Heteroskedasticity and Clustered Covariances from a Bayesian Perspective

Lewis, Gabriel

University of Massachusetts, Amherst

14 December 2022

Working Paper: Heteroskedasticity and Clustered Covariances from a Bayesian Perspective

Gabriel Lewis*

December 14, 2022

Abstract

We show that $\sqrt{n}$-consistent heteroskedasticity-robust and cluster-robust regression estimators and confidence intervals can be derived from fully Bayesian models of population sampling. In our model, the vexed question of how and when to "cluster" is answered by the sampling design encoded in the model: simple random sampling implies a heteroskedasticity-robust Bayesian estimator, and clustered sampling implies a cluster-robust Bayesian estimator, providing a Bayesian parallel to the work of Abadie et al. (2017). Our model is based on the Finite Dirichlet Process (FDP), a well-studied population sampling process that apparently originates with R.A. Fisher, and our findings may not be surprising to readers familiar with the frequentist properties of the closely related Bayesian Bootstrap, Dirichlet Process, and Efron "pairs" or "block" bootstraps. However, our application of FDP to robust regression is novel, and it fills a gap concerning Bayesian cluster-robust regression. Our approach has several advantages over related methods: we present a full probability model with clear assumptions about a sampling design, one that does not assume that all possible data-values have been observed (unlike many bootstrap procedures); and our posterior estimates and credible intervals can be regularized toward reasonable prior values in small samples, while achieving the desirable frequency properties of a bootstrap in moderate and large samples. However, our approach also illustrates some limitations of "robust" procedures.

1 Introduction

Economists usually expect a regression's unobserved residuals to contain *heteroskedasticity*, where the variance is a function of the regressors; or to contain *clustered covariances*, where correlations exist within, but not between, known groups or "clusters" of residuals. Heteroskedasticity and clustered covariances do not bias the ordinary least squares (OLS) coefficient estimator. But they do render OLS inefficient, and they invalidate the classical covariances and confidence intervals automatically produced by most regression programs, which assume *homoskedasticity* — the absence of heteroskedasticity or clustered covariances. To draw valid inferences about the regression coefficients, economists often combine OLS with a "sandwich"[1] covariance estimator such as the Eicker (1967), Huber (1967), and White (1980a) heteroskedasticity-robust estimator ("HC0") or the Liang and Zeger (1986) cluster-robust estimator ("LZ"), or various bootstrap methods that are first-order

*PhD candidate at University of Massachusetts at Amherst. Email: gdlewis@umass.edu

[1]So called because of their shape, as we'll see.

equivalent to these.(Lancaster, 2006) These robust covariance estimators have transformed applied econometrics, (MacKinnon, 2019) allowing researchers to construct confidence intervals that capture the true regression coefficients at the nominal rate (e.g. 95% of the time) under relatively weak regularity assumptions while making essentially no modeling assumptions about the pattern of covariance between residuals, except to group observations into clusters.[2]

But why and how, exactly, should a researcher group observations into clusters? And which covariance estimator should a researcher choose, among the dozens that have been invented?[3] More fundamentally, does using robust covariance estimators merely shirk the necessary task of more fully modeling the scientific phenomenon at hand — including the residual covariances — as some have forcefully argued? (Freedman, 2006; King and Roberts, 2015; Leamer, 2010; Meng and Xie, 2014; Pelenis, 2012; Sims, 2010). These questions are important: Bertrand et al. (2004) established that different approaches to clustering or modeling residual covariances can yield substantially different economic conclusions from real panel data, and some approaches appear to be more reliable than others; in our examples, we will show much the same.

Many of the above questions have been answered authoritatively. However, authoritative answers have disagreed. Our brief and necessarily incomplete overview will largely focus on a literature where clustering has been particularly relevant: the labor economic literature concerning the effects of U.S. state policies as measured in monthly or annual U.S. census data on individuals or households. In this context, Bertrand et al. (2004) use simulations and heuristics to argue that attempts to parametrically model the residual covariance often fail badly, whereas results are satisfactory if one simply clusters by the cross-section (U.S. state), using the LZ or similar covariance estimators; this remains a conventional rule of thumb (Cameron and Miller, 2015). Abadie et al. (2017) and Abadie et al. (2020), in contrast, propose a more formal design-based framework and derive a different estimator in which uncertainty comes from randomized sampling and treatment assignment within a fixed population; they conclude that one should usually cluster at the level at which the treatment is assigned, with the caveat that clustering must be adjusted to account for whether one has observed the entire population of clusters. J. MacKinnon and Webb (2019) strongly object to Abadie et al.'s finite population framework, while largely endorsing Abadie et al.'s advice (and that of Bertrand et al. (2004)), while also saying to the contrary that one should generally cluster at "at the broadest feasible level" (p.11) — using fewer, larger clusters — another conventional rule of thumb. J. MacKinnon and Webb (2019) also add that one could try clustering by both cross-section *and* time in panel data (p.12). At least in some cases, these different pieces of advice, or the different lines of reasoning that support them, conflict.

Some of the above advice is simply ambiguous. It may seem pragmatic to cluster "broadly," following the intuition that allowing for more observations to be correlated produces wider confidence intervals, and so is in a sense "conservative." However, in economics there is usually a plausible argument for clustering even more broadly (are observations from different U.S. states *really* independent?); so when should the researcher stop broadening? Broader clustering brings its own concerns: the fewer the clusters, the more tenuous are the asymptotic approximations which underpin both sandwich estimators and bootstrap methods (Carter et al., 2017); and from a decision-theoretic perspective, a policy of increasing the widths of one's confidence intervals is not necessarily conservative at all, if there is a cost to professing ignorance about the true parameter.(Rice et al., 2008) With the latitude that such ambiguity affords, a researcher may gravitate

[2]Provided that an asymptotic approximation holds, which it does in many applications, even in moderate samples. The crucial assumption is that fourth mixed moments of all variables can be consistently estimated.

[3]See Abadie et al. (2017), Cameron and Miller (2015), and MacKinnon (2019) for reviews.

toward choosing whatever kind of clustering preserves the statistical significance (or nonsignificance) that he may desire.

In the above debate about how (and if) to apply sandwich methods, a key source of disagreement is the very reason that such methods are so useful and popular in econometrics: they are "model agnostic" — they are not typically derived from a model (a likelihood function, for example) which specifies the probability distribution of the data-generating process (perhaps up to some normalizing constant) given unknown parameters. (Buja et al., 2019) Given such a model, there would be (arguably) less room for debate about estimators: the pattern of covariance is determined by the model's parameters, and choosing the parameter-values that maximize the likelihood usually produces both estimators and confidence intervals that are consistent and asymptotically efficient from a frequentist perspective;(van der Vaart, 1998) alternatively, placing a prior probability distribution over the unknown parameters produces a Bayesian posterior distribution that satisfies various proposed axioms of rationality.(Ramsey, 1926) Moreover, the maximum likelihood confidence intervals and the Bayesian credible intervals would usually tend to converge asymptotically, known as a Bernstein-von Mises theorem,(van der Vaart, 1998) providing relatively strong guarantees that whichever method one uses, one's inferences are approximately correct from both Bayesian and frequentist perspectives, at least with sufficient data. Accordingly, in this model-guided paradigm, the debate would be less about the covariance estimator, than about the scientific question of how, precisely, the data was generated. The challenge, in the model-guided paradigm, is to develop a plausible model that delivers valid inferences in the presence of heteroskedasticity or clustered covariances, without pretending to implausible *a priori* knowledge about the covariance structure.

In this paper we present a Bayesian model-based perspective which may help clarify why and how to cluster one's data in both Bayesian and non-Bayesian contexts. Within a simple Bayesian model of population sampling, we define the regression coefficients β as the "best" (squared-error-minimizing) linear approximation to the population distribution, following an approach outlined by Buja et al. (2019) and Szpiro et al. (2010), similar to that of White (1980b). Inference about these coefficients is necessarily clustered according to sampling: if one's model specifies simple random sampling, then the Bayes estimate and credible intervals asymptotically converge to OLS with the usual heteroskedasticity-robust confidence intervals; and if one's model specifies clustered sampling, then the Bayes estimate and credible intervals converge to OLS with the usual cluster-robust confidence intervals. Much has been made of the difficulty or subjectivity of eliciting the prior, or indeed of fully modeling the data beyond its first moments; (e.g. see Buja et al., 2019) but in our case the prior vanishes asymptotically, while the direct logical implication from sampling model to posterior distribution essentially solves the problem of clustering.

Here is a summary of our main theoretical result. We have the following convergence in posterior distribution, almost surely under the true data-generating process as the number of sampled clusters $n \longrightarrow \infty$:

$$\sqrt{n}\left(\beta_n - \hat{\beta}_n\right) | \mathbf{y}_n, \mathbf{X}_n \xrightarrow{d} \mathcal{N}\left(0, \lim_{n\to\infty} n\hat{\mathbb{V}}_n\right) \tag{1}$$

where

$$\hat{\mathbb{V}}_n = \left(\sum_{c=1}^{n} X_i^\top X_i\right)^{-1} \left[\sum_{i=1}^{n} X_i^\top \hat{\epsilon}_i \hat{\epsilon}_i^\top X_i\right] \left(\sum_{i=1}^{n} X_i^\top X_i\right)^{-1} \tag{2}$$

Informally, the above expression says that the Bayesian regression coefficients β_n given data $(\mathbf{y}_n, \mathbf{X}_n)$ are asymptotically distributed as a Gaussian random variable whose mean is the OLS

coefficients $\hat{\beta}_n$, and whose covariance $\hat{\mathbb{V}}_n$ is a standard robust covariance estimator constructed from OLS residuals $\hat{\epsilon}$ and regressors X_i. Here, i indexes the sampling units (not necessarily individual rows of data) and so the form of sampling directly implies the form of the clustered covariance.

Convergence in distribution does not imply convergence of moments, but under stronger regularity assumptions much like those in White (1980b), one finds consistency of the posterior mean and posterior covariance of β_n.

1.1 Related Bayesian literature

The conventional Bayesian approach to heteroskedasticity and clustered observations is either to ignore them (Gelman, Hill, et al., 2021, p.154) or to build a *conditional model* in which the regressors are taken as given (ancillary), and the distribution of the outcome given the regressors has some parametric distribution, making strong assumptions about the functional dependence of the covariance on the regressors (Gelman, Carlin, et al., 2020, Chapter 14). Alternatively, Norets (2015) and Zhao (2015) show that certain classes of Bayesian nonparametric heteroskedasticity models do produce efficient coefficient estimators and CIs with good coverage rates; unfortunately, their methods cannot be extended to clustered covariances, and their results critically depend on correctly estimating the smoothness of the true covariance function, making them substantially more assumption-heavy than OLS with robust covariance estimators. Conditional modeling has considerable merits, but we set it aside in this paper. A critical review with simulations appears in the Appendix.

Various authors have developed quasi-Bayesian interpretations or analogues of robust covariance estimators. Hoff and Wakefield (2012), Startz (2012), and Yin (2009) achieve their results by collapsing the data into its first and second sample moments and modeling only these moments. Because this approach no longer provides a model of the full data-generating process, we set it aside as well.

In our model, the regression observations (y_i, X_i) are jointly drawn from a Finite Dirichlet Process (FDP). The FDP was first described in a modern Bayesian context by Pitman (1996), who attributes the model to R.A. Fisher; important work was subsequently done by Ishwaran and Zarepour (2002) and Muliere and Secchi (2003), but not in a regression context and without reference to heteroskedasticity. Our theoretical results may not be surprising to readers familiar with the asymptotic frequency properties of the Dirichlet Process (introduced by Ferguson, 1973) and the Bayes Bootstrap (due to Rubin, 1981), to which the FDP is closely related and converges in certain cases. It has been known since at least the work of Hjort (1994) and A. Lo (1987b) that both the Bayes Bootstrap and Dirichlet Process converge in various senses to the Efron (1979) "pairs" bootstrap, which is known to have heteroskedasticity robustness properties (Freedman, 1981).

Our work is particularly indebted to the relatively small literature in which the Bayesian bootstrap and related methods are specifically considered for heteroskedasticity robust regression (Aitkin, 2008; Chamberlain and Imbens, 2003; Karabatsos, 2016; Lancaster, 2006; Poirier, 2011; Szpiro et al., 2010). Of these, Karabatsos (2016) is perhaps our nearest precedent, but this paper uses a mixture-of-Dirichlet-Processes model, and when discussing robustness it focuses on an improper prior in which the model becomes equivalent to the Bayes Bootstrap. No one in this literature discusses cluster-robust regression in any detail.

Our novel contribution is to show how a fully Bayesian population sampling model (which the Bayes Bootstrap is not) can deliver inferences which are equivalent to OLS with either heteroskedasticity-robust or cluster-robust covariance estimators (which the aforementioned Dirichlet Process liter-

ature does not discuss). This allows us to provide a relatively elementary explanation (one that does not require empirical process theory) from a Bayesian perspective of how one can address both heteroskedasticity and clustered correlations, while also having frequency guarantees. Since the Bayesian Bootstrap and the Dirichlet Process are limiting cases of our model, they are covered by our results. Other systematic sampling schemes, and formal causal analysis, can also be incorporated into this framework.

1.2 Notation and Aims

In this section we clarify what we are trying to accomplish and so must introduce some notation. We will omit measure-theoretic details. Throughout this paper, we will focus on the simplest kinds of linear regressions, where observations $z_i = (y_i, X_i)$, with $i = 1, 2, \ldots, n$, are drawn independently and identically from some (true) data-generating probability distribution P^* with expectations $\mathbb{E}^*$ that we may think of as an unknown population about which we are seeking inferences; the distribution of a sequence of datasets is given by an infinite product measure $\mathbb{P}^{*(\infty)}$ whose properties we will stipulate when we need them. To simplify discussion, we will sometimes refer to a generic observation $(y_i, X_i) = (y, X)$.

1.2.1 Notation for heteroskedasticity and clustering

In the heteroskedasticity scenario, we have y_i taking values in $\mathbb{R}$ and X_i in $\mathbb{R}^k$ (as a row-vector), so that (y_i, X_i) can be considered a "row" of observations. In this case, the population P^* consists of $j = 1, \ldots, M$ points $(\mathring{y}_j, \mathring{X}_j)$ with associated proportions θ_j.

In the cluster-robustness scenario, each y_i is a column-vector with some dimension N_i, and each X_i is an $N_i \times k$ matrix, so that each observation(y_i, X_i) is a "block" of data. In this case, n is the number of clusters, not the sum of all rows in all clusters.[4] We will sometimes stack n observations in the usual way, writing $\mathbf{y} := (y_1, \ldots, y_n)^\top$, and $\mathbf{X} := (X_1^\top, \ldots X_n^\top)^\top$.

1.2.2 Our estimand

In both the clustering and the heteroskedasticity case, we hope to draw inferences about a squared-error-minimizing linear approximation:

$$\beta^* := \arg\min_\beta \mathbb{E}^* \left[(y - X\beta)^\top (y - X\beta) \right] = \mathbb{E}^*[X^\top X]^{-1} \mathbb{E}^*[Xy] \tag{3}$$

Throughout, we will assume that $\mathbb{E}^*[X^\top X]$ is indeed invertible. In regular cases, an equivalent definition is $\beta^* := \arg\min_\beta \mathbb{E}^* \left[(\mathbb{E}^*[y|X] - X\beta)^\top (\mathbb{E}^*[y|X] - X\beta) \right]$, clarifying that we are approximating the conditional expectation function $\mathbb{E}^*[y|X]$.

