SYSTEMS PLANNING AND ANALYSIS, INC.

Distribution Free Uncertainty for CERs

William King

Shaun Irvin

May 4th 2023

Conformal Prediction

- **Conformal prediction is a technique that generates prediction intervals with rigorous statistical coverage guarantees and without distributional assumptions**
  - Applies to any machine learning algorithm or “black-box” model
  - Applies to both regression and classification problems
  - Only assumes the exchangeability of data (a weaker assumption than independence)
- **For regression, the basic idea with Conformal Prediction is to**
  1. Use a previously trained model to predict unseen calibration data
  2. Find the quantile of the calibration residuals corresponding to your level of significance ($\hat{q}_\alpha$)
  3. Apply that residual quantile to generate intervals around new predictions: ($\hat{y} \pm \hat{q}_\alpha$)
- **Remarkably, this simple procedure yields statistical coverage guarantees given the exchangeability of the underlying data**

Conformal Intuition

- **Suppose we have a model, $\hat{f}(X_i)$, and define residuals $\varepsilon_i = \left|Y_i - \hat{f}(X_i)\right|$**
  - Further, suppose we know residuals $\varepsilon_1, \ldots, \varepsilon_4$ where $\varepsilon_1 \leq \varepsilon_2 \leq \varepsilon_3 \leq \varepsilon_4$
  - Now consider a new residual ε_5: if the ε_i are independent and identically distributed (i.i.d), what's the probability of ε_5 falling between ε_2 and ε_3?
    - With the assumption of i.i.d, it's equally likely for ε_5 to fall within any interval:

$$20\% \quad 20\% \quad 20\% \quad 20\% \quad 20\%$$

$$\varepsilon_1 \qquad \varepsilon_2 \qquad \varepsilon_3 \qquad \varepsilon_4$$

    - For example, $P(\varepsilon_5 \leq \varepsilon_3) = 60\%$ which implies that $P\left(\left|Y_5 - \hat{f}(X_5)\right| \leq \varepsilon_3\right) = 60\%$ which implies that $P\left(Y_5 \in \hat{f}(X_5) \pm \varepsilon_3\right) = 60\%$
      - Thus, a 60% prediction interval for Y_5 is the interval $\left[\hat{f}(X_5) - \varepsilon_3, \hat{f}(X_5) + \varepsilon_3\right]$
      - Similarly, an 80% prediction interval for Y_5 is the interval $\left[\hat{f}(X_5) - \varepsilon_4, \hat{f}(X_5) + \varepsilon_4\right]$
    - In general terms, we can utilize the appropriate quantile of the residuals to formulate a prediction interval
  - A natural question arises in how we obtain residuals ε_i that realistically represents how $\hat{f}(X_i)$ will perform on new, unseen data (especially since the training residuals will tend to be artificially small due to overfitting)

How do we Quantify Uncertainty of ML Models?

- **A "naïve" approach involves calculating Prediction Intervals using the residuals on the training data, $\left|\boldsymbol{Y_i} - \hat{\boldsymbol{f}}(\boldsymbol{X_i})\right|$**
  - $\hat{f}(X_i) \pm$ the $(1-\alpha)$ quantile of $\left|Y_1 - \hat{f}(X_1)\right|, \dots, \left|Y_n - \hat{f}(X_n)\right|$
  - Leads to artificially narrow prediction intervals with overfitting (performs well on trained data, but not new data)
  - Does **Not** account for variability of residuals across the input space
  - Does **Not** guarantee predefined coverage (later methods address this)
- **"Leave-One-Out" cross-validation for Prediction Intervals (Jackknife) using residuals of the held-out test point, $\left|\boldsymbol{Y_i} - \hat{\boldsymbol{f}}_{-\boldsymbol{i}}(\boldsymbol{X_i})\right|$**
  - $\hat{f}(X_i) \pm$ the $(1-\alpha)$ quantile of $\left|Y_1 - \hat{f}_{-1}(X_1)\right|, \dots, \left|Y_n - \hat{f}_{-n}(X_n)\right|$
  - Leads to slightly wider prediction intervals that are more robust than the naïve approach to overfitting
  - Does **Not** account for variability of residuals across the input space
  - Does **Not** guarantee predefined coverage (later methods address this)

