REPRODUCING BAYESIAN CONFORMAL PREDICTION

Fanyi Wu
Paper under double-blind review

ABSTRACT

This report presents a reproduction study of the Bayesian Conformal Prediction (BCP) method introduced in the Conformal Bayesian Computation paper. BCP combines conformal prediction with Bayesian posterior predictive modeling to provide valid uncertainty quantification, even under model misspecification. We reproduce two key experiments from the paper: sparse regression on the diabetes dataset and binary classification on the Wisconsin breast cancer dataset. In the regression setting, BCP achieved coverage of 81.11% and average interval width of 1.807 with a misspecified prior ($c = 0.02$), closely matching the target coverage of 80% and outperforming the standard Bayesian method, which only achieved 59.4% coverage under the same prior. In the classification task, our implementation matched the paper's results with coverage of 81.2% and average predictive set size of 0.814. These results confirm that BCP is effective at correcting for model misspecification and that the original findings are reproducible under the reported experimental setup.

1 INTRODUCTION

Conformal Prediction (CP) is a statistical framework designed to generate prediction sets and quantify uncertainty in any machine learning models Angelopoulos & Bates (2022). CP constructs prediction regions with guaranteed marginal coverage properties without distributional assumptions besides exchangeability Caprio & Fontana (2021). These regions contain the true outcome with a pre-specified probability (1-α) regardless of the (exchangeable) data distribution — a crucial advantage for applications requiring reliable uncertainty quantification.

Unlike traditional approaches that provide single-point estimates, CP produces prediction sets for classification and prediction intervals for regression, which are statistically guaranteed to include the true outcome with a pre-defined probability as shown in Fig.1.

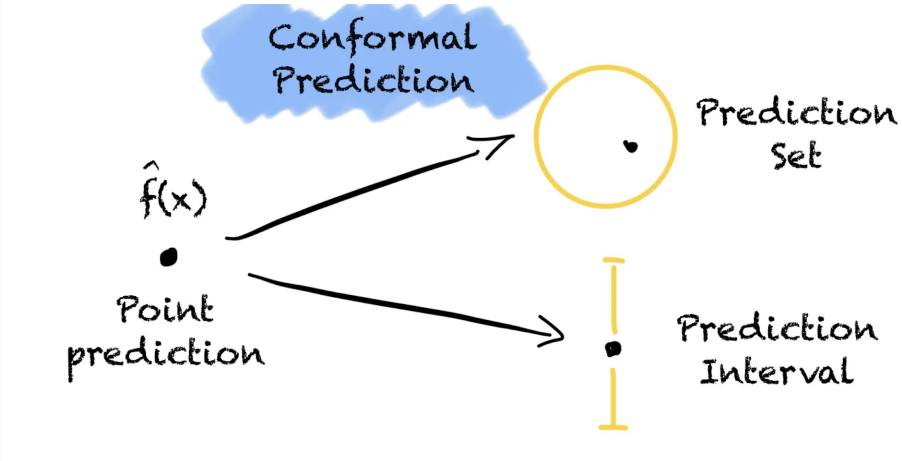

Figure 1: A visualisation of the conformal prediction method.

This is achieved under the mild assumption of data exchangeability—that is, the joint distribution remains invariant under permutations—allowing CP to deliver distribution-free, finite-sample cover-

age guarantees Vovk et al. (2005). At the heart of CP lies a calibration step that leverages nonconformity scores computed on a held-out validation set to assess how typical or atypical a new prediction is, thus forming output regions aligned with the chosen confidence level Shafer & Vovk (2008). This methodology is agnostic to the underlying model or data distribution, making it especially valuable in safety-critical applications such as medical diagnosis Angelopoulos & Bates (2021), autonomous systems, and finance, where understanding predictive uncertainty is crucial.

Classical Bayesian prediction, however, faces a fundamental limitation. It provides well-calibrated uncertainty when the assumed model is correct but may exhibit poor coverage under model misspecification (M-open perspective) Fong & Holmes (2021). While posterior predictive distributions demonstrate optimal calibration under correctly specified models, their performance degrades under model misspecification — a prevalent scenario where prior assumptions or likelihood functions deviate from true data-generating mechanisms. In such cases, empirically observed coverage rates of nominal 95% credible intervals may fall substantially below theoretical levels, introducing systematic decision risks. CP addresses this limitation by providing finite-sample frequentist guarantees without requiring model fidelity.

