Nonparametric regression, the Bayesian way

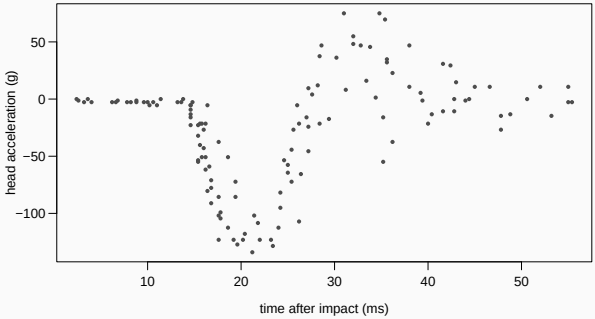
Polynomials, free-knot splines, Bayesian CART, BART & friends on the `mcycle` data

Hedibert Freitas Lopes

Insper – Instituto de Ensino e Pesquisa

Disclaimer: these slides, the R code and the numerical results were prepared with the assistance of an AI model (Claude, Anthropic), under the author's direction. All numbers are produced by the accompanying R scripts.

The data: `mcycle` (Silverman, 1985)



- Head acceleration (g) vs. time (ms) after impact, simulated motorcycle crash.
- Classic benchmark: flat, sharp dip, rebound, slow decay.
- Strongly heteroskedastic: noise grows after $\approx$ 15 ms.
- As in the course script: first $n = 132$ observations; x and y standardized.

Eight competitors

	Model	Prior / tuning	Bayes factor?
0	Polynomial, degree $d \in \{0, \ldots, 20\}$	g-prior, $d \sim U\{0..20\}$, BMA	✓
1	Piecewise linear, unknown knots	g-prior, $k \sim U\{0..7\}$, knots on grid	✓
2	Piecewise quadratic, unknown knots	idem	✓
3	Bayesian CART (single tree)	CGM tree prior, $\alpha = .95$, $\beta = 1$	–
4	BART	200 trees, default CGM (2010) prior	–
5	Random forest *(not Bayesian)*	500 trees, `randomForest`	–
6	Boosting *(not Bayesian)*	`gbm`, depth 2, $\nu = 0.01$, #trees by 5-fold CV	–
7	Neural network	1 hidden layer, size/decay by 5-fold CV	–

- Models 0–2: closed-form marginal likelihoods $\Rightarrow$ exact Bayes factors.
- Models 0–4: full posterior predictive $\Rightarrow$ out-of-sample log predictive density.
- All eight: point predictions $\Rightarrow$ MSE and MAD.

Linear-in-parameters models and exact Bayes factors

A common prior for models 0, 1 and 2

For any basis X_M (polynomial or spline), with X_c its centered version (p columns):

$$y = \alpha \mathbf{1} + X_c \beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I), \qquad p(\alpha, \sigma) \propto 1/\sigma, \qquad \beta \mid \sigma \sim N\big(0,\ g\,\sigma^2 (X_c' X_c)^{-1}\big).$$

Zellner's g-prior with $g = n$ (unit information). Then (Liang et al., 2008, JASA)

$$\log p(y \mid M) \;=\; C \;+\; \tfrac{n-1-p}{2} \log(1+g) \;-\; \tfrac{n-1}{2} \log\big(1 + g(1 - R^2_M)\big).$$

- C is the same for every model $\Rightarrow$ Bayes factors across polynomials *and* splines.
- Posterior: $\beta \mid \sigma, y \sim N\big(\tfrac{g}{1+g}\hat{\beta},\ \tfrac{g}{1+g}\sigma^2 (X_c' X_c)^{-1}\big)$, $\sigma^2 \mid y \sim IG\big(\tfrac{n-1}{2},\ \tfrac{S}{2}\big)$, $S = y_c' y_c \big(1 - \tfrac{g}{1+g} R^2_M\big)$.
- Predictive: Student-t_{n-1} with location $\bar{y} + \tfrac{g}{1+g} x_{0c}' \hat{\beta} \Rightarrow$ exact log predictive densities.