Conformal Variants

- **Full Conformal Prediction**
  - PI: $\{y : |y - \hat{f}_y(x_{n+1})| \leq Q_{1-\alpha}(R_1, \ldots, R_n, R_{n+1})\}$
    - Where $\hat{f}_y$ is the model trained as if (x_{n+1}, y) were a new data point, $R_i = |y_i - \hat{f}_y(x_i)|$ and $Q_{1-\alpha}$ is the $1-\alpha$ quantile of the residuals
    - Does not require a calibration dataset, but requires re-fitting the model for every possible value of y whenever a new prediction is made
      - Since this is infeasible in practice, usually a finite grid of y-values are selected and evaluated, but this can be very computationally expensive even with small datasets
- **Split Conformal Prediction**
  - Partition data into training (size m) and calibration (size $n-m$) sets:
  - PI: $\hat{f}_{train}(x_{n+1}) \pm Q_{1-\alpha}\left(R_1^C, \ldots, R_{n-m}^C\right)$
    - Where $\hat{f}_{train}$ is the model trained on the m training data points, $R_i^C = |y_i - \hat{f}_{train}(x_i)|\ \forall\ i$ in the calibration set, and $Q_{1-\alpha}$ is defined as above
    - Requires sacrificing data to the calibration set, but only needs to be fit once
      - Calibration data can be hard to come by (≈1000 calibration data points are needed to achieve coverage between 88-92% at a 90% confidence level)

Conformal Variants

- **CV+ for K-fold cross-validation**
  - Partition data into K non-overlapping subsets: $S_1, \dots, S_k$
  - PI: $\left[Q_\alpha\left(\hat{f}_{-S_{k(i)}}(x_{n+1}) - R_i^{CV}\right), Q_{1-\alpha}\left(\hat{f}_{-S_{k(i)}}(x_{n+1}) + R_i^{CV}\right)\right]$
    - Where $\hat{f}_{-S_{k(i)}}$ is the model trained with the k-th subset removed, $k(i)$ indicates the subset that includes the i-th data point, $R_i^{CV} = \left|y_i - \hat{f}_{-S_{k(i)}}(x_i)\right|$ is the absolute value of the out-of-fold residual, and Q_α is defined as before
    - Does not require a separate calibration data set and only requires fitting subsets of the data K times
      - The out-of-fold residuals stand in proxy for the calibration dataset, since they are unseen at the time each model is trained during cross-validation
      - If you are already performing cross-validation, then you are already training these models and calculating their out-of-fold residuals
        - » The only extra things you need to do is to save each $\hat{f}_{-S_{k(i)}}$ model and the association of out-of-fold residuals to subsets $k(i)$
    - Note: CV+ where $K = n$ is called the Jackknife+ (a form of Leave-One-Out cross-validation)

Conformal Variants

- **CV+ for K-fold cross-validation (Example)**
  - Example of 2-fold cross-validation with 4 data points

Training Data

i	S_k	$k(i)$	$\hat{f}_{-S_{k(i)}}$	R_i^{CV}
1	S_1	1	$\hat{f}_{-S_1}$	$\lvert y_1 - \hat{f}_{-S_1}(x_1)\rvert$
2		1	$\hat{f}_{-S_1}$	$\lvert y_2 - \hat{f}_{-S_1}(x_2)\rvert$
3	S_2	2	$\hat{f}_{-S_2}$	$\lvert y_3 - \hat{f}_{-S_2}(x_3)\rvert$
4		2	$\hat{f}_{-S_2}$	$\lvert y_4 - \hat{f}_{-S_2}(x_4)\rvert$

