Bayesian Conformal Prediction

Fanyi Wu[1], Michele Caprio[1], Jorge Loria[2], Samuel Kaski[1,2]

[1]Department of Computer Science, University of Manchester

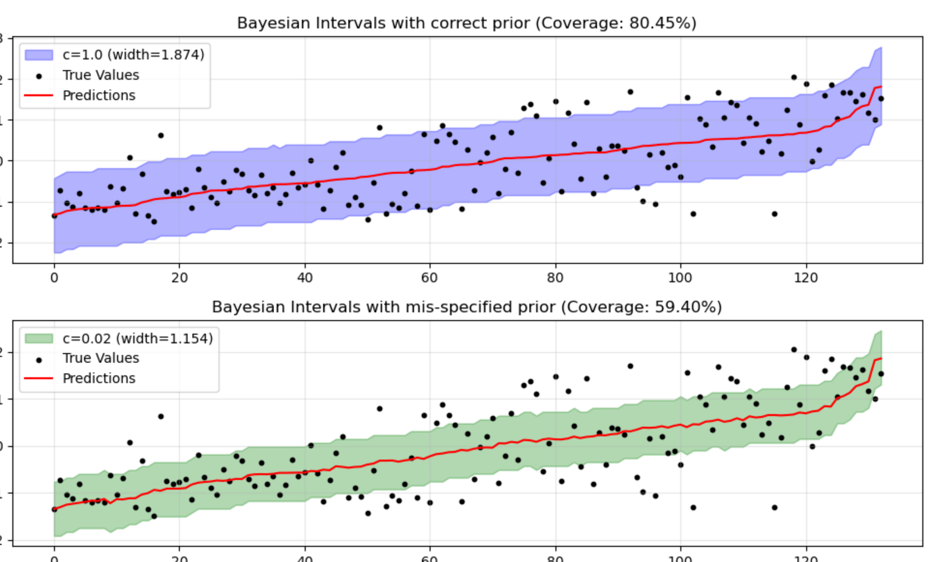
[2]Department of Computer Science, Aalto University

1. Bayes vs. Conformal Prediction

Bayesian prediction can be highly informative and well-calibrated **when the model is correct**, but its coverage can break down under **model misspecification** (the M-open view).

Conformal prediction (CP) builds prediction regions that achieve the desired marginal coverage 1−α under nothing more than **exchangeability**—no other distributional assumptions. These regions **contain the true outcome with probability 1−α** for any exchangeable data-generating process, providing rock-solid uncertainty quantification.

(Bayesian Conformal Prediction) BCP **fixes this vulnerability** by guaranteeing coverage **even when the Bayesian model is wrong**, but off-the-shelf CP scores can be conservative, leading to wide sets.



Motivation: develop **Bayesian-inspired non-conformity scores** that retain the exact finite-sample guarantees of conformal prediction **while producing narrower, more efficient prediction regions**.

2. Bayesian non-conformity score

- **Non-conformity Score (posterior–predictive density)**

$$\sigma_i = p(Y_i \mid X_i,\, Z_{1:n+1})$$

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

- **AOI importance estimate** (no model re-fit). For a candidate y,

$$\widehat{p}(y \mid x_{n+1}) \;=\; \sum_{t=1}^{T} \tilde{w}^{(t)}\, f_{\theta^{(t)}}(y \mid x_{n+1}), \qquad w^{(t)} = f_{\theta^{(t)}}(y \mid x_{n+1}), \qquad \tilde{w}^{(t)} = \frac{w^{(t)}}{\sum_{t'} w^{(t')}}.$$

3. Validity in Bayesian Conformal Prediction

Traditional conformal prediction (CP) satisfies Type-2 validity.

For exchangeable data, the full conformal set

$$C_\alpha(X_{n+1}) = \{\, y : r(y) > \alpha \,\}$$

satisfies: $\mathbb{P}\{Y_{n+1} \in C_\alpha(X_{n+1})\} \geq 1-\alpha$

Why Type-2 validity matters.

It controls the overall error probability:

$$\mathbb{P}\{Y_{n+1} \notin C_\alpha(X_{n+1})\} \leq \alpha$$

in finite samples, **without any modelling assumptions beyond exchangeability**.

Bayesian Conformal Prediction (BCP) also satisfies Type-2 validity.

