The matryoshka doll prior – principled penalization in Bayesian selection

Andrew Womack*,§, Daniel Taylor-Rodríguez†§¶, and Claudio Fuentes‡

Abstract. This paper introduces a general and principled construction of model space priors with a focus on regression problems. The proposed formulation regards each model as a “local” null hypothesis whose alternatives are the set of models that nest it. A simple proportionality principle yields a natural isomorphism of model spaces induced by conditioning on predictor inclusion before or after observing data. This isomorphism produces the Poisson distribution as the unique limiting distribution over model dimension under mild assumptions. We compare this model space prior theoretically and in simulations to widely adopted Beta-Binomial constructions and show that the proposed prior yields a “just-right” penalization profile.

MSC2020 subject classifications: Primary 62F15; secondary 62J99.

Keywords: Model space prior, Bayesian variable selection, Regression models, infinite exchangeability.

1 Introduction

The problem of identifying suitable sets of predictors in regression models of the form

$$y = X_0 \beta_0 + X \beta + \epsilon,$$

where $$\epsilon \sim N_n(0, I/\tau),$$ (1)

has been extensively studied in the statistical literature. In (1), the matrix $$X_0$$ corresponds to $$p_0$$ covariates that are necessarily included in the mean structure of $$y$$ (one of which must be the intercept term). The matrix $$X$$ contains the additional $$p$$ covariates whose importance is to be tested, leading to $$2^p$$ models. Given the increasing number of research questions in which $$p$$ is large or growing with $$n$$ (especially the case when $$p >> n$$), there has been a rekindled interest in the selection/testing problem. In this paradigm, it is imperative to identify strategies that yield consistent selection while remaining computationally viable.

Our chief interest is to design a principled model space prior that appropriately controls the probability of the set of models containing $$k$$ covariates as $$p$$ increases. The learning rates associated with Bayes’ factors work in conjunction with the model space prior to produce posterior inference. To illustrate, assume that we are comparing

*Department of Statistics, Rice University, aw88@rice.edu
†Department of Mathematics and Statistics, Portland State University, dantayrod@pdx.edu
‡Department of Statistics, Oregon State University, fuentesc@stat.oregonstate.edu
§First coauthors
¶Corresponding author
The matryoshka doll model space prior

a finite “true” model $M$ to another finite model $M'$ using local priors on the model-specific parameters.\footnote{For examples using local priors, see \cite{17, 11, 2, 3, 12, 6, 4, 16, 7, 5}. For examples of non-local priors, see \cite{8, 9, 1}.} If $M$ is not nested in $M'$, then the learning rate of $M$ versus $M'$ is exponential in $n$, which makes learning $M$ in the pair-wise comparison relatively easy. However, if $M$ is nested in $M'$, then the Bayes’ Factor learning rate of $M$ against $M'$ is a power of $n$, which makes it difficult for $M$ to beat $M'$ even with large sample sizes. Therefore, when considering the entire model space, the combinatorial size of the number of models nesting the true model hinders learning when model complexity is not properly penalized.

The literature discussing priors over the model space is sparse, but important progress has been made. Ley & Steel \cite{10} and Scott & Berger \cite{13} have advocated for Beta-Binomial($a, b$) priors on indicator variables for covariate inclusion. In both of these works the authors recommend setting $a = b = 1$, which provides a Bayesian version of a multiplicity correction. Unfortunately, as Johnson & Rossell \cite{9} note, this multiplicity correction lacks the strength to induce posterior concentration on a finite true model when local priors are used. This concentration issue arises because infinitely exchangeable Beta-Binomial priors place zero mass on finite models as $p$ increases. To correct this undesirable posterior behavior, Wilson et al. \cite{15} set $a = 1$ and $b = \lambda p$, which yields (a priori) a finite expected number of included covariates as $p$ increases. Alternatively, Castillo et al. \cite{5} obtain promising selection and reconstruction results by requiring a slightly faster than exponential decrease in the prior probability of the number of covariates in the model. This is achieved by taking $b = p^u$ for some $u > 1$. The Beta-Binomial priors in \cite{15} and \cite{5} sacrifice infinite exchangeability in order to produce complexity penalization.

Our strategy is to treat each model as a local null and the set of models that nest it as its set of local alternatives. We will require that the probability of a model and its local alternatives are comparable regardless of model size $k$ or total number of predictors $p$. In Section 2, we detail the model space prior construction that is produced by using a simple proportionality principle. In Section 3 we describe the properties of the prior, showing that the limiting distribution on model dimension is Poisson and demonstrating a self-similarity that motivated naming the prior the matryoshka doll. In Section 4, we discuss further discuss this self-similarity and show the ubiquity of a simple generalization of our construction. In Section 5 we compare the behavior of the matryoshka doll prior to the Beta-Binomial based prior constructions mentioned above. In Section 6, we empirically investigate the posterior performance of the various model space priors through simulations in both finite model spaces and model spaces with an increasing number of covariates. We close the manuscript with a short discussion of ongoing research extending the proposed prior to more complex model spaces.

2 Local null hypotheses and the matryoshka doll prior

The model space $\mathcal{M}_p$ associated with $X$ in (1) can be viewed in different ways. For instance, it has been introduced as the powerset of $\{1, \ldots, p\}$ and is isomorphic to
\{0,1\}^p$, the set of binary $p$-dimensional vectors. This binary space corresponds to the space generated by all possible binary vectors made up of inclusion indicator variables $\gamma_j$ assigned to each $j \in \{1,\ldots,p\}$. In previous constructions of model space priors, the $\gamma_j$ are viewed as random variables and given a prior distribution, inducing a prior on the model space. While this seems reasonable, it requires considerable prior elicitation and specification on the model space to obtain distributions that behave sensibly. This behavior is typically assessed through the prior odds between models or groups of them.

Here, we opt for the opposite approach, first defining a guiding principle (control over the prior odds) and then building the prior following this principle. We accomplish this by viewing each model as a local null hypothesis with respect to the set of models that nest it. This specification provides a straightforward criterion to penalize complexity, which is to define the prior probability of each model as a function of the prior probability of the set of models that nest it.