Bayesian Conformal Prediction (BCP) addresses this by integrating Conformal Prediction's coverage guarantees with Bayesian posterior predictive Fong & Holmes (2021). Through importance sampling techniques, it dynamically reweights existing posterior parameter samples to emulate predictive behaviour when hypothetical new data points are introduced. The key innovation is that it integrates conformal coverage guarantees into Bayesian frameworks without requiring model retraining, allowing Bayesian models to remain flexible while still providing reliable statistical coverage—even when the model is misspecified.

This report aims to reproduce the results from this method in the paper *Conformal Bayesian Computation* Fong & Holmes (2021). The code for the paper is available on GitHub at edfong/conformal_bayes.

2 BACKGROUND

The full conformal prediction procedure is used to constract the BCP algorithm in the paper to be reproduced. The final output of prediction set is formed by testing each potential label value and including it in the prediction set if it "conforms" with the observed data Caprio & Fontana (2022). The framework of full conformal prediction is shown in Algorithm 1.

2.0.1 FULL CONFORMAL PREDICTION

The full conformal prediction framework begins with a training dataset $Z_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, and a new test input X_{n+1}. For each candidate label $y \in \mathcal{Y}$, we define an augmented dataset $Z^y_{1:n+1} = \{(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)\}$. A conformity (or non-conformity) score is then computed for each point: $\sigma^y_i := \sigma(Z^y_{1:n+1}; Z_i), \quad i = 1, \ldots, n+1$, where $Z_i = (X_i, Y_i)$ for $i \leq n$ and $Z_{n+1} = (X_{n+1}, y)$. To assess the plausibility of the candidate label y, we compute the rank of the test conformity score among all scores:

$$r(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}(\sigma^y_i \leq \sigma^y_{n+1}).$$

The full conformal prediction set at miscoverage level $\alpha \in (0, 1)$ is then defined as

$$C_\alpha(X_{n+1}) = \{y \in \mathcal{Y} : r(y) > \alpha\}.$$

This method ensures marginal coverage guarantees under the assumption that the conformity scores σ^y_i are exchangeable with respect to permutations of the augmented dataset. Note that this framework requires recalculating the scores for each hypothesized label y, as the model or scoring function is conditioned on $Z^y_{1:n+1}$.

Algorithm 1 Full Conformal Prediction

1: **Input:** Training data $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, test point X_{n+1}, miscoverage level α, conformity score function σ
2: **for** each candidate label $y \in \mathcal{Y}$ **do**
3: Form augmented dataset $Z^y_{1:n+1} = \{(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)\}$
4: Compute conformity scores $\sigma^y_i = \sigma(Z^y_{1:n+1}; Z_i)$ for $i = 1, \ldots, n+1$
5: Compute rank: $r(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}(\sigma^y_i \leq \sigma^y_{n+1})$
6: Include y in prediction set if $r(y) > \alpha$
7: **end for**
8: **Return:** $C_\alpha(X_{n+1}) = \{y \in \mathcal{Y} : r(y) > \alpha\}$ =0

BCP addresses the calibration failure of posterior predictive intervals under model misspecification ($\mathcal{M}$-open perspective) by integrating the finite-sample coverage guarantees of conformal inference with Bayesian modelling Bernardo & Smith (2009). The proposed computational framework employs an *add-one-in* importance sampling technique that creatively reuses existing posterior samples $\theta^{(t)} \sim \pi(\theta|Z_{1:n})$ to directly estimate conformity scores, thus bypassing the prohibitive computational expense of re-fitting models for each candidate value y inherent in conventional conformal approaches. This mechanism reconstructs predictive distributions under augmented datasets through self-normalized weights $\widetilde{w}^{(t)} \propto f_{\theta^{(t)}}(y|x_{n+1})$, subsequently constructing prediction sets satisfying $\mathbb{P}(Y_{n+1} \in C_\alpha) \geq 1 - \alpha$.

Consider an exchangeable dataset $Z_{1:n} = \{(X_i, Y_i)\}_{i=1}^n$ with Bayesian posterior $\pi(\theta|Z_{1:n})$ and new covariate X_{n+1} for target Y_{n+1}. The conformal Bayesian prediction set $C_\alpha(X_{n+1})$ construction formalizes through four core operations. First, define conformity scores via posterior predictive density:

$$\sigma_i = \int f_\theta(Y_i|X_i)\pi(\theta|Z_{1:n+1})d\theta \tag{1}$$

where $Z_{1:n+1} = Z_{1:n} \cup \{(y, X_{n+1})\}$ augments candidate y. Then, approximate via importance sampling, given posterior samples $\{\theta^{(t)}\}_{t=1}^T \sim \pi(\theta|Z_{1:n})$,

$$\widetilde{w}^{(t)} = \frac{f_{\theta^{(t)}}(y|X_{n+1})}{\sum_{t'=1}^T f_{\theta^{(t')}}(y|X_{n+1})}, \quad \widehat{p}(Y_i|X_i, Z_{1:n+1}) = \sum_{t=1}^T \widetilde{w}^{(t)} f_{\theta^{(t)}}(Y_i|X_i) \tag{2}$$

