Bayesian Rashomon Sets for Model Uncertainty: A Critical Comparison

Aparajithan Venkateswaran
Microsoft
Redmond, WA 98052
apara.vnkat@gmail.com

Anirudh Sankar
Department of Economics
Stanford University
Stanford, CA 94305
anirudh.sankar91@gmail.com

Arun G. Chandrasekhar
Department of Economics
Stanford University
Stanford, CA 94305
arungc@stanford.edu

Tyler H. McCormick
Department of Statistics
University of Washington
Seattle, WA 98101
tylermc@uw.edu

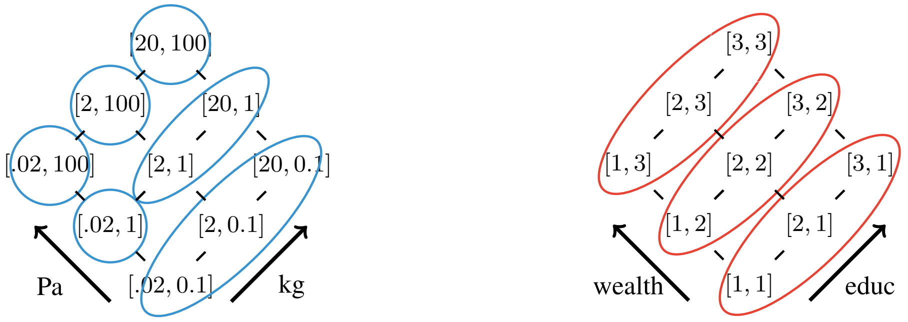
Abstract

In statistical analyses, both observational and experimental, understanding how outcomes vary with covariates is crucial. Traditional methods like Bayesian and frequentist regression, regression trees, and model averaging partition data into homogeneous pools to summarize outcomes. However, these methods either focus on a single optimal partition or sample from all possible partitions, often missing high-quality ones or including low-support partitions. A recently developed Bayesian approach, Rashomon Partition Sets (RPSs), enumerates partitions with posterior densities close to the *maximum a posteriori (MAP)* partition, capturing uncertainty among high-evidence partitions. RPSs adhere to two principles, scientific coherence, and simplicity, by using a minimax optimal ℓ_0 prior without additional dependence assumptions. In this paper, we critically compare the RPS approach with three commonly used alternatives: Bayesian Model Averaging, Bayesian/frequentist regularization, and Causal Random Forests.

1 Introduction

Take a basic question: how does a (continuous or discrete) outcome vary with combinations of the covariates? There are many examples: how do various health interventions affect health outcomes, how does technology adoption depend on incentives and demographics, how does the performance of a material depend on environmental conditions, or how do human reactions to a situation vary based on background and cognitive state? There are many potential explanations of heterogeneity (models) and the covariates of interest can interact in complex and unpredictable ways (e.g. a material is resilient to heat, except in the presence of a particular chemical; each additional piece of information helps a person make a better decision until too much becomes overwhelming). Existing approaches generally fall into two categories. They either (i) search for a single "optimal" model under some assumptions about the association between covariates (e.g. LASSO) or (ii) attempt to *sample* from the *entire* set of possible models (e.g. Bayesian Model Averaging, BMA, or random forests). Both these approaches ignore the reality that many models will be indistinguishable from a statistical perspective, especially with correlation structure in covariates, despite very different implications for policy or science. We evaluate an alternative using the idea of Rashomon Sets from Leo Breiman's 2001 "Statistical modeling: The two cultures" paper (and the highly related "Occam's Razor" philosophy

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) e.g., $\mathbb{E}[y_i|(2,1)] = \mathbb{E}[y_i|(20,1)]$

(b) e.g., $\mathbb{E}[y_i|(2,3)] \neq \mathbb{E}[y_i|(2,2)]$

Figure 1: Two partitions, each representing a distinct model for heterogeneity in the outcome, y_i. The left panel shows heterogeneity in acceleration of a cube measured after dropping it a uniform gravitational field with drag as the mass of the cube and external pressure changes. The right panel shows Banerjee and Duflo (2010)'s model for interest rates as a function of borrower's wealth and education when there are high administrative costs relative to loan amount. There are two features, each with three levels. Each circle represents a combination unique combination. Removing edges corresponds to creating disjoint components in the graph.

from Madigan and Raftery's seminal 1994 paper for BMA)–to enumerate and explore a small number of high (posterior) probability models, called the Rashomon Partition Set (RPS) because each item in the RPS partitions the factorial space of covariates using a tree-like geometry (Breiman, 2001; Madigan and Raftery, 1994; Venkateswaran et al., 2024).