Free-knot splines: estimating the break points

Bases (continuous at the knots $\tau_1 < \cdots < \tau_k$):

$$\text{PW-linear: } \{x,\ (x-\tau_j)_+\}_{j=1}^{k}, \qquad \text{PW-quadratic: } \{x,\ x^2,\ (x-\tau_j)_+^2\}_{j=1}^{k} \ \ (\text{also } C^1).$$

Prior on the break points

- number of knots $k \sim U\{0, 1, \ldots, 7\}$;
- locations: $\{\tau_1, \ldots, \tau_k\}$ uniform over the $\binom{G}{k}$ subsets of a grid of $G = 25$ equally spaced interior points (≈ 2.1 ms apart).

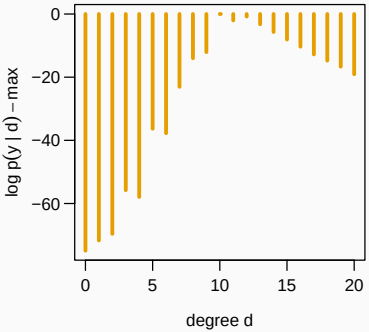
Posterior by complete enumeration: $\sum_{k=0}^{7} \binom{25}{k} = 726{,}206$ configurations per family, each with closed-form $p(y \mid k, \tau)$ (Rcpp, a few seconds).

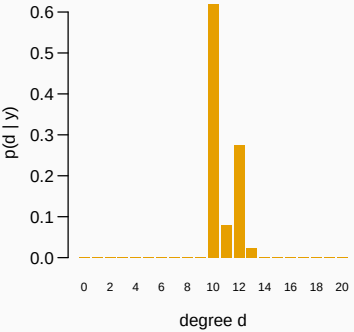
$$p(y \mid \text{family}) = \sum_{k} \sum_{\tau} p(k)\, p(\tau \mid k)\, p(y \mid k, \tau), \qquad p(\tau_j \in \text{knots} \mid y), \qquad p(k \mid y)$$

are all exact (no MCMC), and predictions average over configurations (BMA).

Full-data fits ($n = 132$)

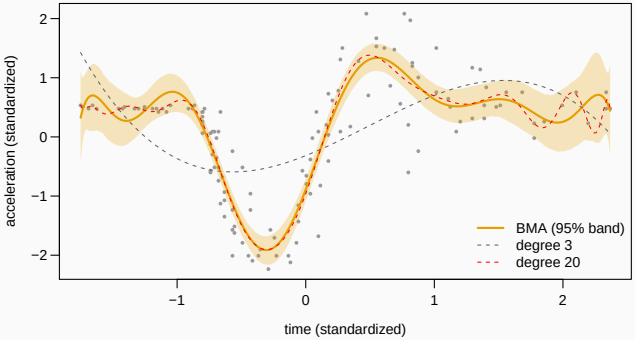
0) Polynomial: which degree?





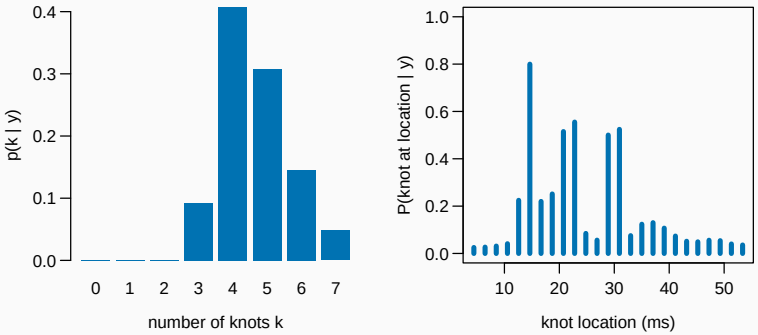
- Uniform prior on $d \in \{0, \ldots, 20\}$. Posterior mode $d = 10$ ($p = 0.62$), then $d = 12$ (0.28), $d = 11$ (0.08).
- Low degrees are hopeless: $\log p(y \mid d = 3)$ is 56 units below the mode.

