Distribution-Free Prediction Sets for Two-Layer Hierarchical Models

Robin Dunn\textsuperscript{a} \textsuperscript{\circ}, Larry Wasserman\textsuperscript{b,c} \textsuperscript{\circ}, and Aaditya Ramdas\textsuperscript{b,c} \textsuperscript{\circ}

\begin{itemize}
  \item \textsuperscript{a}Advanced Methodology and Data Science, Novartis Pharmaceuticals Corporation, East Hanover, NJ
  \item \textsuperscript{b}Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA
  \item \textsuperscript{c}Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA
\end{itemize}

ABSTRACT

We consider the problem of constructing distribution-free prediction sets for data from two-layer hierarchical distributions. For iid data, prediction sets can be constructed using the method of conformal prediction. The validity of conformal prediction hinges on the exchangeability of the data, which does not hold when groups of observations come from distinct distributions, such as multiple observations on each patient in a medical database. We extend conformal methods to a hierarchical setting. We develop CDF pooling, single subsampling, and repeated subsampling approaches to construct prediction sets in unsupervised and supervised settings. We compare these approaches in terms of coverage and average set size. If asymptotic coverage is acceptable, we recommend CDF pooling for its balance between empirical coverage and average set size. If we desire coverage guarantees, then we recommend the repeated subsampling approach.

Supplementary materials for this article are available online.

1. Introduction

Let $Y_1, Y_2, \ldots, Y_n$ be $n$ independent and identically distributed (iid) observations from a distribution $P$. Suppose $1 - \alpha$ is the user-specified confidence level and $Y$ denotes a new observation drawn from $P$. In set-valued, unsupervised prediction, we want to find a set-valued function $C$ such that

$$P(Y \in C(\alpha)) \geq 1 - \alpha.$$  \hfill (1)

(We should really write $P^{n+1}(Y \in C(\alpha)) \geq 1 - \alpha$ since the randomness is over $Y$ and the training data. We have suppressed the superscript for simplicity.) Vovk, Gammerman, and Shafer (2005) created the method of conformal prediction to construct $C$ such that (1), or $P(Y \in C(X;\alpha)) \geq 1 - \alpha$ in the supervised case, holds for all distributions $P$. In other words, conformal methods yield distribution-free prediction sets.

A fundamental assumption of the usual conformal method is that the data are iid (or, at least, exchangeable). We extend conformal methods to the following hierarchical model where the iid assumption fails. Let $P_1, P_2, \ldots, P_k$ be random distributions drawn from $\Pi$. In the unsupervised setting, let $D_j = \{Y_{j1}, Y_{j2}, \ldots, Y_{jm}\}$ be $n_j$ iid observations drawn from $P_j$ for $j = 1, \ldots, k$. It is helpful to imagine that we have $k$ subjects and $D_j$ represents $n_j$ observations on subject $j$. We know the identity of the group (from 1 to $k$) to which each observation belongs. We assume that the values of $n_j$ are fixed prior to data collection.

There are two tasks to consider:

1. Task 1: Predicting an observation on a new subject. Let $P_{k+1} \sim \Pi$ denote a new draw from $\Pi$ (a new subject) and let $Y \sim P_{k+1}$. The goal is to construct a prediction set for $Y$ using the training data $D_1, \ldots, D_k$.

2. Task 2: Predicting a new observation on one of the current subjects. Let $Y$ denote a new draw from one of the distributions $P_j, 1 \leq j \leq k$. We want a prediction set for $Y$ based on the training data.

In Task 1, the new $Y$ is not exchangeable with the observations from any single observed distribution. In addition, if the training data contains multiple observations from at least one of $P_1, \ldots, P_k$, then the new $Y$ is not exchangeable with the full training data either. Rather than applying standard conformal methods, this setting requires novel approaches that build on the exchangeability of the distributions and the exchangeability of the observations from a given distribution. By contrast, one valid approach to Task 2 is to construct conformal sets using only the data from the subject of interest. We consider that method, but we also incorporate shrinkage and borrowing strength into a second conformal approach. By leveraging data across subjects, the latter method may produce smaller sets.

We have described these tasks in the unsupervised setting, but we also consider Task 1 in the supervised setting. One example of a supervised two-layer hierarchical model is the random effects working model $Y_{ji} = \beta_{0j} + \beta_{1j}X_{ji} + \epsilon_{ji}$, where $\epsilon_{ji} \sim N(0, \sigma^2)$. Here $P_j$ denotes the true underlying distribution of $(X, Y)$ for group $j$. Suppose this random effects model represents the true relationship between $X$ and $Y$, and suppose $X \sim N(0,1)$. Then drawing $(X_j, Y_j) \sim P_j$ amounts to drawing $X_j \sim N(0,1)$ and $Y_j \sim N(\beta_{0j} + \beta_{1j}X_j, \sigma^2)$. Furthermore, suppose that $P_\beta$ represents the distribution of $(\beta_{0j}, \beta_{1j})$ over the full population. Then drawing $P_j \sim \Pi$ reduces to drawing...
(β_{0j}, β_{1j}) \sim P_β.\) As discussed in Section 2, it is possible to use a parametric working model to get valid prediction sets even if the model is wrong.

1.1. Related Work

Key early references on conformal prediction include Vovk, Gammerman, and Shafer (2005) and Shafer and Vovk (2008). The literature on conformal prediction is quickly growing in several overlapping directions. Developments on conformal prediction include connections to traditional statistical methods, extensions to flexible settings, and implementations that are computationally efficient. Work in these directions includes interpolations between marginal and conditional coverage (Lei and Wasserman 2014; Barber et al. 2021b), extensions to multiclass set-valued classification (Sadinle, Lei, and Wasserman 2018), Mondrian conformal approaches that ensure validity within categories (Vovk, Gammerman, and Shafer 2008), anti-conservative bounds on coverage, methods for variable importance, and computationally efficient sample-splitting methods (Lei et al. 2018). Many open problems remain in extending conformal methods to new contexts.

Random effects models are common examples of two-layer hierarchical models. Laird and Ware (1982) provide foundational work on the structure and estimation of random effects models for repeated-measures data. The authors note that random effects allow researchers to model both within- and between-subject variation, often using parameters that have natural interpretations. For instance, random effects models are used for prediction by some researchers in parametric settings (Cairns and Sedransk 1991; Booth and Hobert 1998; Schofield et al. 2015). As an alternative to the random effects parametric assumptions, Craggett, Xie, and Tian (2014) develop methods for inference on the quantiles of study-level parameters without distributional assumptions on these parameters. Thus, researchers have developed some approaches for inference and prediction in random effects parametric settings and for inference on study-specific parametric quantities without distributional assumptions. To the best of our knowledge, there are no papers on valid distribution-free prediction for two-layer hierarchical settings.

1.2. Paper Outline

Section 2 reviews conformal prediction. Sections 3–5 each present methods and simulations for conformal prediction in the two-layer hierarchical setting. Section 3 considers unsupervised prediction on a new distribution. Section 4 considers supervised prediction on a new distribution. Section 5 considers unsupervised prediction on an observed distribution. Section 6 implements our supervised prediction methods on data from a sleep deprivation study. Section 7 provides concluding remarks. In the supplementary material, Appendix A contains proofs and Appendix B contains additional simulations. Code is available at https://github.com/RobinMDunn/ConformalTwoLayer.

2. Background on Conformal Prediction

Conformal prediction is a general method for obtaining distribution-free prediction sets with confidence guarantees. Here, we review some background on conformal prediction.

The Unsupervised Case. Let \(Y_1, \ldots, Y_n \in \mathcal{Y}\) be iid observations from a distribution \(P\), and let \(Y_{n+1}\) denote a new draw from \(P\). The goal of conformal prediction is to construct a set \(C(\alpha)\) based on the training data \(Y_1, \ldots, Y_n\) such that \(P(Y_{n+1} \in C(\alpha)) \geq 1 - \alpha\) for every distribution \(P\). When \(\mathcal{Y} \subseteq \mathbb{R}\) or, more generally, \(\mathcal{Y}\) is a linearly ordered set, Theorem 1 provides one valid construction based on order statistics. We say that a method can produce nontrivial sets if \(C(\alpha)\) may be a strict subset of \(\mathcal{Y}\).

