Scores for Multivariate Distributions and Level Sets

Xiaochun Meng
University of Sussex Business School, University of Sussex, UK, xiaochun.meng@sussex.ac.uk

James W. Taylor
Saïd Business School, University of Oxford, UK, james.taylor@sussex.ac.uk

Souhaib Ben Taieb
Department of Computer Science, University of Mons, Belgium, souhaib.bentaieb@umons.ac.be

Siran Li*
School of Mathematical Sciences & IMA-Shanghai, Shanghai Jiao Tong University, Shanghai, China, sl4025@nyu.edu

Forecasts of multivariate probability distributions are required for a variety of applications. Scoring rules enable the evaluation of forecast accuracy, and comparison between forecasting methods. We propose a theoretical framework for scoring rules for multivariate distributions, which encompasses the existing quadratic score and multivariate continuous ranked probability score. We demonstrate how this framework can be used to generate new scoring rules. In some multivariate contexts, it is a forecast of a level set that is needed, such as a density level set for anomaly detection or the level set of the cumulative distribution as a measure of risk. This motivates consideration of scoring functions for such level sets. For univariate distributions, it is well-established that the continuous ranked probability score can be expressed as the integral over a quantile score. We show that, in a similar way, scoring rules for multivariate distributions can be decomposed to obtain scoring functions for level sets. Using this, we present scoring functions for different types of level set, including density level sets and level sets for cumulative distributions. To compute the scores, we propose a simple numerical algorithm. We perform a simulation study to support our proposals, and we use real data to illustrate usefulness for forecast combining and CoVaR estimation.

Key words: decision analysis; probabilistic forecasts; scoring functions; multivariate probability distributions; level sets; quantiles.

1. Introduction

Forecasts should be probabilistic, as this allows forecasters to communicate their uncertainty and hence better support decision-making. In many applications, a forecast of a multivariate probability distribution is needed. For example, Berrocal et al. (2010) consider the impact of extreme weather

* The research of Siran Li is supported by NSFC (National Natural Science Foundation of China) Grant No. 12201399 and Shanghai Frontier Research Institute for Modern Analysis.

risk on road maintenance by estimating the joint distribution of temperature and precipitation to enable the prediction of ice formation on road surfaces. Danaher and Smith (2011) show how accurate estimation of the joint distribution of website visit duration and expenditure can lead to improved distributional forecasts for the latter. Jeon and Taylor (2012) forecast the joint distribution of wind velocity in perpendicular directions, which they use to generate distributional forecasts of wind power. Diks et al. (2014) model the joint distribution of sets of exchange rates to capture tail dependence and its impact on common extreme appreciation and depreciation of the currencies.

To support such applications, Gneiting and Katzfuss (2014) describe the need for decision-theoretically principled methods for evaluating forecasts of multivariate distributions. Distributional forecast accuracy should be evaluated by maximizing sharpness subject to calibration. Sharpness relates to the concentration of the probabilistic forecast, while calibration concerns its statistical consistency with the data. A score summarizes both calibration and sharpness, and can be used to compare forecasts from competing methods. In addition, a score can be used as the objective function in model estimation (Jose 2017).

The continuous ranked probability score (CRPS) is widely used for univariate distributions (Grushka-Cockayne et al. 2017). For multivariate distributions, the energy score of Gneiting and Raftery (2007) has become popular as a multivariate generalization of the CRPS. An alternative generalization is the multivariate CRPS (MCRPS), introduced by Gneiting and Raftery (2007). This score has only been considered further by Yuen and Stoev (2014), who use it for estimation in extreme value theory. Another proposal is the quadratic score, which can be used for univariate and multivariate densities, but is distinct from the log score (see Gneiting and Raftery 2007).

Often the object of interest of a probability distribution is a statistic, or a property, such as a quantile of a univariate distribution. Scores for quantiles have been thoroughly studied (see, for example, Grushka-Cockayne et al. 2017, Jose and Winkler 2009). However, there is a far less developed literature on scores for level sets, which can be viewed as multivariate generalizations of quantiles. Roughly speaking, for a function g, an α level set contains all the points in $\mathbb{R}^d$ where g is greater than or equal to α. For example, if g is a cumulative distribution function (CDF), the α

level set of g is called an α CDF level set. To the best of our knowledge, the only score available for any form of multivariate level set is the excess mass score for density level sets (see, for example, Hartigan 1987). Estimates of such level sets are important for a variety of applications.

In the context of financial and insurance risk management, Embrechts and Puccetti (2006) propose the boundary of a CDF level set as a multivariate extension of value at risk (VaR). VaR helps institutions to decide the quantity of assets required to cover potential losses. The alternative risk measure of Cousin and Di Bernardino (2013) is the expectation of the underlying portfolio of risks, conditional on it lying on the boundary of the level set.

CDF level sets have been used frequently in environmental risk management, particularly in relation to hydrology and coastal management. For example, Corbella and Stretch (2012) use CDF level sets of the joint distribution of storm duration and wave height to quantify risk in relation to coastal erosion. For similar data, Salvadori et al. (2014) show how the CDF level set can be used to find critical values for storm duration and wave height, which are then used to decide the appropriate weight of concrete units in the design of a breakwater. In a bivariate flood hazard analysis, Moftakhari et al. (2017) use the boundaries of CDF level sets to assess risks associated with the compounding effects of river flooding and sea level rise. Each boundary consists of pairs of values of coastal water level and fluvial flow for which there is a common probability of exceedance by at least one of the two variables. This is viewed as a failure probability, which can be used to classify future hazards in a warming climate. CDF level sets are also central to the four definitions of Salvadori et al. (2016) for different forms of hazard scenarios in extreme natural phenomena. Newman et al. (2017) review decision support systems targeted at reducing natural hazard risk. CDF level sets have also been used in the judgemental estimation of bivariate distributions (Abbas et al. 2010). The joint distribution is constructed from CDF level sets judgmentally estimated with the aid of a sequence of binary choices.

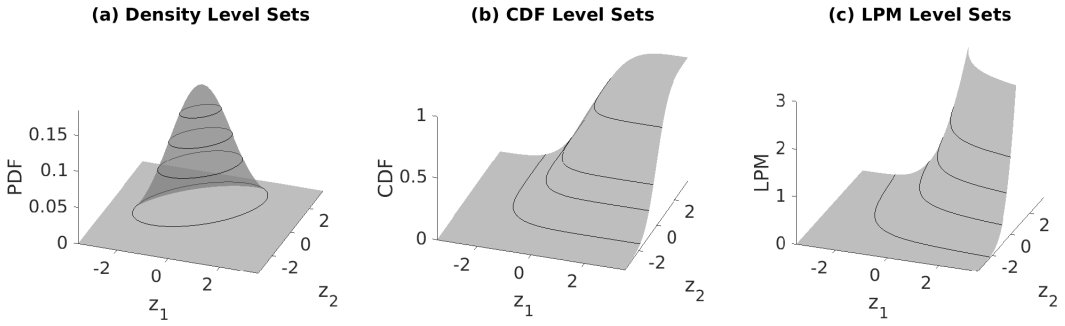
Density level sets, which are associated with probability density functions (PDFs), have been used for anomaly detection and cluster analysis (Chen et al. 2017). Steinwart et al. (2005) explain that a natural way to identify anomalies is to use density level sets because they precisely describe

the concentration of the probability distribution. An observation is deemed to be an anomaly if it lies outside the level set corresponding to a chosen value of the density, which serves as the anomaly tolerance level (Rigollet and Vert 2009). Another common use for density level sets is cluster analysis, which is used extensively for preliminary data analysis, classification and data reduction (Menardi 2016). For a chosen value of the density, the observations that lie within the same connected component of the level set can be considered to be homogeneous (see Hartigan 1975, Rinaldo and Wasserman 2010, Rinaldo et al. 2012). This form of clustering is closely related to modal clustering (see Menardi 2016), and indeed density level sets have also been used as the basis for tests of multimodality (see Müller and Sawitzki 1991).

In this paper, we aim to extend the literature on the evaluation of forecasts of multivariate distributions and their level sets. For a univariate distribution, the CRPS can be expressed as an integral over a quantile score (Laio and Tamea 2007). We generalize this to the multivariate context. In doing so, we make a number of contributions. First, we propose a theoretical framework that links the quadratic score, CRPS and MCRPS. This framework can be used to generate new scores for multivariate distributions, and we demonstrate this by developing a score based on lower partial moments (LPMs) (Price et al. 1982, Briec and Kerstens 2010, Anthonisz 2012). Second, we show that by decomposing the quadratic score, MCRPS, and our new score, we obtain new scores for density, CDF and LPM level sets, respectively. These scores encompass the existing scores for density level sets, and the full class of quantile scores. Finally, to compute the different scores, we propose a simulation-based numerical approach.

To accompany our development of the level set scores, we use a running illustrative example based on a bivariate normal distribution with zero means, unit variances, and covariance 0.5. We refer to this as the data generating process (DGP). For the DGP, Figure 1 presents 3-D illustrative examples of density, CDF and LPM level sets. These level sets will be formally defined in Section 4.

Figure 1 3-D illustration of the density, CDF and LPM level sets for the bivariate normal distribution with zero means, unit variances, and covariance 0.5.



We note that scores for vector-valued properties have been extensively studied (see, for example, Abernethy and Frongillo 2012, Frongillo and Kash 2015, Lambert 2019). Examples of vector-valued properties include the mean and covariance. That literature is distinct from our consideration of level sets, because level sets are set-valued, rather than vector-valued: a level set can be any Borel set, e.g., an ellipsoid or hyperspace, which cannot be described by a vector.

Section 2 describes notations and conventions used in this paper. In Section 3, we present our new theoretical framework for building scoring rules for multivariate distributions, and we demonstrate how this framework can be used to generate new scoring rules. Section 4 shows that the scores presented in Section 3 can be decomposed to obtain scores for different types of level sets. In Section 5, after discussing how to compute the scores, we provide support for the new scores using simulated data. Section 6 provides two practical applications of the new scores. First, we use our new score for distributions to estimate weights in a combination of distributional forecasts taken from the ECB Survey of Professional Forecasters. The new score is particularly convenient for this application as it enables the combining weights to be estimated using a quadratic optimization algorithm, which brings both theoretical and computational advantages. Second, using financial returns data, we show how CDF level sets computed with our scores provide better estimates of conditional value at risk (CoVaR), a widely-used measure of systemic risk. Section 7 concludes the paper.

2. Preliminaries

In this section, we explain some notations and conventions used in the subsequent parts of the paper. We identify a vector in $\mathbb{R}^d$ with a $d \times 1$ (column) matrix. The boldface lower-case letters

$\mathbf{z}, \mathbf{s}, \mathbf{t}, \ldots$ designate points in $\mathbb{R}^d$; we use z_j to denote each coordinate; and $\| \bullet \|$ is the Euclidean modulus. For $\mathbf{z} \in \mathbb{R}$ or $\mathbb{C}$ we write $\|\mathbf{z}\| \equiv |\mathbf{z}|$. Given any Borel set $A \subseteq \mathbb{R}^d$, ∂A denotes its topological boundary. Let λ be a Borel measure on $\mathbb{R}^d$. We say that a function $g : \mathbb{R}^d \to \mathbb{R}^m$ is integrable with respect to λ if $\int_{\mathbb{R}^d} \|g(\mathbf{z})\| \, d\lambda(\mathbf{z}) < \infty$. Let $\delta_{\mathbf{y}}$ be the Dirac delta measure with the mass at $\mathbf{y}$, and let $\mathcal{L}^d$ be the d-dimensional Lebesgue measure. The L^2 norm of a function g with respect to the measure λ is defined as $\int_{\mathbb{R}^d} \|g(\mathbf{z})\|^2 \, d\lambda(\mathbf{z})$. For two univariate functions g_1 and g_2, i.e. $m = 1$, the convolution of g_1 and g_2 is defined as $(g_1 * g_2)(\mathbf{z}) = \int_{\mathbb{R}^d} g_1(\mathbf{s}) g_2(\mathbf{z} - \mathbf{s}) d\mathbf{s}$.

We use capital letters $X, Y, V, \ldots$ for univariate random variables, and boldface capital letters $\mathbf{X}, \mathbf{Y}, \mathbf{V}, \ldots$ for multivariate random variables. As in the literature (see, for example, Gneiting and Raftery 2007), we denote by $P_{\mathbf{X}}, P_{\mathbf{Y}}, P_{\mathbf{V}} \ldots$ the probability measures of $\mathbf{X}, \mathbf{Y}, \mathbf{V}, \ldots$, and by $\mathcal{V}^d$ a convex class of probability measures. For distributions $P_{\mathbf{X}}, P_{\mathbf{Y}}, P_{\mathbf{V}}, \ldots$, denote by $F_{\mathbf{X}}, F_{\mathbf{Y}}, F_{\mathbf{V}}, \ldots$ their CDFs, and denote by $f_{\mathbf{X}}, f_{\mathbf{Y}}, f_{\mathbf{V}}, \ldots$ their generalized probability density functions.

For continuous distributions, $f_{\mathbf{X}}, f_{\mathbf{Y}}, f_{\mathbf{V}}, \ldots$ are the PDFs; for discrete distributions, $f_{\mathbf{X}}, f_{\mathbf{Y}}, f_{\mathbf{V}}, \ldots$ are the probability mass functions; and $f_{\mathbf{X}}, f_{\mathbf{Y}}, f_{\mathbf{V}}, \ldots$ can also represent the probability measures for linear combinations of continuous and discrete distributions. For simplicity, here and hereafter, we refer to $f_{\mathbf{X}}, f_{\mathbf{Y}}, f_{\mathbf{V}}, \ldots$ simply as the PDFs. We write $q_Y(\alpha) = \inf\{z \in \mathbb{R} | F_Y \geq \alpha\}$ for the α quantile of a univariate random variable Y for $\alpha \in [0, 1]$.