$\hat{f}_{-S_1}$ is trained on data points 3 & 4

$\hat{f}_{-S_2}$ is trained on data points 1 & 2

Predicting New Data Point

i	$\hat{f}_{-S_{k(i)}}$	Low	$High$
n+1	$\hat{f}_{-S_1}(x_{n+1})$	$\hat{f}_{-S_1}(x_{n+1}) - R_1^{CV}$	$\hat{f}_{-S_1}(x_{n+1}) + R_1^{CV}$
		$\hat{f}_{-S_1}(x_{n+1}) - R_2^{CV}$	$\hat{f}_{-S_1}(x_{n+1}) + R_2^{CV}$
	$\hat{f}_{-S_2}(x_{n+1})$	$\hat{f}_{-S_2}(x_{n+1}) - R_3^{CV}$	$\hat{f}_{-S_2}(x_{n+1}) + R_3^{CV}$
		$\hat{f}_{-S_2}(x_{n+1}) - R_4^{CV}$	$\hat{f}_{-S_2}(x_{n+1}) + R_4^{CV}$
		α quantile of these 4 values is the low bound of the PI	$1-\alpha$ quantile of these 4 values is the high bound of the PI

  - Note: there is a R_i^{CV} residual for each data point (even though there are fewer models than data points)
  - To implement prediction intervals for new predictions, we just need to save each of the sub-models $\hat{f}_{-S_{k(i)}}$ and the CV+ residuals R_i^{CV}

Conformal Variants

- **Conformal Variant Comparison:**

Variant	Training Cost	Calibration Data	Coverage Guarantee	Empirical Coverage	Notes
Full	∞	No	$\geq 1 - \alpha$	$\approx 1 - \alpha$	Infeasible even with small datasets
Split	1	Yes	$\geq 1 - \alpha$	$\approx 1 - \alpha$	Good when you have lots of calibration data or a computationally expensive model; stronger statistical guarantees than CV+
K-fold CV+	K	No	$\geq 1 - 2\alpha$	$\gtrapprox 1 - \alpha$	Good when you have less data, or a very complex model; substantially computationally cheaper than Full conformal, but more costly than split conformal

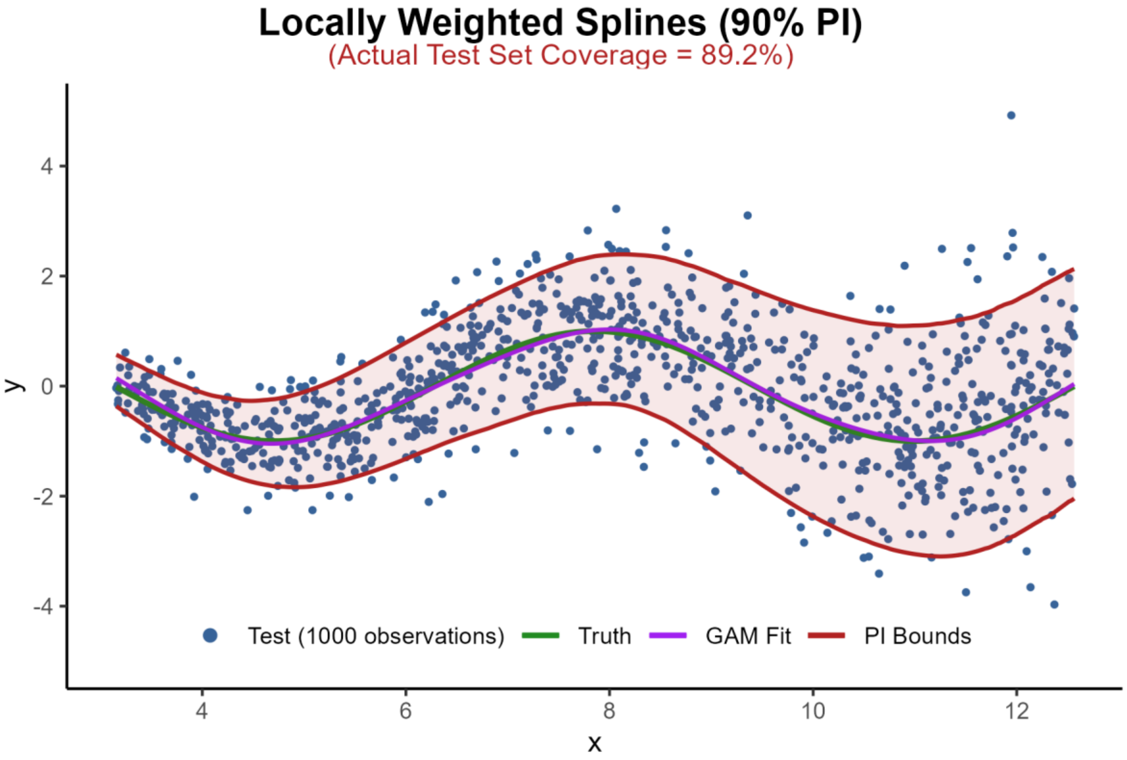
  - K-Fold CV+ offers a balance between the computational cost of Full Conformal and the calibration data size requirements for Split Conformal
  - If you’re already performing cross-validation, CV+ is computationally free (you just need to save the sub-models and residuals you are already calculating)

