Bayesian Conformal Prediction as a Decision Risk Problem Using Bayesian Quadrature

Fanyi Wu[1], **Veronika Lohmanova**[1], **Samuel Kaski**[2], **Michele Caprio**[1]

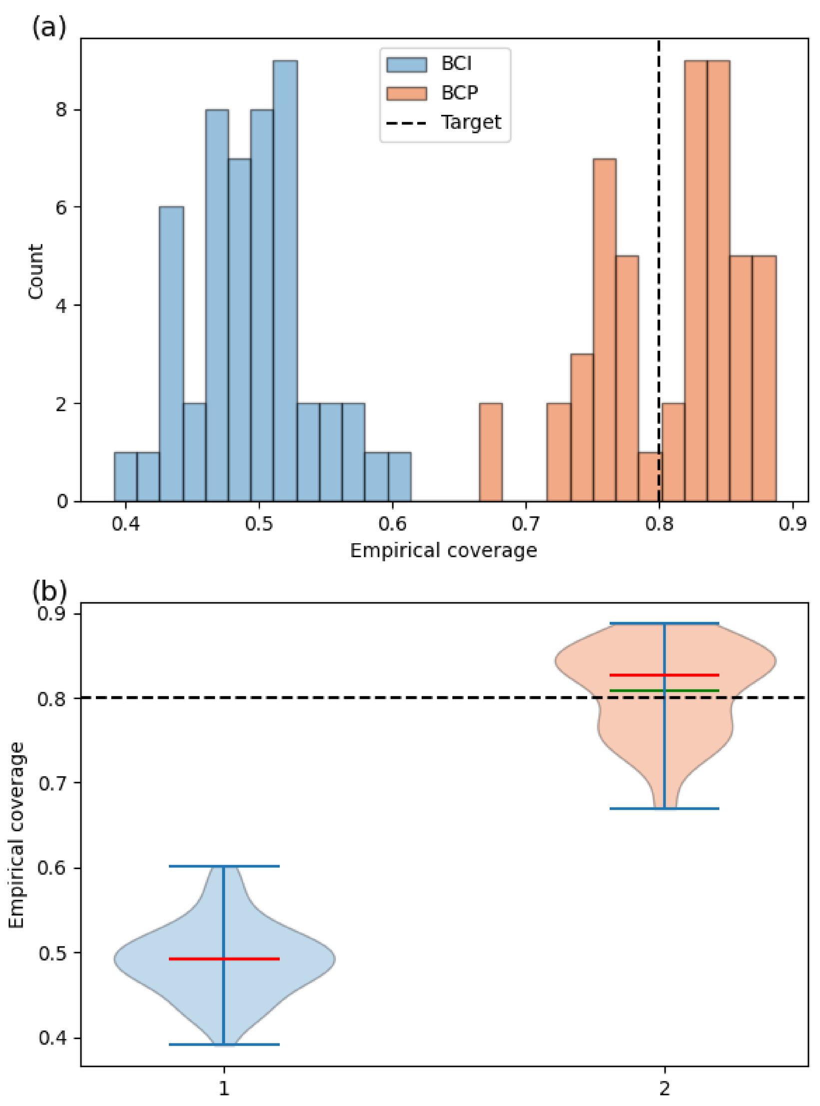
[1]Department of Computer Science, University of Manchester, UK
[2]Department of Computer Science, Aalto University, Espoo, Finland

1. Introduction to BCP

- **Bayesian prediction** (further BCI) can be highly informative and well-calibrated **when the model is correct**, but its coverage can break down under **model misspecification**. An example is demonstrated below.
- **Conformal prediction (CP)**[1] builds prediction regions

$$\mathcal{C}_{\text{cp}}(x_{n+1}) = \{y \in \mathcal{Y} : s(x_{n+1}, y) \leq \lambda_{\text{cp}}\}$$

that achieve the desired **marginal coverage $1-\alpha$** criterion $\mathbb{P}[y_{\text{true}} \in \mathcal{C}(x_{\text{new}})] \geq 1-\alpha$ under nothing more than **exchangeability assumption**.