If the true conditional expectation does happen to be linear, then $\mathbb{E}^*[y|X] = X\beta^*$, and so β^* quantifies the "slope" of the conditional expectation, making β^* clearly a useful estimand in this case. However, we need not (and our model does not) assume linearity in order to fit a linear regression. This perspective is advocated in an recent paper by Buja et al. (2019) and is implicit in

[4] If N_i may vary in the population from 1 to some maximum size $\bar{N}$, then the marginal draws of y_i take values in $\mathcal{Y} = \mathbb{R}^1 \cup, \ldots, \cup \mathbb{R}^{\bar{N}}$, and X_i in $\mathcal{X} = \mathbb{R}^{1\times k} \cup, \ldots, \cup \mathbb{R}^{\bar{N}\times k}$; and jointly, (y_i, X_i) takes values in $\mathcal{Y} \times \mathcal{X}$. The true conditional expectation $\mathbb{E}^*[y_i|X_i]$ can be viewed as mapping an $N_i \times k$ matrix $X_i \in \mathcal{X}$, to a vector in $\mathbb{R}^{N_i} \subset \mathcal{Y}$; the true conditional variance function $\mathbb{V}^*(y_i|X_i)$ maps X_i to a positive semidefinite matrix in $\mathbb{R}^{N_i \times N_i} \subset \mathcal{Y} \times \mathcal{Y}$.

a pivotal paper on the Bayes Bootstrap by Chamberlain and Imbens (2003). This view of regression perhaps originates in the seminal work of White (1980b), in which heteroskedastic-robust estimators are presented as generally robust to model misspecification.

1.2.3 Frequentist estimation

In the frequentist literature, a key attraction of choosing β^* for one's estimand is that β^* is relatively easy to draw inferences about, at least when k remains much smaller than n. As is well-known, the sample OLS coefficients $\hat{\beta} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ are consistent, in the sense that $\hat{\beta} \longrightarrow \beta^*$ in probability as $n \longrightarrow \infty$, under relatively mild regularity conditions. In particular, consistency does not require that the conditional distribution of y given X be Gaussian or homoskedastic.

Under slightly stronger regularity conditions, but still without assuming that the conditional mean or conditional variance functions have any particular functional form, one can construct consistent confidence intervals for $\hat{\beta}$: that is, nominal $(1-\alpha)\%$ intervals around $\hat{\beta}$ that (under hypothetically repeated sampling) capture the true β^* at a rate rapidly approaching $(1-\alpha)\%$ as $n \longrightarrow \infty$. To make such confidence intervals, one uses an estimator of the covariance of $\hat{\beta}$ such as that attributed to Eicker (1967), Huber (1967), and White (1980b):

$$\hat{\mathbb{V}}^{HC0}(\hat{\beta}) = (\mathbf{X}^\top\mathbf{X})^{-1}\left[\sum_{i=1}^{n} X_i^\top \hat{\epsilon}_i^2 X_i\right](\mathbf{X}^\top\mathbf{X})^{-1} \tag{4}$$

where $\hat{\epsilon}_i = y_i - X_i\hat{\beta}$ are the OLS residuals. In the case of clustered covariances with $c = 1, \ldots, C$ clusters, using our notation the Liang-Zeger covariance estimator (Liang and Zeger, 1986) is nearly identical:

$$\hat{\mathbb{V}}^{LZ}(\hat{\beta}) = (\mathbf{X}^\top\mathbf{X})^{-1}\left[\sum_{c=1}^{C} X_c^\top \hat{\epsilon}_c\hat{\epsilon}_c^\top X_c\right](\mathbf{X}^\top\mathbf{X})^{-1} \tag{5}$$

where the N_c-vector $\hat{\epsilon}_c$ and the $N_c \times k$ matrix X_c correspond to the N_c rows of observations in cluster c. In our discussion, we will sometimes use $\hat{\mathbb{V}}$ to generically denote a robust covariance estimator. A large literature is devoted to variants of the above covariance estimators, with various "finite-sample" and "degree-of-freedom" corrections; for a review see Cameron and Miller (2015).

1.2.4 When heteroskedasticity matters

To gain a concrete understanding of how heteroksedasticity can cause problems, we briefly consider the bivariate regression case where $X_i = (1, w_i)$ and w is univariate. Focusing on the slope coefficient $\hat{\beta}_2$ (usually the parameter of interest),[5]

[5] See Buja et al. (2019, p.26) for the first equality.

$$
\begin{aligned}
\mathbb{V}^{\hat{HC0}}(\hat{\beta}_2) &= \frac{1}{n} \frac{\frac{1}{n}\sum_i \hat{\epsilon}_i^2 (w_i - \bar{w})^2}{\left(\frac{1}{n}\sum_i (w_i - \bar{w})^2\right)^2} \\
&= \frac{1}{n} \frac{\frac{1}{n}\sum_i \left(\hat{\epsilon}_i^2 - \frac{1}{n}\sum_j \hat{\epsilon}_j^2\right)\left((w_i - \bar{w})^2 - \frac{1}{n}\sum_l (w_l - \bar{w})^2\right)}{\left(\frac{1}{n}\sum_i (w_i - \bar{x})^2\right)^2} + \frac{1}{n} \frac{\frac{1}{n}\left(\sum_i \hat{\epsilon}_i^2\right)\left(\frac{1}{n}\sum_i (w_i - \bar{w})^2\right)}{\left(\frac{1}{n}\sum_i (w_i - \bar{w})^2\right)^2} \\
&= \frac{1}{n} \frac{\hat{\text{Cov}}(\hat{\epsilon}^2, (w - \bar{w})^2)}{\hat{\text{Var}}(x)^2} + \hat{\mathbb{V}}^{OLS}(\hat{\beta}_2) \qquad (6)
\end{aligned}
$$

We recognize the second term as the classical (homoskedastic) OLS variance estimator, written $\hat{\mathbb{V}}^{OLS}(\hat{\beta}_2)$. Since $\hat{\mathbb{V}}_{HC}(\hat{\beta})$ is (usually) consistent, we can conclude that $\hat{\mathbb{V}}^{OLS}(\hat{\beta}_2)$ is too small whenever the first term is positive: that is, when the squared residuals are positively correlated with $(w - \bar{w})^2$, the squared distance of the regressor from its mean. In other words, if the linear model's fit worsens (either due to increasing variance or nonlinearity) as W gets farther from its mean, confidence intervals based on the classical $\hat{\mathbb{V}}^{OLS}(\hat{\beta}_2)$ would capture the true parameter at less than the nominal 95% rate.

1.2.5 Bayesian inference

A Bayesian model can be built from 1) a "sampling model" $\mathbb{P}(d\mathbf{z}_n|\lambda) = \prod_{i=1}^n P(dz_i|\lambda)$, in which the n datapoints $\mathbf{z}_n$ are independently and identically distributed $P(\cdot|\lambda)$ with unknown parameters λ in some set Θ, and 2) a prior distribution $\mathbb{P}(d\lambda)$ over the parameters. Together, these produce a joint distribution $\mathbb{P}(d\mathbf{z}_n, d\lambda)$, and hence a posterior distribution of the parameters λ given the data $\mathbf{z}_n$. [6].

Within the Bayesian modeling framework, we will write probabilities $\mathbb{P}$, expectations $\mathbb{E}$, and variances $\mathbb{V}$ without asterisks, taking care not to confuse these with their true (data-generating) counterparts that have asterisks. We take the view that a Bayesian model represents beliefs or inferences about ("or models") some true unknown P^*, setting aside philosophical concerns about the well-foundedness of P^*, P or $\mathbb{P}$ for the purposes of this paper.

We will define our Bayesian linear regression coefficients just as frequentist econometricians define their estimand, with

$$
\boldsymbol{\beta}_n := \arg\min_{\beta} \mathbb{E}\left[(y - X\beta)^\top (y - X\beta) \,|\lambda\right] \qquad (7)
$$

, where y and X are distributed according to a distribution $P(dy, dX|\lambda)$ that we will describe in greater detail in subsequent sections.

2 Our model: Finite Dirichlet Process

2.1 Sampling model

Suppose $\mathring{\mathbf{z}} = (\mathring{z}_1, \ldots, \mathring{z}_M) = ((\mathring{y}_1, \mathring{X}_1), \ldots, (\mathring{y}_M, \mathring{X}_M))$ is a set of $M < \infty$ distinct unknown values with respective unknown proportions $\boldsymbol{\theta} = \theta_1, \ldots, \theta_M$ in some population about which we seek

[6] This holds even when Bayes' Theorem does not apply. (Polpo et al., 2015, Ch.1)

inferences. Under simple random sampling with replacement, each observation z_i is drawn independently and identically from a categorical distribution on the given points $\mathring{\mathbf{z}} = \mathring{z}_1, \ldots, \mathring{z}_M$ with given probabilities $\boldsymbol{\theta} = \theta_1, \ldots, \theta_M$, which is described by the following probability transition kernel:

$$P(dz_i|\mathring{\mathbf{z}}, \boldsymbol{\theta}) := \sum_{j=1}^{M} \theta_j \delta_{\mathring{z}_j}(dz_i) \tag{8}$$

Recall that the Dirac delta measure $\delta_{\mathring{z}}(dz)$ simply places all prbability at point $\mathring{z}$, and behaves as$\int \delta_{\mathring{z}}(dz) f(z) = f(\mathring{z})$ for any measurable f.

Here, P represents the unknown population. One can think of each $\mathring{z}_j$ as a collection of numerical quantities associated with a distinct category j of sampling unit in the population, and $\theta_1, \ldots, \theta_M$ as proportions of the total population in each category. Unlike basic categorical models, the values $(\mathring{z}_1, \ldots, \mathring{z}_M)$ are not all known before sampling occurs.

In the standard heteroskedasticity scenario, each $\mathring{y}_j$ takes values in $\mathcal{Y} = \mathbb{R}$ and each $\mathring{X}_j$ take values in $\mathcal{X} = \mathbb{R}^k$, so $\mathring{z}_j$ in $\mathcal{Y} \times \mathcal{X}$ is essentially a row of regression-variables that would be sampled as a single unit, perhaps associated with an individual in a population. Naturally, observations z_i also occur in $\mathcal{Y} \times \mathcal{X}$.

In the cluster-robustness scenario, things are similar, except that each $\mathring{y}_j$ is a vector of N_j values, and each $\mathring{X}_j$ is a matrix of $\mathring{N}_j \times k$ values, so that each $(\mathring{y}_j, \mathring{X}_j)$ is a "cluster" of rows of regression-variables that would be sampled from the population as a single unit. To be more precise, if $\mathring{N}_j$ can be any integer from 1 to some maximum cluster size $\bar{N}$, then each $\mathring{y}_j$ takes values in $\mathcal{Y} = \mathbb{R}^1 \cup, \ldots, \cup \mathbb{R}^{\bar{N}}$, and $\mathring{X}_j$ in $\mathcal{X} = \mathbb{R}^{1\times k} \cup, \ldots, \cup \mathbb{R}^{\bar{N}\times k}$.[7]

2.1.1 What constitutes a cluster?

In this model, "clustering" is determined by our population of interest and how it is being sampled. If we are conducting a simple random cross-sectional sample of people from a population, then each $(\mathring{y}_j, \mathring{X}_j)$ is a single "row" of data for person j. If are randomly cross-sectionally sampling entire households from the population and collecting data on each member of the household, then each household is a cluster of size $\mathring{N}_j$ with observations$(\mathring{y}_j, \mathring{X}_j)$.

Data-generating processes like the above seem to describe a wide variety of social-science scenarios where we are sampling individuals or groups of individuals from a finite population. We conjecture that our model may also approximate sampling without replacement in large-population cases, similar to the closely related models by A. Lo (1987a) and Aitkin (2008), an extension that we may consider in a later paper.

2.2 Regression functional

The parameters $\mathring{\mathbf{z}}, \boldsymbol{\theta}$ themselves are not necessarily of great interest, but they can be used to define a wide range of quantities of interest. Here, we focus on the usual least squares regression functional, assuming throughout that $\sum_{j=1}^{M} \theta_j \mathring{X}_j^\top \mathring{X}_j$ is (almost surely) invertible:

[7]Probability distributions on $\mathcal{Y}\times\mathcal{X}$ are not hard to construct from standard probability distributions: for example, first draw $\mathring{N}_j$ from some probability distribution on $1, \ldots \bar{N}$, then draw the $\mathring{N}_j$ "rows" of $(\mathring{y}_j, \mathring{X}_j)$ independently from some standard $1 + k$-dimensional distribution. Since these are finite unions of standard Borel spaces, they do not raise any concern about measurability.

$$\beta(\mathring{\mathbf{z}}, \boldsymbol{\theta}) := \arg\min_{\beta} \mathbb{E}\left[(y - X\beta)^\top (y - X\beta) \,|\mathring{\mathbf{z}}, \boldsymbol{\theta}\right] = \left[\sum_{j=1}^{M} \theta_j \mathring{X}_j^\top \mathring{X}_j\right]^{-1} \sum_{j=1}^{M} \theta_j \mathring{X}_j^\top \mathring{y}_j \tag{9}$$

The prior and posterior distributions of $\mathring{\mathbf{z}}, \boldsymbol{\theta}$ induce prior and posterior distributions of β. The exact distribution of β is hard to characterize in closed form, but easy to draw samples from, since both the prior and the posterior $\mathring{\mathbf{z}}, \boldsymbol{\theta}$ are easy to sample from.

Other parameters could be derived by minimizing other objective functions, in a Bayesian parallel to frequentist "M-estimators" (see Chamberlain and Imbens 2003). We would expect such parameters to have similar robustness properties to this one.

2.3 Priors

In Bayesian reasoning, all unknowns need prior probability distributions. We assign the vector of unknown population proportions $\boldsymbol{\theta}$ the usual conjugate Dirichlet distribution on the $M-1$-dimensional simplex:

$$\mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_M) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{M} \theta_k^{\alpha_k - 1} d\boldsymbol{\theta} \tag{10}$$

with the Dirichlet distribution parameters $\alpha_j > 0$ all fixed at $\alpha_j = {}^{\alpha}/_{M}$; here,$\Gamma$ represents the Gamma function.

The unknown population-values are given independent and identical priors:

$$\mathring{z}_j \overset{iid}{\sim} F \tag{11}$$

for some distribution F on $\mathcal{Y} \times \mathcal{X}$ that has no atoms and has positive density everywhere.[8] Concretely, to draw a single $\mathring{z}_j$ from F, the analyst might start with a fixed prior "best guess" of the regression slope $\mathring{\beta}_0$, then draw $\mathring{N}_j$ from some prior (fix $\mathring{N}_j = 1$ in the heteroskedasticity case), then draw the $\mathring{N}_j$ rows of $\mathring{X}_j$ independently from some suitable k-dimensional distribution with full support, then draw $\mathring{\epsilon}_j$ from some distribution (independently of X_j, for simplicity) with full support on $\mathbb{R}^{N_j}$ and mean zero, then define the $\mathring{N}_j$ elements of $\mathring{y}_j$ as $\mathring{y}_j = X_j\beta_0 + \mathring{\epsilon}_j$. Other schemes could be devised; the main requirement is that they have full support on $\mathcal{Y} \times \mathcal{X}$.

