Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Nonparametric Fusion Learning for Multiparameters: Synthesize Inferences From Diverse Sources Using Data Depth and Confidence Distribution

Dungang Liu, Regina Y. Liu & Min-ge Xie

To cite this article: Dungang Liu, Regina Y. Liu & Min-ge Xie (2022) Nonparametric Fusion Learning for Multiparameters: Synthesize Inferences From Diverse Sources Using Data Depth and Confidence Distribution, Journal of the American Statistical Association, 117:540, 2086-2104, DOI: 10.1080/01621459.2021.1902817

To link to this article: https://doi.org/10.1080/01621459.2021.1902817

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 540, 2086–2104: Theory and Methods
https://doi.org/10.1080/01621459.2021.1902817

Nonparametric Fusion Learning for Multiparameters: Synthesize Inferences From Diverse Sources Using Data Depth and Confidence Distribution

Dungang Liu[a], Regina Y. Liu[b], and Min-ge Xie[b]

[a]Department of Operations, Business Analytics and Information Systems, University of Cincinnati Lindner College of Business, Cincinnati, OH; [b]Department of Statistics, Rutgers University, New Brunswick, NJ

ABSTRACT

Fusion learning refers to synthesizing inferences from multiple sources or studies to make a more effective inference and prediction than from any individual source or study alone. Most existing methods for synthesizing inferences rely on parametric model assumptions, such as normality, which often do not hold in practice. We propose a general nonparametric fusion learning framework for synthesizing inferences for multiparameters from different studies. The main tool underlying the proposed framework is the new notion of *depth confidence distribution (depth-CD)*, which is developed by combining data depth and confidence distribution. Broadly speaking, a *depth-CD* is a data-driven nonparametric summary distribution of the available inferential information for a target parameter. We show that a *depth-CD* is a powerful inferential tool and, moreover, is an omnibus form of confidence regions, whose contours of level sets shrink toward the true parameter value. The proposed fusion learning approach combines *depth-CD*s from the individual studies, with each *depth-CD* constructed by nonparametric bootstrap and data depth. The approach is shown to be *efficient*, *general* and *robust*. Specifically, it achieves high-order accuracy and Bahadur efficiency under suitably chosen combining elements. It allows the model or inference structure to be different among individual studies. And, it readily adapts to heterogeneous studies with a broad range of complex and irregular settings. This last property enables the approach to use indirect evidence from incomplete studies to gain efficiency for the overall inference. We develop the theoretical support for the proposed approach, and we also illustrate the approach in making combined inference for the common mean vector and correlation coefficient from several studies. The numerical results from simulated studies show the approach to be less biased and more efficient than the traditional approaches in nonnormal settings. The advantages of the approach are also demonstrated in a *Federal Aviation Administration* study of aircraft landing performance. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2019
Accepted February 2021

KEYWORDS
Confidence distribution; Data depth; Evidence synthesis; Fusion learning; Heterogeneous studies; Indirect evidence; Multiparameter meta-analysis; *p*-Value function

1. Introduction

Powerful data acquisition technologies have greatly enabled the simultaneous collection of data from different sources in many domains. It is often impossible or inappropriate to simply aggregate all the data to draw inference, due to concerns over storage, privacy, cost constraints, or the desire to enhance inference by incorporating external or publicly available data sources, etc. Instead, one would need to combine the inference results from individual sources to draw an overall inference. Fusion learning refers to synthesizing inferences from multiple sources or studies to provide a more effective inference than that from any individual source or study alone.

1.1. A motivating example

We begin with an example to illustrate the need of efficient fusion learning. This example arose from a research project sponsored by the Federal Aviation Administration (FAA). The FAA, as the regulatory agency for air transportation safety, establishes guidelines for all air operations. For example, to ensure safe aircraft landings, FAA analysis has set guidelines recommending that the height of the aircraft at the crossing of runway threshold be around 15.85m and touchdown distance be around 432m from runway threshold. To help assess whether aircraft landings generally follow these guidelines, we can simply test the hypothesis $H_0 : \boldsymbol{\mu} = (15.85, 432)'$, where $\boldsymbol{\mu}$ is the mean vector for the height and distance. A sample of 2796 landing records (820 from Airbus and 1976 from Boeing) yields a combined sample mean of $(15.86, 432.95)'$, and a p-value of 0.942 from Hotelling's T^2 test. The finding would lead to the conclusion that aircraft landings generally comply with the FAA guidelines. Surprisingly, this conclusion appears to contradict the conclusion that we would draw from the two separate individual tests from Airbus and Boeing, with respective p-values 0.006 and 0.167. A closer examination of the scatter plots in Figure 1, of the two individual studies for Airbus and Boeing, indicates that the two samples do not appear to follow the same distribution and neither follows an elliptical distribution, and that the Boeing sample seems to be truncated on the right. This casts doubt on the aforementioned conclusion of landing operations meeting the FAA guidelines, and suggests the need of a nonparametric test for the hypothesis that both landing operations from Airbus and Boeing meet the FAA

CONTACT Min-ge Xie mxie@stat.rutgers.edu Department of Statistics, Rutgers University, New Brunswick, NJ 08903.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

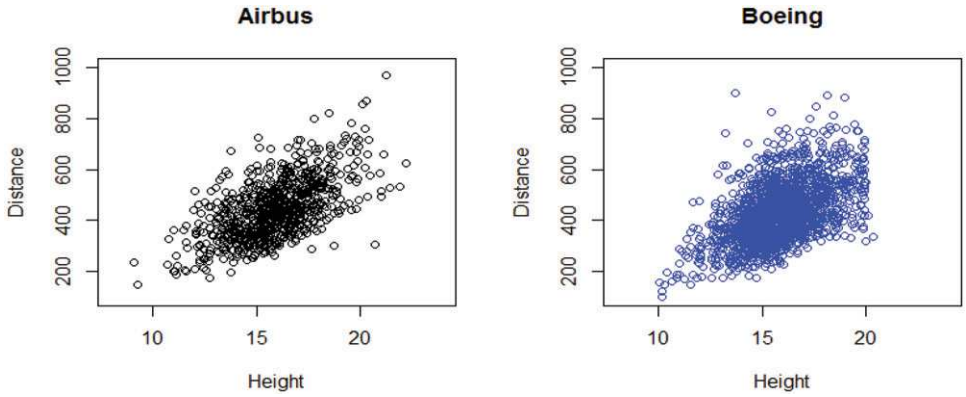

Figure 1. Scatterplots of Distance versus Height for the two aircraft makes, Airbus and Boeing.

guidelines, that is, $H : \mu_{\text{Airbus}} = \mu_{\text{Boeing}} = (15.85, 432)$. More importantly, this example shows that blindly aggregating data from different data sources may not necessarily yield correct overall inferences. This example is discussed further in Section 7.

1.2. Outline of Proposed Approach and Highlights of Results

In this article, we develop an efficient nonparametric approach for fusion learning to make inference for the common features or parameters shared by different studies. This approach consists of two key parts: (a) We develop a general nonparametric inference procedure to ascertain a valid inference for each individual study by applying the notion of depth confidence distribution (depth-CD) and its associated depth confidence curve (depth-CV). Specifically in this article, we construct a depth-CD using data depth and bootstrap and show the depth-CD as a comprehensive summary distribution of all the inferential information for the target parameter; (b) We derive an overall combined inference by suitably combining the depth-CVs from the individual studies. Our proposed approach for individual-study inference and that for the combined inference are completely nonparametric and data driven, and broadly applicable without any model assumptions. For instance, this can substantially broaden the scope of the existing meta-analysis and evidence synthesis, where common practice routinely requires parametric models, often the normality assumption, see, for example, Normand (1999) and Sutton and Higgins (2008).

The proposed fusion learning framework is established based on depth confidence distribution (depth-CD), which is a new powerful inference tool developed in this article by using three distinct concepts: confidence distribution (e.g., Xie and Singh 2013; Schweder and Hjort 2016), data depth (e.g., Liu 1990; Liu, Parelius, and Singh 1999; Zuo and Serfling 2000), and bootstrap (Efron 1979). Simply put, a depth-CD is a sample-dependent distribution function defined on the space of the target parameter, which summarizes all the information from the data that is relevant for the inference of the parameter. Based on the evidence in the given data, a depth-CD can also be viewed as a reference function that reflects the plausibility or "confidence" associated with each possible parameter value in the parameter space. We investigate general properties of depth-CD, in particular the following three, in Sections 3.2-3.4,

- ($\mathcal{P}$-1) a depth-CD is an omnibus form of confidence regions at all confidence levels;
- ($\mathcal{P}$-2) a depth-CD is an omnibus form of p-values for testing any parameter value on the entire parameter space;
- ($\mathcal{P}$-3) the contours of the level sets of a depth-CD shrink toward the true value of the parameter as the sample size increases.

These properties show that a depth-CD is useful in yielding all inference outcomes commonly sought in practice, and also that it is a versatile tool for nonparametric fusion learning.

Under the proposed general depth-CD fusion learning framework, we develop an efficient nonparametric fusion learning approach by fusing the depth-CDs from individual studies where the depth-CD of each study is constructed from data depth and nonparametric bootstrap as described in Section 4. The fused output, similar to the individual input, remains a distribution function on the parameter space, which now depicts the level of "confidence" in assuming each possible parameter value in view of the totality of all available evidence gathered from all studies. This combined depth-CD, following $\mathcal{P}$-1, 2, 3 above, can readily provide an overall inference such as confidence regions, p-values, or consistent point estimators.

The proposed fusion approach is shown to be efficient, general and robust. More precisely, it is efficient, as it achieves high-order accuracy and Bahadur efficiency under suitably chosen combining elements, as shown in Section 5.1. It is general, as (a) it covers multiparameter settings, (b) it is nonparametric, and (c) it permits flexible choices of mappings of input functions, weighting schemes and methods for deriving each individual depth-CD, across all studies. Such choices are often needed to account for the different circumstances or degrees of trustworthiness surrounding each individual study. It is robust, as it adapts efficiently to the fusion of heterogeneous studies, covering a wide range of complex and irregular studies, as investigated in Section 5.2. In fact, our fusion approach covers the particularly challenging setting where the target parameter may not be even estimable in some subset of studies, such as in the case of incomplete studies. Although the target parameter vector may not be estimable in incomplete studies, those studies often contain information from their data that can contribute to

the overall inference of the target parameter, as the information among different component parameters is often correlated; see, for example, Liu, Liu, and Xie (2015). This data information from incomplete studies is often regarded as *indirect evidence*. Therefore, our fusion approach can incorporate both direct and indirect evidence, all in a nonparametric manner. This is a desirable property since it gains efficiency in the overall inference, as shown in Section 6.1.

We present an extensive comparison study of our fusion method with several existing methods in the setting of making inference for a common mean vector, in three data scenarios. The results can be summed up as three advantages of our method, namely, in the absence of the normality assumption: (i) it preserves inference accuracy in hypothesis testing and confidence regions; (ii) its point estimator has less bias and is more efficient; and (iii) it achieves a gain of efficiency in the presence of heterogeneous studies. We also present a numerical study of our method in meta-analysis of correlation coefficients in Section 6.2. There we observe that traditional methods may yield misleading conclusions, while ours remains valid in both normal and nonnormal cases.

The remaining article is organized as follows. Section 2 gives a general setup for fusion learning. Section 3 covers a brief review of confidence distribution (CD) and data depth, and then the development of depth-CD and depth confidence curve (depth-CV) for multiparameter inference. Section 4 provides a concrete procedure for constructing a depth-CV by using bootstrap and data depth. Section 5 develops nonparametric fusion learning by combining the depth CDs derived from individual studies. Section 6 covers all simulation studies. Section 7 applies our fusion learning approach to conduct the FAA study of aircraft landing performance. Section 8 contains more comments and discussions.

2. A General Problem Setting for Fusion Learning

We consider the problem of fusion learning in a general setting. Suppose that K independent studies are available for analysis to address the same scientific or business question. Let

$$\boldsymbol{X}_{k,1}, \boldsymbol{X}_{k,2}, \ldots, \boldsymbol{X}_{k,n_k}, \text{ iid } \sim \mathcal{F}_k, \tag{1}$$

be the sample from the kth study, $k = 1, \ldots, K$, where $\mathcal{F}_k$ is an unknown p_k-dimensional multivariate distribution. Assume that the parameter of interest $\boldsymbol{\theta}_k$ is a finite-dimensional functional of $\mathcal{F}_k$, which can be scalar or vector-valued. Assume that

$$\boldsymbol{\theta} \equiv \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_K. \tag{2}$$

The goal is to make an efficient inference for $\boldsymbol{\theta}$ by fusing the information from all K studies, without assuming specific parametric forms of the distributions $\mathcal{F}_k(\boldsymbol{\theta}_k)$. This setting covers:

Example 1 (common mean inference). Let $\boldsymbol{\theta}_k = \int \mathbf{x}\, \mathrm{d}\mathcal{F}_k(\mathbf{x})$ be the mean of the distribution $\mathcal{F}_k$, $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_K$ is the (p-dimensional) common mean of the K unknown distributions. We are interested in constructing a confidence region for $\boldsymbol{\theta}$ or testing the hypothesis $\Sigma_0 : \boldsymbol{\theta} = \boldsymbol{\mu}$ versus $\Sigma_1 : \boldsymbol{\theta} \neq \boldsymbol{\mu}$ for a particular value $\boldsymbol{\mu}$.

Example 2 (correlation inference). Consider the correlation coefficient of any two components of the p-dimensional distribution $\mathcal{F}_k$. Let $\boldsymbol{\theta}_k$ include all such pairwise correlation coefficients. Then $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_K$ is a parameter vector of dimension $p(p-1)/2$. We are interested in testing the hypothesis $\Sigma_0 : \boldsymbol{\theta} = \mathbf{0}$ versus $\Sigma_1 : \boldsymbol{\theta} \neq \mathbf{0}$.

These two examples will be used throughout the development and simulations to illustrate the proposed fusion approach. In various scenarios these examples help showcase the merits of our approach, described briefly as efficient, general and robust in the Introduction. To elaborate further, our fusion framework is general because it requires no specific parametric forms of the underlying distributions $\mathcal{F}_k$. It is also robust because it permits a broad range of heterogeneity among studies: (i) the individual studies do not have be homogeneous in terms of their designs, reporting formats, models, and inference methods; (ii) the studies can have different types of data (e.g., continuous, binary or ordinal responses); (iii) the studies can be analyzed using different models, such as linear regression models for continuous outcomes in some studies and logistic regression models for binary outcomes in others, and (iv) the individual studies can even use different inference methods, for instance, estimating the population location by the sample mean, the trimmed mean or the median as dictated by the specific situation of each study; and v) our fusion framework does not require that the parameter $\boldsymbol{\theta}_k$ be estimable in all studies.

To be more precise with the last point, our CD fusion approach applies even if the parameter of interest $\boldsymbol{\theta}_k$ is not estimable in some studies, as long as there exists a known continuous mapping from the parameter space $\boldsymbol{\Theta}$ (of $\boldsymbol{\theta}$) to a lower-dimensional space $\boldsymbol{\Theta}_k$ such that

$$\tilde{\boldsymbol{\theta}}_k = \boldsymbol{f}_k(\boldsymbol{\theta}_k) \tag{3}$$

is estimable. Similar formulation of partially estimable parameters also arose in the applications in (Sutton and Higgins 2008; Liu, Liu, and Xie 2015). Obviously, when all $\boldsymbol{f}_k$'s are identifiable mappings, this setting reduces to the case where all $\boldsymbol{\theta}_k$'s are estimable. Our fusion approach is thus adaptable to indirect evidence. Two numerical examples in Sections 6.1 and 7 illustrate how our approach gains efficiency in the final combined inference from incorporating indirect evidence.

3. Depth-CD and Depth-CV for Multiparameter Inference

3.1. Reviews: Confidence Distribution (CD) and Data Depth

3.1.1. Confidence Distribution (CD) and Confidence Curve (CV) for Scalar Parameter

The idea of the confidence distribution (CD) is borne out of the wish to use a sample-dependant distribution function, rather than a point estimate or an (confidence) interval estimate, to estimate an unknown parameter. For a scalar parameter $\theta \in \Theta$, a function $H_n(\cdot) \equiv H_n(\boldsymbol{X}_n, \cdot)$ is said to be a CD function for θ if it meets these two requirements: (i) given a sample $\boldsymbol{X}_n$, it is a distribution function on Θ; and (ii) at the true parameter value $\theta = \theta^o$, $H_n(\theta^o) \equiv H_n(\boldsymbol{X}_n, \theta^o)$, as a function of the sample

$\boldsymbol{X}_n$, follows the uniform distribution on (0,1) (Schweder and Hjort 2002; Singh, Xie, and Strawderman 2005). Essentially, (i) says that a CD function is a *'distribution estimate'* dependent on the observed sample, and (ii) ensures that a CD function carries frequentist properties in terms of repeated sampling. For instance, under (ii), $(-\infty, H_n^{-1}(1-\alpha))$ is a $(1-\alpha)$ confidence interval, and also $H_n(\theta^o)$ can be used as a p-value for testing the hypotheses $\Omega_0 : \theta \leq \theta^o$ versus $\Omega_1 : \theta > \theta^o$. More precisely, given a dataset and an inferential procedure, a CD function represents a set of confidence intervals for all possible confidence levels. It describes a *distribution of confidence* associated with each θ value in Θ.