**Theorem 1.** Define \(C(\alpha) = \{Y_{(r)}, Y_{(u)}\}\), where \(r = \lceil (n + 1)(\alpha/2)\rceil\) and \(s = \lfloor (n + 1)(1 - \alpha/2)\rfloor\). (If \(r < 1\) and \(s > n\), set \(Y_{(r)} = \min[Y]\) and \(Y_{(u)} = \max[Y]\).) Then for every distribution \(P\), \(P(Y_{n+1} \in C(\alpha)) \geq 1 - \alpha\). This method can produce nontrivial sets if \(n \geq 2/\alpha - 1\).

Theorem 1 relies on the exchangeability of the original sample’s order statistics. For a proof, see Appendix A, supplementary materials. Alternative valid constructions rely on the exchangeability of conformal residuals constructed from the sample. For any \(u \in \mathcal{Y} \subseteq \mathbb{R}^d\), let \(A(u) = (Y_1, \ldots, Y_n, u)\), which can be thought of as the training data augmented with a guess that \(Y_{n+1} = u\). Define the residual (or nonconformity score) \(R_i(u) = \phi(Y_i, A(u))\) where \(\phi : \mathcal{Y} \times \mathcal{Y}^{n+1} \to \mathbb{R}\) is any function that is invariant under permutations of the elements of \(A(u)\). We wish to test the hypothesis \(H_0 : Y_{n+1} = u\). The set of all \(u\) for which we do not reject \(H_0\) at level \(1 - \alpha\) will provide the 100(1 - \(\alpha\))% prediction set. Assuming \(Y_{n+1} = u\), we define

\[
\pi(u) = \frac{1}{n + 1} \sum_{i=1}^{n+1} I(R_i(u) \geq R_{n+1}(u))
\]

which is the \(p\)-value for testing this hypothesis. Intuitively, the \(p\)-value for a given \(u\) is small if the residuals at most of \(Y_1, \ldots, Y_n\) are smaller than the residual at \(u\) (i.e., the \(p\)-value is small if \(Y_{n+1} = u\) does not “conform” to the original sample). \(\pi(u)\) is a valid \(p\)-value because under \(H_0\), \(\pi(u)\) follows a super-uniform distribution over \(t \in [0, 1]\). That is, \(P(\pi(u) \leq t) = P(\pi(u) \leq \lfloor (t(n + 1))/(n + 1) \rfloor) \leq \lfloor (t(n + 1))/(n + 1) \rfloor \leq t\). Often \(P\) is a continuous distribution and \(P(\phi(Y_i, A(u)) = \phi(Y_i, A(u))) = 0\) for \(i \neq j\). In this case, \(\pi(u)\) is uniformly distributed over the set \([1/(n + 1), 2/(n + 1), \ldots, 1]\). We invert the test to define \(C(\alpha) = \{u : \pi(u) \geq \alpha\}\).

**Theorem 2.** For \(C(\alpha)\) as given above, \(P(Y_{n+1} \in C(\alpha)) \geq 1 - \alpha\) for every distribution \(P\). For this method to produce nontrivial sets, it must hold that \(n > 1/\alpha - 1\).

See Vovk, Gammerman, and Shafer (2005) for a proof. For the nontrivial condition, note that \(\pi(u) \geq 1/(n + 1)\) for any \(u\). Hence, if \(n \leq 1/\alpha - 1\), then \(\pi(u) \geq 1/(n + 1) \geq \alpha\) for all \(u\).
There is great flexibility in the choice of nonconformity score \( \phi \). Every choice leads to a prediction set with valid coverage, but different choices may lead to smaller sets. Thus, the choice of \( \phi \) can affect the efficiency of the prediction set but not its validity; see Lei, Robins, and Wasserman (2013).

As an example, let \( R_i(u) = |Y_i - \bar{Y}(u)| \) where \( \bar{Y}(u) = (u + \sum_{i=1}^{n} Y_i)/(n + 1) \) is the mean of the augmented data. Then \( \pi(u) = (n + 1)^{-1} \sum_{i=1}^{n+1} I(Y_i - \bar{Y}(u) \leq |u - \bar{Y}(u)|) \). Another useful nonconformity score is \( \tau_i(u) = 1/\widehat{P}_n(Y_i) \) where \( \widehat{P}_n \) is a density estimator based on the augmented data. Lei, Robins, and Wasserman (2013) showed that this choice is minimax optimal when some conditions hold. Finally, the density estimator could be based on a working parametric model such as \( Y = (Q_\theta : \theta \in \Theta) \). For example, we could use \( R_i(u) = 1/q\widehat{P}(Y_i) \), where \( \widehat{P}(u) \) is the maximum likelihood estimate based on \((Y_1, \ldots, Y_n, u)\). Importantly, this choice of residual is valid even if \( P \) is not in \( Q \).

### The Supervised Case

In this case the data are \((X_1, Y_1), \ldots, (X_n, Y_n) \sim P\). Let \((X, Y) \sim P\) be a new observation. We want a set \( C(X; \alpha) \) such that \( P(Y \in C(X; \alpha)) \geq 1 - \alpha \) for all \( P \). As one possibility, fix \((x, y)\) and let \( \widehat{m}_{(x,y)} \) be a regression estimator based on the augmented data \((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})\) with \((X_{n+1}, Y_{n+1}) = (x, y)\). Where \( R_i(x, y) = |Y_i - \widehat{m}_{(x,y)}(X_i)| \), let

\[
\pi(x, y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_i(x, y) \geq R_{n+1}(x, y))
\]

and

\[C(x; \alpha) = \{ y : \pi(x, y) \geq \alpha \} .\]

Then \( \inf_P P(Y_{n+1} \in C(X_{n+1}; \alpha)) \geq 1 - \alpha \).

A second useful choice of conformal residual is \( R_i(x, y) = 1/\widehat{P}(X_i, Y_i) \) where \( \widehat{P} \) is a joint density estimate based on the augmented data. See Lei, Robins, and Wasserman (2013) for more details.

### Methods for Each Task

We have described an order statistic method and a residual method that are valid under minimal assumptions. In the one-dimensional unsupervised case, the order statistic approach is valid if the data are exchangeable. This method is a simpler construction that does not require data augmentation or the choice of a nonconformity score. Alternatively, the residual approach affords more flexibility through the construction of a nonconformity score, and it extends beyond the one-dimensional and unsupervised setting. This method relies on the exchangeability of the residuals, which holds for any permutation-invariant \( \phi \) when the underlying data are exchangeable. To construct prediction sets for a new observation on a new subject (Task 1), we use methods based on the original sample's order statistics in the unsupervised setting, and we use the residual method in the supervised setting. To construct prediction sets for a new observation on an existing subject (Task 2), we use the residual method. In the Task 2 setting, the residual method allows us to implement a nonconformity score based on a shrinkage estimator.

### 3. Unsupervised Prediction for a New Distribution

To develop valid prediction sets in the two-layer hierarchical setting, we start with the unsupervised version. Recall that the data come in groups \( D_1, \ldots, D_k \) and each group has iid data \( D_j = \{Y_{j1}, \ldots, Y_{jm_j}\} \sim P_j \), where \( Y_{j1}, \ldots, Y_{jm_j} \in Y \subseteq \mathbb{R} \) and \( P_1, \ldots, P_k \sim \Pi \). More generally, if \( Y \) is a linearly ordered set, the unsupervised methods that do not require continuous CDFs will still hold. (These are Methods 0, 2, and 3, which we will describe in this section.) Assuming a new distribution \( P_{k+1} \sim \Pi \) and \( Y \sim P_{k+1} \), we want a prediction region for \( Y \). We can construct prediction sets \( C(\alpha) \) such that \( Y \) is contained in \( C(\alpha) \) with probability at least \( 1 - \alpha \), over the randomness in the initial sample and \( Y \sim \tilde{\Pi} \), where \( \tilde{\Pi} = \int Pd\Pi(P) \). More formally, for \( y \in Y \) and \( y_j \in Y_{nj} \), we define the distribution over these sources of randomness as

\[
\tilde{\Pi}(y, y_1, y_2, \ldots, y_k) = \left\{ \int P(Y \leq y)d\Pi(P) \right\} \times \left\{ \prod_{j=1}^{k} \left\{ \int P(Y_j \leq y_j)d\Pi(P) \right\} \right\} .
\]

We overload notation slightly by allowing \( \tilde{\Pi} \) to refer to both the probability measure and its CDF. We construct prediction sets \( C(\alpha) \) that satisfy \( \tilde{\Pi}(Y \in C(\alpha)) \geq 1 - \alpha \). Method 0 requires equal \( n_j \) across groups, while the other methods allow varying \( n_j \). For validity, the non-asymptotic methods (Methods 0, 2, and 3) only require \( k \geq 1 \) and \( n_j \geq 1 \), \( j = 1, \ldots, k \). On the other hand, some methods place requirements on \( k \) and \( n_j \) for nontrivial sets. We note these requirements in the theorems associated with each method.

