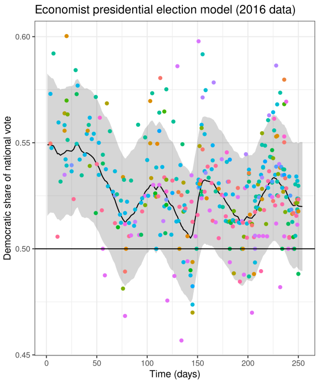
Approximate data deletion and replication with the Bayesian influence function

Ryan Giordano (rgiordano@berkeley.edu, UC Berkeley), Tamara Broderick (MIT)

Duke Statistics Seminar Oct 2024

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

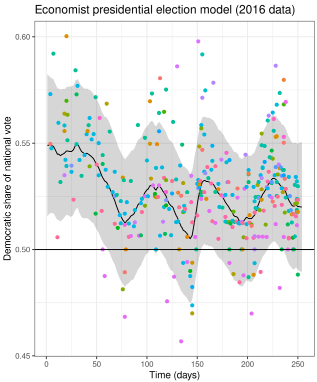
Model:

- $X = x_1, \ldots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \ldots, x_N$ = Polling data ($N = 361$).
- θ = Lots of random effects (day, pollster, etc.)
- $f(\theta)$ = Democratic % of vote on election day

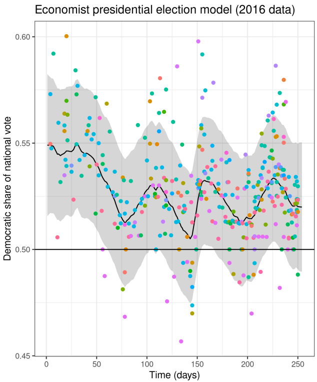
Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

The people who responded to the polls were randomly selected.
If we had selected a different random sample, how much would our estimate have changed?

Idea: Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \ldots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

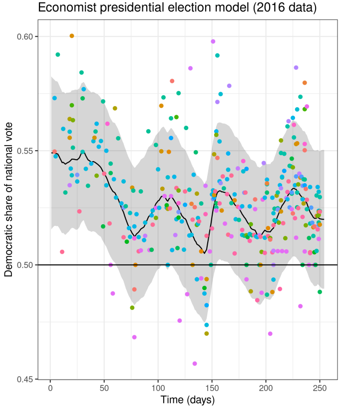
The people who responded to the polls were randomly selected.
If we had selected a different random sample, how much would our estimate have changed?

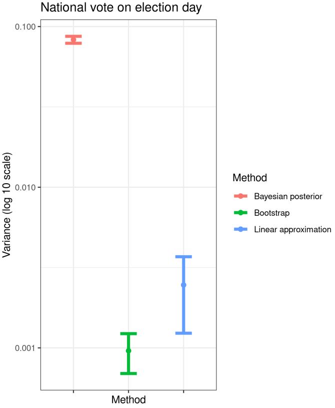
Idea: Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

Problem: Each MCMC run takes about 10 hours (Stan, six cores).

Results

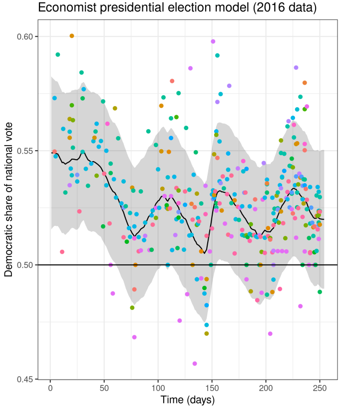
Proposal: Use full–data posterior draws to form a linear approximation to *data reweightings*.

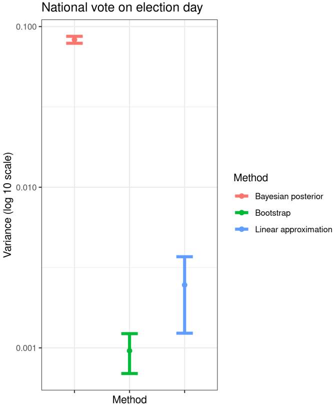




Results

Proposal: Use full–data posterior draws to form a linear approximation to *data reweightings*.





Compute time for 100 bootstraps: 51 days

Compute time for the linear approximation: Seconds
(But note the approximation has some error)

Outline

- Data reweighting
    - Write the change in the posterior expectation as linear component + error
    - The linear component can be computed from a single run of MCMC

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component $+$ error
  - The linear component can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
  - As $N \to \infty$, the linear component provides an arbitrarily good approximation
  - Consistent variance estimates (the errors vanish in aggregate)

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component + error
  - The linear component can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
  - As $N \to \infty$, the linear component provides an arbitrarily good approximation
  - Consistent variance estimates (the errors vanish in aggregate)
- High-dimensional problems
  - The linear component is the same order as the error
  - Even for parameters which concentrate *a posteriori*, even as $N \to \infty$
  - Argue that the variance estimates are inconsistent, but maybe not too bad in practice

Outline

- Data reweighting
  - Write the change in the posterior expectation as linear component $+$ error
  - The linear component can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
  - As $N \to \infty$, the linear component provides an arbitrarily good approximation
  - Consistent variance estimates (the errors vanish in aggregate)
- High-dimensional problems
  - The linear component is the same order as the error
  - Even for parameters which concentrate *a posteriori*, even as $N \to \infty$
  - Argue that the variance estimates are inconsistent, but maybe not too bad in practice
- Some implications and future work

Data re-weighting.

Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)]$.

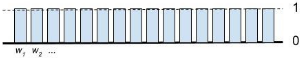
$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta; w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

Data re-weighting.

Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta; w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

Original weights:

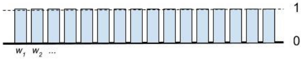


Data re-weighting.

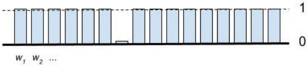
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta; w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

Original weights:



Leave-one-out weights:

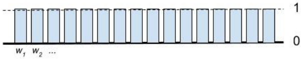


Data re-weighting.

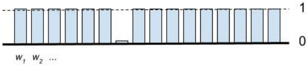
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta;w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

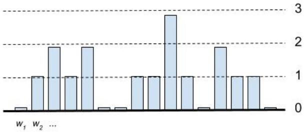
Original weights:



Leave-one-out weights:



Bootstrap weights:

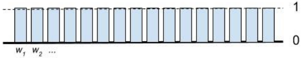


Data re-weighting.