- [2] Theorem 2.10 implies that among all valid predictors, there always exists a conformal predictor that achieves **equal coverage** with **smaller or equal** prediction regions. Hence, defining the **efficiency** of prediction sets as $|\mathcal{C}(x)|$.
- **Conformal Bayesian Computation (CB)**[3] solves the vulnerability of Bayesian prediction by guaranteeing coverage even when the Bayesian model is wrong, but off-the-shelf CP scores can be conservative, leading to wide sets.
- **Bayesian quadrature (BQ) optimisation**[4] provides a sample-efficient estimator of expected prediction-set size, allowing us to optimise the conformal threshold.

BCP = CB + BQ

2. Bayesian non-conformity score

- **Non-conformity score (posterior-predictive density)**

$$s(x, y) = -\log \hat{p}(y \mid x, \mathcal{D})$$

is **exchangeable** across the augmented sample, so the usual rank test still guarantees finite-sample coverage.

- **AOI importance estimate** (to avoid model re-training) for a candidate y:

$$\tilde{w}^{(t)}(x, y) = \frac{f_{\theta^{(t)}}(y \mid x)}{\sum_{t'=1}^{T} f_{\theta^{(t')}}(y \mid x)} \qquad \widehat{p}(y \mid x, \mathcal{D}_{\text{tr}}) = \sum_{t=1}^{T} \tilde{w}^{(t)}(x, y)\, f_{\theta^{(t)}}(y \mid x)$$

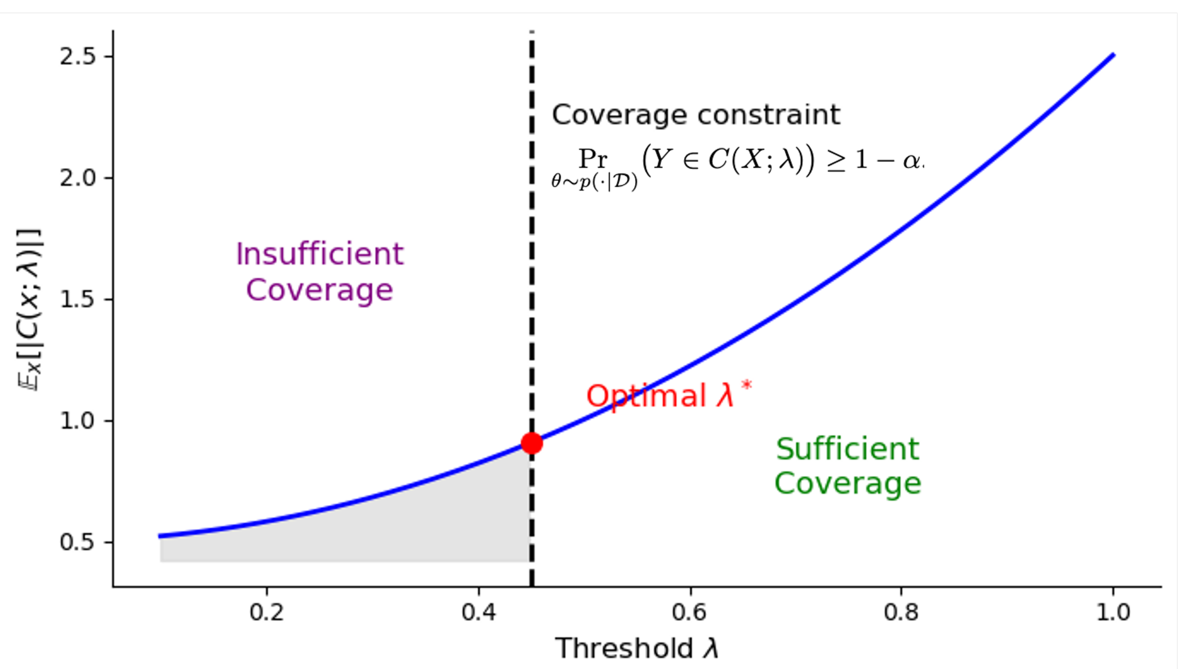
3. Conformal prediction as a decision risk problem

