Conformal Bayesian Computation

Edwin Fong

University of Oxford
The Alan Turing Institute

Joint work with Chris Holmes (Oxford, Turing)

NeurIPS 2021

Motivation

Consider the regression setting, with $Z_i = \{X_i, Y_i\}$ for an outcome of interest Y_i and covariates X_i. Assume that $Z_{1:n+1} \overset{\text{iid}}{\sim} P$, and we observe $Z_{1:n}$.

We are concerned with *prediction* in Bayesian regression:

$$\underbrace{\pi(\theta \mid Z_{1:n})}_{\text{Posterior}} \propto \prod_{i=1}^{n} \underbrace{f_\theta(Y_i \mid X_i)}_{\text{Likelihood}} \underbrace{\pi(\theta)}_{\text{Prior}}$$

The posterior predictive density is:

$$p(Y_{n+1} \mid X_{n+1}, Z_{1:n}) = \int f_\theta(Y_{n+1} \mid X_{n+1})\, \pi(\theta \mid Z_{1:n})\, d\theta\,.$$

Motivation

We can summarize the predictive distribution with $1 - \alpha$ credible intervals:

$$C_\alpha(X_{n+1}) = [P^{-1}(\alpha/2), P^{-1}(1 - \alpha/2)]$$

where P^{-1} is the quantile function of $P(Y_{n+1} \mid X_{n+1}, Z_{1:n})$.

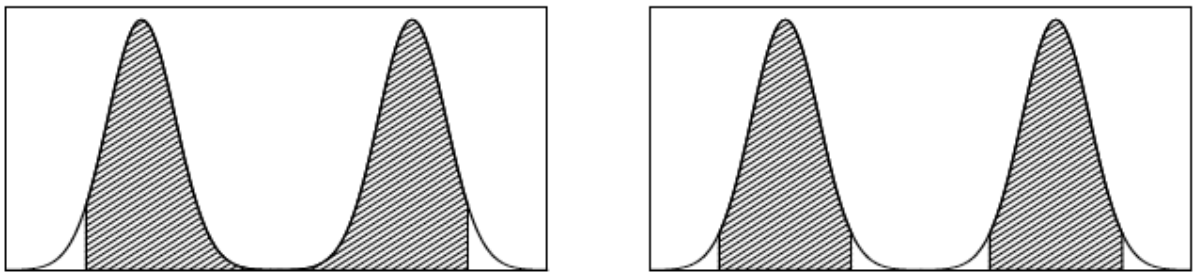

Figure 1: Central (left) and highest density (right) credible intervals; taken from [Gelman et al., 2013]

- Bayesian intervals have a nice interpretation, but no frequentist guarantees if the model is misspecified (which it always is)

Conformal Prediction [Vovk et al., 2005]

Assume $Z_{1:n+1} \overset{\text{iid}}{\sim} P$, we can estimate

$$E\left[Y_{n+1} \mid X_{n+1}\right] \approx \hat{\mu}(X_{n+1})$$

with your predictive algorithm of choice.

Conformal prediction gives us a confidence interval $C_\alpha(X_{n+1})$ that satisfies

$$P\left(Y_{n+1} \in C_\alpha(X_{n+1})\right) \geq 1 - \alpha.$$

Note: P is over $Z_{1:n+1}$.

There are no assumptions on P, and the average width of C_α depends on the quality of $\hat{\mu}$.

Conformity Measures

A conformity measure is a function $\sigma : \mathbb{Z}^{n+1} \times \mathbb{Z} \to \mathbb{R}$, that measures how similar a datum is to a **bag** (unordered set) of data:

$$\sigma_i := \sigma(\underbrace{Z_1, \ldots, Z_{n+1}}_{\text{Bag}}; \underbrace{Z_i}_{\text{Datum}}).$$

The most common choice is the (negative) residual:

$$\sigma_i = -|Y_i - \hat{\mu}(X_i)|$$

where $\hat{\mu}$ is computed from $Z_{1:n+1}$ and is permutation-invariant.

Key: If $Z_{1:n+1}$ are exchangeable, then $\sigma_{1:n+1}$ are exchangeable[1].

[1] To see this, imagine swapping Z_i, Z_j - this will only swap σ_i, σ_j.

Full Conformal Prediction - Algorithm

Algorithm 1: Conformal Prediction

We have observed $Z_{1:n}$ and X_{n+1}.
Select miscoverage level α.

for $y \in \mathbb{R}$ **do**

- Fit regression model $\hat{\mu}$ with *augmented* data $\{Z_{1:n}, \{y, X_{n+1}\}\}$
- Compute $\sigma_{1:n}$ and σ_{n+1}
- Compute $r(y) = \frac{\text{Rank}(\sigma_{n+1})}{n+1}$ among $\sigma_{1:n+1}$

end

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

The above set satisfies $P\left(Y_{n+1} \in C_\alpha(X_{n+1})\right) \geq 1 - \alpha$.

Note: In practice, we need to select a grid[2] $y \in \mathcal{Y}$ to approximate the algorithm.

[2] See [Chen et al., 2016, Chen et al., 2018]

Conformal Prediction - Summary

Conformal prediction gives us guaranteed confidence intervals $C_\alpha(X_{n+1})$:

- We only require exchangeability of the data $\{X_i, Y_i\}_{i=1:n+1}$.
- We fit the model to the *augmented* data set $\{Z_{1:n}, \{y, X_{n+1}\}\}$.