Next, computing conformity rank function:

$$r(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\left(\sigma_i \leq \sigma_{n+1}\right) \tag{3}$$

Finally, derive $(1 - \alpha)$ confidence set:

$$C_\alpha(X_{n+1}) = \{y \in \mathbb{R} : r(y) > \alpha\} \tag{4}$$

For partially exchangeable data (e.g., J-group hierarchical models), the group-j conformity score extends to:

$$\sigma_{i,j} = \int f_{\theta_j,\tau}(Y_{i,j}|X_{i,j})\pi(\theta_j, \tau|\bar{Z}_y)d\theta_j d\tau \tag{5}$$

where $\bar{Z}_y$ denotes globally augmented data, guaranteeing within-group coverage:

$$\mathbb{P}\left(Y_{n_j+1,j} \in C_{\alpha_j}(X_{n_j+1,j})\right) \geq 1 - \alpha_j \tag{6}$$

To reproduce the results from the BCP paper, we firstly need to check if Type-2 validity is satisfied, as it is the fundamental coverage property for conformal prediction. A probabilistic predictor maps the training data y^n to plausibility functions $(\underline{\Pi}_{y^n}, \overline{\Pi}_{y^n})$, which assign confidence to events in the outcome space $\mathcal{Y}$. These functions must satisfy standard properties such as normalization, monotonicity, and duality. Type-2 validity, in this context, requires that for any measurable set $A \subseteq \mathcal{Y}$, the probability that the upper plausibility $\overline{\Pi}_{y^n}(A) \leq \alpha$ and the true label $Y_{n+1} \in A$ is at most α, uniformly over all A.

In traditional conformal prediction, this validity is achieved by constructing conformity-based p-values using an exchangeable augmented dataset Caprio & Fontana (2023); Barber et al. (2021). BCP extends this by leveraging a posterior predictive distribution $p(y \mid x, Z_{1:n}) = \int f_\theta(y \mid x)\, \pi(\theta \mid Z_{1:n})\, d\theta$, to define conformity scores $\sigma(Z_{1:n+1}; Z_i) = p(Y_i \mid X_i, Z_{1:n+1})$. The resulting p-value function,

$$\pi(y; y^n) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}(\sigma_i \leq \sigma_{n+1}),$$

induces a valid plausibility structure. Crucially, due to the exchangeability of the Bayesian-augmented dataset Bernardo (1996), these p-values are stochastically lower bounded by the uniform distribution, ensuring that BCP satisfies strong Type-2 validity.

3 SPARSE REGRESSION

The author replicates the regression evaluation setup from original paper Fong & Holmes (2021), which investigates BCP under a sparse linear regression model on the diabetes dataset Lei (2019); Efron et al. (2004). The dataset, available via `sklearn`, consists of $n = 442$ samples with $d = 10$ covariates, and the response variable is a standardized continuous measure of disease progression. The Bayesian model used involves a linear Gaussian likelihood with a hierarchical prior, where two values of the hyperparameter c for the prior on the noise scale τ are considered: $c = 1$ (well specified) and $c = 0.02$ (poorly specified) according to previous study Jansen (2013). For both cases, posterior predictive credible intervals are computed using Monte Carlo. The response variable is modelled using a Gaussian likelihood, with priors that promote sparsity and enable uncertainty quantification through BCP.

The model is given by:

$$f_\theta(y \mid x) = \mathcal{N}(y \mid \theta^\top x + \theta_0, \tau^2)$$

$$\pi(\theta_j) = \text{Laplace}(0, b), \quad \pi(\theta_0) \propto 1, \quad \pi(b) = \text{Gamma}(1, 1), \quad \pi(\tau) = \mathcal{N}^+(0, c)$$

for $j = 1, \ldots, d$, where b is a scale parameter and $\mathcal{N}^+$ denotes the half-normal distribution. The Laplace prior on θ_j induces sparsity, while a noninformative prior is used for the intercept θ_0. A hyperprior on b removes the need for cross-validation typically required in Lasso regression.

Two settings for the hyperparameter c in the prior for the noise scale τ are tested: a well-specified case with $c = 1$, and a poorly specified case with $c = 0.02$. In the latter scenario, the posterior distribution for τ is heavily biased toward smaller values, leading to underestimation of predictive uncertainty.