3. A New Framework for Scoring Rules

Let $P_{\mathbf{Y}} \in \mathcal{V}^d$ be the unknown distribution of a random variable $\mathbf{Y}$ we want to study. A key problem is to evaluate the accuracy of $P_{\mathbf{X}}$, an estimate of $P_{\mathbf{Y}}$, given a finite collection of realizations of $\mathbf{Y}$, labeled as $\{\mathbf{y}^i\}_{i=1,2,\ldots T}$. A central tool developed for this purpose is the *scoring rule*:

$$S(P_{\mathbf{X}}, \mathbf{y}) : \mathcal{V}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}.$$

It is said to be *proper* if $\mathbb{E}_{P_{\mathbf{Y}}}[S(P_{\mathbf{X}}, \bullet)]$, the expectation of $S(P_{\mathbf{X}}, \bullet)$ with respect to $P_{\mathbf{Y}}$, is minimized by $P_{\mathbf{Y}}$, and *strictly proper* if $P_{\mathbf{Y}}$ is the unique minimizer (Gneiting and Raftery 2007).

It can be intuitive to understand a score S through its divergence, which is defined as $\mathbb{E}_{P_{\mathbf{Y}}}[S(P_{\mathbf{X}}, \bullet)] - \mathbb{E}_{P_{\mathbf{Y}}}[S(P_{\mathbf{Y}}, \bullet)]$. The divergence is always non-negative, and is equal to 0 if $P_{\mathbf{X}} = P_{\mathbf{Y}}$.

Hence, it essentially defines a measure of "distance" between distributions. Scores that have the same divergence lead to the same measure of distance, so are often viewed as *equivalent*.

In the next section, we propose a class of L^2 scoring rules for distributions. In Section 3.2, we show that the new class of scoring rules encompasses the quadratic score, CRPS and MCRPS. We also show that this framework can easily be used to generate other scoring rules, and we demonstrate this by proposing a scoring rule based on lower partial moments.

3.1. A Novel Method for Constructing Scoring Rules

The quadratic score is a well-known proper scoring rule, defined as $\int_{\mathbb{R}^d} f_{\mathbf{X}}^2(\mathbf{z})\, d\mathbf{z} - 2f_{\mathbf{X}}(\mathbf{y})$, whose divergence function is the L^2 distance between PDFs, i.e. $\int_{\mathbb{R}^d} \left(f_{\mathbf{X}}(\mathbf{z}) - f_{\mathbf{Y}}(\mathbf{z})\right)^2 d\mathbf{z}$. Despite its popularity, the quadratic score has the notable weakness that sometimes it cannot distinguish the relative accuracy between misspecified estimates. For example, if $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$ have disjoint support, then the divergence is simply $\int_{\mathbb{R}^d} f_{\mathbf{X}}^2(\mathbf{z})d\mathbf{z} + \int_{\mathbb{R}^d} f_{\mathbf{Y}}^2(\mathbf{z})d\mathbf{z}$. Consider the case where $f_{\mathbf{Y}}(\mathbf{z}) = \delta_{\mathbf{0}}(\mathbf{z})$ representing all the probability mass at 0, and $f_{\mathbf{X}}(\mathbf{z}) = \delta_{\mathbf{x}_0}(\mathbf{z})$ representing all the probability mass at $\mathbf{x}_0 \neq 0$, then the divergence equals 2 regardless of the value of $\mathbf{x}_0$.

To address this issue, we generalize the divergence of the quadratic score by computing the L^2 distance between smoothed PDFs. Specifically, we consider the convolutions between the PDFs and a piecewise smooth function w, i.e. $f_{\mathbf{X}} * w$ and $f_{\mathbf{Y}} * w$. The L^2 divergence between these is given by

$$\int_{\mathbb{R}^d} \left(\left(f_{\mathbf{X}} * w\right)(\mathbf{z}) - \left(f_{\mathbf{Y}} * w\right)(\mathbf{z})\right)^2 h(\mathbf{z})\, d\mathbf{z}, \tag{1}$$

where h is a weight function needed to ensure the convergence of the integral, which is not necessarily finite with respect to the Lebesgue measure. In some cases, these convolutions have intuitive interpretations. When w is the PDF of a continuous random variable $\mathbf{W}$, such as a Gaussian random variable, the convolutions $f_{\mathbf{X}} * w$ and $f_{\mathbf{Y}} * w$ represent the PDFs of the random variables $\mathbf{X} + \mathbf{W}$ and $\mathbf{Y} + \mathbf{W}$, respectively; when w is the Heaviside function, i.e. $w(\mathbf{z}) = \prod_{j=1}^d \mathbb{1}\{z_j \geq 0\} =: \mathbb{1}\{\mathbf{z} \geq \mathbf{0}\}$, the convolutions lead to the CDFs.

Recall the example with $f_{\mathbf{Y}}(\mathbf{z}) = \delta_{\mathbf{0}}(\mathbf{z})$, $f_{\mathbf{X}}(\mathbf{z}) = \delta_{\mathbf{x}_0}(\mathbf{z})$ and $\mathbf{x}_0 \neq 0$ that we discussed earlier in this section. The weakness that we highlighted with the quadratic score can be addressed by setting

$w(\mathbf{z})$ as the Heaviside function in (1). It can be shown that our L^2 divergence in (1) is then equal to $|\mathbf{x}_0|$, which does give credit to more accurate estimates. In addition, the convolutions $f_{\mathbf{X}} * w$ and $f_{\mathbf{Y}} * w$ in (1) are piecewise smooth for any distribution if w is piecewise smooth. This is both theoretically and numerically appealing.

We can see that (1) is nonnegative and is equal to 0 if $f_{\mathbf{X}} = f_{\mathbf{Y}}$ $\mathcal{L}^d$-a.e.. Hence, scores with this divergence function are by definition proper. Moreover, (1) is equal to 0 if and only if $(f_{\mathbf{X}} * w)(\mathbf{z})h(\mathbf{z}) = (f_{\mathbf{Y}} * w)(\mathbf{z})h(\mathbf{z})$ $\mathcal{L}^d$-a.e.. In this case, if h is nonzero $\mathcal{L}^d$-a.e and $f_{\mathbf{Y}} * w$ uniquely determines any $P_{\mathbf{Y}} \in \mathcal{V}^d$ (i.e., $f_{\mathbf{X}} * w = f_{\mathbf{Y}} * w$ $\mathcal{L}^d$-a.e. implies $f_{\mathbf{X}} = f_{\mathbf{Y}}$ $\mathcal{L}^d$-a.e.), then the scores associated with (1) are strictly proper.

Next, we introduce our L^2 scoring rule whose divergence is precisely (1). The idea is to view a realization $\mathbf{y}$ as a Dirac delta measure located at $\mathbf{y}$. The convolution between this Dirac delta measure and w leads to $w(\mathbf{z}-\mathbf{y})$. Thus, our L^2 scoring rule can be written as

$$S(P_{\mathbf{X}}, \mathbf{y}; w, h) = \int_{\mathbb{R}^d} \left((f_{\mathbf{X}} * w)(\mathbf{z}) - w(\mathbf{z}-\mathbf{y}) \right)^2 h(\mathbf{z})\, d\mathbf{z}. \tag{2}$$

By expanding the integrand in (2), we can observe that $\int_{\mathbb{R}^d} w^2(\mathbf{z}-\mathbf{y})h(\mathbf{z})\, d\mathbf{z}$ only depends on $\mathbf{y}$, hence has the same value regardless of $P_{\mathbf{X}}$. Therefore, omitting this term has no impact on the L^2 divergence in (1). Hence, we can define another scoring rule that is equivalent to (2) as

$$S'(P_{\mathbf{X}}, \mathbf{y}; w, h) = \int_{\mathbb{R}^d} (f_{\mathbf{X}} * w)^2(\mathbf{z})h(\mathbf{z})\, d\mathbf{z} - 2\int_{\mathbb{R}^d} (f_{\mathbf{X}} * w)(\mathbf{z})w(\mathbf{z}-\mathbf{y})h(\mathbf{z})\, d\mathbf{z}. \tag{3}$$

We summarize the results in the following Theorem. The proof is postponed to the appendix.

THEOREM 1. *Under the conditions in Assumption 1 in the appendix, the L^2 scoring rules in* (2) *and* (3) *are proper. They are strictly proper if h is nonzero $\mathcal{L}^d$-a.e. and $f_{\mathbf{y}} * w$ uniquely characterizes any $P_{\mathbf{Y}} \in \mathcal{V}^d$. Both of the L^2 scoring rules in (2) and (3) have the divergence function given in* (1).

Recall that our motivation was to use piecewise smooth functions w to smooth the PDFs via a convolution to improve upon the quadratic score. From a purely theoretical perspective, the requirement that w is piecewise smooth can be relaxed. For any local Borel measure w and weight function h, if the convolutions are well-defined and the integrals in (1) and (2) are finite, then (2) is

a proper scoring rule. Similarly, if the convolutions are well-defined and the integrals in (1) and (3) are finite, then (3) is a proper scoring rule. We summarize the technical conditions in Assumption 1. The weight function h can be anything, but the most convenient form is a PDF, because PDFs decay rapidly, and so usually guarantee the convergence of (1). In view of this, we set h to be a PDF in our empirical studies. The function h can also be used to emphasize regions of interest. Gneiting and Ranjan (2011) put more weight on the tails for inflation data, and Grushka-Cockayne et al. (2017) consider non-constant weight functions as a way to align the score with the cost function in a given business context. Moreover, as we show in Section 5.1, choosing h to be a PDF is also convenient for numerical computation of the scores.

The proposed L^2 score is not a single score but a class of scores. In practice, we want to consider specific w and h so that $\mathcal{V}^d$ can be sufficiently general that important distributions are not excluded. For this purpose, we reiterate our recommendation that w is a piecewise smooth function and h is a PDF. With these choices, the L^2 scores can be used with $\mathcal{V}^d$ that covers almost all standard distributions, e.g. the distributions with bounded PDFs.

In the literature, there are studies regarding *local scoring rules,* which depend solely on the PDF or the derivatives of the PDF purely at the realization $\mathbf{y}$, where a well-known example is the log score $\log(f_{\mathbf{X}}(\mathbf{y}))$ (see, for example, Ehm et al. 2012, Parry et al. 2012). Because the L^2 scoring rules involve integration over $\mathbb{R}^d$, they are fundamentally different from local scoring rules. Hence, the proposed L^2 scoring rules do not cover the local scoring rules, and are not exhaustive. The non-locality of the L^2 scoring rules is important for our derivation of the level set scores in Section 4, because level sets are not local statistical objects in the sense that their specifications require information regarding the entire domain. Thus, local scoring functions are not within the scope of this study, and so we do not discuss them further.

Remark 1. Using the Plancherel identity, we can also construct L^2 scoring rules based on characteristic functions of distributions. In particular, this approach can lead to the well-known energy score. We discuss this further in the appendix.

3.2. Particular Cases of L^2 Scoring Rules

In this section, we first demonstrate the generality of our proposed L^2 scoring rules by showing that the quadratic score, CRPS, and MCRPS emerge as special cases of the general framework laid down in Section 3.1. We then show how our framework naturally generates other scoring rules by developing a score based on lower partial moments. We emphasize that, throughout the examples in this section, the key issue is the specification of w.

3.2.1. Quadratic Score Let us first note that the convolution between any PDF $f_{\mathbf{X}}(\mathbf{z})$ and $\delta_{\mathbf{0}}(\mathbf{z})$ is simply $f_{\mathbf{X}}(\mathbf{z})$ itself. In view of this, setting $w(\mathbf{z}) = \delta_{\mathbf{0}}(\mathbf{z})$ in (3) gives the following score,

$$\mathrm{DQS}^{'}(P_{\mathbf{X}}, \mathbf{y}; h) = S^{'}(P_{\mathbf{X}}, \mathbf{y}; \delta_{\mathbf{0}}, h) = \int_{\mathbb{R}^d} f_{\mathbf{X}}^2(\mathbf{z}) h(\mathbf{z})\, d\mathbf{z} - 2 f_{\mathbf{X}}(\mathbf{y}) h(\mathbf{y}). \tag{4}$$

When $h \equiv 1$, the score in (4) reduces to the well-known quadratic score. For simplicity, from now on, we refer to (4) as the quadratic score. Note that, for this case, we have used (3) rather than (2) because the latter involves the multiplication of two Dirac delta masses and hence is not well-defined. The quadratic score can be used with discrete or continuous distributions that have finite PDFs.

3.2.2. CRPS & MCRPS Recall that $u(\mathbf{z}) = \prod_{j=1}^{d} \mathbb{1}\{z_j \geq 0\} =: \mathbb{1}\{\mathbf{z} \geq \mathbf{0}\}$ is the Heaviside function. As discussed in Section 3.1, the convolution of a PDF $f_{\mathbf{X}}$ and u gives the CDF $F_{\mathbf{X}}$. In fact, we can write

$$(f_{\mathbf{X}} * u)(\mathbf{z}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(s_1, \ldots, s_d) \prod_{j=1}^{d} \mathbb{1}\{z_j \geq s_j\}\, ds_1 \ldots ds_d =: F_{\mathbf{X}}(z_1, \ldots, z_d).$$

By Theorem 1, and if we use $w(\mathbf{z}) = u(\mathbf{z})$ in (2), we obtain the following proper score:

$$\mathrm{MCRPS}(P_{\mathbf{X}}, \mathbf{y}; h) = S(P_{\mathbf{X}}, \mathbf{y}; u, h) = \int_{\mathbb{R}^d} \left(F_{\mathbf{X}}(\mathbf{z}) - \mathbb{1}\left\{\mathbf{z} \geq \mathbf{y}\right\}\right)^2 h(\mathbf{z})\, d\mathbf{z}. \tag{5}$$

This score is precisely the MCRPS considered by Gneiting and Raftery (2007). For univariate distributions, (5) is simply the well-known CRPS for $h \equiv 1$, and the threshold weighted CRPS for a general function h (Gneiting and Ranjan 2011, Grushka-Cockayne et al. 2017).

Following (3), we obtain an equivalent expression for the MCRPS, given by

$$\mathrm{MCRPS}^{'}(P_{\mathbf{X}}, \mathbf{y}; h) = \int_{\mathbb{R}^d} F_{\mathbf{X}}^2(\mathbf{z}) h(\mathbf{z})\, d\mathbf{z} - 2 \int_{\mathbb{R}^d} F_{\mathbf{X}}(\mathbf{z}) \mathbb{1}\left\{\mathbf{z} \geq \mathbf{y}\right\} h(\mathbf{z})\, d\mathbf{z}. \tag{6}$$

This expression will be useful in our consideration of level sets in Section 4. Both MCRPS and MCRPS$'$ scores are strictly proper if h is nonzero $\mathcal{L}^d$-a.e., because the CDF uniquely characterizes a distribution. If h is a PDF, both MCRPS and MCRPS$'$ scores can be used with all distributions.