0) Polynomial: BMA fit



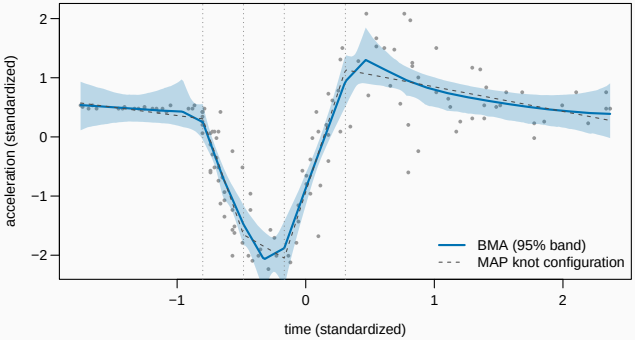
- BMA over degrees, 95% band for $f(x)$.
- Captures dip and rebound, but wiggles in the flat parts.
- Bands flare at both ends: the classic polynomial edge effect.
- Degree 20 (red) is close to the BMA: the g-prior shrinks the extra terms.

1) Piecewise linear: how many knots, and where?



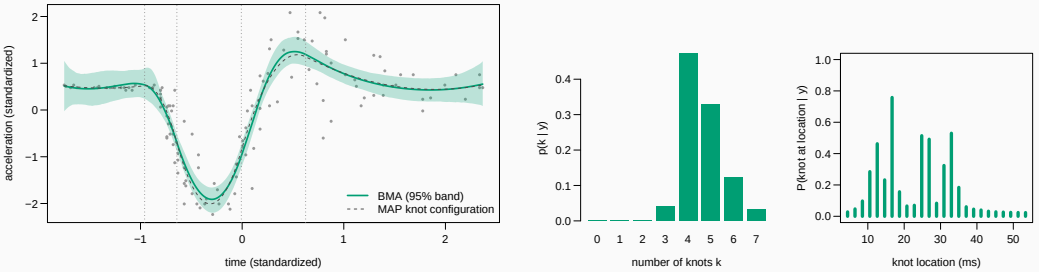
- $p(k \mid y)$: $k = 4$ (0.41), $k = 5$ (0.31), $k = 6$ (0.14), $k = 3$ (0.09), $k = 7$ (0.05).
- A break at 14.6 ms is near certain (0.80); the next two are split between neighbours: 20.8 or 22.8 ms (≈ 0.5 each), 28.9 or 30.9 ms (≈ 0.5 each). Little mass after 35 ms.
- MAP configuration: knots at 14.6, 18.7, 22.8, 28.9 ms.

1) Piecewise linear: BMA fit



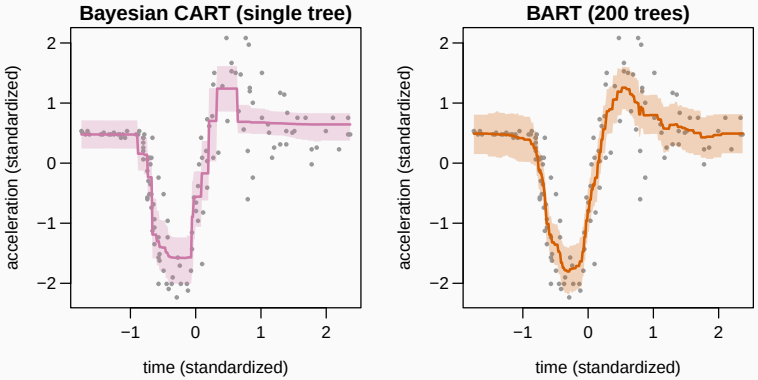
- Flat before impact, then a sharp V and a linear decay.
- Averaging over knot locations rounds the corners of the BMA curve.
- Dotted lines: MAP knots. Dashed: fit given the MAP knots.

2) Piecewise quadratic (C^1 spline)



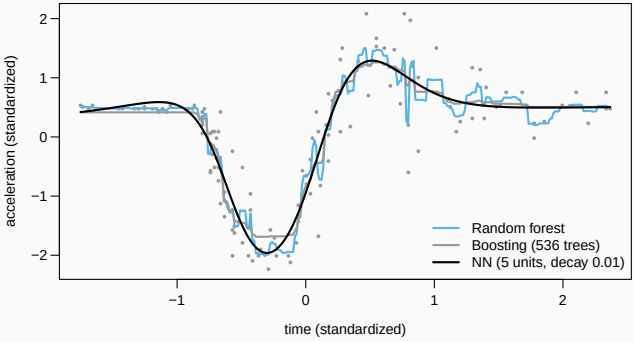
- $p(k \mid y)$: $k = 4$ (0.47), $k = 5$ (0.33), $k = 6$ (0.12). MAP knots: 12.6, 16.7, 24.8, 33.0 ms.
- Most probable break: 16.7 ms (0.76); others near 12.6, 24.8/26.9 and 33.0 ms. Smooth curve, same shape as the linear spline.