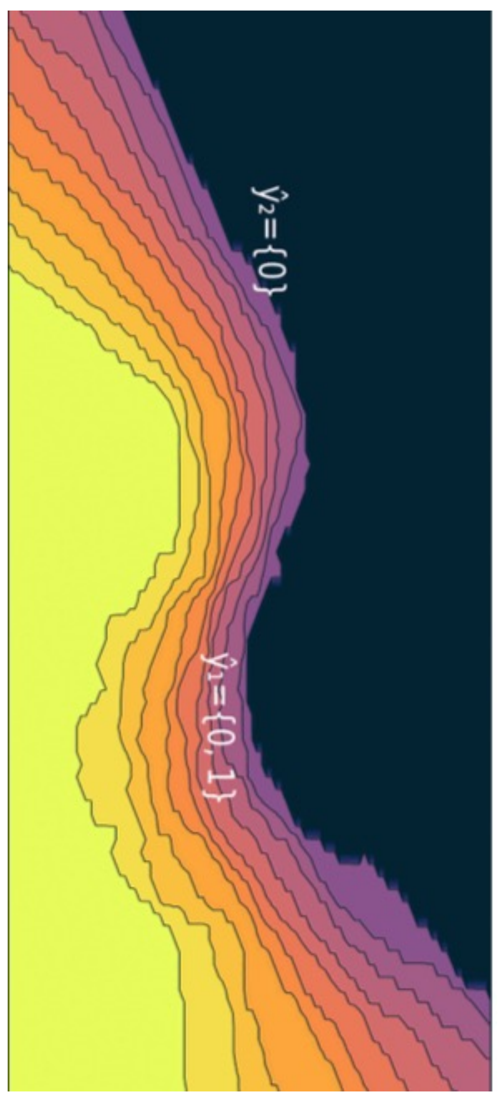
Using the posterior predictive density as an exchangeable conformity score:

$$\sigma_i = p(Y_i \mid X_i,\, Z_{1:n+1})$$

BCP inherits the same guarantee:

$$\mathbb{P}\{Y_{n+1} \in C_\alpha^{\mathrm{BCP}}(X_{n+1})\} \geq 1-\alpha$$

Hence BCP preserves CP's marginal coverage --Bayesian procedure can possess exact frequentist validity.