Putting it all together, we will write the prior distribution as follows:

$$\Pi_0(d\boldsymbol{\theta}, d\mathring{\mathbf{z}}) = \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_M) \prod_{k=1}^{M} F(d\mathring{z}_k) \tag{12}$$

These priors bear some explanation. As we will see from the posterior distribution, the parameters α_j can be interpreted as "counts" of (hypothetical) observations of the unknown population-values $\mathring{z}_j$. It makes sense to set $\alpha_j = \frac{\alpha}{M}$, since this is symmetric and keeps the sum of our prior counts $\sum_j \alpha_j$ invariant with respect to the number of latent points M; we do not want our "total prior data" to automatically increase with M.

[8] In the cluster case, I have in mind that every cluster-size N_j up to some $\bar{N}$ is given positive probability, and for any given cluster size the observations $(\mathring{y}_j, \mathring{X}_j)$ are drawn iid from a standard distribution with a (Lebesgue-dominated) density that is positive everywhere.

The fixed parameter α essentially governs the "concentration" of the population; the smaller it is, the greater prior certainty we have that the population is concentrated at fewer values (a few θ_j are relatively large while the rest are small).

The distribution F describes one's prior expectations about how the population values are distributed, as we will see from the prior predictive distribution. The non-atomicity of F is simply a logical requirement: if $\mathring{z}_j$ are independently and identically distributed as F but we also want them to be distinct (almost surely), then their distribution can have no atoms.

The postulated number of distinct values M could be endowed with a prior, but for simplicity we consider it fixed at some sufficiently large value in any given analysis. We will discuss this more in our section on asymptotics, showing that the precise value of M matters little.

2.4 Prior Predictive Distribution

The prior predictive distribution is helpful for understanding the prior and is essential for defining the posterior distribution. We will derive it here.

For a single predicted value z, the prior predictive distribution is defined (Mazzi and Spizzichino in Polpo et al. (chapter by Mazzi and Spizzichino, 2015)) as any probability measure $P_0(dz)$ that satisfies

$$\int P_0(dz)h(z) = \int \Pi_0(d\boldsymbol{\theta}, d\mathring{\mathbf{z}}) \int P(dz|\mathring{\mathbf{z}}, \boldsymbol{\theta})h(z) \tag{13}$$

for any arbitrary measurable function $h(z)$ taking values in $\mathbb{R}$. Here,

$$\int P_0(dz)h(z) = \int \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_M) \int \prod_{k=1}^{M} F(d\mathring{z}_k) \sum_j \theta_j \int \delta_{\mathring{z}_j}(dz)h(z) \tag{14}$$

$$\sum_j \frac{\alpha_j}{\sum_k \alpha_k} \int F(d\mathring{z}_j)h(\mathring{z}_j) \tag{15}$$

$$= \int F(d\mathring{z}_j)h(\mathring{z}_j) \tag{16}$$

using the fact that $\alpha_j = \frac{\alpha}{M}$.

Thus, the prior predictive distribution is simply F. This is why we say that F represents one's prior "best guess" about the population distribution.

2.5 Posterior Distribution

Given observed data $\mathbf{z}_n = z_1, \ldots, z_n$ consisting of m distinct values $\tilde{z}_1, \ldots, \tilde{z}_m$ with respective multiplicities $n_1, \ldots, n_m$ such that $\sum_{j=1}^{m} n_j = n$ and $m \leq M$, we learn that the "first" m population values $\mathring{z}_1, \ldots, \mathring{z}_m$ are fixed at the observed values $\tilde{z}_1, \ldots, \tilde{z}_m$, respectively, which we represent using Dirac delta distributions $\delta_{\tilde{z}_1}(d\mathring{z}_1), \ldots, \delta_{\tilde{z}_m}(d\mathring{z}_m)$. We learn nothing about the remaining unobserved population values $\mathring{z}_{m+1}, \ldots, \mathring{z}_M$, each of which retains the distribution F, independently and identically. Given the data $\mathbf{z}_n$ and $\mathring{\mathbf{z}}$, the conditional posterior distribution of $\boldsymbol{\theta}$ is Dirichlet, with the observed counts added to the "prior counts," much as in a standard Multinomial-Dirichlet categorical model.

Putting it all together, we may write the posterior distribution as

$$\Pi_n(d\boldsymbol{\theta}, d\mathring{\mathbf{z}}|\mathbf{z}_n) = \text{Dir}(d\boldsymbol{\theta}|\alpha/M + n_1, \ldots \alpha/M + n_m, \alpha/M, \ldots, \alpha/M) \prod_{j=1}^{m} \delta_{\tilde{z}_j}(d\mathring{z}_j) \prod_{k=m+1}^{M} F(d\mathring{z}_k) \quad (17)$$

In (arbitrarily) picking the "first" m population-values specifically, we have rather freely used a property of our model called *exchangeability* to make the above expression as simple as possible, and arguably simpler. The true expression contains expressions like the above, but within a sum of all possible ways to assign population values to observed points. All of these assignments are equivalent in a precise sense, allowing us to pick one assignment arbitrarily. In the technical appendix, we discuss exchangeability and show that our simplification does not affect any meaningful posterior inferences.

A more formal argument

As Ishwaran and Zarepour (2002) write, Corollary 20 in Pitman (1996) states that for the Fisher model that we use, given distinct observed values $\tilde{z}_1, \ldots, \tilde{z}_m$ with multiplicities $n_1, \ldots, n_m$, the kernel $P(dz)$ can be represented in the following way:

$$P|\mathbf{z}_n = \sum_{j=1}^{m} \pi_j \delta_{\tilde{z}_j}(dz) + \pi_{m+1} \mathring{P}(dz) \quad (18)$$

where

$$(\pi_1, \ldots, \pi_m, \pi_{m+1}) \sim \text{Dir}(\alpha/M + n_1, \ldots, \alpha/M + n_m, (M-m)\,\alpha/M) \quad (19)$$

independently of

$$\mathring{P}(dz) := \sum_{k=1}^{M-m} \lambda_k \delta_{\mathring{z}_j}(\cdot) \quad (20)$$

with iid $\mathring{z}_k \sim F(\cdot)$ and $(\lambda_1, \ldots, \lambda_{M-m}) \sim \text{Dir}(\alpha/M, \ldots, \alpha/M)$.

Picking up where Isharan and Zarepour leave off, we observe that $\lambda_1, \ldots, \lambda_{M-m}$ are essentially proportions that subdivide or "decimate" the variable π_{m+1}. Since both $\lambda_1, \ldots, \lambda_{M-m}$ and $(\pi_1, \ldots, \pi_m, \pi_{m+1})$ are jointly Dirichlet, and moreover π_{m+1} corresponds to parameter $(M-m)\,\alpha/M = \sum_{k=1}^{M-m} \alpha/M$, we may apply the "decimation" or "expansion" rule for Dirichlet distributions, (see Zhang (2008)), yielding:

$$(\pi_1, \ldots, \pi_m, \pi_{m+1}\lambda_1, \ldots, \pi_{m+1}\lambda_{M-m}) \sim \text{Dir}(\alpha/M + n_1, \ldots, \alpha/M + n_m, \alpha/M, \ldots, \alpha/M) \quad (21)$$

So in the posterior distribution,

$$P(dz|\boldsymbol{\theta}, \mathring{\mathbf{z}}) = \sum_{j=1}^{m} \theta_j \delta_{\tilde{z}_j}(dz) + \sum_{k=m+1}^{M} \theta_k \delta_{\mathring{z}_j}(dz) \quad (22)$$

with

$$\boldsymbol{\theta} \sim \text{Dir}(\alpha/M + n_1, \ldots \alpha/M + n_m, \alpha/M, \ldots, \alpha/M) \quad (23)$$

and iid $\mathring{z}_k \sim F(\cdot)$ for $k = m+1, \ldots, M$.

2.6 Posterior Predictive distribution

The posterior predictive distribution is helpful for understanding the posterior distribution, and also the effect (and therefore the meaning) of the prior distribution. It also allows us to draw connections to other closely-related models, and to begin considering our model's asymptotic properties.

Let h be any integrable real-valued function of a single prediction z. The posterior predictive distribution is fully characterized by the following expectation:

$$\mathbb{E}[h(z)|\mathbf{z}_n] \quad = \quad \int \Pi_n(d\boldsymbol{\theta}, d\mathring{\mathbf{z}}|\mathbf{z}_n) \int \sum_{l=1}^{M} \theta_l \delta_{\mathring{z}_l}(dz) h(z)$$

Using the posterior Π_n that we have just derived,

$$\mathbb{E}[h(z)|\mathbf{z}_n] = \int \mathrm{Dir}(d\boldsymbol{\theta}|\alpha/M+n_1, \ldots \alpha/M+n_m, \alpha/M, \ldots, \alpha/M) \left[\sum_{l=1}^{m} \theta_l h(\tilde{z}_j) + \sum_{l=m+1}^{M} \theta_l \int F(d\mathring{z}_l) h(\mathring{z}_l) \right] \tag{24}$$

Taking the expected value of the Dirichlet vector and rearranging,

$$\mathbb{E}[h(z)|\mathbf{z}_n] = \frac{n}{n+\alpha} \left[\frac{1}{n} \sum_{j=1}^{m} n_j h(\tilde{z}_j) \right] + \frac{\alpha}{n+\alpha} \frac{m}{M} \left[\frac{1}{m} \sum_{j=1}^{m} h(\tilde{z}_j) \right] + \frac{\alpha}{n+\alpha} \frac{M-m}{M} \mathbb{E}_0[h(\mathring{z})] \tag{25}$$

So the posterior expectation of $h(z)$ is a weighted average of, from left to right, 1) the sample mean, 2) the mean of the distinct observed values, and 3) the prior expectation, with weights that sum to one.

If the model "saturates" and $m = M$, then then the posterior expectation comprises the first two terms only:

$$\frac{n}{n+\alpha} \left[\frac{1}{n} \sum_{j=1}^{m} n_j h(\tilde{z}_j) \right] + \frac{\alpha}{n+\alpha} \left[\frac{1}{m} \sum_{j=1}^{m} h(\tilde{z}_j) \right] \tag{26}$$

This is the posterior predictive mean of a basic multinomial model supported on M known points with a symmetric Dirichlet prior with parameter α/M, to which the model reduces in this case. In comparison to the basic multinomial, the posterior mean of the FDP generally takes $\frac{M-m}{M}$ away from the weight of the mean of the population points and gives it to the prior mean, reflecting the lack of prior certainty about the unknown points.

The mean of distinct values, $\frac{1}{m}\sum_{j=1}^{m} h(\tilde{z}_j)$ in the second term, introduces a bias toward "more uncertain" (i.e. higher-entropy) distributions on the observed points. Desirable or not, this bias is usually small. As m approaches n, the mean of distinct values approaches the simple mean, so the term introduces less bias as its weight $\frac{\alpha}{n+\alpha}\frac{m}{M}$ increases. Furthermore, in many applications $M >> m$, making the term's contribution quite small. In fact, taking $M \longrightarrow \infty$, the second term of (25) vanishes and we have the posterior predictive mean of a Dirichlet Process (DP):

$$\frac{n}{n+\alpha} \left[\frac{1}{n} \sum_{j=1}^{m} n_j h(\tilde{z}_j) \right] + \frac{\alpha}{n+\alpha} \mathbb{E}_0[h(\mathring{z})] \tag{27}$$

Because in the "$M \longrightarrow \infty$" case the prior predictive distribution of the FDP is simply F, also as in a DP, and furthermore the DP is the *only* distribution with these prior and posterior predictive distributions (A. Y. Lo, 1991), this observation functions as a simple demonstration that the FDP converges weakly (see Muliere and Secchi, 2003)to the DP as $M \longrightarrow \infty$. Indeed, this gives us license to consider $M \longrightarrow \infty$ as a well-defined Bayesian model. In comparison to the DP, the posterior mean of the FDP takes $\frac{m}{M}$ from the weight of the prior mean and adds it to weight of the mean of the distinct observed values, making the FDP slightly more responsive to the data than the DP.

If $\alpha \longrightarrow 0$ (an improper prior), then the posterior distribution reduces to a Bayes Bootstrap, with a posterior mean equal to the sample mean. In this case, F has no effect on inferences; for this reason, this may be considered a "noninformative prior."

2.7 Posterior Distribution of the Regression Functional

Recall that our main quantity of interest is the regression functional

$$\beta(\mathring{\mathbf{z}}, \boldsymbol{\theta}) = \left[\sum_{j=1}^{M} \theta_j \mathring{X}_j^\top \mathring{X}_j\right]^{-1} \sum_{j=1}^{M} \theta_j \mathring{X}_j^\top \mathring{y}_j \tag{28}$$

, and that in the posterior distribution,

$$(\theta_1, \ldots, \theta_m, \theta_{m+1}, \ldots, \theta_M) \sim \mathrm{Dir}(\alpha/\mathrm{M} + \mathrm{n}_1, \ldots, \alpha/\mathrm{M} + \mathrm{n}_\mathrm{m}, \alpha/\mathrm{M}, \ldots, \alpha/\mathrm{M}) \tag{29}$$

while the variables $(\mathring{X}_j, \mathring{y}_j)$ for $1 < j \leq m$ are fixed at the observed values $(\tilde{X}_j, \tilde{y}_j)$, and the remaining $(\mathring{X}_k, \mathring{y}_k) \overset{iid}{\sim} F(d\mathring{X}_k, d\mathring{y}_k)$ for $m < k \leq M$.

2.7.1 A convenient representation using Gamma-distributed variables

The above construction with Dirichlet variables is inconvenient mathematically because the θ_j are correlated, and inconvenient computationally because many θ_j may be very small when n is large, causing numerical underflow. Like Lancaster (2006), we will use the well-known fact that any Dirichlet vector $(\theta_1, \ldots, \theta_M) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_\mathrm{M})$ can be equivalently defined as $\theta_j = \frac{g_j}{\sum_{k=1}^{M} g_k}$ using independent Gamma-distributed random variables $g_k \sim \mathrm{Gamma}(\alpha_k, 1)$[9]. We divide out the denominators $\sum_{k=1}^{M} g_k$ in the regression functional, which will render the terms of the sums independent and simplify our analysis considerably.

Then the posterior regression functional is:

$$\beta = \left[\sum_{j=1}^{m} g_j \tilde{X}_j^\top \tilde{X}_j + \sum_{k=m+1}^{M} g_k \mathring{X}_k^\top \mathring{X}_k\right]^{-1} \left[\sum_{j=1}^{m} g_j \tilde{X}_j^\top \tilde{y}_j + \sum_{k=m+1}^{M} g_k \mathring{X}_k \mathring{y}_k\right] \tag{30}$$

with $g_j \sim \mathrm{Gamma}(n_j + \alpha/M, 1)$ for $1 \leq j \leq m$ and $g_k \sim \mathrm{Gamma}(\alpha/M, 1)$ for $m < k \leq M$, with $(\mathring{X}_k, \mathring{y}_k) \overset{iid}{\sim} F(d\mathring{X}_k, d\mathring{y}_k)$, and with $\tilde{X}, \tilde{y}$ fixed.