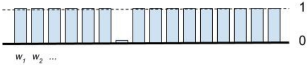
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta; w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

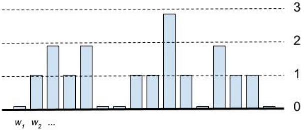
Original weights:

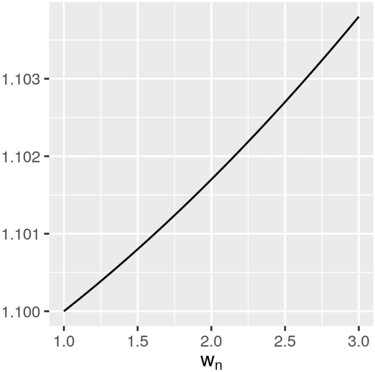


Leave-one-out weights:



Bootstrap weights:



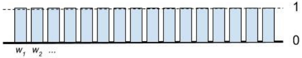


Data re-weighting.

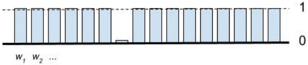
Augment the problem with *data weights* $w_1, \dots, w_N$. We can write $\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta; w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

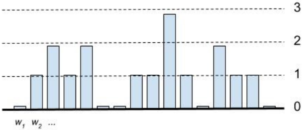
Original weights:

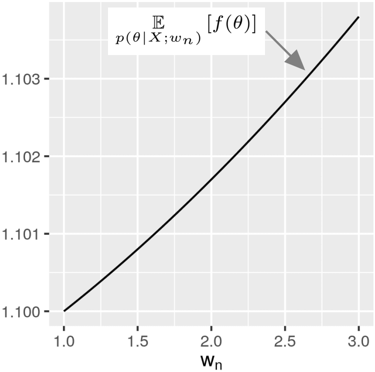


Leave-one-out weights:



Bootstrap weights:



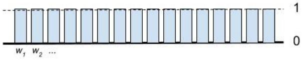


Data re-weighting.

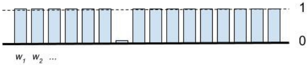
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta;w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

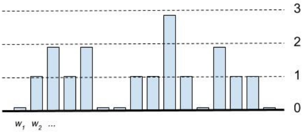
Original weights:

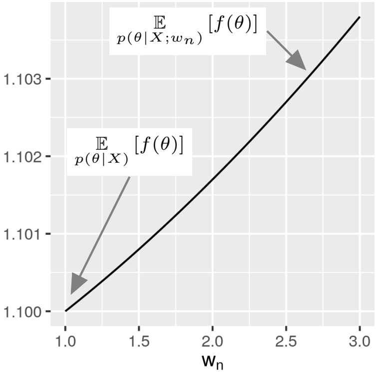


Leave-one-out weights:



Bootstrap weights:



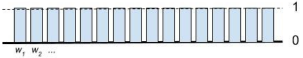


Data re-weighting.

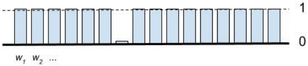
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta;w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

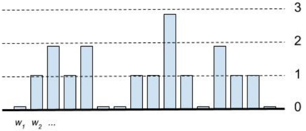
Original weights:

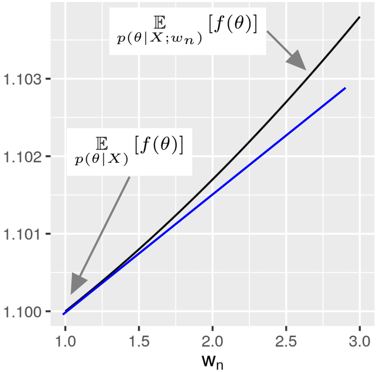


Leave-one-out weights:



Bootstrap weights:



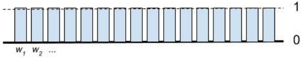


Data re-weighting.

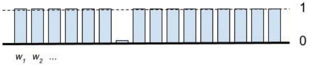
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\mathbb{E}_{p(\theta|X;w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta; w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

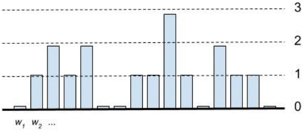
Original weights:

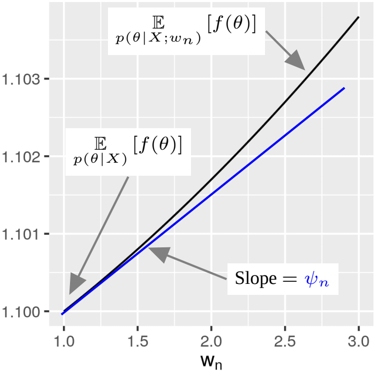


Leave-one-out weights:



Bootstrap weights:

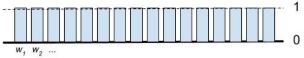




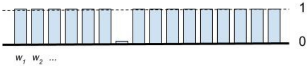
Augment the problem with *data weights* $w_1, \dots, w_N$. We can write $\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta;w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

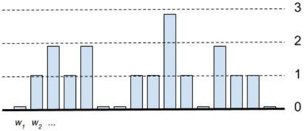
Original weights:

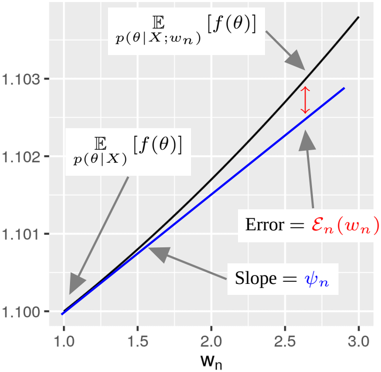


Leave-one-out weights:



Bootstrap weights:



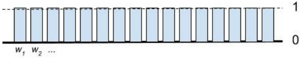


Data re-weighting.

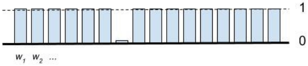
Augment the problem with *data weights* $w_1, \ldots, w_N$. We can write $\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad\qquad \log p(X|\theta; w) = \sum_{n=1}^{N} w_n \ell_n(\theta)$$

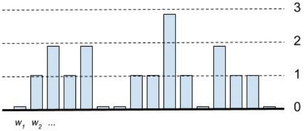
Original weights:

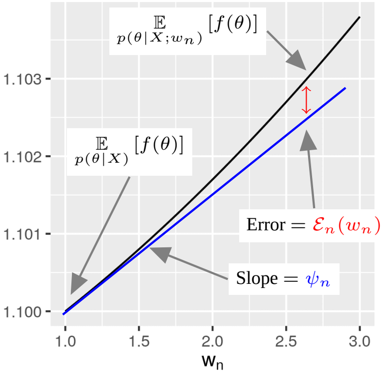


Leave-one-out weights:



Bootstrap weights:





The re-scaled slope $N\psi_n$ is known as the “influence function” at data point x_n.

$$\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}_n(w)$$

Data re-weighting.

How can we use the approximation?

Assume the slope is computable and error is small.

$$\mathop{\mathbb{E}}_{p(\theta|X;w)}[f(\theta)] - \mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}_n(w)$$

Data re-weighting.

How can we use the approximation?