3–4) Bayesian CART and BART



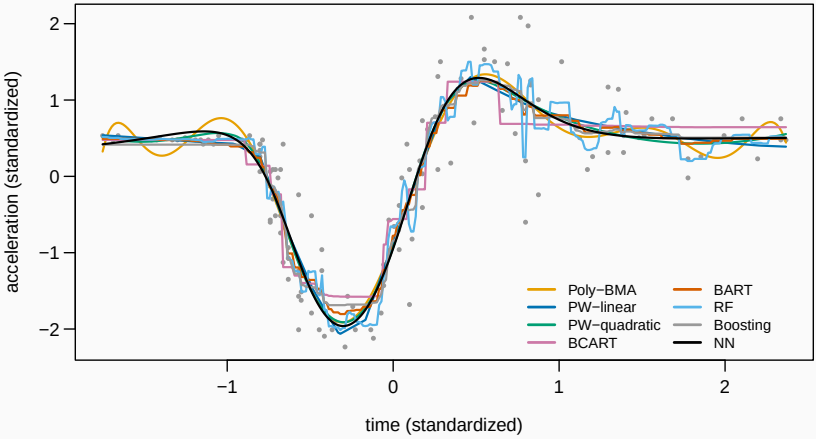
- Both via `BART::wbart`; Bayesian CART = one tree ($\alpha = 0.95$, $\beta = 1$), 5,000 draws after 2,000 burn-in.
- One tree $\Rightarrow$ step function; posterior averaging over trees softens the steps.
- BART: sum of 200 shrunk trees $\Rightarrow$ smooth-ish mean with honest bands.

5–7) Random forest, boosting, neural network



- RF: bagged deep trees, follows the noise.
- Boosting: 536 trees chosen by 5-fold CV.
- NN: 5 hidden units, decay 0.01 chosen by 5-fold CV; average of 5 random starts.
- Weight decay = Gaussian prior on weights: the NN fit is a MAP estimate, not a posterior.
- Point predictions only $\Rightarrow$ no Bayes factor, no predictive density.

All eight posterior means / fits



Bayes factors

Bayes factors: models 0, 1, 2 (full data)

Family	$\log p(y \mid \text{family})$	$\log BF$ vs. polynomial	$\log p(y \mid \text{best member})$
Polynomial (d = 0..20)	72.38	0.00	74.95
PW-linear (k = 0..7)	79.51	7.13	88.34
PW-quadratic (k = 0..7)	78.68	6.29	87.47

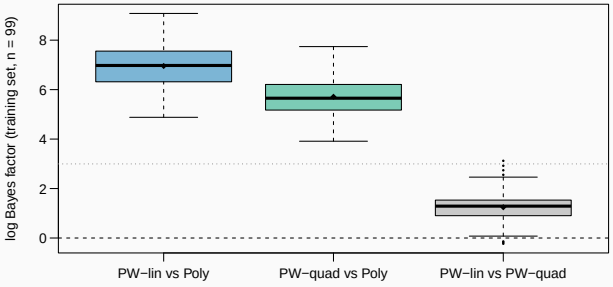
- PW-linear vs. polynomial: $\log BF = 7.13$, $BF \approx 1{,}250$ (very strong, Kass–Raftery).
- PW-quadratic vs. polynomial: $\log BF = 6.29$, $BF \approx 540$.
- PW-linear vs. PW-quadratic: $\log BF = 0.84$, $BF \approx 2.3$: barely worth a mention.
- **Occam at work**: the best single knot configuration beats its own family by ≈ 9 log units. Averaging over 726,206 configurations pays for the search over break points.

How sensitive are the Bayes factors to g?