Note that, conditional on the observed sample $\mathbf{X}_n$, a CD function $H_n(\theta) \equiv H_n(\boldsymbol{X}_n, \theta)$ is a distribution function on the parameter space Θ. Let θ^* be a random variable following the distribution $H_n(\cdot)$. We refer to θ^* as a *CD-random variable*. Conditional on the given data, we can used simulated samples θ^*'s from $H_n(\cdot)$ to carry out inference, as discussed in Section 4.

To illustrate the CD inference approach, we consider simple example with a sample $\boldsymbol{x} = \{x_i, i = 1, \ldots, n\}$ from $N(\theta, 1)$, where the mean θ is the parameter of interest. A natural CD for θ is $\mathcal{N}(\bar{x}_n, 1/n)$ or equivalently its cumulative distribution function $H_n(\theta) = \Phi(\sqrt{n}(\theta - \bar{x}_n))$. Given a sample $\boldsymbol{x}$, the function $H_n(\theta)$ is a distribution function on the parameter space Θ, and it carries all commonly used inference outcomes. For instance, $(H_n^{-1}(\alpha/2), H_n^{-1}(1-\alpha/2)) = (\bar{x}_n + \Phi^{-1}(\alpha/2)/\sqrt{n}, \bar{x}_n + \Phi^{-1}(1-\alpha/2)/\sqrt{n})$ is a $(1-\alpha)$ confidence interval for θ, for any $0 < \alpha \leq 1$; the mean/median of $(H_n^{-1}(.5)\ (= \bar{x}_n)$ is a point estimate for θ; and the tail mass $H_n(b) = \Phi(\sqrt{n}(b - \bar{x}_n)))$ is a p-value for testing the one-sided hypothesis $K_0 : \theta \leq b$ versus $K_1 : \theta > b$. The curve in Figure 2(a) is a CD function given a random sample of size $n = 20$. The dashed lines there help illustrate all types of inference outcomes from a CD function, including a point estimate of 0.11, a 90% confidence interval of (-0.26, 0.48), and a p-value of 0.31 for testing $\Omega_0 : \theta \leq 0$ versus $\Omega_1 : \theta > 0$.

For the ease of visualization of confidence intervals of different levels, the distributional form of a CD $H_n(\cdot) \equiv H_n(\boldsymbol{X}_n, \cdot)$ seen in in Figure 2(a) can be expressed alternatively as a confidence curve (CV) seen in Figure 2(b) which is defined as

$$CV_n(\theta) = 1 - 2|H_n(\theta) - 0.5| = 2\min\{H_n(\theta), 1 - H_n(\theta)\}, \quad (4)$$

(see Xie and Singh 2013, pp. 29 and 31; Schweder and Hjort 2016, pp. 10–14). While the CD function $H_n(\theta)$ represents the upper limits of one-sided confidence intervals, the confidence curve $CV_n(\theta)$ gives an omnibus form of the limits of two-sided confidence intervals. In Figure 2(b), the two limits of a 90% confidence interval identified by the two points on the confidence curve at the height $\alpha = 0.1$ are exactly the same as those obtained from Figure 2(a). Furthermore, following the duality between confidence intervals and hypothesis testing, $CV_n(\theta^0)$ can serve as a p-value function for the two-sided hypothesis testing, $\Omega_0 : \theta = \theta^o$ versus $\Omega_1 : \theta \neq \theta^o$, for any $\theta^o \in \Theta$. Also, the confidence curve peaks at the median of the CD function, that is, $CV_n^{-1}(1) = 0.11$ as shown in Figure 2(b), which yields a median-unbiased estimate for θ.

Without linking to CD, the concept of CV has actually been explored in Birnbaum (1961), Blaker (2000), and Blaker and Spjøtvoll (2000) for a scalar parameter θ. In fact, Blaker and Spjøtvoll (2000) interpreted a CV as a summary of "how each parameter value is ranked in view of the data" from the peak decreasing gradually along the tails. This ranking interpretation of the CV in fact suggests a natural extension of the CD to the multiparameter setting by incorporating the notion of data depth, which has been developed to establish a center-outward ordering of multivariate observations. We will develop this extension after the brief review of data depth and its properties.

3.1.2. Data Depth and Center-outward Ordering of Multivariate Data

Data depth is a way to measure how deep or central a given point is with respect to a multivariate sample cloud, say $\{\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_m\} \sim F \in \boldsymbol{R}^p$, or to its underlying distribution F. It naturally yields a measure of "outlyingness" and thus also a center-outward ordering of these multivariate points. Common depth functions include, Mahalanobis depth (MD) (Mahalanobis 1936), half-space depth (HD) (Hodges 1955; Tukey 1975), simplicial depth (SD) (Liu 1990), among others.

Using simplicial depth as an example, the depth at point $\mathbf{z} \in \boldsymbol{R}^p$ with respect to F is $D_F(\mathbf{z}) = P_F\{\mathbf{z} \in s[\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{p+1}]\}$, where $s[\boldsymbol{\xi}_{i_1}, \ldots, \boldsymbol{\xi}_{i_{p+1}}]$ is the p-dimensional simplex with vertices $\{\boldsymbol{\xi}_{i_1}, \ldots, \boldsymbol{\xi}_{i_{p+1}}\}$. The empirical version of $D_F(\mathbf{z})$ is $D_{\hat{F}}(\mathbf{z}) = \sum \mathbf{1}_{\{\mathbf{z} \in s[\boldsymbol{\xi}_{i_1}, \ldots, \boldsymbol{\xi}_{i_{p+1}}]\}} \big/ \binom{m}{p+1}$. In R^2,

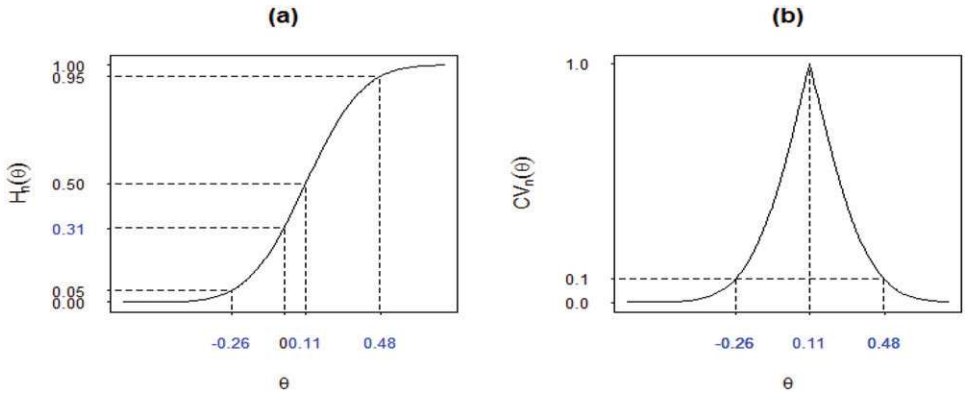

Figure 2. The curves represent a confidence distribution function (a) and the corresponding confidence curve (b) for the mean parameter θ in the normal distribution $N(\theta, 1)$. They are obtained based on a sample $\boldsymbol{x} = \{x_i, i = 1, \ldots, 20\}$ from $N(0, 1)$. Illustrated is how to draw commonly used inferential outcomes such as a point estimate of 0.11, a 90% confidence interval of $(-0.26, 0.48)$, and a p-value of 0.31 for testing the hypothesis $\Omega_0 : \theta \leq 0$ versus $\Omega_1 : \theta > 0$.

$D_{\hat{F}}(\mathbf{z}) = \sum_{i,l,k} \mathbf{1}_{\{\mathbf{z}\in\Delta(\boldsymbol{\xi}_i,\boldsymbol{\xi}_l,\boldsymbol{\xi}_k)\}} \big/ \binom{m}{3}$, the fraction of the triangles $\Delta(\boldsymbol{\xi}_i, \boldsymbol{\xi}_l, \boldsymbol{\xi}_k)$ generated from the sample that contains $\mathbf{z}$ inside. Clearly, a point with a larger depth value indicates that it lies more central within the data cloud or closer to the center of the distribution.

By computing the depth values for all data points $\boldsymbol{\xi}_i$'s and then ordering $\boldsymbol{\xi}_i$'s by their descending depth values, we can obtain the depth order statistics $\{\boldsymbol{\xi}_{[1]}, \ldots, \boldsymbol{\xi}_{[m]}\}$ with an ordering from the deepest (or most central) point $\boldsymbol{\xi}_{[1]}$ to the most outlying $\boldsymbol{\xi}_{[m]}$. This center-outward ordering naturally gives rise to nested central regions expanding with increasing levels of probability coverage. The convex region spanning the deepest $(1-\alpha)n$ sample points is referred to as the $(1-\alpha)$-*central region*. Formally, the population and empirical versions of $(1-\alpha)$-central region can be expressed respectively as

$$A_{(1-\alpha);F} = \{\mathbf{z} : C_F(\mathbf{z}|D_F) \geq \alpha\} \quad \text{and}$$
$$A_{(1-\alpha);\hat{F}} = \{\mathbf{z} : C_{\hat{F}}(\mathbf{z}|D_{\hat{F}}) \geq \alpha\}, \quad 0 < \alpha < 1. \tag{5}$$

Here, $C_F(\mathbf{z}|D)$ and $C_{\hat{F}}(\mathbf{z}|D_{\hat{F}})$ are referred to as *centrality functions* with, respectively,

$$C_F(\mathbf{z}|D) = P_F\{\boldsymbol{\xi} : D_F(\boldsymbol{\xi}) \leq D_F(\mathbf{z})\} \quad \text{and}$$
$$C_{\hat{F}}(\mathbf{z}|D_{\hat{F}}) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\{D_{\hat{F}}(\boldsymbol{\xi}_i) < D_{\hat{F}}(\mathbf{z})\}}. \tag{6}$$

The central regions $A_{(1-\alpha);\hat{F}}$ are data driven and nonparametric, and are shown to be particularly useful for inference under asymmetric underlying distributions or nonstandard asymptotics.

Lemma 1 below shows important properties of centrality functions. Its proof is in the appendix (supplementary material).

Lemma 1. Let $\boldsymbol{\eta}$ be a random vector following a p-dimensional distribution F. The centrality function in Equation (6) satisfies these properties:
(a) (Uniform transformation) The transformed variable $C_F(\boldsymbol{\eta}|D_F)$ satisfies $C_F(\boldsymbol{\eta}|D_F) \sim U(0,1)$, provided that the depth contours $\{\boldsymbol{\eta} : D_F(\boldsymbol{\eta}) = t\}$ all have probability zero w.r.t. F for any $t > 0$.
(b) (Affine-invariance) Let A be a $p \times p$ nonsingular matrix and $\mathbf{b}$ a $p \times 1$ constant vector. If both $F(\cdot)$ and $D_F(\cdot)$ are affine invariant, that is, for any point $\mathbf{z} \in R^p$, $\tilde{F}(A\mathbf{z} + \mathbf{b}) = F(\mathbf{z})$ and $D_F(\mathbf{z}) = D_{\tilde{F}}(A\mathbf{z} + \mathbf{b})$, then so is the centrality function $C_F(\cdot)$, that is, $C_F(\mathbf{z}|D_F) = C_{\tilde{F}}(A\mathbf{z} + \mathbf{b}|D_{\tilde{F}})$.

Typically, depth functions have been used to rank *sample points* and provide a center-outward ordering of sample points in the *sample space*, as reviewed above. In this article, a depth function will be used instead to rank *parameter values* and provide an ordering of all parameter values in the *parameter space*. Specifically, instead of applying depth ordering to the sample $\boldsymbol{\xi}_i$'s drawn from the distribution $F(\cdot)$, we apply it to the sample CD-random variables $\boldsymbol{\theta}_i^*$'s drawn from the confidence distribution $H_n(\cdot)$. This center-outward ordering in the parameter space can be interpreted as the plausibility of each parameter value relative to the others. This line of interpretation underlies the proposed CD fusion learning framework and justifies the resulting inferences, for example, using the central regions formed by $\boldsymbol{\theta}^*$'s as confidence regions for the parameter of interest $\boldsymbol{\theta}$. This is elaborated further in Sections 3.2–3.4.

3.2. Depth-CD and Depth-CV: an Omnibus Form of Confidence Regions

The definition of a CD as a sample-dependent distribution function on the parameter space that can represent confidence regions for all possible confidence levels applies to a scalar parameter (as seen in Section 3.1.2) as well as a vector parameter. However, mathematical rigor for multiparameter CDs has so far been elusive, since the region created by the inversion of a multivariate cumulative distribution function may not be unique or suitable for providing any natural form of inference. To this end, (Singh, Xie, and Strawderman 2007; Xie and Singh 2013; Schweder and Hjort 2016) have proposed to limit confidence regions within a certain subclass. In this article, we propose to consider the set of center-outward nested confidence regions derived from using data depth, which we refer to as depth-CDs. The depth-CDs provide a natural extension of the CD concept from the scalar setting to the multiparameter setting.

As discussed in Section 3.1.1, a confidence curve (CV), as plotted in Figure 2(b), can provide two-sided confidence intervals for a scalar parameter of all levels, with the intervals expanding outward to two tails as the level of confidence increases. The CV in Figure 2(b) clearly ranks parameter values in the parameter space from the center outward as the level α decreases. In fact, the CV defined in Equation (4) can be re-expressed, using data depth and its associated centrality function, as

$$C_{H_n}(\theta|D_{\text{HS}}) = P_{H_n}\{\xi : D_{\text{HS}}(\xi) \leq D_{\text{HS}}(\theta)\} \tag{7}$$
$$= 2\min\{H_n(\theta), 1 - H_n(\theta)\} = \text{CV}_n(\theta), \tag{8}$$

where $D_{\text{HS}}(\boldsymbol{\vartheta}) = \inf_E\{P_{H_n}(E) : E \text{ is a closed half-space in } \mathbb{R}^p \text{ and } \boldsymbol{\vartheta} \in E\}$ is the half-space depth when $p = 1$ and P_{H_n} is the probability measure corresponding to the CD $H_n(\cdot)$ on the parameter space, that is, $P_{H_n}\{(-\infty, t]\} = H_n(t)$.

By extending Equation (7), we can directly define a CV for a parameter vector $\boldsymbol{\vartheta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ as

$$\text{CV}_n(\boldsymbol{\vartheta}) =: C_{H_n}(\boldsymbol{\vartheta}|D) = P_{H_n}\{\boldsymbol{\xi} : D(\boldsymbol{\xi}) \leq D(\boldsymbol{\vartheta})\}, \tag{9}$$

where D is a depth function with the associated probability measure P_{H_n} from a multivariate depth-CD $H_n(\cdot)$ on $\boldsymbol{\Theta}$. Formally, we define multivariate depth- CD and CV as follows:

Definition 1 (depth-CD and depth-CV). (A) A function $H_n(\cdot) \equiv H_n(\boldsymbol{X}_n, \cdot)$ on $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is called a depth confidence distribution (depth-CD) associated with depth function D for a vector-valued parameter $\boldsymbol{\theta}$, if (i) it is a distribution function on the parameter space $\boldsymbol{\Theta}$ for any fixed sample set $\boldsymbol{X}_n$; and (ii) the $(1-\alpha)$ "central region" of the distribution $H_n(\cdot)$, $\mathcal{R}_{1-\alpha}(H) = \{\boldsymbol{\vartheta} \in \boldsymbol{\Theta} : C_{H_n}(\boldsymbol{\vartheta}) \geq \alpha\}$, is a confidence region for $\boldsymbol{\theta}$ with a coverage probability of $(1-\alpha)$. Here, the centrality function associated with depth D and CD $H_n(\cdot)$, that is, $C_{H_n}(\boldsymbol{\vartheta})$, is also referred to as depth confidence curve (depth-CV).
If the statements in (ii) holds only asymptotically, then we refer to H_n and C_{H_n} as asymptotic depth-CD and asymptotic depth-CV, respectively.

Continuing with half-space depth in the scalar setting, we see that the result in Lemma 1(a) resembles $2\min\{G(Z), 1 - G(Z)\}$ for a univariate random variable Z with its cumulative

distribution function G. This result ensures that $\{\theta : \mathrm{CV}_n(\theta) \geq \alpha\} = [H_n^{-1}, H_n^{-1}(b)]$, with $a = \min(\alpha/2, 1-\alpha/2)$ and $b = \max(\alpha/2, 1-\alpha/2)$, is a $(1-\alpha)$ confidence interval θ. Similarly, by Lemma 1(a), the $(1-\alpha)$ "central region" of depth-CD $H_n(\cdot)$ or depth-CV $CV_n(\cdot)$

$$\mathcal{R}_{1-\alpha}(H_n) = \{\boldsymbol{\vartheta} \in \boldsymbol{\Theta} : C_{H_n}(\boldsymbol{\vartheta}) \geq \alpha\} \tag{10}$$

leads to a $(1-\alpha)$ confidence region for a multiparameter $\boldsymbol{\theta}$. In conclusion, Lemma 1 ensures that the depth-CD and depth-CV (defined in Definition 1) can provide valid center-outward confidence regions of all levels.

For convenience, we use the familiar bivariate normal to illustrate the above framework of depth-CD inference, where the depth contours have explicit expressions, although the normality assumption is not required in our general framework.

