An Introduction into Bayesian Modeling (with application to marketing)

(Some) Bayesian Essentials

Goethe University Frankfurt
otter@marketing.uni-frankfurt.de

October 19, 2022

Links to reference material and software

- https://www.perossi.net/home/bsm-1
- https://link.springer.com/referenceworkentry/10.1007/978-3-319-05542-8_24-1
- https://CRAN.R-project.org/package=bayesm
- https://mc-stan.org/
- I teach a semester long course on Bayesian modeling in Goethe U.'s Ph.D. program. Remote participation is possible.

The Bayesian take-off in Marketing

- Demand creating conditions:
  1. Relatively short time series, many heterogeneous units (e.g. panel data, conjoint experiments)
  2. Limited dependent variables
  3. Optimal actions, e.g. profit maximizing product lines, are nonlinear functions of model parameters

The Bayesian take-off in Marketing

Demand creating conditions	**Bayesian answer**
Relatively short time series, many heterogeneous units (e.g. panel data, conjoint experiments)	Adaptive shrinkage or pooling of information
Limited dependent variables	Data augmentation / Metropolis Hastings algorithm
Optimal actions, e.g. profit maximizing product lines, are nonlinear functions of model parameters	Represent posterior through a finite sample; compute optimal actions for arbitrarily non-linear loss-functions

The Bayesian take-off in Marketing

- Industry loves how Bayes helps here
- Demand from academia (on top)
- Basic interest in formulating and estimating quantitative models tailored to existing theory
- Decision support from counterfactual simulations taking account of uncertainty

Truths about Bayesian Analysis

- **Bayesian analysis:**
  - ... reverses the data generating process exactly, therefore is coherent (implies that one can and should test one's own and others' algorithms for Bayesian inference using simulated data)
  - ... is clear about uncertainty due to limited data
  - ... is fundamentally tied to decision making
  - ... is facilitated by a set of relatively simple, powerful algorithms that work on difficult surfaces

The Goal of Inference

- Make **inferences** about **unknown quantities** using available **information**.
  - **inferences** - make probability statements
  - **unknown quantities** - parameters, functions of parameters, states or latent variables, *future* outcomes, outcomes conditional on an action ("counterfactuals")
  - **information**
    - data-based
    - non data-based
    - theories of behavior; *subjective views* there is an underlying structure
    - parameters are finite or in some range

Benefits and Costs of Bayes Inference

- Benefits
  - finite sample answer to right question
  - full accounting for uncertainty
  - integrated approach to inference and decision making
- Costs
  - computational (true any more?)
  - prior (cost or benefit?) *esp. with many parameters (hierarchical/non-parametric problems)*

Bayesian inference builds on the likelihood principle

$$p(D|\theta) \equiv \ell(\theta)$$

- **LP**: the likelihood contains all information relevant for inference. That is, as long as I have same likelihood function, I should make the same inferences about the unknowns.
- Implies analysis is done conditional on the data, in contrast to the frequentist approach, where the sampling distribution is determined through hypothetical replications of the data
- Implies efficiency, but requires specification
- Allows for straightforward inference from adaptive designs (see e.g., Liu et al., MarkSci 2007)
- *Note: any function proportional to data density can be called the likelihood.*

Bayes theorem

$$p(\theta|D) = \frac{p(D,\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{p(D)}$$

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

- Modern Bayesian computing - simulation methods for generating draws from the posterior distribution $p(\theta|D)$

Summarizing the posterior

- Output from Bayesian inference: $p(\theta|D)$
  - A high dimensional distribution
- Summarize this object via simulation:
  - marginal distributions of $\theta, h(\theta)$
  - don't just compute $\mathrm{E}(\theta|D), \mathrm{Var}(\theta|D)$
  - directly compute $\mathcal{L}(a|D, \mathcal{M})$ taking all posterior uncertainty into account (e.g., to determine the optimal price)
- Contrast with Sampling Theory:
  - point estimate with standard error
  - summary of irrelevant distribution
  - bad summary (normal)
  - limitations of asymptotics