Assume the slope is computable and error is small.

$$\mathop{\mathbb{E}}_{p(\theta|X;w)}[f(\theta)] - \mathop{\mathbb{E}}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}_n(w)$$

Bootstrap. Draw bootstrap weights $w \sim p(w) = \text{Multinomial}(N, N^{-1})$.

$$\text{Bootstrap variance} = \mathop{\text{Var}}_{p(w)}\left(\mathop{\mathbb{E}}_{p(\theta|X;w)}[f(\theta)]\right)$$

Data re-weighting.

How can we use the approximation?

Assume the slope is computable and error is small.

$$\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}_n(w)$$

Bootstrap. Draw bootstrap weights $w \sim p(w) = \text{Multinomial}(N, N^{-1})$.

$$\begin{aligned}
\text{Bootstrap variance} &= \underset{p(w)}{\text{Var}} \left(\underset{p(\theta|X;w)}{\mathbb{E}}[f(\theta)] \right) \\
&= \underset{p(w)}{\text{Var}} \left(\sum_{n=1}^{N} \psi_n (w_n - 1) + \mathcal{E}_n(w) \right) \\
&= \frac{1}{N^2} \sum_{n=1}^{N} \left(\psi_n - \overline{\psi} \right)^2 + \text{Term involving } \mathcal{E}_n(w) \text{ for } n = 1, \ldots, N \\
&\approx \frac{1}{N} \underbrace{\left(\frac{1}{N} \sum_{n=1}^{N} \left(\psi_n - \overline{\psi} \right)^2 \right)}_{\text{"Infinitesimal jackknife variance estimate"}}
\end{aligned}$$

Expressions for the slope and error

How to compute the slopes ψ_n? How large is the error $\mathcal{E}(w)$?

For simplicity, let us consider a single weight for the moment.

$$\underset{p(\theta|X;w_n)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}_n(w)$$

Expressions for the slope and error

How to compute the slopes ψ_n? How large is the error $\mathcal{E}(w)$?

For simplicity, let us consider a single weight for the moment.

$$\underset{p(\theta|X;w_n)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}_n(w)$$

Let an overbar denote "posterior–mean zero." For example, $\bar{f}(\theta) := f(\theta) - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\psi_n = \underbrace{\underset{p(\theta|X)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Estimatable with MCMC!}} \qquad \mathcal{E}_n(w) = \frac{1}{2}\underbrace{\underset{p(\theta|X;\tilde{w}_n)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Cannot compute directly (don't know } \tilde{w})}(w_n - 1)^2$$

Expressions for the slope and error

How to compute the slopes ψ_n? How large is the error $\mathcal{E}(w)$?

For simplicity, let us consider a single weight for the moment.

$$\underset{p(\theta|X;w_n)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}_n(w)$$

Let an overbar denote "posterior–mean zero." For example, $\bar{f}(\theta) := f(\theta) - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\psi_n = \underbrace{\underset{p(\theta|X)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Estimatable with MCMC!}} \qquad \mathcal{E}_n(w) = \frac{1}{2}\underbrace{\underset{p(\theta|X;\tilde{w}_n)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Cannot compute directly (don't know } \tilde{w}\text{)}}(w_n - 1)^2$$

$= O_p(N^{-1})$ under posterior concentration

$= O_p(N^{-2})$ under posterior concentration

Expressions for the slope and error

How to compute the slopes ψ_n? How large is the error $\mathcal{E}(w)$?

For simplicity, let us consider a single weight for the moment.

$$\underset{p(\theta|X;w_n)}{\mathbb{E}}[f(\theta)] - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}_n(w)$$

Let an overbar denote "posterior–mean zero." For example, $\bar{f}(\theta) := f(\theta) - \underset{p(\theta|X)}{\mathbb{E}}[f(\theta)]$.

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\psi_n = \underbrace{\underset{p(\theta|X)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Estimatable with MCMC!}} \qquad \mathcal{E}_n(w) = \frac{1}{2}\underbrace{\underset{p(\theta|X;\tilde{w}_n)}{\mathbb{E}}\left[\bar{f}(\theta)\bar{\ell}_n(\theta)\bar{\ell}_n(\theta)\right]}_{\text{Cannot compute directly (don't know } \tilde{w})}(w_n - 1)^2$$

$= O_p(N^{-1})$ under posterior concentration $\qquad = O_p(N^{-2})$ under posterior concentration

The scaling $O_p(N^{-2})$ for the error is classical for a *particular weight* [Kass et al., 1990].

For variance estimation, we need (and prove) conditions under which the $O_p(N^{-2})$ scaling applies sufficiently uniformly in *all the weights*.

Variance consistency theorem

How do the results for a single weight translate into variance estimates?

$$\operatorname*{Var}_{p(w)}\left(\mathop{\mathbb{E}}_{p(\theta|X,w)}[f(\theta)]\right) = \frac{1}{N^2}\sum_{n=1}^{N}\left(\psi_n - \overline{\psi}\right)^2 + \text{Term involving } \mathcal{E}_n(w) \text{ for } n = 1, \ldots, N$$

Variance consistency theorem

How do the results for a single weight translate into variance estimates?

$$\underset{p(w)}{\mathrm{Var}}\left(\underset{p(\theta|X,w)}{\mathbb{E}}\left[f(\theta)\right]\right) = \frac{1}{N^2}\sum_{n=1}^{N}\left(\psi_n - \overline{\psi}\right)^2 + \text{Term involving } \mathcal{E}_n(w) \text{ for } n = 1, \ldots, N$$

- Assume (sketch): A well–behaved MAP *maximum a posteriori* estimator $\hat{\theta}$ exists.
  - The dimension of θ is fixed as $N \to \infty$
  - The expected log likelihood has a strict maximum at θ_∞
  - The observed log likelihood statisfies $\hat{\theta} \to \theta_\infty$
  - The expected log likelihood Hessian is negative definite at θ_∞
- Assume (sketch): We can apply standard asymptotics.
  - The log prior and log likelihood are four times continuously differentiable
  - The prior is proper, and a technical set of prior expectations are finite
  - The log likelihood and its derivatives are dominated by a square–integrable envelope function for all θ in a neighborhood of θ_∞.

Variance consistency theorem

How do the results for a single weight translate into variance estimates?

$$\underset{p(w)}{\mathrm{Var}} \left(\underset{p(\theta|X,w)}{\mathbb{E}} [f(\theta)] \right) = \frac{1}{N^2} \sum_{n=1}^{N} \left(\psi_n - \overline{\psi} \right)^2 + \text{Term involving } \mathcal{E}_n(w) \text{ for } n = 1, \ldots, N$$

- Assume (sketch): A well–behaved MAP *maximum a posteriori* estimator $\hat{\theta}$ exists.
  - The dimension of θ is fixed as $N \to \infty$
  - The expected log likelihood has a strict maximum at θ_∞
  - The observed log likelihood statisfies $\hat{\theta} \to \theta_\infty$
  - The expected log likelihood Hessian is negative definite at θ_∞
- Assume (sketch): We can apply standard asymptotics.
  - The log prior and log likelihood are four times continuously differentiable
  - The prior is proper, and a technical set of prior expectations are finite
  - The log likelihood and its derivatives are dominated by a square–integrable envelope function for all θ in a neighborhood of θ_∞.