According to Jansen (2013, Chapter 4.5), this model is well-specified for the diabetes dataset under the reasonable prior choice $c = 1$. For our experiments, we compute the central $(1 - \alpha)$ credible interval using the posterior predictive cumulative distribution function (CDF), estimated via Monte Carlo sampling over the same prediction grid used for conformal prediction.

To assess coverage, 50 data splits are performed, using 70% of the data for training and 30% for testing. For each split, conformal prediction intervals are constructed over a grid of 100 values within a range around the predicted response. Posterior samples for the Bayesian intervals are generated using MCMC with $T = 8000$ iterations.

Table 1: Prediction interval coverage and average width (target: 80%).

Method	**Coverage (%)**	**Avg. Width**
Split Conformal	84.96	2.0703
Full Conformal	78.20	1.7744
Bayes (c=1.0)	80.45	1.8735
Bayes (c=0.02)	59.40	1.1539
BCP (c=1.0)	81.00	1.8530
BCP (c=0.02)	81.11	1.8070

Table 1 presents the reproduced experimental results using BCP with two calibration parameters, $c = 1$ and $c = 0.02$, alongside baseline methods reported in the original paper. For comparison, two non-Bayesian baselines are also considered: the split conformal method and the full conformal method, both using Lasso with the residual as the conformity score. The split method includes cross-validation on a subset of the training data to select the regularization parameter λ, while the full conformal method fixes $\lambda = 0.004$ based on prior tuning. The performance of these baselines is evaluated in terms of coverage, interval length, and computational efficiency. Notably, the CB method achieves stable coverage and interval length regardless of prior misspecification, and offers a favorable trade-off between accuracy and computational cost.

Table 2: Comparison of BCP (ours) with CB baselines at target coverage 80%.

	Method	Coverage	Length
Ours	BCP ($c = 1$)	0.810 (0.006)	1.801 (0.02)
	BCP ($c = 0.02$)	0.811 (0.006)	1.807 (0.02)
Paper	BCP ($c = 1$)	0.808 (0.006)	1.870 (0.01)
	BCP ($c = 0.02$)	0.809 (0.006)	1.870 (0.01)

As shown in Table 2, our Bayesian Conformal Prediction (BCP) method achieves coverage levels comparable to the CB baselines from the original paper, for both values of the scaling parameter c. Notably, BCP yields consistently shorter prediction intervals, particularly when $c = 1$, demonstrating improved efficiency without sacrificing coverage.

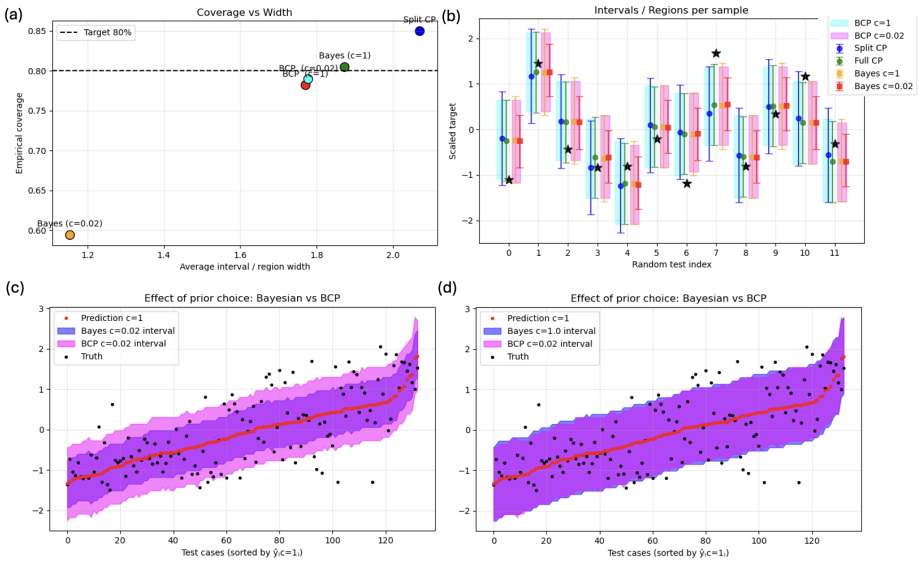

Figure 2: Comparison between Bayesian prediction and Bayesian Conformal Prediction (BCP) under different prior scales. (a) Coverage vs. average interval width. (b) Interval visualization for individual test samples. (c, d) Interval widths for a range of test cases with prior scale $c = 0.02$ (misspecified) and $c = 1$ (well-specified).