Prediction

- See D, compute: $p(\tilde{D}|D)$ *Predictive Distribution*

$$p(\tilde{D}|D) = \int p(\tilde{D}|\theta)p(\theta|D)d\theta$$

$$(\neq p(\tilde{D}|\hat{\theta})!!!)$$

$$\text{assumes} \quad p(\tilde{D}, D|\theta) = p(\tilde{D}|\theta)p(D|\theta)$$

Decision theory

- Loss: $L(a, \theta)$, where $a =$ action, $\theta =$ state of nature
- Bayesian decision theory:

$$\min_a \{\bar{L}(a) = \mathrm{E}_{\theta|D}[L(a, \theta)] = \int L(a, \theta) p(\theta|D) d\theta\}$$

- Estimation problem is a special case:

$$\min_{\hat{\theta}} \{\bar{L}(\hat{\theta})\}; \text{typically}, L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)' A (\hat{\theta} - \theta)$$

- Accounting for model uncertainty:

$$\mathcal{L}(a|D, \mathcal{M}_1, \ldots, \mathcal{M}_K) = \sum_k p(D|\mathcal{M}_k) \Pr(\mathcal{M}_k) \int \mathcal{L}(a, \boldsymbol{\theta}) p(\boldsymbol{\theta}|D, \mathcal{M}_k) d\boldsymbol{\theta}$$

Identification

$$R = \{\theta : p(Data|\theta) = k\}$$
$$\text{where} \quad k = \max_{\theta} p(\text{Data}|\theta)$$

- If dim(R) > 1, then we have an *identification* problem. That is, there are a set of **observationally equivalent** values of the model parameters. The likelihood is *flat* or constant over R.
- Practical implications
  - likelihood can have flats or ridges
  - Issue for both the Bayesian (is it?) and non-Bayesian

Identification

- Is this a problem?
  - no, I have a proper prior
  - no, I don't maximize
- Classical solution:
  - impose enough constraints so that constrained parameter space is identified
- Bayesian solution:
  - use proper prior and recognize that some functions of θ are determined entirely by prior
    - essentially always the case when learning from data occurs sequentially through time
    - or when learning from data involves (partial) pooling over smaller, individually ill-conditioned data sets
  - simulation based inference will allow you to "see" and directly investigate identification from the data and the lack thereof
  - set-identification no problem for Bayesian inference

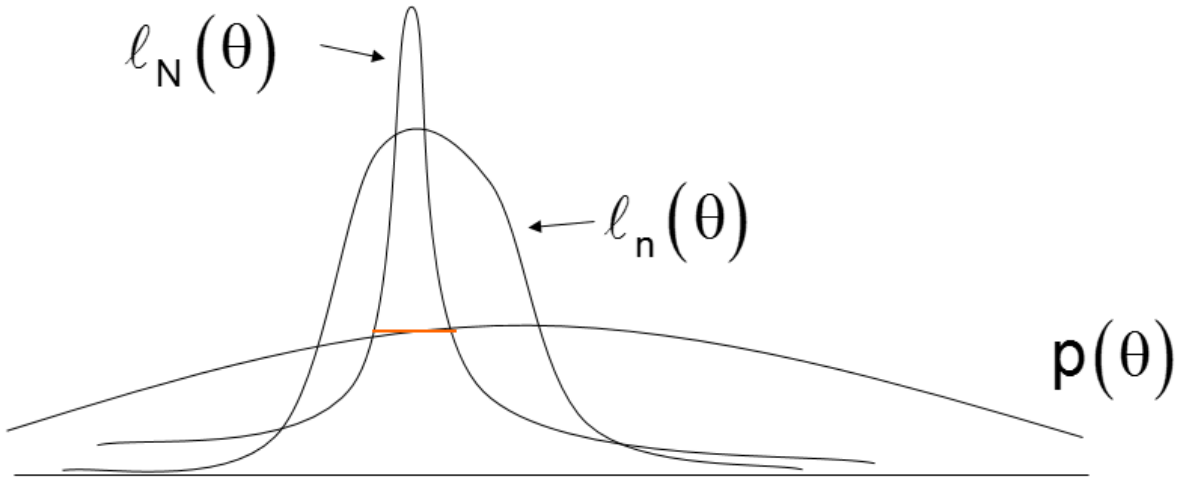
Bayes Inference: Summary

- Bayesian Inference delivers an integrated approach to:
  - Inference - including *estimation* and *testing*
  - Prediction - with a full account of uncertainty
  - Decision - with likelihood and loss (these are distinct!)
- Bayesian Inference is conditional on available info
- The right answer to the right question
- Bayes estimators are admissible. All admissible estimators are Bayes (Complete Class Thm). Which Bayes estimator?