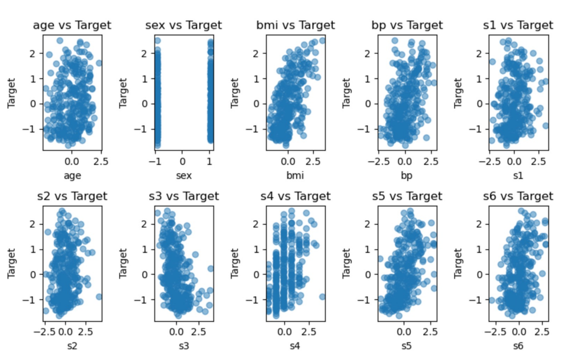


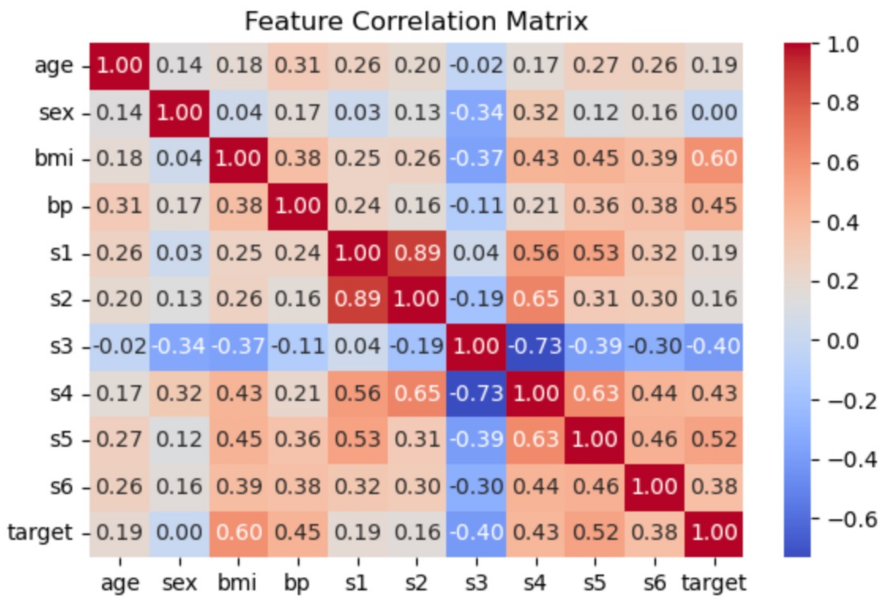
4. Diabetes Regression Example

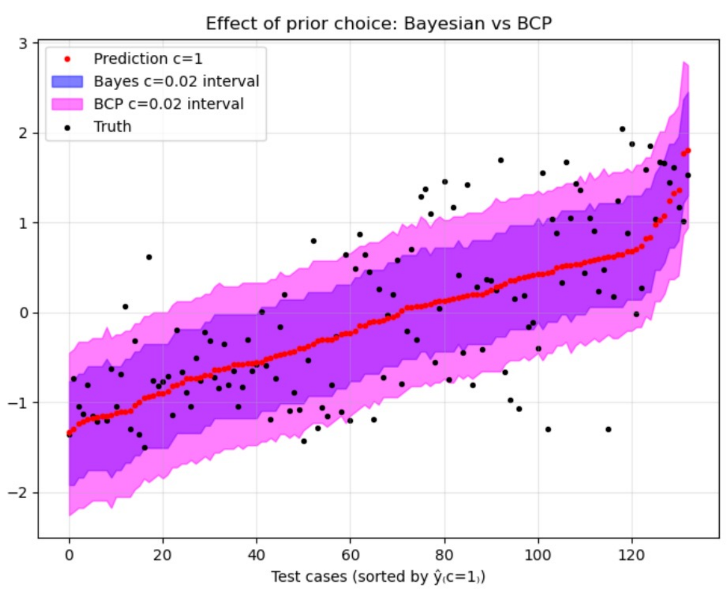
- **Dataset** — Diabetes (Efron *etal.* 2004), $n = 442$, $d = 10$; all predictors and the response are standardised.
- **Bayesian model**

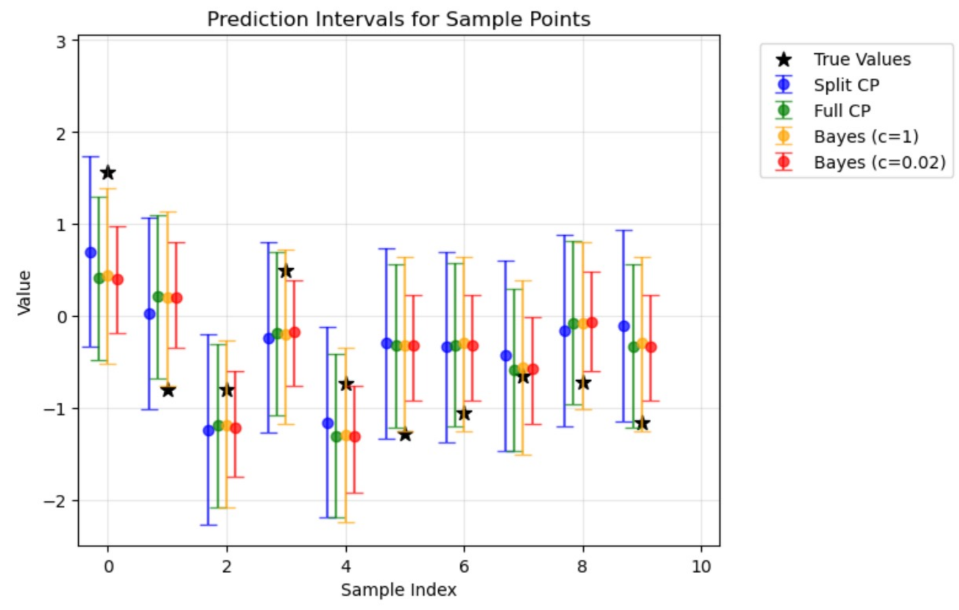
$$f_\theta(y \mid x) = \mathcal{N}\big(y \mid \theta^\top x + \theta_0,\, \tau^2\big), \qquad \pi(\theta_j) = \mathrm{Laplace}(0,b),\ \pi(\tau) = \mathcal{N}^+(0,c).$$

- **Two priors on** τ
  - $c = 1$: well–specified
  - $c = 0.02$: misspecified









		Bayes	CB	Split	Full ($\lambda = 0.004$)
Coverage	$c = 1$	0.806 (0.005)	0.808 (0.006)	0.809 (0.006)	0.808 (0.006)
	$c = 0.02$	**0.563** (0.006)	0.809 (0.006)	/	/
Length	$c = 1$	1.84 (0.01)	1.87 (0.01)	1.91 (0.02)	1.86 (0.01)
	$c = 0.02$	1.14 (0.00)	1.87 (0.01)	/	/
Run-time (secs)	$c = 1$	0.488 (0.107)	0.702 (0.019)	0.065 (0.001)	11.529 (0.232)
	$c = 0.02$	0.373 (0.002)	0.668 (0.003)	/	/

5. The 'Optimal' Non-conformity Score

Vovk et al. (2005, Theorem 2.10) establishes that conformal predictors are at minimum equivalent to conservatively valid predictors.

