it fails to be positive homogeneous, and therefore Theorem 5 is not applicable. Nevertheless, the Gelbrich risk evaluation problem can still be reformulated as a tractable second-order cone program.

THEOREM 6—Gelbrich mean-variance risk of linear loss functions: *Suppose that $\{\mathcal{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathcal{M}}$ is a family of mean-variance risk measures with coefficient $\beta > 0$. If $\widehat{\Sigma} \succ 0$, then the Gelbrich risk and the Wasserstein risk of any portfolio loss function $\ell(\xi) = -w^\top \xi$ coincide and are equal to the optimal value of the second-order cone program*

$$
\begin{array}{ll}
\inf & \gamma\rho^2 - \widehat{\mu}^\top w + \frac{1}{4} y + \beta z \\
\text{s.t.} & \gamma \in \mathbb{R}_+,\ y \in \mathbb{R}_+,\ z \in \mathbb{R}_+ \\
& \left\| \begin{pmatrix} 2\widehat{\Sigma}^{\frac{1}{2}} w \\ z + \beta y - 1 \end{pmatrix} \right\| \leq z - \beta y + 1,\ \left\| \begin{pmatrix} 2w \\ y - \gamma \end{pmatrix} \right\| \leq y + \gamma,\ \beta y \leq 1.
\end{array}
\tag{27}
$$

PROOF OF THEOREM 6: The equivalence of the Gelbrich risk and the Wasserstein risk for portfolio loss functions follows from Proposition 1 and the observation that the mean-variance risk measure depends on the distribution of the asset returns only through its first and second moments. It remains to be shown that the Gelbrich risk coincides with (27).

If $w = 0$, then the Gelbrich risk evaluates to 0, and problem (27) is solved by $\gamma = y = z = 0$. Thus, the claim is trivially satisfied. From now on we may assume without loss of generality that $w \neq 0$. Using the decomposition (7a), the Gelbrich risk can be recast as

$$\mathcal{R}_{\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})}(\ell) = \sup_{(\mu,\Sigma)\in\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q}\in\mathcal{C}(\mu,\Sigma)} \mathcal{R}_{\mathbb{Q}}(-w^\top\xi) = \sup_{(\mu,\Sigma)\in\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \left\{-w^\top\mu + \beta w^\top\Sigma w\right\},$$

where the second equality uses the definition of the mean-variance risk measure. Note that the last optimization problem in the above expression evaluates the support function of $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ at the point $(-w, \beta w w^\top)$, and by Proposition 13 we thus have

$$\sup_{\mathbb{Q}\in\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})} \mathcal{R}_{\mathbb{Q}}(-w^\top\xi) = \begin{cases} \inf \ -\widehat{\mu}^\top w + \tau + \gamma\big(\rho^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + \mathrm{Tr}\left[Z\right] \\ \text{s.t.}\ \gamma\in\mathbb{R}_+,\ \tau\in\mathbb{R}_+,\ Z\in\mathbb{S}_+^n \\ \qquad \begin{bmatrix} \gamma I - \beta w w^\top & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \left\| \begin{pmatrix} -w \\ \tau - \gamma \end{pmatrix} \right\| \leq \tau + \gamma. \end{cases}$$

Fix now any feasible solution (γ, τ, Z) of the resulting semidefinite program. The first matrix inequality implies that $\gamma I \succeq \beta w w^\top$. As $\beta > 0$ and $w \neq 0$, this is only possible if $\gamma > 0$, which in turn implies that $\gamma\widehat{\Sigma}^{\frac{1}{2}} \succ 0$. By [Bernstein, 2009, Corollary 8.2.2], the first matrix inequality in the above semidefinite program therefore also implies that $\gamma I - \beta w w^\top \succ 0$, which is equivalent to $\gamma > \beta\|w\|^2$. It is easy to verify that, at optimality, τ coincides with $\|w\|^2/(4\gamma)$ and Z coincides with its Schur complement $\gamma^2(\gamma I - \beta w w^\top)^{-1}\widehat{\Sigma}$, which is well-defined because $\gamma I - \beta w w^\top \succ 0$. Thus, τ and Z can be eliminated together with their constraints to obtain

$$\begin{aligned} &\sup_{\mathbb{Q}\in\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})} \mathcal{R}_{\mathbb{Q}}(-w^\top\xi) \\ &= \inf_{\gamma>\beta\|w\|^2} -\widehat{\mu}^\top w + \frac{\|w\|^2}{4\gamma} + \gamma\big(\rho^2 - \mathrm{Tr}\big[\widehat{\Sigma}\big]\big) + \gamma^2\,\mathrm{Tr}\big[(\gamma I - \beta w w^\top)^{-1}\widehat{\Sigma}\big] \\ &= \inf_{\gamma>\beta\|w\|^2} \gamma\rho^2 - \widehat{\mu}^\top w + \frac{\|w\|^2}{4\gamma} + \beta(1 - \beta\gamma^{-1}\|w\|^2)^{-1} w^\top\widehat{\Sigma} w, \end{aligned}$$