### 3.1. Method 0: Double Conformal

The hierarchical set-up involves two levels of randomness. At the level of group \( j \), we have independent observations from a distribution \( P_j \). At the distribution level, each distribution is sampled from \( \Pi \). A “double conformal” method is one natural approach that incorporates this hierarchical structure when all \( n_j \) values are equal, \( j = 1, \ldots, k \). If the samples are not equally sized, then we could work with \( \min n_j \) observations per group, sampled uniformly at random without replacement. This method first constructs a prediction set within each group and then uses those sets to construct a final prediction set across groups.

At the group level, let \( C_j(\alpha/2) = \{y_{ij}, u_j\} \) be the 100(1 - \( \alpha/2 \))% prediction set obtained by applying the method in Theorem 1 at level \( \alpha/2 \) to group \( j = 1, \ldots, k \). We construct a vector of \( k \) lower bounds \( \{\ell_1, \ldots, \ell_k\} \) and \( k \) upper bounds \( \{u_1, \ldots, u_k\} \). Using the order statistics from those vectors, we set \( C^d\{(\alpha)\} = [\ell_1, u_1, \ldots, \ell_k, u_k] \), where \( r = [(k+1)\alpha/4] \) and \( s = [(k+1)(1-\alpha/4)] \). If \( r < 1 \) and \( s > k \), let \( \ell_1 = \min \{Y\} \) and \( u_k = \max \{Y\} \). By Theorem 3, \( C^d\{(\alpha)\} \) is a valid 100(1 - \( \alpha \))% prediction set for a new \( Y \) from a new group.

### Theorem 3

If all groups have an equal number of observations \( n_1 = n_2 = \cdots = n_k \), then \( \tilde{\Pi}(Y \in C^d\{(\alpha)\}) \geq 1 - \alpha \) for \( C^d\{(\alpha)\} \) as defined above. This method can produce nontrivial sets if \( k \geq 4\alpha/\alpha - 1 \) and \( n_1 \geq 4\alpha/\alpha - 1 \).
For a proof, see Appendix A, supplementary materials. While this method is valid, our results show that this method overcovers. Thus, we turn to several methods that are better choices.

3.2. Method 1: Pooling CDFs

To produce smaller prediction sets, we construct an empirical CDF within each group. We average these CDFs across groups, and we determine the prediction set bounds based on the quantiles of the average of CDFs. If \( Y \) has a continuous distribution, this method is asymptotically valid as \( k \to \infty \) for any values of \( n_j \geq 1, j = 1, \ldots, k \).

Formally, for any group \( j \), the empirical CDF is defined as

\[
\hat{F}_j(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} I(Y_{ji} \leq t).
\]

We set

\[
\widehat{q}_k(\alpha) = \inf \left\{ t \in \mathcal{Y} : \frac{k}{1} \sum_{j=1}^{k} \hat{F}_j(t) \geq \alpha \right\}.
\]

Then an asymptotic \( 1 - \alpha \) prediction set is \( C_{\text{poolCDF}}(\alpha) = [\widehat{q}_k(\alpha/2), \widehat{q}_k(1 - \alpha/2)] \). For a proof of Theorem 4, see Appendix A, supplementary materials.

Theorem 4. Assume that \( Y \) has a continuous distribution. For \( C_{\text{poolCDF}}(\alpha) \) as defined above, \( \Pi(Y \in C_{\text{poolCDF}}(\alpha)) \to 1 - \alpha \) as \( k \to \infty \). This method can produce nontrivial sets for any \( k \geq 1 \) and \( n_j \geq 1, j = 1, \ldots, k \).

3.3. Method 2: Subsampling Once

While the previous method is asymptotically valid under a CDF condition, we may desire a method without the continuous distribution requirement and with both reasonable coverage and finite sample validity. To achieve these goals, we propose a method based on subsampling. Draw one observation uniformly at random from each group. Then the data consist of \( k \) iid observations \( Y_1, Y_2, \ldots, Y_k \) from \( \mathcal{F} = \int \mathcal{F} \Pi(d\mathcal{F}) \). We define a prediction set \( C_{\text{sub}}(\alpha) = [Y_{(\ell)}, Y_{(s)}], \) where \( r = \lfloor (k + 1)(\alpha/2) \rfloor \) and \( s = \lfloor (k + 1)(1 - \alpha/2) \rfloor \). If \( k < 1 \) and \( s > k \), set \( Y_{(s)} = \min(Y) \) and \( Y_{(r)} = \max(Y) \). Since the subsample contains \( k \) iid draws from \( \Pi \), Theorem 5 follows from Theorem 1.

Theorem 5. For \( C_{\text{sub}}(\alpha) \) as defined above, \( \Pi(Y \in C_{\text{sub}}(\alpha)) \geq 1 - \alpha \). This method can produce nontrivial sets if \( k \geq 2/\alpha - 1 \) and \( n_j \geq 1, j = 1, \ldots, k \).

3.4. Method 3: Repeated Subsampling

The single subsampling approach is simple and valid, but it ignores most of the data. Due to the use of a single subsample, the results may be insufficiently reproducible. We address these shortcomings by incorporating \( B \) subsamples of a single observation from each of the \( k \) groups. Gupta, Kuchibhotla, and Ramdas (2020) developed the method of constructing conformal prediction sets through repeated subsampling in the case of exchangeable data. Suppose \( Y_{(1)}^b, Y_{(2)}^b, \ldots, Y_{(k)}^b \) are the ordered observations from the \( b \)th subsample. Conformal prediction is implicitly testing \( H_0 : Y_{k+1} = u \) versus \( H_1 : Y_{k+1} \neq u \), and the level \( 1 - \alpha \) conformal prediction set is the set of values at which we would not reject \( H_0 \) under the given construction. Thus, within the \( b \)th subsample, the \( p \)-value at \( u \in \mathcal{Y} \) is

\[
\pi_b(u) = \begin{cases} 
1 & \text{if } u \in \left[ Y_{(\lfloor (k+1)/2 \rfloor)}^b, Y_{(\lfloor (k+1)/2 \rfloor)}^b \right] \\
\inf \{ \alpha : u \notin [Y_{(r)}^b, Y_{(s)}^b] \} & \text{otherwise} 
\end{cases},
\]

where \( r = \lfloor (k + 1)(\alpha/2) \rfloor \) and \( s = \lfloor (k + 1)(1 - \alpha/2) \rfloor \). We define a prediction set \( C_{\text{rep}}(\alpha) = \left\{ u : B^{-1} \sum_{b=1}^{B} \pi_b(u) \geq \alpha \right\} \).

Theorem 6. For \( C_{\text{rep}}(\alpha) \) as defined above, \( \Pi(Y \in C_{\text{rep}}(\alpha)) \geq 1 - 2\alpha \). This method can produce nontrivial sets if \( k > 2/\alpha - 1 \) and \( n_j \geq 1, j = 1, \ldots, k \).