For more compact notation, we can collect $g_j \sim \mathrm{Gamma}(n_j + \alpha/M, 1)$ for $1 \leq j \leq m$ into the diagonal matrix $\tilde{G}$, and collect $g_k \sim \mathrm{Gamma}(\alpha/M, 1)$ for $m < k \leq M$ into the diagonal matrix $\mathring{G}$. Writing $\tilde{\mathbf{X}} = (\tilde{X}_1^\top, \ldots, \tilde{X}_m^\top)^\top$, $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_m)^\top$, and similarly for the $\mathring{X}_k, \mathring{y}_k$ variables, we have:

[9] We use the "shape-scale" parametrization, so that $\mathbb{E}[g_k] = \mathbb{V}(g_k) = \alpha_k$.

$$\beta = \left[\tilde{\mathbf{X}}^\top \tilde{G} \tilde{\mathbf{X}} + \mathring{\mathbf{X}}^\top \mathring{G} \mathring{\mathbf{X}}\right]^{-1} \left[\tilde{\mathbf{X}}^\top \tilde{G} \tilde{\mathbf{y}} + \mathring{\mathbf{X}}^\top \mathring{G} \mathring{\mathbf{y}}\right] \tag{31}$$

2.7.2 The posterior is a weighted average of the prior and the data

The following quantity behaves much like a Bayes Bootstrap OLS estimator (see Lancaster, 2006) and ultimately carries the influence of the data, and the desirable asymptotic properties of our model:

$$\tilde{\beta} := \left[\tilde{\mathbf{X}}^\top \tilde{G} \tilde{\mathbf{X}}\right]^{-1} \tilde{\mathbf{X}}^\top \tilde{G} \tilde{\mathbf{y}} \tag{32}$$

Let us show how β is a weighted average of $\tilde{\beta}$ and a prior regression parameter $\mathring{\beta}_0$. Without loss of generality, we may suppose that the prior is of the form we suggested earlier, with $\mathring{y}_j = \mathring{X}_j \mathring{\beta}_0 + \mathring{\epsilon}_j$ for some fixed "prior best guess" slope parameter $\mathring{\beta}_0$ and some error term $\mathring{\epsilon}_j$ such that $\mathbb{E}[\mathring{\epsilon}_j | \mathring{\mathbf{X}}, \mathring{G}] = 0$. Then

$$\begin{aligned}
\beta &= \left[\frac{1}{n}\tilde{\mathbf{X}}^\top \tilde{G} \tilde{\mathbf{X}} + \frac{1}{n}\mathring{\mathbf{X}}^\top \mathring{G} \mathring{\mathbf{X}}\right]^{-1} \left(\frac{1}{n}\tilde{\mathbf{X}}^\top \tilde{G} \tilde{\mathbf{X}}\right) \tilde{\beta} + & (33)\\
&\quad \left[\frac{1}{n}\tilde{\mathbf{X}}^\top \tilde{G} \tilde{\mathbf{X}} + \frac{1}{n}\mathring{\mathbf{X}}^\top \mathring{G} \mathring{\mathbf{X}}\right]^{-1} \left(\frac{1}{n}\mathring{\mathbf{X}}^\top \mathring{G} \mathring{\mathbf{X}}\right) \mathring{\beta}_0 + & (34)\\
&\quad \left[\frac{1}{n}\tilde{\mathbf{X}}^\top \tilde{G} \tilde{\mathbf{X}} + \frac{1}{n}\mathring{\mathbf{X}}^\top \mathring{G} \mathring{\mathbf{X}}\right]^{-1} \frac{1}{n}\mathring{\mathbf{X}}^\top \mathring{G} \mathring{\boldsymbol{\epsilon}} & (35)
\end{aligned}$$

We see that the posterior β is a weighted average of 1) the bootstrap-like $\tilde{\beta}$ which contains the contribution of the data, 2) the prior expectation $\mathring{\beta}_0$, and 3) a mean-zero error term $\mathring{\mathbf{X}}^\top \mathring{G} \mathring{\boldsymbol{\epsilon}}$. The weights $\tilde{\mathbf{X}}^\top \tilde{G} \tilde{\mathbf{X}}$ and $\mathring{\mathbf{X}}^\top \mathring{G} \mathring{\mathbf{X}}$ have the form of the classical least squares precision, and so can be interpreted naturally as describing the informativess of the data and the prior, respectively.

2.7.3 Computation of the posterior distribution

Equation (30) suggests a straightforward weighted bootstrap procedure for simulating directly from the posterior distribution. Repeat the following $s = 1, \ldots, S$ times:

1. Draw $g_j^{(s)} \sim \text{Gamma}(n_j + \alpha/M, 1)$ for $1 \leq j \leq m$, and $g_k^{(s)} \sim \text{Gamma}(\alpha/M, 1)$ for $m < k \leq M$, all independently.

2. Draw $(\mathring{X}_k, \mathring{y}_k)^{(s)} \overset{iid}{\sim} F(d\mathring{X}_k, d\mathring{y}_k)$ for $m < k \leq M$, for some user-specified F.

3. Construct $\beta^{(s)}$ as in 30. This is a simple weighted least squares computation, where the weights are given by the Gamma variables g_i, and some of the "data" are $(\mathring{X}_k, \mathring{y}_k)^{(s)}$ drawn from the prior.

To approximate the posterior expectation $\int h(\beta) \Pi_n(d\beta | \mathbf{z}_n)$ for an arbitrary function h, one can use the Monte Carlo estimate $\frac{1}{S}\sum_{s=1}^{S} h(\beta^{(s)})$.

3 Asymptotics

We will consider infinite sequences of datasets $(\mathbf{z}_n)$ of n observations, with each observation drawn independently and identically from some unknown distribution P^*. The sequence of datasets is drawn from some true product distribution $\mathbb{P}^{*\infty}$. For each dataset $\mathbf{z}_n$, we consider a single posterior distribution $\Pi_n(d\boldsymbol{\theta}_n, d\mathring{\mathbf{z}}_n|\mathbf{z}_n)$ of the parameters $\boldsymbol{\theta}_n, \mathring{\mathbf{z}}_n$, as defined above, inducing a posterior distribution of the least squares functional $\boldsymbol{\beta}_n = \boldsymbol{\beta}_n(\boldsymbol{\theta}_n, \mathring{\mathbf{z}}_n)$. We will not assume that the model is correct; there need not exist any $\lambda^* \in \Theta$ such that $P(dz_i|\lambda^*) = P^*$. Our general strategy is as follows:

1. Observe that under plausible regularity conditions such as those in (White, 1980b), the sequence of datasets$(\mathbf{z}_n)$ has certain limit properties $\mathbb{P}^{*\infty}$-almost surely, namely that the OLS coefficients $\hat{\beta}_n$ and the (scaled) robust covariance estimator $n\hat{\mathbb{V}}_n$ are well-defined and consistent estimators. Assume that these conditions hold.

2. Show that for any given sequence $(\mathbf{z}_n)$ with the above limiting properties, the sequence of posterior distributions converges: $\sqrt{n}\left(\beta_n - \hat{\beta}_n\right)|\mathbf{z}_n \xrightarrow{d} \mathcal{N}(0, \lim_{n\to\infty} n\hat{\mathbb{V}}_n)$, under some mild regularity assumptions on the prior. So the regression coefficients β_n converge in posterior distribution to a Gaussian random variable with mean $\hat{\beta}_n$ and covariance $\hat{\mathbb{V}}_n$.

3. [Not yet done] We would like to formally conclude that the posterior mean $\mathbb{E}[\beta_n|\mathbf{z}_n]$ is a $\sqrt{n}$-consistent estimator. As described in van der Vaart (1998, p.17), asymptotic uniform integrability of $\sqrt{n}\beta_n|\mathbf{z}_n$[10] is the necessary and sufficient condition to ensure that the posterior expectation of $\sqrt{n}\beta_n|\mathbf{z}_n$ also converges, $\sqrt{n}\mathbb{E}[\beta_n|\mathbf{z}_n] - \sqrt{n}\hat{\beta}_n \longrightarrow 0$. Since $\hat{\beta}_n$ is $\sqrt{n}$-consistent, we could then conclude that $\sqrt{n}\mathbb{E}[\beta_n|\mathbf{z}_n] - \sqrt{n}\beta^* \longrightarrow 0$.

4. [Not yet done] As above, under further regularity conditions the sequence $\left(n\left(\beta_n - \hat{\beta}_n\right)\left(\beta_n - \hat{\beta}_n\right)^\top\right)$ is asymptotically uniformly integrable, ensuring that that $n\mathbb{V}[\beta_n|\mathbf{z}_n] - n\hat{\mathbb{V}}_n \longrightarrow 0$. This allows us to conclude that the posterior covariance does converge as suggested above.

3.1 Asymptotic regimes

We will consider several asymptotic regimes in which data within each dataset are drawn independently and identically from some unknown distribution P^*, for the usual "triangular array" of data with increasing n. Recall that our model supposes that observations can take up to $M < \infty$ distinct values; m_n represents the number of distinct values in the sample.

- **Regime 0**: As n increases, the number of distinct observations in the sample, m_n, eventually exceeds our model's assumed total number of distinct values M and cause the model to fail.

- **Regime 1**: As n increases, the nondecreasing m_n eventually reaches some true number of distinct values $M^* \leq M$ in the population and stops increasing.

[10] A sequence (W_n) of random variables is asymptotically uniformly integrable if $\lim_{b\to\infty} \limsup_{n\to\infty} \mathbb{E}[|W_n| \mathbf{1}[|W_n| > b]] < \infty$. This essentially requires that the tails of the distribution of W_n contribute little to its mean for sufficiently large n. This is not especially difficult to obtain in realistic settings; a sufficient condition for uniform integrability (and hence asymptotic uniform integrability) is that the sequence be (uniformly) L_p-bounded for $p > 1$ (Cinlar, pp. 72); that is, $\mathbb{E}[|W_n|^p] < \infty$ for all n.

- **Regime 2**: As n increases, the nondecreasing m_n continues increasing without bound; however we also allow M_n to increase with n, setting $M_n = bn$ for some $b \geq 1$.
- **Regime 3**: Informally, one could consider first taking $M_n \longrightarrow \infty$, yielding a Dirichlet Process (Muliere and Secchi, 2003), then allowing n to increase, with m_n nondecreasing however we like.

Regime 0 is unlikely to pose a problem in practice. If, as is often the case in economics, observations are sampled from a population that is large but finite, then it always possible in principle to set M large enough that it will never be exceeded, particularly for sample sizes that economists will actually see. Consequently, we will set Regime 0 aside.

Regime 1 is not problematic for our model. However, a non-increasing m_n may not be particularly relevant to many economic studies, in which the true number of possible data-values M^* is much larger than n, and so m_n increases with n. Here, the problem is similar to that described in Abadie et al. (2017), who also consider increasingly large samples in a finite population. They argue that in most social science applications, the relevant asymptotic regime is not the one where the entire population is sampled.

Regime 2 is similar to that of Abadie et al. (2017), who consider a sequence of increasingly large populations as n increases. Regime 2 need not be taken as an un-Bayesian procedure to increase the prior parameter M as n increases in any given analysis; rather, it is a way of considering a sequence of analyses of increasingly large datasets, none of which has yet exhausted the set of distinct values in the model. Another interpretation of Regime 2 is that it approximates the Dirichlet Process model, a fully Bayesian model to which our model converges as M increases without bound (Muliere and Secchi, 2003), indicating that this approach is valid. Since the Dirichlet Process model is necessarily simulated using finite models like ours as approximations, our analysis will give us a concrete idea of how the simulated regression functional depends on the size (M) of our approximating model and sample size.

Regime 3 considers the case where we essentially use the Dirichlet Process itself (Muliere and Secchi, 2003). The following discussion applies to this case as well.

3.1.1 Lemma: we can focus on the bootstrap-like $\tilde{\beta}$ in any asymptotic regime

Here, we will show that $\left(\frac{1}{n}\mathring{\mathbf{X}}^\top \mathring{G}\mathring{\mathbf{X}}\right)$ and $\frac{1}{\sqrt{n}}\mathring{X}^\top \mathring{G}\mathring{\boldsymbol{\epsilon}}$ in (34) and (35) vanish asymptotically, for any data-generating distribution and any asymptotic regime. This allows us to focus attention on $\tilde{\beta}$.

Consider $\frac{1}{\sqrt{n}}\mathring{X}^\top \mathring{G}\mathring{X} = \frac{1}{\sqrt{n}}\sum_{k=m_n+1}^{M_n} g_k \mathring{X}_k^\top \mathring{X}_k$, where $g_k \overset{iid}{\sim} \text{Gamma}(\alpha/M_n, 1)$ and $(\mathring{y}_k, \mathring{X}_k) \overset{iid}{\sim} F$. All of the following expectations and variances are with respect to the posterior distribution. Notice that $\mathring{X}_k^\top \mathring{X}_k$ is always a $k \times k$ matrix, even in the cluster case. Suppose that in the prior distribution, $\mathbb{E}[\mathring{X}_k^\top \mathring{X}_k]$ and $\mathbb{V}\left(\left(\mathring{X}_k^\top \mathring{X}_k\right)_{i,j}\right)$ are finite and constant across k for $m_n < k \leq M_n$. We will explicitly subscript m_n and M_n to emphasize that in some regimes they increase with n. Then

$$\mathbb{E}[\frac{1}{\sqrt{n}}\sum_{k=m_n+1}^{M_n} g_k \mathring{X}_k^\top \mathring{X}_k] = \frac{1}{\sqrt{n}}(M_n - m_n)\frac{\alpha}{M_n}\mathbb{E}[\mathring{X}_k^\top \mathring{X}_k] \tag{36}$$

Above, we see that the mean diminishes to zero as n increases in Regimes 1, 2, and 3.

For the variance, we consider the i, j^{th} element of the matrix $\mathring{X}_k^\top \mathring{X}_k$:

$$\mathbb{V}\left(\frac{1}{\sqrt{n}}\sum_{k=m_n+1}^{M_n} g_k\left(\mathring{X}_k^\top \mathring{X}_k\right)_{i,j}\right) = \frac{1}{n}\mathbb{V}\left(\mathbb{E}\left[\sum_{k=m_n+1}^{M_n} g_k\left(\mathring{X}_k^\top \mathring{X}_k\right)_{i,j}|\mathring{\mathbf{X}}\right]\right) + \frac{1}{n}\mathbb{E}\left[\mathbb{V}\left(\sum_{k=m_n+1}^{M_n} g_k\left(\mathring{X}_k^\top \mathring{X}_k\right)_{i,j}|\mathring{\mathbf{X}}\right)\right]$$

$$\frac{1}{n}\mathbb{V}\left(\sum_{k=m_n+1}^{M_n}\frac{\alpha}{M}\left(\mathring{X}_k^\top \mathring{X}_k\right)_{i,j}\right) + \frac{1}{n}\mathbb{E}\left[\sum_{k=m_n+1}^{M_n}\frac{\alpha}{M}\left(\mathring{X}_k^\top \mathring{X}_k\right)_{i,j}^2\right] \tag{37}$$

$$\frac{1}{n}\left(\frac{\alpha}{M_n}\right)^2(M_n-m_n)\mathbb{V}\left(\left(\mathring{X}_k^\top \mathring{X}_k\right)_{i,j}\right) + \frac{1}{n}\frac{\alpha}{M_n}(M_n-m_n)\mathbb{E}\left[\left(\mathring{X}_k^\top \mathring{X}_k\right)_{i,j}^2\right] \tag{38}$$

Similar to above, we see that the covariance matrix converges elementwise in probability to zero as n increases, in Regimes 1, 2, and 3.