Example 3 (Bivariate normal distribution). Given a random sample $\{\boldsymbol{Y}_i\}_{i=1}^n$ from a bivariate normal distribution $\mathrm{BN}(\boldsymbol{\theta}, \Sigma)$, we consider making inference for the mean parameter $\boldsymbol{\theta}$. Let $\bar{\boldsymbol{Y}}_n = \sum_{i=1}^n \boldsymbol{Y}_i/n$. Assuming that Σ is known, then the bivariate normal distribution $\mathrm{BN}(\bar{\boldsymbol{Y}}_n, \Sigma)$ is a depth-CD for $\boldsymbol{\theta}$, since: I) $\mathrm{BN}(\bar{\boldsymbol{Y}}_n, \Sigma)$ is a sample-dependent distribution function of the parameter space of $\boldsymbol{\theta}$, and II) the depth contours of $\mathrm{BN}(\bar{\boldsymbol{Y}}_n, \Sigma)$, using any depth mentioned in Section 3.1.2, provide valid center-outward confidence regions of all levels.

For a given simulated sample of size $\{\boldsymbol{Y}_i\}_{i=1}^{20}$ under $\boldsymbol{\theta} = (1,1)$, and $\Sigma = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, using Mahalanobis depth, we obtain the depth-CD on the parameter space $\boldsymbol{\Theta}$ as a 3D-surface plot in Figure 3(a). Projecting this 3D plot to $\boldsymbol{\Theta}$ (the two-dimensional plane below) gives depth contours in a gray-color heat map in Figure 3(b), where the brighter the region, the larger the depth value. A depth contour in Figure 3(b) connects the points in $\boldsymbol{\Theta}$ which have the same depth value $D_H(\cdot)$. Corresponding to Figure 3(b), a similar projection of the 3D-surface plot of the depth-CV results in Figure 3(c) showing the contours which connect the points in $\boldsymbol{\Theta}$ with the same centrality value $C_{H_n}(\cdot|D)$. For instance, the peak of depth-CV corresponds to the deepest (or most central) point in Figure 3(c), which also corresponds to the highest point in the depth-CD in Figure 3(a) as well as the deepest point in Figure 3(b).The depth-CV in Figure 3(c) ranks the plausibility of each possible value of the bivariate parameter space $\boldsymbol{\Theta}$. For instance, the black round dot being on the contour with centrality value 0.9 implies that this particular parameter value is deeper than 90% of all the possible parameter values w.r.t. the confidence distribution $H(\cdot)$ or more plausible than 90% of all the possible parameter values in $\boldsymbol{\Theta}$.

Inferences about $\boldsymbol{\theta}$ can be derived from the depth-CD or depth-CV with its contours in Figure 3. For example, the largest elliptical region within the contour of centrality value .1 (labeled with a solid triangle) in Figure 3(c) is a 90% confidence region for $\boldsymbol{\theta}$. The deepest point in all three plots (0.94, 0.92), marked by a cross, can be considered the most plausible parameter value and thus a suitable point estimate for $\boldsymbol{\theta}$. This point estimate is shown to be consistent later in Section 3.4.

When Σ is unknown, the $\mathrm{BN}(\bar{\boldsymbol{Y}}_n, \hat{\Sigma})$ can be shown to be a depth-CD for $\boldsymbol{\theta}$ asymptotically. Here $\hat{\Sigma}$ is the sample covariance matrix. Similar illustration plots and asymptotic inferences can be drawn accordingly.

3.3. Depth-CD and Depth-CV: An Omnibus Form of p-Values

To show how depth-CD and depth-CV can give rise to an omnibus form of p-values, we first justify that, for a given $\boldsymbol{\vartheta} \in \boldsymbol{\Theta}$, the depth CV $C_{H_n(\boldsymbol{X}_n,\cdot)}(\boldsymbol{\vartheta})$ is a *limiting p-value* for testing the hypothesis $\Omega_0 : \boldsymbol{\theta} = \boldsymbol{\vartheta}$ versus $\Omega_1 : \boldsymbol{\theta} \neq \boldsymbol{\vartheta}$. Liu and Singh (1997) defines a sequence of statistics p_n to be a *limiting p-value* if $p_n \in [0,1]$ and p_n satisfies

(a) $\limsup_{n\to\infty} P_F\{p_n \leq t\} \leq t$ for all $F \in \Omega_0$ and any $t \in [0,1]$; and
(b) $p_n \to 0$ in probability for all $F \in \Omega_1$, as $n \to \infty$.

Note that (a) is required to ensure the testing size and (b) to ensure that the asymptotic test power goes to 1, as the sample size increases.

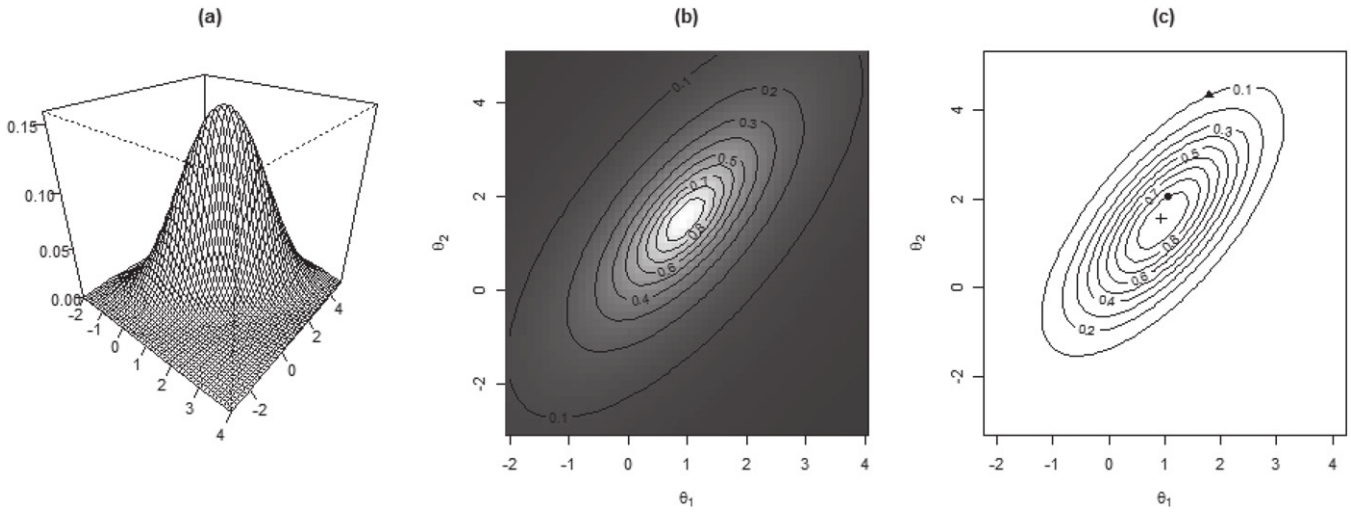

Figure 3. Illustrations of a depth-CD for the mean parameter $\boldsymbol{\theta}$ in $\mathrm{BN}(\boldsymbol{\theta}, \Sigma)$: (a) a 3D-surface plot for the depth-CD; (b) a gray-color heat map for the depth contours with Mahalanobis depth values. The contours in (c) in the parameter space connect the parameter values of the same centrality value. (c) illustrates the utility of the depth-CV for drawing confidence regions, p-values, and a point estimate. The plots are based on a simulated sample of size $n = 20$ and $\boldsymbol{\theta} = (1,1)$.

To see why $C_{H_n(\boldsymbol{X}_n,\cdot)}(\boldsymbol{\vartheta})$ is a limiting p-value, we need the simple but useful result below:

Proposition 1. The statement that $\mathcal{R}_{1-\alpha}(H_n(\boldsymbol{X}_n,\cdot))$ is a confidence region for $\boldsymbol{\theta}$ with a coverage probability of $(1-\alpha)$ (Requirement (ii) in Definition 1) is equivalent to the statement that $C_{H_n}(\boldsymbol{\theta}^o) \equiv C_{H_n(\boldsymbol{X}_n,\cdot)}(\boldsymbol{\theta}^o)$, as a function of the sample $\boldsymbol{X}_n$, follows the uniform distribution on [0,1], where $\boldsymbol{\theta}^o$ is the true value of $\boldsymbol{\theta}$.

In view of Proposition 1, $C_{H_n(\boldsymbol{X}_n,\cdot)}(\boldsymbol{\vartheta})$ is a limiting p-value as long as $C_{H_n(\boldsymbol{X}_n,\cdot)}(\boldsymbol{\vartheta}) \to 0$ in probability for all $\boldsymbol{\vartheta} \neq \boldsymbol{\theta}^o$, which is a mild condition that holds generally. Such a CD p-value and the classical p-value share a similar idea in their approaches, as they both try to assess the degree of inconsistency between the given data and the target null hypothesis by comparing a fixed *value* w.r.t. a *reference distribution*. But they are fundamentally different, since

- A classical p-value is derived by comparing the observed value $\boldsymbol{t}$ of a statistic $\boldsymbol{T}$ with a reference distribution over the *sample space*, that is, the null distribution of $\boldsymbol{T}$, say $F_{T,0}$;
- A CD p-value is derived by comparing a hypothesized value $\boldsymbol{\vartheta}$ of the parameter $\boldsymbol{\theta}$ with a reference distribution over the *parameter space*, that is, the depth-CD $H_n(\boldsymbol{X}_n,\cdot)$.

The key difference is that the assessment of statistical significance, namely, measuring the outlyingness of the value ($\boldsymbol{t}$ or $\boldsymbol{\vartheta}$) w.r.t. the reference distribution ($F_{T,0}$ or $H_n(\boldsymbol{X}_n,\cdot)$), is performed in different (sample or parameter) spaces.

The CD p-value has several advantages, including:
(A1) Given an inference procedure, the reference distribution $H_n(\boldsymbol{X}_n,\cdot)$ is determined solely by the sample $\boldsymbol{X}_n$ and it does not depend on the specified value in the null hypothesis. This is different from the classical p-value method where the reference distribution must satisfy the null constraint and thus may vary depending on the null value.
(A2) Since the CD method does not need to rely on a test statistic, the CD p-value essentially serves simultaneously as a test statistic and a p-value. Thus, it compress the usual three-step test procedure in the classical p-value approach into just one, bypassing: (i) construct an explicit test statistic, and then (ii) establish or approximate its sampling distribution. This advantage will be elaborated further in Section 4 where bootstrap is used to devise depth-CD functions.
(A3) The CD reference distribution $H_n(\boldsymbol{X}_n,\cdot)$ carries infinitely many p-values $\{C_{H_n(\boldsymbol{X}_n,\cdot)}(\boldsymbol{\vartheta}) : \boldsymbol{\vartheta} \in \boldsymbol{\Theta}\}$ for a set of hypothesis testing problems $\{\{\Omega_0 : \boldsymbol{\theta} = \boldsymbol{\vartheta} \text{ versus } \Omega_1 : \boldsymbol{\theta} \neq \boldsymbol{\vartheta}\} : \boldsymbol{\vartheta} \in \boldsymbol{\Theta}\}$. This implies that $C_{H_n}(\cdot)$ provides a distribution of p-values over $\boldsymbol{\Theta}$; see for example Figure 3(c), where the contours of centrality values can be used as p-value contours. As a p-value in testing $\boldsymbol{\theta} = \boldsymbol{\vartheta}$ is generally viewed as the strength of evidence from the data in support of the assumption $\boldsymbol{\theta} = \boldsymbol{\vartheta}$, $C_{H_n}(\boldsymbol{\vartheta})$ can be viewed as a measure of the plausibility of assuming $\boldsymbol{\theta} = \boldsymbol{\vartheta}$. The smaller the value of $C_{H_n}(\boldsymbol{\vartheta})$, the less plausible $\boldsymbol{\theta} = \boldsymbol{\vartheta}$. In Figure 3(c), for example, the parameter value marked by the solid triangle is much less plausible than the one marked by the solid round dot. To sum up, a depth-CD provides a simple but comprehensive summary of data evidence in the sense that a *single* reference distribution $H_n(\boldsymbol{X}_n,\cdot)$ can express the plausibility of every $\boldsymbol{\theta}$ value for the entire parameter space $\boldsymbol{\Theta}$. In contrast, the reference distribution $F_{T,0}$ used in the classical p-value method expresses only the plausibility of the specific parameter value under the null hypothesis. For instance, it does not simultaneously provide the plausibility of $\boldsymbol{\theta}$ values in the alternative parameter space.

Following along Example 3, for a given sample there, the centrality value $C_{H_n}(\boldsymbol{\theta}_0)$ in the depth-CV as in Figure 3(c) would be a p-value for testing $\Omega_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $\Omega_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

3.4. Depth-CD: Its Deepest Point As a Consistent Estimator

Given a dataset $\boldsymbol{X}_n$ and a depth-CD $H_n(\boldsymbol{X}_n,\cdot)$, we propose to use the deepest point of the depth-CD $H_n(\boldsymbol{X}_n,\cdot)$ or equivalently the maximum point of the centrality function $C_{H_n(\boldsymbol{X}_n,\cdot)}(\cdot)$, denoted by $\hat{\boldsymbol{\theta}}_n^{\text{MCE}}$, as a point estimate for the parameter of interest $\boldsymbol{\theta}$. That is,

$$\hat{\boldsymbol{\theta}}_n^{\text{MCE}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} C_{H_n(\boldsymbol{X}_n,\cdot)}(\boldsymbol{\theta}). \tag{11}$$

This estimate is referred to as a *maximum centrality estimate* (MCE). Note that, in $\boldsymbol{R}^1$, the MCE corresponds to the highest value of CV in Figure 2(b) or, equivalently, the median (also central most point) of the CD in Figure 2(a). The estimate $\hat{\boldsymbol{\theta}}_n^{\text{MCE}}$ extends to general multiparameter settings the idea of using the "median" or deepest point of a CD function for point estimation. We show below that $\hat{\boldsymbol{\theta}}_n^{\text{MCE}}$ is a consistent estimator under some mild conditions.

Proposition 2. Assume that for any $\epsilon > 0$, as $n \to \infty$,

$$\Delta_n(\epsilon) = \max_{\boldsymbol{\vartheta}_i,\boldsymbol{\vartheta}_j\in\{\boldsymbol{\vartheta}:C_{H_n(\boldsymbol{X}_n,\cdot)}(\boldsymbol{\vartheta})=\epsilon\}} ||\boldsymbol{\vartheta}_i - \boldsymbol{\vartheta}_j|| \to 0$$

in probability. Then, $\hat{\boldsymbol{\theta}}_n^{\text{MCE}} \to \boldsymbol{\theta}^o$ in probability. Furthermore, if $\Delta_n(\epsilon) = O_p(a_n)$ for a nonnegative sequence $a_n \to 0$, then $\hat{\boldsymbol{\theta}}_n^{\text{MCE}} - \boldsymbol{\theta}^o = O_p(a_n)$.

The condition $\Delta_n(\epsilon) \to 0$ basically requires that the depth contours $\{\boldsymbol{\vartheta} : C_{H_n(\boldsymbol{X}_n,\cdot)}(\boldsymbol{\vartheta}) = \epsilon\}$ (e.g., the contours in Figure 3(c)) shrink to a single point (e.g., the cross in Figure 3(c)) as the sample size $n \to \infty$. For scalar parameters, $\Delta_n(\epsilon)$ is the distance between the two intersection points of the CV and the horizontal dashed line in Figure 2(b). The condition $\Delta_n(\epsilon) \to 0$ means that as information increases ($n \to \infty$), the depth-CD concentrates onto a shrinking area of the parameter space whose measure decreases to zero. This is a mild condition which holds often in practice. Under this condition, Proposition 2 justifies that the estimate $\hat{\boldsymbol{\theta}}_n^{MCE}$ converges to the true value $\boldsymbol{\theta}^o$. Thus, we have established Property ($\mathcal{P}$-3) of depth-CDs stated in the Section 1.

4. Construct Depth-CDs From Nonparametric Bootstrap

This section provides a concrete approach of using nonparametric bootstrap to construct a depth-CD and derive inferences for the target parameter vector in an individual study. Given the

sample $\{X_1, X_2, \ldots, X_n\}$, assume that $\hat{\boldsymbol{\theta}}_n$ is an estimate of the target parameter $\boldsymbol{\theta}$, where

$$\hat{\boldsymbol{\theta}}_n \equiv \hat{\boldsymbol{\theta}}_n(X_1, X_2, \ldots, X_n). \tag{12}$$

Let $\hat{\boldsymbol{\theta}}^*_n = \hat{\boldsymbol{\theta}}^*_n(X^*_1, X^*_2, \ldots, X^*_n)$ be a bootstrap estimate of $\boldsymbol{\theta}$, where $\{X^*_1, X^*_2, \ldots, X^*_n\}$ is a bootstrap sample drawn independently from $\{X_1, X_2, \ldots, X_n\}$ with replacement. Let B_n and B^*_n denote the sampling distribution of $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\theta}}^*_n$, respectively. Theorem 1 below shows that the bootstrap distribution B^*_n is a depth-CD for $\boldsymbol{\theta}$ asymptotically under the following regularity conditions:

(C1) Let L_n be the distribution of $a_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ for some positive sequence $a_n \to \infty$, as $n \to \infty$. Assume that L_n converges D-regularly to a distribution L, namely, (i) L_n converges weakly to L as $n \to \infty$; and (ii) $\lim_{n\to\infty} \sup_{x\in\mathbb{R}^p} |D(L_n, x) - D(L, x)| = 0$.
(C2) Let L^*_n be the distribution of $a_n(\hat{\boldsymbol{\theta}}^*_n - \hat{\boldsymbol{\theta}}_n)$. Assume that L^*_n converges D regularly to the distribution L almost surely.
(C3) The distribution L is continuous and symmetric around 0.
(C4) The distribution of $D(L, l)$ is continuous, where the random variable $l \sim L$.