For a conformal predictor with significance level $\epsilon \in (0,1)$, there exists a conformity measure that produces confidence regions that are valid and minimal. This is formalized through the non-conformity measure:

$$A(B,z) := \inf\{\epsilon : z \in S(B,\epsilon)\}$$

where $S(B,\epsilon)$ is the set of elements that would cause an error at significance level ϵ. The theorem guarantees that the resulting conformal predictor is at least as good as any given predictor Γ.

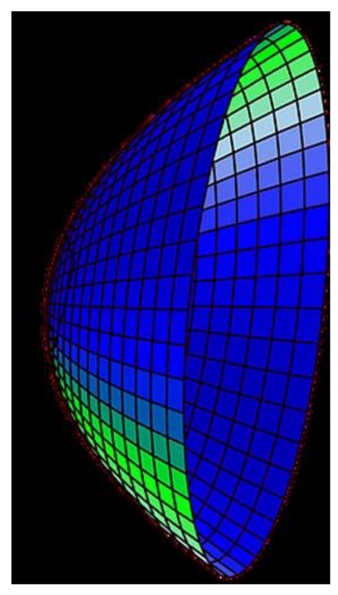
For any $\alpha \in [0,1]$ and any non-conformity measure $\Psi \in \mathcal{F}$, there exists another non-conformity measure $\Psi' \neq \Psi$ such that:

$$\mathcal{R}_\alpha^{\Psi'}(\mathbf{y}^n) \supset \mathcal{R}_\alpha^{\Psi}(\mathbf{y}^n)$$

with strict inclusion for some value of α. Additionally:

$$P[Y_{n+1} \in \mathcal{R}_\alpha^{\Psi'}(\mathbf{y}^n)] \geq 1-\alpha$$

uniformly in probability distribution P and sample size n.



LABLE Method uses probabilistic models to determine smallest prediction sets:

$$\begin{aligned}
\text{Set-valued classifier} &: H(x) = \{y : p(y|x) \geq t_\alpha\} \\
\text{Error constraint (total)} &: \mathbb{P}\{Y \notin H(X)\} \leq \alpha \\
\text{Error constraint (class-specific)} &: \mathbb{P}\{Y \notin H(X) | Y = y\} \leq \alpha_y \text{ for all } y \\
\text{Optimization objective} &: \min_H \mathbb{E}\{|H(X)|\} \text{ subject to error constraints}
\end{aligned}$$

6. Discussion

Feature	Conformal Prediction	Bayesian Conformal Prediction	LABEL Classifiers
Core idea	Distribution-free uncertainty quantification	Using Bayesian posterior as non-conformity score	Optimizing prediction set size under error constraints
Theoretical guarantee	Valid coverage regardless of distribution	Valid coverage with improved efficiency	Valid coverage with minimal prediction set size
Assumes knowledge of	Only exchangeability of data	Prior distribution and likelihood	True conditional probabilities
Computational approach	Ranking-based non-conformity scores	Posterior sampling (MCMC) + importance sampling	Thresholding conditional probabilities
Key advantage	Distribution-free guarantees	Efficiency through Bayesian modeling	Minimal prediction set size
Key limitation	Can be inefficient without good conformity measure	Requires accurate Bayesian model	Requires calibrated probabilities

There are many ways to suggest and investigate for the non-conformity score that gives the minimal prediction region by considering Vovk's Theorem 2.10 and LABEL method. For example, LABEL methods can be **thresholding the posterior-predictive probabilities.**

- **Define a posterior–predictive score**

$$\sigma\big((x,y)\big) \;=\; -\hat{p}(y \mid x, \mathcal{D}),$$

where $\hat{p}(y \mid x, \mathcal{D})$ is the Bayesian posterior predictive density.

- **LABEL thresholding** Construct the set-valued classifier

$$H_\alpha(x) = \Big\{\, y : \hat{p}(y \mid x, \mathcal{D}) \geq t_\alpha \Big\},$$

with the threshold t_α chosen to satisfy LABEL's error-rate constraint.

- **Bayesian Conformal Prediction (BCP)** Apply Vovk's full/split conformal rank test to the augmented exchangeable scores $\{\sigma_1, \ldots, \sigma_{n+1}\}$.

Hence LABEL drives the set toward the highest-density (smallest-volume) region, while the conformal wrapper guarantees exact finite-sample coverage, giving BCP both tight and valid prediction sets.

July 16, 2025, University of Cambridge, Cambridge, UK