Locally Robust Semiparametrically Efficient Bayesian Inference

Ulrich K. Müller and Andriy Norets
Princeton University and Brown University

July 2025

Appeal of Bayesian Inference

- Complete class theorem: A decision rule is admissible if and only if it is Bayes

  $\Rightarrow$ Rules generated by other reasoning are either inadmissible, or implicitly Bayes

- If there are multiple objectives, Bayes actions are automatically coherent

- Ability to express soft constraints on parameters via prior

- Bayes descriptions of uncertainty make sense conditional on data

  $\Rightarrow$ Even optimal frequentist confidence intervals sometimes do not (Fisher (1956), Buehler (1959), Wallace (1959), Cornfield (1969), ..., Müller and Norets (2016))

Major Practical Concern: How to Specify the Likelihood?

- Linear regression: Need to specify the distribution of the disturbances conditional on the regressors
- In practice, Bayesian inference typically based on convenient parametric assumptions (say, student-t regression) that seems a reasonable fit to data
- Concern about small/local misspecification

  $\Rightarrow$ Local misspecification leads in general to bias of same order as posterior standard deviation

  $\Rightarrow$ Potentially highly misleading posterior inference unless sure that parametric assumption is correct

This Paper

- "Augment" parametric model in a way that eliminates asymptotic bias under local misspecification

  $\Rightarrow$ Requires that parameter of interest has interpretation in encompassing semiparametric model

  $\Rightarrow$ Augmented model has additional parameter δ of the same dimension as the parameter of interest

  $\Rightarrow$ Form of augmentation intimately linked to theory of semiparametric efficient estimation

  $\Rightarrow$ Augmented model is a model, so continue to do real Bayes inference

- Theory: Posterior in augmented model is asymptotically Gaussian, centered on semiparametrically efficient estimator, with variance equal to semiparametric efficiency bound

- Practice: Generic suggestion for MCMC sampler for augmented model (not obvious, as it involves an intractable constant of integration)

Alternative Approaches

1. Limited Information Bayes

   - Example: Treat OLS estimator and its approximate normal distribution as (only) observation, proceed to be Bayesian with that normal likelihood

   - Often good idea, but lose coherency of multiple objectives and requires buying into (frequenitist) asymptotic approximations

2. Semiparametric/Nonparametric Bayes

   - Difficult to do

   - Typically not known whether semiparametric efficiency bound is achieved, so still concern about misspecification

Outline of Talk

1. Theory
2. Implementation
3. Empirical Illustration
4. Conclusion

Standard Theory for Baseline Model

- $Y_i \sim iid\mathbb{P}_\theta$ with density $p_\theta(\cdot|\theta)$, scalar parameter of interest is θ (for expositional ease; see paper for theory with nuisance parameters), prior $\pi(\theta)$

- Standard results:

$$\text{MLE} \quad : \quad \sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} I_\theta^{-1} \sum_{i=1}^{n} \dot{\ell}_\theta(Y_i) + o_{\mathbb{P}_\theta}(1) \Rightarrow_\theta \mathcal{N}(0, I_\theta^{-1})$$

$$\text{BvM} \quad : \quad \sqrt{n}(\theta - \hat{\theta})|\{Y_i\}_{i=1}^n \overset{a}{\sim} \mathcal{N}(0, I_\theta^{-1})$$

  where $\dot{\ell}_\theta(Y_i) = \partial \log p(Y_i|\theta)/\partial\theta$ and $I_\theta = \mathbb{E}[\dot{\ell}_\theta \dot{\ell}_\theta']$

- Example: inference for mean θ of student-t observations with known degrees of freedom ν and known scale $\sigma = 1 - 2/\nu$ (so $\mathbb{E}[Y_i^2] = 1$)

$$\dot{\ell}_\theta(y) = \frac{(\nu + 1)(y - \theta)}{\nu - 2 + (y - \theta)^2}$$

Local Misspecification

- Consider semiparametric model $\mathbb{P}_{\theta,\eta}$ where $\eta \in H$ is nonparametric and $\mathbb{P}_\theta = \mathbb{P}_{\theta,\eta_0}$