Theorem 1. Under the regularity conditions (C1)–(C4), the distribution of $\hat{\boldsymbol{\theta}}^*_n$, conditional on the sample $\{X_1, X_2, \ldots, X_n\}$, is a depth-CD for the parameter $\boldsymbol{\theta}$ asymptotically as $n \to \infty$.

Theorem 1 shows that the bootstrap distribution B^*_n is a depth-CD, and hence justifies the validity of using B^*_n to make inferences about the parameter vector $\boldsymbol{\theta}$. For example, the deepest point of the distribution B^*_n can be used as a point estimate of $\boldsymbol{\theta}$, and the central region $\mathcal{R}_{1-\alpha}(B^*_n)$, as defined in Equation (10), as a $(1-\alpha)$ confidence region for $\boldsymbol{\theta}$. Moreover, for testing the hypothesis $\boldsymbol{\theta} = \boldsymbol{\vartheta}$ versus $\boldsymbol{\theta} \neq \boldsymbol{\vartheta}$, the value of the centrality function at $\boldsymbol{\vartheta}$, that is, $C_{B^*_n}(\boldsymbol{\vartheta})$, can be used as a p-value.

This p-value approach for hypothesis testing is fundamentally different from traditional approaches, as mentioned in Section 3.3. First, the reference distribution here is depth-CD B^*_n, which is fully determined by the sample. Once the sample is given, it does not vary, unlike the traditional approaches. Second, depth-CD B^*_n is a single reference distribution and provides a p-value for testing each parameter value in the entire parameter space $\boldsymbol{\Theta}$. Third, in the derivation of the p-value, $C_{B^*_n}(\boldsymbol{\vartheta})$ does not rely on any test statistic. Essentially, $C_{B^*_n}(\boldsymbol{\vartheta})$ now simultaneously serves as a test statistic and as a p-value. This actually compresses into a single step, namely calculating the centrality of $\boldsymbol{\vartheta}$ w.r.t. depth-CD B^*_n, the usual three steps in traditional testing procedures, namely identifying a test statistic, establishing its sampling distribution, and then calculating the p-value. Fourth, the reference distribution depth-CD B^*_n is obtained by resampling directly from the empirical distribution, rather than from the null distribution that is usually restricted by parametric assumptions. This also explains why our CD inference here can be obtained without distributional assumptions of the sample.

The idea of connecting bootstrap to data depth for multiparameter inference is not new. For example, it has been used in Liu and Singh (1997) for hypothesis testing and in Yeh and Singh (1997) for deriving confidence regions. These two inference methods can be viewed as special cases in our depth-CD inference framework, since Theorem 1 implies that the bootstrap distribution is a depth-CD.

Theorem 2 is a direct consequence of Lemma 1 and Proposition 1, and it provides a procedure for constructing depth-CDs from pivot statistics.

Theorem 2. Assume that $A_n(\boldsymbol{X}_n)$ is a nonsingular matrix such that $A_n(\boldsymbol{X}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ follows a distribution Q_n that is free of all unknown parameters. Also assume that $\boldsymbol{\eta}_n$ is a random vector, independent of the sample $\boldsymbol{X}_n$, following the distribution Q_n. Then, conditional on $\boldsymbol{X}_n$, the distribution function of $(\hat{\boldsymbol{\theta}}_n - A_n(\boldsymbol{X}_n)^{-1}\boldsymbol{\eta}_n)$ is a depth-CD for $\boldsymbol{\theta}$, under the following conditions
(i) The depth D is affine-invariant; and
(ii) The depth contours $\{\boldsymbol{\eta}_n : D_{Q_n}(\boldsymbol{\eta}_n) = t\}$ all have probability zero w.r.t. Q_n.

Theorem 2 shows that when the statistic $A_n(\boldsymbol{X}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ is a pivot, a depth-CD can be easily derived using the inverse probability function. Returning to Example 3 where we make inference for the mean parameter of a bivariate normal distribution. In this case, we know that $\Sigma^{-1/2}(\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta})$ is a pivot following a bivariate standard normal distribution and that the three depths mentioned in Section 3.2 are affine-invariant. Thus, by Theorem 2, the bivariate normal distribution $BN(\bar{\boldsymbol{Y}}_n, \Sigma)$ is a depth-CD for $\boldsymbol{\theta}$. When Σ is not known, $BN(\bar{\boldsymbol{Y}}_n, \hat{\Sigma})$ is a depth-CD for $\boldsymbol{\theta}$ asymptotically. In this example, the distribution of the pivot $\Sigma^{-1/2}(\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta})$, namely Q_n in Theorem 2, is structured under certain distributional assumptions, that is, Q_n is bivariate standard normal. But generally, as long as $A_n(\boldsymbol{X}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ can be structured to have (approximately) a parameter-free distribution, Theorem 2 can be applied to construct depth-CDs and draw all forms of inference accordingly, as seen in Section 3.

With the distribution Q_n as a prerequisite, Theorem 2 may be perceived as applicable only for deriving depth-CDs in the setting of parametric inference, but its precise formulation can actually shed light on the bootstrap approach in Theorem 1 and other general nonparametric approaches for deriving depth-CDs. Here is how Theorem 2 explains intuitively why the nonparametric bootstrap distribution is indeed a depth-CD. To avoid making assumptions about the distribution Q_n of the statistic $A_n(\boldsymbol{X}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$, a natural choice is to use the bootstrap distribution of $A_n(\boldsymbol{X}_n)(\hat{\boldsymbol{\theta}}^*_n - \hat{\boldsymbol{\theta}}_n)$ to approximate Q_n, that is, set $\boldsymbol{\eta}_n = A_n(\boldsymbol{X}_n)(\hat{\boldsymbol{\theta}}^*_n - \hat{\boldsymbol{\theta}}_n)$ in Theorem 2. Assuming that this approximation is appropriate (e.g., under conditions (C1)-(C4)), Theorem 2 shows that the distribution of $(2\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}^*_n)$ $(= \hat{\boldsymbol{\theta}}_n - A_n(\boldsymbol{X}_n)^{-1}A_n(\boldsymbol{X}_n)(\hat{\boldsymbol{\theta}}^*_n - \hat{\boldsymbol{\theta}}_n))$ is a depth-CD for $\boldsymbol{\theta}$. Note that $(2\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}^*_n)$ and $\hat{\boldsymbol{\theta}}^*_n$ have an identical distribution (provided that the distribution of $(\hat{\boldsymbol{\theta}}^*_n - \hat{\boldsymbol{\theta}}_n)$ is symmetric, see condition (C3)), since when centering around $\hat{\boldsymbol{\theta}}_n$, the reflection image of $(2\hat{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}^*_n)$ is exactly $\hat{\boldsymbol{\theta}}^*_n$. Hence, the bootstrap distribution of $\hat{\boldsymbol{\theta}}^*_n$ is also a depth-CD.

5. Fusion Learning Using Depth-CVs

5.1. Combining Depth-CVs

We have shown that the very form of the depth-CD being an all-encompassing distributional function estimate, rather than a

mere point or interval estimate, is the key feature that leads to the omnibus form of all inferences of a parameter. This feature will also be shown to underlie the great flexibility that makes depth-CDs particularly suited for combining inferences from different and even heterogeneous studies.

For each individual study, we can obtain a depth-CD $H_{k,n_k}(\cdot) \equiv H_{k,n_k}(\boldsymbol{X}_k;\cdot)$ and its corresponding depth-CV $C_{H_{k,n_k}}(\cdot)$ (see Definition 9) for the parameter $\boldsymbol{\theta}$, $k = 1, 2, \ldots, K$, from K independent studies. Here we propose a general formula in Equation (13) for synthesizing those K individual inference results to draw an overall and efficient inference for the parameter $\boldsymbol{\theta}$,

$$C_{(c)}(\boldsymbol{\theta}) = G_c\left(g_c(C_{H_{1,n_1}}(\boldsymbol{\theta}), C_{H_{2,n_2}}(\boldsymbol{\theta}), \ldots, C_{H_{K,n_K}}(\boldsymbol{\theta}))\right). \quad (13)$$

Here, $g_c(u_1, u_2, \ldots, u_K)$ is a continuous mapping from $[0,1]^K$ to $\mathbb{R}$ which is increasing in each coordinate, and $G_c(t) \equiv P\{g_c(U_1, U_2, \ldots, U_K) \le t\}$ where U_k's are iid random variables following U[0,1] distribution. A special yet important case of Equation (13) to which we will return often later is

$$C_{(c)}(\boldsymbol{\theta}) = F_{(c)}\Big\{w_1\varphi(C_{H_{1,n_1}}(\boldsymbol{\theta})) + w_2\varphi(C_{H_{2,n_2}}(\boldsymbol{\theta})) + \cdots w_K\varphi(C_{H_{K,n_K}}(\boldsymbol{\theta}))\Big\}, \quad (14)$$

with $g_c(u_1, u_2, \ldots, u_K) = w_1\varphi(u_1) + w_2\varphi(u_2) + \cdots w_K\varphi(u_K)$, where $\varphi(\cdot)$ is a monotonic increasing "mapping function" and $w_k > 0$ is the weight assigned to the k-th study.

Fusion formulas similar to Equations (13) and (14) have been used in Singh, Xie, and Strawderman (2005), Xie, Singh, and Strawderman (2011), and Liu, Liu, and Xie (2014) to combine CDs for a scalar parameter. However, these do not apply directly to combining depth-CDs for multiparameter inference. If the combining formula Equation (13) were applied directly, the resulting function $G_c(g_c(H_{1,n_1}(\boldsymbol{\theta}), H_{2,n_2}(\boldsymbol{\theta}), \ldots, H_{K,n_K}(\boldsymbol{\theta})))$ would not yield any valid statistical inference. To mitigate this shortcoming, our proposal in Equations (13) and (14) combines the depth-CVs through their corresponding centrality functions $C_{H_{k,n_k}}(\boldsymbol{\theta})$'s to obtain $C_{(c)}(\boldsymbol{\theta})$ (cf. Equation (13)) and the desired overall inference.

Theorem 3. Given the individual depth-CDs $H_{k,n_k}(\boldsymbol{\theta})$, $k = 1, \ldots, K$, the following forms of inference for the common parameter $\boldsymbol{\theta}$ derived from the combined depth CV function $C_{(c)}(\boldsymbol{\theta})$ in Equation (13) are valid.

(a) (Hypothesis testing) For testing the null hypothesis

$$\Omega_0 : \boldsymbol{\theta} = \boldsymbol{\vartheta} \quad \text{versus} \quad \Omega_1 : \boldsymbol{\theta} \neq \boldsymbol{\vartheta},$$

$C_{(c)}(\boldsymbol{\vartheta})$ is a limiting p-value, as discussed in Section 3.3, provided that $C_{H_{k,n_k}}(\boldsymbol{\vartheta}) \to 0$ in probability for all $\boldsymbol{\vartheta} \neq \boldsymbol{\theta}^o$. Here $\boldsymbol{\theta}^o$ is the true parameter value.

(b) (Confidence region) A $(1-\alpha)$ confidence region for $\boldsymbol{\theta}$ is

$$\mathcal{R}^{(c)}_{1-\alpha}(H_{1,n_1}, H_{2,n_2} \ldots, H_{K,n_K}) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : C_{(c)}(\boldsymbol{\theta}) \ge \alpha\}.$$

(c) (Point estimation) Assume that $C_{(c)}(\boldsymbol{\theta})$ achieves its maximum at $\hat{\boldsymbol{\theta}}_{(c)}$, that is,

$$\hat{\boldsymbol{\theta}}_{(c)} = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} C_{(c)}(\boldsymbol{\theta}). \quad (15)$$

Then, $C_{(c)}(\boldsymbol{\theta})$ is a consistent estimator for $\boldsymbol{\theta}^o$. Specifically, as $n_1, n_2, \ldots, n_K \to \infty$, $\hat{\boldsymbol{\theta}}_{(c)} \to \boldsymbol{\theta}^o$ in probability, provided that $C_{H_{k,n_k}}(\cdot)$ is continuous and

$$\Delta_{k,n_k}(\epsilon) = \max_{\boldsymbol{\vartheta}_i,\boldsymbol{\vartheta}_j\in\{\boldsymbol{\vartheta}:C_{H_{k,n_k}}(\boldsymbol{X}_{k,n_k},\cdot)(\boldsymbol{\vartheta})=\epsilon\}} ||\boldsymbol{\vartheta}_i - \boldsymbol{\vartheta}_j|| \to 0$$

in probability, for $k = 1, 2, \ldots, K$.

Theorem 3 justifies that the overall inferences based on the combined centrality function $C_{(c)}(\boldsymbol{\theta})$ can be made in ways similar to those based on centrality functions from individual studies. For example, Theorem 3(a) shows that, similar to each individual centrality function, the combined function $C_{(c)}(\boldsymbol{\theta})$ is a single invariant (under the given samples) function defined on the parameter space and it provides infinitely many p-values for testing all $\boldsymbol{\theta}$ values in the entire parameter space. It expresses the relative ranking or level of plausibility of each $\boldsymbol{\theta}$ value w.r.t. the totality of evidence collected from all studies. This expression of relative ranking of plausibility adapts readily to the common interpretation of a p-value. Theorem 3(b) describes a $(1-\alpha)$ confidence region for $\boldsymbol{\theta}$ as the collection of parameter values whose $C_{(c)}(\boldsymbol{\theta})$ is no less than α. Theorem 3(c) shows that the maximizer of the combined $C_{(c)}(\cdot)$ is a valid point estimator.

Our fusion learning does not rely on parametric assumptions, if Equations (13) or (14) is applied to depth-CDs from nonparametric approaches, such as bootstrap. This fusion approach is broadly applicable. It is valid as long as the input functions $H_{k,n_k}(\boldsymbol{\theta})$'s are depth-CDs (or asymptotically).

5.1.1. Higher-Order Accuracy of $C_{(c)}(\cdot)$ and its Inference Results

The depth-CDs obtained by bootstrap are not *exact*, in the sense they only satisfy asymptotically the requirement (ii) in Definition 1 or (ii)′ in Proposition 1. To see how this approximation accuracy affects the accuracy of the inference results, we consider the example of a univariate common mean problem, where a CD obtained by the regular bootstrap in Section 4 is $H_{k,n_k}(\theta) = P^*\{\bar{X}^*_k \le \theta\}$. Such a CD can yield confidence regions whose coverage probability approximates the nominal value. A better accuracy can be achieved by using the bootstrap t (Efron and Tibshirani 1994). This bootstrap method generates a second-order accurate CD $H^{(t)}_{k,n_k}(\theta) = P^*\left\{(\bar{X}^*_k - \bar{X}_k)/S^* \ge (\bar{X}_k - \theta)/S\right\}$, where S is an estimate of the standard deviation. It would be interesting to know whether or not the improved accuracy in the input CDs is carried over to the combined outcome. It is worth noting that our fusion approach in Equation (14) generally does preserve the order of accuracy of the individual depth-CDs, even if they are not *exact*. To state the result, we first define the order of accuracy for a depth-CD.

A depth-CD function $H_n(\cdot) \equiv H_n(\boldsymbol{X}_n, \cdot)$ on $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is said to be jth-order accurate, if the random variable $C_{H_n}(\boldsymbol{\theta}^o) \equiv C_{H_n(X_n,\cdot)}(\boldsymbol{\theta}^o)$, where $\boldsymbol{\theta}^o$ is the true value of $\boldsymbol{\theta}$, converges in distribution to the uniform distribution on (0,1) at the order of $n^{-j/2}$, that is, $P\left\{C_{H_n}(\boldsymbol{\theta}^o) \le a\right\} - a = O(n^{-j/2})$ for any $a \in (0,1)$. If a depth-CD function $H_n(\cdot)$ is jth order accurate, the coverage probability of $\mathcal{R}_{1-\alpha}(H_n)$ in Equation (10) converges to its nominal level at the rate of $O(n^{-j/2})$.

Theorem 4. (Accuracy of $C_{(c)}(\cdot)$). For the kth study ($k = 1, 2, \ldots, K$), assume that n_k/n converges to a constant a_k and also that its depth-CD function $H_{k,n_k}(\boldsymbol{\theta})$ is jth-order accurate uniformly, in the sense that $P\left\{C_{H_{k,n_k}}(\boldsymbol{\theta}^o) \le a\right\} - a = O(n^{-j/2})$ uniformly for all $a \in (0,1)$ as $n \to \infty$. Then the combined function $C_{(c)}(\boldsymbol{\theta})$ in its general form Equation (13) is also jth-order accurate.

Our numerical studies in Section 6 show that, even in small-sample cases, the overall inferences are quite accurate, when the input CD functions $H^{(t)}_{k,n_k}(\boldsymbol{\theta})$'s are obtained by the bootstrap t.

5.1.2. Bahadur Efficiency of $C_{(c)}(\cdot)$

The fusion formula in Equation (13) provides a general class of fusion approaches for synthesizing nonparametric or parametric inferences. We show here that among this general class, a specific form of Equation (14) with $w_k = 1$ for all k and $\varphi(t) = \log(t)$ yields the most efficient combination in terms of achieving Bahadur efficiency. Following the ideas in Littell and Folks (1973) and Singh, Xie, and Strawderman (2005), we define the concept of *Bahadur slope* for a depth-CV.