3.2.3. A New Score Based on Lower Partial Moments Let $u^{*k}(\mathbf{z}) \equiv \underbrace{u * u * \cdots * u}_{k \text{ times}}(\mathbf{z}) = \prod_{j=1}^{d} \frac{1}{k!} z_j^k \mathbb{1}\{z_j \geq 0\}$ denote the k^{th} convolution power of the Heaviside function, where $k = 1, 2, 3....$ The convolution $f_{\mathbf{X}} * u^{*k}$ gives the LPM of order k, which is given by

$$f_{\mathbf{X}} * u^{*k}(\mathbf{z}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(s_1, \ldots, s_d) \prod_{j=1}^{d} \frac{1}{k!}(z_j - s_j)^k \mathbb{1}\{z_j \geq s_j\} \, ds_1 \ldots ds_d := \text{LPM}_{\mathbf{X},k}(\mathbf{z}). \quad (7)$$

The LPM for univariate distributions has been widely considered for systemic risk (Price et al. 1982), asset pricing (Anthonisz 2012), and portfolio management (Briec and Kerstens 2010). However, to the best of our knowledge, the LPM for multivariate distributions has not been considered in the literature.

By Theorem 1, and if we use $w(\mathbf{z}) = u^{*k}(\mathbf{z})$ in (2), we obtain a new proper scoring rule based on the LPM of order k:

$$\text{LPMS}(P_{\mathbf{X}}, \mathbf{y}; k, h) = S(P_{\mathbf{X}}, \mathbf{y}; u^{*k}, h) = \int_{\mathbb{R}^d} \left(\text{LPM}_{\mathbf{X},k}(\mathbf{z}) - \prod_{j=1}^{d} \frac{1}{k!}(z_j - y_j)^k \mathbb{1}\{z_j \geq y_j\} \right)^2 h(\mathbf{z}) \, d\mathbf{z}.$$

We may infer from (3), the following equivalent score:

$$\text{LPMS}'(P_{\mathbf{X}}, \mathbf{y}; k, h) = \int_{\mathbb{R}^d} \text{LPM}_{\mathbf{X},k}^2(\mathbf{z}) h(\mathbf{z}) \, d\mathbf{z} - 2 \int_{\mathbb{R}^d} \text{LPM}_{\mathbf{X},k}(\mathbf{z}) \prod_{j=1}^{d} \frac{1}{k!}(z_j - y_j)^k \mathbb{1}\{z_j \geq y_j\} h(\mathbf{z}) \, d\mathbf{z}. \quad (8)$$

This form of the score will be useful in our consideration of level sets in Section 4. Both LPMS and LPMS$'$ scores are strictly proper if h is nonzero $\mathcal{L}^d$-a.e., because the lower partial moment uniquely characterizes a distribution. A proof is presented in the appendix. If h is a PDF with bounded support or exponential decay, both LPMS and LPMS$'$ scores can be used with distributions that have finite PDFs.

4. New Scoring Functions for Level Sets

Sometimes we are more interested in a specific region of a distribution than the entire domain. For example, tails of a distribution are often of great interest to various applications, with

quantiles widely studied as an important risk measure (see, for example, Jose and Winkler 2009, Grushka-Cockayne et al. 2017). Quantiles have also been considered in the context of model averaging (Lichtendahl Jr et al. 2013).

In this paper we consider level sets, which can be viewed as multivariate generalizations of quantiles (Abbas et al. 2010, Cousin and Di Bernardino 2013). Let $g : \mathbb{R}^d \to \mathbb{R}$ be a univariate function, and $\alpha \in \mathbb{R}$. We define the α level set of g as

$$L(g;\alpha) := \{\mathbf{z} \in \mathbb{R}^d : g(\mathbf{z}) \geq \alpha\}. \tag{9}$$

For simplicity, we denote $\{\mathbf{z} \in \mathbb{R}^d : g(\mathbf{z}) \geq \alpha\}$ as $\{g \geq \alpha\}$ in the rest of the paper. In the literature, sometimes $L(g;\alpha)$ is referred to as the upper level set, and $\partial\{L(g;\alpha)\}$ as the level set (see, for example, Chen et al. 2017). Our terminology is consistent with Cadre (2006) and Singh et al. (2009).

More specifically, in this section, we consider $g = f_{\mathbf{Y}} * w$, as explained in Section 3, and we derive consistent scoring functions for the level set $L(f_{\mathbf{Y}} * w;\alpha) := \{f_{\mathbf{Y}} * w \geq \alpha\}$. As a special case, for a univariate random variable Y, and if w is the Heaviside function, the α level set $L(F_Y;\alpha)$ can be identified with the α quantile as follows: $q_Y(\alpha) \longleftrightarrow [q_Y(\alpha), \infty) = L(F_Y;\alpha)$.

Similar to the spirit of a proper scoring rule, we can define scoring functions for level sets. Let $l(A, \mathbf{y}; w, h)$ be an $\mathbb{R}$-valued function whose arguments consist of a Borel set $A \subseteq \mathbb{R}^d$, an estimate of $L(f_{\mathbf{Y}} * w;\alpha)$, and a realization $\mathbf{y} \in \mathbb{R}^d$. We say that l is a consistent scoring function of the level sets if $\mathbb{E}_{P_{\mathbf{Y}}}[l(A, \bullet; w)]$ is minimized by $L(f_{\mathbf{Y}} * w;\alpha)$, and strictly consistent if $L(f_{\mathbf{Y}} * w;\alpha)$ is the only minimizer. In this paper, to distinguish scores for distributions and level sets, we adopt the convention that the scores for level sets are termed (consistent) scoring functions (see, for example, Fissler et al. 2021). To the best of our knowledge, scoring functions for general level sets have not been studied, with the only exceptions being a score for density level sets $L(f_{\mathbf{Y}};\alpha)$ and scores for quantiles of univariate distributions.

In this section, we propose a systematic approach for constructing scoring functions for level sets. Our inspiration stems from the well-known result that the CRPS can be decomposed as the integral of the quantile scores (see, for example, Gneiting and Raftery 2007, Grushka-Cockayne et al. 2017):

$$\mathrm{CRPS}(P_X, y) := \int_{-\infty}^{\infty} \Big(F_X(z) - \mathbb{1}\{z \geq y\}\Big)^2 dz = 2\int_0^1 \Big(\alpha - \mathbb{1}\{y < q_X(\alpha)\}\Big)\Big(y - q_X(\alpha)\Big)\, d\alpha, \tag{10}$$

where the integrand in the last integral is the well-known quantile score, also known as the "pinball loss" or "tick-loss". The equivalence can be established via a change of variables (see, for example, Laio and Tamea 2007). In spite of its simplicity, this algebraic manipulation does not extend in a straightforward way to other types of scoring functions, and so, in the multivariate case, it is not trivial to obtain consistent scoring functions for level sets from proper scoring rules.

Our idea is to use the "layer cake representation" to decompose the scoring rule $S^{'}(P_{\mathbf{X}}, \mathbf{y}; w, h)$ in (3) into an integral of scoring functions for level sets of $f_{\mathbf{Y}} * w$. This provides a unified approach for constructing scoring functions for different types of level sets, including the scores for density level sets and quantiles. In addition, our approach provides insight into the relationship between distributions and their level sets.

In the next section, we formally describe our approach for constructing scoring functions for level sets. We then consider specific examples in Section 4.2.

4.1. Scoring Functions for Level Sets

The key tool for our further developments is the elementary but useful *layer cake representation*. Its proof can be found in Lieb and Loss (2001), and the name "layer cake" refers to the level set structure. Roughly speaking, the layer cake representation decomposes the L^2 scoring rules in (3) into an integral of "layers", which are precisely the level set scores. Theorem 2 summarizes the theoretical results. The derivation and proof are postponed to the appendix.

THEOREM 2. *Let $\mathcal{V}^d$, h, and w satisfy Assumption 1 in the appendix, and w is nonnegative. Let $\alpha > 0$, and let $A \subseteq \mathbb{R}^d$ be an arbitrary Borel set. Suppose $(S')^{\Gamma}(A, \bullet; w, h, \alpha)$ is finite $\mathcal{L}^d$-a.e. and $(S')^{\Gamma}(A, \bullet; w, h, \alpha)\, f_{\mathbf{Y}}(\bullet)$ is integrable. Then the following expression defines a consistent scoring function for the α level set $L(f_{\mathbf{Y}} * w; \alpha)$:*

$$(S^{'})^{\Gamma}\Big(A, \mathbf{y}; w, h, \alpha\Big) = \int_{\mathbb{R}^d} (\alpha - w(\mathbf{z} - \mathbf{y}))\, \mathbb{1}\{\mathbf{z} \in A\}\, h(\mathbf{z})\, d\mathbf{z}. \tag{11}$$

If h is nonzero $\mathcal{L}^d$-a.e. and $L(f_{\mathbf{Y}} * w; \alpha)$ equals the closure of $\{f_{\mathbf{Y}} * w > \alpha\}$, it is strictly consistent. Moreover, with $A = L(f_{\mathbf{X}} * w; \alpha)$, the integral of (11) with respect to α is precisely half of the L^2 score in (3), i.e., $\frac{1}{2} S^{'}(P_{\mathbf{X}}, \mathbf{y}; w, h) = \int_0^{\infty} (S^{'})^{\Gamma}\Big(L(f_{\mathbf{X}} * w; \alpha), \mathbf{y}; w, h, \alpha\Big)\, d\alpha$.

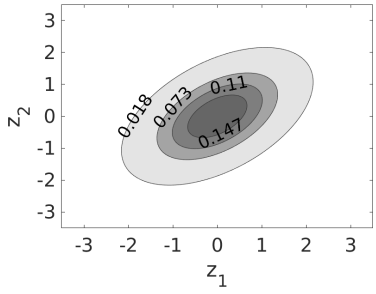
The superscript $^{\Gamma}$ emphasizes that $(S')^{\Gamma}$ is a scoring function for a level set, which is derived from the L^2 scoring rule $S'(P_{\mathbf{X}}, \mathbf{y}; w, h)$ in (3). Theorem 2 provides important insight. It generalizes the decomposition of the CRPS in (10) into multivariate settings, namely, the L^2 scoring rule for $P_{\mathbf{Y}}$ is equal to two times the integral with respect to α of the scoring functions for the level sets $L(f_{\mathbf{Y}} * w; \alpha)$. Since level sets represent different regions, the level set scores effectively enable us to focus on a specific region by considering specific values of α.

4.2. Particular Cases of Scoring Functions for Level Sets

In this section, we first show how our approach allows us to derive the scoring functions for density level sets (Hartigan 1987, Chen et al. 2017). We then develop two new scores: the scoring functions for CDF level sets $L(F_{\mathbf{Y}}; \alpha)$ and LPM level sets $L(\mathrm{LPM}_{\mathbf{Y},k}; \alpha)$. In particular, we show that, in the univariate context, the scoring functions for CDF level sets induce the full class of quantile scores studied by Gneiting (2011) and Komunjer (2005).

4.2.1. Density Level Set Scores Recall that we consider $w(\mathbf{z}) \equiv \delta_{\mathbf{0}}(\mathbf{z})$ for the quadratic score DQS' in Section 3.2.1. For this w, $f_{\mathbf{Y}} * w$ reduces to the PDF $f_{\mathbf{Y}}$, and by (9), the α density level set is defined as $L(f_{\mathbf{Y}}; \alpha) = \{f_{\mathbf{Y}} \geq \alpha\}$. Figure 2 shows the 2-D projections of the density level sets of the DGP that were shown in 3-D in Figure 1(a). The numerical values within the plot indicate the value α of the PDF $f_{\mathbf{Y}}$ at points on the boundary of the α density level set.

Figure 2 2-D projections of the density level sets shown in Figure 1 (a). The numerical values within the plot indicate the value of α for each level set. Darker coloring corresponds to a higher value of α.



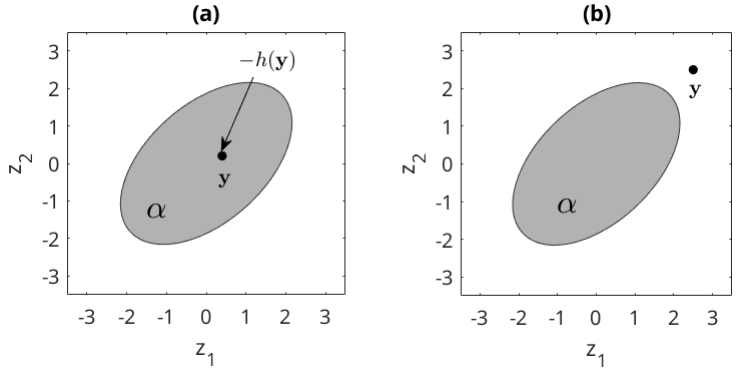
By Theorem 2, and using $w(\mathbf{z}) \equiv \delta_{\mathbf{0}}(\mathbf{z})$ in (11), we obtain the following consistent scoring function for the α density level set,

$$(\mathrm{DQS}')^{\Gamma}(A, \mathbf{y}; h, \alpha) = \int_{\mathbb{R}^d} (\alpha - \delta_{\mathbf{0}}(\mathbf{z} - \mathbf{y}))\, \mathbb{1}\{\mathbf{z} \in A\}\, h(\mathbf{z})\, d\mathbf{z}$$

$$= \alpha \int_{\mathbb{R}^d} \mathbb{1}\{\mathbf{z} \in A\} h(\mathbf{z}) d\mathbf{z} - \mathbb{1}\{\mathbf{y} \in A\} h(\mathbf{y}). \tag{12}$$

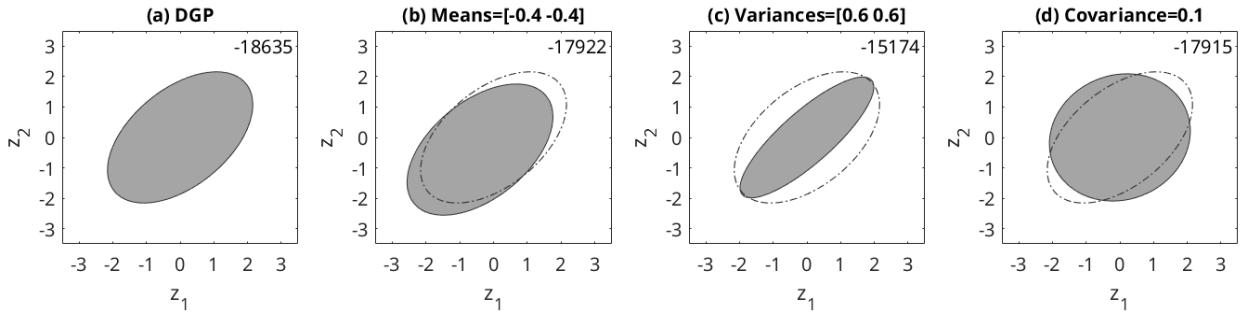
This is precisely the scoring function for density level sets, also known as the *excess mass score* (see, for example, Hartigan 1987, Chen et al. 2017). The score (12) has an intuitive graphical interpretation, which we illustrate in Figure 3.