Theorem 2 [Giordano and Broderick, 2023]: Under the above assumptions,

$$\sqrt{N} \left(\underset{p(\theta|X)}{\mathbb{E}} [g(\theta)] - g(\theta_\infty) \right) \xrightarrow[N \to \infty]{dist} \mathcal{N}\left(0, V^g\right) \quad \text{[Kleijn and Van der Vaart, 2012]}$$

and $$V^{\mathrm{IJ}} := \frac{1}{N} \sum_{n=1}^{N} \left(\psi_n - \overline{\psi} \right)^2 \xrightarrow[N \to \infty]{prob} V^g. \quad \text{(Our contribution)}$$

Negative binomial experiment

Example: Negative binomial models with an unknown parameter γ.

For $n = 1, \ldots, N$ let $x_n|\gamma \overset{iid}{\sim} \text{NegativeBinomial}\,(r, \gamma)$ for fixed r.

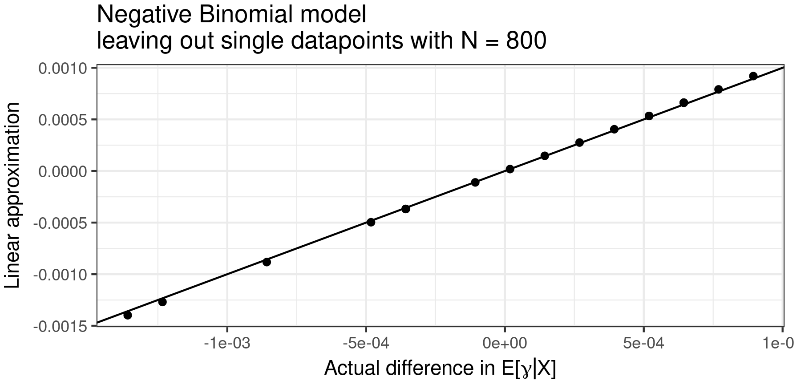
$$\text{Write } \log p(X|\gamma, w) = \sum_{n=1}^{N} w_n \ell_n(\gamma).$$

Negative binomial experiment

Example: Negative binomial models with an unknown parameter γ.

For $n = 1, \ldots, N$ let $x_n | \gamma \overset{iid}{\sim} \mathrm{NegativeBinomial}(r, \gamma)$ for fixed r.

Write $\log p(X|\gamma, w) = \sum_{n=1}^{N} w_n \ell_n(\gamma)$.



Data Analysis Using Regression and Multilevel/Hierarchical Models.

We ran `rstanarm` on 56 different models on 13 different datasets from Gelman and Hill [2006], including Gaussian and logistic regression, fixed and mixed-effects models.

Across all models, we estimate 799 distinct covariances (regression coefficients and log scale parameters).

Using the bootstrap as ground truth, compute the relative errors:

$$\frac{V_{\mathrm{Bayes}} - V_{\mathrm{Boot}}}{|V_{\mathrm{Boot}}|} \quad \text{and} \quad \frac{V_{\mathrm{IJ}} - V_{\mathrm{Boot}}}{|V_{\mathrm{Boot}}|}.$$

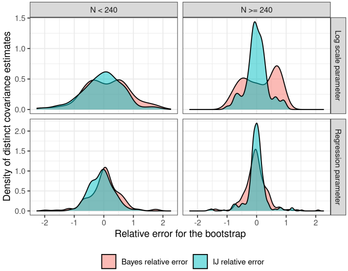


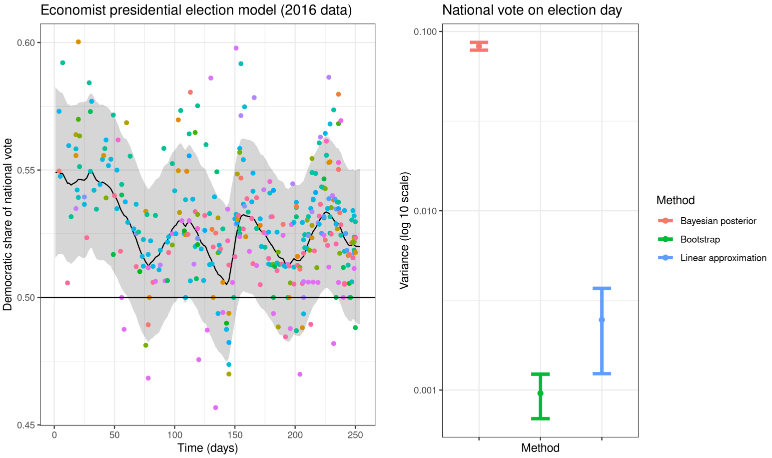
Figure 1: The distribution of the relative errors. Log scale parameters include all variances or covariances that involve at least one log scale parameters.

Total compute time for all models:

Initial fit:	1.6 hours
Bootstrap:	381.5 hours
Linear approximation:	A few minutes

How to connect to the election data?

Problem: MCMC is only interesting when the posterior doesn't concentrate.



High dimensional problems

Example: Exponential families with random effects (REs) λ and fixed effects γ.

If the observations per random effect remains bounded as $N \to \infty$, then

- Parameter λ (“local”) grows in dimension with N.
- Parameter γ (“global”) is finite-dimensional.
- Marginally $p(\lambda|X)$ does not concentrate.
- Marginally, $p(\gamma|X)$ concentrates.

High dimensional problems

Example: Exponential families with random effects (REs) λ and fixed effects γ.

If the observations per random effect remains bounded as $N \to \infty$, then

- Parameter λ ("local") grows in dimension with N.
- Parameter γ ("global") is finite-dimensional.
- Marginally $p(\lambda|X)$ does not concentrate.
- Marginally, $p(\gamma|X)$ concentrates.

In general, we cannot hope for an asymptotic analysis of $\underset{p(\lambda,\gamma|X)}{\mathbb{E}}[f(\lambda)]$.

Can we save the approximation when *some* parameters concentrate?

Does the residual vanish asymptotically for $w_n \mapsto \underset{p(\gamma|X;w_n)}{\mathbb{E}}[f(\gamma)]$?