Definition 2. A nonnegative function $S_{\boldsymbol{\lambda}}(b) \equiv S_{\boldsymbol{\lambda}}(b; \boldsymbol{\theta}^o)$ is said to be the Bahadur slope for the depth-CV function $C_{H_n}(\cdot)$ along the direction $\boldsymbol{\lambda}$, where $\boldsymbol{\lambda} \in \mathbb{R}^p$ and $||\boldsymbol{\lambda}|| = 1$, if $S_{\boldsymbol{\lambda}}(b) \equiv \lim_{n\to\infty} -\log\{C_{H_n}(\boldsymbol{\theta}^o + b\boldsymbol{\lambda})\}/n$ almost surely for any non-zero $b \in \mathbb{R}$.

The Bahadur slope $S_{\boldsymbol{\lambda}}(b)$ defined above reflects the rate, in an exponential scale, at which $C_{H_n}(\boldsymbol{\theta}^o + b\boldsymbol{\lambda})$ decays toward zero as the sample size increases. The larger the slope, the more efficient the depth-CV in Bahadur's sense. In the multiparameter case where the depth-CD $H_n(\boldsymbol{\theta})$ is a multivariate distribution, we need Bahadur slope functions $S_{\boldsymbol{\lambda}}(b)$ along each direction $\boldsymbol{\lambda}$ to characterize how fast the tails of the distribution decay to zero.

The Bahadur slope provides a means assessing the efficiency of the proposed fusion method Equation (13). Specifically, the theorem below establishes an upper bound of the Bahadur slope (i.e., the fastest possible rate of tail decay) for the combined function $C_{(c)}(\boldsymbol{\theta})$. It also suggests a specific combination formula for achieving exactly this bound.

Theorem 5. Under $\boldsymbol{\theta} = \boldsymbol{\theta}^o$ and $n_k = \{a_k + o(1)\}n$, as $n \to \infty$, the following inequality holds for any fused function $C_{(c)}(\boldsymbol{\theta})$ as defined in the general fusion formula Equation (13)

$$\limsup_{n\to\infty} -\log\{C_{(c)}(\boldsymbol{\theta}^o + b\boldsymbol{\lambda})\}/n \le \sum_{k=1}^{K} a_k S_{k,\boldsymbol{\lambda}}(b). \tag{16}$$

Furthermore, let $C^{log}_{(c)}$ denote the fused function in its specific form Equation (14) when $w_k = 1$ for all k and $\varphi(t) = \log(t)$. Then,

$$\lim_{n\to\infty} -\log\{C^{log}_{(c)}(\boldsymbol{\theta}^o + b\boldsymbol{\lambda})\}/n = \sum_{k=1}^{K} a_k S_{k,\boldsymbol{\lambda}}(b). \tag{17}$$

Theorem 5 states that the Bahadur slope of any combined function $C_{(c)}(\cdot)$ derived from Equation (13) has an upper bound, and that this upper bound can be achieved by taking $w_k = 1$ for all k and $\varphi(t) = \log(t)$ in Equation (14). In this case, the explicit formula for combining depth-CVs is

$$C^{\log}_{(c)}(\boldsymbol{\Theta}) = P\left\{\chi^2_{2K} \ge -2\sum_{k=1}^{K} \log(C_{H_{k,n_k}}(\boldsymbol{\theta}))\right\}. \tag{18}$$

This formula turns out to be the same as Fisher's method used for combining p-values. The optimality of this particular choice does not rely on the direction $\boldsymbol{\lambda}$. Thus, an interesting implication is that if we use Equation (18) to combine depth-CVs, the highest Bahadur slope (or the fastest rate of tail decay) will be achieved along *every* direction (i.e., the line spanned by $\boldsymbol{\theta}^o + b\boldsymbol{\lambda}$ as b varies). The optimality established in Theorem 5 is a global, rather than merely directional, property of Equation (18).

5.2. Fusion of Heterogeneous Studies

Our fusion framework is general and can cover complex and irregular settings containing heterogeneous studies. Study heterogeneity arises often in practice, due to different study designs, populations or outcomes, as seen in the applications in (Chen, Chatterjee, and Carroll 2013; Yang et al. 2014; Liu, Liu, and Xie 2015; Chatterjee et al. 2016; Gao and Carroll 2017). In the presence of heterogeneous studies, the parameter of interest may not be estimable in some studies. These studies are often excluded from conventional analyses, which can result in a nonnegligible loss of information. Our fusion method Equation (13) can be extended to incorporate heterogenous studies in the analysis. The theoretical results established in previous sections remain valid and applicable as well.

To accommodate heterogeneous studies, we give up the assumption that $\boldsymbol{\theta}_k$ is estimable in each study. Instead, we assume only that a certain mapping of $\boldsymbol{\theta}_k$, denoted by $\tilde{\boldsymbol{\theta}}_k (= \boldsymbol{f}_k(\boldsymbol{\theta}_k))$ as in Equation (3), is estimable and its corresponding depth-CD $H_{k,n_k}(\tilde{\boldsymbol{\theta}}_k)$ for $\tilde{\boldsymbol{\theta}}_k$ can be derived, say, using bootstrap. With a minor modification, the general fusion formula Equation (13) is still applicable to combining depth-CDs from different $\tilde{\boldsymbol{\theta}}_k$'s for making the overall inference about the common parameter of interest $\boldsymbol{\theta}$. More specifically,

$$C^{\text{Het}}_{(c)}(\boldsymbol{\theta}) = G_c\left(g_c(C_{H_{1,n_1}}(\tilde{\boldsymbol{\theta}}_1), C_{H_{2,n_2}}(\tilde{\boldsymbol{\theta}}_2), \ldots, C_{H_{K,n_K}}(\tilde{\boldsymbol{\theta}}_K))\right). \tag{19}$$

Similar to Equation (14), a special case is

$$C^{\text{Het}}_{(c)}(\boldsymbol{\theta}) = F_{(c)}\Big\{w_1\varphi(C_{H_{1,n_1}}(\tilde{\boldsymbol{\theta}}_1)) + w_2\varphi(C_{H_{2,n_2}}(\tilde{\boldsymbol{\theta}}_2)) + \cdots w_K\varphi(C_{H_{K,n_K}}(\tilde{\boldsymbol{\theta}}_K))\Big\}. \tag{20}$$

Theorem 6 shows how to use $C^{\text{Het}}_{(c)}(\boldsymbol{\theta})$ in Equations (19) or (20) to make valid combined inference about $\boldsymbol{\theta}$. Here, we require that $\boldsymbol{\theta}$ be identifiable in the combined function $C^{\text{Het}}_{(c)}(\boldsymbol{\theta})$. Following Rothenberg (1971) and Little, Heidenreich, and Li (2010), we say that $\boldsymbol{\theta}$ is (locally) *identifiable* if for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there is no $\boldsymbol{\vartheta} \ne \boldsymbol{\theta}$ (in a neighborhood of $\boldsymbol{\theta}$) such that $C^{\text{Het}}_{(c)}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K; \boldsymbol{\theta}) = C^{\text{Het}}_{(c)}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K; \boldsymbol{\vartheta})$ almost surely.

Theorem 6. Consider the setup in Equation (3) and the given depth-CDs $H_{k,n_k}(\tilde{\boldsymbol{\theta}}_k)$ for $\tilde{\boldsymbol{\theta}}_k$, $k = 1, \ldots, K$. Assume that the

parameter $\boldsymbol{\theta}$ is identifiable in the combined function $C_{(c)}^{\text{Het}}(\boldsymbol{\theta})$ in its general form Equation (19). Then, the following inferences derived from $C_{(c)}^{\text{Het}}(\boldsymbol{\theta})$ are valid.

(a) (Hypothesis testing) For testing the null hypothesis

$$\Omega_0 : \boldsymbol{\theta} = \boldsymbol{\vartheta} \quad \text{versus} \quad \Omega_1 : \boldsymbol{\theta} \neq \boldsymbol{\vartheta},$$

$C_{(c)}^{\text{Het}}(\boldsymbol{\vartheta})$ is a limiting p-value, in the sense as discussed in Section 3.3, provided that $C_{H_{k,n_k}}(\boldsymbol{\vartheta}_k) \to 0$ in probability for all $\boldsymbol{\vartheta}_k \neq \boldsymbol{\theta}_k^o$.

(b) (Confidence region) A $(1-\alpha)$ confidence region for $\boldsymbol{\theta}$ is

$$\mathcal{R}_{1-\alpha}^{(c)}(H_{1,n_1}, H_{2,n_2} \ldots, H_{K,n_K}) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : C_{(c)}^{\text{Het}}(\boldsymbol{\theta}) \geq \alpha\}.$$

(c) (Point estimation) Assume that $C_{(c)}^{\text{Het}}(\boldsymbol{\theta})$ achieves its maximum at $\hat{\boldsymbol{\theta}}_{(c)}$, that is,

$$\hat{\boldsymbol{\theta}}_{(c)} = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} C_{(c)}^{\text{Het}}(\boldsymbol{\theta}).$$

Then, $C_{(c)}(\boldsymbol{\theta})$ is a consistent estimator for $\boldsymbol{\theta}^o$. Specifically, as $n_1, n_2, \ldots, n_K \to \infty$, $\hat{\boldsymbol{\theta}}_{(c)} \to \boldsymbol{\theta}^o$ in probability, provided that $C_{H_{k,n_k}}(\cdot)$ is continuous and

$$\Delta_{k,n_k}(\epsilon) = \max_{\boldsymbol{\vartheta}_i, \boldsymbol{\vartheta}_j \in \{\boldsymbol{\vartheta} : C_{H_{k,n_k}}(\boldsymbol{X}_{k,n_k,\cdot})(\boldsymbol{\vartheta}) = \epsilon\}} ||\boldsymbol{\vartheta}_i - \boldsymbol{\vartheta}_j|| \to 0$$

in probability, for $k = 1, 2, \ldots, K$.

Theorem 6 justifies the validity of using the modified combined depth-CV to draw overall inferences from heterogenous studies, which is the counterpart of Theorem 3 in the case of homogeneous studies. In fact, the counterparts of Theorems 4-5 can also be established to obtain the same theoretical results of high-order accuracy and Bahadur efficiency under heterogenous studies. In short, the combining in Equation (19) preserves the order of accuracy of each individual study, and it achieves Bahadur efficiency when $w_k = 1$ for all k and $\phi(t) = \log(t)$.

Compared to the meta-analysis of heterogeneous studies in Liu, Liu, and Xie (2015), our fusion method here is more general. Liu, Liu, and Xie (2015) requires normality of the distribution of summary statistics. Our fusion method does not require a parametric form of distributional assumptions. If each individual depth-CD is derived using the nonparametric bootstrap, then the inference drawn from the combined function is also nonparametric. When it is reasonable to make an assumption of the underlying distribution, we can derive depth-CDs using Theorem 2 and our fusion method is still applicable and yields valid inferences.

6. Simulation Studies

To demonstrate the theoretical advantages of our fusion method, we conduct simulation studies for the common mean problem and meta-analysis of correlation coefficients.

6.1. The Common Mean Problems

Making inference on the common mean parameter of multiple populations is referred to as the common mean problem. This problem has been investigated extensively, see, for example, Lin, Lee, and Wang 2007; Pal et al. 2007, and the references therein. Traditional approaches rely on the assumption that the sample of each study is drawn from a normal distribution. The normality assumption however is often unrealistic in practice, and it can be hardly justified when the sample size is small. To the best of our knowledge, there have not been any systematic investigation of the common mean problem in general and nonnormal situations.

Our framework of fusion learning readily applies to the common mean problem, in both normal and nonnormal settings. In this section, we examine its numerical performance, in comparison with that of several existing methods associated with the well-known Graybill-Deal estimator (Graybill and Deal 1959). The numerical results show that without the normality assumption, our fusion method has the following advantages: (i) it preserves inference accuracy in hypothesis testing/confidence regions; (ii) its point estimator has less bias and is more efficient; and (iii) it achieves a gain of efficiency in the presence of heterogeneous studies.

In the multiparameter setting, the Graybill-Deal estimator is

$$\hat{\boldsymbol{\mu}}_{\text{GD}} = \left\{\sum_{k=1}^{K} n_k \boldsymbol{S}_k^{-1}\right\}^{-1} \sum_{k=1}^{K} n_k \boldsymbol{S}_k^{-1} \bar{\boldsymbol{X}}_k,$$

where $\bar{\boldsymbol{X}}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} \boldsymbol{X}_{k,j}$ and $\boldsymbol{S}_k = \frac{1}{n_k-1}\sum_{j=1}^{n_k}(\boldsymbol{X}_{k,j} - \bar{\boldsymbol{X}}_k)(\boldsymbol{X}_{k,j} - \bar{\boldsymbol{X}}_k)'$. This estimator yields confidence intervals and p-values by considering the statistic (Lin, Lee, and Wang 2007)

$$\sum_{k=1}^{K} w_k T_k^2 = \sum_{k=1}^{K} w_k n_k (\bar{\boldsymbol{X}}_k - \boldsymbol{\mu}_0)' \boldsymbol{S}_k^{-1} (\bar{\boldsymbol{X}}_k - \boldsymbol{\mu}_0). \tag{21}$$

Assume that $\boldsymbol{X}_{k,j}$ follows a multivariate normal distribution, then T_k^2's are Hotelling's T^2 statistics and $\frac{n_k - p}{p(n_k-1)} T_k^2 \sim F_{p,n_k-p}$. Thus, the statistic in Equation (21) follows a weighted convolution of multiple F distributions. We evaluate (21) in the construction of confidence regions and hypothesis testing when $w_k \equiv 1$ (referred to as the GD method) and $w_k = \text{var}(T_k^2)^{-1} = \{2p(n_k-1)^2(n_k-2)\}/\{(n_k-p-2)^2(n_k-p-4)\}$ (referred to as the KJ method, Jordan and Krishnamoorthy 1995). If the normality assumption holds, both the GD and KJ methods are exact in the sense that the test (or confidence region) achieves the nominal Type I error (or coverage probability), since the exact distribution of Equation (21) is known. We also consider a method based on the central limit theorem (CLT). This method needs a weaker assumption, namely that $\bar{\boldsymbol{X}}_k$ only approximately follows a normal distribution. The inference relies on the statistic

$$\left\{\sum_{k=1}^{K} n_k \hat{\boldsymbol{\Sigma}}_k^{-1} (\bar{\boldsymbol{X}}_k - \boldsymbol{\mu}_0)\right\}' \left\{\sum_{k=1}^{K} n_k \hat{\boldsymbol{\Sigma}}_k^{-1}\right\}^{-1} \left\{\sum_{k=1}^{K} n_k \hat{\boldsymbol{\Sigma}}_k^{-1} (\bar{\boldsymbol{X}}_k - \boldsymbol{\mu}_0)\right\},$$

where $\hat{\boldsymbol{\Sigma}}_k = (n_k - 1)\boldsymbol{S}_k / n_k$. This statistic follows χ^2 distribution with p degrees of freedom.

To implement our CD fusion method Equation (14), we use half-space depth and Bahadur-efficient combination in Equation (18). The bootstrap-t is used, when applicable, with 2000 bootstrap replicates in each run. We compare our method with

the GD, JK and CLT methods under the following scenarios. Without loss of generality, we set $K = 2$ and consider bivariate distributions.

Scenario 1 (Normal distribution) Let $\boldsymbol{X} = (Z_1, Z_2)'$ follow a bivariate normal distribution with $\boldsymbol{\mu}_0 = E(\boldsymbol{X}) = (0,0)'$, $\sigma(Z_1) = 1$, $\sigma(Z_2) = 2$, and $Corr(Z_1, Z_2) = \rho$. In Study 1, $\boldsymbol{X}_{1,j}$ iid $\sim$ $\boldsymbol{X}$ with $\rho = 0.8$, and in Study 2: $\boldsymbol{X}_{2,j}$ iid $\sim$ $\boldsymbol{X}$ with $\rho = 0.3$.

Scenario 2 (χ^2 distribution) Let $\boldsymbol{X} = (Z_1^2, Z_2^2)'$ where $(Z_1, Z_2)'$ follows the same bivariate normal distribution as in Scenario 1. The true value $\boldsymbol{\mu}_0 = E(\boldsymbol{X}_{1,j}) = E(\boldsymbol{X}_{2,j}) = (1,4)'$.

Scenario 3 (Cauchy distribution) Let $\boldsymbol{X} = (Z_1, Z_2)'$ follow a bivariate Cauchy distribution where Z_1 and Z_2 are independent. The scale parameters $\sigma(Z_1) = 1$ and $\sigma(Z_2) = 2$ in Study 1, and $\sigma(Z_1) = 4$ and $\sigma(Z_2) = 2$ in Study 2. The location parameters $\boldsymbol{\mu}_0 = (0,0)'$ in both studies.