Figure 3 Graphical illustration of the density level set score in (12). The shaded areas represent the α predictive density level set and $\mathbf{y}$ represents a realization. In (a), $\mathbf{y}$ is inside the level set, so the density level set score is the Borel measure (induced by h) for the shaded region multiplied by α, minus $h(\mathbf{y})$; in (b), $\mathbf{y}$ is outside the level set, so $\mathbf{y}$ has no contribution to the score, i.e., the density level set score is simply the Borel measure for the shaded region multiplied by α.



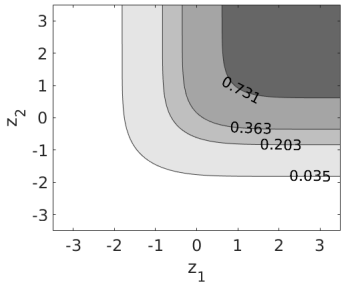
We note that competing density forecasts can be compared in terms of their accuracy for different regions of the density by evaluating the level sets for different choices of α. Focusing on $\alpha = 0.018$, Figure 4 compares the density level set of the DGP and that of three misspecified distributions, along with the mean of the density level set scores computed in our simulation study of Section 5.2. The DGP receives the lowest mean score, which shows the consistency of the score.

Figure 4 For $\alpha = 0.018$, comparison of the density level set of the DGP in (a) and the following three misspecified distributions: (b) misspecified means $[-0.4, -0.4]$; (c) misspecified variances $[0.6, 0.6]$; and (d) misspecified covariance 0.5. The dotted lines in (b), (c) and (d) indicate the density level set of the DGP in (a). The numerical value at the top right corner of each plot is the average score ($\times 10^6$) computed in the simulation study of Section 5.2.



4.2.2. New CDF Level Set Score Recall that we consider $w(\mathbf{z}) = u(\mathbf{z}) = \mathbb{1}\{\mathbf{z} \geq \mathbf{0}\}$ in Section 3.2.2 to derive MCRPS and MCRPS$'$. For this w, $f_{\mathbf{Y}} * w$ reduces to the CDF $F_{\mathbf{Y}}$, and by (9), the α CDF level set is defined as $L(F_{\mathbf{Y}}; \alpha) = \{F_{\mathbf{Y}} \geq \alpha\}$. Figure 5 shows the 2-D projections of the CDF level sets of the DGP that were shown in 3D in Figure 1(b). The numerical values within the plot indicate the value α of the CDF $F_{\mathbf{Y}}$ at points on the boundary of the α CDF level set, i.e. α is a probability.

Figure 5 2-D projections of the CDF level sets shown in Figure 1 (b). The numerical values within the plot indicate the value of α for each level set. Darker coloring corresponds to a higher value of α.

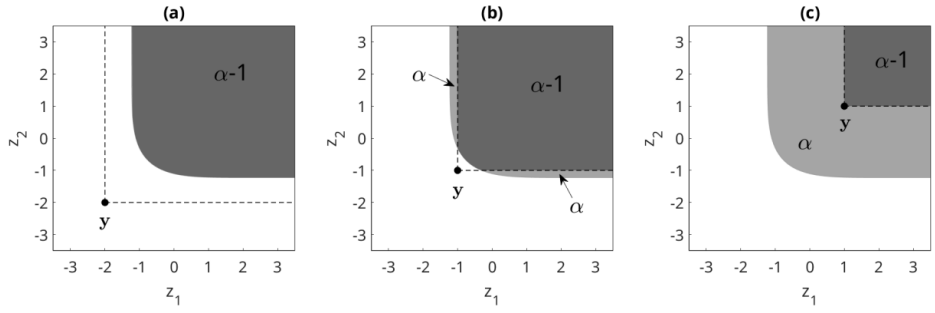


By Theorem 2, and using $w(\mathbf{z}) = u(\mathbf{z})$ in (11), we obtain a consistent scoring function for the α CDF level set,

$$(\text{MCRPS}')^{\Gamma}(A, \mathbf{y}; h, \alpha) = \int_{\mathbb{R}^d} \left(\alpha - \mathbb{1}\{\mathbf{z} \geq \mathbf{y}\}\right) \mathbb{1}\{\mathbf{z} \in A\} h(\mathbf{z})\, d\mathbf{z}. \tag{13}$$

The score (13) also has an intuitive graphical interpretation, which we illustrate in Figure 6.

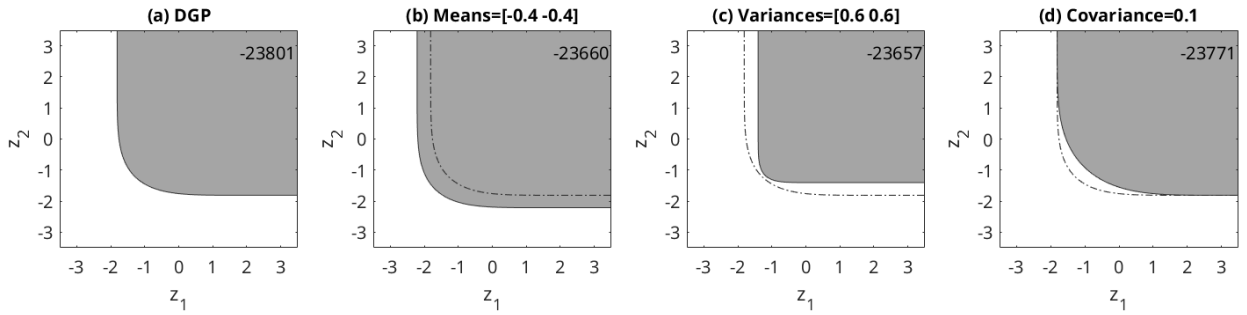
Figure 6 Graphical illustration of the CDF level set score in (13). The curves represent the α CDF level set and $\mathbf{y}$ represents a realization, and α or $\alpha - 1$ is the weight for the corresponding shaded region. Three scenarios are presented depending on the position of $\mathbf{y}$ relative to the level set. The CDF level set score is the weighted sum of the Borel measures (induced by h) for the shaded regions.



Focusing on $\alpha = 0.035$, Figure 7 compares the CDF level set of the DGP and that of three misspecified distributions, along with the average scores computed in our simulation study of

Section 5.2. The DGP has the lowest average score, showing the consistency of the CDF level set score.

Figure 7 For $\alpha = 0.035$, comparison of the CDF level set of the DGP in (a) and the following three misspecified distributions: (b) misspecified means $[-0.4, -0.4]$; (c) misspecified variances $[0.6, 0.6]$; and (d) misspecified covariance 0.5. The dotted lines in (b), (c) and (d) indicate the CDF level set of the DGP in (a). The numerical value at the top right corner of each plot is the average score ($\times 10^5$) computed in the simulation study of Section 5.2.



REMARK 2. When $d = 1$, (13) is equivalent to the full class of quantile scores. To see this, let us first identify a quantile estimate $q(\alpha)$ with a Borel set $A = [q(\alpha), \infty)$ and then add to (13) the term $(1-\alpha)\int_{-\infty}^{\infty} \mathbb{1}\{z \geq y\} h(z)\, dz$, which depends only on y, as follows:

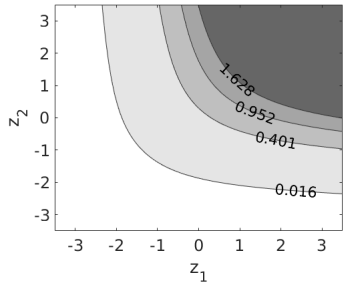
$$
\begin{aligned}
&\int_{-\infty}^{\infty} \big(\alpha - \mathbb{1}\{z \geq y\}\big)\, \mathbb{1}\{z \geq q(\alpha)\} h(\mathbf{z})\, d\mathbf{z} + (1-\alpha) \int_{-\infty}^{\infty} \mathbb{1}\{z \geq y\} h(\mathbf{z})\, d\mathbf{z} \\
&= \alpha \int_{-\infty}^{\infty} \big(\mathbb{1}\{z \geq q(\alpha)\} - \mathbb{1}\{z \geq y\}\big)\, h(\mathbf{z})\, d\mathbf{z} + \int_{-\infty}^{\infty} \big(1 - \mathbb{1}\{z \geq q(\alpha)\}\big)\, \mathbb{1}\{z \geq y\} h(\mathbf{z})\, d\mathbf{z} \\
&= \alpha \int_{q(\alpha)}^{y} \lambda\,(d\mathbf{z}) - \mathbb{1}\{y < q(\alpha)\} \int_{q(\alpha)}^{y} h(\mathbf{z})\, d\mathbf{z} \\
&= \big(\alpha - \mathbb{1}\{y < q(\alpha)\}\big)\Big(H(y) - H\big(q(\alpha)\big)\Big),
\end{aligned}
$$

where H is an anti-derivative of h, hence is non-decreasing (since h is non-negative). The expression in the final line is the full class of quantile scores (see, for example, Gneiting 2011, Komunjer 2005).

4.2.3. New LPM Level Set Score Recall that we consider $w(\mathbf{z}) = u^{*k}(\mathbf{z}) = \prod_{j=1}^{d} \frac{1}{k!} z_j^k \mathbb{1}\{z_j \geq 0\}$ in Section 3.2.3 to derive the novel L^2 scoring rules LPMS and LPMS$'$. For this w, $f_{\mathbf{Y}} * w$ reduces to the k-th lower partial moment function of the distribution $P_{\mathbf{Y}}$, and by (9) the α level set of the k-th LPM is defined by $L(\mathrm{LPM}_{\mathbf{Y},k}; \alpha) = \{\mathrm{LPM}_{\mathbf{Y},k} \geq \alpha\}$.

Figure 8 shows the 2-D projections of the LPM level sets for $k=1$ of the DGP that were shown in 3D in Figure 1(c). The numerical values within the plot indicate the value α of the $\text{LPM}_{\mathbf{Y},1}$ at points on the boundary of the α level set of the $\text{LPM}_{\mathbf{Y},1}$.

Figure 8 2-D projections of the LPM level sets for $k=1$ shown in Figure 1 (c). The numerical values within the plot indicate the value of α for each level set. Darker coloring corresponds to a higher value of α.

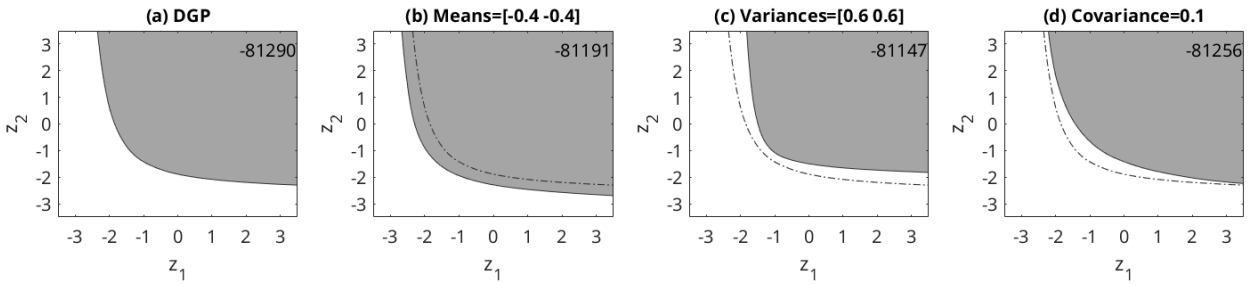


By Theorem 2, and using $w(\mathbf{z})=u^{*k}(\mathbf{z})$ in (11), we obtain a consistent scoring function for the α level set of the k-th LPM:

$$(\text{LPMS}')^{\Gamma}(A,\mathbf{y};h,\alpha,k) := \int_{\mathbb{R}^d} \left(\alpha - \mathbb{1}\{\mathbf{z}\geq\mathbf{y}\}\prod_{j=1}^{d}\frac{(z_j-y_j)^k}{k!}h(\mathbf{z})\right)\mathbb{1}\{\boldsymbol{z}\in A\}\,d\mathbf{z}. \tag{14}$$

We note that, unlike the density and CDF level set scores, the LPM level set score does not have an intuitive graphical interpretation. For $k=1$ and $\alpha=0.016$, Figure 9 compares the LPM level set of the DGP and that of three misspecified distributions, along with the average scores computed in the simulation study of Section 5.2. The DGP receives the lowest average score, showing the consistency of the LPM level set score.

Figure 9 For $k=1$ and $\alpha=0.016$, comparison of the LPM level set of the DGP in (a) and the following three misspecified distributions: (b) misspecified means $[-0.4,-0.4]$; (c) misspecified variances $[0.6,0.6]$; and (d) misspecified covariance 0.5. The dotted lines in (b), (c) and (d) indicate the LPM level set of the DGP in (a). The numerical value at the top right corner of each plot is the average score ($\times 10^5$) in the simulation study in Section 5.2.