Theorem 6 holds because \( (2/B) \sum_{b=1}^{B} \pi_b(u) \) (double the test statistic) is a valid \( p \)-value for the stated test (Rüschendorf 1982; Meng 1994; Barber et al. 2021a; Vovk and Wang 2020). In practice, however, \( C_{\text{rep}}(\alpha) \) has close to 100% coverage. The guaranteed level 1 − 2\( \alpha \) coverage and empirical level \( 1 - \alpha \) coverage is analogous to the coverage of the jackknife+ method (Barber et al. 2021a), which constructs conformal sets through leave-one-out prediction. Furthermore, Tian et al. (2021) show that the multiplicative correction is not always necessary when averaging \( p \)-values. Their Corollary 1 proves that for small enough \( \alpha \), the average of \( p \)-values on one-dimensional normal variables with arbitrary positive correlation (and certain degrees of negative correlation) is a valid \( p \)-value. For a proof of the nontrivial condition, see Appendix A, supplementary materials.

3.5. Unsupervised New Distribution Simulations

To understand the performance of these methods, we consider a study simulation. We begin by generating data from \( k \) distributions. We draw \( \theta_1, \ldots, \theta_k \sim N(0, 1) \). Then we simulate \( Y_{1j_1}, \ldots, Y_{nj_j} \sim N(\theta_j, 1) \) for \( j = 1, \ldots, k \). We use \( n_j = 100 \) observations per group. We vary the number of groups \( k \) from 5 to 100 in increments of 5 and from 200 to 1000 in increments of 100. The repeated subsampling sets use \( B = 100 \) sub-samples. Each simulation generates a data sample, draws a new \( \theta_{k+1} \sim N(0, 1) \) and \( Y \sim N(\theta_{k+1}, 1) \), constructs a prediction set \( C(\alpha) \), determines the size of the prediction set, and checks whether \( Y \in C(\alpha) \). The coverage is the proportion of simulations for which \( Y \in C(\alpha) \). We set \( \alpha = 0.1 \), and we perform 1000 simulations at each value of \( k \).

Figure 1 displays the empirical coverage and average set length from the four unsupervised methods. The double conformal method consistently overcovers, with coverage close to 1. CDF pooling undercovers at small to moderate values of \( k \) (e.g., \( k \leq 35 \)) but has approximately 1 − \( \alpha \) coverage for larger \( k \). Single and repeated subsampling tend to overcover for small to moderate \( k \) and have approximately 1 − \( \alpha \) coverage for large \( k \).

The pooling method has the smallest sets, the single subsampling and repeated subsampling methods have the next smallest sets (mostly on par), and double conformal has the largest sets. (The right panel of Figure 1(b) excludes the double conformal sets, which have average lengths between 8.4 and 8.6 for \( k \geq 200 \).) Appendix B contains simulations that produce similar
behavior on normal data at $n_j \in \{40, 1000\}$ and on nonnormal data.

Figure 1 has considered cases with balanced numbers of observations in each group. We now consider a highly unbalanced case: one group has 200 times as many observations as each of the other groups, and the between-group variation exceeds the within-group variation by three orders of magnitude. We take $Y_{ji} \sim N(\theta_j, \sigma^2 = 0.1)$ where $\theta_j \sim N(0, \tau^2 = 100), n_1 = 1000$, and $n_j = 5$ for $2 \leq j \leq k$. We let $k$ vary from 5 to 100 in increments of 5. Figure 2 shows that the single subsample and repeated subsample methods typically have at least nominal coverage. CDF pooling undercovers for small $k$. For $k \geq 20$, all three methods produce finite prediction intervals, and CDF pooling only undercovers by about 0.05. CDF pooling produces the smallest prediction sets, and single and repeated subsampling have similar average prediction set lengths. Thus, the behavior we observe in this highly unbalanced case is similar to the balanced case.

Overall, CDF pooling appears to be the best choice, with a few caveats. CDF pooling consistently produces the smallest prediction sets, and it achieves nominal or approximately nominal coverage. As a drawback, this method often slightly undercovers for small to moderate $k$, with more notable undercoverage in the small $k$ unbalanced setting. In addition, this method only guarantees coverage asymptotically as $k \to \infty$ and for continuous $Y$. Hence, if we desire a method with consistent simulated coverage, with theoretical guarantees on coverage, or without the CDF requirement, the subsampling conformal methods are better choices. Between the two subsampling methods, we recommend repeated subsampling. This method has guaranteed coverage at level $1 - 2\alpha$, but in practice it tends to cover at level $1 - \alpha$. Furthermore, for moderate $k$, its prediction sets are about the same size as the single subsample method’s sets. Favorably, repeated subsampling yields more reproducible prediction intervals than single subsampling. To understand the variation in these two methods, we consider a single dataset with $k = 100$ groups and $n_j = 100$ observations per group, using the same setup as Figure 1. Based on Figure 1, these methods have similar coverage and size at these parameters. Through 1000 repetitions, we construct 90% prediction intervals using the two subsampling methods. Across simulations, single subsampling has lower bounds between $-3.6$ and $-1.7$, while repeated subsampling has lower bounds between $-2.7$ and $-2.4$. Similarly, single subsampling has upper bounds between $1.5$ and $3.5$, while repeated subsampling has upper bounds between $2.4$ and $2.7$. Thus, as expected, we see less variation in the average prediction set size.
in the prediction intervals constructed through repeated subsampling.

4. Supervised Prediction for a New Distribution

In the supervised case, each group \( D_1, \ldots, D_k \) has iid data \( D_j = \{ (X_{ji}, Y_{ji}), \ldots, (X_{jmi}, Y_{jmi}) \} \sim P_j \), where \( Y_{j1}, \ldots, Y_{jm} \in \mathbb{Y} \subseteq \mathbb{R} \) and \( P_1, \ldots, P_k \sim \Pi \). Each \( X_{ji} \) is a \( p \)-dimensional vector given by \( X_{ji} = (X_{ji}^{(1)}, X_{ji}^{(2)}, \ldots, X_{ji}^{(p)}) \). Suppose we have a new distribution \( P_{k+1} \sim \Pi \) and \( (X, Y) \sim P_{k+1} \). Then \( (X, Y) \sim \tilde{\Pi} \), where \( \tilde{\Pi} = \int Pd\Pi(P) \). Assuming that we only observe \( X = x \), we want a prediction region for \( \tilde{\Pi}(Y = y | X = x) \). To define a distribution, we use \( x \in \mathbb{R}^p \), \( y \in \mathbb{Y} \), \( x_j \in \mathbb{R}^{n_jp} \), and \( y_j \in \mathbb{Y}_j \). Similar to the unsupervised setting, we define a distribution over the randomness in the initial sample and the new \( (X, Y) \sim \tilde{\Pi} \) as

\[
\tilde{\Pi}(x, y, x_1, y_1, x_2, y_2, \ldots, x_k, y_k) = \frac{1}{k} \prod_{j=1}^{k} \left\{ P(X^{(1)}_j \leq x^{(1)}_j, \ldots, X^{(p)}_j \leq x^{(p)}_j, Y \leq y) d\pi(P) \right\}.
\]

We construct sets \( C(x; \alpha) \) such that \( \tilde{\Pi}(Y \in C(X; \alpha)) \geq 1 - \alpha \) for \( (X, Y) \sim \tilde{\Pi} \). As in the unsupervised case, the non-asymptotic, subsampling methods are valid for \( k \geq 1 \) and \( n_j \geq 1 \), \( j = 1, \ldots, k \). We note requirements on \( k \) and \( n_j \) for nontrivial sets.

4.1. Method 1: Pooling CDFs

Similar to the unsupervised setting, we consider methods that average empirical CDFs across groups. We first consider a sample-splitting method that is asymptotically valid as \( k \to \infty \), regardless of the choice of model. Let \( [k] = \{ 1, \ldots, k \} \). We start by using the observations from some strict subset \( k_0 \subseteq [k] \) of the \( k \) groups to fit any model \( \hat{\mu}(X) \) as an estimator of \( \mathbb{E}[Y | X] \).

For instance, \( \hat{\mu}(X) \) could be a single model based on the pooled observations or an average of \( k_0 \) models fit on the individual groups. Importantly, \( \hat{\mu}(X) \) must stay fixed as \( k \) grows. Since \( \hat{\mu}(X) \) will be the center of the prediction set, it is best if \( \hat{\mu}(X) \) is a good approximation to \( \mathbb{E}[Y | X] \). We use the remaining groups to fit the residuals \( R_{j1} = Y_{ji} - \hat{\mu}(X_{ji}), j \in [k] \setminus k_0, i = 1, \ldots, n_j \).