More formally, let $M_A$ for $A \subseteq \{1,\ldots,p\}$ denote the model with mean structure defined by $X_0\beta_0 + X_A\beta_A$, where $X_A$ is the submatrix of $X$ containing the columns indexed by $a \in A$. The precision parameter in model $M_A$ is denoted by $\tau_A$ and $M_A$ comes equipped with a prior distribution $\pi(\beta_0,\beta_A,\tau_A|M_A)$. The model space $M_p$ consists of models $M_A$ for all possible $A \subseteq \{1,\ldots,p\}$. A model $M_A$ is said to be nested in $M_B$ ($M_A \subset M_B$) whenever $A \subset B$. Though (1) corresponds to Gaussian regression, the model space prior constructed in this paper applies to regression models in general.

Given that the prior odds are the metric typically used to evaluate prior behavior, the simple rule we propose for building the prior is to assume that the relative odds for any model $M_A \in M_p$ (i.e., the local null hypothesis) with respect to the set of models that nest it $\{M_C : C \subseteq \{1,\ldots,p\} \text{ with } A \subset C\}$ (i.e., the corresponding local alternative) is a fixed quantity, which we denote by $\eta$. That is, for any $A \subset \{1,\ldots,p\}$

$$\eta = \frac{P(M_A|M_p, \eta)}{\sum_{C \subseteq \{1,\ldots,p\}} P(M_C|M_p, \eta)},$$

which completely determines the recursion needed to calculate model prior probabilities for all models in $M_p$, given by

$$P(M_A|M_p, \eta) = \eta \sum_{C \subseteq \{1,\ldots,p\}} P(M_C|M_p, \eta), \quad (2)$$

for all models $M_A \in M_p$, with $A \subset \{1,\ldots,p\}$.

Simply put, Equation (2) states that the probability of a model is proportional to the sum of the probabilities of all the models that nest it, with the proportionality constant given by $\eta$. In practice, all that is needed to calculate the prior probabilities over the entire model space is selecting a value for $\eta$, which denotes the prior odds of a model against all models that nest it. To understand how the recursion is carried out, it helps to represent the model space as a directed acyclic graph, where the nodes correspond to models, and the edges between them define their nesting relationships (e.g., see the
left graph in Figure 1(d)). Then, starting from the largest element in the graph (i.e., the full model), it is trivial to calculate the prior probability for each node by assigning to each of the edges of the graph a probability according to expression (2).

To fix ideas, consider the model space made up by the sixteen models obtained from all possible model combinations available with four predictors, with the intercept only model being the smallest model in the space. In this model space, there are five types of models determined by the number of terms they include (i.e., models with 0, 1, 2, 3 or 4 terms). Because of the symmetry of the model space, all models of a particular type have equal prior probability, as such, calculating the probability of one model per model type is all required to determine the probability distribution over the entire space. In Figure 1 we step through the recursive process (powered by our simple rule) used to calculate the prior over the model space, where the probability in each orange-colored node, corresponds to a proportion of the sum of the probabilities in the blue nodes that sit above them.

Here we provide the first few steps of the recursion to guide the reader on how the it proceeds. First, choose the prior odds of a model to the set of models that nest it, \( \eta \), and let \( \psi = \Pr(M_{\{1,2,3,4\}}|M_4, \eta) \) denote the prior probability of the full model. Going down the graph (Figure 1(d)) note that model \( M_{\{1,2,3\}} \) is only nested in \( M_{\{1,2,3,4\}} \), hence our simple proportionality rule implies that \( \Pr(M_{\{1,2,3\}}|M_4, \eta) = \eta \psi \). Now, because \( M_{\{1,2\}} \) is exclusively nested in models \( M_{\{1,2,3\}} \), \( M_{\{1,2,4\}} \), and \( M_{\{1,2,3,4\}} \), its prior is given by \( \Pr(M_{\{1,2\}}|M_4, \eta) = \eta(\eta \psi + \eta \psi + \psi) = \eta \psi (2\eta + 1) \). The remaining steps of the recursion proceed in a similar fashion and are illustrated in Figure 1.

In general, the rule proposed in (2) is quite simple, intuitive and only uses the structure of the model space induced by the nesting of models. The subset relation, \( \subseteq \), induces a partial ordering on \( M_p \) and the prior probability of a model is required to be proportional to the sum of the probabilities of the models that are greater than it in the partial order. Enforcing this proportionality throughout induces a multiplicity correction against local alternatives that leads to a more balanced behavior as we will evidence in later sections of the paper. The rule is also generalizable to any model space with a poset structure. For a given \( \eta \), the model space prior is uniquely defined whenever the poset has a greatest element (i.e. a model that nests all other models in \( M_p \)). When the poset does not have a unique greatest element, the relative probabilities of the set of greatest elements must be also be set to define the model space prior. An example would be a regression model space that places an upper bound (strictly less than \( p \)) on the number of predictors allowed in a model. That being said, here we restrict our discussion to properties of the prior on \( M_p \) induced by (2).

## 3 Theoretical properties of the matryoshka doll prior

In this section, we provide general properties of the prior induced by the rule given in (2), and identify the limiting distribution of the prior.
Figure 1: Recursive calculation for the prior probability of one model of each type in the model space. Above $\xi = 2(\eta \psi) + \psi$, $\kappa = 3(\eta \xi) + 3(\eta \psi) + \psi$, and $\delta = 4(\eta \kappa) + 6(\eta \xi) + 4(\eta \psi) + \psi$. The probability for the full model, $\psi$, is obtained through normalization.
3.1 Finite exchangeability and conditioning isomorphism

Before going into details, we describe the results briefly.