- Let $\mathcal{C}(x; \lambda)$ denote the CP prediction set produced by a **decision rule $\boldsymbol{\lambda}$** for input x and define the **loss** for missing true label y as $L(y, \mathcal{C}(x; \lambda)) = \mathbb{I}\{y \notin \mathcal{C}(x; \lambda)\}$.
- Classical CP constructs a prediction set by **tresholding a non-conformity score $\boldsymbol{s(x, y)}$** that is fixed, i.e., **independent of model parameters $\boldsymbol{\theta}$** and decision variable λ.

BCP seeks λ such that mis-coverage risk $R(\lambda) = \mathbb{E}\left[L(Y, \mathcal{C}(X; \lambda))\right]$ does not exceed a user-specified α. Hence, leading to **the constrained optimisation**:

$$\min_{\lambda} \quad \mathbb{E}_X\big[\,|\mathcal{C}(X; \lambda)|\,\big]$$
$$\text{s.t.} \quad \mathbb{P}_{\mathcal{D}}\Big[\mathbb{P}_{(X,Y)}(Y \notin \mathcal{C}(X; \lambda)) \leq \alpha\Big] \geq 1-\beta$$

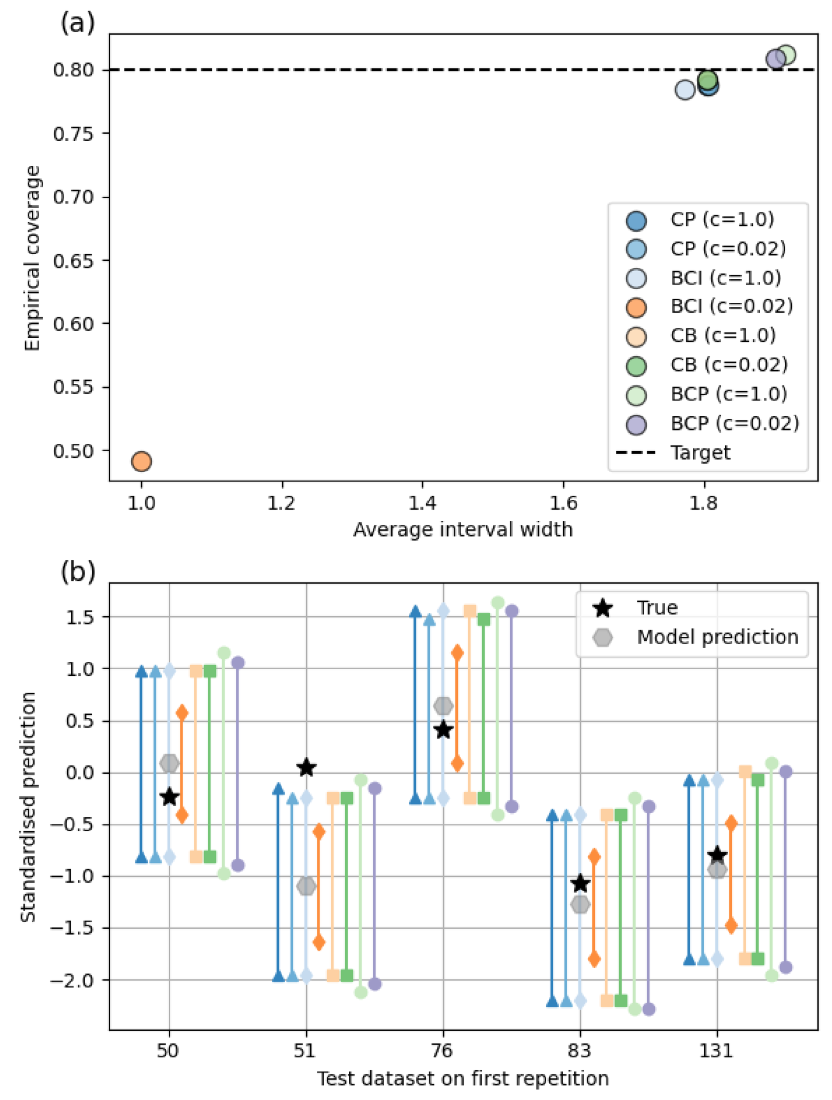
- To estimate $\mathbb{E}_X[\cdot]$, BQ is used: $\mathbb{E}_X[|\mathcal{C}(X; \lambda)|] = \int |\mathcal{C}(x; \lambda)|\, p(x)\, dx$. Therefore, enabling **efficient optimisation of $\boldsymbol{\lambda}$ under posterior uncertainty**.
- Trade-off between threshold and efficiency is illustrated below.