Now for each \( j \in [k] \setminus k_0 \), we define group \( j \)’s empirical CDF of the residuals

\[
\hat{F}_j(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} I(R_{ji} \leq t).
\]

We define

\[
\hat{q}_k(\alpha) = \inf \left\{ t \in \mathbb{R} : \frac{1}{|[k] \setminus k_0|} \sum_{j \in [k] \setminus k_0} \hat{F}_j(t) \geq \alpha \right\}.
\]

For continuous \( Y \), \( C_{\text{poolCDF}}(x; \alpha) = [\hat{\mu}(x) - \hat{q}_k(1 - \alpha), \hat{\mu}(x) + \hat{q}_k(1 - \alpha)] \) is an asymptotic \( 1 - \alpha \) prediction set. For a proof of Theorem 7, see Appendix A, supplementary materials.

Theorem 7. Fit a model \( \hat{\mu}(X) \) as an estimator of \( \mathbb{E}[Y | X] \) using the observations in groups \( k_0 \subseteq [k] \). \( \hat{\mu}(X) \) stays fixed as \( k \) grows.) If \( Y \) has a continuous distribution, then \( \hat{\Pi}(Y \in C_{\text{poolCDF}}(X; \alpha)) \to 1 - \alpha \) as \( k \to \infty \). At any \( x \), this method can produce nontrivial sets for \( k \geq 1 \) and \( n_j \geq 1, j = 1, \ldots, k \).

Under stronger assumptions on the agreement between the true and estimated models, we consider a second asymptotically valid approach \( (k \to \infty) \) that does not require sample splitting. If \( (X_j, Y_j) \sim P_j \) then suppose \( Y_j = \mu(P_j(x)) + \epsilon \), where \( \epsilon \) has a zero-mean distribution. For each group \( j \in [1, \ldots, k] \), we use the observations \( X_{ji}, X_{j2}, \ldots, X_{jmi} \) to fit a model \( \hat{\mu}_j \). At any given \( x \), we define a pooled model \( \hat{\mu}(x) = \int \mu(x) d\pi(P) \) and an estimated pooled model \( \hat{\mu}(x) = k^{-1} \sum_{j=1}^{k} \hat{\mu}_j(x) \).

Under \( \mu \) and \( \hat{\mu} \), we have the residuals \( R_{ji} = Y_{ji} - \hat{\mu}(X_{ji}) \) and \( R_{ji} = [\mu(X_{ji}) - Y_{ji}] \). The residuals have empirical CDFs

\[
\hat{F}_{j,\mu}(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} I(R_{ji}(\mu) \leq t)
\]

\[
\hat{F}_{j,\hat{\mu}}(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} I(R_{ji}(\hat{\mu}) \leq t).
\]

We obtain sample quantiles

\[
\hat{q}_k(\hat{\mu}; \alpha) = \inf \left\{ t \in \mathbb{R} : \int \hat{F}_{j,\hat{\mu}}(t) \geq \alpha \right\}.
\]

Under the assumptions on \( \mu, \hat{\mu}, \) and \( \hat{\Pi} \) stated in Theorem 8, an asymptotic \( 1 - \alpha \) prediction set is

\[
C_{\text{modAvg}}(x; \alpha) = [\hat{\mu}(x) - \hat{q}_k(\hat{\mu}; 1 - \alpha), \hat{\mu}(x) + \hat{q}_k(\hat{\mu}; 1 - \alpha)].
\]
Theorem 8. Suppose $Y \sim \tilde{\Pi}$ has a continuous distribution. If $(X_i, Y_i) \sim P_{\theta}$, suppose $Y_i = \mu(P, X_i) + \epsilon$, where $\epsilon$ has a zero-mean distribution. Assume $\tilde{\mu}$ satisfies

$$
1 \over k \sum_{j = 1}^{k} \sup_{t} \left| \tilde{\mu}_{i}(t) - \tilde{\mu}_{j}(t) \right| \to 0
$$

as $k \to \infty$, and assume that for $\delta > 0$, $\lim_{k \to \infty} \tilde{\Pi}(|\tilde{\mu}(X) - \mu(X)| > \delta) = 0$. For $C_{\text{modAvg}}(x; \alpha)$ as defined above, $\lim_{k \to \infty} \tilde{\Pi}(Y \in C_{\text{modAvg}}(X; \alpha)) = 1 - \alpha$. At any value of $x$, this method can produce nontrivial sets for $k \geq 1$ and $n_j \geq 1$, $j = 1, \ldots, k$.

4.2. Method 2: Subsampling Once

If we desire a method with finite sample coverage guarantees, a conformal method based on subsampling can achieve that goal. As in the unsupervised setting, we randomly select one observation from each of the $k$ groups. This creates a sample of $k$ pairs of iid observations $(X, Y)$. Suppose we have a new data point $(X_{k+1}, Y_{k+1}) \sim P_{k+1}$, but we only observe $X_{k+1}$. Letting $X_{k+1} = x$, we have an augmented $X$ sample $(X_1, \ldots, X_k, X_{k+1})$. For each possible $y$, we test $H_0 : Y_{k+1} = y$ at a $1 - \alpha$ confidence level using the following procedure: Assume $Y_{k+1} = y$, giving an augmented $Y$ sample of $(Y_1, \ldots, Y_k, Y_{k+1})$. Using the sample augmented with $(x, y)$ as training data, fit a model $\tilde{\mu}_{(x,y)}(X)$ as an estimator of $E[Y|X]$. Then compute nonconformity scores $R_i(x, y) = |\tilde{\mu}_{(x,y)}(X_i) - Y_i|, i = 1, \ldots, \hat{k} + 1$. The $p$-value for the test of $H_0 : Y_{k+1} = y$ is $\pi(x, y) = (k + 1) - \sum_{i=1}^{\hat{k}+1} I(R_i(x, y) \geq R_{k+1}(x, y))$. The $1 - \alpha$ conformal prediction set is $C_{\text{sup}}(x; \alpha) = \{y \in \mathbb{R} : \pi(x, y) \geq \alpha\}$.

Theorem 9. For $C_{\text{sub}}(x; \alpha)$ as defined above, $\tilde{\Pi}(Y \in C_{\text{sub}}(X; \alpha)) \geq 1 - \alpha$. For $C_{\text{sub}}(x; \alpha)$ to be nontrivial, it must hold that $k > 1/\alpha - 1$ and $n_j \geq 1, j = 1, \ldots, k$.

Since the subsample of $k$ observations is an iid sample, Section 2 justifies this method. Note that for any $(x, y)$, $\pi(x, y) \geq 1/(k + 1)$. Hence, if $k \leq 1/\alpha - 1$, then $\pi(x, y) \geq 1/(k + 1) \geq \alpha$ will hold for all $(x, y)$.

4.3. Method 3: Repeated Subsampling

Since a single subsample ignores most of the data, it may be inadequately reproducible. To improve the reproducibility, we modify the previous method to incorporate $B$ subsamples of a single observation from each of the $k$ groups. For the $b$th subsample, $(X_{bs}, Y_{bs}), \ldots, (X_{bs}, Y_{bs})$ contains one observed pair from each of the $k$ groups. Suppose $X_{k+1} = x$ is a new observation from a new group. Conformal prediction is implicitly testing $H_0 : Y_{k+1} = y$ versus $H_1 : Y_{k+1} \neq y$, and the level $1 - \alpha$ conformal prediction set is the set of values at which we would not reject $H_0$ under the given construction. Using the $b$th subsample augmented with $(x, y)$, we construct residuals $R_{bs}(x, y)$ in the same manner as Section 4.2. Then $\pi_{bs}(x, y) = (k + 1)^{-1} \sum_{i=1}^{\hat{k}+1} I(R_{bs}(x, y) \geq R_{b,k+1}(x, y))$ is a valid $p$-value for the stated test. We construct $\pi_{bs}(x, y)$ for $B$ subsamples. We define $C_{\text{rep}}(x; \alpha) = \{y : B^{-1} \sum_{b=1}^{B} \pi_{bs}(x, y) \geq \alpha\}$.