First, for each $p$ and integer $0 \leq k \leq p$, the set of models $M_A$ with $|A| = k$ forms a complexity class and each model in the class has the same prior probability. For $|A| = k \leq p$ we can write $P(M_A|M_p, \eta) = \pi_p(k|\eta)\binom{k}{\ell}^{-1}$, where $\pi_p(k|\eta) = P(|A| = k|M_p, \eta)$ is the prior on model complexity. This is equivalent to finite exchangeability on the equality constraints on the priors over complexity classes for different values of $p$.

Second, if we let $B \subseteq \{1, \ldots, p\}$ and condition on models $M_A$ such that $A \supseteq B$, then the conditional prior $P(M_A|A \supseteq B, M_p, \eta)$ behaves as if we had moved the covariates corresponding to $B$ into $X_0$. That is, the conditional prior is isomorphic to the prior that would have been constructed on a model space with $p - |B|$ test covariates. This isomorphism comes directly from the construction in (2) and provides a number of equality constraints on the priors over complexity classes for different values of $p$.

We now provide the formal statements and proofs. The arguments are essentially about a backwards recursion. If we have $p$ covariates to consider for testing, then we begin with the model space that has no predictors to be tested (all $p$ are assumed to be in the null model). We then add one covariate at a time to the test space. After we have added $k$ covariates to the test space and computed the prior probability on the model space for $k$ covariates, a recursive argument using (2) provides a means of updating model prior probabilities when we add another covariate to the test space. The prior on model complexity is derived through careful book-keeping.

**Proposition 1.** Assume that (2) holds and let $\psi_p(\eta) = P(M_{\{1, \ldots, p\}}|M_p, \eta)$. Then the following hold:

1. If $A \subseteq \{1, \ldots, p\}$ with $|A| = p - \ell$, then $P(M_A|M_p, \eta) = f_{\ell}(\eta)\psi_p(\eta)\binom{p}{\ell}$ where $f_0(\eta) = 1$ and for $\ell > 0$ we have $f_{\ell}(\eta) = \eta \sum_{j=0}^{\ell-1} f_j(\eta)\binom{p}{j}$.

2. Letting $\pi_p(k|\eta) = f_{p-k}(\eta)\psi_p(\eta)\binom{p}{k}$ for $k = 0, \ldots, p$, we have $P(M_B|M_p, \eta) = \pi_p(k|\eta)\binom{p}{k}$ for $B \subseteq \{1, \ldots, p\}$ with $|B| = k$.

3. For $k = 0, \ldots, p - 1$, we have $\pi_p(k|\eta) = \eta \sum_{j=1}^{p-k} \pi_p(k+j|\eta)\binom{k+j}{k}$.

4. $\psi_p(\eta)^{-1} = \sum_{j=0}^{p} f_{p-j}(\eta)\binom{p}{j}$.

The first claim in Proposition 1 is proven using induction and the rest follows directly from it. The strong result from Proposition 1 is that the probability of a model of dimension $p - \ell$ only depends on $p$ through $\psi_p(\eta)$. This result allows us to make some nice equality constraints between the probabilities of sets of models of different sizes for differing numbers of potential covariates. The following corollary follows directly from Proposition 1.
Corollary 2. Let $\pi_p(k|\eta) = f_{p-k}(\eta)\psi_p(\eta)^{p-k}$ for $k = 0, \ldots, p$ and $p \in \mathbb{N}$. Let $k, \ell$, and $j$ be integers satisfying $0 \leq k < p$, $0 < \ell \leq p - k$, and $j > 0$, then

$$\frac{\pi_p(k + \ell|\eta)}{\pi_p(k|\eta)} = \left(\frac{j + k + \ell}{\ell}\right)^{k + \ell} \frac{\pi_{p+j}(j + k + \ell|\eta)}{\pi_{p+j}(j + k|\eta)}. \quad (3)$$

In particular, we have

$$\frac{\pi_p(j + 1|\eta)}{\pi_p(j|\eta)} = \frac{1}{j + 1} \frac{\pi_{p-j}(1|\eta)}{\pi_{p-j}(0|\eta)} \quad (4)$$

for $j = 0, 1, 2, \ldots, p - 1$.

Another consequence of Proposition 1 is the value of $\pi_p(0|\eta)$.

Corollary 3. For all $p \in \mathbb{N}$ we have $\pi_p(0|\eta) = \frac{\eta}{1 + \eta}$.

Proof. From result 3 in Proposition 1, we have

$$1 = \sum_{j=0}^{p} \pi_p(j|\eta) = \pi_p(0|\eta) + \sum_{j=1}^{p} \pi_p(j|\eta) = \pi_p(0|\eta) + \pi_p(0|\eta)/\eta$$

from which the claim follows.

Proposition 1 also suggests the following isomorphism theorem.

Theorem 4. Fix $B \subseteq \{1, \ldots, p\}$ with $|B| = k < p$, then

$$P(M_A|A \supseteq B, \mathcal{M}_p, \eta) = \pi_{p-k}(\ell|\eta) \left(\frac{p-k}{\ell}\right)^{p-k-1}. \quad (5)$$

for $B \subseteq A \subseteq \{1, \ldots, p\}$ with $|A| = k + \ell \leq p$. A relative probability preserving isomorphism between the set of models $\{M_A \in \mathcal{M}_p : A \supseteq B\}$ and the model space given by $\mathcal{M}_{p-k}$ is obtained by mapping such a set $A$ to the set of the ranks of the elements of $A \setminus B$ in $\{1, \ldots, p\} \setminus B$.

Proof. The latter claim follows directly from the former. The former claim can be proven directly using the results of Proposition 1. In particular

$$P(M_A|B \subseteq A, \mathcal{M}_p, \eta) = \frac{P(M_A|\mathcal{M}_p, \eta)}{P(A \supseteq B|\mathcal{M}_p, \eta)} = \frac{f_{p-k-\ell}(\eta)\psi_p(\eta)}{\sum_{j=0}^{p-k} f_{p-k-j}(\eta)\psi_p(\eta)^{p-k-j}} \frac{\pi_{p-k}(\ell|\eta)}{\pi_{p-k}(0|\eta)^{p-k-j}}$$

$$= \frac{f_{p-k-\ell}(\eta)}{\sum_{j=0}^{p-k} f_{p-k-j}(\eta)\psi_p(\eta)^{p-k-j}} = f_{p-k-\ell}(\eta)\psi_{p-k}(\eta) = \pi_{p-k}(\ell|\eta) \left(\frac{p-k}{\ell}\right)^{p-k-1}. \quad \square$$
The matryoshka doll model space prior

The isomorphism theorem motivates naming this prior the matryoshka doll. When one reduces the model space by deciding that some variables should be in the base model, one obtains a smaller version of the same model space. Just like opening a matryoshka doll and finding a smaller version of the same doll inside.