2 Rashomon Partition Sets

Both modern and classical statistical tools, either implicitly or explicitly, address this problem using *partitions*. They partition observations into "pools" where outcomes are similar within the pool but differ across pools, then compute a summary (or fit a model) to pool. Some models, such as Bayesian or frequentist tree models, are explicit about these partitions. Others, however, do so implicitly (e.g. a regression with a single binary covariate posits heterogeneity between people in one group versus the other, and homogeneity otherwise). Banerjee et al. (2021) introduced Hasse diagrams as a geometric representation of partitions. Figure 1 gives an example of two partitions. The Hasse defines partitions by removing (splitting on) edges to form disjoint connected components, which guarantees that all sets in a partition contain only "connected" feature combinations. Distinguishing meaningful from spurious heterogeneity then amounts to evaluating partitions. Estimation strategies and algorithms privilege different partitions in search of the partition that captures meaningful complexity without sacrificing power, but the partitions capture the root of the heterogeneity directly.

In a recent working paper, we propose *Rashomon Partition Sets (RPSs)* (Venkateswaran et al., 2024). The RPS consists of all partitions that are close to the *maximum a posteriori (MAP)* partition in terms of posterior density. Venkateswaran et al. (2024) bound the difference between posterior quantities computed using the entire posterior and using only the RPS and use an ℓ_0 prior, which is minimax optimal but does not impose additional restrictions on the association between covariates. Restrictions on the universe of partitions ensure that each partition corresponds to a scientifically plausible explanation. For experiments, policymakers can then weigh additional considerations (e.g., cost, equity, privacy) in choosing which policies from the RPS to implement. RPSs also yield insights to generate new scientific theories. Looking across models in the RPS, one can build an archetype of feature combinations that appear consistently and have consistent effects on the outcome, regardless of the structure imposed on other covariates by other high posterior partitions.

More formally, suppose that there are n units (or individuals) and each has M features. The feature matrix is given by $\mathbf{X}_{1:n,1:M}$ and outcomes are $\mathbf{y} \in \mathbb{R}^n$. Every feature has R possible values, partially ordered. Let $\mathcal{K}$ be the set of all $K = R^M$ unique feature combinations. Each combination of features $k \in \mathcal{K}$ can be represented in a dummy binary matrix $\mathbf{D}$ with entries $D_{ik} = 1$ if and only if observation i has feature combination k. The dataset is $\mathbf{Z} := (\mathbf{y}, \mathbf{X})$. So, $\mathbf{y} = \mathbf{D}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\beta_k = \mathbb{E}[Y_i \mid D_{ik} = 1]$ is the expected outcome in the population given the feature combination and ϵ_i

is some idiosyncratic mean-zero residual. A *partition*, Π, in the space of all partitioning models, $\mathcal{P}$, is a model of heterogeneity such that for every pool $\pi \in \Pi$, possibly a singleton, if feature combinations $k, k' \in \pi$, then $\beta_k = \beta_{k'}$. The posterior given the data $\mathbf{Z}$ is $\mathbb{P}(\Pi \mid \mathbf{Z})$. Let $\mathcal{P}^\star \subseteq \mathcal{P}$ be the set of permissible partitions that obey some permissibility rules.