Bayes/Classical Estimators

- Does MLE obey LP? YES
- Does theory of maximum likelihood estimator obey LP? No! who cares about an infinite amount of irrelevant data!
- Is there any relationship?
- Investigate asymptotic behavior of the posterior

Bayes/Classical Estimators



- Prior washes out - locally uniform!!! Bayes is consistent unless you have dogmatic prior.

$$p(\theta|D) \sim N(\hat{\theta}_{MLE}, [-H_{\theta=\hat{\theta}_{MLE}}]^{-1})$$

Beta-Binomial model

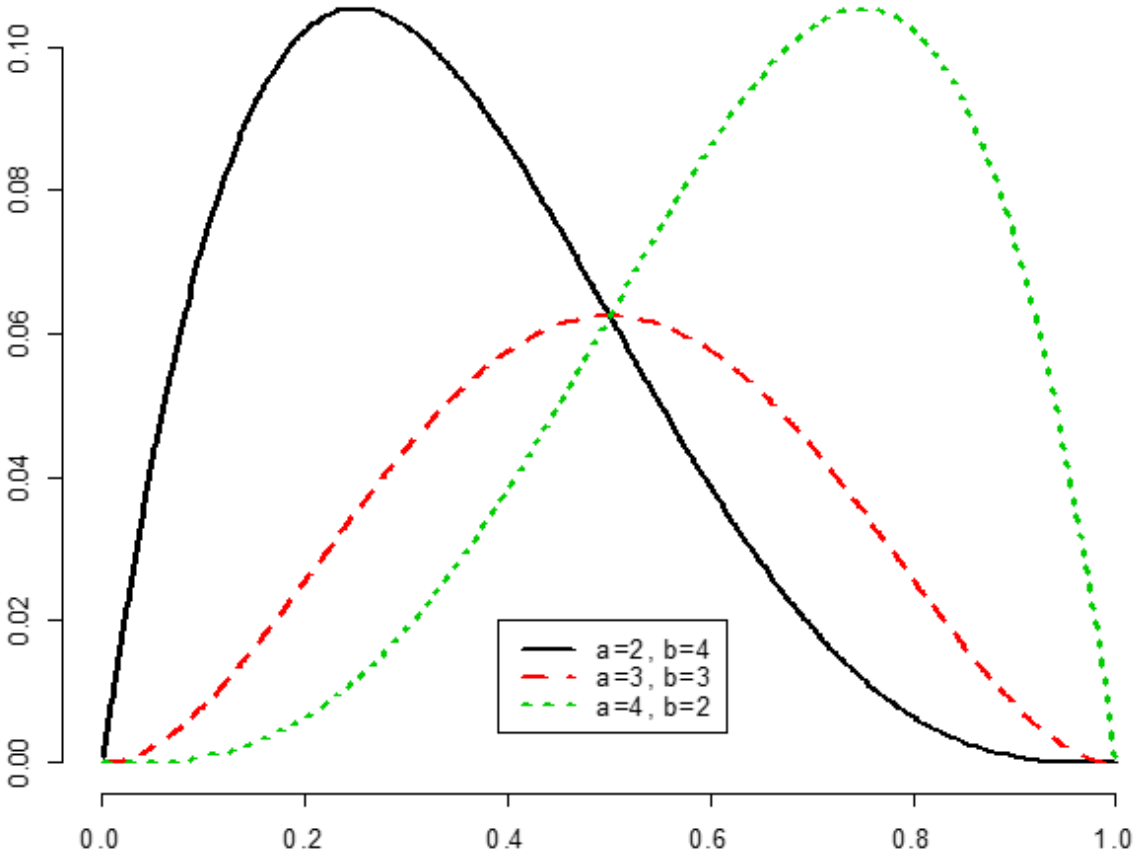
$$y_i \sim Bern(\theta)$$

$$\ell(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

$$= \theta^{y}(1-\theta)^{n-y} \text{ where } y = \sum_{i=1}^{n} y_i$$

$p(\theta|y) =?$ Need a prior!