A partial converse to Theorem 4 is true. That is, assuming a relative probability preserving isomorphism plus the assumption that \( \pi_k(0|\eta) = \eta/(1+\eta) \) for all \( k = 1, \ldots, p \) implies expression (2) for a given \( p \). We fully characterize the priors exhibiting the isomorphism condition in Section 4.1.

Notice that Theorem 4 provides an important difference between the matryoshka doll and infinitely exchangeable prior distributions on \( \mathcal{M}_p \). Infinite exchangeability arises from an exchangeability and marginalization condition. In essence, we have to envision a universe with an infinite set of potential covariates to include in our models and make our inference on the proposed set of covariates by marginalizing out all but a finite number of the infinite covariates. In contrast, the matryoshka doll asks us to live in a world where the following two reasoners (with the same finite set of \( p \) covariates) draw the same conclusions: reasoner one knows about the inclusion of \( B \) before making an inference, and reasoner two decides to condition on the set of models containing \( B \) after making an inference in \( \mathcal{M}_p \).

### 3.2 Limiting distribution as \( p \) increases

One final implication of expression (2) is that the limit of \( \pi_p(k|\eta) \) as \( p \) increases is given by \( \theta^k \exp(-\theta)/k! \) where \( \theta = \log(1+1/\eta) \), which provides a limiting Poisson distribution on model size. This is proven using the results of Proposition 1 and Corollary 2.

**Theorem 5.** Suppose that expression (2) holds, then

\[
\lim_{p \to \infty} \pi_p(k|\eta) = \frac{\theta^k \exp(-\theta)}{k!}
\]

where \( \theta = \log(1 + 1/\eta) \).

**Proof.** First, from (4), we have

\[
\frac{\pi_p(k+1|\eta)}{\pi_p(0|\eta)} = \frac{1}{(k+1)!} \prod_{j=0}^{k} \frac{\pi_{p-j}(1|\eta)}{\pi_{p-j}(0|\eta)}.
\]

Letting

\[
\theta = \lim_{\ell \to \infty} \frac{\pi_{\ell}(1|\eta)}{\pi_{\ell}(0|\eta)}
\]

we have

\[
\lim_{p \to \infty} \frac{\pi_p(k+1|\eta)}{\pi_p(0|\eta)} = \frac{\theta^{k+1}}{(k+1)!}.
\]
for each fixed \( k \). We also have

\[
1 + \frac{1}{\eta} = 1 + \sum_{k=0}^{p-1} \frac{\pi_p(k+1|\eta)}{\pi_p(0|\eta)} = 1 + \sum_{k=0}^{p-1} \frac{1}{(k+1)!} \prod_{j=0}^k \frac{\pi_p(j|\eta)}{\pi_p(j|0)}.
\]

Applying the Bounded Convergence Theorem and taking the limit inside of the sum establishes

\[
1 + \frac{1}{\eta} = \sum_{j=0}^{\infty} \frac{\theta^j}{j!} = \exp(\theta),
\]

which provides \( \theta = \log(1 + 1/\eta) \).

\[\square\]

In Section 5, we compare this limiting Poisson distribution to the limiting distributions obtained by the prior constructions recommended by Scott & Berger [13], Wilson et al. [15], and Castillo et al. [5].

4 Generalized Matryoshka Doll

In order to provide a complete picture of the matryoshka doll prior and its limiting Poisson distribution, in this section we present a generalization of (2) and its equivalence to the isomorphism condition.

4.1 Full characterization of the isomorphism condition

For a finite, non-empty subset \( D \) of the natural numbers, let \( \mathcal{P}_D \) be the powerset of \( D \). Define \([p] = \{1, \ldots, p\}\) for all natural numbers \( p \). Suppose that \(|D| = m\) and define the rank map \( \phi_D : \mathcal{P}_D \to \mathcal{P}_{[m]} \) by defining \( \phi_D(\{d\}) = \{\text{rank of } d \text{ in } D\} \) for \( d \in D \), \( \phi_D(\emptyset) = \emptyset \), and \( \phi_D(E) = \bigcup_{e \in E} \phi_D(\{e\}) \) for \( E \in \mathcal{P}_D \). The Isomorphism Condition states that for any natural number \( p \), a fixed \( B \subset [p] \), and every \( A \) such that \( B \subset A \subset [p] \) we have

\[
P(M_A | A \supseteq B, \mathcal{M}_p, I) = P\left(M_{\phi_D^{-1}(A)} | \mathcal{M}_{p-|B|}, I\right)
\]

where \( I \) is any relevant information for prior construction. Note that (8) along with priors on \( \mathcal{M}_k \) for \( k = 1, \ldots, p - 1 \) is not enough to uniquely specify a prior distribution on \( \mathcal{M}_p \) because there is no way from (8) to determine a probability for \( M_\emptyset \) in \( \mathcal{M}_p \).

In order to complete the construction, we need to specify a sequence of positive real numbers \( \eta_k \) for \( k \in \mathbb{N} \) and define

\[
P(M_\emptyset | \mathcal{M}_k, I) = \frac{\eta_k}{1 + \eta_k}.
\]

This sequence would be subsumed into the relevant information \( I \).
A weaker requirement than (2) is to allow the proportionality constant to depend on model size in some way. In particular, we define the Proportionality Condition for a sequence of positive real numbers $\eta_k$ to be

$$P(M|\mathcal{M}, I) = \eta_{p-|M|} \sum_{M' \subseteq [p]} \sum_{M' \subseteq M} P(M'|\mathcal{M}, I)$$

for $M \subseteq [p]$. We refer to the priors induced by (10) as generalized Matryoshka Doll priors and obtain the following theorem.