  $\Rightarrow$ Importantly, parameter of interest in $\mathbb{P}_{\theta,\eta}$ remains θ

  $\Rightarrow$ In example: $Y_i \sim iid\eta(\theta, 1)$

- Smooth one-dimensional path $\eta_t \in H$ for $t \in [0, \infty)$ characterized by score

$$\log \prod_{i=1}^{n} \frac{d\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}}{d\mathbb{P}_\theta}(Y_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(Y_i) - \tfrac{1}{2}\mathbb{E}[g(Y_i)^2] + o_{\mathbb{P}_\theta}(1)$$

- Consider effect of such local misspecification $\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}$ on $\hat{\theta}$ (and thus posterior of θ)

  $\Rightarrow$ Cannot be detected from data with probability one, even asymptotically

Local Misspecification II

- Under $\mathbb{P}_\theta$

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}-\theta) \\ \log \prod_{i=1}^n \frac{d\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}}{\mathbb{P}_\theta}(Y_i) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} I_\theta^{-1} \sum_{i=1}^n \dot{\ell}_\theta(Y_i) \\ \frac{1}{\sqrt{n}} \sum_{i=1}^n g(Y_i) - \frac{1}{2}\mathbb{E}[g(Y_i)^2] \end{pmatrix} + o_{\mathbb{P}_\theta}(1)$$

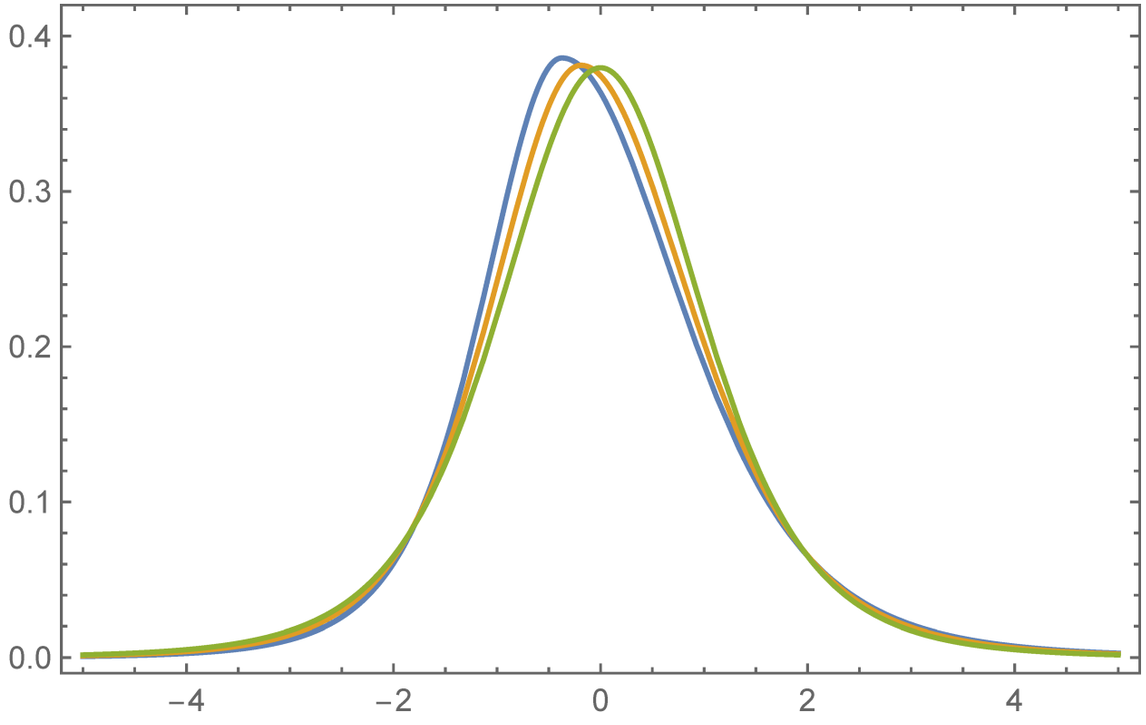
$$\Rightarrow \; {}_\theta\mathcal{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}\mathbb{E}[g(Y_i)^2] \end{pmatrix}, \begin{pmatrix} I_\theta^{-1} & \cdot \\ I_\theta^{-1}\mathbb{E}[\dot{\ell}_\theta(Y_i)g(Y_i)] & \mathbb{E}[g(Y_i)^2] \end{pmatrix} \right)$$