REMARK 3. We remark that our level set scores are related to the exhaustive scores for the "set-valued functionals" studied by Fissler et al. (2021). They utilize a "bottom-up" approach that constructs the scores from identification functions, while we consider a "top-down" approach, where the level set scores are obtained through the decomposition of the L^2 scores. In some cases, similar results can be obtained from the two approaches. For example, the CDF level set can be identified with a Vorob'ev quantile of the random set $\{\mathbf{z} \geq \mathbf{Y}\}$. As a result, for this particular example, our CDF level set score in (13) coincides with the Vorob'ev quantile scores in (5.2) in Fissler et al. (2021).

5. Numerical Illustration

In this section, we show how to compute the different scores we have previously introduced using a simulation-based method. We also perform a simulation study to demonstrate the use of these scores. We implement the L^2 scoring rules for multivariate distributions, which we presented in Section 3. These are the DQS' in (4), the MCRPS' in (6), and the LPMS' in (8). For each L^2 scoring rule, we also implement the corresponding level set scoring functions described in Section 4: the $(\text{DQS}')^\Gamma$ for density level sets in (12), the $(\text{MCRPS}')^\Gamma$ for CDF level sets in (13), and the $(\text{LPMS}')^\Gamma$ for LPM level sets in (14) for $k = 1$. Recall that all these L^2 scoring rules and level set scoring functions are particular cases of our score S' in (3) and $(S')^\Gamma$ in (11), respectively.

5.1. Computation of Scores

Since the computation of the L^2 scoring rule S' in (3) and the level set score $(S')^\Gamma$ in (11) essentially involves integrations over $\mathbb{R}^d$, one can adopt any numerical integration approach. If we assume h is a probability measure, the integrals can be simply computed as expectations with respect to the distribution induced by h. More specifically, let $\{\mathbf{z}^i\}_{i=1}^M$ be a random sample drawn from the distribution with PDF h. The L^2 scoring rule in (3) can be computed as

$$\overline{S'}(P_{\mathbf{X}}, \mathbf{y}; w, h) = \frac{1}{M} \sum_{i=1}^{M} \left\{ \left(f_{\mathbf{X}} * w\right)^2 (\mathbf{z}^i) - 2 \left(f_{\mathbf{X}} * w\right)(\mathbf{z}^i) w(\mathbf{z}^i - \mathbf{y}) \right\}, \tag{15}$$

and the level set score in (11) can be computed as[1]

$$\overline{(S')^{\Gamma}}\{A,\mathbf{y};w,h,\alpha\}=\frac{1}{M}\sum_{i=1}^{M}\left\{\left(\alpha-w(\mathbf{z}^i-\mathbf{y})\right)\mathbb{1}\left\{\mathbf{z}^i\in A\right\}\right\}. \tag{16}$$

Following the Central Limit Theorem, the error of this numerical approach decays at the rate of $O\left(M^{-\frac{1}{2}}\right)$ (see Caflisch 1998, Robert and Casella 2013). It is quick to simulate random values using methods such as MCMC, and the numerical integration boils down to matrix operation, which can be computed swiftly using parallel computing. In our numerical studies, we primarily chose h to be a uniform distribution on the bounded region $[a,b]^d$.

When comparing different models, it is the relative difference between the scores that is important, not the absolute values. In reporting numerical results in this paper, we set one model as the reference, and then subtract the score of this reference model from the score for each other model. This leads to a clearer presentation of the results, and has no impact on the ranking of the performance of the models.

5.2. Simulation Study

We use the scores to compare the fit of candidate distributions and their level sets to simulated data. We generate $T=2\times10^5$ observations from the DGP of our running example, i.e., a bivariate normal distribution with zero means, unit variances, and covariance 0.5.

For CDF level sets, α has range $[0,1]$ for all distributions. For density and LPM level sets, however, α does not have a universal range for all distributions. To select a set of values of α for each type of level set, we first recorded the value of α for the level set on which each of the T simulated observations was located. We then chose α to be the $0.1, 0.2, ..., 0.9$ quantiles of these values. The resultant values of α for each type of level set are presented in the second rows of Tables 1-5.

We compared 13 candidate distributions in terms of their fit to the simulated observations. All were bivariate normal. One had no misspecification, i.e. it was the DGP; four had the following misspecified means [-2,-2], [-0.4,0.4], [0.4,0.4], [2,2], and no misspecification in the variances and

[1] For the quadratic score $\mathrm{DQS}'(P_{\mathbf{X}},\mathbf{y};h)$ in (4) and the density level set score in (12), the terms $f(\mathbf{y})h(\mathbf{y})$ and $\mathbb{1}\{\mathbf{y}\in\mathbf{A}\}h(\mathbf{y})$ do not require numerical integration.

covariance; four had the following misspecified variances [0.6,0.6], [0.8,0.8], [1.2,1.2], [1.4,1.4], and no misspecification in the means and covariance; four had the following misspecified covariance -0.5, 0.1, 0.7, and 0.9, and no misspecification in the means and variances. We calculated the scores using (15) and (16), as described in Section 5.1, with h chosen to be the PDF of a uniform distribution on $[-3,3]^2$, and $M = 10^5$ values sampled from this distribution. For each of the T observations, we computed the scoring function for each type of level set and the corresponding L^2 scoring rule. We chose the DGP as the reference, and subtracted its scores from the scores for each candidate. For each method, lower values of this score difference are preferable.

Table 1 presents the score differences for the density level set score $(\text{DQS}')^{\Gamma}$ and its corresponding L^2 scoring rule DQS'; Table 2 presents the results for the CDF level set score $(\text{MCRPS}')^{\Gamma}$ and the scoring rule MCRPS'; and Table 3 presents the results for the LPM level set score $(\text{LPMS}')^{\Gamma}$ and the scoring rule LPMS'. In each table, all entries for the misspecified candidates are strictly positive, indicating that the scores are able to identify the DGP. This supports our assertion that the new scoring rule LPMS' is proper and that $(\text{DQS}')^{\Gamma}$, $(\text{MCRPS}')^{\Gamma}$, and $(\text{LPMS}')^{\Gamma}$ are consistent scoring functions for density level sets, CDF level sets, and LPM level sets, respectively. It also supports the more general theoretical result that the L^2 scoring rule S' in (3) is proper, and $(S')^{\Gamma}$ in (11) is consistent for level sets.

Next, we show that the properness of our L^2 scoring rules and the consistency of the level set scores also hold for other choices of h. Tables 4 and 5 present results for the density level set score $(\text{DQS}')^{\Gamma}$ and its corresponding L^2 scoring rule DQS' with h chosen as the PDF of the bivariate normal distributions $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ and $\mathcal{N}(\mathbf{0}, 2 \times \mathbf{I}_2)$, respectively, where $\mathbf{I}_2$ denotes the identity matrix. To save space, we do not present the results for the other scores. It can be seen that all values for the misspecified candidates are strictly positive, indicating that the DGP receives the lowest scores. Tables 1-5 were based on a single run of the simulation study (with $T = 2 \times 10^5$ observations). We repeated the study 1000 times, and found the DGP almost always received the lowest scores. This shows that the proposed scores can reliably distinguish the DGP from the misspecified distributions.

To conclude this section, we briefly discuss the role of h as a weight function. In the literature, there are the two major weighting schemes, threshold weighting and quantile weighting, which have

been applied to scores for univariate distributions (Gneiting and Ranjan 2011, Grushka-Cockayne et al. 2017). The function h we consider in our paper enables threshold weighting. For example, the uniform PDF considered in Tables 1-3 ignore all the regions outside $[-3,3]^d$, while the bivariate normal PDFs considered in Tables 4-5 assign more weight to the center of the distributions. The function h cannot enable quantile weighting, and so, for example, cannot be used to emphasize regions of a univariate distribution related to VaR for a chosen probability level. For this, one could use either the quantile score or the quantile weighted CRPS. Our level set scores can be viewed as generalizations of the quantile score to the multivariate setting, which enable a complementary and useful way to emphasize regions of interest. The consistency of the level set score holds for any h that is a PDF, which means the regions of interest in the level set scores are governed by the level sets and not by h. In practice, one could consider different h to compare two misspecified estimates. If one candidate consistently outperforms the other for a large family of h, then we could conclude that the former is *dominating* the latter. For example, Ehm et al. (2016) essentially consider different h to produce a so-called "Murphy diagram" to graphically compare two quantile estimates for univariate distributions.

Table 1 For the simulated data, comparison of candidate distributions using the density level set score $(\mathrm{DQS}')^{\Gamma}$ and L^2 score DQS' ($\times 10^6$). The scores were computed with h chosen as the PDF of the uniform distribution on $[-3,3]^2$. Each value is the score for a candidate distribution minus the score for DGP. Lower values are better.

	$(\mathrm{DQS}')^{\Gamma}$									DQS'
α	0.018	0.037	0.055	0.073	0.092	0.110	0.129	0.147	0.165	
DGP	0	0	0	0	0	0	0	0	0	0
Misspecified means										
[-2,-2]	13297	13840	13395	12558	11502	10082	7611	5281	2802	3497
[-0.4,-0.4]	713	950	1071	1055	1026	1061	774	566	341	280
[0.4, 0.4]	657	941	1096	1024	919	908	614	502	299	261
[2, 2]	13270	13780	13326	12432	11349	9966	7500	5109	2602	3459
Misspecified variances										
[0.6,0.6]	3461	2342	1447	899	608	633	542	725	1093	1981
[0.8,0.8]	375	256	134	57	30	118	118	252	354	145
[1.2,1.2]	127	101	38	3	106	370	496	797	288	86
[1.4,1.4]	429	250	107	213	566	1487	1601	797	288	220
Misspecified covariance										
-0.5	3709	3649	3204	2622	1912	1342	641	336	172	693
0.1	720	726	619	478	319	283	215	319	288	154
0.7	458	376	307	207	137	157	107	105	174	99
0.9	3766	2890	2090	1631	1303	1239	1139	1199	1324	1334

Table 2 For the simulated data, comparison of candidate distributions using the CDF level set score $(\text{MCRPS}')^{\Gamma}$ and L^2 score MCRPS' ($\times 10^5$). The scores were computed with h chosen as the PDF of the uniform distribution on $[-3,3]^2$. Each value is the score for a candidate distribution minus the score for DGP. Lower values are better.

	$(\text{MCRPS}')^{\Gamma}$									MCRPS'
α	0.035	0.082	0.138	0.203	0.277	0.363	0.462	0.579	0.731	
DGP	0	0	0	0	0	0	0	0	0	0
Misspecified means										
[-2,-2]	975	3046	5854	8342	10464	12425	14161	15309	15071	21754
[-0.4,-0.4]	141	269	374	476	532	565	597	548	433	817
[0.4, 0.4]	204	347	431	501	555	547	526	437	288	721
[2, 2]	8296	10002	10447	10145	9348	8151	6543	4638	2330	11033
Misspecified variances										
[0.6,0.6]	144	137	102	67	39	24	27	51	103	156
[0.8,0.8]	27	27	19	13	7	5	5	11	22	31
[1.2,1.2]	18	18	15	9	5	3	4	9	16	22
[1.4,1.4]	61	66	50	31	17	10	16	34	62	77
Misspecified covariance										
-0.5	320	323	278	232	187	129	85	46	15	222
0.1	31	39	40	34	31	27	19	13	5	37
0.7	6	8	10	10	9	8	7	5	3	11
0.9	19	30	35	40	38	38	35	27	16	49

Table 3 For the simulated data, comparison of candidate distributions using the LPM level set score $(\text{LPMS}')^{\Gamma}$ and L^2 score LPMS' ($\times 10^5$) for $k = 1$. The scores were computed with h chosen as the PDF of the uniform distribution on $[-3,3]^2$. Each value is the score for a candidate distribution minus the score for DGP. Lower values are better.

	$(\text{LPMS}')^{\Gamma}$									LPMS'
α	0.016	0.047	0.094	0.163	0.261	0.401	0.611	0.952	1.628	
DGP	0	0	0	0	0	0	0	0	0	0
Misspecified means										
[-2,-2]	381	1588	3745	7050	11755	17914	25674	36623	55301	3278906
[-0.4,-0.4]	99	232	408	624	865	1154	1538	2033	2669	52604
[0.4, 0.4]	123	317	541	804	1118	1470	1864	2297	2820	32690
[2, 2]	12288	20694	27612	33659	38675	42980	45236	45084	36098	261370
Misspecified variances										
[0.6,0.6]	143	226	272	277	279	263	224	156	76	725
[0.8,0.8]	19	36	50	57	57	55	52	36	21	178
[1.2,1.2]	27	45	54	56	57	54	47	32	16	143
[1.4,1.4]	76	132	170	189	199	193	172	133	72	577
Misspecified covariance										
-0.5	560	877	1130	1351	1532	1714	1785	1821	1691	16643
0.1	34	73	108	143	181	210	248	270	282	2835
0.7	11	18	25	32	40	48	52	59	56	585
0.9	27	50	80	112	140	173	202	241	241	2591

Table 4 For the simulated data, comparison of candidate distributions using the density level set score $(\mathrm{DQS}^{'})^{\Gamma}$ and L^2 score $\mathrm{DQS}^{'}$ ($\times 10^5$). The scores were computed with h chosen as the PDF of $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$. Each value is the score for a candidate distribution minus the score for DGP. Lower values are better.

	$(\mathrm{DQS}^{'})^{\Gamma}$									$\mathrm{DQS}^{'}$
α	0.018	0.037	0.055	0.073	0.092	0.110	0.129	0.147	0.165	
DGP	0	0	0	0	0	0	0	0	0	0
Misspecified means										
[-2,-2]	4473	4483	3801	2976	2172	1454	862	431	119	803
[-0.4,-0.4]	37	94	158	198	255	275	266	233	139	61
[0.4, 0.4]	35	100	173	213	253	264	246	192	90	58
[2, 2]	4472	4498	3809	2978	2171	1457	860	424	114	803
Misspecified variances										
[0.6,0.6]	847	696	490	314	208	134	131	234	395	915
[0.8,0.8]	67	54	37	6	10	21	65	134	172	70
[1.2,1.2]	16	17	10	3	25	96	211	357	85	29
[1.4,1.4]	42	39	23	38	195	583	751	357	85	80
Misspecified covariance										
-0.5	342	540	644	641	596	474	324	191	38	141
0.1	61	105	138	119	116	77	94	129	85	35
0.7	79	79	68	36	43	38	64	62	75	33
0.9	920	791	607	436	320	234	217	288	389	494

Table 5 For the simulated data, comparison of candidate distributions using the density level set score $(\mathrm{DQS}^{'})^{\Gamma}$ and L^2 score $\mathrm{DQS}^{'}$ ($\times 10^5$). The scores were computed with h chosen as the PDF of $\mathcal{N}(\mathbf{0}, 2 \times \mathbf{I}_2)$. Each value is the score for a candidate distribution minus the score for DGP. Lower values are better.