**Theorem 6.** Let $\eta_k$ be a sequence of positive real numbers. Then (8) and (9) hold if and only if (10) holds. Moreover, in the context of (10), we have $\lim_{k \to \infty} \eta_k = \eta > 0$ if and only if $\lim_{p \to \infty} \pi_{p-0}(j|I) = \theta_j \exp(-\theta_j) j!$ for fixed $j$ where $\theta_j = \log(1 + 1/\eta_k)$.

The most important implication of Theorem 6 is that if we want some version of the Isomorphism condition to hold, and for the sequence $\pi_{p}(0|I)$ to converge to a fixed number in the open interval $(0, 1)$, then the limiting probability distribution on model complexity is Poisson. Of course, we can consider other kinds of sequences for $\pi_{p}(0|I)$, such as those that converge to 0 or 1 or those that do not have a unique limit as $p$ increases. These give rise to different kinds of limits (or no limit) for the distribution on model complexity. We do not discuss these possibilities here.

### 4.2 Representations of the Generalized Matryoshka Doll

In considering Theorem 6 and the implications of (8)-(10), it is illuminating to consider the nature of different representations for the same prior distribution on the model space. These representations are driven by the implication of the isomorphism condition in (8) that

$$\pi_p(k|I) = \pi_p(j+k|I) \frac{\pi_{p-k}(0|I)}{\pi_{p-k}(j|I)} \binom{k+j}{j}$$

for $j = 1, \ldots, p-k$. Setting $b_1 + \ldots + b_{p-k} = 1$ provides

$$\pi_p(k|I) = \sum_{j=0}^{p-k} b_j \pi_p(j+k|I) \frac{\pi_{p-k}(0|I)}{\pi_{p-k}(j|I)} \binom{k+j}{j}.$$  

(12)

As an alternative to (10), one could require the probability of a model, say $M$ in $\mathcal{M}_p$, to be $\mu_{p-k}$ times the sum of the probabilities of the models obtained by adding a single covariate to $M$ (the so-called children models of $M$). Because of the isomorphism condition, this is equivalent to

$$\pi_{p-k}(0|I) = \mu_{p-k} \pi_{p-k}(1|I)$$

(13)

for $k = 0, \ldots, p-1$. We get an equivalence between (13) and (10) by choosing

$$b_j = \frac{ \frac{1}{j} \pi_{p-k-j}(j-1) }{ \sum_{j=1}^{p-k} \frac{1}{j} \pi_{p-k-j}(j-1) \pi_{p-k-j}(0) }$$

and

$$\eta_{p-k} = \frac{ \mu_{p-k} }{ \sum_{j=1}^{p-k} \frac{1}{j} \pi_{p-k-j}(j-1) \pi_{p-k-j}(0) }.$$  

(14)
If one chooses $\mu_k = \mu$ in (13) for all $k$, then a Poisson($\mu^{-1}$) distribution truncated to $\{0, \ldots, p\}$ is immediately obtained for model complexity in $\mathcal{M}_p$ and we get

$$\eta_p = \left[\sum_{j=1}^{p} \frac{\mu_j}{j!}\right]^{-1}$$

for the prior odds of $M_\emptyset$ in $\mathcal{M}_p$.

The representation of the model space prior can be extended using different $b_j$ sequences that sum to one. An illustrative use of this is to make the most generic representation of the isomorphism condition by defining an array $\eta_{i,j}$ for $j = 1, \ldots, i$ and $i = 1, \ldots, p$ and setting

$$\pi_{p-k}(0|I) = \sum_{j=1}^{p-k} \eta_{p-k,j} \pi_{p-k}(j|I).$$

This produces the model space prior induced by (13) with

$$\mu_1^{-1} = \frac{\pi_1(0|I)}{1 - \pi_1(0|I)} \quad \text{and} \quad \mu_k^{-1} = \eta_{k,1} + \sum_{j=2}^{k} \left(\frac{\eta_{k,j}}{j!} \prod_{i=k-j+1}^{k-1} \mu_i^{-1}\right)$$

for $k = 2, \ldots, p$ and the prior induced by (10) with

$$\eta_{p-k} = \left[\sum_{j=1}^{p-k} \prod_{i=1}^{j} \frac{\eta_{p-k-j+i}}{j!}\right]^{-1}.$$  

Any representation of a model space prior satisfying isomorphism condition provides a Generalized Matryoshka Doll prior with a sequence of probabilities for the null model from (9). The limiting Poisson is inescapable so long as the sequence of $\eta_k$s converges to a positive real number.

5 Comparisons to previous constructions

In this section, we compare the properties of the matryoshka doll prior to distributions based on widely used previously used Beta-Binomial constructions [see 13, 15, 5]. The prior probability of the set of models with $k$ covariates is

$$\pi_p^{BB}(k|a,b) = \frac{\Gamma(a+b)\Gamma(p+1)\Gamma(k+a)\Gamma(p-k+b)}{\Gamma(b)\Gamma(a)\Gamma(k+1)\Gamma(p-k+1)\Gamma(p+a+b)},$$

where the superscript $BB$ denotes the Beta-Binomial prior construction. Similarly, we will use the notation $\pi_p^{MD}(k|\eta)$ to denote the matryoshka doll prior from Section 2.

5.1 Limiting behavior of $\pi_p(k)$

Complexity penalization can be investigated through the ratio $\pi_p(k+1)/\pi_p(k)$, which is motivated by Castillo et al. [5]; their detection and reconstruction results rely on

$$c_1 p^{-c_2} \leq \frac{\pi_p(k+1)}{\pi_p(k)} \leq c_3 p^{-c_4},$$

(19)
In general, for a Beta-Binomial prior, we have

\[
\frac{\pi_p^{BB}(k+1|a,b)}{\pi_p^{BB}(k|a,b)} = \binom{k+a}{k} \left(\frac{p-k}{p-k-1+b}\right).
\]  

(20)

We are interested in the different limits for (20) as \( p \) increases for fixed \( k \) when \( b \) is a function of \( p \). For ease of discussion, we will take \( a = 1 \).