6.1.1. Inference Accuracy in Hypothesis Testing/Confidence Regions

To assess inference accuracy, we present the null distribution of p-values in Figures 4–6 (based on 10,000 simulation replications). The deviation of this distribution from the U(0,1) distribution depicts the difference between the actual and nominal Type I error rates in hypothesis testing, or equivalently, the difference between the actual and the nominal coverage probabilities of confidence regions. When the sample distribution is normal, Figure 4 shows that the null distribution of p-values aligns well with the U(0,1) distribution for all the methods considered, except that the CLT method is slightly off the target line. However, when the sample distribution is nonnormal, such as χ^2, Figure 5 shows a notable deviation for GD, JK, and CLT methods. More details on those deviations can be seen from the empirical values reported in Table 1 for a set of specific points. The numerical values in the table can also be viewed as the (nominal or actual) Type I error rates. Boldfaced are the values with a notable deviation from their nominal levels. For example, when the nominal probability (or the Type I error rate) is 0.05, the actual probability is 0.18, 0.18, and 0.25, respectively, for GD, JK, and CLT methods. Such a substantial deviation indicates a non-negligible loss of inference accuracy and raises serious concerns on using those methods for inference. Only our CD method yields a null distribution following very closely the target distribution. This example shows that our CD method, due to its nonparametric nature, is robust against the violation of the normality assumption. In Scenario 3, we sample from a bivariate Cauchy distribution, whose mean does not exist, and our inference is on the location parameter instead. Since the moments of Cauchy distributions do not exist, it is not surprising to see in Figure 6 that GD, JK and CLT methods all exhibit an appreciable loss of inference accuracy. Again, our method remains approximately accurate, when using the median in Equation (12) to construct depth-CDs. The advantage of CD method seen in Figure 6 is also confirmed numerically in Table 1, where the actual Type I error rates are quite close to the nominal levels. This example highlights the flexibility of our method in adapting easily to irregular situations where moments of the distribution do not exist.

6.1.2. Bias and Efficiency in Point Estimation

We compare our CD point estimator in Equation (15) and Graybill-Deal estimator $\hat{\boldsymbol{\mu}}_{\text{GD}}$ in estimating the common mean (or location) parameter $\boldsymbol{\mu} = (\mu_1, \mu_2)'$. The distribution of estimates (based on 1000 simulation replications) is presented

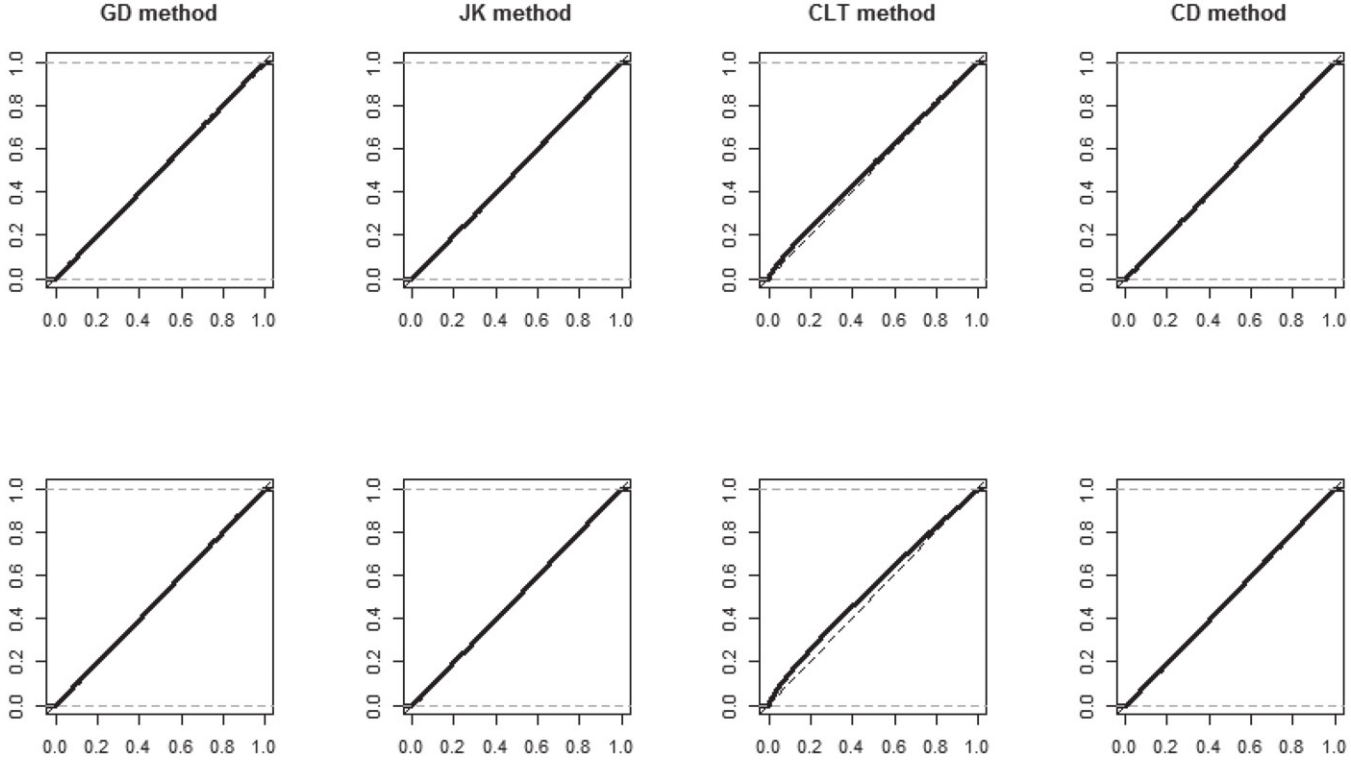

Figure 4. The null distributions of p-values derived from an individual study (upper row) and from the combined inference (lower row) for the common mean. The sample of size $n = 30$ in each individual study are drawn from a bivariate normal distribution.

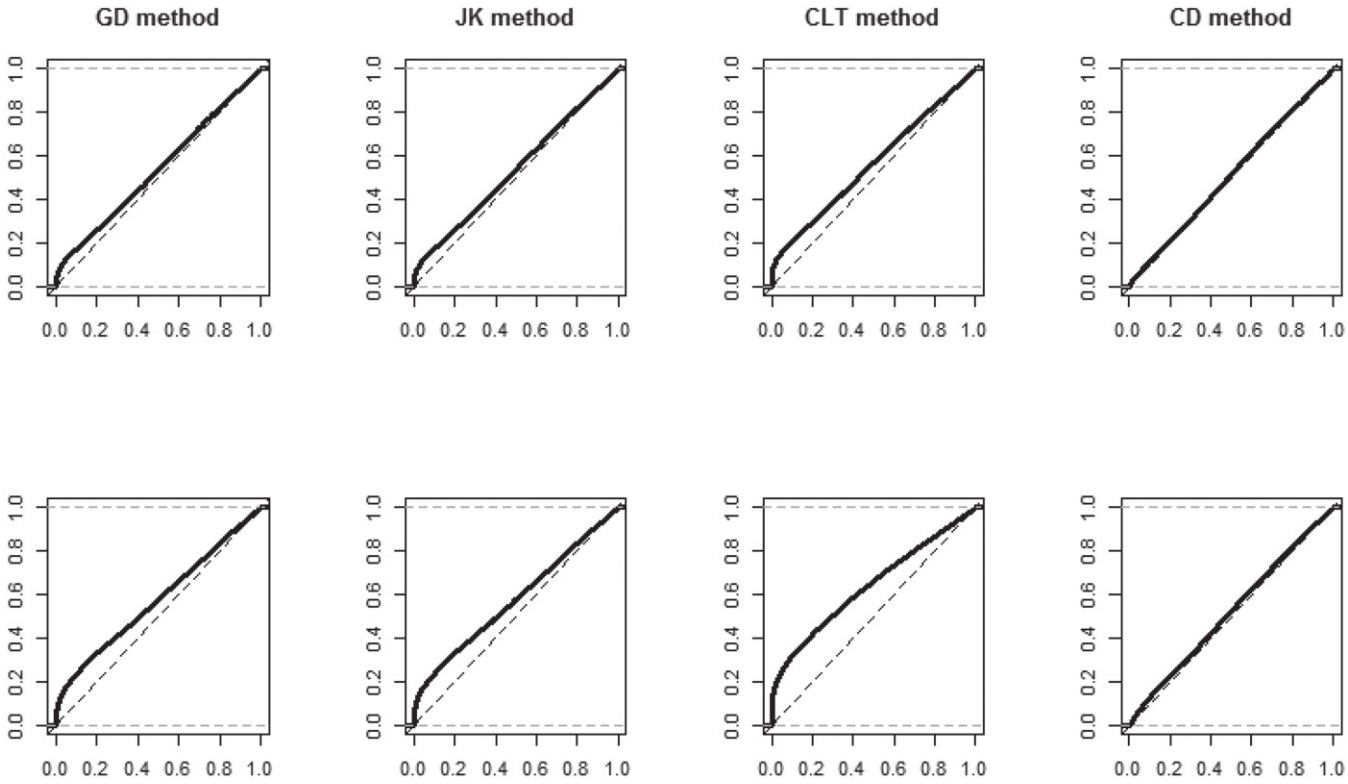

Figure 5. The null distributions of *p*-values derived from an individual study (upper row) and from the combined inference (lower row) for the common mean. The sample of size $n = 30$ in each individual study are drawn from a bivariate χ^2 distribution.

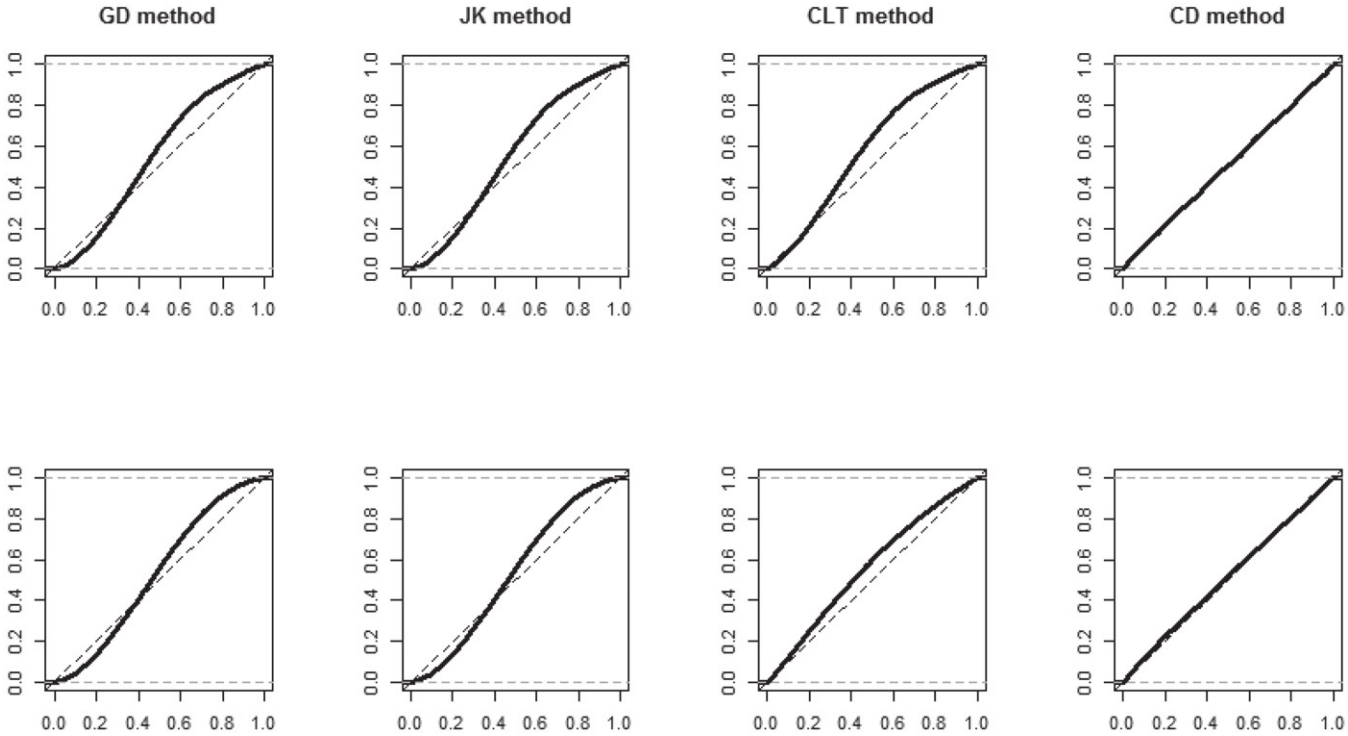

Figure 6. The null distributions of *p*-values derived from an individual study (the upper row) and the combined inference (the lower row) for the common mean. The sample of size $n = 30$ in each individual study are drawn from a bivariate Cauchy distribution.

as boxplots in Figure 7. When the sample distribution is normal, it shows in the first column that both estimators (i) are unbiased; and more interestingly, (ii) have comparable variabilities. More precisely, the standard errors of GD and CD estimates are 0.125 and 0.126 for μ_1, and 0.258 and 0.261 for μ_2, respectively. This observation implies that although the CD method is nonparametric, it sustains negligible efficiency loss compared to the GD method which does make use of the parametric assumption.

Table 1. Empirical distribution of the *p*-values at the null for the common mean problem.

Scenario 1. (Normal distribution)										
Nominal Probs	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
GD method	0.05	0.10	0.20	0.30	0.39	0.49	0.60	0.70	0.80	0.90
JK method	0.05	0.10	0.20	0.30	0.39	0.49	0.60	0.70	0.80	0.90
CLT method	**0.09**	**0.15**	0.26	0.36	0.46	0.55	0.65	0.74	0.83	0.91
CD method	0.04	0.10	0.19	0.29	0.39	0.50	0.60	0.70	0.80	0.90
Scenario 2. (χ^2 distribution)										
GD method	**0.18**	**0.24**	**0.33**	**0.41**	**0.49**	0.58	0.66	0.75	0.84	0.92
JK method	**0.18**	**0.24**	**0.33**	**0.41**	**0.49**	0.58	0.66	0.75	0.84	0.92
CLT method	**0.25**	**0.32**	**0.42**	**0.50**	**0.59**	**0.66**	**0.73**	**0.80**	0.87	0.94
CD method	0.07	0.12	0.23	0.32	0.42	0.52	0.62	0.72	0.82	0.92
Scenario 3. (Cauchy distribution)										
GD method	**0.01**	0.04	0.14	0.26	0.40	0.56	**0.70**	**0.82**	**0.91**	0.97
JK method	**0.01**	0.04	0.14	0.26	0.40	0.56	**0.70**	**0.82**	**0.91**	0.97
CLT method	0.05	0.12	0.25	0.37	**0.49**	**0.60**	**0.69**	0.78	0.86	0.93
CD method	0.06	0.12	0.22	0.32	0.42	0.52	0.61	0.71	0.80	0.90

Note: Boldfaced are those values with notable deviations from the nominal value.

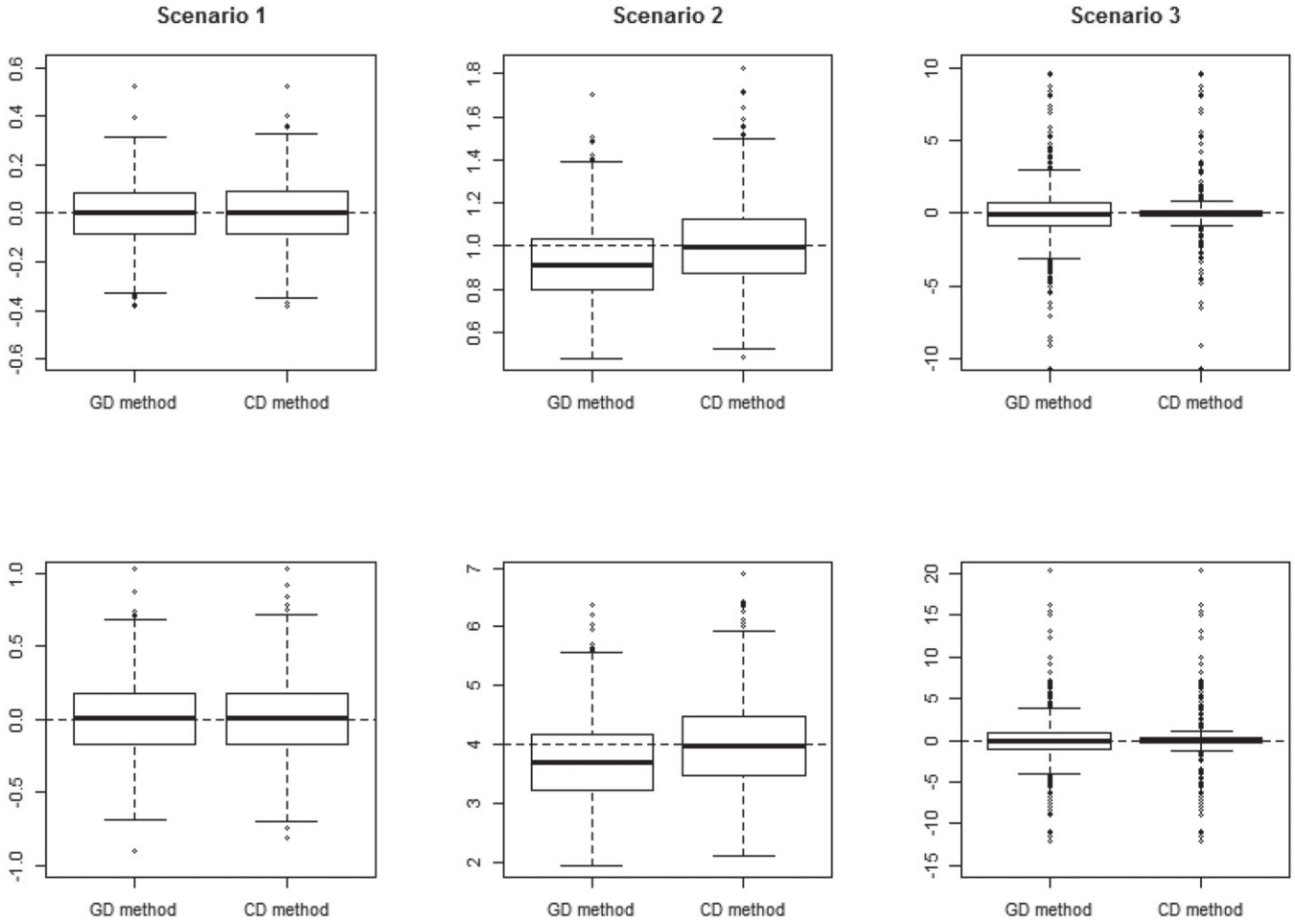

Figure 7. Boxplots of Graybill-Deal estimates and CD estimates for inferring the common mean (or location) vector $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ (the upper row for μ_1 and the lower row for μ_2). The true values $\boldsymbol{\mu}_0$ are drawn in the dashed lines. The sample of size $n = 30$ in each individual study are drawn from Scenario 1 (bivariate normal), Scenario 2 (χ^2), and Scenario 3 (Cauchy).