By Chebyshev's inequality, $\frac{1}{\sqrt{n}}\mathring{X}^\top \mathring{G}\mathring{X} \longrightarrow \mathbf{0}$ in posterior probability.

The same reasoning applies to $\frac{1}{\sqrt{n}}\mathring{X}^\top \mathring{G}\mathring{\boldsymbol{\epsilon}}$.

By the continuous mapping theorem, we have rapid convergence in posterior probability:

$$\sqrt{n}\left(\beta_n - \tilde{\beta}_n\right) \xrightarrow{p} 0 \tag{39}$$

for *any* data sequence with $(\mathbf{z}_n)$ with m_n and M_n behaving as specified in Regimes 1,2, or 3. This will allow us to derive asymptotic results for $\tilde{\beta}_n$ that we can transfer to β_n.

3.1.2 A Bernstein von Mises theorem

Here, we derive a central limit theorem for the Bayesian posterior distribution given a fixed sequence of datasets $((\mathbf{y},\mathbf{X})_n)$. For conciseness, we use $\mathbb{E}[\cdot]$ and $\mathbb{V}(\cdot)$ to denote expectations and variances under the posterior distribution, leaving it implicit that we condition on $\mathbf{y},\mathbf{X}$.

We begin by relating the Bayesian $\tilde{\beta} := \left[\tilde{\mathbf{X}}^\top \tilde{G}\tilde{\mathbf{X}}\right]^{-1}\tilde{\mathbf{X}}^\top \tilde{G}\tilde{\mathbf{y}}$ to the OLS $\hat{\beta}$. First define the residuals of the distinct observed values. $\hat{\tilde{\boldsymbol{\epsilon}}} := \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\beta}$. Then $\tilde{\beta} = \hat{\beta} + \left(\tilde{\mathbf{X}}^\top \tilde{G}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^\top \tilde{G}\hat{\tilde{\boldsymbol{\epsilon}}}$, so we can construct the variable that will have a normal limiting distribution:

$$\sqrt{n}\left(\tilde{\beta} - \hat{\beta}\right) = \left(\frac{1}{n}\tilde{\mathbf{X}}^\top \tilde{G}\tilde{\mathbf{X}}\right)^{-1}\frac{1}{\sqrt{n}}\tilde{\mathbf{X}}^\top \tilde{G}\hat{\tilde{\boldsymbol{\epsilon}}} \tag{40}$$

Taking this apart, we begin by considering sequences of $\frac{1}{\sqrt{n}}\tilde{\mathbf{X}}^\top \tilde{G}\hat{\tilde{\boldsymbol{\epsilon}}} = \frac{1}{\sqrt{n}}\sum_{j=1}^{m_n}\tilde{g}_j \tilde{X}_j^\top \hat{\tilde{\epsilon}}_j$, with $\tilde{g}_j \sim \text{Gamma}(n_j + \alpha/M_n, 1)$. For a central limit theorem, we prefer to work with sums of n terms, so we use the fact that each Gamma-distributed $\tilde{g}_j$ is equal in distribution to the sum $\mathring{g}_j + \sum_{i=1}^{n_j} g_{ji}$ for independent $\mathring{g}_j \sim \text{Gamma}(\alpha/M_n, 1)$ and $g_{ji} \sim \text{Gamma}(1,1)$, so with a slight abuse of notation,

$$\frac{1}{\sqrt{n}}\tilde{\mathbf{X}}^\top \tilde{G}\hat{\tilde{\boldsymbol{\epsilon}}} = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} g_j X_j^\top \hat{\epsilon}_j + \frac{1}{\sqrt{n}}\sum_{j=1}^{m_n}\mathring{g}_j \tilde{X}_j^\top \hat{\tilde{\epsilon}}_j \tag{41}$$

Keeping the data fixed, this has following posterior mean:

$$\bar{\mu}_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j^\top \hat{\epsilon}_j + \frac{\alpha}{M_n} \frac{1}{n} \sum_{i=1}^{m_n} \tilde{X}_i^\top \hat{\tilde{\epsilon}}_i = \frac{\alpha}{M_n} \frac{1}{n} \sum_{i=1}^{m_n} \tilde{X}_i^\top \hat{\tilde{\epsilon}}_i \tag{42}$$

where the first term vanishes due to the orthogonality of X and the residuals. The posterior variance is:

$$\bar{Q}_n = \frac{1}{n} \sum_{i=1}^{n} X_i^\top \hat{\epsilon}_i \hat{\epsilon}_i^\top X_i + \frac{\alpha}{M_n} \frac{1}{n} \sum_{i=1}^{m_n} \tilde{X}_i^\top \hat{\tilde{\epsilon}}_i \hat{\tilde{\epsilon}}_i^\top \tilde{X}_i \tag{43}$$

Notice that the first term of (43) is the center of the robust covariance estimator; we need the second term to vanish.

We assume that we are Regimes 1, 2, or 3 , and that the following hold:

1. Asymptotic non-colinearity: $\frac{1}{n} \sum_{i=1}^{n} X_i^\top X_i \longrightarrow \mathbb{E}^*[X^\top X]$, where $\mathbb{E}^*[X^\top X]$ is positive definite.
2. Bounded mixed fourth moments of X: $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(X_i^\top X_i\right)_{j,k}^2 < \infty$ for all j , k.
3. Second moments of distinct residuals vanish when multiplied by $\frac{\alpha}{M_n} \frac{m_n}{n}$:

$$\left(\frac{\alpha}{M_n} \frac{m_n}{n}\right) \left(\frac{1}{m_n} \sum_{i=1}^{m_n} \tilde{X}_i^\top \hat{\tilde{\epsilon}}_i\right) \longrightarrow 0 \tag{44}$$

4. Fourth mixed moments of distinct residuals vanish when multiplied by $\frac{\alpha}{M_n} \frac{m_n}{n}$:

$$\left(\frac{\alpha}{M_n} \frac{m_n}{n}\right) \left(\frac{1}{m_n} \sum_{i=1}^{m_n} \tilde{X}_i^\top \hat{\tilde{\epsilon}}_i \hat{\tilde{\epsilon}}_i^\top \tilde{X}_i\right) \longrightarrow 0 \tag{45}$$

5. Asymptotic variance is well defined: $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i^\top \hat{\epsilon}_i \hat{\epsilon}_i^\top X_i = Q$, where Q is positive definite and finite.
6. Levy condition: For all i, $\lim_{n\to\infty} \left(n\bar{Q}_n\right)^{-1} \mathbb{V}(w_i) = \mathbf{0}$

Assumption 3) implies $\bar{\mu}_n \longrightarrow 0$. Assumption 4) implies $\bar{Q}_n \longrightarrow Q$. Then Assumptions 3-6 provide the conditions for the multivariate Lindberg-Levy Central Limit Theorem (Greene, 2012), so that as $n \longrightarrow \infty$,

$$\frac{1}{\sqrt{n}} \tilde{\mathbf{X}}^\top \tilde{G} \hat{\tilde{\boldsymbol{\epsilon}}} | \mathbf{y}_n, \mathbf{X}_n \xrightarrow{d} \mathcal{N}(0, Q) \tag{46}$$

Now we turn to the other constituent of $\tilde{\beta}$ and show that it converges in posterior probability. Assumption 1) implies $\mathbb{E}[\frac{1}{n} \tilde{\mathbf{X}}^\top \tilde{G} \tilde{\mathbf{X}}] \longrightarrow \mathbb{E}^*[X^\top X]$. Since $\mathbb{V}\left(\sum_{i=1}^{m} \left(\frac{1}{n} \tilde{\mathbf{X}}^\top \tilde{G} \tilde{\mathbf{X}}\right)_{j,k}\right) = \frac{1}{n^2} \sum_{i=1}^{n} \left(X_i^\top X_i\right)_{j,k}^2 + \frac{1}{n^2} \frac{\alpha}{M_n} \sum_{i=1}^{m_n} \left(\tilde{X}_i^\top \tilde{X}_i\right)_{j,k}^2$, and moreover $0 \leq \sum_{i=1}^{m_n} \left(\tilde{X}_i^\top \tilde{X}_i\right)_{j,k}^2 \leq \sum_{i=1}^{m_n} n_j \left(\tilde{X}_i^\top \tilde{X}_i\right)_{j,k}^2 = \sum_{i=1}^{n} \left(X_i^\top X_i\right)_{j,k}^2$, Assumption 2) implies $\mathbb{V}\left(\sum_{i=1}^{m} \left(\frac{1}{n} \tilde{\mathbf{X}}^\top \tilde{G} \tilde{\mathbf{X}}\right)_{j,k}\right) \longrightarrow$

0. Then by Chebyshev's inequality, 1) and 2) imply the following convergence in posterior probability as $n \longrightarrow \infty$:

$$\frac{1}{n}\tilde{\mathbf{X}}^\top \tilde{G}\tilde{\mathbf{X}} \xrightarrow{p} \mathbb{E}^*[X^\top X] \tag{47}$$

By Slutsky's Lemma and the continuous mapping theorem, (46) and (47) imply:

$$\sqrt{n}\left(\tilde{\beta} - \hat{\beta}\right)|\mathbf{y}_n, \mathbf{X}_n \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}^*[X^\top X]^{-1} Q \mathbb{E}^*[X^\top X]^{-1}\right) \tag{48}$$

By (39), we can substitute β for $\tilde{\beta}$, above.

Discussion of the conditions

Assumption 1), which ensures non-collinearity, Assumption 2, which assumes that the fourth mixed moments of the X distribution are finite, and Assumption 5), which ensures a well-defined asymptotic variance, are all standard assumptions for frequentist robust regression, holding $\mathbb{P}_*^\infty$-almost surely under regularity conditions used by White (1980b). Assumptions 3) and 4) are not standard, but they clearly hold $\mathbb{P}_*^\infty$-almost surely under the White (1980b) regularity conditions in Regime 1 (where m_n eventually reaches an upper bound less than M), Regime 2 (where $M_n = bn$ where $b \geq 0$) and Regime 3 (the Dirichlet Process case). Given assumptions 3) and 4), Assumption 6 is simply the standard Levy condition that each variance is vanishingly small compared to the sum of variances, also holding $\mathbb{P}_*^\infty$-almost surely under standard assumptions. This allows us to conclude that (48) holds $\mathbb{P}_*^\infty$-almost surely.

4 Simulations

4.1 Heteroskedasticity

In very small samples, using our method with a well-informed (i.e. fairly accurate concerning the true slope) prior may substantially improve the coverage rate of Bayesian CIs in comparison to White's HC0 intervals. A poorly-informed prior would probably worsen the coverage rate.

Table 1: Coverage Rates: Bayes vs. HC0 - Small sample, good prior

Coeff	Coverage Bayes Boot	Coverage Hc0
1	1	0.839
2	1	0.833

Using 10000 simulated datasets of size n = 10. Prior slope = 0.9; true slope = 1

In small-to-moderate samples, there may be worse coverage, perhaps due to higher-order asymptotics.

In moderate-to-large samples, even poorly informed priors are washed out and the posterior coefficient estimates and standard errors are numerically very close to OLS with HC0, which has good coverage rates.

Table 2: Coverage Rates: FDM vs. HC0 - Small sample, good prior

Coeff	Coverage Bayes Boot	Coverage Hc0
1	0.907	0.931
2	0.905	0.927

Using 10000 simulated datasets of size n = 100. Prior slope = 0.9; true slope = 1

Table 3: Bayes vs. HC0 - one moderate sample

Coeff	OLS Estimate	Bayes Estimate	HC0 SE	Bayes SE
1	0.009	0.011	0.045	0.044
2	1.014	1.009	0.072	0.072

Using 1 simulated datasets of size n = 500. Prior slope = 0; true slope = 1

5 Discussion

Our assumed sampling process tells us how to handle clustering: simple random sampling implies a heteroskedasticity-robust Bayesian estimator, and clustered sampling implies a cluster-robust Bayesian estimator. This provides a plausible phyiscal explanation of why one would have independent error terms (in the heteroskedasticity case), or error terms that are correlated within clusters but independent between clusters (in the clustered case). We assume that we know very little about the shape of the conditional variance function and even the conditional mean function, which is in many cases more plausible than assuming linearity and homoskedasticity.

The discreteness and finiteness of our model may be viewed as disadvantageous, but they are simple consequences of the assumption that we are sampling from a finite, discrete population, which is usually the case in the social sciences. Many other sampling designs could be contemplated in a framework similar to this one, and we may pursue some of these in a later paper. Perhaps most relevant among these is sampling without replacement, which we have strong reason to believe will yield similar results to the above model, owing to the well-known convergence of the hypergeometric to the multinomial as the number of categories increases (A. Lo, 1987a).

Our findings can be read as shining a critical light on common uses of robust standard errors. Economists frequently use sandwich estimators or their equivalents in cases where sampling is far more complex than the simple random sampling or clustered sampling described here, or in panel data with complex time dependencies in sampling or treatment assignment; or indeed where sampling is not systematic at all. Narrowly construed, the arguments we develop here do not support any particular clustering scheme in these cases; rather, one may need to construct a more complex model of how the data actually came to be.

Authors with Bayesian and other model-based perspectives have raised thoughtful objections to the use of robust standard errors. King and Roberts (2015) have argued that "Robust Standard Errors Expose Methodological Problems They Do Not Fix," to quote their paper's title (see also Freedman, 2006; Leamer, 2010; Meng and Xie, 2014; Pelenis, 2012; Sims, 2010). These authors point out various ways that robust standard errors, whenever they substantially differ from classical standard errors (and traditional Bayesian credible intervals), provide evidence that one's model is

inefficient at best, and possibly misspecified to the point of being grossly misleading about the scientific questions at stake. We agree, as did White (1980a, p.828) in the paper that introduced robust standard errors to many economists. Like White, however, we argue that robust standard errors and, we add, their Bayesian equivalents, can be part of a larger workflow in which the model is critiqued and amended. The Bayesian framework we use here is particularly amenable to this perspective, since we may decide that the vector of squared-error-minimizing linear coefficients β^* is not a useful estimand, or that its Bayesian counterpart β is not a useful summary of the posterior distribution, without changing the underlying Finite Dirichlet Process model, which remains weakly consistent for the true P^* under relatively broad conditions (Muliere and Secchi, 2003). Indeed, the decomposition (6) shows us that if $\mathbb{V}^{\hat{HC0}}(\hat{\beta}_2) > \hat{\mathbb{V}}^{OLS}(\hat{\beta}_2)$, then the linear model's fit worsens specifically in the tails of the X-distribution, either due to nonlinearity or to heteroskedasticity (or both): if we seek ways to improve on the linear model, then robust standard errors (Bayesian or otherwise) show us a place to start.

However, not every model can be a mere stepping-stone to another model; at some point, one must draw inferences from a model that one considers good enough. In many cases, a fully Bayesian model which provides$\sqrt{n}$-consistent heteroskedasticity-robust and cluster-robust regression estimators and confidence intervals, even if it is inefficient, is good enough.