Beta distribution



$$\text{Beta}(\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$E[\theta] = \frac{\alpha}{\alpha+\beta}$$

Posterior

$$
\begin{aligned}
p(\theta|D) &\propto p(D|\theta)p(\theta) \\
&= [\theta^{y}(1-\theta)^{n-y}] \times [\theta^{\alpha-1}(1-\theta)^{\beta-1}] \\
&= \theta^{\alpha+y-1}(1-\theta)^{n-y+\beta-1} \\
&\sim \text{Beta}(\alpha+y, n-y+\beta)
\end{aligned}
$$

Modern Bayesian inference I

- relies on computationally intensive algorithms to generate draws from (generally analytically intractable) joint posterior distributions
- **Gibbs sampling** generates draws by cycling through a complete set of conditional posterior distributions that uniquely characterize the joint posterior distribution (by the Hammersley-Clifford theorem9.
- requires that conditionals are tractable ("recognized as known distributions"); **data augmentation** can help
- **Metropolis-Hastings sampling** can generate draws from distributions only known up to an (intractable) normalizing constant ("the posterior is proportional to the likelihood times the prior").
- in its simplest form requires that likelihood and prior are analytically tractable
- generally, these sampling techniques only yield non-iid samples
- therefore, some burn-in from arbitrary starting values is required before summarizing draws for valid inference

Modern Bayesian inference II

- **Hamiltonian-Monte-Carlo sampling** as implemented in Stan is based on automatic derivatives of the log-posterior and very effectively navigates the posterior
- again requires that likelihood and prior are analytically tractable
- **Pseudo-marginal MH-sampling** can help when the likelihood is intractable but can be "forward-simulated" (see Andrieu and Roberts 2009, Annals of Statistics)
- The **exchange algorithm** can help when the normalizing constant of the likelihood is intractable (see Kosyakova et al. 2020, MarkSci for an example)
- However, in a recent project with Tetyana Kosyakova, Max Pachali, and Adam Smith we find that the requirement for iid exchange proposals is not innocuous

A note on software, programming environments

- I am currently relying heavily on `R` in combination with `Rcpp` in my own research.
- However, the No U-turn Sampler (NUTS) as implemented in `Stan` is a major break-through towards the goal of focusing on the specification of models (almost) exclusively.
- I have heard tremendous things about `Julia` (`https://julialang.org/`) from colleagues at the Frankfurt Institute for Advanced Studies (FIAS) but still have to use it myself
- `SPSS`, `STATA`, or `SAS` have started to include options for Bayesian estimation of well established "standard" statistical models such as ANOVA and generalized linear regression models.
- `WinBUGS` (`OpenBUGS`) is an earlier example of an attempt to automate Bayesian inference, with the idea that the user should be able to concentrate on the specification of a model

Binomial Probit example

$$
\begin{aligned}
p\left(y_i = 1|\beta\right) &= \int_{-\infty}^{\mathbf{x}_i'\beta} \mathcal{N}\left(z|0,1\right) dz = \int_{0}^{\infty} \mathcal{N}\left(z|\mathbf{x}_i'\beta,1\right) dz \\
p\left(y_i = 0|\beta\right) &= \int_{\mathbf{x}_i'\beta}^{\infty} \mathcal{N}\left(z|0,1\right) dz = \int_{-\infty}^{0} \mathcal{N}\left(z|\mathbf{x}_i'\beta,1\right) dz
\end{aligned}
$$

$$
p\left(\beta|\boldsymbol{y},\beta^{\mathbf{0}},\boldsymbol{\Sigma}^{\mathbf{0}}\right) \propto
$$

$$
\exp\left(-\frac{1}{2}\left(\beta-\beta^{0}\right)'\left(\boldsymbol{\Sigma}^{0}\right)^{-1}\left(\beta-\beta^{0}\right)\right)\prod_{i=1}^{n}\left(\Phi\left(\mathbf{x}_i'\beta\right)\right)^{y_i}\left(\Phi\left(-\mathbf{x}_i'\beta\right)\right)^{1-y_i}
$$