Definition 1 (Rashomon Partition Set (RPS))**.** *For some posterior probability threshold $\tau \in (0,1)$, define the* Rashomon Partition Set *relative to a reference partition Π_0, $\mathcal{P}_\tau(\Pi_0)$, as*

$$\mathcal{P}_\tau(\Pi_0) = \{\Pi \in \mathcal{P}^\star : \ \mathbb{P}(\Pi \mid \mathbf{Z}) \geq (1-\tau) \cdot \mathbb{P}(\Pi_0 \mid \mathbf{Z})\}. \tag{1}$$

The RPS relative to Π_0 is the set of partitions that have a similar or higher posterior value than the reference. If Π^{MAP} is the *maximum a posteriori (MAP)* partition, then define $\mathcal{P}_\tau(\Pi^{\text{MAP}})$. That is, take the RPS as the set of partitions that are sufficiently close to the posterior of the MAP partition. To construct the RPS, begin with an initialization partition, then enumerate all models with posteriors at least as high as this initialization partition (which by definition includes Π^{MAP}). Then, construct the RPS by moving down the list of partitions, ordered by posterior value, until reaching $(1-\tau)\mathbb{P}(\Pi^{\text{MAP}}|\mathbf{Z})$.

Given a posterior over the partition models, the posterior over the effects of various feature combinations, possibly pooled, on the outcome of interest conditional on the partition models in this set. So $\mathbb{P}(\boldsymbol{\beta} \mid \mathbf{Z}, \mathcal{P}_\tau) = \sum_{\Pi \in \mathcal{P}_\tau} \mathbb{P}(\boldsymbol{\beta} \mid \mathbf{Z}, \Pi)\mathbb{P}(\Pi \mid \mathbf{Z}, \mathcal{P}_\tau)$, and analogously for measurable functions of $\boldsymbol{\beta}$. The posterior for $\boldsymbol{\beta}$ restricted to the RPS is, of course, not the same as the distribution over all possible partitions, $P_{\boldsymbol{\beta}|\mathbf{Z}}(\boldsymbol{\beta})$. Venkateswaran et al. (2024) characterize the uniform approximation error of the posterior distribution of $\boldsymbol{\beta}$, and measurable functions of it, restricting to the RPS. This result does not depend on a specific prior over partitions, but a prior is required to find RPS in practice. Venkateswaran et al. (2024) propose an ℓ_0 prior that assumes only sparsity in heterogeneity and is robust to any potential correlation structure between covariates. Venkateswaran et al. (2024) show that the ℓ_0 prior is minimax optimal and calculate bounds on the size of the RPS and enumerate the entire RPS.

3 Comparison of RPSs to existing approaches

Bayesian Model Averaging Our goal is to represent uncertainty amongst scientifically plausible explanations of heterogeneity in an outcome. Scientifically plausible explanations should be supported by the observed data. RPSs, therefore, are *enumerative*, meaning that they consist of a list of *all* partitions with posterior density close to the MAP partition. In their seminal paper developing an *Occam's Window* approach to Bayesian Model Averaging (BMA) for graphical models, Madigan and Raftery (1994) express a similar philosophy:

> [standard BMA] does not accurately represent model uncertainty. Science is an iterative process in which competing models of reality are compared on the basis of how well they predict what is observed; models that predict much less well than their competitors are discarded. Most of the models in [standard BMA] have been discredited [...] so they should be discarded.

Madigan and Raftery (1994) use this logic to justify an approach that favors high posterior, simple models but constructs this set by sampling from the posterior. Sampling includes models that are consistent with observed data and others that are highly unlikely, with the latter comprising most of the posterior mass (Moulton, 1991). RPSs break from the literature that relies on sampling, instead using the MAP as an anchor and listing models with similar posterior density. Further, sampling explores the space of partitions/models, whereas the domain of science is explanations. Without restrictions on the space, there could be multiple partitions that correspond to a single explanation, and the number of partitions corresponding to each explanation will depend on the complexity of the explanation. Sampling in the space of partitions, then, weights scientific explanations differentially based on the number of partitions corresponding to that explanation.