- Under $\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}$, by LeCam's Third Lemma

$$\sqrt{n}(\hat{\theta}-\theta) \Rightarrow_{\theta,\eta_{1/\sqrt{n}}} \mathcal{N}(I_\theta^{-1}\mathbb{E}[\dot{\ell}_\theta(Y_i)g(Y_i)], I_\theta^{-1})$$

$\Rightarrow$ MLE (and thus posterior center) biased unless $\mathbb{E}[\dot{\ell}_\theta(Y_i)g(Y_i)] = 0$

Example



Density of Hansen's (2004) skewed student-t distribution for $\nu = 5$ and skewness parameter $t \in \{0, 0.1, 0.2\}$

Here $\mathbb{E}[\dot{\ell}_\theta(Y_i)g(Y_i)] \neq 0$, so $t = O(n^{-1/2})$ induces local bias in location estimator of student-t model

Semiparametric Efficient Estimation

- Estimation θ in semiparametric model $\mathbb{P}_{\theta,\eta}$ at least as hard as in two-dimensional parametric model (θ, δ), where δ parametrizes particular smooth submodel of η

- By usual MLE expansion

$$\sqrt{n}\begin{pmatrix} \hat{\theta}^s - \theta \\ \hat{\delta} - \delta \end{pmatrix} = \frac{1}{\sqrt{n}} \begin{pmatrix} \mathbb{E}[\dot{\ell}_\theta(Y_i)^2] & \cdot \\ \mathbb{E}[\dot{\ell}_\theta(Y_i)\dot{\ell}_\delta(Y_i)] & \mathbb{E}[\dot{\ell}_\delta(Y_i)^2] \end{pmatrix}^{-1} \sum_{i=1}^n \begin{pmatrix} \dot{\ell}_\theta(Y_i) \\ \dot{\ell}_\delta(Y_i) \end{pmatrix} + o_{\mathbb{P}_\theta}(1)$$

  so

$$\sqrt{n}(\hat{\theta}^s - \theta) \Rightarrow_\theta \mathcal{N}(0, \mathbb{E}[\dot{\ell}_{\theta\perp\delta}(Y_i)^2]^{-1}), \quad \dot{\ell}_{\theta\perp\delta}(y) = \dot{\ell}_\theta(y) - \frac{\mathbb{E}[\dot{\ell}_\theta(Y_i)\dot{\ell}_\delta(Y_i)]}{\mathbb{E}[\dot{\ell}_\theta(Y_i)^2]}\dot{\ell}_\delta(y)$$

- *Least favorable submodel* maximizes $\mathbb{E}[\dot{\ell}_{\theta\perp\delta}(Y_i)^2]^{-1}$ and yields *efficient score* $\dot{\ell}_{\theta\perp\delta}(y) = \tilde{\ell}(y)$ that is orthogonal to $\dot{\ell}_\delta$ of all such one-parameter submodels. Corresponding nuisance score is

$$\dot{\ell}_\delta(y) = \dot{\ell}_\theta(y) - \tilde{\ell}_\theta(y)$$

  $\Rightarrow$ In example, $\tilde{\ell}(y) = y - \theta$, so $\dot{\ell}_\delta(y) = \frac{(\nu+1)(y-\theta)}{\nu-2+(y-\theta)^2} - (y - \theta)$

Model Augmentation

- Augmented model has density

$$q(y|\theta,\delta) = c(\theta,\delta)p(y|\theta)k_0(\delta\dot{\ell}_\delta(y))$$

where $c(\theta,\delta)^{-1} = \int p(y|\theta)k_0(\delta\dot{\ell}_\delta(y))dy$ and $k_0(0) = k_0'(0) = 1$, such as $k_0(z) = \frac{2}{1+e^{-2z}}$ (Example 26.16 in van der Vaart (1998))