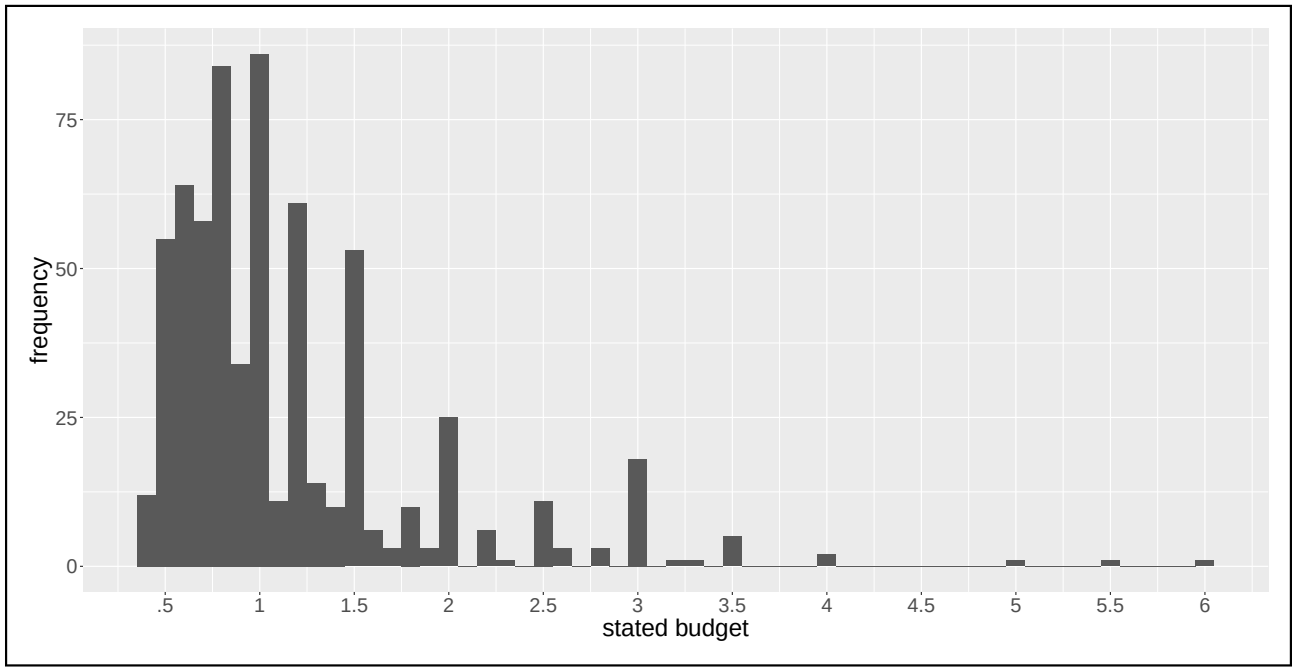
Data generation: $y_i = I(z_i > 0)$ where $z_i \sim \mathcal{N}\left(\mathbf{x}_i'\beta,1\right)$

Bayesian models excel at pooling information from different sources

Example: Pachali et al. (2022), "Omitted Budget Constraint Bias and Implications for Competitive Pricing", available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3044553`

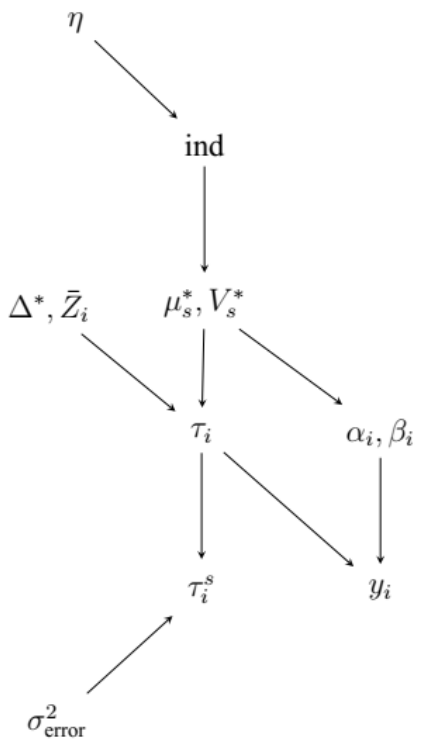
Bayesian models excel at pooling information from different sources

Figure: Distribution of respondents' stated budgets.



Notes: Respondents were asked in the survey about the maximum amount able to spend for their next laptop purchase prior to evaluating the choice tasks. Budget axis in $1,000$ EUR.