Theorem 10. For $C_{\text{rep}}(x; \alpha)$ as defined above, $\tilde{\Pi}(Y \in C_{\text{rep}}(X; \alpha)) \geq 1 - 2\alpha$. For $C_{\text{rep}}(x; \alpha)$ to be nontrivial, it must hold that $k > 1/\alpha - 1$ and $n_j \geq 1, j = 1, \ldots, k$.

Similar to Theorem 6 in the unsupervised case, Theorem 10 is true because $(2/B) \sum_{b=1}^{B} \pi_{bs}(x, y)$ is a valid $p$-value for the stated test. As in the unsupervised case, $C_{\text{rep}}(x; \alpha)$ has empirical coverage of approximately $1 - \alpha$. The nontrivial condition in Theorem 10 holds for the same reason as the nontrivial condition in Theorem 9.

4.4. Supervised New Distribution Simulations

We explore the supervised prediction methods through a simulation study. To generate data from $k$ distributions, we draw $\theta_1, \ldots, \theta_k \sim N(\mu, \tau^2), X_1, \ldots, X_{nj} \sim N(0, 1)$, and $\epsilon_{1j}, \ldots, \epsilon_{nj} \sim N(0, 1)$. We let $Y_{ij} = \theta_jX_{ij} + \epsilon_{ij}, j = 1, \ldots, k$, $i = 1, \ldots, nj$. Then we draw a new $X \sim N(0, 1)$, $\theta_{k+1} \sim N(\mu, \tau^2)$, and response $Y \sim N(\theta_{k+1}X, 1)$. Treating $\theta_{k+1}$ and $Y$ as unknown, we wish to predict $Y$ from the observed $X = x$. For the CDF pooling method, we use the approach justified by Theorem 7. We pool the observations from $k_0 = \lfloor k/2 \rfloor$ groups to fit a one-parameter linear regression model $\hat{\mu}(X) = \hat{\theta}X$. Then we use the remaining groups for quantile estimation. For the subsampling methods, we fit $\hat{\mu}_{(x,y)}(X; \hat{\theta}) = \hat{\theta}X$ using subsamples of one observation per group, augmented with $(x, y)$. For repeated subsampling, we use $B = 100$ subsamples.

We draw $n_1 = 100$ observations per group. We vary the number of groups $(k)$ from 20 to 100 in increments of 5 and from 200 to 1000 in increments of 100. To draw the $\theta$ parameters, we set $\mu = 0$ and $\tau^2 = 1$. In Appendix B, we see similar behavior for $n_1 \in \{20, 1000\}$ and for $\mu = 1$ and $\tau^2 = 0.1$. We perform 1000 simulations at each $k$. We set $\alpha = 0.1$. Each simulation generates a data sample, draws a new $(X, Y)$ from a new distribution, constructs a prediction set $C(X; \alpha)$, determines the size of the set, and checks whether $Y \in C(X; \alpha)$.

Figure 3 shows the coverage and average length of the prediction sets from these supervised methods. The coverage is the proportion of simulations for which $Y \in C(X; \alpha)$. All three methods have coverage close to $1 - \alpha$ for all $k$. For small $k$, repeated subsampling often overcovers by up to 0.05. The pooling sets are the smallest, followed by the single subsampling sets, and the repeated subsampling sets are the largest. CDF pooling appears to be the best choice in this setting, with the caveats that it has asymptotic coverage and $Y$ is continuous. If choosing between the subsampling methods, we recommend repeated subsampling. While single subsampling often produces slightly smaller prediction sets, the average size differs by less than 1% for $k \geq 300$ in our simulations. Furthermore, the results from repeated subsampling are more reproducible. Similar to the unsupervised case, we use the setup from Figure 3 to create a single dataset with $k = 100$ groups and $nj = 100$ observations per group. In 1000 repetitions, we construct 90% prediction intervals at $X = 1$ using the two subsampling methods. Across simulations, single subsampling has lower bounds between $-3.7$ and $-1.2$, while repeated subsampling has lower bounds between $-2.7$ and $-2.4$. Similarly, single subsampling has upper bounds between 1.6 and 3.6, while repeated subsampling has upper bounds between 2.4 and 2.7.
Again, the repeated subsampling intervals have less variation than the single subsampling intervals.

5. Unsupervised Prediction for an Observed Group

Task 2 considers unsupervised prediction of a new observation on an existing subject. For each subject \( j \), we observe \( n_j \) iid real-valued samples \( Y_{j1}, \ldots, Y_{jn_j} \sim P_j \). We assume \( P_1, \ldots, P_k \sim \Pi \). We assume without loss of generality that we wish to predict a new observation from subject 1. Hence, this setting’s probability distribution \( \Pi^{(1)} \) accounts for the randomness in both the training sample and this new observation \( Y_{1,n_1+1} \sim P_1 \). For \( y_1 \in \mathbb{R}^{n_1+1} \) and \( y_j \in \mathbb{R}^{n_j}, 2 \leq j \leq k \), we define

\[
\Pi^{(1)}(y_1, y_2, \ldots, y_k) = \left\{ \prod_{i=1}^{n_1+1} P(Y_{1i} \leq y_{1i}) d\Pi(P) \right\} \times \left\{ \prod_{j=2}^{k} \left[ \prod_{i=1}^{n_j} P(Y_{ji} \leq y_{ji}) d\Pi(P) \right] \right\}.
\]

Our conformal prediction sets \( C(\alpha) \) satisfy \( \Pi^{(1)}(Y_{1,n_1+1} \in C(\alpha)) \geq 1 - \alpha \).

We explore two conformal methods to capture the new observation from subject 1. The first method is a standard conformal procedure using subject 1’s data. The second method “borrows” information from other subjects to obtain a shrinkage estimator of the mean of subject 1’s data. Then it performs conformal prediction using this shrinkage estimator. The validity of either method follows from the usual theory described in Section 2, but the shrinkage approach may lead to smaller prediction sets. While validity holds for \( n_1 \geq 1 \), both methods require \( n_1 > 1/\alpha - 1 \) for nontrivial sets.

5.1. Method 1: Isolate Single Group

The simplest approach constructs conformal prediction sets for subject 1 using only subject 1’s observed data. We propose a new \( y \), and we wish to test \( H_0 : Y_{1,n_1+1} = y \) at a \( 1 - \alpha \) confidence level. Letting \( Y_{1,n_1+1} = y \), we have an augmented data vector \((Y_{11}, \ldots, Y_{1,n_1}, Y_{1,n_1+1})\) for subject 1. We define \( Y_{1i} = (1/(n_1 + 1)) \sum_{j=1}^{n_1+1} Y_{1j} \). Then we calculate nonconformity scores \( R_i = |Y_{1j} - Y_{1i}|, i = 1, \ldots, n_1 + 1 \). The \( p \)-value for the test of \( H_0 : Y_{1,n_1+1} = y \) is \( \pi(y) = (1/(n_1 + 1)) \sum_{i=1}^{n_1+1} I(R_i \geq R_{n_1+1}) \).

We invert this test to obtain a \( 1 - \alpha \) conformal prediction set \( C_{\text{isolate}}(\alpha) = \{ y : \pi(y) \geq \alpha \} \). Since this approach uses con-
formal methods on iid observations from a single distribution, Theorem 2 justifies Theorem 11.

**Theorem 11.** For \( C_{\text{Isolate}}(\alpha) \) as defined above, \( \prod^{(1)}(Y_{1,n_1+1} \in C_{\text{Isolate}}(\alpha)) \geq 1 - \alpha \). For \( C_{\text{Isolate}}(\alpha) \) to be nontrivial, it must hold that \( n_1 > 1/\alpha - 1 \).

### 5.2. Method 2: James–Stein Shrinkage

A conformal method that borrows strength across distributions may yield tighter prediction intervals, especially if the k distributions are “close” to each other. To use the data from all subjects, we work with a conformal residual based on shrinkage. Again, we propose a new value of \( y \), and we wish to test \( H_0 : Y_{1,n_1+1} = y \) at a \( 1 - \alpha \) confidence level. We define \( \tilde{Y}_1 = (1/(n_1 + 1)) \sum_{i=1}^{n_1+1} Y_{i,i} \). Then for \( j = 2, \ldots, k \), we define \( \tilde{Y}_j = \sqrt{n_j}^{-1} \sum_{i=1}^{n_j} Y_{i,j} \). Let \( v = k^{-1} \sum_{j=1}^{k} \tilde{Y}_j \), and let \( \hat{\sigma}^2 \) be the sample variance of \( (\tilde{Y}_1, \ldots, Y_{1,n_1}, Y_{1,n_1+1}) \).