**Proposition 7.** Suppose that \( k \) is finite and \( a = 1 \) in (20), then:

1. If \( b \) is constant as a function of \( p \), then (20) tends to 1 as \( p \) increases.

2. If \( b = \lambda p \), then (20) tends to \( 1/(\lambda + 1) \) as \( p \) increases.

3. If \( b = p^u \) for \( u > 1 \), then (20) is approximated by \( p^{1-u} \) as \( p \) increases.

In more detail, Proposition (7) implies that if \( b \) is constant, the prior on model complexity exhibits mass loss; the prior probability of models with dimension less than any fixed \( k \) decreases like \( k/p \) as \( p \) increases. Also, taking \( b = \lambda p \) provides some penalization and a limiting geometric distribution on model complexity. Unfortunately, this penalization is not enough to get good signal detection and mean structure reconstruction results. The prior does not penalize enough to reduce false positives as the size of the true model increases, which motivated Castillo et al. [5] to require (19). Nevertheless, if \( b = p^u \) for \( u > 1 \), the prior degenerates to a point mass on the empty model.

In contrast, the matryoshka doll prior provides

\[
\lim_{p \to \infty} \frac{\pi_p^M(k+1|\eta)}{\pi_p^M(k|\eta)} = \frac{\theta}{k+1},
\]  

(21)

where \( \theta = \log(1 + 1/\eta) \), which is essentially an adaptive version of (19). The limiting additional penalization for adding a covariate does not depend on the total number of covariates under consideration, rather it depends on local model complexity. When \( k \) increases to be a power of \( p \), then (19) holds for the matryoshka doll prior, implying a penalty adaptation that meets the requirements specified in Castillo et al. [5] without degenerating to a single point mass as \( p \) increases. In this sense, the matryoshka doll prior is a “just right” prior on model complexity and perhaps even Goldilocks would approve of its use.

### 5.2 Comparison to children models

In the variable selection problem the hardest models for the true model to beat are those with all of the true positives and just one false positive. These are the children models of the true model in the poset of models. Thus, studying the prior odds between a model \( M \) and a model \( M' \supset M \) with \(|M| + 1\) predictors provides insights about the effect of the model space prior on posterior inference.
Proposition 8. Let $M$ be a model and define the children set of models to be $C(M) = \{M' \supset M \text{ and } |M'| = |M| + 1\}$. For $M' \in C(M)$, the prior odds under any finitely exchangeable prior is
\[
\text{odds}(M', M) = \frac{P(M'|M_p)}{P(M|M_p)} = \binom{p}{k} \pi_p(k + 1) \pi_p(k) = \frac{k + 1}{p - k} \pi_p(k) \pi_p(k) \tag{22}
\]
and the prior odds versus the set of children models is
\[
\text{odds}(C(M), M) = \sum_{M' \in C(M)} \text{odds}(M', M) = (k + 1) \times \frac{\pi_p(k + 1)}{\pi_p(k)}. \tag{23}
\]

Table 1 displays the values for (22) and (23) attained by the different prior constructions and limiting situations for model complexity. We investigate finite model size, model size growing as $\log(p)$, model size growing as $p^q$ for $0 < q < 1$, and model size growing as a proportion of $p$. Though this list is not exhaustive, it provides enough variation to observe the influence of the prior on posterior inference. There are a couple of important things to note in this table.

First we discuss finite model size, $|M| = k < \infty$. The beta-binomial with $b = p^u$ with $u > 1$ provides odds($C(M), M$) that converge to 0 at a $k/p$ rate as $p$ increases for any $k$. The beta-binomial with constant $b$ or $b = \lambda p$ provides odds($C(M), M$) that converge to a linear function of $k$ as $p$ increases. The slope for constant $b$ is one and the slope for $b = \lambda p$ is less than one. The matryoshka doll provides constant odds($C(M), M$) regardless of $k$ as $p$ increases.

Second we discuss increasing model size. The beta-binomial odds follow the same patterns as before for each regime for $b$. Essentially, the $k$ for finite models is replaced with the appropriate growth rate for $k$ as a function of $p$. This can cause issues with false positives for the beta-binomial with constant $b$ or $b = \lambda p$, and issues with false negatives for $b = p^u$ for $u > 1$ when $|M|$ increases with $p$. Once again, the matryoshka doll provides constant odds versus the set of children models regardless of the growth rate of $k$ with $p$. Interestingly, the beta-binomial with $b = p^u$ requires $u \geq 2$ to produce the same (or more) penalization as the matryoshka doll when $k \propto p$. The choice of $u \geq 2$ provides very strong penalization for finite $k$ and could lead to a dramatic increase false negatives for finite $k$.

Figure 2 provides heatmaps of the odds of a single child model and the set of children models versus a model as a function of model complexity and growth rate with $p$. Some additional insights can be gleaned from the figure that are harder to understand just using Table 1. First, the beta-binomial with constant $b$ or $b = p$ displays its inability to properly penalize the set of children models even when $k$ is finite. Second, the beta-binomial with $b = p$ fails to appropriately penalize the set of children models when $k = \xi \log(p)$ and $\xi$ is not close to 0. Third, the beta-binomial with $b = p^2$ over-penalizes the set of children models in all scenarios except when $k = \xi p$ with $\xi$ values away from 0 or 1. As expected from its construction, the matryoshka doll provides appropriate penalization of the set of children models regardless of the size of the model or its growth rate with $p$. 
The matryoshka doll model space prior