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$\underset{p(\gamma,\lambda|X;w_n)}{\mathbb{E}}[\gamma] - \underset{p(\gamma,\lambda|X)}{\mathbb{E}}[\gamma] =$$

$$\psi_n(w_n - 1) \qquad + \mathcal{E}_n(w)$$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$\underset{p(\gamma,\lambda|X;w_n)}{\mathbb{E}}[\gamma] - \underset{p(\gamma,\lambda|X)}{\mathbb{E}}[\gamma] =$$

$$\psi_n(w_n - 1) \qquad + \mathcal{E}_n(w)$$

$$= \underset{p(\gamma,\lambda|X)}{\mathbb{E}}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)\right](w_n - 1) \qquad + \frac{1}{2}\underset{p(\gamma,\lambda|X;\tilde{w}_n)}{\mathbb{E}}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)^2\right](w_n - 1)^2$$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{p(\gamma,\lambda|X;w_n)}[\gamma] - \mathop{\mathbb{E}}_{p(\gamma,\lambda|X)}[\gamma] = \\
&\quad \psi_n(w_n - 1) && + \mathcal{E}_n(w) \\
&= \mathop{\mathbb{E}}_{p(\gamma,\lambda|X)}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)\right](w_n - 1) && + \frac{1}{2}\mathop{\mathbb{E}}_{p(\gamma,\lambda|X;\tilde{w}_n)}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)^2\right](w_n - 1)^2 \\
&= \mathop{\mathbb{E}}_{p(\gamma|X)}\Bigg[\bar{\gamma}\underbrace{\mathop{\mathbb{E}}_{p(\lambda|\gamma,X)}\left[\bar{\ell}_n(\gamma,\lambda)\right]}_{F_1(\gamma)}\Bigg](w_n - 1) && + \frac{1}{2}\mathop{\mathbb{E}}_{p(\gamma|X;\tilde{w}_n)}\Bigg[\bar{\gamma}\underbrace{\mathop{\mathbb{E}}_{p(\lambda|X,\gamma;\tilde{w}_n)}\left[\bar{\ell}_n(\gamma,\lambda)^2\right]}_{F_2(\gamma)}\Bigg](w_n - 1
\end{aligned}
$$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{p(\gamma,\lambda|X;w_n)}[\gamma] - \mathop{\mathbb{E}}_{p(\gamma,\lambda|X)}[\gamma] = \\
&\quad \psi_n(w_n - 1) && + \mathcal{E}_n(w) \\
&= \mathop{\mathbb{E}}_{p(\gamma,\lambda|X)}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)\right](w_n - 1) && + \frac{1}{2}\mathop{\mathbb{E}}_{p(\gamma,\lambda|X;\tilde{w}_n)}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)^2\right](w_n - 1)^2 \\
&= \mathop{\mathbb{E}}_{p(\gamma|X)}\left[\bar{\gamma}\underbrace{\mathop{\mathbb{E}}_{p(\lambda|\gamma,X)}\left[\bar{\ell}_n(\gamma,\lambda)\right]}_{F_1(\gamma)}\right](w_n - 1) && + \frac{1}{2}\mathop{\mathbb{E}}_{p(\gamma|X;\tilde{w}_n)}\left[\bar{\gamma}\underbrace{\mathop{\mathbb{E}}_{p(\lambda|X,\gamma;\tilde{w}_n)}\left[\bar{\ell}_n(\gamma,\lambda)^2\right]}_{F_2(\gamma)}\right](w_n - 1 \\
&= \underbrace{\mathop{\mathbb{E}}_{p(\gamma|X)}\left[\bar{\gamma}F_1(\gamma)\right](w_n - 1)}_{\substack{O_p(N^{-1}) \\ \text{(by } p(\gamma|X) \text{ concentration)}}} && + \frac{1}{2}\underbrace{\mathop{\mathbb{E}}_{p(\gamma|X;\tilde{w}_n)}\left[\bar{\gamma}F_2(\gamma)\right](w_n - 1)^2}_{\substack{O_p(N^{-1}) \\ \text{(by } p(\gamma|X) \text{ concentration)}}}
\end{aligned}
$$

$\Rightarrow$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$\underset{p(\gamma,\lambda|X;w_n)}{\mathbb{E}}[\gamma] - \underset{p(\gamma,\lambda|X)}{\mathbb{E}}[\gamma] =$$

$$\psi_n(w_n - 1) \qquad + \mathcal{E}_n(w)$$

$$= \underset{p(\gamma,\lambda|X)}{\mathbb{E}}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)\right](w_n - 1) \qquad + \frac{1}{2}\underset{p(\gamma,\lambda|X;\tilde{w}_n)}{\mathbb{E}}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)^2\right](w_n - 1)^2$$

$$= \underset{p(\gamma|X)}{\mathbb{E}}\left[\bar{\gamma}\underbrace{\underset{p(\lambda|\gamma,X)}{\mathbb{E}}\left[\bar{\ell}_n(\gamma,\lambda)\right]}_{F_1(\gamma)}\right](w_n - 1) \qquad + \frac{1}{2}\underset{p(\gamma|X;\tilde{w}_n)}{\mathbb{E}}\left[\bar{\gamma}\underbrace{\underset{p(\lambda|X,\gamma;\tilde{w}_n)}{\mathbb{E}}\left[\bar{\ell}_n(\gamma,\lambda)^2\right]}_{F_2(\gamma)}\right](w_n - 1$$

$$= \underbrace{\underset{p(\gamma|X)}{\mathbb{E}}\left[\bar{\gamma}F_1(\gamma)\right](w_n - 1)}_{\substack{O_p(N^{-1}) \\ \text{(by } p(\gamma|X) \text{ concentration)}}} \qquad + \frac{1}{2}\underbrace{\underset{p(\gamma|X;\tilde{w}_n)}{\mathbb{E}}\left[\bar{\gamma}F_2(\gamma)\right]}_{\substack{O_p(N^{-1}) \\ \text{(by } p(\gamma|X) \text{ concentration)}}}(w_n - 1)^2$$