ℓ_1 Regularization. We compare RPSs to widely used regularization-based approaches, such as Bayesian or frequentist Lasso, which use the ℓ_1 penalty, compared to ℓ_0 in the RPS. First, the ℓ_1 penalty requires irrepresentability: that there is limited correlation between the regressors so that the support may be consistently recovered (Zhao and Yu, 2006). The conditions required for the ℓ_1 to consistently select the "right" partition are likely not met in most substantively interesting settings

Table 1: Each cell shows the fraction of CRF trees inside the RPS (within parentheses are absolute counts).

	# trees = 20	# trees = 50	# trees = 100
	(# permissible = 1.29)	(# permissible = 2.67)	(# permissible = 10.32)
$\epsilon = 0.1$ ($\lvert\mathcal{P}_\theta\rvert = 7.46$)	0% (0)	0% (0)	0% (0)
$\epsilon = 0.2$ ($\lvert\mathcal{P}_\theta\rvert = 46.6$)	0% (0)	0% (0)	0% (0)
$\epsilon = 0.3$ ($\lvert\mathcal{P}_\theta\rvert = 126.54$)	0.41% (0.52)	0.91% (1.15)	3.35% (4.24)
$\epsilon = 0.5$ ($\lvert\mathcal{P}_\theta\rvert = 823.81$)	0.16% (1.29)	0.32% (2.67)	1.25% (10.32)

(e.g. with unknown and potentially large correlation between variables). Second, the Bayesian lasso means that the ℓ_1 penalty corresponds to priors $\mathbb{P}(\boldsymbol{\alpha})$ that are i.i.d. Laplace on every dimension k. The main philosophical problem is that there is no reason to place the meta-structure that the marginal differences between adjacent variants should come from an i.i.d. distribution. In fact, one might think that the basic science or social science dictates *exactly the opposite*. Independence means that a marginal increase in dosage of drug A, holding fixed B and C at some level, is thought to be *independent* of increasing A holding fixed B and C at (potentially very similar) different levels.

Take a setting with four features. Each feature takes on four ordered factors including the control (which corresponds to zero, when the feature is inactive), $\{0, 1, 2, 3\}$. There are sixteen different feature profiles: $2^4 = 16$ possible combinations of active and inactive features. The control corresponds to the profile where all features are inactive. Our data-generating process groups all feature combinations in a given profile into the same pool. We will assume that the following profiles have a non-zero outcome: $\beta_{(0,0,0,1)} = 4.4$, $\sigma^2_{(0,0,0,1)} = 1$, $\beta_{(0,1,0,0)} = 4.3$, $\sigma^2_{(0,1,0,0)} = 1$, $\beta_{(0,1,0,1)} = 4.45$, $\sigma^2_{(0,1,0,1)} = 1$, $\beta_{(1,0,1,0)} = 4.5$, $\sigma^2_{(1,0,1,0)} = 1.5$, $\beta_{(1,1,1,1)} = 4.35$, $\sigma^2_{(1,1,1,1)} = 1$. All other feature profiles have outcome $\beta = 0$ and variance $\sigma^2 = 1$. The feature profile $(1, 0, 1, 0)$ is the best, however, the other four profiles listed above are very close. We generated data with $n_k = 30$ data points per feature combination. The outcomes were drawn from $\mathcal{N}(\beta_i, \sigma_i^2)$. We averaged the results over $r = 100$ simulations. Figure 2 tells us how often the true best feature profile is present in RPS as a function of the threshold ϵ. Lasso selects the data-generating model only about half the time, though as we increase ϵ we see that the true data-generating model is nearly always included in the RPS.

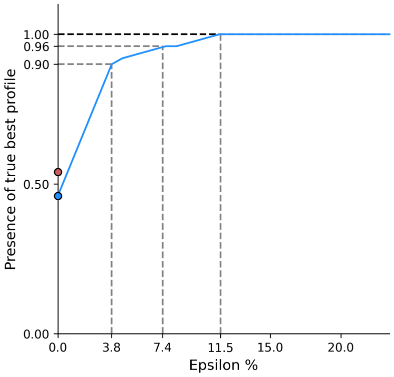


Figure 2: Simulation results. The plot shows often the true best profile is discovered as we increase the Rashomon threshold in the blue curve. With just $\epsilon \approx 0.038$, we recover the true best profile in the Rashomon set about 90% of the time. The red dot corresponds to how often Lasso recovers the true best.