| Prior odds of a single child model versus a model | k finite | \( \lim \frac{k}{p} = \xi \) | \( \lim \frac{k}{p^q} = \xi \) | \( \lim \frac{k}{p} = \xi \) |
|-----------------------------------------------|---------|-----------------|---------------|-----------------|
| \( MD(\theta) \) | \( \frac{\theta}{p-k} \) | \( \frac{\theta}{p} \) | \( \frac{\theta}{p} \) | \( \frac{\theta}{p^{1-q}} \) |
| \( BB(a, b) \) | \( \frac{k+a}{p-k-1+b} \) | \( \frac{k+a}{p} \) | \( \frac{k+a}{p} \) | \( \frac{\xi}{p^q} \) |
| \( BB(a, \lambda) \) | \( \frac{k+a}{p-k-1+\lambda p} \) | \( \frac{k+a}{p} \) | \( \frac{k+a}{p} \) | \( \frac{\xi}{p^q} \) |
| \( BB(a, p^u) \) | \( \frac{k+a}{p-k-1+p^n} \) | \( \frac{k+a}{p} \) | \( \frac{k+a}{p} \) | \( \frac{\xi}{p^q} \) |

Table 1: Prior odds of children models versus a model with \( k \) predictors under different limiting regimes for \( k \) as \( p \) increases. \( \xi > 0 \) is a constant (with \( \xi < 1 \) if \( k \propto p \)), \( u > 1 \), \( 0 < q < 1 \), and \( \theta = \log \left( 1 + \frac{q}{q} \right) \).

5.3 Comparison to nesting models

The matryoshka doll construction provides constant prior odds for a model versus the set of nesting models. In this subsection, we compare this to the prior odds obtained through the various Beta-Binomial prior constructions. Once again, assume a model \( M \) of dimension \( k \). The prior odds of the models that nest \( M \) versus \( M \) is given by

\[
\sum_{M' \supseteq M} \frac{P(M'|M_p)}{P(M|M_p)} = \sum_{j=1}^{p-k} \frac{\pi_p(k+j)}{\pi_p(k)} \binom{k+j}{j}, \tag{24}
\]

Of course, because of (2), the matryoshka doll prior provides a value of \( 1/\eta \) for (24) regardless of \( p \) and \( k \). When \( \pi_p(k) \) comes from the Beta-Binomial \((a, b)\) construction, then (24) is given by

\[
\sum_{M' \supseteq M} \frac{P^{BB}(M'|M_p, a, b)}{P^{BB}(M|M_p, a, b)} = \sum_{j=1}^{p-k} \frac{(p-k)! \Gamma(k+j+a) \Gamma(p-k-j+b)}{j! (p-k-j)! \Gamma(k+a) \Gamma(p-k+b)}. \tag{25}
\]

When \( a \) and \( b \) are constants, the sum in (25) tends to \( \infty \) as \( p \) increases for any fixed \( k \). When \( b = \lambda p \), the limit of (25) for finite \( k \) is given by

\[
\lim_{p \to \infty} \frac{\sum_{M' \supseteq M} P^{BB}(M'|M_p, a, b = \lambda p)}{P^{BB}(M|M_p, a, b = \lambda p)} = \sum_{j=1}^{\infty} \frac{\Gamma(k+j+a)}{\Gamma(k+a)} \left( \frac{1}{1+\lambda} \right)^j. \tag{26}
\]
The series from (26) can be written in terms of probabilities from a negative-binomial distribution. In particular, if \( Q \) follows a negative-binomial distribution counting success until \( k+a \) failures with individual trial probability of success \( 1/(1+\lambda) \), then (26) becomes

\[
\lim_{p \to \infty} \sum_{M' \supseteq M} P^{BB}(M'|M, a, b = \lambda p) = \left(1 + \frac{1}{\lambda}\right)^{k+a} \times P(Q > 0|k, a, \lambda). \tag{27}
\]

For each fixed \( k \), the prior proposed by Wilson et al. \[15\] does have a finite limit. However, that limit increases exponentially as a function of \( k \). This produces poor control of false positives if the dimension of the true model increases, as discussed in Castillo et al. \[5\].

Letting \( b = p^u \) for \( u > 1 \), (25) tends to 0 as \( p \) increases for any fixed \( k \), showing the overly strong penalization of this prior. Further, if \( u \geq 2 \), then the limit of (25) is 0 even if \( k \) increases with \( p \) (except when \( k \propto p \) and \( u = 2 \)). This overly strong penalization, combined with Bayes’ Factor learning rates could lead to undesired false negatives.

Once again, the matryoshka doll strikes a balance between under- and over-penalizing the set of nesting models. Because (24) is a constant for all \( k \) and \( p \), the prior helps...
The matryoshka doll model space prior

The Bayes’ factors overcome the combinatorial complexity of the model space without over-penalizing. In Section 6 and in the online supplement, we provide empirical evidence that constant $b$ or $b \propto p$ leads to an inflated number of false positives and lack of concentration of the posterior. Similarly, we provide empirical evidence that the degeneration of the $BB(a, p^n)$ prior to a point mass at the base model controls far too tightly false positives, leading to increased false negatives and unwarranted posterior concentration on models that exclude covariates that explain meaningful variation in the response variable.

6 Performance using synthetic data sets

To substantiate the claims made in previous sections, we now assess empirically the performance of the matryoshka doll prior and compare it to that of the Beta-Binomial constructions we have discussed so far. In the simulations we consider the matryoshka doll prior ($MD(\theta = 1)$, with $\theta = \log(1 + 1/\eta)$), and for the competing prior specifications we assumed the $BB(1, p)$ and the $BB(1, p^2)$. We generate synthetic datasets under multiple scenarios assuming a finite true model, in a model space that can be fully enumerated. In these experiments we vary the sample size, the number of predictors in the full model, and the rate at which the signal in the predictors decays to better understand how the different priors process signals of varying strength. We also performed a larger set of simulations with model spaces where the number of predictors $p$ grows with $n$; however, the description and results of this additional set of experiments are deferred to the online supplement.