$$\Rightarrow \psi_n = O_p(N^{-1}) \qquad \mathcal{E}_n(w) = O_p(N^{-1})$$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{p(\gamma,\lambda|X;w_n)}[\gamma] - \mathop{\mathbb{E}}_{p(\gamma,\lambda|X)}[\gamma] = \\
&\quad \psi_n(w_n - 1) && + \mathcal{E}_n(w) \\
&= \mathop{\mathbb{E}}_{p(\gamma,\lambda|X)}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)\right](w_n-1) && + \frac{1}{2}\mathop{\mathbb{E}}_{p(\gamma,\lambda|X;\tilde{w}_n)}\left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)^2\right](w_n-1)^2 \\
&= \mathop{\mathbb{E}}_{p(\gamma|X)}\Bigg[\bar{\gamma}\underbrace{\mathop{\mathbb{E}}_{p(\lambda|\gamma,X)}\left[\bar{\ell}_n(\gamma,\lambda)\right]}_{F_1(\gamma)}\Bigg](w_n-1) && + \frac{1}{2}\mathop{\mathbb{E}}_{p(\gamma|X;\tilde{w}_n)}\Bigg[\bar{\gamma}\underbrace{\mathop{\mathbb{E}}_{p(\lambda|X,\gamma;\tilde{w}_n)}\left[\bar{\ell}_n(\gamma,\lambda)^2\right]}_{F_2(\gamma)}\Bigg](w_n-1 \\
&= \underbrace{\mathop{\mathbb{E}}_{p(\gamma|X)}\left[\bar{\gamma}F_1(\gamma)\right]}_{\substack{O_p(N^{-1}) \\ \text{(by } p(\gamma|X) \text{ concentration)}}}(w_n-1) && + \frac{1}{2}\underbrace{\mathop{\mathbb{E}}_{p(\gamma|X;\tilde{w}_n)}\left[\bar{\gamma}F_2(\gamma)\right]}_{\substack{O_p(N^{-1}) \\ \text{(by } p(\gamma|X) \text{ concentration)}}}(w_n-1)^2 \\
&\quad \Rightarrow \psi_n = O_p(N^{-1}) && \mathcal{E}_n(w) = O_p(N^{-1})
\end{aligned}
$$

Corollary [Giordano and Broderick, 2023]:

In general, $w_n \mapsto N\left(\mathop{\mathbb{E}}_{p(\gamma|X;w_n)}[\gamma] - \mathop{\mathbb{E}}_{p(\gamma|X)}[\gamma]\right)$ remains non-linear as $N \to \infty$.

Bayesian von–Mises Expansion

How can we apply the single–weight result to variance computations?

Bayesian von–Mises Expansion

How can we apply the single–weight result to variance computations?

Define the "generalized posterior" functional (θ may be growing in dimension):

$$T(\mathbb{G}, N) := \frac{\int g(\theta) \exp\left(N \int \ell(x_0|\theta)\mathbb{G}(dx_0)\right) \pi(\theta) d\theta}{\int \exp\left(N \int \ell(x_0|\theta)\mathbb{G}(dx_0)\right) \pi(\theta) d\theta}.$$

Let $\mathbb{F}_N$ denote the empirical distribution over x_n. Then

$$\mathop{\mathbb{E}}_{p(\theta|X)} [g(\theta)] = \frac{\int g(\theta) \exp\left(N \frac{1}{N} \sum_{n=1}^N \ell(x_n|\theta)\right) \pi(\theta) d\theta}{\int \exp\left(N \frac{1}{N} \sum_{n=1}^N \ell(x_n|\theta)\right) \pi(\theta) d\theta} = T(\mathbb{F}_N, N).$$

Bayesian von–Mises Expansion

How can we apply the single–weight result to variance computations?

Define the "generalized posterior" functional (θ may be growing in dimension):

$$T(\mathbb{G}, N) := \frac{\int g(\theta) \exp\left(N \int \ell(x_0|\theta)\mathbb{G}(dx_0)\right) \pi(\theta)d\theta}{\int \exp\left(N \int \ell(x_0|\theta)\mathbb{G}(dx_0)\right) \pi(\theta)d\theta}.$$

Let $\mathbb{F}_N$ denote the empirical distribution over x_n. Then

$$\mathop{\mathbb{E}}_{p(\theta|X)} [g(\theta)] = \frac{\int g(\theta) \exp\left(N \frac{1}{N} \sum_{n=1}^{N} \ell(x_n|\theta)\right) \pi(\theta)d\theta}{\int \exp\left(N \frac{1}{N} \sum_{n=1}^{N} \ell(x_n|\theta)\right) \pi(\theta)d\theta} = T(\mathbb{F}_N, N).$$

Let $\mathbb{F}$ denote the true distribution of x_n, and let $\mathbb{F}_N^t = t\mathbb{F}_N + (1-t)\mathbb{F}$.

We can study the *von Mises expansion*:

$$\sqrt{N}\left(\mathop{\mathbb{E}}_{p(\theta|X)} [g(\theta)] - T(\mathbb{F}, N)\right) = \sqrt{N} \left.\frac{\partial T(\mathbb{F}_N^t, N)}{\partial t}\right|_{t=0} (\mathbb{F}_N - \mathbb{F}) \qquad + \mathcal{E}(\tilde{t})$$

$$= \underbrace{\sqrt{N} \sum_{n=1}^{N} (\psi_n - \overline{\psi})}_{\text{Infinitesimal jackknife estimator}} + o_p(1) \qquad + \mathcal{E}(\tilde{t}).$$

Inconsistency is suggested if $\mathcal{E}(\tilde{t})$ fails to vanish.

Bayesian von–Mises Expansion Results

Theorem 3 [Giordano and Broderick, 2023] (sketch):
(Consistency of the von–Mises expansion in finite dimensions)
Under slightly stronger conditions our original finite–dimensional posterior consistency result,

$$\sup_{\tilde{t}\in[0,1]} |\mathcal{E}(\tilde{t})| \to 0 \quad \text{in the Bayesian von–Mises expansion.}$$

Bayesian von–Mises Expansion Results

Theorem 3 [Giordano and Broderick, 2023] (sketch):
(Consistency of the von–Mises expansion in finite dimensions)
Under slightly stronger conditions our original finite–dimensional posterior consistency result,

$$\sup_{\tilde{t}\in[0,1]} |\mathcal{E}(\tilde{t})| \to 0 \quad \text{in the Bayesian von–Mises expansion.}$$

Theorem 4 [Giordano and Broderick, 2023] (sketch, not yet on arxiv):
(Inconsistency of the von–Mises expansion in high dimensional exponential families)
Assume that x_n comes with a equiprobable group assignment $g_n \in 1, \ldots, G$.
Conditional on g, x_n is modeled as a finite-dimensional exponential family given λ, γ:

$$\log p(x_n | g_n = g, \gamma, \lambda) = \tau(x_n)^\intercal \eta_g(\gamma, \lambda) + \text{Constant}.$$

Define the average product of second moments:

$$\mathcal{V}_N(\gamma) := \frac{1}{G} \sum_{g=1}^{G} \text{tr}\left(\underset{\mathbb{F}(x_n)}{\mathbb{E}} \left[\tau(x_n)\tau(x_n)^\intercal\right] \underset{p(\lambda|\gamma,\mathbb{F})}{\text{Cov}} \left(\eta_g(\gamma, \lambda)\right) \right).$$

If $N \underset{p(\gamma|\mathbb{F})}{\mathbb{E}} \left[\bar{f}(\gamma)\mathcal{V}_N(\gamma)\right]$ is strictly bounded away from 0 as $N \to \infty$, then

$$\sup_{\tilde{t}\in[0,1]} |\mathcal{E}(\tilde{t})| \to \infty \quad \text{in the Bayesian von–Mises expansion.}$$