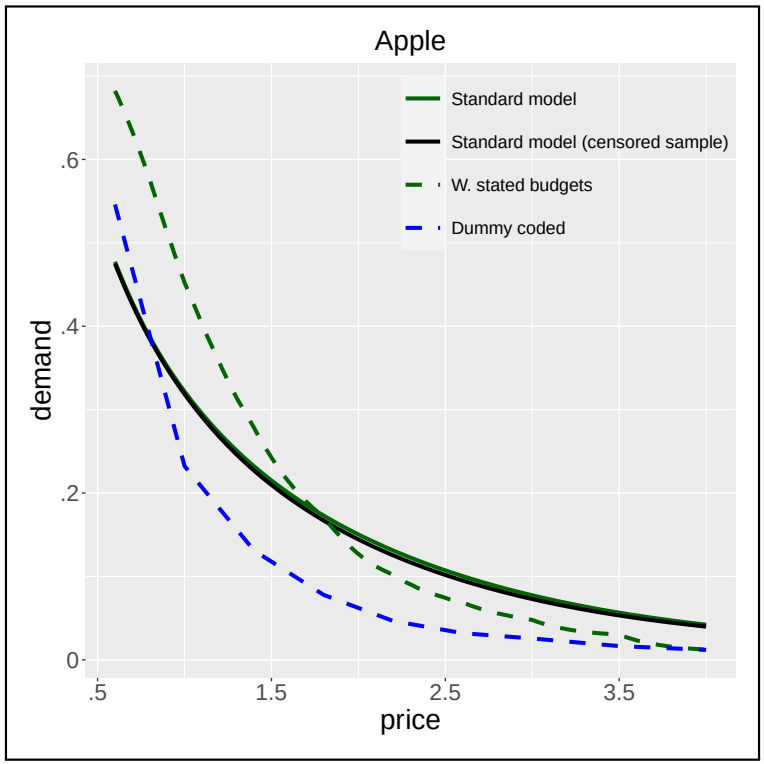
Figure 1: Inferring budgets from choice data, stated budgets and financial demographics DAG.



Notes: The arrow between Δ^* and α_i, β_i is removed to keep the graph simple. The picture also does not explicitly show the decomposition of μ_s^*, V_s^* as in Pachali et al. (2020) for the same reason.

Bayesian models excel at pooling information from different sources

Figure: Posterior predictive demand curve for the Apple laptop.



Notes: "W. stated budgets" refers to the model including stated budgets in the estimation. Censored sample drops 2% of respondents flagged by marginal likelihoods lower than random choice. Model-based "demand" is computed as a mean of posterior predictive choice probabilities. Competitors' prices are fixed at $4,000$ EUR. Price axis in $1,000$ EUR.

Bayes Theorem is awesome

Example: De Bruyn, Arnaud and Thomas Otter (2022), "Bayesian Consumer Profiling: How to Estimate Consumer Characteristics from Aggregate Data," Journal of Marketing Research, 59(4), 755-774.

- Can we determine the prevailing political preference in a target list with shared preferences for political candidates based on Zip code information only?
- Solution: Use reference list (population information): Detailed results of the elections for each of the 36,239 voting districts (Interior Minister's Web site), and apply Bayes Theorem

Who was supported by list members?

	National Averages	Simple Count Method	Bayesian Profiling	
			Estimate	*Post S.D.*
ABSTENTION	0.162	0.140	0.000	*< 0.001*
BLANK	0.012	0.010	0.001	*< 0.001*
SARKOZY	**0.257**	**0.303**	**0.999**	*< 0.001*
ROYAL	0.214	0.205	0.000	*< 0.001*
BAYROU	0.153	0.165	0.000	*< 0.001*
LE PEN	0.086	0.076	0.000	*< 0.001*
BESANCENOT	0.034	0.028	0.000	*< 0.001*
VILLIERS	0.018	0.017	0.000	*< 0.001*
BUFFET	0.016	0.013	0.000	*< 0.001*
VOYNET	0.013	0.013	0.000	*< 0.001*
BOVE	0.011	0.010	0.000	*< 0.001*
LAGUILLER	0.011	0.009	0.000	*< 0.001*
NIHOUS	0.009	0.008	0.000	*< 0.001*
SCHIVARDI	0.003	0.002	0.000	*< 0.001*

Table 11 - Abstentions, blank votes, and valid votes per candidate in the first round of the 2007 French presidential elections: national averages (leftmost column) versus target list estimates using the simple count and the Bayesian profiling methods. Although actual voting behavior of the target list is unknown, list members are expected to be extremely loyal to the candidate Sarkozy, a phenomenon that is predicted with striking accuracy by the Bayesian method.