Now in place of \( \tilde{Y}_1 \) in the nonconformity scores, we use the James–Stein shrinkage estimator:

\[
\tilde{y}_1 = \left( 1 - \frac{(k - 2)\hat{\sigma}^2/(n_1 + 1)}{\sum_j (\tilde{Y}_j - v)^2} \right) + (\tilde{Y}_1 - v) + v,
\]

where \((x)_+ = \max(x, 0)\). \( \tilde{y}_1 \) is defined when \( k \geq 2 \). The rest of the procedure mirrors the previous method. We calculate nonconformity scores \( R_i = |Y_{1,i} - \tilde{Y}_i|, i = 1, \ldots, n_1 + 1 \). For the proposed \( y \), we obtain a p-value \( \pi(y) = (1/(n_1 + 1)) \sum_{i=1}^{n_1+1} I(R_i \geq R_{n_1+1}) \). The 1 - \( \alpha \) conformal prediction set is \( C_{\text{Shrinkage}}(\alpha) = \{ y : \pi(y) \geq \alpha \} \). Theorem 2 justifies Theorem 12.

**Theorem 12.** For \( C_{\text{Shrinkage}}(\alpha) \) as defined above, \( \prod^{(1)}(Y_{1,n_1+1} \in C_{\text{Shrinkage}}(\alpha)) \geq 1 - \alpha \). For \( C_{\text{Shrinkage}}(\alpha) \) to be nontrivial, it must hold that \( n_1 > 1/\alpha - 1 \) and \( k \geq 2 \).

### 5.3. Unsupervised Observed Group Simulations

We compare these methods under two data generation processes. We draw subject-specific means \( \theta_1, \ldots, \theta_k \sim N(0, 1) \). Then for \( j = 1, \ldots, k \), we generate \( Y_{1,j}, Y_{2,j}, \ldots, Y_{n_j} \sim N(\theta_j, \sigma^2) \).

We consider \( \sigma^2 = 1 \) and \( \sigma^2 = 100 \). Across all simulations, we set \( n_j = 20 \). We vary \( k \) from 5 to 1000 in increments of 5. At each choice of \( k \), we perform 1000 simulations at \( \alpha = 0.1 \). Each simulation generates a data sample, draws another observation \( Y_{1,n_1+1} \sim N(\theta_1, \sigma^2) \) from subject 1’s distribution, constructs a prediction set \( C(\alpha) \) for subject 1, determines the size of the prediction set, and checks whether \( Y_{1,n_1+1} \in C(\alpha) \).

Figure 4 shows the empirical coverage at \( 1 - \alpha = 0.9 \), when \( \sigma^2 = 1 \) or \( \sigma^2 = 100 \). The coverage is typically about 0.9, with no clear difference between methods. Figure 5 plots the average size of the conformal sets in both data set-ups. When \( \sigma^2 = 1 \), the two methods produce sets with similar length. When \( \sigma^2 = 100 \), shrinkage consistently produces smaller sets than using only group 1’s observations. This shows that shrinkage is especially beneficial when the within-group variance is high, relative to the between-group variance. There does not appear to be a trend in set size as the number of groups increases.

### 6. Data Example

We now consider a data example from a sleep deprivation study (Balkin et al. 2000; Belenky et al. 2003). This study evaluates 18 commercial vehicle drivers on a series of tests after 0, 1, 2, ..., 9 nights of restriction to 3 hr of sleep. On each day, subjects take a series of reaction time tests, and the experimenters record each subject’s average reaction time. The data are available in the sleepstudy dataset of R’s lme4 package (Bates et al. 2015).

We restructure the data to fit regressions that predict average sleep-deprived reaction time (Y) from number of days of sleep deprivation (X₁) and the subject’s baseline (Day 0) average reaction time under their normal sleep amount (X₂). For each individual i, we observe nine triplets (X₁j, X₂j, Yj). For the purpose of this demonstration, we treat each \( (X_{1j}, X_{2j}, Y_j) \) as a random draw from a subject-specific distribution \( P_j \). (Alternatively, we could treat \( X_{1j} \) as fixed, \( X_{2j} \) as random, and \( Y_j \) as a random draw from \( P_j|X \). These methods are valid as long as the nonconformity scores are exchangeable, as discussed below.) The variable \( X_{1j} \) ranges from 1 to 9 days, and the baseline time \( X_{2j} \) is measured once for each subject j. Across subjects, \( X_{2j} \) ranges from 199 to...

---

**Figure 4.** Coverage of conformal methods for a new observation from an observed group. Loess smoothing for visualization. Both methods have approximately nominal coverage.
Figure 5. Average size of conformal sets for a new observation from an observed group. Loess smoothing for visualization. At \( \sigma^2 = 100 \), within-group variance is high relative to between-group variance, and a shrinkage-based conformal method produces smaller sets.

322 ms, and \( Y \) ranges from 194 to 466 ms. Our fitted regression models have the form \( \hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \). We have also considered a model that includes an intercept. This does not make much of a difference when assessing whether the residuals appear to be exchangeable.

Suppose we observe \((X_1, X_2) = x\) on a 19th individual, and we want to predict the associated \( Y \). We construct prediction sets \( C(X_1, X_2; \alpha) \) such that \( \hat{p}(Y \in C(X_1, X_2; \alpha)) \geq 1 - \alpha \). We use the constructions from Section 4, and we use nonconformity scores of \( R_k(x, y) = |Y_i - \hat{Y}_i| \). CDF pooling uses the process justified by Theorem 7. We fit the regression model \( \mu \) on the pooled observations of 9 of the 18 individuals, and we estimate the quantiles from the remaining individuals. Single subsampling randomly selects one observation per individual. We augment the subsample with \((X_1, X_2, y)\) for the observed \((X_1, X_2)\) and some proposed \( y \). We fit the regression model on this augmented sample of size 19. Repeated subsampling averages \( p \)-values across \( B = 100 \) repetitions of single subsampling using the same \((X_1, X_2, y)\). CDF pooling is asymptotically valid \((k \to \infty)\) if the nonconformity scores are exchangeable across all observations used for quantile estimation and if \( Y \) is continuous. The subsampling methods are valid if the nonconformity scores are exchangeable for all subsamples of one observation per subject. From visual inspection (not shown), the CDF pooling exchangeability assumption may not be met. Several subjects have particularly high or particularly low absolute residuals on all of their observations. Also, we have \( k = 18 \) subjects, and CDF pooling is an asymptotic method. The subsampling methods’ exchangeability assumption is reasonable, based on plots of the absolute residuals when we fit and evaluate the model on one observation per subject.

Figure 6 shows the prediction sets and their size at \( \alpha = 0.10 \). The left panel shows the sets at \( X_1 = \{1, 5, 9\} \) and at \( X_2 = \{200, 230, 260, 290, 320\} \). For most \((X_1, X_2)\) combinations, all three sets have similar centers. Interestingly, the Day 1/5 prediction sets and most Day 9 prediction sets contain the Day 0 reaction time. This suggests it is plausible to maintain the baseline reaction time despite sleep deprivation. The right panel compares the length of the three sets over an expanded set of \((X_1, X_2)\) combinations. CDF pooling produces the smallest sets in most cases, but this method is only asymptotically valid \((k \to \infty)\). Repeated subsampling produces smaller sets than single subsampling in about half of the cases, and it has the least variation in set lengths across \( X_2 \) for a given \( X_1 \).