Locally Weighted CV+

- **The conformal variants previously discussed tend to generate prediction intervals with constant width**
- **This makes sense with additive errors, but not with the multiplicative errors we tend to see with cost data**
- **Luckily, conformal prediction works with any non-conformity measure**
  - Previously we used the absolute value of the calibration residuals as the non-conformity measure
  - Scaling the absolute value of the residuals by an estimate of the residual spread is still a valid non-conformity measure
    - Before, we defined $R_i^{CV} = \left|y_i - \hat{f}_{-S_{k(i)}}(x_i)\right|$, now we consider $R_i^{LW} = \frac{\left|y_i - \hat{f}_{-S_{k(i)}}(x_i)\right|}{\hat{\rho}_{-S_{k(i)}}(x_i)}$
      - Where $\hat{\rho}_{-S_{k(i)}}(x_i)$ is the estimate of the conditional mean absolute deviation of the residuals from $\hat{f}_{-S_{k(i)}}$ (note this involves fitting two models at each step of cross-validation)
  - $PI: \left[Q_\alpha\left(\hat{f}_{-S_{k(i)}}(x_{n+1}) - R_i^{CV} * \hat{\rho}_{-S_{k(i)}}(x_i)\right), Q_{1-\alpha}\left(\hat{f}_{-S_{k(i)}}(x_{n+1}) + R_i^{CV} * \hat{\rho}_{-S_{k(i)}}(x_i)\right)\right]$