Bayesian Inference can deal with otherwise intractable problems

Example: Kosyakova, Tetyana, Thomas Otter, Sanjog Misra, and Christian Neuerburg (2020), "Exact MCMC for Choices from Menus – Measuring Substitution and Complementarity among Menu Items," Marketing Science, 39, (2), 427-447.

$$\mathbf{Y}_i = \{Y_{i,1}, \ldots, Y_{i,k}, \ldots, Y_{i,K}\},$$

$$\Pr(\mathbf{Y}_i) = \frac{\exp(U(\mathbf{Y}_i; \mathbf{X}, \Psi_i))}{\sum_{Y' \in \mathcal{Y}} \exp(U(\mathbf{Y}'_i; \mathbf{X}, \Psi_i))} = \frac{\exp(U(\mathbf{Y}_i; \mathbf{X}, \Psi_i))}{\mathcal{Z}(\mathbf{X}, \Psi_i)}$$

For, say $K = 20$, the sum in the denominator has $2^{20} = 1,048,576$ terms.

Bayesian inference can deal with otherwise intractable problems

$$\alpha_{exchange}\left(\Psi_i \rightarrow \Psi_i^c, \mathbf{Y}_{i,t}^c\right) =$$

$$= min\left(1, \frac{p\left(\Psi_i^c\right) q\left(\Psi_i\right)}{p\left(\Psi_i\right) q\left(\Psi_i^c\right)} \prod_{t=1}^{T} \frac{\Pr\left(\mathbf{Y}_{i,t}|\Psi_i^c\right)}{\Pr\left(\mathbf{Y}_{i,t}|\Psi_i\right)} \frac{\Pr\left(\mathbf{Y}_{i,t}^c|\Psi_i\right)}{\Pr\left(\mathbf{Y}_{i,t}^c|\Psi_i^c\right)}\right)$$

$$= min\left(1, \frac{p\left(\Psi_i^c\right) q\left(\Psi_i\right)}{p\left(\Psi_i\right) q\left(\Psi_i^c\right)} \prod_{t=1}^{T} \frac{\ell^*\left(\mathbf{Y}_{i,t}|\Psi_i^c\right)}{\mathcal{Z}_t\left(\Psi_i^c\right)} \frac{\mathcal{Z}_t\left(\Psi_i\right)}{\ell^*\left(\mathbf{Y}_{i,t}|\Psi_i\right)} \frac{\ell^*\left(\mathbf{Y}_{i,t}^c|\Psi_i\right)}{\mathcal{Z}_t\left(\Psi_i\right)} \frac{\mathcal{Z}_t\left(\Psi_i^c\right)}{\ell^*\left(\mathbf{Y}_{i,t}^c|\Psi_i^c\right)}\right)$$

$$= min\left(1, \frac{p\left(\Psi_i^c\right) q\left(\Psi_i\right)}{p\left(\Psi_i\right) q\left(\Psi_i^c\right)} \prod_{t=1}^{T} \frac{\ell^*\left(\mathbf{Y}_{i,t}|\Psi_i^c\right)}{\ell^*\left(\mathbf{Y}_{i,t}|\Psi_i\right)} \frac{\ell^*\left(\mathbf{Y}_{i,t}^c|\Psi_i\right)}{\ell^*\left(\mathbf{Y}_{i,t}^c|\Psi_i^c\right)}\right)$$

$$= min\left(1, \frac{p\left(\Psi_i^c\right) q\left(\Psi_i\right)}{p\left(\Psi_i\right) q\left(\Psi_i^c\right)} \prod_{t=1}^{T} \frac{\ell^*\left(\mathbf{Y}_{i,t}|\Psi_i^c\right)}{\ell^*\left(\mathbf{Y}_{i,t}^c|\Psi_i^c\right)} \frac{\ell^*\left(\mathbf{Y}_{i,t}^c|\Psi_i\right)}{\ell^*\left(\mathbf{Y}_{i,t}|\Psi_i\right)}\right)$$

Takeaways for someone entering the field

- Bayes theorem immensely useful, even if you don't fully convert to Bayesian inference
- Priors are your friends in high-dimensional inference problems (subjectively informed regularization)
- The Bayesian approach uniquely supports decision making based on a (necessarily) imperfect understanding of how "the world works"