References

Abadie, A., Athey, S., Imbens, G., & Wooldridge, J. (2017). When Should You Adjust Standard Errors for Clustering? *arXiv:1710.02926 [econ, math, stat]*. arXiv: 1710.02926 [econ, math, stat]

Abadie, A., Athey, S., Imbens, G., & Wooldridge, J. (2020). Sampling-Based versus Design-Based Uncertainty in Regression Analysis. *Econometrica*, *88*(1), 265–296. doi:10.3982/ECTA12675

Aitkin, M. (2008). Applications of the Bayesian Bootstrap in Finite Population Inference. *Journal of Official Statistics*, *24*(1), 21–51.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics*, 27.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., ... Zhang, K. (2019). Models as Approximations I: Consequences Illustrated with Linear Regression. *arXiv:1404.1578 [stat]*. arXiv: 1404.1578 [stat]

Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372. doi:10.3368/jhr.50.2.317

Carter, A. V., Schnepel, K. T., & Steigerwald, D. G. (2017). Asymptotic Behavior of a t -Test Robust to Cluster Heterogeneity. *The Review of Economics and Statistics*, *99*(4), 698–709. doi:10.1162/REST_a_00639

Chamberlain, G., & Imbens, G. (2003). Nonparametric Applications of Bayesian Inference. *Journal of Business & Economic Statistics*, *21*(1), 12–18. doi:10.1198/073500102288618711

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. doi:10.1007/978-1-4612-4380-9_41

Eicker, F. (1967). Limit Theorems for Regressions with Unequal and Dependent Errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, p. 24). Berkeley: University of California Press.

Ferguson, T. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, *1*(2), 209–230.

Freedman, D. (1981). Bootstrapping regression models.pdf. *The Annals of Statistics*, *9*(6), 1218–1228.

Freedman, D. (2006). On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician*, *60 (Nov)*, 299–302. doi:10.1198/000313006X152207

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2020). *Bayesian Data Analysis* (Third). Boca Raton, Florida: CRC Press.

Gelman, A., Hill, J., & Vehtari, A. (2021). *Regression and Other Stories* (First). Cambridge, U.K.: Cambridge University Press.

George E.P. Box, & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis* (Wiley Classics Library). Delhi: Wiley.

Geweke, J. (1993). Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics*, *8*(S1), S19–S40. doi:10.1002/jae.3950080504

Greene, W. H. (2012). *Econometric Analysis* (7 (International)). Essex, England: Pearson Education Limited.

Hjort, N. L. (1994). Bayesian approaches to non- and semiparametric density estimation. In *Fifth Valencia. International Meeting on Bayesian Statistics*, Alicante, Spain.

Hoff, P., & Wakefield, J. (2012). Bayesian sandwich posteriors for pseudo-true parameters. *arXiv:1211.0087 [stat]*. arXiv: 1211.0087 [stat]

Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 5.1, pp. 221–223). Berkeley: University of California Press.

Ishwaran, H., & Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, *30*(2), 269–283. doi:10.2307/3315951

Karabatsos, G. (2016). A Dirichlet process functional approach to heteroscedastic-consistent covariance estimation. *International Journal of Approximate Reasoning*, *78*, 210–222. doi:10.1016/j.ijar.2016.07.008

King, G., & Roberts, M. E. (2015). How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It. *Political Analysis*, *23*(2), 159–179. doi:10.1093/pan/mpu015

Kleijn, B. J. K., & van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, *34*(2), 837–877. doi:10.1214/009053606000000029

Lancaster, T. (2006). *A note on bootstraps and robustness*. doi:10.1920/wp.cem.2006.0406

Leamer, E. E. (2010). Tantalus on the Road to Asymptopia. *Journal of Economic Perspectives*, *24*(2), 31–46. doi:10.1257/jep.24.2.31

Liang, K.-Y., & Zeger, S. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, *71*(1), 13–22.

Lo, A. (1987a). A Bayesian Bootstrap for Finite Population. *The Annals of Statistics*, *16*(4), 1685–1695.

Lo, A. (1987b). A Large Sample Study of the Bayesian Bootstrap. *Annals of Statistics*, *15*(1), 16.

Lo, A. Y. (1991). A characterization of the Dirichlet process. *Statistics & Probability Letters*, *12*(3), 185–187. doi:10.1016/0167-7152(91)90075-3

MacKinnon. (2019). How cluster-robust inference is changing applied econometrics. *Canadian Journal of Economics/Revue canadienne d'économique*, *52*(3), 851–881. doi:10.1111/caje.12388

MacKinnon, J., & Webb, M. (2019). When and how to deal with clustered errors in regression models. *Queen's Economics Department Working Paper*, *1421*.

Meng, X.-L., & Xie, X. (2014). I Got More Data, My Model is More Refined, but My Estimator is Getting Worse! Am I Just Dumb? *Econometric Reviews*, *33*(1-4), 218–250. doi:10.1080/07474938.2013.808567

Muliere, P., & Secchi, P. (2003). Weak Convergence of a Dirichlet-Multinomial Process. *gmj*, *10*(2), 319–324. doi:10.1515/GMJ.2003.319

Norets, A. (2015). Bayesian regression with nonparametric heteroskedasticity. *Journal of Econometrics*, *185*(2), 409–419. doi:10.1016/j.jeconom.2014.12.006

Pelenis, J. (2012). *Bayesian Semiparametric Regression* (tech. rep. No. 285). Institut für Höhere Studien. Wien.

Pelenis, J. (2014). Bayesian regression with heteroscedastic error density and parametric mean function. *Journal of Econometrics*, *178*, 624–638. doi:10.1016/j.jeconom.2013.10.006

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Institute of Mathematical Statistics Lecture Notes - Monograph Series* (pp. 245–267). doi:10.1214/lnms/1215453576

Poirier, D. J. (2011). Bayesian Interpretations of Heteroskedastic Consistent Covariance Estimators Using the Informed Bayesian Bootstrap. *Econometric Reviews*, *30*(4), 457–468. doi:10.1080/07474938.2011.553542

Polpo, A., Louzada, F., Rifo, L. L. R., Stern, J. M., & Lauretto, M. (Eds.). (2015). *Interdisciplinary Bayesian Statistics: EBEB 2014*. doi:10.1007/978-3-319-12454-4

Ramsey, F. (1926). Truth and Probability. In *The Foundations of Mathematics and other Logical Essays* (p.156–198). New York: Harcourt, Brace and Company.

Rice, K., Lumley, T., & Szpiro, A. (2008). Trading bias for precision: Decision theory for intervals and sets. *UW Biostatistics Working Paper Series*, (336), 29.

Robert, C. P., Chopin, N., & Rousseau, J. (2009). Harold Jeffreys's Theory of Probability Revisited. *Statistical Science*, *24*(2). doi:10.1214/09-STS284

Rubin, D. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, *9*(1), 130–134.

Shalizi, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, *3*(none). doi:10.1214/09-EJS485

Sims, C. (2010). Understanding Non-Bayesians.

Stan. (2022). Stan Development Team.

Startz, R. (2012). Bayesian Heteroskedasticity-Robust Standard Errors. *UC Santa Barbara Working Papers*.

Szpiro, A. A., Rice, K. M., & Lumley, T. (2010). Model-robust regression and a Bayesian "sandwich" estimator. *The Annals of Applied Statistics*, *4*(4). doi:10.1214/10-AOAS362. arXiv: 1101.1402

van der Vaart, A. (1998). *Asymptotic Statistics* (First). New York: Cambridge University Press.

West, M. (1984). Outlier Models and Priors in Bayesian Linear Regression. *Journal of the Royal Statistical Society*, *46*(3), 431–439.

White, H. (1980a). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817–838.

White, H. (1980b). Using Least Squares to Approximate Unknown Regression Functions. *International Economic Review*, *21*(1), 149. doi:10.2307/2526245

Wood, S. (2017). Ch2: Linear Mixed Models. In *Generalized Additive Models* (Second). Chapman and Hall/CRC.

Yin, G. (2009). Bayesian generalized method of moments. *Bayesian Analysis*, *4* (2). doi:10.1214/09-BA407

Zhang, X. (2008). *A Very Gentle Note on the Construction of Dirichlet Process*. Australian National University.

Zhao, Y. (2015). Bayesian Linear Regression with Conditional Heteroskedasticity. *UC3M Working Papers*, 24.

Appendix A:

Review of current Bayesian thinking about heteroskedasticity

In current Bayesian thinking about heteroskedasticity, two views appear to be prevalent: first, that it is usually safe to assume homoskedasticity, since heteroskedasticity usually does not matter in practice; and second, that if heteroskedasticity does matter in practice, then the true conditional variance $\mathbb{V}^*(y|X)$ must be modeled directly (using some model $\mathbb{V}(y|X,\lambda)$), as part of modeling the entire conditional distribution $\mathbb{P}^*(dy|X)$ (using some model $\mathbb{P}(dy|X,\lambda)$). In both approaches, the distribution of X is usually considered ancillary and therefore is not modeled. Our model fits within a third, smaller Bayesian literature in which we account for heteroskedasticity, but as a byproduct of modeling the joint distribution of y and X.

We will show using simulations and some theory that assuming homoskedasticity when heteroskedasticity is present can lead to credible intervals that are spuriously narrow — professing greater precision or certainty than a frequentist model, and not due to a better model or judicious priors, but rather due to an incorrect model that ignores the uncertainty coming from heteroskedasticity. We are more sympathetic to the view that the conditional variance function must be modeled (we will call this the "conditional modeling" view), but note that existing Bayesian conditional models make assumptions about heteroskedasticity that can be difficult to defend, and that can have the same consequences as ignoring heteroskedasticity if incorrect, as we show. Admittedly, we do not consider ways in which these models might be patched if their deficiencies became apparent to the researcher.

For the rest of this section we write $X_i^\top = \mathbf{x}_i$ as a vector, to clarify that we refer to the heteroskedastic case. Since clustered covariances usually include heteroskedasticity, our criticisms below also apply to them, but clustered covariances also present complexities of their own that we will not consider in this section for lack of space.

5.1 Ignoring or downplaying heteroskedasticity

According to a popular regression textbook with a Bayesian perspective, heteroskedasticity "does not affect what is typically the most important aspect of a regression model, which is the information that goes into the predictors and how they are combined..." and consequently heteroskedasticity is generally a minor issue. (Gelman, Hill, et al., 2021, p.154)

To examine this common view, let us consider a deceptively simple Bayesian model of het-

eroskedasticity to which we will return in later sections. For now, assume the variance function is fixed, $\mathbb{V}(y_i|\mathbf{x}_i) = \sigma^2(\mathbf{x}_i)$, and observations $(y_i, \mathbf{x}_i)$ are independently Normally distributed:

$$[y_i|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\gamma}] \sim \mathcal{N}\left(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2(\mathbf{x}_i|\boldsymbol{\gamma})\right) \tag{49}$$

For illustrative purposes, we use an improper uniform prior on the regression coefficients $\boldsymbol{\beta}$. This impropriety will not affect our point; proper priors will yield increasingly similar conclusions as data increases.[11] Collecting the variances into a diagonal matrix $\Omega = \text{diag}\{\sigma^2(\mathbf{x}_i)\}$, standard manipulations (George E.P. Box and Tiao, 1973) show that the posterior distribution of $\boldsymbol{\beta}$ given $\mathbf{y}, X, \Omega$ is a Normal distribution with mean

$$\hat{\beta}_{GLS} := \left(\mathbf{X}^\top \Omega^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \Omega^{-1} \mathbf{y} \tag{50}$$

and covariance $\hat{V}_{GLS} := \left(\mathbf{X}^\top \Omega^{-1} \mathbf{X}\right)^{-1}$; that is, the posterior distribution is centered on the (frequentist) Generalized Least Squares estimator, which happens to be asymptotically efficient. [12]

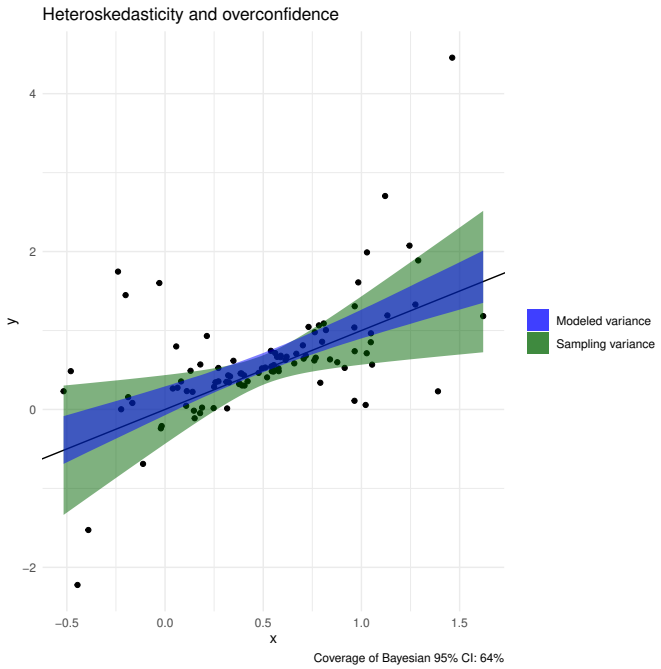
Crucially, we see that in this model, both the posterior mean and the posterior variance of $\boldsymbol{\beta}$ depend on Ω, which encodes the heteroskedasticity. Heteroskedasticity therefore affects not only the width of one's credible intervals, but also one's point estimates. As emphasized by Leamer (2010), this is not a minor effect, and it runs counter to Gelman and Hill's claim that heteroskedasticity does not affect important aspects of the model. Indeed, as we shall see in the next section, in cases where Ω is not known and must be modeled using a prior distribution, a poorly-chosen model for Ω can cause serious problems.

Still, in the above model, we also see that assuming homoskedasticity, so that $\Omega = \sigma^2 I_n$ with some variance σ^2, results in the classical OLS coefficients $\hat{\beta} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ and classical covariance $\hat{\mathbb{V}}_{OLS} := \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$; for our argument, little changes if σ^2 is given a prior and estimated rather than known. Since $\hat{\boldsymbol{\beta}}$ is usually consistent for $\boldsymbol{\beta}^*$ even when heteroskedasticity exists, a pragmatic Bayesian might still insist that little is to be lost from ignoring known heteroskedasticity, except perhaps efficiency. This is not the case: one also risks overconfidence.

As we have shown in (6), classical confidence intervals that use $\hat{\mathbb{V}}_{OLS}$ can be spuriously narrow in the presence of heteroskedasticity and clustered correlations. To illustrate an extreme case, we show the results of a Bayesian OLS (homoskedastic) regression with a diffuse prior when data are in fact strongly heteroskedastic.

[11] As usual, the distribution of $\mathbf{x}$ is assumed to be ancillary and is not modeled. This is not the same as assuming that $\mathbf{x}$ is non-random.

[12] As Norets (2015) discusses, this model's Normality assumption often leads to excellent frequency properties even when the data are not in fact normally distributed.



Above, we see that the Bayesian 95% probability region for $\mathbb{E}[y|\mathbf{x}, \beta] = \mathbf{x}^\top \beta$, (blue) is much narrower than the frequentist 95% sampling band for the Bayes estimator $\mathbf{x}^\top \hat{\beta}$ (green), and it happens that the Bayesian 95% posterior credible intervals capture the true slope coefficient only 64% of the time.[13] Here, Bayesian estimates that ignore heteroskedasticity are asymptotically consistent, but Bayesian inferences have much greater inferential certainty than a frequentist would think is warranted. There may be times when such certainty from a Bayesian model is justified, but this is not one of them, we argue. Here, the Bayesian model's extra certainty comes not from a better model or judicious use of prior information, but rather from simply ignoring an important source of uncertainty: heteroskedasticity.