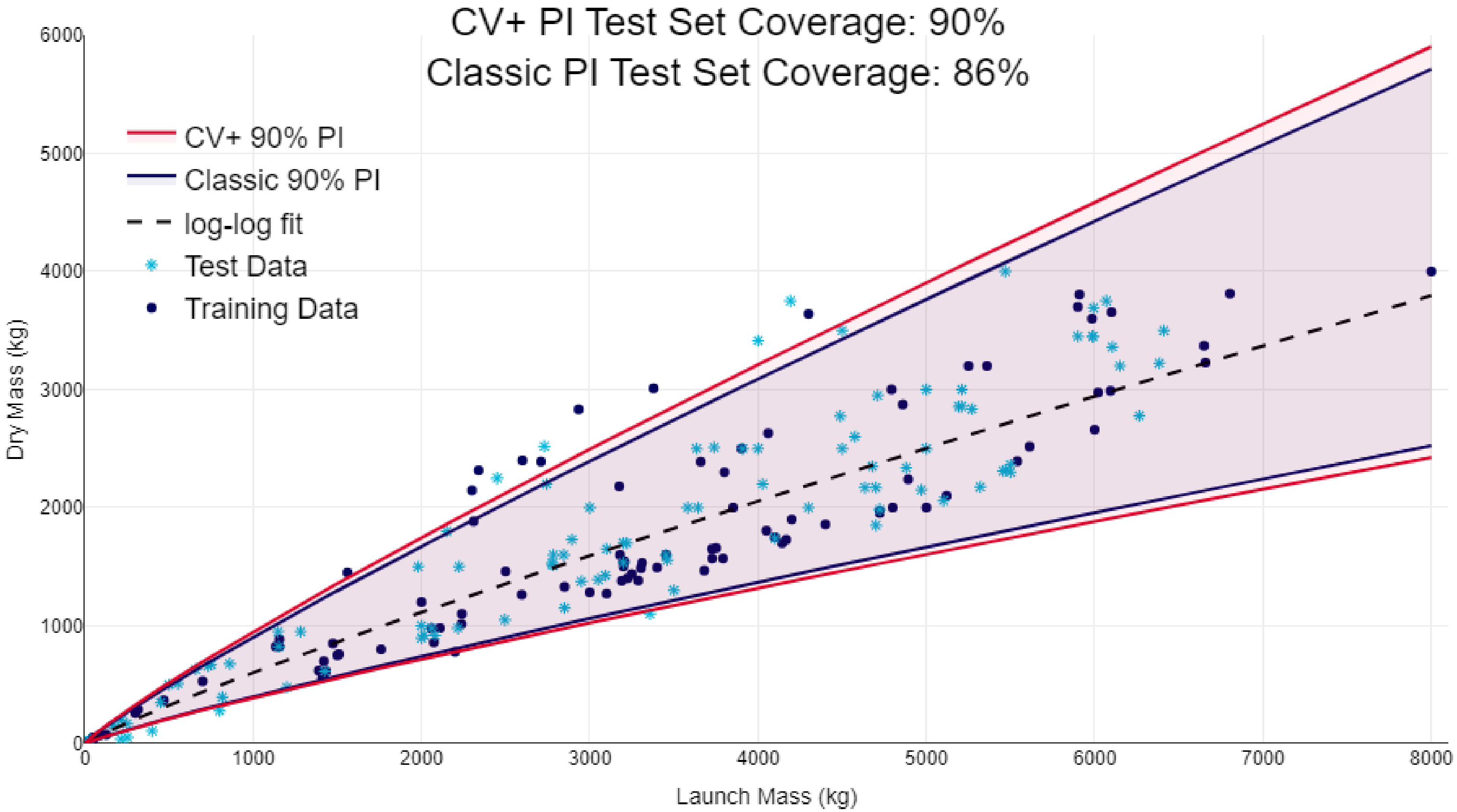
Applications to Regression

- **Create datasets to train/calibrate a ML model using CV+ method**
- **Fit a RF model on training data and plot predictions of test data**
- **Use CV+ method to determine 90% Prediction Interval bounds**
- **Explore other CV+ variants**
  - Locally Weighted
  - Generalized Additive Model (GAM) using splines



Applications to Regression

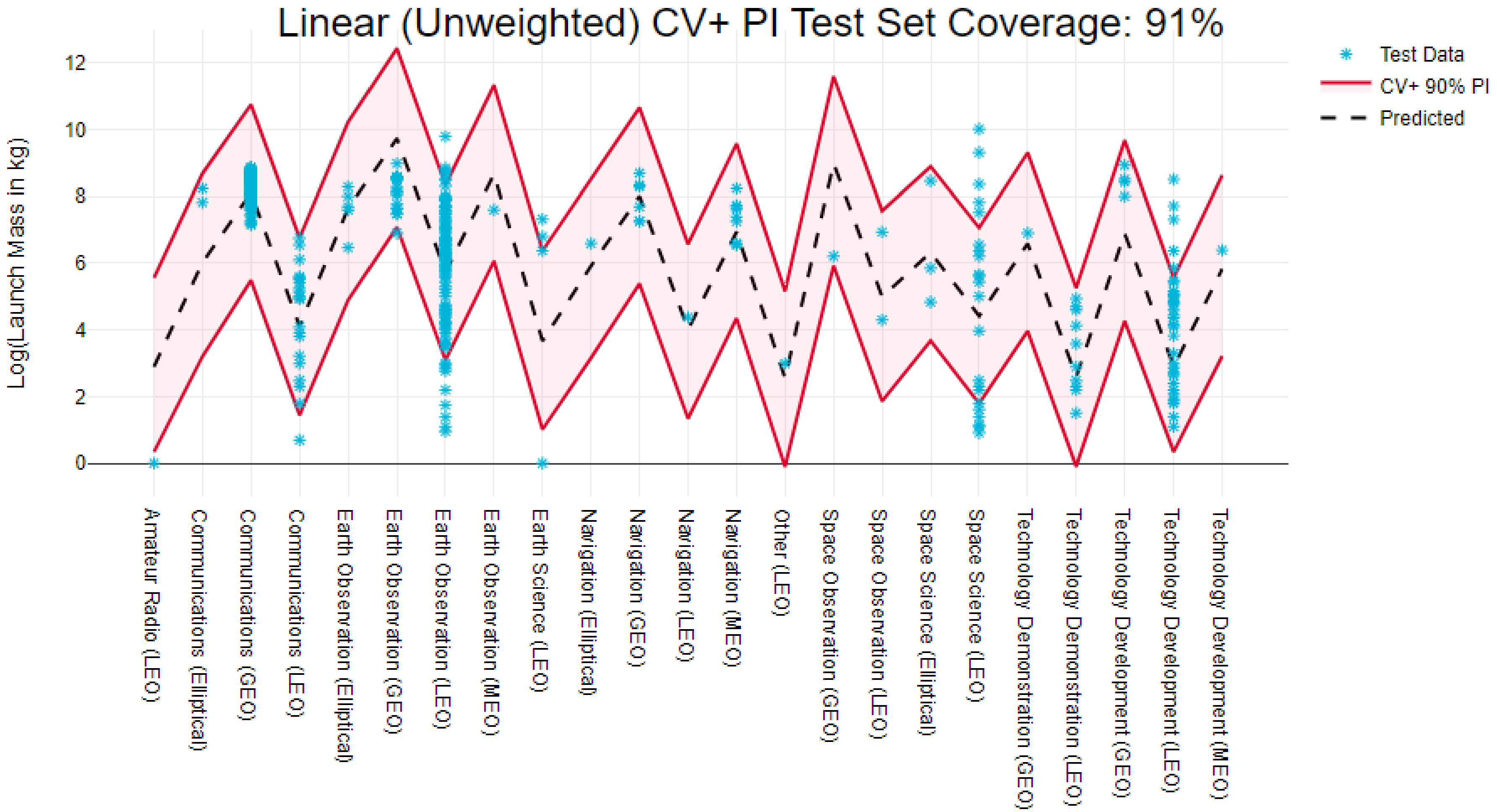
- **Predicting SV Dry Mass from SV Launch Mass**
  - Data split approximately in half (½ for Training, ½ to evaluate coverage)
  - Log-log linear fit to the training data