	$(\mathrm{DQS}^{'})^{\Gamma}$									$\mathrm{DQS}^{'}$
α	0.018	0.037	0.055	0.073	0.092	0.110	0.129	0.147	0.165	
DGP	0	0	0	0	0	0	0	0	0	0
Misspecified means										
[-2,-2]	2751	2688	2300	1826	1360	960	621	342	116	506
[-0.4,-0.4]	58	103	150	162	169	165	163	118	33	42
[0.4, 0.4]	57	115	166	173	170	168	159	122	60	44
[2, 2]	2756	2703	2316	1844	1381	977	636	351	121	511
Misspecified variances										
[0.6,0.6]	639	484	333	200	113	86	106	164	251	527
[0.8,0.8]	60	44	38	3	11	18	50	80	107	44
[1.2,1.2]	20	14	9	8	21	40	112	164	24	15
[1.4,1.4]	53	34	23	35	118	314	376	164	24	43
Misspecified covariance										
-0.5	419	526	546	476	382	285	198	96	16	112
0.1	82	101	113	94	69	48	48	61	24	24
0.7	69	64	62	30	27	16	41	38	42	24
0.9	681	554	418	295	214	171	173	217	280	307

6. Practical Applications

In this section, we apply the L^2 scoring rules and level set scores to two practical applications. In Section 6.1, we use the L^2 scoring rules to estimate weights for combining distributional forecasts, and in Section 6.2, we use the CDF level set score to estimate conditional Value-at-Risk (CoVaR).

6.1. Using L^2 Scores to Combine Distributional Forecasts

Combining is a popular and pragmatic way to improve forecast accuracy. In recent years, interest has increased in methods for combining probabilistic forecasts (Aastveit et al. 2018, Winkler et al. 2019). We show that, for a combination of distributional forecasts, the L^2 scoring rules enable us to estimate the combining weights as a quadratic optimization algorithm, which is both theoretically and computationally appealing. Let $\{P_{\mathbf{X}_{it}}\}_{i=1}^{N}$ be N distributional forecasts for $\mathbf{Y}_t$, where $\mathbf{X}_{it}$ has density $f_{\mathbf{X}_{it}}$ for $t = 1, 2, \ldots, T$. The most common method for combining distributional forecasts is to produce a mixture distribution of the form $\sum_{i=1}^{N} \eta_i P_{\mathbf{X}_{it}}$, which has PDF $\sum_{i=1}^{N} \eta_i f_{\mathbf{X}_{it}}$, where $\sum_{i=1}^{N} \eta_i = 1$ and $\eta_i \geq 0$. The standard approach to estimating the combining weights η_i is to minimize the log score, which can be viewed as maximum likelihood estimation (see, for example, Hall and Mitchell 2007, Gneiting and Ranjan 2013). Our alternative proposal is to estimate the weights by minimizing an L^2 scoring rule.

Let us consider the form of the L^2 scores in (3). If we expand the square bracket and rearrange the terms in (3), we find that, with T in-sample observations, the average L^2 score, $\frac{1}{T}\sum_{t=1}^{T} S'\left(\sum_{i=1}^{N} \eta_i P_{\mathbf{X}_{it}}, \mathbf{y}_t; w, h\right)$, can be simplified into a quadratic function of the weights η_i. In fact, it can be equivalently written, as $\boldsymbol{\eta}^{'} B \boldsymbol{\eta} + \mathbf{c}^{'} \boldsymbol{\eta}$, where $\boldsymbol{\eta} = [\eta_1, \eta_2, \ldots, \eta_N]^{'}$, B is an $N \times N$ positive semi-definite matrix with $B_{i,j} = \frac{1}{T}\sum_{t=1}^{T} \int_{\mathbb{R}^d} (f_{\mathbf{X}_{it}} * w)(\mathbf{z})(f_{\mathbf{X}_{jt}} * w)(\mathbf{z}) h(\mathbf{z}) d\mathbf{z}$ and $\mathbf{c} = [c_1, c_2, \ldots, c_N]^{'}$ with $c_i = -\frac{2}{T}\sum_{t=1}^{T} \int_{\mathbb{R}^d} (f_{\mathbf{X}_{it}} * w)(\mathbf{z}) w(\mathbf{z} - \mathbf{y}) h(\mathbf{z}) d\mathbf{z}$. Hence, to find the optimal weights, we simply need to solve a quadratic optimization problem with linear constraints $\mathbf{1}_N^{'} \boldsymbol{\eta} = 1$ and $\boldsymbol{\eta} \geq 0$, where $\mathbf{1}_N$ is a column vector of ones.

Our proposed approach has several desirable properties when compared with estimating the combining weights by minimizing the log score. First, the L^2 scores offer greater flexibility. This is

because, unlike the log score, the L^2 scores allow the PDFs to have zero values. Furthermore, by contrast with the log score, the CDF-based MCRPS and MCRPS$'$, can be used for estimation in applications where PDFs are not available, such as when the distributional forecast is produced via simulation (see, for example, Taylor and Jeon 2018). Second, the quadratic optimization problem is computationally efficient, as it can be solved numerically in polynomial time (Ye and Tse 1989). Third, the matrix B reveals useful information regarding each individual forecast. Any individual forecast that is perfectly correlated with the others in the sense of the matrix B can be viewed as redundant, as it has no impact on the minimal value of the L^2 score. By removing the corresponding rows (columns), we can assume that the matrix B is positive definite, which will induce a unique solution for the optimal combining weights. Finally, our proposed approach generalizes the seminal work of Makridakis and Winkler (1983) on point forecasting. Hence, recent developments for combining point forecasts (see, for example, Soule et al. 2021) can be incorporated in our approach for combining probabilistic forecasts. Our work also includes, as a special case, the work of Hora and Kardeş (2015), who use the quadratic score to estimate combining weights.

To illustrate our proposal, we use forecasts provided by multiple experts in the ECB Survey of Professional Forecasters. The data consists of rolling year-ahead predictions for quarterly inflation, GDP and unemployment in the euro area for the 91 quarters up to the third quarter of 2021. The distributional forecasts are submitted in the form of probabilities for the economic variable falling in pre-specified bins. There were many cases where forecasters submitted zero probability for at least one bin, implying that the PDFs took zero values. This meant that the log score could not be directly used as the basis for estimating combining weights. Although there were more than 100 individual forecasters, many did not provide forecasts for every quarter. We kept only the individual forecasters for which forecasts were missing for eight or fewer quarters. This left six forecasters for GDP, seven for inflation, and five for unemployment. For these forecasters, any missing distributional forecasts were replaced by the average of all the other available forecasts for that quarter.

We considered three combining methods: the simple average, and the proposed combining approach with weights estimated using the DQS$'$ and CRPS$'$. To compute the L^2 scores, for

each series, $M = 10^4$ values were sampled from h, which we chose to be the uniform distribution between the minimum and maximum values for the entire period. We repeatedly re-estimated model parameters using a rolling window of 48 quarters, which delivered 43 out-of-sample forecasts. We also used DQS$'$ and CRPS$'$ for evaluation. We present the results in Table 6, where the scores corresponding to the benchmark simple average, which is used as the reference, have been subtracted from each candidate. It can be seen that all the scores associated with DQS$'$ and CRPS$'$ are negative, indicating that they outperformed the simple average.

Table 6 Comparison of combining methods. Each value is a score for a method minus the score for the simple average ($\times 10^3$). Lower values are better. † and * indicate significantly less than zero at 10% and 5% significance levels.

	GDP			Inflation			Unemployment		
	Combining method			Combining method			Combining method		
	Simple Avg	DQS$'$	CRPS$'$	Simple Avg	DQS$'$	CRPS$'$	Simple Avg	DQS$'$	CRPS$'$
DQS$'$	0	-42	-97^*	0	-128	-258^*	0	-630^*	-651^*
CRPS$'$	0	-42^*	-5	0	-164^*	$-31^\dagger$	0	-290^*	-301^*

6.2. Using the CDF Level Set Score for CoVaR Estimation

CoVaR is a popular measure of systemic risk, which assesses the probability that a market incurs a heavy loss given that an individual stock is already in distress (see, for example, Tobias and Brunnermeier 2016, Dimitriadis and Hoga 2022). Let Y_{1t} and Y_{2t} denote returns for a market index and an individual stock, respectively. The CoVaR of Y_{1t} conditional on Y_{2t}, at probability level θ, is defined as the scalar, $\text{CoVaR}_{Y_{1t}|Y_{2t}}$, that satisfies $\mathbb{P}\left(Y_{1t} < \text{CoVaR}_{Y_{1t}|Y_{2t}} | Y_{2t} < q_{Y_{2t}}(\theta)\right) = \theta$, where $q_{Y_{2t}}(\theta)$ denotes the θ quantile of Y_{2t}. In this paper, to forecast $\text{CoVaR}_{Y_{1t}|Y_{2t}}$, we consider a standard approach in the literature, which is to use a GARCH-Copula model (see, for example, Patton 2012). The approach involves two steps. First, marginal distributions of Y_{1t} and Y_{2t} are estimated using GARCH models, and then a copula is used to capture the dependence between the marginal distributions.

We note that $\mathbb{P}\left(Y_{1t} < \text{CoVaR}_{Y_{1t}|Y_{2t}}, Y_{2t} < q_{Y_{2t}}(\theta)\right) = \theta^2$. This indicates that the pair $\left(\text{CoVaR}_{Y_{1t}|Y_{2t}}, q_{Y_{2t}}(\theta)\right)$ is precisely on the θ^2 CDF level set of (Y_{1t}, Y_{2t}). This prompts us to propose the estimation of the copula parameters by minimizing the score for the θ^2 CDF level set of

(Y_{1t}, Y_{2t}). We argue that this will deliver better forecasting accuracy, in comparison with the standard maximum likelihood approach, because the CDF level set score enables improved fit in the tails.

To illustrate the proposed CoVaR estimation method based on the CDF level set score, we set Y_{1t} as the S&P 500 index, and for Y_{2t}, we consider, in turn, the three stocks of the S&P 500 that were analyzed by Diks and Fang (2020): Alcoa (AA), MacDonald's (MCD), and Merck (MRK). We used the 5000 daily log returns recorded between 21 February 2002 and 31 December 2021. Using a rolling window of 2000 observations, we repeatedly re-estimated model parameters to generate 3000 out-of-sample day-ahead forecasts for the joint distribution. We used a GARCH(1,1) model with Student-t distribution for the marginal distributions, and a Gaussian copula to model the dependence. The CoVaR probability level θ was set to 10%. We estimated the copula parameters first using maximum likelihood estimation, and then using our method based on the score for the θ^2=1% CDF level set. This CDF level set score was computed using 10^4 values sampled from the PDF h, which we set as the uniform distribution on $[-0.3, 0.3]^2$.

Table 7 compares the CoVaR forecasting performance of the two approaches using three measures. First, we computed the CoVaR coverage as the proportion of the days for which both of the following were true: $Y_{2t} < q_{Y_{2t}}(\theta)$ and $Y_{1t} < \mathrm{CoVaR}_{Y_{1t}|Y_{2t}}$. In the table, the coverage values for the proposed method are closer to the desired value of 1% ($=\theta^2$) compared to the maximum likelihood approach. Second, as $\mathrm{CoVaR}_{Y_{1t}|Y_{2t}}$ is the θ quantile of Y_{1t} given $Y_{2t} < q_{Y_{2t}}(\theta)$, we report the θ quantile score for $\mathrm{CoVaR}_{Y_{1t}|Y_{2t}}$ for periods in which $Y_{2t} < q_{Y_{2t}}(\theta)$. Third, we report the 1% CDF level set score. For both these scores, lower values are better. For clarity, the scores corresponding to the benchmark maximum likelihood approach are chosen as the reference and subtracted from those for the proposed method. In all three rows, the proposed method has negative values for the scores, indicating it outperforms the benchmark in estimating the CoVaR for all three stocks.

Table 7 Comparison of CoVaR forecasting methods. Coverage closer to 1% is preferred. For the CoVaR quantile and level set scores, each value is the score for the proposed approach minus the score for the maximum likelihood benchmark ($\times 10^7$). Lower values are better. † and * indicate significantly less than zero at 10% and 5% significance levels.

	Coverage (%)		CoVaR quantile score		Level set score	
	Max likelihood	Proposed	Max likelihood	Proposed	Max likelihood	Proposed
AA & SP500	1.67	1.37	0.00	−6.04	0.00	−13.92
MRK & SP500	1.77	1.37	0.00	$-9.88^\dagger$	0.00	−40.29
MCD & SP500	1.60	1.27	0.00	-8.10^*	0.00	−12.10

7. Conclusion

Forecasts of multivariate distributions and level sets are needed to support decision making in a variety of contexts. In this paper, we have studied the scoring rules for multivariate distributions and scoring functions for level sets. The paper has several novel contributions. Firstly, we propose the class of L^2 scoring rules for multivariate distributions, for which the existing quadratic score and MCRPS are specific examples. The L^2 scoring functions can easily generate new scoring rules for multivariate distributions, and we demonstrate this with the introduction of the LPMS, a new scoring rule based on the lower partial moments. Secondly, by decomposing the L^2 scoring rules, we obtain a unified approach for generating scoring functions for level sets, including the scoring functions for density level sets, CDF level sets, and LPM level sets. Thirdly, we propose a simple numerical approach for computing the L^2 scoring rules and the scoring functions for level sets. Finally, we performed a simulation study to provide support for the theoretical properties of our new scores, and we used real data to illustrate their practical usefulness for forecast combining and CoVaR estimation.