Causal Random Forests (Wager and Athey, 2018). There are two fundamental differences. The first is geometric. Causal Random Forests (CRFs) use regression trees, which impose a hierarchy between variables that is not supported by the (partially ordered) data and, thus, are not interpretable as explanations. Hasse diagrams, in contrast, are the natural geometry for partially ordered sets.

Second, CRFs bootstrap samples over the data propagate this uncertainty. The trees sampled as part of this process create a "forest" are, by definition, random draws given the data. Given a different set of data, the distribution of likely trees would change. They are also not guaranteed to be optimal or nearly optimal. If the goal is to identify interpretable explanations of heterogeneity, then sampling randomly is very unlikely to produce high quality trees. With RPS, by definition, all models in the set are of high posterior.

We simulate data with four features, the first being a binary treatment variable. The second feature takes on 3 ordered levels and the last two features have 4 ordered levels: $\beta_{(1,1,1:2,1:3)} = 2$, $\beta_{(1,1,1:2,4)} = 4$, $\beta_{(1,1,3:4,1:3)} = 2$, $\beta_{(1,1,3:4,4)} = 0$, $\beta_{(1,2,1:2,1:3)} = 3$, $\beta_{(1,2,1:2,4)} = 5$, $\beta_{(1,2,3:4,1:3)} = 7$, $\beta_{(1,2,3:4,4)} = 1$, $\beta_{(1,3,1:2,1:3)} = 1$, $\beta_{(1,3,1:2,4)} = -1$, $\beta_{(1,3,3:4,1:3)} = -1$, $\beta_{(1,3,3:4,4)} = -2$. We generate $n_k = 10$ data points per feature combination. In the treatment group, we drew outcomes from a $\mathcal{N}(\beta_k, 1)$ distribution.The results are presented in Table 1. The vast majority of partitions sampled by CRFs are not scientifically coherent (permissible) partitions and thus cannot be interpreted as plausible explanations. This result is not specific to CRFs and would hold for any algorithm using unrestricted trees. Consequently, the number of trees that are in the RPS is also very small, meaning that, although averaging over trees has appealing asymptotic properties, the trees included in particular sample are unlikely to be high-quality explanations. This result is, again, not specific to CRFs but highlights the value of exploring uncertainty by enumerating high-quality explanations compared to sampling.

4 Conclusion

In summary, methods such as Bayesian or frequentist regression search for a single optimal model under some assumptions about the covariate structure. However, in large datasets, many models may be statistically indistinguishable even though their scientific implications are very different. RPSs overcome this limitation in two ways. First, it uses an ℓ_0 prior that is agnostic to any covariate dependence. Second, it enumerates all near-optimal models, making it feasible to infer the true scientific model. On the other hand, methods such as random forests or BMA sample from the entire set of models. While this avoids the “single model” paradigm, there is no guarantee that the sampled models are the near-optimal models. In fact, Moulton (1991) suggests that models not supported by the data account for most of the posterior. In contrast, RPSs, by definition, consist of all near-optimal models.

References

Abhijit Banerjee, Arun G Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew O Jackson, Harini Kannan, Francine N Loza, Anirudh Sankar, Anna Schrimpf, et al. Selecting the most effective nudge: Evidence from a large-scale experiment on immunization. Technical report, National Bureau of Economic Research, 2021.

Abhijit V Banerjee and Esther Duflo. Giving credit where it is due. *Journal of Economic Perspectives*, 24(3):61–80, 2010.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

David Madigan and Adrian E Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428): 1535–1546, 1994.

Brent R Moulton. A Bayesian approach to regression selection and estimation, with application to a price index for radio services. *Journal of Econometrics*, 49(1-2):169–193, 1991.

Aparajithan Venkateswaran, Anirudh Sankar, Arun G Chandrasekhar, and Tyler H McCormick. Robustly estimating heterogeneity in factorial data using rashomon partitions. *arXiv preprint arXiv:2404.02141*, 2024.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.