Source: UCS Satellite Database (5/1/2022)

Applications to Regression

- **Predicting SV Launch Mass from Mission and Orbit**



Source: UCS Satellite Database (5/1/2022)

Conclusions

- **Conformal prediction enables distribution free uncertainty with for any machine learning algorithm**
  - Only requirement is the exchangeability of the data (a weaker form of the i.i.d. assumption we make with classical approaches)
  - We get a rigorous statistical coverage guarantee regardless of how well the underlying model fits the data
  - As we embrace more accurate regression techniques that are less interpretable than classical techniques, we don't have to sacrifice predictive uncertainty
- **CV+ is a conformal technique that balances computational cost with the need for lots of calibration data**
  - If you're already performing cross-validation, CV+ is computationally free
  - CV+ offers guaranteed coverage of at least $1 - 2 * \alpha$ with empirical coverage often close to $1 - \alpha$ (examples we've shown have had coverage between $89 - 93\%$ with $\alpha = 10\%$)

Future Research

- **Hierarchical classification for WBS normalization**
  - In a classification setting, conformal prediction produces prediction sets, that are guaranteed to contain the true label with some measure of statistical certainty (where larger prediction sets indicate more uncertainty)
  - Applying conformal prediction to hierarchical classification for WBS normalization can direct human intervention to elements with large prediction sets (i.e., where there the algorithm is highly uncertain)
- **Many packages in R and Python to facilitate conformal prediction**
  - MAPIE (Model Agnostic Prediction Interval Estimator) for Python
  - conformalInference for R (not available on CRAN)
    - Note: We had to write our own wrappers for CaReT to perform CV+ and its locally weighted variant

Backup

Contact Information

- **William King**
  - Email: wking@spa.com
- **Shaun Irvin**
  - Email: sirvin@spa.com

References

- *Predictive Inference with the Jackknife+*
  - Barber, Candes, Ramdas and Tibshirani, 2021, The Annuals of Statistics
- *Distribution-Free Predictive Inference for Regression*
  - Lei, G’Sell, Rinaldo, Tibshirani and Wasserman, 2018, Journal of the American Statistical Association
- *A Tutorial on Conformal Prediction*
  - Shafer and Vovk, 2008, Journal of Machine Learning Research
- *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*
  - Angelopoulos and Bates, 2022, arXiv:2107.07511v6