Acknowledgments

The authors are grateful to the Area Editor, the Associate Editor and three reviewers for providing very helpful comments. The authors are also grateful to Tobias Fissler and participants at the INFORMS Advances in Decision Analysis conference held at Bocconi University in Milan in 2019 for insightful feedback.

References

Aastveit, K. A., J. Mitchell, F. Ravazzolo, and H. K. Van Dijk (2018). The evolution of forecast density combinations in economics. Technical report.

Abbas, A. E., D. V. Budescu, and Y. Gu (2010). Assessing joint distributions with isoprobability contours. *Management Science 56*(6), 997–1011.

Abernethy, J. D. and R. M. Frongillo (2012). A characterization of scoring rules for linear properties. In *Conference on Learning Theory*, pp. 27–1.

Anthonisz, S. A. (2012). Asset pricing with partial-moments. *Journal of Banking & Finance 36*(7), 2122–2135.

Baringhaus, L. and C. Franz (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis 88*(1), 190–206.

Berrocal, V. J., A. E. Gelfand, and D. M. Holland (2010). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics 15*(2), 176–197.

Briec, W. and K. Kerstens (2010). Portfolio selection in multidimensional general and partial moment space. *Journal of Economic Dynamics and Control 34*(4), 636–656.

Cadre, B. (2006). Kernel estimation of density level sets. *Journal of Multivariate Analysis 97*(4), 999–1023.

Caflisch, R. E. (1998). Monte carlo and quasi-monte carlo methods. *Acta Numerica 1998*, 1–49.

Chen, Y.-C., C. R. Genovese, and L. Wasserman (2017). Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association 112*(520), 1684–1696.

Corbella, S. and D. Stretch (2012). Multivariate return periods of sea storms for coastal erosion risk assessment. *Natural Hazards and Earth System Sciences 12*(8), 2699–2708.

Cousin, A. and E. Di Bernardino (2013). On multivariate extensions of value-at-risk. *Journal of Multivariate Analysis 119*, 32–46.

Danaher, P. J. and M. S. Smith (2011). Modeling multivariate distributions using copulas: Applications in marketing. *Marketing Science 30*(1), 4–21.

Diks, C. and H. Fang (2020). Comparing density forecasts in a risk management context. *International Journal of Forecasting 36*(2), 531–551.

Diks, C., V. Panchenko, O. Sokolinskiy, and D. Van Dijk (2014). Comparing the accuracy of multivariate density forecasts in selected regions of the copula support. *Journal of Economic Dynamics and Control 48*, 79–94.

Dimitriadis, T. and Y. Hoga (2022). Dynamic co-quantile regression. *arXiv preprint arXiv:2206.14275*.

Ehm, W., T. Gneiting, et al. (2012). Local proper scoring rules of order two. *The Annals of Statistics 40*(1), 609–637.

Ehm, W., T. Gneiting, A. Jordan, and F. Krüger (2016). Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(3), 505–562.

Embrechts, P. and G. Puccetti (2006). Bounds for functions of multivariate risks. *Journal of Multivariate Analysis 97*(2), 526–547.

Fissler, T., R. Frongillo, J. Hlavinová, and B. Rudloff (2021). Forecast evaluation of quantiles, prediction intervals, and other set-valued functionals. *Electronic Journal of Statistics 15*(1), 1034–1084.

Frongillo, R. and I. A. Kash (2015). Vector-valued property elicitation. In *Conference on Learning Theory*, pp. 710–727.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association 106*(494), 746–762.

Gneiting, T. and M. Katzfuss (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application 1*, 125–151.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Gneiting, T. and R. Ranjan (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business and Economic Statistics 29*(3), 411–422.

Gneiting, T. and R. Ranjan (2013). Combining predictive distributions. *Electronic Journal of Statistics 7*, 1747–1782.

Grushka-Cockayne, Y., K. C. Lichtendahl Jr, V. R. R. Jose, and R. L. Winkler (2017). Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. *Operations Research 65*(3), 712–728.

Hall, S. G. and J. Mitchell (2007). Combining density forecasts. *International Journal of Forecasting 23*(1), 1–13.

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

Hartigan, J. A. (1987). Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association 82*(397), 267–270.

Hora, S. C. and E. Kardeş (2015). Calibration, sharpness and the weighting of experts in a linear opinion pool. *Annals of Operations Research 229*(1), 429–450.

Jeon, J. and J. W. Taylor (2012). Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association 107*(497), 66–79.

Jose, V. R. R. (2017). Percentage and relative error measures in forecast evaluation. *Operations Research 65*(1), 200–211.

Jose, V. R. R. and R. L. Winkler (2009). Evaluating quantile assessments. *Operations Research 57*(5), 1287–1297.

Komunjer, I. (2005). Quasi-maximum likelihood estimation for conditional quantiles. *Journal of Econometrics 128*(1), 137–164.

Kong, L. and I. Mizera (2012). Quantile tomography: using quantiles with multivariate data. *Statistica Sinica 22*, 1589–1610.

Laio, F. and S. Tamea (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences 11*(4), 1267–1277.

Lambert, N. S. (2019). Elicitation and evaluation of statistical forecasts. *Working paper*.

Lichtendahl Jr, K. C., Y. Grushka-Cockayne, and R. L. Winkler (2013). Is it better to average probabilities or quantiles? *Management Science 59*(7), 1594–1611.

Lieb, E. H. and M. Loss (2001). *Analysis, Graduate Studies in Mathematics, vol. 14*.

Lions, J.-L. (1951). Supports de produits de composition. *Les Comptes rendus de l'AcadÃ©mie des sciences 232*(18), 1622–1624.

Makridakis, S. and R. L. Winkler (1983). Averages of forecasts: Some empirical results. *Management Science 29*(9), 987–996.

Menardi, G. (2016). A review on modal clustering. *International Statistical Review 84*(3), 413–433.

Moftakhari, H. R., G. Salvadori, A. AghaKouchak, B. F. Sanders, and R. A. Matthew (2017). Compounding effects of sea level rise and fluvial flooding. *Proceedings of the National Academy of Sciences 114*(37), 9785–9790.

Müller, D. W. and G. Sawitzki (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association 86*(415), 738–746.

Newman, J. P., H. R. Maier, G. A. Riddell, A. C. Zecchin, J. E. Daniell, A. M. Schaefer, H. van Delden, B. Khazai, M. J. O'Flaherty, and C. P. Newland (2017). Review of literature on decision support systems for natural hazard risk reduction: Current status and future research directions. *Environmental Modelling & Software 96*, 378–409.

Parry, M., A. P. Dawid, and S. Lauritzen (2012). Proper local scoring rules. *The Annals of Statistics 40*(1), 561–592.

Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis 110*, 4–18.

Price, K., B. Price, and T. J. Nantell (1982). Variance and lower partial moment measures of systematic risk: some analytical and empirical results. *The Journal of Finance 37*(3), 843–855.

Rigollet, P. and R. Vert (2009). Optimal rates for plug-in estimators of density level sets. *Bernoulli 15*(4), 1154–1178.

Rinaldo, A., A. Singh, R. Nugent, and L. Wasserman (2012). Stability of density-based clustering. *Journal of Machine Learning Research 13*(Apr), 905–948.

Rinaldo, A. and L. Wasserman (2010). Generalized density clustering. *The Annals of Statistics 38*(5), 2678–2722.

Robert, C. and G. Casella (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Salvadori, G., F. Durante, C. De Michele, M. Bernardi, and L. Petrella (2016). A multivariate copula-based framework for dealing with hazard scenarios and failure probabilities. *Water Resources Research 52*(5), 3701–3721.

Salvadori, G., G. Tomasicchio, and F. D'Alessandro (2014). Practical guidelines for multivariate analysis and design in coastal and off-shore engineering. *Coastal Engineering 88*, 1–14.

Singh, A., C. Scott, and R. Nowak (2009). Adaptive hausdorff estimation of density level sets. *The Annals of Statistics 37*(5B), 2760–2782.

Soule, D., Y. Grushka-Cockayne, and J. R. Merrick (2021). A heuristic for combining correlated experts when there is little data. *Available at SSRN 3680229*.

Steinwart, I., D. Hush, and C. Scovel (2005). A classification framework for anomaly detection. *Journal of Machine Learning Research 6*(Feb), 211–232.

Taylor, J. W. and J. Jeon (2018). Probabilistic forecasting of wave height for offshore wind turbine maintenance. *European Journal of Operational Research 267*(3), 877–890.

Tobias, A. and M. K. Brunnermeier (2016). Covar. *The American Economic Review 106*(7), 1705.

Winkler, R. L., Y. Grushka-Cockayne, K. C. Lichtendahl Jr, and V. R. R. Jose (2019). Probability forecasts and their combination: A research perspective. *Decision Analysis 16*(4), 239–260.

Ye, Y. and E. Tse (1989). An extension of karmarkar's projective algorithm for convex quadratic programming. *Mathematical programming 44*(1), 157–179.

Yuen, R. and S. Stoev (2014). CRPS M-estimation for max-stable models. *Extremes 17*(3), 387–410.

Appendix. Assumption in Theorem 1

ASSUMPTION 1. *(a) Let λ be a Borel measure on $\mathbb{R}^d$ whose Radon–Nikodym derivative with respect to the Lebesgue measure is a non-negative integrable function h, i.e., $d\lambda(\mathbf{z}) = h(\mathbf{z})d\mathbf{z}$. Let w be a local Borel measure such that $f_{\mathbf{Y}} * w$ is a well-defined local Borel measure $\forall P_{\mathbf{Y}} \in \mathcal{V}^d$.*

(b) If $f_{\mathbf{Y}} * w$ and w are square-integrable with respect to λ, and $\int_{\mathbb{R}^d} w^2(\mathbf{z} - \bullet)h(\mathbf{z})f_{\mathbf{Y}}(\bullet)\, d\mathbf{z}$ is integrable with respect to the PDF $f_{\mathbf{Y}}$, then (2) is well-defined and finite.

(c) If $(f_{\mathbf{X}} * w)(\bullet)w(\bullet - \mathbf{y})$ is integrable with respect to λ, and $\int_{\mathbb{R}^d} (f_{\mathbf{X}} * w)(\mathbf{z})w(\mathbf{z} - \bullet)h(\mathbf{z})f_{\mathbf{Y}}(\bullet)\, d\mathbf{z}$ is integrable for any distribution $P_{\mathbf{Y}} \in \mathcal{V}^d$, then (3) is well-defined and finite.

Appendix. Proofs of Theorems

Proof of Theorem 1 Consider the divergence $\Delta := \mathbb{E}_{P_{\mathbf{Y}}}[S(P_{\mathbf{X}}, \bullet; w)] - \mathbb{E}_{P_{\mathbf{Y}}}[S(P_{\mathbf{Y}}, \bullet; w)]$. One may compute it as follows:

$$
\begin{aligned}
\Delta &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Big[(f_{\mathbf{X}} * w)(\mathbf{z}) - w(\mathbf{z} - \mathbf{s}) \Big]^2 f_{\mathbf{Y}}(\mathbf{s}) h(\mathbf{z}) d\mathbf{z} d\mathbf{s} - \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Big[(f_{\mathbf{Y}} * w)(\mathbf{z}) - w(\mathbf{z} - \mathbf{s}) \Big]^2 f_{\mathbf{Y}}(\mathbf{s}) h(\mathbf{z}) d\mathbf{z} d\mathbf{s} \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Big\{ (f_{\mathbf{X}} * w)^2(\mathbf{z}) - (f_{\mathbf{Y}} * w)^2(\mathbf{z}) - 2\Big[(f_{\mathbf{X}} * w)(\mathbf{z}) - (f_{\mathbf{Y}} * w)(\mathbf{z}) \Big] w(\mathbf{z} - \mathbf{s}) \Big\} f_{\mathbf{Y}}(\mathbf{s})\, d\mathbf{s}\, h(\mathbf{z})\, d\mathbf{z} \\
&= \int_{\mathbb{R}^d} \Big\{ (f_{\mathbf{X}} * w)^2(\mathbf{z}) - (f_{\mathbf{Y}} * w)^2(\mathbf{z}) - 2\Big[(f_{\mathbf{X}} * w)(\mathbf{z}) - (f_{\mathbf{Y}} * w)(\mathbf{z}) \Big] (f_{\mathbf{Y}} * w)(\mathbf{z}) \Big\} h(\mathbf{z})\, d\mathbf{z} \\
&= \int_{\mathbb{R}^d} \Big\{ f_{\mathbf{X}} * w - f_{\mathbf{Y}} * w \Big\}^2 (\mathbf{z}) h(\mathbf{z})\, d\mathbf{z} \geq 0,
\end{aligned}
$$

where the penultimate equality holds by Fubini's theorem and the definition of convolution.

For strict properness, we assume $\Delta = 0$. If h is non-zero $\mathcal{L}^d$-a.e., we obtain that $f_{\mathbf{X}} * w = f_{\mathbf{Y}} * w$ $\mathcal{L}^d$-a.e.. If $f_{\mathbf{Y}} * w$ uniquely characterizes $P_{\mathbf{Y}}$ $\forall P_{\mathbf{Y}} \in \mathcal{V}^d$, we obtain $P_{\mathbf{X}} = P_{\mathbf{Y}}$. $\square$

Derivation and proof of Theorem 2 (i) Derivation of (11) in Theorem 2. The layer cake representation states that if μ is a Borel measure on $\mathbb{R}^d$ and $g:\mathbb{R}^d\to[0,\infty)$ is a μ-measurable function, then for any $p\in[1,\infty)$, there holds

$$\int_{\mathbb{R}^d} g(\mathbf{z})^p\, d\mu(\mathbf{z}) = \int_0^\infty p\alpha^{p-1}\mu\big\{g\geq\alpha\big\}\, d\alpha = \int_0^\infty p\alpha^{p-1}\mu\left\{L(g;\alpha)\right\}\, d\alpha,$$

where $\mu\{\bullet\}$ denotes the measure of a set under μ.