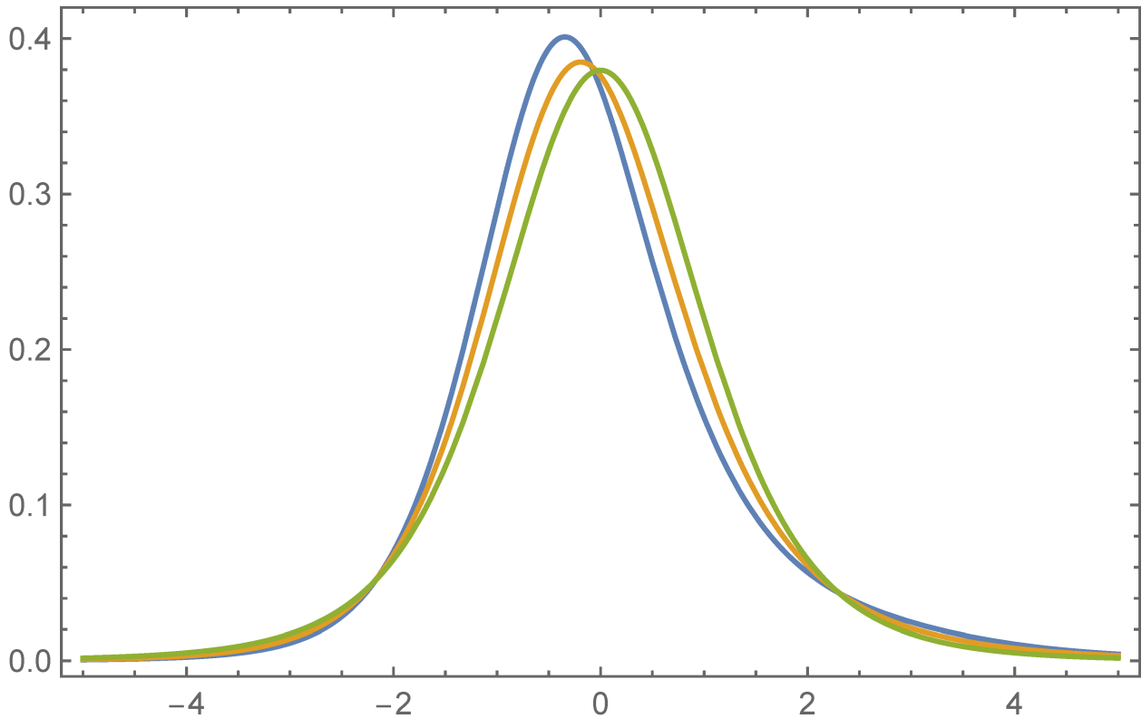
- This augmentation induces $\dot{\ell}_{\theta\perp\delta}(y) = \tilde{\ell}(y)$. MLE $\hat{\theta}^a$ immune to local misspecification, since $\mathbb{E}[\tilde{\ell}_\theta(Y_i)g(Y_i)] = 0$

- **Theorem:** Under regularity conditions, with d_{TV} the total variation distance and $\Pi^a(\theta|Y)$ the posterior distribution of θ in augmented model,

$$d_{TV}(\Pi(\theta|Y), \mathcal{N}(\hat{\theta}^a, n^{-1}\mathbb{E}[\tilde{\ell}(Y_i)^2]^{-1})) \overset{\mathbb{P}_{\theta,\eta_{1/\sqrt{n}}}}{\to} 0$$

- Implementation only requires knowledge of efficient score to construct $\dot{\ell}_\delta(y) = \dot{\ell}_\theta(y) - \tilde{\ell}_\theta(y)$