Experiments

Example: Poisson regression with Gamma-distributed random effects

$$\text{For } g = 1, \ldots, G,\ \lambda_g \stackrel{iid}{\sim} \text{Gamma}(\alpha, \beta) \text{ for fixed } \alpha, \beta$$

$$\text{For } n = 1, \ldots, N,\ g_n \stackrel{iid}{\sim} \text{Categorical}(1, \ldots, G),\ y_n | \lambda_n, \gamma, g_n \stackrel{iid}{\sim} \text{Poisson}(\gamma \lambda_{g_n}).$$

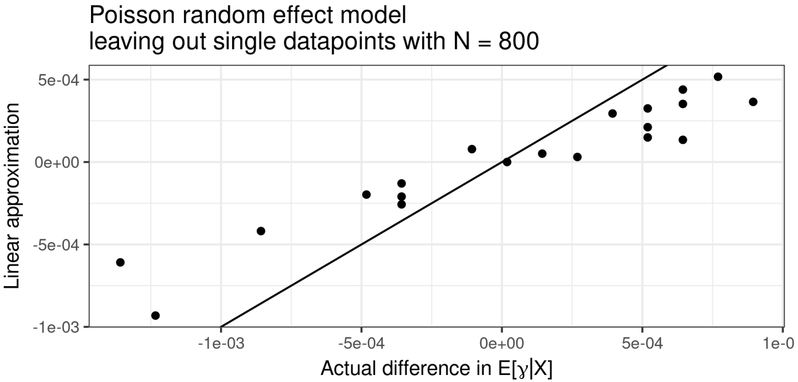
$x_n = (y_n, g_n)$ are IID given λ, γ. Write $\log p(X|\lambda, \gamma; w) = \sum_{n=1}^{N} w_n \ell_n(\lambda, \gamma)$.

Example: Poisson regression with Gamma-distributed random effects

For $g = 1, \ldots, G,\ \lambda_g \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$ for fixed α, β

For $n = 1, \ldots, N,\ g_n \overset{iid}{\sim} \text{Categorical}(1, \ldots, G),\ y_n | \lambda_n, \gamma, g_n \overset{iid}{\sim} \text{Poisson}(\gamma \lambda_{g_n})$.

$x_n = (y_n, g_n)$ are IID given λ, γ. Write $\log p(X | \lambda, \gamma; w) = \sum_{n=1}^{N} w_n \ell_n(\lambda, \gamma)$.



More experimental results for Gamma–Poisson mixtures

We ran simulations of the Gamma–Poisson mixture with different ratios of N/G (average observations per group).

- When N/G is small:
    - IJ is biased significantly downwards
    - Bootstrap is biased somewhat downwards
- When N/G is larger:
    - Both improve
    - Both remain somewhat biased
    - The IJ and bootstrap perform similarly

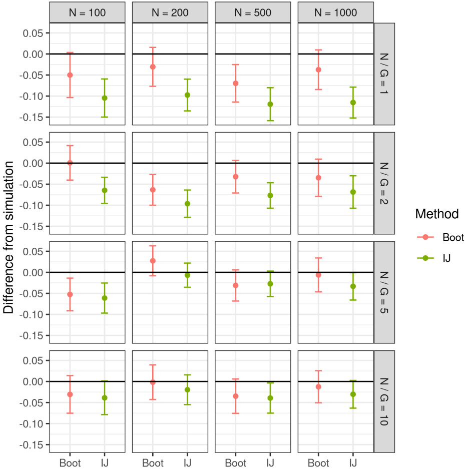


Figure 2: The error of the IJ and bootstrap covariances for different values of N and G. The y-axis shows the difference between $N(V - \hat{V}_{\text{sim}})$, where V is either $\hat{V}_{\text{IJ}}$ or $\hat{V}_{\text{Boot}}$.

Exchangeable units. (A contradiction?)

Negative binomial observations.

Asymptotically linear in w.

Poisson observations with random effects.

Asymptotically non-linear in w.

Exchangeable units. (A contradiction?)

Negative binomial observations.

Asymptotically linear in w.

Poisson observations with random effects.

Asymptotically non-linear in w.

With Gamma REs, one RE per observation, and appropriate prior parameters, these are the same model, with the same $p(\gamma|X)$.

Is $\underset{p(\gamma|X;w)}{\mathbb{E}}[\gamma]$ linear in the data weights or not?

Exchangeable units. (A contradiction?)

Negative binomial observations.

Asymptotically linear in w.

Poisson observations with random effects.

Asymptotically non-linear in w.

With Gamma REs, one RE per observation, and appropriate prior parameters, these are the same model, with the same $p(\gamma|X)$.

Is $\underset{p(\gamma|X;w)}{\mathbb{E}}[\gamma]$ linear in the data weights or not?

Exchangeable units. (A contradiction?)

Negative binomial observations.

Asymptotically linear in w.

Poisson observations with random effects.

Asymptotically non-linear in w.

With Gamma REs, one RE per observation, and appropriate prior parameters, these are the same model, with the same $p(\gamma|X)$.

Is $\underset{p(\gamma|X;w)}{\mathbb{E}}[\gamma]$ **linear in the data weights or not?**

Trick question! We weight a log likelihood contribution, not a datapoint.

$$\log p(X|\gamma; w^m) = \sum_{n=1}^{N} w_n^m \log p(x_n|\gamma) \qquad \log p(X|\gamma, \lambda; w^c) = \sum_{n=1}^{N} w_n^c \log p(x_n|\lambda, \gamma)$$

The two weightings are not equivalent in general.

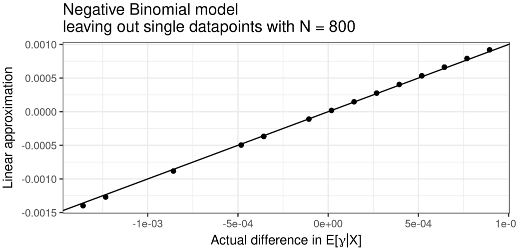
What is the right exchangeable unit for a particular problem?