Let us define two Borel measures, λ as $d\lambda(\mathbf{z}) := h(\mathbf{z})\, d\mathbf{z}$, and $\mu_{\mathbf{y},w}$ as $d\mu_{\mathbf{y},w}(\mathbf{z}) := w(\mathbf{z}-\mathbf{y})h(\mathbf{z})d\mathbf{z}$. Then, we can express (3) in terms of the measures λ and $\mu_{\mathbf{y},w}$ as follows,

$$S^{'}(P_{\mathbf{X}},\mathbf{y};w,h) = \int_{\mathbb{R}^d} \left(f_{\mathbf{X}} * w\right)^2 d\lambda(\mathbf{z}) - 2\int_{\mathbb{R}^d} \left(f_{\mathbf{X}} * w\right)(\mathbf{z})\, d\mu_{\mathbf{y},w}(\mathbf{z}).$$

Because a nonnegative w implies $f_{\mathbf{Y}} * w$ is also nonnegative, we apply the layer cake representation to each term in the above expression to obtain the following,

$$\frac{1}{2}S^{'}(P_{\mathbf{X}},\mathbf{y};w,h) = \int_0^\infty \left(\alpha\lambda\Big\{L(f_{\mathbf{X}} * w;\alpha)\Big\} - \mu_{\mathbf{y},w}\Big\{L(f_{\mathbf{X}} * w;\alpha)\Big\}\right) d\alpha. \tag{17}$$

In the integrand on the right-hand side of (17), $L(f_{\mathbf{X}} * w;\alpha)$ is viewed as an estimate of $L(f_{\mathbf{Y}} * w;\alpha)$. Replacing $L(f_{\mathbf{X}} * w;\alpha)$ by any Borel set A leads to (11).

(ii) Proof of Theorem 2. We need to prove that for each number α and each Borel set $A\subseteq\mathbb{R}^d$,

$$\Delta' := \mathbb{E}_{P_{\mathbf{Y}}}\Big[(\mathrm{S}^{'})^{\Gamma}\big(A,\bullet;w,h,\alpha\big) - (\mathrm{S}^{'})^{\Gamma}\big(L(f_{\mathbf{Y}} * w;\alpha),\bullet;w,h,\alpha\big)\Big] \geq 0. \tag{18}$$

For notational convenience, for any $P_{\mathbf{X}}\in\mathcal{V}^d$ we write $\vartheta(\mathbf{y},\mathbf{z}) := w(\mathbf{z}-\mathbf{y})$, and $\Xi[\mathbf{Y}](\mathbf{z}) := \left(f_{\mathbf{Y}} * w\right)(\mathbf{z})$. We can express (18) by $\Delta' = \mathbb{E}_{P_{\mathbf{Y}}}[\Delta'']$, where

$$\Delta'' = \underbrace{\alpha\Big(\lambda\left\{A\right\} - \lambda\left\{\Xi[\mathbf{Y}]>\alpha\right\}\Big)}_{B_1} + \underbrace{\mu_{\mathbf{y},w}\big\{\Xi[\mathbf{Y}]>\alpha\big\} - \mu_{\mathbf{y},w}\big\{A\big\}}_{B_2}.$$

Let us partition $\mathbb{R}^d$ as $\mathbb{R}^d = \Sigma_{++}\sqcup\Sigma_{+-}\sqcup\Sigma_{-+}\sqcup\Sigma_{--}$, where $\sqcup$ is the disjoint union, and

$$\begin{aligned}
\Sigma_{++} &:= A\cap\big\{\Xi[\mathbf{Y}]>\alpha\big\}, \qquad \Sigma_{+-} := A\cap\big\{\Xi[\mathbf{Y}]\leq\alpha\big\},\\
\Sigma_{-+} &:= A^c\cap\big\{\Xi[\mathbf{Y}]>\alpha\big\}, \qquad \Sigma_{--} := A^c\cap\big\{\Xi[\mathbf{Y}]\leq\alpha\big\},
\end{aligned}$$

where A^c represents the complement set of A. By construction, each of these four sets is Borel measurable. Using the definitions of $d\mu_{\mathbf{y},w}$, $d\lambda$, and Fubini's theorem, we get

$$\begin{aligned}
B_1 &= \alpha\int_{\mathbb{R}^d}\left[\mathbb{1}_{\Sigma_{+-}}(\mathbf{z}) - \mathbb{1}_{\Sigma_{-+}}(\mathbf{z})\right] h(\mathbf{z})\, d\mathbf{z},\\
B_2 &= \int_{\mathbb{R}^d}\vartheta(\mathbf{y},\mathbf{z})\left[\mathbb{1}_{\Sigma_{-+}}(\mathbf{z}) - \mathbb{1}_{\Sigma_{+-}}(\mathbf{z})\right] h(\mathbf{z})\, d\mathbf{z},\\
\Delta'' &= \int_{\mathbb{R}^d}\left\{\Big(\alpha-\vartheta(\mathbf{y},\mathbf{z})\Big)\left[\mathbb{1}_{\Sigma_{+-}} - \mathbb{1}_{\Sigma_{-+}}\right]\right\} h(\mathbf{z})\, d\mathbf{z},
\end{aligned}$$

where $\mathbb{1}_B(\bullet)$ denotes the indicator function on the set B. Notice that $\mathbb{E}_{P_{\mathbf{Y}}}[\vartheta(\bullet,\mathbf{z})] = \Xi[\mathbf{Y}](\mathbf{z})$, which implies that

$$\Delta' = \mathbb{E}_{P_{\mathbf{Y}}}[\Delta''] = \int_{\mathbb{R}^d} \left\{ \Big(\alpha - \Xi[\mathbf{Y}](\mathbf{z})\Big) \Big[\mathbb{1}_{\Sigma_{+-}}(\mathbf{z}) - \mathbb{1}_{\Sigma_{-+}}(\mathbf{z})\Big] \right\} h(\mathbf{z})\, d\mathbf{z}. \tag{19}$$

On Σ_{+-} one has $\mathbb{1}_{\Sigma_{+-}} - \mathbb{1}_{\Sigma_{-+}} = 1$ and $\alpha - \Xi[\mathbf{Y}] \geq 0$, and on Σ_{-+}, $\mathbb{1}_{\Sigma_{+-}} - \mathbb{1}_{\Sigma_{-+}} = -1$ and $\alpha - \Xi[\mathbf{Y}] \leq 0$. Therefore, the integrand of Δ' is pointwise non-negative, so $\Delta' \geq 0$.

Regarding the strict consistency, we adopt the approach considered by Fissler et al. (2021). If A and $L(f_{\mathbf{Y}} * w; \alpha)$ differ by a set B that is not $\mathcal{L}^d$-null, then

$$\Delta' = \int_{B} \left\{ \Big(\alpha - \Xi[\mathbf{Y}](\mathbf{z})\Big) \Big[\mathbb{1}_{\Sigma_{+-}}(\mathbf{z}) - \mathbb{1}_{\Sigma_{-+}}(\mathbf{z})\Big] \right\} h(\mathbf{z})\, d\mathbf{z}. \tag{20}$$

However, if $L(f_{\mathbf{Y}} * w; \alpha)$ coincides with the closure of $\{f_{\mathbf{Y}} * w > \alpha\}$, the set $\{f_{\mathbf{Y}} * w = \alpha\}$ is $\mathcal{L}^d$-null. It follows that the integrand on the right-hand side of (20) is strictly positive on B modulo an $\mathcal{L}^d$-null set. But by Assumption 1(c), one has $h > 0$ $\mathcal{L}^d$-a.e.; so Δ' is strictly positive. This proves the strict consistency of $(S')^{\Gamma}$. $\square$

Appendix. Discussion of LPMs in Section 4.2.2

We first show that LPMs uniquely characterize distributions. To see this, we can take the k^{th} partial derivatives with respect to all z_j,

$$\begin{aligned}\left[\frac{\partial^k}{\partial z_1^k} \cdots \frac{\partial^k}{\partial z_d^k}(f_{\mathbf{X}} * u^{*k})\right](\mathbf{z}) &= \int_{-\infty}^{z_d} \cdots \int_{-\infty}^{z_1} \frac{\partial^k}{\partial s_1^k} \cdots \frac{\partial^k}{\partial s_d^k} u^{*k}(\mathbf{s}) f_{\mathbf{X}}(s_1, \ldots, s_d)\, ds_1 \cdots ds_d \\ &= \int_{-\infty}^{z_d} \cdots \int_{-\infty}^{z_1} f_{\mathbf{X}}(s_1, \ldots, s_d)\, ds_1 \cdots ds_d = f_{\mathbf{X}} * u(\mathbf{z}) = F_{\mathbf{X}}(\mathbf{z}).\end{aligned}$$

Therefore, if $\mathrm{LPMS}'(P_{\mathbf{X}}, \mathbf{y}; k, h) = \mathrm{LPMS}'(P_{\mathbf{Y}}, \mathbf{y}; k, h)$, then their CDFs must coincide, which means $P_{\mathbf{X}} = P_{\mathbf{Y}}$. We also note that when $k = 0$, $\mathrm{LPM}_{\mathbf{X},0} = F_{\mathbf{X}}$ is simply the CDF of $P_{\mathbf{X}}$, and in this case LPMS and LPMS$'$ are simply the MCRPS and MCRPS$'$ discussed in Section 3.2.2.

In addition to the three scores in Section 3.2, the following two conditions can also guarantee that $f_{\mathbf{Y}} * w$ uniquely characterizes $P_{\mathbf{Y}}$:

(i) if w and every distribution of $\mathcal{V}^d$ are compactly supported: it follows from J.-L. Lions' generalisation of the Titchmarsh convolution theorem (Lions 1951) that $f_{\mathbf{X}} = f_{\mathbf{Y}}$ $\mathcal{L}^d$-a.e..

(ii) Let w and/or its certain weak derivatives have well-defined Fourier transforms (denoted with a hat) that are nonzero $\mathcal{L}^d$-a.e.. First, if $\hat{w}$ exists and is nonzero $\mathcal{L}^d$-a.e., then $f_{\mathbf{X}} * w = f_{\mathbf{Y}} * w$ $\mathcal{L}^d$-a.e. implies $f_{\mathbf{X}} = f_{\mathbf{Y}}$ $\mathcal{L}^d$-a.e.. Next, let v be any given weak derivative of w and $\hat{v} \neq 0$ $\mathcal{L}^d$-a.e.. Then, the corresponding derivatives of $f_X * w$ and $f_Y * w$ are precisely $f_X * v$ and $f_Y * v$, respectively, thanks to the fundamental theorem of calculus. Finally, for $f_{\mathbf{X}} * w = f_{\mathbf{Y}} * w$ (or, using the fact that convolution

intertwines with multiplication, we obtain $\hat{f}_{\mathbf{X}}\hat{w} = \hat{f}_{\mathbf{Y}}\hat{w}$. Because $\hat{w} \neq 0$ $\mathcal{L}^d$-a.e., we conclude $\hat{f}_{\mathbf{X}} = \hat{f}_{\mathbf{Y}}$ $\mathcal{L}^d$-a.e., which implies $P_{\mathbf{X}} = P_{\mathbf{Y}}$. The same argument holds for v.

Appendix. Discussion on Remark 1

For a function $g : \mathbb{R}^d \to \mathbb{C}$, its Fourier transform $\hat{g} \equiv \mathcal{F}(g) : \mathbb{R}^d \to \mathbb{C}$ is given by $\hat{g}(\mathbf{t}) := \int_{\mathbb{R}^d} e^{2\pi i \mathbf{z}^T \mathbf{t}} g(\mathbf{z})\, d\mathbf{z}$, where $\mathbf{t} \in \mathbb{R}^d$ and i is the imaginary unit. When g is a PDF of a distribution, $\hat{g}$ is called the characteristic function of the distribution. The definition of the Fourier transform can be extended to certain generalized functions including Dirac delta masses.

Assume that the Plancherel identity holds. Then, using the fact that convolution intertwines with multiplication, we can obtain a Fourier representation for the L^2 score in (2),

$$S(P_{\mathbf{X}}, \mathbf{y}; \hat{w}, \hat{h}) = \int_{\mathbb{R}^d} \left| \left(\hat{f}_{\mathbf{X}} \hat{w} - \hat{\delta}_{\mathbf{y}} \hat{w} \right) * \hat{h} \right|^2 (\mathbf{t})\, d\mathbf{t}, \tag{21}$$

where $\hat{w}, \hat{h}, \hat{f}_{\mathbf{X}}, \hat{\delta}_{\mathbf{y}}$ denote the Fourier transforms of $w, h, f_{\mathbf{X}}, \delta_{\mathbf{y}}$, respectively.

Note that (21) can remain valid even if the Plancherel identity does not hold, provided that $S(P_{\mathbf{X}}, \mathbf{y}; \hat{w}, \hat{h})$ is finite for all $\mathbf{y}$ and that $S(P_{\mathbf{X}}, \bullet; \hat{w}, \hat{h}) f_{\mathbf{Y}}(\bullet)$ is integrable for any $P_{\mathbf{Y}} \in \mathcal{V}^d$. In general, there is no symmetry between (2) and (21). In Examples 3.2.2 and 3.2.3, $|w|$ is not integrable hence does not have Fourier transform, hence (21) cannot be defined; conversely, if we consider $\hat{w}$ with no well-defined inverse Fourier transform, then one does not have (2).

If we consider $\hat{w}_{\mathrm{ES}}(\mathbf{t}) = \left(\|\mathbf{t}\|^{\frac{d+1}{2}} \right)^{-1}/2$ and $\hat{h}(\mathbf{t}) = \delta_{\mathbf{0}}(\mathbf{t})$, (21) leads to the well-known energy score (see, for example, Baringhaus and Franz 2004, Gneiting and Raftery 2007). For the level set score, as the inverse Fourier transform $w = \mathcal{F}^{-1}(\hat{w})$ does not exists, we cannot directly apply Theorem 2. Observing that $\hat{w}_{\mathrm{ES}}(\mathbf{t})$ is a radial function, we can consider an alternative "projective" approach, where we take the Fourier transform of $\hat{w}$ and apply Theorem 2 "projectively" along different directions over the unit sphere (see, for example, Baringhaus and Franz 2004). This leads to the scoring function for the so-called projective quantile (Kong and Mizera 2012). However, this approach is essentially straightforward, as it only relies on the decomposition of the multivariate distributions into univariate projections. This is not the focus of the paper, and so we do not discuss this further.