$\log p(y \mid \text{family})$	$g = 10$	$g = n = 132$	$g = 1000$
Polynomial (d = 0..20)	68.52	72.38	63.47
PW-linear (k = 0..7)	69.94	79.51	75.75
PW-quadratic (k = 0..7)	69.94	78.68	73.88

- Splines $\succ$ polynomial for every g; the margin grows with g (Bartlett/Lindley: big g punishes the 10+ polynomial terms).
- With $g = 10$ the two spline families tie, and $p(k \mid y)$ shifts to $k = 6$–7: a weak penalty asks for more knots, so $K_{\max}$ starts to matter.
- Lesson: Bayes factors are prior-sensitive; report a sensitivity table.

Bayes factors on the 100 training sets ($n = 99$)



- Same exact computation on each training set; dotted lines at $\pm \log 20$.
- PW-linear vs. poly: mean $\log BF = 6.95$; $> \log 20$ in 100/100.
- PW-quad vs. poly: mean 5.71; $> \log 20$ in 100/100.
- PW-linear vs. PW-quad: mean 1.24; positive in 97/100, but strong only in 1/100.

Out-of-sample comparison: 100 random 75/25 splits

Design

- 100 random splits: 99 training, 33 test observations; identical splits for all models.
- All tuning happens inside the training set (degree/knots by the posterior; GBM trees and NN size/decay by inner 5-fold CV).
- Metrics on the test set (point prediction = posterior mean for Bayesian models):

$$\text{MSE} = \frac{1}{33}\sum (y_i - \hat{y}_i)^2, \qquad \text{MAD} = \frac{1}{33}\sum |y_i - \hat{y}_i|,$$

$$\text{LPD} = \frac{1}{33}\sum \log p(y_i \mid y_{\text{train}}) \quad \text{(Bayesian models only)}.$$

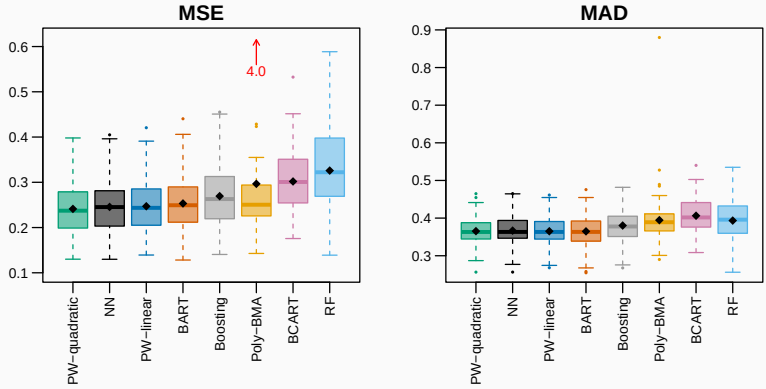
- LPD: exact t-mixtures for 0–2; $\log \frac{1}{S}\sum_s N(y_i \mid f^{(s)}(x_i), \sigma^{2(s)})$ for BCART/BART.
- Also: coverage of 95% posterior predictive intervals.

Results: averages over 100 splits (sd in parentheses)

Model	MSE	MAD	LPD (↑)	Cover. 95% (%)	Avg. rank (MSE)	# best (MSE)
Poly-BMA	0.297 (0.38)	0.395 (0.06)	-0.755 (0.12)	93.0	4.2	17
PW-linear	0.247 (0.06)	0.365 (0.04)	-0.733 (0.12)	93.4	3.4	14
PW-quadratic	**0.241 (0.06)**	0.365 (0.04)	**-0.726 (0.12)**	93.0	2.6	25
BCART	0.302 (0.07)	0.406 (0.05)	-0.820 (0.12)	93.3	6.4	7
BART	0.253 (0.06)	**0.365 (0.04)**	-0.735 (0.13)	92.3	3.8	16
RF	0.326 (0.09)	0.393 (0.05)	–	–	7.2	1
Boosting	0.269 (0.07)	0.380 (0.04)	–	–	5.2	4
NN	0.246 (0.06)	0.367 (0.04)	–	–	3.1	16

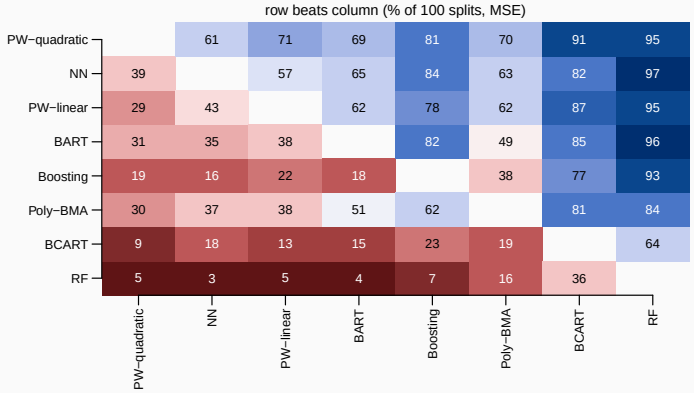
Outcome standardized: MSE $=$ 1 is the variance of y. Bold = best in column.