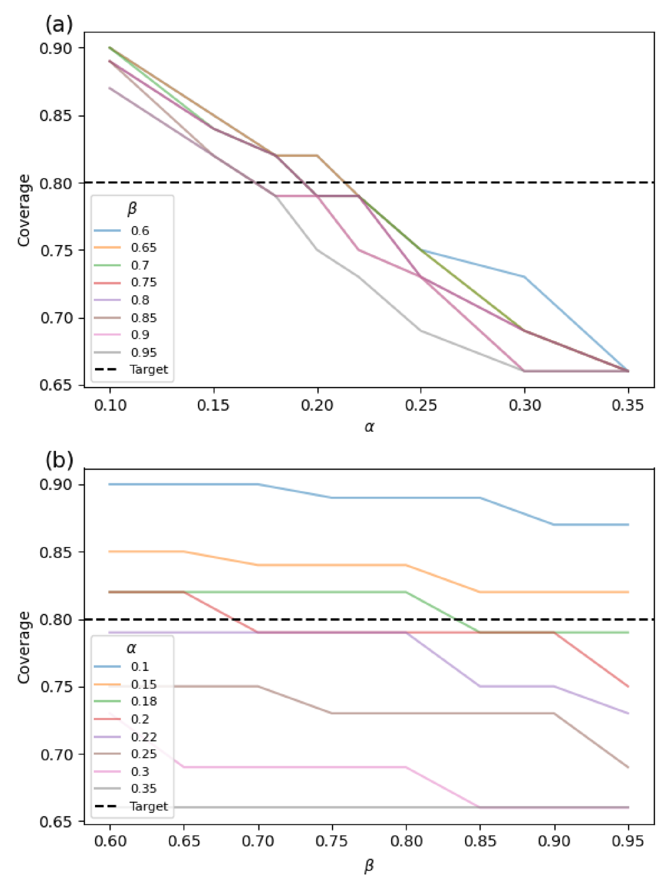


4. Regression and binary classification examples

- **Dataset:** Diabetes[5], $n = 442$, $d = 10$; all predictors and the response are standardised.
- **Bayesian sparse linear regression model:** $f_\theta(y \mid x) = \mathcal{N}(y \mid \theta^\top x + \theta_0, \tau^2)$.
- With hierarchical priors:
$\pi(\theta_j) = \text{Laplace}(0, b)$,
$\pi(b) = \text{Gamma}(1, 1)$
and $\pi(\tau) = \mathcal{N}^+(0, c)$.
- **Two priors** on τ: $c = 1.0$ as **well-specified** (realistic error rate) and $c = 0.02$ as **misspecified** (overconfident error scale).

Method	c	Cov. (%)	Width
Split-CP	1.0	78.75	1.80
Split-CP	0.02	78.75	1.81
BCI	1.0	78.47	1.77
BCI	0.02	49.17	1.00
CB	1.0	79.26	1.80
CB	0.02	79.20	1.80
BCP	1.0	**81.25**	**1.92**
BCP	0.02	**80.83**	**1.90**