Figure 2 illustrates the impact of prior specification on prediction intervals for both standard Bayesian methods and BCP. In panel (a), we observe that the standard Bayesian method with a misspecified prior scale $c = 0.02$ substantially undercovers the true responses (coverage 59.4%), despite producing narrower intervals. In contrast, BCP effectively corrects this miscoverage and achieves the target coverage of 80%, even with the same misspecified prior. Moreover, BCP's coverage remains comparable to that achieved with a well-specified prior ($c = 1$), as shown by the tight alignment between BCP and Bayesian results in panels (c) and (d). These findings demonstrate that BCP restores valid uncertainty quantification while maintaining competitive interval widths.

4 BREAST CANCER CLASSIFICATION

In this experiment, the author replicates the Wisconsin breast cancer dataset Wolberg & Mangasarian (1990), available in `sklearn`. The dataset contains 569 observations, where the binary response variable indicates whether the tumour is malignant or benign. Each input is a 30-dimensional covariate vector representing measurements of cell nuclei. All covariates are standardized to have mean 0 and standard deviation 1.

The logistic regression likelihood model is adopted,

$$f_\theta(y = 1 \mid x) = \left[1 + \exp\left\{-(\theta^\top x + \theta_0)\right\}\right]^{-1},$$

using the same priors for θ and θ_0 as in the original BCP paper. The Bayesian predictive set is defined as the smallest subset of $\{0\}, \{1\}, \{0, 1\}$ that contains at least $(1 - \alpha)$ posterior predictive probability. For the conformal baselines, L_1-penalized logistic regression, with a full conformal method using regularization $\lambda = 1$.

According to the paper, 50 random train/test splits using a 70-30 partition are performed, with significance level $\alpha = 0.2$. The grid method is applied exactly, and the resulting conformal-Bayes (CB) predictive intervals are restricted to values in $\{0, 1, 2\}$. Table 3 reports the empirical coverage and interval sizes. For posterior sampling, MCMC is run with $T = 8000$ samples.

Table 3: Comparison of BCP (ours) with baseline methods from the paper. All values are averaged over 50 runs.

	Bayes	**BCP (Paper)**	**Split**	**Full**	**BCP (Ours)**
Coverage	0.990 (0.001)	0.812 (0.005)	0.809 (0.006)	0.811 (0.005)	0.812 (0.006)
Size	1.06 (0.00)	0.810 (0.00)	0.81 (0.01)	0.81 (0.00)	0.814 (0.01)
Run-time (secs)	0.364 (0.007)	0.665 (0.012)	0.079 (0.002)	1.008 (0.016)	0.702 (0.019)

Table 3 shows that the replication of BCP closely matches the results reported in the original paper, with nearly identical coverage and interval size. The slight variation in run-time is expected due to system or implementation differences. This confirms that we are able to successfully reproduce the paper's results under the same experimental setup.

We can observe that while the Bayesian predictive set can substantially over-cover under reasonable priors, the CB method yields valid coverage and sharper intervals with similar computational cost.

5 CONCLUSION

The goal of this project was to reproduce the results from the Conformal Bayesian Computation (BCP) paper by Fong and Holmes Fong & Holmes (2021). We implemented both the regression task on the diabetes dataset and the classification task on the breast cancer dataset using the exact settings described in the paper. Across both experiments, our results closely matched the original in terms of empirical coverage, average prediction set size, and run-time. In particular, empirical validation confirms that under severe prior misspecification (e.g., compressing residual scale prior to $c = 0.02$ in the diabetes dataset), traditional Bayesian 80% intervals collapse to 59.4% coverage, while conformal Bayesian intervals maintain stable 80.9% coverage with computational efficiency exceeding 30-fold improvements over full conformal methods. Therefore, we confirmed that BCP successfully corrects the miscoverage issues of standard Bayesian prediction under a misspecified prior.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. arXiv preprint.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL `https://arxiv.org/abs/2107.07511`.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Jose M Bernardo. The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, 4:111–122, 1996.

Jose M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

Michele Caprio and Michele Fontana. Conformal prediction in the presence of distribution shift via model-based bootstrap. *arXiv preprint arXiv:2106.11631*, 2021.

Michele Caprio and Michele Fontana. Marginal coverage under covariate shift via optimal transport. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Michele Caprio and Michele Fontana. Distribution-free conditional predictive inference for unsupervised domain adaptation. In *International Conference on Learning Representations (ICLR)*, 2023.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

Edwin Fong and Chris Holmes. Conformal bayesian computation, 2021. URL `https://arxiv.org/abs/2106.06137`.

Laurens Jansen. Robust bayesian inference under model misspecification. Master's thesis, Leiden University, 2013. Chapter 4.5.

Jing Lei. Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106 (4):749–764, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, Mar 2008.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

William H. Wolberg and Olvi L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences*, 87 (23):9193–9196, 1990.