Augmented Model in Example



Augmented Student-t density for $\nu = 5$ and $\delta \in \{0, 0.4, 0.8\}$

MCMC Algorithm

- Posterior in augmented model is

$$\begin{aligned}\pi(\theta,\delta|Y) &\propto \pi(\theta,\delta)\prod_{i=1}^{n} q(Y_i|\theta,\delta) \\ &= \pi(\theta,\delta)\prod_{i=1}^{n} c(\theta,\delta)p(Y_i|\theta)\frac{2}{1+e^{-2\delta\dot{\ell}_\delta(Y_i)}} \qquad \dot{\ell}_\delta(y)=\dot{\ell}_\theta(y)-\tilde{\ell}_\theta(y)\end{aligned}$$

- Standard MCMC does not require $p(Y)$, but would require $c(\theta,\delta)$

- Following Rao, Lin and Dunson (2016), use auxiliary latent variables and acceptance sampling

  $\Rightarrow$ See paper for details

  $\Rightarrow$ Generic Matlab code for Hamiltonian Monte Carlo

Regression with Student-t Errors

- Model: $y_i = x_i'\beta + \varepsilon_i$, ε_i/σ has student-t density with ν degrees of freedoms

  $\Rightarrow$ (Subset of) β is parameter of interest, (σ, ν) are nuisance parameters

  $\Rightarrow$ Recommended for heavy tailed data in most Bayesian textbooks (Koop (2003), Geweke (2005), Greenberg (2012))

- OLS is semiparametrically efficient with efficient score

$$\tilde{\ell}_\gamma = \frac{x_i(y_i - x_i'\beta)}{\mathsf{Var}[\varepsilon_i]}$$

- In Monte Carlo, set $x_1 = 1$, $x_2 \sim iid\mathcal{N}(0,1)$, $\nu = 2.5$ and $n = 1000$, independent Gaussian priors on $\beta, \delta, \ln\sigma, \ln(\nu - 2)$

- Consider performance under local misspecification: true distribution is skewed student-t