When the sample distribution is χ^2, the second column of Figure 7 shows that the variabilities of the two estimators are still comparable, but the GD estimator now shows a notable bias, whereas the CD estimator remains unbiased. When the sample distribution is Cauchy, the third column of Figure 7 shows that both estimators are unbiased, but the CD estimator has much smaller variability than the GD estimator, which indicates that the CD method is more efficient. To summarize, in the absence of normality, the CD estimator outperforms the GD estimator in terms of both bias and efficiency.

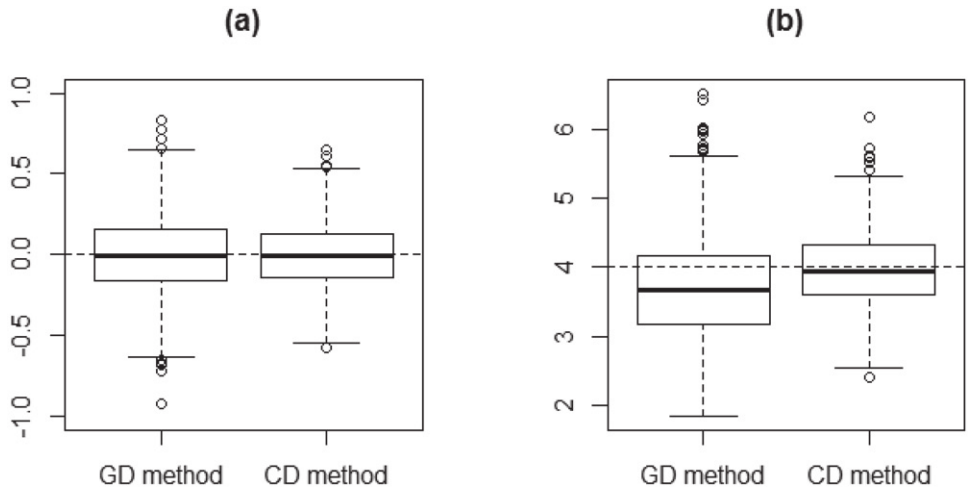

Figure 8. Boxplots of Graybill-Deal estimates and CD estimates for inferring μ_2 in the common mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)'$. The underlying distribution of $\boldsymbol{X}$ is bivariate normal in (a) and χ^2 in (b), with the true value of μ_2 being 0 and 4, respectively (drawn in the dashed lines). The sample size in each individual study is 30.

6.1.3. Gain of Efficiency in the Presence of Heterogeneous Studies

We consider a setting of heterogeneous studies by replicating the two studies in Scenario 1 (bivariate normal) and assuming that the two replicated studies are irregular, in that only the sum of the two components of the random vector $\boldsymbol{X}$ is observed. We are interested in combining inferences from all four studies. Here, neither of the two marginal means μ_1 and μ_2 is estimable in all studies, but the sum $\mu_1 + \mu_2$ is. The GD estimator $\hat{\boldsymbol{\mu}}_{\text{GD}}$ can combine only the two regular studies but discard the other two, whereas our CD estimator in Equation (20) can incorporate the two irregular studies as well. This same simulation is repeated under Scenario 2 (χ^2). To visualize the gain of efficiency in combining the inferences from all four studies, we present in Figure 8 the boxplots of the GD and CD estimates of μ_2 (based on 1000 simulation replications). The boxplots show that in both normal and nonnormal cases, our CD estimator, by combining all studies, is less variable and thus achieves a greater efficiency. Moreover, our CD estimator still remains almost unbiased in the nonnormal case. This phenomenon highlights again the flexibility of our fusion method in accommodating a broad class of study heterogeneity.

6.2. Meta-analysis of Correlation Coefficients

In social and behavioral sciences, correlation coefficients, being invariant to the measuring scale, are often used to represent the size of an effect. The meta-analysis of such an effect size has long been used as a tool to draw a more comprehensive conclusion on the bivariate association; see Schulze (2004) for an in-depth discussion. Classical meta-analysis inference methods for correlations, such as Fisher's z-transformation, assume that the samples of $(X_k, Y_k), k = 1, \ldots, K$, all follow bivariate normal distributions. When such an assumption is violated, inference outcomes could be invalid. In what follows, we show that our CD fusion method readily applies to meta-analysis of correlation coefficients, without requiring any parametric assumptions.

To illustrate the CD fusion method, we use the Pearson sample correlation r as an estimate of the correlation coefficient ρ in Equation (12), and apply regular bootstrap (with 2000 replicates) to construct a depth-CD in each study. To combine depth-CDs, we use half-space depth and Bahadur-efficient combination Equation (18). We compare our method with a naive method and the Hedges-Olkin (HO) method (Schulze 2004). The naive method merges the datasets as if all the data are from a single source. It then calculates the sample correlation r and applies Fisher's z-transformation $z = \frac{1}{2}\log(\frac{1+r}{1-r})$, where z follows approximately a normal distribution with mean 0 and variance $1/(n-3)$. The HO method obtains Fisher's z-transformed statistic z_k from each study, and combines them using $\bar{z} = \sum_{k=1}^{K}(n_k - 3)z_k / \sum_{k=1}^{K}(n_k - 3)$. The inference is based on that $\bar{z}\sqrt{\sum_{k=1}^{K}(n_k - 3)}$ follows approximately the $N(0,1)$ distribution. Figure 9 compares the three methods by examining the null distribution of p-values for testing the hypothesis $H_0 : \rho = 0$. When the samples of (X_k, Y_k) indeed follow a bivariate normal distribution, the upper row of Figure 9 shows that the distribution of each p-value approximates the U(0,1) distribution quite well. This observation indicates that all three methods lead to valid inference in normal cases. In the absence of normality, we let $X_k = Z_k$ and $Y_k = Z_k^2$, where $Z_k \sim N(0,1)$. The lower row of Figure 9 shows that the p-value distributions of the naive and HO methods deviate substantially from the $U(0,1)$ distribution. More specifically, the Type I error rates ($\alpha = 0.05$) are 0.38 and 0.37, respectively. The results indicate that these two methods may lead to invalid inference in nonnormal cases. The p-value distribution of CD method remains very close to the $U(0,1)$ distribution, which is indicative of its robustness to changing distribution assumptions.

7. Case Study: Analysis of Aircraft Landing Performance

Recall from the Introduction the motivating example from the FAA project on investigating whether or not aircraft landing operations generally comply with the FAA recommendation that the height of the aircraft at the crossing of runway threshold be around 15.85m and touchdown distance be around 432 m from runway threshold. This question can be addressed by

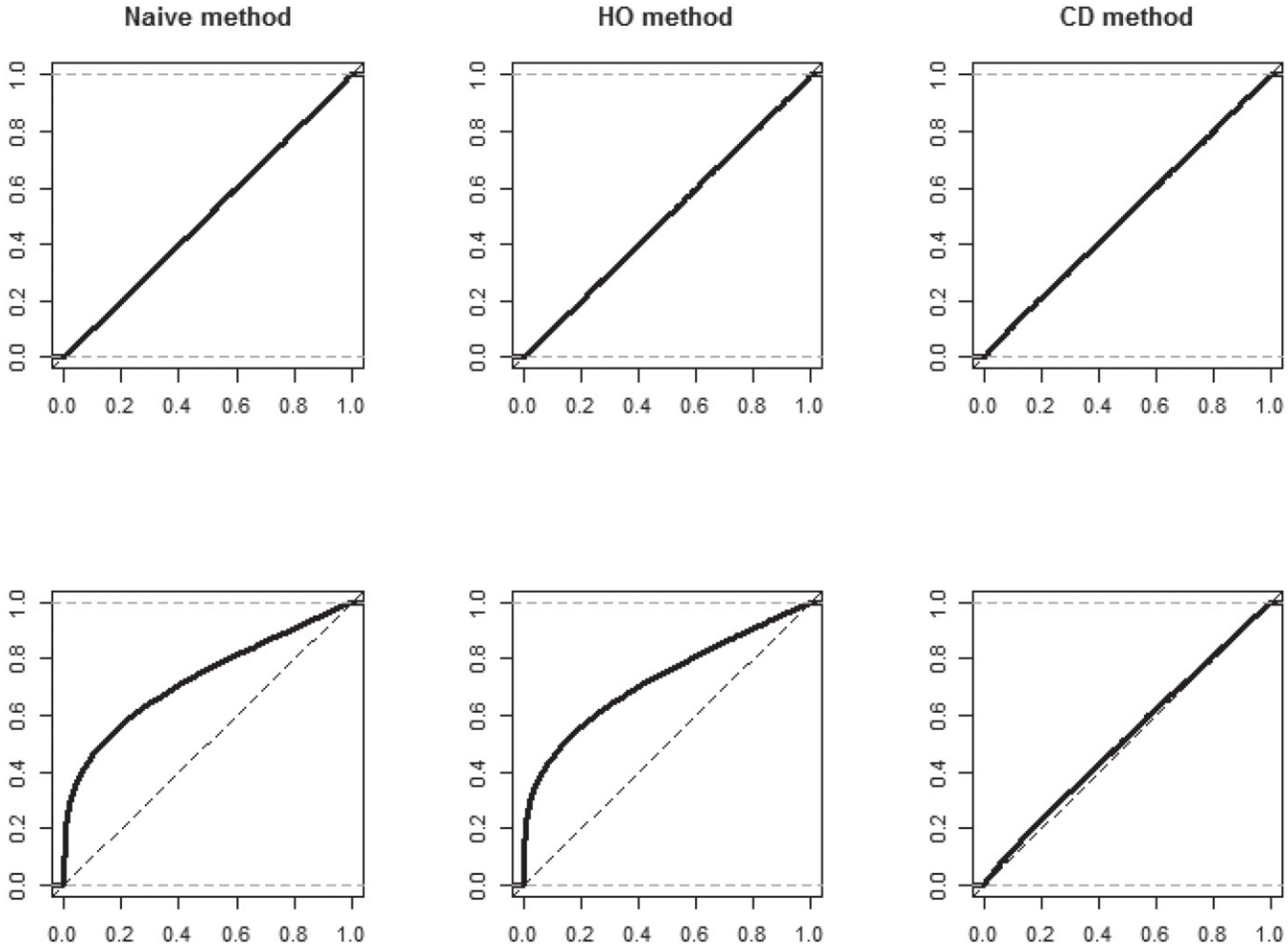

Figure 9. The null distributions of p-values for meta-analysis of correlation coefficients. The sample of size $n = 200$ in each individual study are drawn from a bivariate normal distribution (upper row) or a bivariate normal-χ^2 distribution (lower row).

testing

$$H_0 : \boldsymbol{\mu} = (15.85, 432)' \quad \text{versus} \quad H_1 : \text{otherwise}, \tag{22}$$

where $\boldsymbol{\mu}$ is the mean vector for the height at the runway threshold and touch down distance.

We are given landing records of two fleets of aircraft, 820 from Airbus and 1976 from Boeing. In view of the large samples, an intuitive approach would be just to apply Hotelling's T^2 test to the entire sample of 2796 landing records, pooling together both fleets. This yields a p-value of 0.942, which would suggest that there is no evidence supporting that the landing performances do not comply with the FAA recommendation. This intuitive approach for combining two studies however is flawed, since it implicitly assumes that the two studies follow the same distribution and thus fails to account for the difference underlying the two studies, shown clearly in Figure 1. After all, it is only natural to expect difference in performance from different aircraft manufactured by different makers or designs.

To accommodate such potential study heterogeneity, our fusion learning method can synthesize evidence from the two studies to provide a valid answer to the question raised. Specifically, this problem setting consists of two independent studies sharing a common bivariate mean parameter $\boldsymbol{\mu}$, that is, $\boldsymbol{\mu}_A = \boldsymbol{\mu}_B = \boldsymbol{\mu}$, where $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ are the means of the Airbus Study and the Boeing Study, respectively. We construct a depth-CD from each study to carry out separately the two tests $\boldsymbol{\mu}_A = \boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_B = \boldsymbol{\mu}_0$ with $\boldsymbol{\mu}_0 = (15.85, 432)'$, and then combine the two test results using Equation (14) to draw the overall inference on testing the hypothesis in Equation (22).

Specifically, we obtain a sample mean $\hat{\boldsymbol{\mu}}_A^*$ based on a bootstrap-t sample of Airbus Study, and replicate this 2000 times to obtain a depth-CD $H_A(\boldsymbol{\mu}_A)$, in this case, namely the empirical distribution of $\{\hat{\boldsymbol{\mu}}_{A,1}^*, \hat{\boldsymbol{\mu}}_{A,2}^*, \ldots, \hat{\boldsymbol{\mu}}_{A,2000}^*\}$. A depth-CD $H_B(\boldsymbol{\mu}_B)$ for the Boeing study can be obtained similarly. We then combine $H_A(\boldsymbol{\mu}_A)$ and $H_B(\boldsymbol{\mu}_B)$ using Equation (14) for testing the hypothesis in (22). Our fusion method yields a p-value of 0.008, indicating that the data provide strong evidence against the null hypothesis that the landings follow the FAA recommendation. This conclusion is drawn without assuming the sample follow any particular (say, normal) distribution.

The seemingly contradictory results between the intuitive method and our fusion method may be best explained visually by the plots of individual depth-CDs for Airbus (blue circles) and Boeing (black crosses) in Figure 10. The depth-CDs here are represented by the empirical distributions of their respective bootstrap estimates. The red triangle marks the null value $\boldsymbol{\mu}_0 = (15.85, 432)$, which is clearly far from the centers of the two depth-CDs (which are the point estimates of their two respective

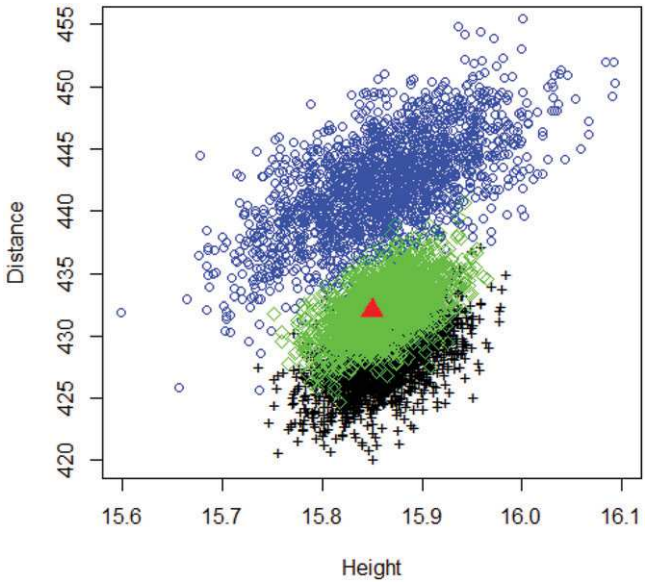

Figure 10. depth-CDs for each individual study, Airbus (blue circles) and Boeing (black crosses), and a depth-CD by aggregating data from the two studies as if they were from the same source (green diamonds). The depth-CDs here are represented by the empirical distributions of bootstrap estimates. The red triangle indicates the target value $\boldsymbol{\mu}_0$ in the null hypothesis $H_0 : \boldsymbol{\mu}_A = \boldsymbol{\mu}_B = \boldsymbol{\mu}_0 = (15.85, 432)'$.

means). The centrality values at $\boldsymbol{\mu}_0$ w.r.t. the two depth-CDs, or equivalently the two individual p-values, are 0.006 and 0.167. This finding implies low plausibilities for the assumption $\boldsymbol{\mu}_A = \boldsymbol{\mu}_0$ or $\boldsymbol{\mu}_B = \boldsymbol{\mu}_0$. Thus, a small p-value (0.008) from our fusion method leading to the rejection of H_0 should be expected.

We also plot in Figure 10 the depth-CD (green diamonds) obtained from the pooled data erroneously assuming the same distribution for the two studies. The red triangle $\boldsymbol{\mu}_0$ is near the center of this depth-CD, which suggests that $\boldsymbol{\mu}_0$ as a plausible target value, as also reflected in a large p-value 0.956. This example shows that ignoring the heterogeneity of data sources or blindly aggregating data may mask important signals and lead to invalid and misleading conclusions.

Finally, to demonstrate the flexibility of our fusion method in handling more challenging situations, we suppose that the recordings of the variable "height" from Airbus aircraft are not available. In this scenario, traditional methods can make inference about "height" only based on the landings of Boeing aircraft. For example, applying Hotelling's T^2 test to Boeing observations yields a p-value of 0.152. This again yields an incorrect conclusion. Unlike traditional methods, our fusion method can efficiently incorporate the information in the incomplete observations from Airbus Study, as shown in Section 5.2. Combining the indirect evidence from Airbus with the direct evidence from Boeing, our method yields a p-value of 0.016. This result suggests strong evidence against the null hypothesis, which is consistent with our conclusion drawn from the complete data from both studies. Our analysis here shows that indirect evidence may contain valuable information (e.g., possibly through the correlation between "height" and "distance" in this case) without which incorrect inference outcome may be reached.