5.2 Conditional modeling of heteroskedasticity

It is clear that if a Bayesian knows there is heteroskedasticity, then her model must incorporate it somehow. In this subsection, we discuss recent work in which which $\mathbb{V}^*(y|\mathbf{x})$ is modeled explicitly, $\mathbb{E}^*[y_i|\mathbf{x}_i]$ is assumed to be linear in $\mathbf{x}$, and the distribution of $\mathbf{x}$ is considered ancillary and not modeled; we call this the "conditional modeling" approach, following (Gelman, Carlin, et al., 2020).

Bayesian conditional models for heteroskedasticity are numerous and we do not attempt to review or categorize them all here. [14] However, we can arrange many Bayesian conditional models

[13] We estimated this with 10,000 simulated regressions. Currently, the coverage rate is calculated for a frequentist OLS model, which we'd expect to behave extremely similarly to a Bayesian OLS model with a diffuse prior, but is much faster to simulate. Different priors could be contemplated, but the usual recommended informative prior centered at a zero slope will not improve the coverage rate here, where the true regression slope is positive.

[14] An extensive Bayesian literature is devoted to modeling heteroskedasticity, particularly in financial time-series, where the conditional variance is known as "volatility," and is essential to asset-pricing models such as the Black-Scholes model. We do not attempt to review this literature here.

along a specrum of how "wiggly"[15] $\mathbb{V}(y|\mathbf{x})$ is. We take readers on a brief tour of three kinds of heteroskedastic models: one that is not wiggly enough, some cutting-edge models that can be made just wiggly enough, and one that is altogether too wiggly.

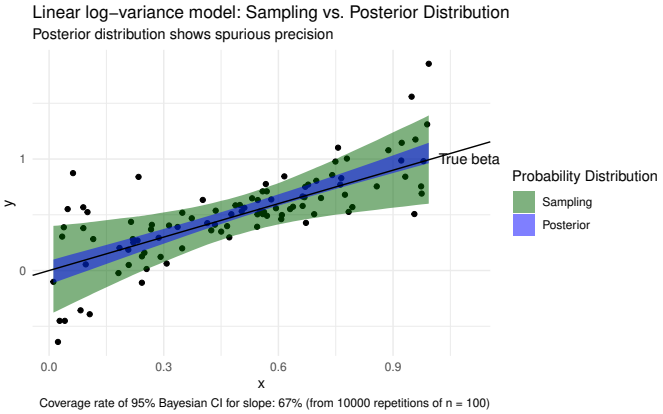
We will show in simulations that Bayesian models that assume too little or too much wiggliness can behave poorly indeed from a frequentist perspective, noting that it can be difficult or impossible in most real-world cases to verify whether one's wiggliness assumptions are correct.

5.2.1 Too little wiggliness

When Ω is not precisely known, but one suspects that heteroskedasticity might be lurking, perhaps the simplest way to extend (49) is with something like the following:

$$\sigma^2(\mathbf{x}_i) = \exp\{\mathbf{x}_i^\top \boldsymbol{\gamma}\} \tag{51}$$

giving the unknown parameters $\boldsymbol{\gamma}$ some prior. We choose this particular model because it appears to be the most common starting-point for modeling heteroskedasticity in the Bayesian (and frequentist) literature, and it is the default in Stan[16], one of the most widely-used Bayesian MCMC engines and modeling platforms. Eliding the details, we can estimate the posterior distribution of this model using Stan's default settings.



Unfortunately, the result hardly looks better than in the previous subsection. The posterior distribution is far narrower than the frequentist sampling distribution, and the Bayesian's 95% credible intervals for the slope coefficients contain the true slope only 67% of the time, whereas the frequentist HC0 intervals have 92% coverage. So at least by the frequentist standard, the Bayesian is still overconfident, although there has been a slight improvement over the case where the Bayesian ignored heteroskedasticity entirely.

The origin of this disagreement is simple: the Bayesian model is wrong about the underlying process. By choosing $\sigma^2(\mathbf{x}_i) = \exp\{\mathbf{x}_i^\top \boldsymbol{\gamma}\}$, we have put all prior (and posterior) probability on a space of variance functions that does not contain the true variance function, which here happens to be $\sigma^{2(*)}(x) = (0.1 + 2(x - 0.5)^2)^2$. Since this is a problem with the likelihood, not the priors, more carefully chosen priors and more data will not generally fix the problem.

[15] Perhaps measured in terms of integrated squared derivatives, as in the spline literature, where "wiggliness" is in fact a techical term.(Wood, 2017)

[16] Stan. (2022). Stan Development Team, Modeling.

5.2.2 Just enough wiggliness

If unsatisfied with the above model, a Bayesian can turn to highly flexible models that use infinitely many parameters (naturally, these are called "nonparametric" models) to describe the conditional variance, in the hope that this much larger model-space will contain the true model, or a sufficiently good approximation of it. To take some leading recent examples of non-parametric or semiparametric models, Norets (2015), Pelenis (2014), and Zhao (2015) essentially extend model (49) with the unknown $\sigma^2(\mathbf{x}_i)$ modeled as a (perhaps transformed) linear combination of infinitely many basis functions $\psi_j(\mathbf{x}_i)$ with coefficients γ_j:

$$\sigma^2(\mathbf{x}_i) = \sum_{j=1}^{\infty} \gamma_j \psi_j(\mathbf{x}_i) \tag{52}$$

For example, in Norets, ψ_j are Bernstein polynomials. The above authors prove that their models produce consistent estimates and asymptotically correct confidence intervals under a very wide range of data-generating processes, even when the true conditional distribution is not at all Gaussian. As Norets explains, the key to such general success is that in large samples, Bayesian posterior distributions tend to concentrate around models that are "close" to the data-generating process in terms of Kullback-Liebler (KL) divergence[17]; and for Gaussian models like (49), the KL divergence is minimized by matching the conditional mean and variance in the model to the conditional mean and variance of the data-generating process — for any data-generating process. So if one is interested in robust Bayesian models of the mean and variance, Gaussian models tend to perform well (Kleijn and van der Vaart, 2006).

There are some difficulties with the above modeling approach. The desirable frequency properties rest heavily on smoothness assumptions about the true $\sigma^{(*)2}(\mathbf{x})$ that are difficult to verify (or even intuit) in most applications and are too technical to state precisely here, although they involve placing stochastic bounds on partial deriviatives of $\sigma^2(\mathbf{x})$ (eg. see Norets, 2015, p.410) in a way that matches the partial derivatives of the unknown true $\sigma^{(*)2}(\mathbf{x})$. These assumptions are far stronger than the regularity conditions that underpin frequentist robust models (compare to White (1980b), where no assumptions about derivatives are required).

Compounding this problem, when $\mathbf{x}_i$ has more than three dimensions or so, the nonparametrically modeled $\sigma^2(\mathbf{x}_i)$ may converge to the truth relatively slowly and be extremely computationally demanding, particularly if smoothness parameters are unknown and must also be given priors.[18] Clustered correlations appear nearly impossible to handle as nonparametric functions in this way, since in this case $\mathbb{V}(\mathbf{y}_c|X_c)$ is a matrix function whose dimension and form may change between clusters c.

5.2.3 Too much wiggliness

Next, we will show how someone who is dissatisfied with the above models and who is searching for arbitrarily flexible models of heteroskedasticity might construct a homoskedastic model as a limiting case. Indeed, the use of homoskedastic Student's-t models for heteroskedasticity (e.g., Geweke, 1993) appears to date back at least to Jeffreys' seminal *Theory of Probability* (Robert et al., 2009) and remains popular; it is currently a first-page Google search result for "Bayesian Heteroskedastic

[17] See Shalizi (2009)

[18] This is part of why we do not present MCMC simulations of these procedures here.

regression."[19][20]. Unfortunately, we will see that in simulations, the Student's t model can perform extremely poorly when faced with realistic patterns of heteroskedasticity. Ultimately, the Student's-t model illustrates the practical dangers not only of using homoskedastic models when the truth is heteroskedastic, but also of making ill-founded (non-)smoothness assumptions concerning $\sigma^2(\mathbf{x}_i)$.

Suppose that as above, each "error term" $\epsilon_i := y_i - \mathbf{x}_i^\top \beta$ is distributed independently as $\epsilon_i \sim N(0, \sigma^2(\mathbf{x}_i))$, and that to model $\sigma^2(\mathbf{x}_i)$ we use a linear combination of basis functions ψ_j with unknown coefficients $\gamma_j > 0$, similar to the semiparametric models in the previous section:

$$\sigma^2(\mathbf{x}_i) = \tau \sum_{j=1}^{J} \gamma_j \psi_j(\mathbf{x}_i) \tag{53}$$

For reasons that will become apparent, we include $\tau > 0$ as a scale factor; for now we will treat τ as fixed. Suppose that our basis consists of indicator functions $\psi_j(\mathbf{x}_i) = 1(\mathbf{x}_i \in A_j)$ for some partition $A_1, \ldots, A_J$ of the $\mathbf{x}$-space. This is the simplest form of B-spline. Suppose we use independent Inverse-Gamma priors on the coefficients, $\gamma_j \sim \text{InvGamma}(\frac{\nu}{2}, \frac{\nu}{2})$ for some fixed $\nu > 0$ (this enforces $\sigma^2(\mathbf{x}_i) > 0$), and because we want to accommodate nearly arbitrarily wiggly forms of heteroskedasticity, we make J much, much larger than any dataset we're likely to see. There is nothing intrinsically wrong with more parameters than data-points in a Bayesian analysis, and an extremely large J might even seem necessary if we want a fine partition and $\mathbf{x}$ has a few dozen dimensions.[21]

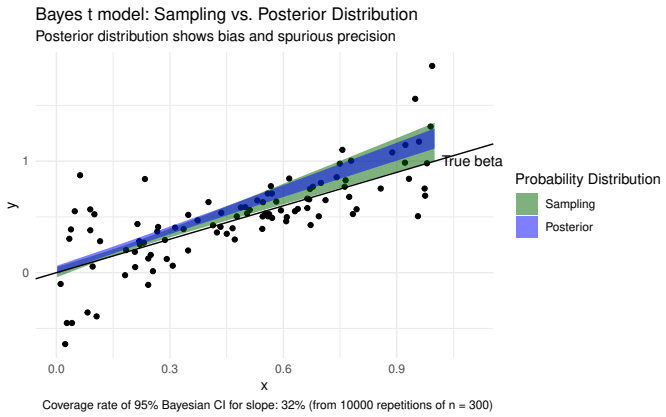
Consider what happens if data arrives and each observed $\mathbf{x}_i$ falls in its own member of the partition, with no member of the partition containing multiple observations (our main point would hold even with multiple observations per member, though the notation and qualifications would proliferate). In this case, each error term ϵ_i for $i = 1, \ldots, n$ is being modeled as $\epsilon_i \sim N(0, \sigma_i^2)$ independently, with each observation seeming to get its own variance $\sigma_i^2 = \tau\gamma_{j_i}$, with $\gamma_{j_i} \sim \text{InvGamma}(\frac{\nu}{2}, \frac{\nu}{2})$ independently.

Interestingly, integrating out each σ_i^2 while holdng the other parameters fixed, one sees that $\epsilon_i \sim t_\nu(\mathbf{x}^\top\beta, \tau^2)$, a Student's t distribution with ν degrees of freedom and a scale parameter equal to τ^2. This means that in this model, which is supposed to be almost arbitrarily heteroskedastic, each each error-term is effectively being modeled as *homoskedastic* for the given dataset. In many cases, this homoskedasticity leads to problems.

[19] As of Feb 2022. Stable link:
https://web.archive.org/web/20220214205338/https://jrnold.github.io/bayesian_notes/heteroskedasticity.html

[20] This model is also sometimes recommended for outlier-robust regression, a different kind of robustness than we are concerned with (see West (1984))

[21] For example, if each vector $\mathbf{x}$ comprises 20 variables, and we wish to partition each variable's values into 10 subcategories, then there are $J = 10^{20}$ elements of the partition in total.



In 10,000 simulations like the one pictured above, we find that 95% credible intervals based on the Student's t model capture the true slope (pictured in black) only 32 % of the time. Meanwhile, the frequentist HC-robust CIs (not pictured) capture the true value 94% of the time. In the picture, we see that the Bayesian estimate is strongly biased upwards, and the posterior distribution is also narrower than the sampling distribution of the Bayes estimate. This bias comes not from the prior (which is very diffuse in this case, although proper), but rather from the poorly-chosen Student's t likelihood, which does not actually model the heteroskedasticity in our chosen error-distribution. A maximum-likelihood estimator based on the same Student's t model would have the same pathologies. Because the problem is with the likelihood, not the prior, it only gets worse with more data.

A short explanation for this particular problem is that the Student's t model is akin to a weighted least squares model, where observations get less weight if they are farther from the regression line; more precisely, in Gibbs sampling the observations are iteratively re-weighted according to their inverse squared distance from the previous iteration's fitted regression line (see West (1984) for details). We made the true error distribution asymmetric as well as heteroskedastic, with a longer tail in the negative direction than in the positive direction, and with both tails of the error distribution getting longer as x increases (keeping the error term mean-zero). With this kind of error distribution, as x increases, the model systematically gives less and less weight to observations in the lengthening tail below the fitted regression line, biasing the regression line upwards. It may be somewhat unfair to test the Student's t model with asymmetric data, but the world is often unfairly asymmetrical; and as far as we are aware, no warnings about asymmetry appear in any of the standard Bayesian texts on the Student's t model.

What we have described is not peculiar to the Student's t model. When faced with heteroskedasticity, one would expect similar pathologies from any homoskedastic model, not just the Student's t.

5.3 Conclusion of review

This ladder of models, from one that ignores heteroskedasticity, to one with insufficient wiggliness of the conditional variance, to one with just enough wiggliness, to one with too much wiggliness, can be seen as a stylized story of how a researcher might expand a model to better fit the data. We have shown what happens when this story goes wrong at various stages: Bayesian 95% credible intervals may capture the true value much less often than 95%, being both biased and professing

much greater inferential certainty than a frequentist would think is warranted — not because the Bayesian model makes better use of information, but simply because the Bayesian model incorrectly accounts for heteroskedasticity.

This leads us to search for other Bayesian models that do not hinge so crucially on assumptions about the wiggliness of conditional variance function.

Technical Appendix: The posterior distribution of a Finite Dirichlet Process

Why we can't just use Bayes' Theorem

Bayes' Theorem applies only when the prior distribution dominates the posterior distribution: when every set that has zero probability under the prior also has probability zero under the posterior. (Polpo et al., 2015) For our model, this doesn't hold. In this subsection, I give a concrete illustration of what goes wrong when one tries to (improperly) apply Bayes' Theorem to the simplest possible non-dominated model.