MC Averages in Student-t Regression

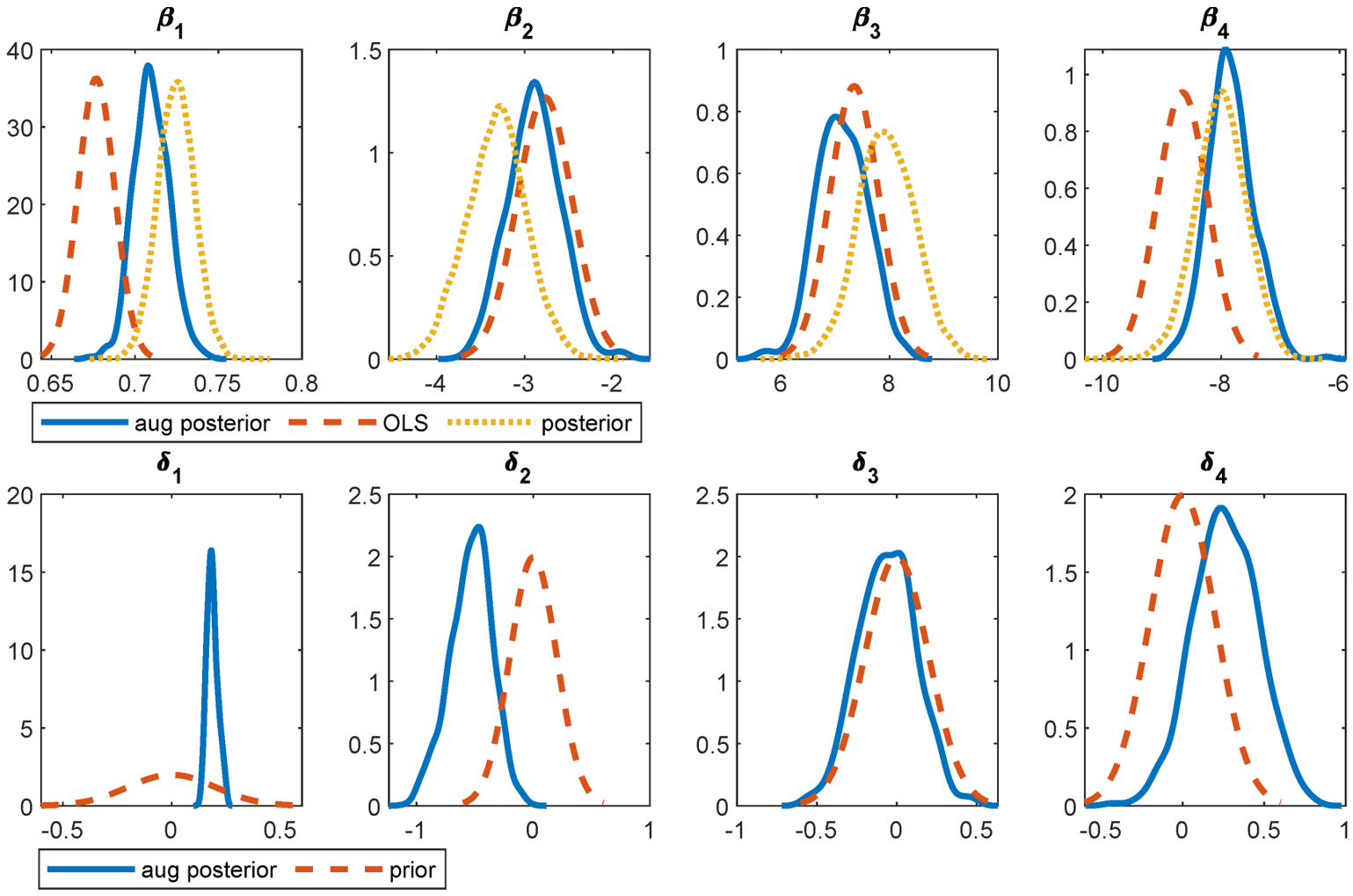
	$t = 0$		$t = n^{-1/2}x_{i2}$		$t = 5n^{-1/2}x_{i2}$	
	base	aug.	base	aug.	base	aug.
$\lvert\hat{\beta}_1 - \mathbb{E}[\beta_1\lvert Y]\rvert$	0.018	0.012	0.022	0.016	0.022	0.015
$\lvert\hat{\beta}_2 - \mathbb{E}[\beta_2\lvert Y]\rvert$	0.019	0.014	0.025	0.014	0.060	0.018
$\mathsf{sd}(\beta_1\lvert Y)/\mathsf{sd}(\hat{\beta}_1)$	0.63	1.38	0.61	1.35	0.60	1.09
$\mathsf{sd}(\beta_1\lvert Y)/\mathsf{sd}(\hat{\beta}_1)$	0.65	1.34	0.62	1.24	0.64	1.12
$\mathbb{E}[\beta_1\lvert Y] - \beta_1$	0.000	-0.001	-0.001	0.002	0.003	0.003
$\mathbb{E}[\beta_2\lvert Y] - \beta_2$	0.001	0.003	-0.011	0.004	-0.061	-0.003

Notes: t is skewness parameter. Results averaged over 100 data sets.

Empirical Illustration

- Jackman (2000) uses Bayesian student-t regression to study the degree of incumbency advantage
- y_i : proportion of two party vote for Democrat in district i
- x_i : proportion of two party vote for Democrat in previous election, previous winning party, indicators for Democratic and Republican incumbency, 20 dummy variables for time effects
- $n = 5090$ observations

Posterior Distributions of Baseline and Augmented Model



Implied Posterior Probability of Democrat Candidate Winning

Prev Dem Vote share	base	aug.	Gaussian
31	0.7	1.3	0.3
33	0.8	1.9	0.5
35	1.3	2.2	0.9
37	1.9	2.3	1.5
39	2.7	3.2	2.5

Notes: Conditional on previous winner is a Republican, Democratic candidate is not incumbent, Republican candidate is incumbent. In percent.

Summary

- Augmented model is generalization of baseline model
- It is parametric—easier to estimate and interpret
- Inference is fully Bayesian
- Semiparametric efficient and robust to local misspecification

Thank you!