8. Discussion

We have used the concept of depth-CD and depth-CV to develop a new framework for fusion learning for multivariate parameters, especially in nonparametric settings. This fusion learning framework imposes no model assumptions on the data or statistics in individual studies. It has been shown to be efficient, general and robust by both theoretical properties and numerical studies. In the nonnormal settings, it can reduce bias and improve efficiency in inference, as observed from simulation studies. In addition, our fusion framework can easily adapt to complex heterogeneous studies settings where existing methods fail. In particular, it can incorporate indirect evidence from heterogeneous studies for which the target parameter is not estimable, and achieve an additional gain of efficiency, as illustrated in both our simulation and case study. The phenomenon of incorporating indirect evidence to gain efficiency has also been observed, though only in the normal or asymptotic normal settings, in for example, Xie et al. (2013), Yang et al. (2014), Liu, Liu, and Xie (2015), Hoff (2019), Chen, Chatterjee, and Carroll (2013), Chatterjee et al. (2016), and Gao and Carroll (2017). The last three combined information from diverse studies through estimating equations, using large sample central limit theorem under parametric models.

Our proposed formula for nonparametric fusion learning is versatile as it permits flexible choices of fusion elements, namely, the depth function $D(\cdot)$, the mapping function $g(\cdot)$ and the weighting scheme w_k's. Unless there are concerns that not all studies are equally trustworthy, we may use equal weights ($w_k = 1$) with the mapping function $\varphi(\cdot) = \log(\cdot)$ and half-space or simplicial depth for general implementations. This set of choices is used in our numerical studies, showing our fusion formula to compete well with the classical approaches in the normal case and outperform in nonnormal cases, in gaining efficiency and reducing bias. This superb performance is rooted in the theoretical results established in Theorems 4 and 5, which ensure the fusion recipe to achieve high-order accuracy and the optimal efficiency in Bahadur's sense.

Fusion approaches derived from geometric depths such as half-space or simplicial depth have the desirable property of being nonparametric, and hence broader applicability. Although efficient exact algorithms for computing half-space and simplicial depths are available thus far for dimensions not higher than 3, as seen in Rousseeuw and Struyf (1998), the random approximation algorithm in Cuesta-Albertos and Nieto-Reyes (2008) is computationally efficient in any dimension. The refined sample geometric depth constructed in Einmahl, Li, and Liu (2015), by incorporating the extreme value theory, can help mitigate the inherent complications in calculating sample depth outside the data region (which by definition is zero without refinement) or in breaking the increasing number of ties in high-dimensional settings. As the computing technology continues its fast advances alongside the competing research effort in the computational geometry community to develop efficient depth computing algorithms, we believe that the concern over depth computational feasibility is likely to lessen gradually.

Similar to a Bayesian posterior distribution, a CD also uses a (sample-dependent) distribution function on the parameter space to estimate the target parameter and it contains the information for all possible inference. But, different from a Bayesian method, a CD method does not need to assume any prior distribution. In most cases when the sample size is sufficiently large, a Bayesian posterior can be shown to be a CD under

suitable regularity conditions (see, e.g., Xie and Singh 2013; Thornton and Xie 2020). In this case, the Bayesian posterior can be used as a CD to draw inferences and also as an input study in the CD fusion learning framework. Although the notion of CD is developed completely within the frequentist framework, it is shown to provide a common platform for unifying Bayesian, frequentist and fiducial approaches and also for making direct comparisons or combinations of inferences across these different paradigms. This also shows that Bayesian, frequentist and fiducial (BFF) inferences are indeed much more congruous than they have been perceived historically in the scientific community; as argued in recent research (Kass 2011; Reid and Cox 2015; Hannig et al. 2016; Shen, Liu, and Xie 2018; Thornton and Xie 2020)

A depth-CV, obtained through depth-CD (see Section 3.2), incorporates the idea of centrality measure from of data depth to form nested central regions expanding with growing probability mass. The capturing of the nested central regions with their associate probability coverages is key in making depth-CV such a versatile and effective multivariate inference tool. This formulation of central regions expanding with growing probability is akin to those referred to as "quality index" and "multivariate spacings" considered in Liu and Singh (1993) and Li and Liu (2008) in the context of assessing the distribution underlying the data for quality control purpose. The depth-CV and the fusion learning method developed in this article can help broaden those two problem settings to make them more practical in reality, especially in multivariate quality control.

The concept of depth-CD plays a key role in developing our nonparametric multivariate fusion framework. As a distribution function embedded in the parameter space, a depth-CD conveys the level of "confidence" on each possible parameter value, w.r.t. the given data. It is an omnibus of all intrinsic inference forms of any parameter, including the common inferences of point estimates, confidence intervals/regions and p-values. This all-inclusive characteristic affords our fusion scheme the desirable theoretical and numerical properties seen in this article. Given that depth-CD is a general multivariate extension of *CD*, many challenging problems in fusion learning in the scalar or normal setting that have been solved by combining *CD*s can be expected to be solved by using depth-CD if they arise in nonparametric multivariate settings. These include robust inference with outlying studies (Xie, Singh, and Strawderman 2011), exact inference for discrete data (Liu, Liu, and Xie 2014; Yang et al. 2016), efficient inference for heterogeneous studies or network meta-analysis (Clagget, Xie, and Tian 2014; Yang et al. 2014; Liu, Liu, and Xie 2015), or with external data (Xie et al. 2013) scalable split-conquer-combine approaches for massive data (Chen and Xie 2014) and individualized inference for a particular study (Shen, Liu, and Xie 2019). Cheng, Liu, and Xie (2017) gave a brief review on fusion learning via CDs.

For the ease of presentation, the article has focused on fusion of independent studies, with iid observations in each study. But we stress that our fusion framework is quite general and can remain applicable even if these assumptions are violated. For example, our fusion extends to possibly non-iid observations within a study. Specifically, since the fusion is on the inference for a common parameter shared by the studies, our method is valid as long as the estimate within each study converges to the common parameter even if the observations are not identically distributed. To this end, in our bootstrap implementation of the method, the method remains valid as long as bootstrap works, as seen in the non-iid setting covered in Liu (1988). Our method can also be extended in the direction of fusing related studies seen in Li, Hung, and Xie (2020).

Our approach is shown to enable the fusion of multivariate inferences from a wide range of data sources, including studies of irregular, incomplete or heterogeneous of various types. The development of depth-CV here may be further extended to cover different data types in the domains of directional data (data on circles/spheres) (Liu and Singh (1992)) and functional data (López-Pintado and Romo 2009; Claeskens et al. 2014; Narisetty and Nair 2016; Fan and Liu 2019), where applications abound, for example, an efficient fusion of the existing different climate or weather forecast approaches. Those extensions would be worth exploring.

Supplementary Material

Supplementary materials containing all the technical proofs for this article are available online.

Funding

Their research is supported in part by the NSF grants DMS1737857, DMS1812048, DMS2015373, and DMS2027855.

References

Birnbaum, A. (1961), "Confidence Curves: An Omnibus Technique for Estimation and Testing Statistical Hypotheses," *Journal of the American Statistical Association*, 56, 246–249. [2089]

Blaker, H. (2000), "Confidence Curves and Improved Exact Confidence Intervals for Discrete Distributions," *Canadian Journal of Statistics*, 28, 783–798. [2089]

Blaker, H., and Spjøtvoll, E. (2000), "Paradoxes and Improvements in Interval Estimation," *The American Statistician*, 54, 242–247. [2089]

Chatterjee, N., Chen, Y.-H., Maas, P., and Carroll, R. J. (2016), "Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-Level Information From External Big Data Sources," *Journal of the American Statistical Association*, 111, 107–117. [2095,2102]

Chen, X., and Xie, M. (2014), "A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data," *Statistica Sinica*, 1655–1684. [2103]

Chen, Y.-H., Chatterjee, N., and Carroll, R. J. (2013), "Using Shared Genetic Controls in Studies of Gene-Environment Interactions," *Biometrika*, 100, 319–338. [2095,2102]

Cheng, J., Liu, R., and Xie, M. (2017), "Fusion Learning," *Wiley StatsRef: Statistics Reference Online*, (Editor W. Piegorsch). [2103]

Claeskens, G., Hubert, M., Slaets, L., and Vakili, K. (2014), "Multivariate Functional Halfspace Depth," *Journal of the American Statistical Association*, 109, 411–423. [2103]

Clagget, B., Xie, M., and Tian, L. (2014), "Meta Analysis With Fixed, Unknown, Study-specific Parameters," *Journal of the American Statistical Association*, 109, 1667–1671. [2103]

Cuesta-Albertos, J. A., and Nieto-Reyes, A. (2008), "The Random Tukey Depth," *Computational Statistics and Data Analysis*, 52, 4979–4988. [2102]

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1–26. [2087]

Efron, B., and Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, CRC press. [2094]

Einmahl, J., Li, J., and Liu, R. (2015), "Bridging Centrality and Extremity: Refining Sample Data Depth Using Extreme Value Statistics," *The Annals of Statistics*, 43, 2738–2765. [2102]

Fan, Y., and Liu, R. (2019), "Antipodal Refection Depth (ARD) for Multivariate and Functional Data and Nonparametric Outlier Detection," *Preprint*. [2103]
Gao, X., and Carroll, R. J. (2017), "Data Integration With High Dimensionality," *Biometrika*, 104, 251–272. [2095,2102]
Graybill, F. A., and Deal, R. (1959), "Combining Unbiased Estimators," *Biometrics*, 15, 543–550. [2096]
Hannig, J., Iyer, H., Lai, R. C. S., and Lee, T. C. M. (2016), "Generalized Fiducial Inference: A Review and New Results," *Journal of American Statistical Association*, 111, 1346–1361. [2103]
Hodges, J. L. (1955), "A Bivariate Sign Test," *The Annals of Mathematical Statistics*, 26, 523–527. [2089]
Hoff, P. (2019), "Smaller p-values Via Indirect Information," Preprint. [2102]
Jordan, S., and Krishnamoorthy, K. (1995), "Confidence Regions for the Common Mean Vector of Several Multivariate Normal Populations," *The Canadian Journal of Statistics*, 23, 283–297. [2096]
Kass, R. E. (2011), "Statistical Inference: The Big Picture," *Statistical Science*, 26, 1–9. [2103]
Li, C., Hung, Y., and Xie, M. (2020), "A Sequential Split-and-Conquer Approach for the Analysis of Big Dependent Data in Computer Experiments," *Canadian Journal of Statistics*, https://doi.org/10.1002/cjs.11559. [2103]
Li, J., and Liu, R. (2008), "Multivariate Spacings Based on Data Depth: I. Construction of Nonparametric Multivariate Tolerance Regions," *The Annals of Statistics*, 36, 1299–1323. [2103]
Lin, S.-H., Lee, J. C., and Wang, R. (2007), "Generalized inferences on the common mean vector of several multivariate normal populations," *Journal of Statistical Planning and Inference*, 137, 2240–2249. [2096]
Littell, R. C., and Folks, J. L. (1973), "Asymptotic Optimality of Fisher's Method of Combining Independent Tests II," *Journal of the American Statistical Association*, 68, 193–194. [2095]
Little, M. P., Heidenreich, W. F., and Li, G. (2010), "Parameter Identifiability and Redundancy: Theoretical Considerations," *PloS ONE*, 5, e8915. [2095]
Liu, D., Liu, R., and Xie, M. (2014), "Exact Meta-analysis Approach for Discrete Data and Its Application to 2 × 2 Tables With Rare Events," *Journal of the American Statistical Association*, 109, 1450–1465. [2094,2103]
——— (2015), "Multivariate Meta-analysis of Heterogeneous Studies Using Only Summary Statistics: Efficiency and Robustness," *Journal of the American Statistical Association*, 110, 326–340. [2088,2095,2096,2102,2103]
Liu, R. (1990), "On a Notion of Data Depth Based on Random Simplices," *The Annals of Statistics*, 18, 405–414. [2087,2089]
Liu, R., Parelius, J., and Singh, K. (1999), "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference," *The Annals of Statistics*, 27, 783–840. [2087]
Liu, R., and Singh, K. (1992), "Ordering Directional Data: Concepts of Data Depth on Circles and Spheres," *The Annals of Statistics*, 1468–1484. [2103]
——— (1993), "A Quality Index Based on Data Depth and Multivariate Rank Tests," *Journal of the American Statistical Association*, 88, 252–260. [2103]
——— (1997), "Notions of Limiting P Values Based on Data Depth and Bootstrap," *Journal of the American Statistical Association*, 92, 266–277. [2091,2093]
Liu, R. Y. (1988), "Bootstrap Procedures Under Some Non-iid Models," *The Annals of Statistics*, 16, 1696–1708. [2103]
López-Pintado, S., and Romo, J. (2009), "On the Concept of Depth for Functional Data," *Journal of the American Statistical Association*, 104, 718–734. [2103]
Mahalanobis, P. (1936), "On the Generalized Distance in Statistics," in *Proceedings of the National Academy of India*, 12, pp. 49–55. [2089]
Narisetty, N. N., and Nair, V. N. (2016), "Extremal Depth for Functional Data and Applications," *Journal of the American Statistical Association*, 111, 1705–1714. [2103]
Normand, S. (1999), "Meta-analysis: Formulating, Evaluating, Combining, and Reporting," *Statistics in Medicine*, 18, 321–359. [2087]
Pal, N., Lin, J. J., Chang, C. H., and Kumar, S. (2007), "A Revisit to the Common Mean Problem: Comparing the Maximum Likelihood Estimator With the Graybill-Deal Estimator," *Computational Statistics & Data Analysis*, 51, 5673–5681. [2096]
Reid, N., and Cox, D. R. (2015), "On Some Principles of Statistical Inference," *International Statistical Review*, 83, 293–308. [2103]
Rothenberg, T. J. (1971), "Identification in Parametric Models," *Econometrica*, 39, 577–591. [2095]
Rousseeuw, P. J., and Struyf, A. (1998), "Computing Location Depth and Regression Depth in Higher Dimensions," *Statistics and Computing*, 8, 193–203. [2102]
Schulze, R. (2004), *Meta-analysis-A Comparison of Approaches*, Lower Saxony, Germany: Hogrefe & Huber Publishers. [2100]
Schweder, T., and Hjort, N. (2002), "Confidence and Likelihood," *Scandinavian Journal of Statistics*, 29, 309–332. [2089]
——— (2016), *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*, New York: Cambridge University Press. [2087,2089,2090]
Shen, J., Liu, R., and Xie, M. (2018), "Prediction With Confidence a General Framework for Predictive Inference," *Journal of Statistical Planning and Inference*, 195, 126–140. [2103]
——— (2019), "iFusion: Individualized Fusion Learning," *Journal of the American Statistical Association*, 115, 1251–1267. [2103]
Singh, K., Xie, M., and Strawderman, W. E. (2005), "Combining Information From Independent Sources Through Confidence Distributions," *Annals of Statistics*, 159–83. [2089,2094,2095]
——— (2007), "Confidence Distribution (CD): Distribution Estimator of a Parameter," in *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond. IMS Lecture Notes-Monograph Series*, eds. Liu, R., Strawderman, W., and Zhang, C.-H., Beachwood, Ohio: Institute of Mathematical Statistics, vol. 54, pp. 132–50. [2090]
Sutton, A. J., and Higgins, J. P. T. (2008), "Recent Developments in Meta-analysis," *Statistical Medicine*, 27, 625–650. [2087,2088]
Thornton, S., and Xie, M. (2020), "Bridging Bayesian, frequentist and fiducial (BFF) inferences using confidence distribution", arXiv preprint arXiv:2012.04464. [2103]
Tukey, J. (1975), "Mathematics and the Picturing of Data," in *Proceedings of the International Congress of Mathematicians*, vol. 2, pp. 523–531. [2089]
Xie, M., Liu, R. Y., Damaraju, C. V., and Olson, W. H. (2013), "Institute of Mathematical Statistics", *The Annals of Applied Statistics*, 7, 342–368. [2102,2103]
Xie, M., and Singh, K. (2013), "Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review" (with discussion), *International Statistical Review*, 81, 2–39. [2087,2089,2090,2103]
Xie, M., Singh, K., and Strawderman, W. E. (2011), "Confidence Distributions and a Unifying Framework for Meta-analysis," *Journal of the American Statistical Association*, 106, 320–33. [2094,2103]
Yang, G., Liu, D., Liu, R. Y., Xie, M., and Hoaglin, D. C. (2014), "Efficient Network Meta-Analysis: A Confidence Distribution Approach," *Statistical Methodology*, 20, 105–125. [2095,2102,2103]
Yang, G., Liu, D., Wang, J., and Xie, M.-g. (2016), "Meta-analysis Framework for Exact Inferences With Application to the Analysis of Rare Events," *Biometrics*, 72, 1378–1386. [2103]
Yeh, A., and Singh, K. (1997), "Balanced Confidence Regions Based on Tukey's Depth and the Bootstrap," *Journal of the Royal Statistical Society*, Series B, 59, 639–652. [2093]
Zuo, Y., and Serfling, R. (2000), "General Notions of Statistical Depth Function," *The Annals of Statistics*, 28, 461–482. [2087]