Main limitation: Refitting the model for each $y \in \mathcal{Y}$ and X_{n+1} is expensive!

The split conformal method (e.g. [Lei et al., 2018]) is one way to avoid this, but produces wider intervals.

Interestingly, Bayesian models provide another possible scalable solution for *full* conformal prediction.

Conformal Bayes - Idea

Given a Bayesian model, a natural conformity score is the "add-one-in" (AOI) posterior predictive density, which we call **conformal Bayes**:

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

where

$$p(y \mid x, Z_{1:n+1}) = \int f_\theta(y \mid x)\, \pi(\theta \mid Z_{1:n+1})\, d\theta.$$

- The predictive is permutation-invariant to $Z_{1:n+1}$.
- A conforming datum $\{Y_i, X_i\}$ will have a high density value.

Refitting Bayes

For each $Z_{n+1} = \{y, X_{n+1}\}$ with $y \in \mathcal{Y}$, we want the updated posterior

$$\pi(\theta \mid Z_{1:n+1}) \propto f_\theta(y \mid X_{n+1}) \times \pi(\theta \mid Z_{1:n}).$$

Bayesian analysis usually involves Markov Chain Monte Carlo (MCMC) to provide samples from the posterior, giving us

$$\theta^{(1:T)} \sim \pi(\theta \mid Z_{1:n}).$$

Key: Use (self-normalized) importance sampling, with $q(\theta) = \pi(\theta \mid Z_{1:n})$ as the proposal for $p(\theta) = \pi(\theta \mid Z_{1:n+1})$.

Here, q usually has heavier tails than p which is helpful for weight stability, unlike when computing leave-one-out (LOO) predictives.

Refitting Bayes

For each $Z_{n+1} = \{y, X_{n+1}\}$ with $y \in \mathcal{Y}$, we want to estimate

$$\sigma_i := \underbrace{p(Y_i \mid X_i, Z_{1:n+1})}_{\text{AOI predictive}} = \int f_\theta(Y_i \mid X_i) \underbrace{\pi(\theta \mid Z_{1:n+1})}_{\text{Updated posterior}} d\theta$$

for $i = 1, \ldots, n+1$.

Given $\theta^{(1:T)} \sim \pi(\theta \mid Z_{1:n})$, compute AOI predictive through:

$$\sigma_i \approx \sum_{t=1}^{T} \tilde{w}^{(t)} f_{\theta^{(t)}}(Y_i \mid X_i)$$

where

$$\tilde{w}^{(t)} \propto \frac{\pi(\theta \mid Z_{1:n+1})}{\pi(\theta \mid Z_{1:n})} \propto f_{\theta^{(t)}}(y \mid X_{n+1})$$

Conformal Bayes - Summary

Conformal Bayes is a cheap post-processing step to produce calibrated intervals $C_\alpha(X_{n+1})$ from MCMC samples:

- Conformity score σ_i is the AOI posterior predictive density
- Refitting is carried out through efficient importance sampling, where the weights are well-behaved (compared to LOO)
- Computational complexity is $\mathcal{O}(nT|\mathcal{Y}|)$, but is very efficient on GPU due to a matrix multiplication
- We also extend to partial exchangeable data with Bayesian hierarchical models

Example

Diabetes dataset [Efron et al., 2004] contains $n = 442$ subjects.
Y is diabetes progression and X is patient readings ($d = 10$).

The Bayesian linear model:

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

Sparse prior:

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

For the variance prior, $c = 1$ is well-specified, and $c = 0.02$ is misspecified.

Example

We carry out 50 repeats of train/test (70/30) splits to evaluate coverage and lengths of intervals on Y_{n+1}.

MCMC overhead required $\approx$ 25s for $T = 8000$. We set the grid to $[y_{\text{min}} - 2, y_{\text{max}} + 2]$ with $|\mathcal{Y}| = 100$.

Table 1: Diabetes; Coverage values *not* within 3 standard errors (in brackets) of the target coverage $(1 - \alpha) = 0.8$ are in **red**; $c = 0.02$ is misspecified.

		Bayes	Conformal Bayes
Coverage	$c = 1$	0.806 (0.005)	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)
	$c = 0.02$	1.14 (0.00)	1.87 (0.01)
Run-time (secs)	$c = 1$	0.488 (0.107)	0.702 (0.019)
	$c = 0.02$	0.373 (0.002)	0.668 (0.003)

Conclusions

Conformal Bayes: Use the Bayesian posterior predictive density as the conformity measure

- Provides guaranteed coverage under model misspecification and can be used to diagnose Bayesian miscalibration
- A general wrapper around MCMC output like Stan, PyMC3, etc. based on importance sampling
- Enables full conformal inference for a wider class of models

Thank you for your attention!

References

Chen, W., Chun, K.-J., and Barber, R. F. (2018).
Discretized conformal prediction for efficient distribution-free inference.
Stat, 7(1):e173.

Chen, W., Wang, Z., Ha, W., and Barber, R. F. (2016).
Trimmed conformal prediction for high-dimensional models.
arXiv preprint arXiv:1611.09933.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004).
Least angle regression.
Annals of statistics, 32(2):407–499.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013).
Bayesian Data Analysis.
CRC Press.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018).
Distribution-free predictive inference for regression.
Journal of the American Statistical Association, 113(523):1094–1111.

Vovk, V., Gammerman, A., and Shafer, G. (2005).
Algorithmic learning in a random world.
Springer Science & Business Media.