Consider the simplest Bayesian model of all: a single error-free observation z of an unknown parameter $\mathring{z}$ in $\mathbb{R}$:

$$z \sim \delta_{\mathring{z}}(dz) \tag{54}$$

with some prior $\mathring{z} \sim \pi(\mathring{z})d\mathring{z}$ that is a continuous distribution with full support on $\mathbb{R}$. Recall that the Dirac delta measure $\delta_{\mathring{z}}(dz)$ simply puts all probability at $\mathring{z}$; it operates on measurable sets A as $\delta_{\mathring{z}}(A) = 1(\mathring{z} \in A)$; or equivalently $\int \delta_{\mathring{z}}(dz)h(z) = h(\mathring{z})$ for any measurable function h. To be clear, $\delta_{\mathring{z}}(dz)$ is a measure, *not* a function on the real domain, and *cannot* be written as $\delta(z - \mathring{z})$ whatever the engineering textbooks may do.

Intuitively, we know what the posterior distribution must be: a Bayesian simply must believe an error-free observation. That is, the posterior of $\mathring{z}$ must put all probability at the observed value z, which we can write as $\mathring{z}|z \sim \delta_z(d\mathring{z})$.

But how can we prove this? Note that the prior is continuous, so for any fixed z, $\pi(\mathring{z} = z) = 0$, so the prior does not dominate the posterior that we have just intuited.

Suppose we naively plug everything into Bayes' Theorem, as though $\delta_{\mathring{z}}(dz)$ were a continuous likelihood function on the real domain. This would give us the following expression, which is nonsense:

$$\text{" } \Pr(\mathring{z} \in A|z) = \frac{\int_A \delta_{\mathring{z}}(dz)\pi(\mathring{z})d\mathring{z}}{\int \delta_{\mathring{z}}(dz)\pi(\mathring{z})d\mathring{z}}\text{"} \tag{55}$$

One problem with the above is that expressions like "$\int_A \delta_{\mathring{z}}(dz)\pi(\mathring{z})d\mathring{z}$" don't clarify what we are integrating or in what order; and it does not make sense to directly integrate $\delta_{\mathring{z}}(dz)$ with respect to $\mathring{z}$. But suppose we forge on, noting that we want the numerator to represent someting like $p(z, \mathring{z})$ so we re-interpret the numerator as the well-defined expression $\int d\mathring{z}\, \pi(\mathring{z}) \int_A \delta_{\mathring{z}}(dz) = \int d\mathring{z}\, \pi(\mathring{z}) 1(\mathring{z} \in A) = \pi(A)$; but by the same logic, the denominator is $\pi(\mathbb{R}) = 1$, so the posterior probability is simply $\pi(A)$ — the same as the prior probability. Nothing has been learned from the error-free data!

Undeterred, we might change our approach and try the discrete form of Bayes' Theorem:

$$\text{" } \Pr(\mathring{z} \in A | z = \tilde{z}) = \frac{\Pr(z = \tilde{z} | \mathring{z} \in A) \Pr(\mathring{z} \in A)}{\Pr(z = \tilde{z})} \text{"} \tag{56}$$

Each factor of the above equation now makes sense, but the whole equation does not, because it divides by zero: for fixed $\tilde{z}$,

$$\Pr(z = \tilde{z}) = \int \pi(d\mathring{z}) \int \delta_{\mathring{z}}(dz) 1(z = \tilde{z}) = \int \pi(d\mathring{z}) 1(\mathring{z} = \tilde{z}) = 0 \tag{57}$$

At this point, the reader may agree that we must give up on applying Bayes' Theorem to this model. If Bayes' Theorem *defined* the posterior distribution, this would all be bad news for Bayesians: it would be impossible to simply tell a Bayesian an error-free fact. It would be especially bad news for this paper, since our model is just a linear combination of delta-functions.

But Bayes' Theorem does not define the Bayesian posterior distribution; rather, it is just a useful way to calculate the Bayesian posterior distribution, whenever the theorem applies — which here, it obviously doesn't. So we must turn to a more general definition of the Bayesian posterior distribution.

The general definition of a posterior distribution

The underlying intuition is this: in elementary probability theory, probability densities are related in the following way:

$$\int_{\theta \in C} \int_{y \in A} p(\theta|y) p(y) dy d\theta = \int_{\theta \in C} \int_{y \in A} p(\theta, y) dy d\theta = \int_{y \in A} \int_{\theta \in C} p(y|\theta) p(\theta) d\theta dy \tag{58}$$

So intuitively, once one has specified the sampling model $p(y|\theta)$ and the prior $p(\theta)$ on the right-hand-side, and therefore also specified $p(y) = \int p(y|\theta)p(\theta)d\theta$, the posterior density $p(\theta|y)$, the only remaining term, can be *defined* as whatever valid probability density function satisfies the above equation. Crucially, this definition avoids dividing by $p(y)$.

A slightly more measure theoretic approach relieves us from assuming that densities exist at all, and therefore covers the model used in this paper. For a formal treatment, see (Polpo et al., 2015); here we merely sketch the main idea.

Suppose our model of the data $\mathbf{z}_n$ given the unknowns $(\boldsymbol{\theta}, \mathring{\mathbf{z}})$ is characterized by the probability transition kernel $P(d\mathbf{z}_n | \boldsymbol{\theta}, \mathring{\mathbf{z}})$, and our prior over the unknowns is a probability measure $\Pi_0(d\boldsymbol{\theta}, d\mathring{\mathbf{z}})$.

The prior predictive distribution is defined as $\hat{P}_0(A) := \int_{(\theta, \mathring{\mathbf{z}}) \in \Theta} \Pi_0(d\boldsymbol{\theta}, d\mathring{\mathbf{z}}) P(A | \boldsymbol{\theta}, \mathring{\mathbf{z}})$.

The posterior distribution of $(\mathring{\mathbf{z}}, \boldsymbol{\theta})$ is then defined as any probability kernel $\Pi_n(d\boldsymbol{\theta}, d\mathring{\mathbf{z}} | \mathbf{z}_n)$ that satisfies the following equality for all $\mathbf{z}_n \in \mathbf{A}, \mathring{\mathbf{z}} \in \mathbf{B}, \boldsymbol{\theta} \in C$, where $\mathbf{A}, \mathbf{B}, C$ are measurable sets:

$$\int_{\mathbf{z}_n \in \mathbf{A}} \hat{P}_0(d\mathbf{z}_n) \int_{\mathring{\mathbf{z}} \in \mathbf{B}, \boldsymbol{\theta} \in C} \Pi_n(d\boldsymbol{\theta}, d\mathring{\mathbf{z}} | \mathbf{z}_n) = \int_{\mathring{\mathbf{z}} \in \mathbf{B}, \boldsymbol{\theta} \in C} \Pi_0(d\boldsymbol{\theta}, d\mathring{\mathbf{z}}) \int_{\mathbf{z} \in \mathbf{A}} P(d\mathbf{z}_n | \boldsymbol{\theta}, \mathring{\mathbf{z}}) \tag{59}$$

6 The posterior distribution of a Finite Dirichlet Process: a brute-force approach

If we guess the form of $\Pi_n(d\boldsymbol{\theta}, d\mathring{\mathbf{z}} | \mathbf{z}_n)$ by heuristic reasoning, we can use the identity (59) to check whether our guess is correct. Our strategy is to test whether a "postulated" measure $\hat{\mathbb{P}}(A, \mathbf{B}, C)$,

corresponding to the left-hand side of (59), equals a "target" $\mathbb{P}(A, \mathbf{B}, C)$ on the right-hand side of (59).

Target integral (single observation, z)

$$\mathbb{P}(A, \mathbf{B}, C) := \int_{\mathring{\mathbf{z}} \in \mathbf{B}, \boldsymbol{\theta} \in C} \Pi_0(d\boldsymbol{\theta}, d\mathring{\mathbf{z}}) \int_{z \in A} P(dz|\boldsymbol{\theta}, \mathring{\mathbf{z}}) \tag{60}$$

$$= \int_{\boldsymbol{\theta} \in C} \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_M) \int_{\mathring{\mathbf{z}} \in \mathbf{B}} \prod_{k=1}^{M} F(d\mathring{z}_k) \int_{z \in A} \sum_j \theta_j \delta_{\mathring{z}_j}(dz) \tag{61}$$

$$= \int_{\boldsymbol{\theta} \in C} \sum_j \theta_j \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_M) \int_{\mathring{\mathbf{z}} \in \mathbf{B}} \prod_{k=1}^{M} F(d\mathring{z}_k) \int_{z \in A} \delta_{\mathring{z}_j}(dz) \tag{62}$$

$$= \int_{\boldsymbol{\theta} \in C} \sum_j \theta_j \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_M) \int_{\mathring{\mathbf{z}} \in \mathbf{B}} \prod_{k=1}^{M} F(d\mathring{z}_k) 1(\mathring{z}_j \in A) \tag{63}$$

using the fact that $\int_A \delta_{x_0}(dx) h(x) = \int \delta_{x_0}(dx) 1(x \in A) h(x) = 1(x_0 \in A) h(x_0)$

$$= \int_{\boldsymbol{\theta} \in C} \sum_j \theta_j \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_M) \left(\prod_{k \neq j}^{M} \int_{B_k} F(d\mathring{z}_k) \right) \int_{\mathring{z}_j \in B_j} F(d\mathring{z}_j) 1(\mathring{z}_j \in A) \tag{64}$$

$$= \int_{\boldsymbol{\theta} \in C} \sum_j \theta_j \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_M) \left(\prod_{k \neq j}^{M} F(B_j) \right) F(B_j \cap A) \tag{65}$$

Note:

$$\theta_j \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_M) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \theta_j^{(\alpha_j + 1) - 1} \prod_{k \neq j} \theta_k^{\alpha_k - 1} d\boldsymbol{\theta} \tag{66}$$

$$= \left(\frac{\Gamma(\sum_k \alpha_k)}{\Gamma(1 + \sum_k \alpha_k)} \frac{\Gamma(\alpha_j + 1)}{\Gamma(\alpha_j)} \right) \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_M) \tag{67}$$

$$= \frac{\alpha_j}{\sum_k \alpha_k} \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_M) \tag{68}$$

using the identity $\frac{\Gamma(x+1)}{\Gamma(x)} = x$.
So,

$$\mathbb{P}(A, \mathbf{B}, C) = \sum_j \frac{\alpha_j}{\sum_k \alpha_k} \mathrm{Dir}(C|\alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_M) \left(\prod_{k \neq j}^{M} F(B_j) \right) F(B_j \cap A) \tag{69}$$

Verifying the postulated posterior given a single observation

We postulate that

$$\Pi_n(d\boldsymbol{\theta}, d\mathring{\mathbf{z}}|z) = \sum_{j=1}^{M} \frac{\alpha_j}{\sum_k \alpha_k} \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_M) \delta_z(d\mathring{z}_j) \prod_{k \neq j}^{M} F(d\mathring{z}_k) \tag{70}$$

The summation of M terms comes from the fact that there are M ways to assign one of the M latent points $\mathring{z}_k$ to the observed value z.

Verify:

$$\hat{\mathbb{P}}(A, \mathbf{B}, C) = \int_{z \in A, \mathring{\mathbf{z}} \in \mathbf{B}, \boldsymbol{\theta} \in C} \sum_{j=1}^{M} \frac{\alpha_j}{\sum_k \alpha_k} \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_M) \delta_z(d\mathring{z}_j) \left(\prod_{k \neq j}^{M} F(d\mathring{z}_k) \right) F(dz) \tag{71}$$

$$= \sum_{j=1}^{M} \frac{\alpha_j}{\sum_k \alpha_k} \int_{\boldsymbol{\theta} \in C} \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_M) \int_{z \in A} F(dz) \int_{\mathring{z}_j \in B_j} \delta_z(d\mathring{z}_j) \prod_{k \neq j}^{M} \int_{\mathring{z}_k \in B_k} F(d\mathring{z}_k) \tag{72}$$

$$= \sum_{j=1}^{M} \frac{\alpha_j}{\sum_k \alpha_k} \int_{\boldsymbol{\theta} \in C} \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_M) \int_{z \in A} F(dz) 1(z \in B_j) \prod_{k \neq j}^{M} F(B_k) \tag{73}$$

$$= \sum_{j=1}^{M} \frac{\alpha_j}{\sum_k \alpha_k} \int_{\boldsymbol{\theta} \in C} \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_M) F(A \cap B_j) \prod_{k \neq j}^{M} F(B_k) \tag{74}$$

So, $\hat{\mathbb{P}} = \mathbb{P}$; our postulated posterior distribution is correct.

Simplifying using exchangeability

Recall that we use a symmetric Dirichlet prior, $\alpha_j = \alpha/M$, and the posterior distribution is

$$\Pi_n(d\boldsymbol{\theta}, d\mathring{\mathbf{z}}|z) = \frac{1}{M} \sum_{j=1}^{M} \mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_M) \delta_z(d\mathring{z}_j) \prod_{k \neq j}^{M} F(d\mathring{z}_k) \tag{75}$$

This summation of M terms represents the M ways to assign a latent point to the observed point; but there is no physical difference between these assignments, so we may (correctly) suspect that this summation is needless bookkeeping. Here we show that we can indeed dispense with the summation by merely picking a single term of the sum to represent our posterior distribution, without altering our posterior inferences in any meaningful way.

In any application, we will draw inferences only about functions of the unknown parameters $\mathring{z}_1, \ldots, \mathring{z}_M$ and $\theta_1, \ldots, \theta_M$ that are ***exchangeable***: functions T such that $T(\mathring{z}_1, \ldots, \mathring{z}_M, \theta_1, \ldots, \theta_M) = T(\mathring{z}_{\varsigma(1)}, \ldots, \mathring{z}_{\varsigma(M)}, \theta_{\varsigma(1)}, \ldots, \theta_{\varsigma(M)})$ for any permutation $\varsigma(\cdot)$ of the indices $1, \ldots, M$. Exchangeability codifies our assumption that the indices themselves do not encode any meaningful information about the latent points (such as an ordering), except to distinguish one from another.

Considering the posterior expectation of $h(T)$ for arbitrary h, it is clear that the distribution of an exchangeable $T(\mathring{z}_1, \ldots, \mathring{z}_M, \theta_1, \ldots, \theta_M)$ induced by the probability measure $\mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_M)\delta_z(d\mathring{z}_j)\prod_{k\neq j}^{M} F(d\mathring{z}_k)$ is identical to the distribution of $T(\mathring{z}_{\varsigma(1)}, \ldots, \mathring{z}_{\varsigma(M)}, \theta_{\varsigma(1)}, \ldots, \theta_{\varsigma(M)})$ induced by $\mathrm{Dir}(d\boldsymbol{\theta}\varsigma|\alpha_{\varsigma(1)}, \ldots, \alpha_{\varsigma(j)} + 1, \ldots, \alpha_{\varsigma(M)})\delta_z(d\mathring{z}_{\varsigma(j)})\prod_{\varsigma(k)\neq\varsigma(j)}^{M} F(d\mathring{z}_{\varsigma(k)})$, where $\boldsymbol{\theta}\varsigma$ denotes a permutation of the elements of $\boldsymbol{\theta}$.

It follows that the posterior distribution of any exchangeable function $T(\mathring{z}_1, \ldots, \mathring{z}_M, \theta_1, \ldots, \theta_M)$ given a single datapoint can be represented by a distribution of the form

$$\mathrm{Dir}(d\boldsymbol{\theta}|\alpha_1 + 1, \ldots, \alpha_j, \ldots, \alpha_M)\delta_z(d\mathring{z}_1)\prod_{k\neq 1}^{M} F(d\mathring{z}_k) \tag{76}$$