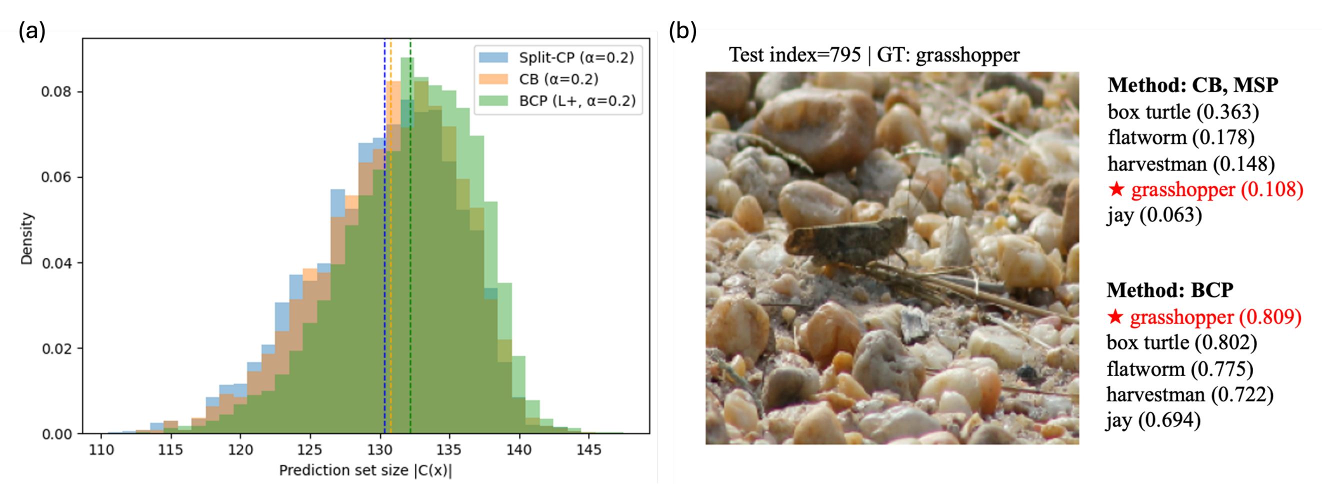
- **Dataset:** Breast cancer[6], $n = 569$, $d = 30$, binary labels (benign vs. malignant).
- **Bayesian logistic model:**
$f_\theta(y{=}1 \mid x) = \big(1 + \exp[-(\theta^\top x + \theta_0)]\big)^{-1}$
with standard Gaussian priors and target miscoverage level $\alpha = 0.2$.

Method	Cov. (%)	Avg. set size
Split CP	**80.54**	**0.81**
BCI	98.91	1.05
CB	80.34	0.81
BCP	81.94	0.82

5. High-dimensional categorical classification example

- **Dataset:** ImageNet-A out-of-distribution dataset[7].
- **Model:** ImageNet-pretrained ResNet-50 backbone without hyperparameter tuning.

Method	Coverage (%)	Cov. std	Mean $\lvert\mathcal{C}(x)\rvert$	P95 $\lvert\mathcal{C}(x)\rvert$
Split CP	77.78	0.4158	122.7 ± 5.1	138
CB	79.34	0.4147	123.1 ± 5.0	138
BCP	**79.51**	**0.4108**	**125.3 ± 4.9**	**139**



References

[1] Anastasios N. Angelopoulos and Stephen Bates. *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification.* arXiv preprint. 2021. arXiv: 2107.07511 [stat.ML].

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

[3] Edwin Fong and Chris Holmes. *Conformal Bayesian Computation.* 2021. arXiv: 2106.06137 [stat.ME]. URL: https://arxiv.org/abs/2106.06137.

[4] Jake C Snell and Thomas L Griffiths. "Conformal Prediction as Bayesian Quadrature". In: *arXiv preprint arXiv:2502.13228* (2025).

[5] Bradley Efron et al. "Least angle regression". In: *Annals of Statistics* 32.2 (2004), pp. 407–499.

[6] William H. Wolberg and Olvi L. Mangasarian. "Multisurface method of pattern separation for medical diagnosis applied to breast cytology". In: *Proceedings of the National Academy of Sciences* 87.23 (1990), pp. 9193–9196.

[7] Dan Hendrycks et al. "Natural Adversarial Examples". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2021, pp. 15262–15271.