MSE and MAD across splits



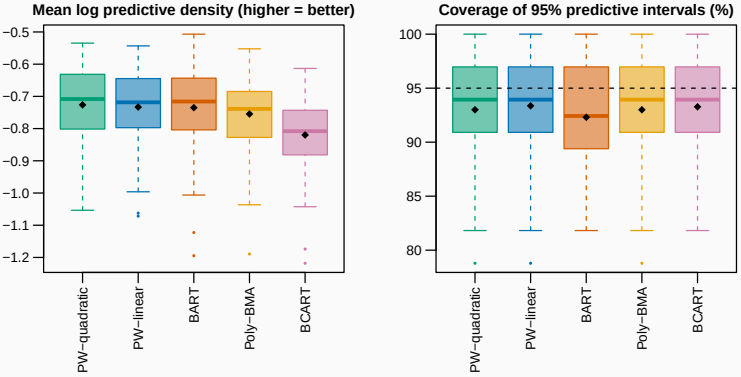
- Top group (PW-quad, NN, PW-linear, BART) within ≈ 0.01 of each other in MSE; MAD is a near tie.
- Polynomial: median MSE 0.250, but one split gives 4.0: the three last time points (55.0–55.4 ms) fell in the test set and the degree-10 polynomial extrapolates beyond 53.2 ms.

Head-to-head: who beats whom?



PW-quadratic beats every competitor in more than half of the splits (NN: 61%, PW-linear: 71%, BART: 69%).

Posterior predictive: log density and coverage



- LPD ranking: PW-quad $>$ PW-linear $\approx$ BART $>$ polynomial $>$ Bayesian CART.
- All coverages are below 95% (92–93%): one common σ^2 cannot fit a noise level that changes with time.

Take-aways

What did we learn?

1. **Structure wins.** Free-knot splines (a handful of break points) beat a degree-10 polynomial by $\log BF \approx 6$–7, and also out of sample.
2. **BF vs. prediction need not agree.** BF slightly prefers PW-linear ($BF \approx 2$); prediction slightly prefers PW-quadratic. Marginal likelihood = *prior* predictive; CV = *posterior* predictive.
3. **Occam's razor is automatic**: averaging over knot locations costs ≈ 9 log units.
4. **BART is a superb default**: no tuning, competitive MSE, and a real predictive distribution. Single-tree Bayesian CART is too rough.
5. **Off-the-shelf ML**: NN (with CV-tuned decay) is competitive in MSE; RF is worst; neither gives a predictive density.
6. **Polynomials extrapolate badly** — one split shows it dramatically.
7. **Every model here is misspecified**: constant $\sigma^2 \Rightarrow$ under-coverage. Next step: heteroskedastic models (e.g., $\log \sigma^2(x)$ spline, heteroskedastic BART).

Reproducibility

- `R/functions.R` — all models (polynomial/spline BMA with Rcpp enumeration, BCART, BART, RF, GBM, NN).
- `R/full_data.R` — full-data fits, exact Bayes factors, g-sensitivity, figures.
- `R/cv_experiment.R` — 100 splits (seed 2026), $\approx$ 7 min on 2 cores.
- `R/summaries.R` — tables and CV figures used in these slides.

Packages: `MASS`, `Rcpp`, `BART`, `randomForest`, `gbm`, `nnet`.

References. Silverman (1985, JRSS-B); Zellner (1986); Liang, Paulo, Molina, Clyde & Berger (2008, JASA); Denison, Mallick & Smith (1998, JRSS-B); Chipman, George & McCulloch (1998, JASA; 2010, AoAS); Breiman (2001); Friedman (2001); Kass & Raftery (1995, JASA).