Exchangeable units: Experimental results revisited

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

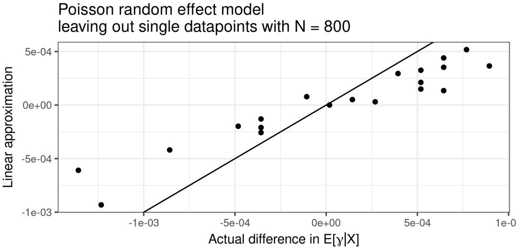
Uses $\log p(x_n|\gamma)$:

$$\psi_n = \mathop{\mathbb{E}}_{p(\gamma|X)} \left[\bar{\gamma}\bar{\ell}_n(\gamma)\right]$$



Uses $\log p(x_n|\gamma,\lambda)$:

$$\psi_n = \mathop{\mathbb{E}}_{p(\gamma,\lambda|X)} \left[\bar{\gamma}\bar{\ell}_n(\gamma,\lambda)\right]$$

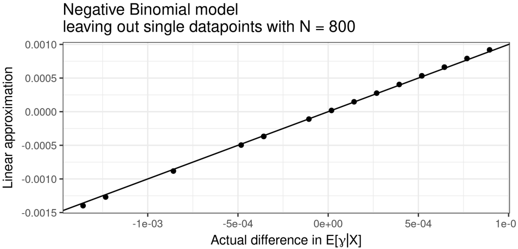


Exchangeable units: Experimental results revisited

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

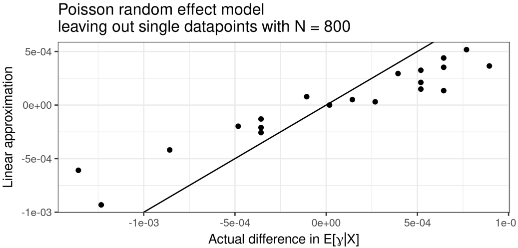
Uses $\log p(x_n|\gamma)$:
$\psi_n = \underset{p(\gamma|X)}{\mathbb{E}} \left[\bar{\gamma}\bar{\ell}_n(\gamma)\right]$

Not easily computable from
$\gamma, \lambda \sim p(\gamma, \lambda|X)$
in general.



Uses $\log p(x_n|\gamma, \lambda)$:
$\psi_n = \underset{p(\gamma,\lambda|X)}{\mathbb{E}} \left[\bar{\gamma}\bar{\ell}_n(\gamma, \lambda)\right]$

Easily computable from
$\gamma, \lambda \sim p(\gamma, \lambda|X)$.

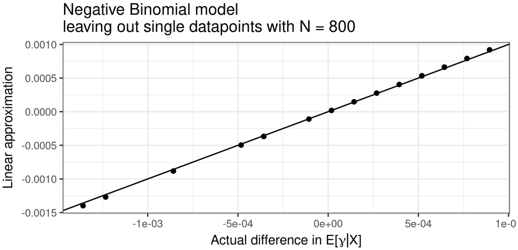


Exchangeable units: Experimental results revisited

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Uses $\log p(x_n|\gamma)$:
$$\psi_n = \mathop{\mathbb{E}}_{p(\gamma|X)}\left[\bar{\gamma}\bar{\ell}_n(\gamma)\right]$$

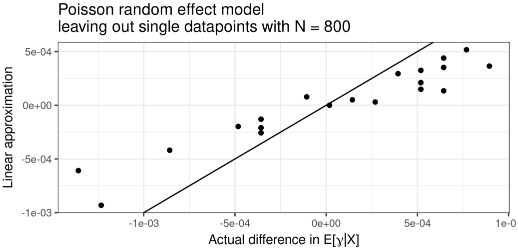
Not easily computable from $\gamma, \lambda \sim p(\gamma, \lambda|X)$ in general.



Uses $\log p(x_n|\gamma, \lambda)$:
$$\psi_n = \mathop{\mathbb{E}}_{p(\gamma,\lambda|X)}\left[\bar{\gamma}\bar{\ell}_n(\gamma, \lambda)\right]$$

Easily computable from $\gamma, \lambda \sim p(\gamma, \lambda|X)$.

May still be useful when $p(\lambda|X)$ is *somewhat* concentrated.



Observations and consequences

- For finite–dimensional models which concentrate asymptotically:
  - Posterior expectations are approximately linear in data weights
  - The linearized variance estimate (infinitesimal jackknife) is consistent
  - The residual of the von Mises expansion vanishes
- For high–dimensional models which marginally concentrate only asymptotically:
  - Posterior expectations are not approximately linear in data weights
  - The linearized variance estimate (infinitesimal jackknife) is inconsistent
  - The residual of the von Mises expansion does not vanish
  - Even if the error $\mathcal{E}(w)$ does not vanish, it can still be small enough in practice.
    ... Especially given the linear approximation's huge computational advantage.

Observations and consequences

- For finite–dimensional models which concentrate asymptotically:
  - Posterior expectations are approximately linear in data weights
  - The linearized variance estimate (infinitesimal jackknife) is consistent
  - The residual of the von Mises expansion vanishes
- For high–dimensional models which marginally concentrate only asymptotically:
  - Posterior expectations are not approximately linear in data weights
  - The linearized variance estimate (infinitesimal jackknife) is inconsistent
  - The residual of the von Mises expansion does not vanish
  - Even if the error $\mathcal{E}(w)$ does not vanish, it can still be small enough in practice.
    ... Especially given the linear approximation's huge computational advantage.

- When the weighting is linear, there are many other applications:
  - Cross-validation
  - Conformal inference
  - Identification of influential subsets
- When the weighting is non–linear, the inconsistency results should apply more widely:
  - The EM algorithm
  - The nonparametric bootstrap
  - Local prior sensitivity measures

Observations and consequences

- For finite–dimensional models which concentrate asymptotically:
  - Posterior expectations are approximately linear in data weights
  - The linearized variance estimate (infinitesimal jackknife) is consistent
  - The residual of the von Mises expansion vanishes
- For high–dimensional models which marginally concentrate only asymptotically:
  - Posterior expectations are not approximately linear in data weights
  - The linearized variance estimate (infinitesimal jackknife) is inconsistent
  - The residual of the von Mises expansion does not vanish
  - Even if the error $\mathcal{E}(w)$ does not vanish, it can still be small enough in practice.
    ... Especially given the linear approximation's huge computational advantage.

- When the weighting is linear, there are many other applications:
  - Cross-validation
  - Conformal inference
  - Identification of influential subsets
- When the weighting is non–linear, the inconsistency results should apply more widely:
  - The EM algorithm
  - The nonparametric bootstrap
  - Local prior sensitivity measures

Preprint: Giordano and Broderick [2023] (`arXiv:2305.06466`)
(Major update in progress, coming soon.)

References

A. Gelman and M. Heidemanns. The Economist: Forecasting the US elections., 2020. URL `https://projects.economist.com/us-2020-forecast/president`. Data and model accessed Oct., 2020.

A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

R. Giordano and T. Broderick. The Bayesian infinitesimal jackknife for variance. *arXiv preprint arXiv:2305.06466*, 2023.

J. Huggins and J. Miller. Reproducible model selection using bagged posteriors. *Bayesian Analysis*, 18(1):79–104, 2023.

R. Kass, L. Tierney, and J. Kadane. The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics*, 1990.

B. Kleijn and A. Van der Vaart. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6: 354–381, 2012.