We also explore the coverage of these methods. CDF pooling is only asymptotically valid \((k \to \infty)\) at level \( 1 - \alpha \), single subsampling is valid at level \( 1 - \alpha \) but has more variance, and repeated subsampling only has guaranteed coverage at level \( 1 - 2\alpha \). We evaluate coverage by holding out 1 of the 18 individuals, selecting a triplet \((X_1, X_2, Y)\) from the held-out individual, fitting a prediction set \( C \) on the remaining 17 subjects, and checking whether \( Y \in C(X_1, X_2; \alpha) \). We perform this procedure \( 18 \times 9 = 162 \) times, using each observation as the test \((X_1, X_2, Y)\) once. The proportion of simulations in which \( Y \in C(X_1, X_2; \alpha) \) is an estimate of the coverage. CDF pooling has algorithmic randomness in the individuals selected for model fitting (eight individuals) versus quantile estimation (nine individuals). The subsampling methods have algorithmic randomness in the observations selected for each subsample. Thus, we repeat this coverage estimation procedure 1000 times.

Table 1 shows the coverage proportions at \( \alpha \in \{0.10, 0.15, 0.20\} \). For each method, Table 1 displays the average coverage, the 2.5th percentile, and the 97.5th percentile over 1000 simulations. On average, CDF pooling undercovers by about 0.02 to 0.03, and the subsampling methods overcover by about 0.03 to 0.06. Compared to single subsampling, repeated subsampling has slightly higher coverage but lower variation in coverage. Overall, repeated subsampling is the best choice in this setting. This method achieves coverage of at least \( 1 - \alpha \) and has lower variation in set size and coverage than the other two methods.

| Method                      | \( 1 - \alpha = 0.90 \) | \( 1 - \alpha = 0.85 \) | \( 1 - \alpha = 0.80 \) |
|-----------------------------|-------------------------|-------------------------|-------------------------|
| 1. CDF pooling              | 0.87 (0.84, 0.90)       | 0.83 (0.80, 0.86)       | 0.78 (0.75, 0.81)       |
| 2. Subsample once           | 0.94 (0.92, 0.97)       | 0.89 (0.86, 0.92)       | 0.83 (0.80, 0.87)       |
| 3. Repeated subsample       | 0.95 (0.94, 0.96)       | 0.91 (0.90, 0.92)       | 0.84 (0.83, 0.85)       |
Figure 6. Prediction sets for sleep-deprived reaction time given baseline reaction time and days of sleep deprivation. All sets have similar centers. CDF pooling has the smallest sets. Repeated subsampling has less variation in set size than single subsampling.

7. Conclusion

We have proposed and compared several methods for constructing distribution-free prediction sets for two-layer hierarchical models. We believe these are the first such methods. We consider a CDF pooling method that is asymptotically valid as \( k \to \infty \), a single subsample method that uses one observation per group, and a repeated subsample method that repeatedly selects one observation per group and averages \( p \)-values over subsamples. The single subsample method is valid at level \( 1 - \alpha \). The repeated subsample method has guaranteed coverage at level \( 1 - 2\alpha \) but tends to have coverage of at least \( 1 - \alpha \) in practice.

Based on our simulations and data example, we recommend CDF pooling if asymptotic coverage is acceptable and the outcome is continuous. CDF pooling typically has the smallest prediction sets, and it yields approximately nominal coverage, especially for large \( k \) and balanced groups. For small to moderate \( k \) or for unbalanced groups, CDF pooling may undercover. If we desire finite sample coverage guarantees, we recommend repeated subsampling. While this method guarantees \( 1 - 2\alpha \) coverage, it achieves coverage close to \( 1 - \alpha \) in simulations. In the sleep example, this method has coverage of at least \( 1 - \alpha \) and has more stable size and coverage than the other methods. Single subsampling is valid at level \( 1 - \alpha \) but ignores most of the data. Repeated subsampling has less algorithmic variation than single subsampling, which makes this method more stable and more reproducible. It is a curiosity that single subsampling often produces slightly smaller prediction sets than repeated subsampling. The asymptotic efficiency of these methods relative to an oracle model remains an open question. In fact, characterizing the asymptotic efficiency of conformal methods is an open question in conformal research more broadly, with some results under additional assumptions in Lei and Wasserman (2014).

The main focus of this article has been the prediction of a new observation on a new subject. In the unsupervised setting, we also considered prediction of a future observation on an existing subject. Future work may consider alternatives to the James–Stein shrinkage residual or may incorporate repeated subsampling into the shrinkage approach. In addition, supervised conformal methods that borrow strength across subjects to construct prediction sets for an existing subject remain an open problem. Space does not permit a thorough investigation of these problems, but we hope to report more on them in a future article.

Supplementary Materials

Appendix A Mathematical Details: This appendix (PDF file) provides proofs of Theorems 1, 3, 4, 6, 7, and 8. Theorems 2, 5, 9, 10, 11, and 12 are justified in the in-text citations.

Appendix B Additional Simulations: This appendix (PDF file) provides simulations for additional values of \( n_j \) (supervised and unsupervised),
nonnormal data (unsupervised), and additional \((\mu, \tau^2)\) parameters (supervised).

**Acknowledgments**

This work made extensive use of the R statistical software \(\text{(R Core Team 2021)}\), as well as the \texttt{data.table} \(\text{(Dowle and Srinivasan 2021)}\), \texttt{formula.tools} \(\text{(Bown 2018)}\), \texttt{gridExtra} \(\text{(Auguie 2017)}\), \texttt{lme4} \(\text{(Bates et al. 2015)}\), \texttt{progress} \(\text{(Csárdi and Fitzjohn 2019)}\), \texttt{R.utils} \(\text{(Bengtsson et al. 2019)}\), and \texttt{tidyverse} \(\text{(Wickham et al. 2019)}\) packages. The authors thank the associate editor and reviewers for helpful feedback that has greatly improved the quality of the paper. The authors also thank Jing Lei and Mauricio Sadinle for helpful discussions.

**Funding**

This work was conducted while RD was at Carnegie Mellon University. RD’s research was supported by the National Science Foundation Graduate Research Fellowship Program under grant nos. DGE 1252522 and DGE 1745016. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work used the Extreme Science and Engineering Discovery Environment \(\text{(XSEDE)}\) \(\text{(Town et al. 2014)}\), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system \(\text{(Nystrom et al. 2015)}\), which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center \(\text{(PSC)}\).

**ORCID**

Robin Dunn \texttt{http://orcid.org/0000-0003-1874-5601}
Larry Wasserman \texttt{http://orcid.org/0000-0003-1874-5601}
Aaditya Ramdas \texttt{http://orcid.org/0000-0003-0497-311X}

**References**

Auguie, B. (2017), \texttt{gridExtra}: Miscellaneous Functions for “Grid” Graphics, R package version 2.3. \[2502\]
Balkin, T., Thome, D., Sing, H., Thomas, M., Redmond, D., Wesensten, N., Williams, J., Hall, S., and Belenky, G. (2000), “Effects of Sleep Schedules on Commercial Motor Vehicle Driver Performance,” Technical report, United States. Department of Transportation. Federal Motor Carrier Safety Administration. \[2499\]
Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021a), “Predictive Inference with the Jackknife+,” \textit{The Annals of Statistics}, 49, 486–507. \[2494\]
——— (2021b), “The Limits of Distribution-Free Conditional Predictive Inference,” \textit{Information and Inference: A Journal of the IMA}, 10, 455–482. \[2492\]
Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015), “Fitting Linear Mixed-Effects Models Using \texttt{lme4},” \textit{Journal of Statistical Software}, 67, 1–48. \[2499,2502\]
Belenky, G., Wesensten, N. J., Thorne, D. R., Thomas, M. L., Sing, H. C., Redmond, D. P., Russo, M. B., and Balkin, T. J. (2003), “Patterns of Performance Degradation and Restoration during Sleep Restriction and Subsequent Recovery: A Sleep Dose-Response Study,” \textit{Journal of Sleep Research}, 12, 1–12. \[2499\]
Bengtsson, H. (2021), \texttt{R.utils}: Various Programming Utilities, R package version 2.11.0. \[2502\]
Booth, J. G., and Hobert, J. P. (1998), “Standard Errors of Prediction in Generalized Linear Mixed Models,” \textit{Journal of the American Statistical Association}, 93, 262–272. \[2492\]
Brown, C. (2018), \texttt{formula.tools}: Programmatic Utilities for Manipulating Formulas, Expressions, Calls, Assignments and Other R Objects, R package version 1.7.1. \[2502\]