where the second equality follows from the Sherman-Morrison formula [Bernstein, 2009, Corollary 2.8.8]. Introducing an auxiliary variable $y \geq 0$ and rewriting the constraint in the last expression as $\|w\|^2/\gamma \leq y \leq \beta^{-1}$ yields

$$\sup_{\mathbb{Q}\in\mathcal{G}_\rho(\widehat{\mu},\widehat{\Sigma})} \mathcal{R}_{\mathbb{Q}}\left(-w^\top\xi\right) = \begin{cases} \inf\ \gamma\rho^2 - \widehat{\mu}^\top w + \frac{y}{4} + \beta(1-\beta y)^{-1} w^\top\widehat{\Sigma} w \\ \text{s.t.}\ \|w\|^2/\gamma \leq y \leq \beta^{-1},\ \gamma > 0. \end{cases}$$

As $\gamma \geq 0$, $y \geq 0$ and $\beta y \leq 1$, we may use [Lobo et al., 1998, Equation (8)] to reformulate the hyperbolic constraint and the quadratic-over-linear term in the objective function in terms of second-order cone constraints, that is,

$$\begin{aligned} \|w\|^2 \leq \gamma y \iff \left\| \begin{pmatrix} 2w \\ y - \gamma \end{pmatrix} \right\| \leq y + \gamma \quad \text{and} \quad \frac{w^\top\widehat{\Sigma} w}{1 - \beta y} \leq z \\ \iff \left\| \begin{pmatrix} 2\widehat{\Sigma}^{\frac{1}{2}} w \\ z + \beta y - 1 \end{pmatrix} \right\| \leq z - \beta y + 1, \end{aligned}$$

where $z \geq 0$ is an auxiliary epigraphical variable. Thus, the claim follows. *Q.E.D.*

In analogy to Proposition 6, we can determine the first and second moments of the worst-case probability distributions that maximize the Gelbrich mean-variance risk of a fixed linear loss function.

PROPOSITION 14—Worst-case moments: *Suppose that* $\{\mathcal{R}_{\mathbb{Q}}\}_{\mathbb{Q}\in\mathcal{M}}$ *is a family of mean-variance risk measures with* $\beta > 0$. *If* $\widehat{\Sigma} \succ 0$, $w \neq 0$ *and* $\rho > 0$, *then any extremal distribution* $\mathbb{Q}^\star$ *that attains the Gelbrich risk of the linear loss function* $\ell(\xi) = -w^\top \xi$ *has the same mean* $\mu^\star \in \mathbb{R}^n$ *and covariance matrix* $\Sigma^\star \in \mathbb{S}^n_+$, *where*

$$\mu^\star = \widehat{\mu} - \frac{w}{2\gamma^\star}, \quad \Sigma^\star = \left(I - \frac{\beta w w^\top}{\gamma^\star}\right)^{-1} \widehat{\Sigma} \left(I - \frac{\beta w w^\top}{\gamma^\star}\right)^{-1}$$

and $\gamma^\star > \beta\|w\|^2$ *is the unique solution of the nonlinear algebraic equation*

$$\frac{\|w\|^2}{4\gamma^2} + \mathrm{Tr}\left[\widehat{\Sigma}\left(I - \gamma(\gamma I - \beta w w^\top)^{-1}\right)^2\right] = \rho^2.$$

We emphasize again that there may be multiple extremal distributions in the Gelbrich ambiguity set $\mathcal{G}_\rho(\widehat{\mu}, \widehat{\Sigma})$ that share the unique extremal mean $\mu^\star$ and covariance matrix $\Sigma^\star$ identified in Proposition 14.

PROOF OF PROPOSITION 14: From the proof of Theorem 6 we know that the first and second moments of any extremal distribution $\mathbb{Q}^\star$ are maximizers of the support function evaluation problem $\sup\{q^\top \mu + \mathrm{Tr}\left[Q\Sigma\right] : (\mu, \Sigma) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})\}$ with $q = -w$ and $Q = \beta w w^\top$. The claim thus follows from Lemma 4, which applies because $w \neq 0$. *Q.E.D.*

Distributionally robust mean-variance portfolio optimization problems with 2-Wasserstein ball ambiguity sets were also studied by Blanchet et al. [2020]. Instead of minimizing a worst-case mean-variance risk measure, however, they minimize the worst-case variance of the portfolio return subject to a lower bound on the worst-case mean, which results in a more conservative model because the extremal distriburtions in the objective function and in the constraints may differ. The additional conservatism enhances analytical tractability.