In all of our simulation scenarios the true model $M_T$ is assumed to have $p_T = 5$ predictors. We let $n \in \{41, 81, 121, 161, 201\}$ for scenarios where the sample size varies, while setting the number of predictors in the full model to $p = 20$. In scenarios where we keep the sample size fixed, we set $n = 101$ and let the number of predictors take values in $p \in \{20, 40, \ldots, 100\}$. To control how the signal is assigned to predictors in the true model $M_T$, we assume the $j$th largest regression coefficient in the true model, $\beta_j$, satisfies $\beta_j^2 \propto \zeta^j$ with $\zeta \in \{1/2, 3/4\}$, such that when $\zeta = 1/2$ the signal is highly concentrated on the first few predictors, whereas with $\zeta = 3/4$, while still decreasing, the signal is more spread out across the true model’s coefficients. We choose the value for $\sigma^2$ such that the series of coefficients produces the mean structure with signal to noise ratio $\sum \beta_j^2 / \sigma^2 = 2$.

With each combination of $n$, $p$ and $\zeta$ we simulated 1000 datasets to evaluate the performance of the different model-space prior specifications in terms of: (i) their signal reconstruction ability, and (ii) how posterior probability mass concentrates under each prior distribution. Finally, we propose and implement a fast Pruned Greedy Binary Tree search algorithm and assess how the search algorithm fares against complete enumeration, this considering that in a large proportion of real problems the model space is too large for enumeration.
6.1 Signal Reconstruction

Under each of the model space priors considered we obtained the distribution of the mean true positive rate (tpr) and for the false discovery rate (fdr). These metrics aim to describe how well each of the priors lead the posterior distribution to concentrate on sets of good models (i.e., those including as many true predictors and as little false predictors as possible), which is essential to produce suitable model averaging results. We also obtained the tpr and fdr of the highest probability model (HPM or mode hereforth), which inform about the correspondence between the modal and the true models whenever selection is of interest.

Let $M \setminus M_T$ denote the set of terms in $M$ but not in the true model $M_T$, and $M \cap M_T$ denotes the set of terms in both $M$ and $M_T$. To approximate the distributions for the mean tpr and the mean fdr, for each of the $\ell = 1, 2, \ldots, 1000$ data sets generated under a particular simulation scenario we calculate

$$
\bar{tpr}_\ell = \sum_{M \in M} tpr(M) \cdot p(M|y_\ell), \text{ with } tpr(M) = \frac{|M \cap M_T|}{|M_T|},
$$

$$
\bar{fdr}_\ell = \sum_{M \in M} fdr(M) \cdot p(M|y_\ell), \text{ where } fdr(M) = \frac{|M \setminus M_T|}{|M|},
$$

where $y_\ell$ represents the response vector from the $\ell$th data set. Using the 1000 values obtained for these two metrics under a particular scenario, we approximate their corresponding distribution. The distributions for the tpr and fdr for the mode are obtained by identifying the modal model $M^*_\ell = \text{argmax}_{M \in M} p(M|y_\ell)$ and calculating $tpr(M^*_\ell)$ and $fdr(M^*_\ell)$ for $\ell = 1, 2, \ldots, 1000$.

Growing $n$ and fixed $p$

The results produced under each simulation scenario for these four metrics are well-aligned with our discussion in Section 5 regarding the behaviors that characterize each of these model priors. Figure 3 displays the results as $n$ ranges from 41 to 201 with $p = 20$. In general, the prevailing pattern is that the BB($1, p^2$) prior is more conservative, providing lower (worse) tpr values in exchange for lower (better) fdr values, with the BB($1, p$) penalizing weakly and therefore producing the completely opposite behavior, and with the MD($\theta = 1$) producing metrics that sit between those of the other two priors. This pattern suggests that the BB($1, p^2$) prior consistently places more probability mass on smaller models including terms with large signals, with the weaker signal terms often being excluded.

The trends described above are exacerbated when $n = 41$ under both signal decay regimes. With the BB($1, p^2$) prior the distributions for both the mean tpr and the HPM’s tpr, concentrate closer to zero than the other priors, indicating that this prior places excessive confidence on smaller models which exclude many of the terms in the true model. Conversely, the other two priors suitably enable propagating the uncertainty that
The matryoshka doll model space prior originates from the limited amount of data to the posterior distribution. When \( n = 201 \) and \( \zeta = 1/2 \), the dominating mode for both \( \text{fdr} \) distributions under the \( \text{BB}(1, p^2) \) is around 0.8, whereas the predominant mode under the other two priors is closer to 1. With the same sample size but with slower signal decay (i.e., \( \zeta = 3/4 \)) the weaker signals are substantially stronger than in the \( \zeta = 1/2 \) scenario, thus making it easier to detect all of the true predictors no matter the prior.

![Figure 3: Signal reconstruction metrics for simulation experiments over finite model spaces with growing \( n \) and fixed \( p = 20 \).](image)

The distributions for the mean \( \text{fdr} \) follows the expected behavior. Under the \( \text{BB}(1, p^2) \) prior the mean \( \text{fdr} \) distribution is tightly concentrated around zero regardless of the sample size, while for the other two priors the distribution gradually concentrates more and more about zero as the sample size increases. As discussed in Section 5, due to it’s limited penalization we evidence a slightly inflated number of false positives with the \( \text{BB}(1, p) \) prior; this weakness presents itself more clearly when \( k \) grows with \( p \) in the simulations provided in the online supplement. In terms of the mean \( \text{fdr} \), the \( \text{MD}(\theta = 1) \) prior again provides a sensible compromise between the \( \text{BB}(1, p^u) \) prior, which penalizes excessively, and the \( \text{BB}(1, p) \) that does not penalize enough. Lastly, the distribution for the HPM’s \( \text{fdr} \) is concentrated around zero no matter the sample size, signal decay scenario, or the prior considered. However, under the \( \text{BB}(1, p) \) we can evidence a very small mode around 0.2 in all instances, being somewhat more pronounced in the \( \zeta = 3/4 \) scenario.
Fixed $n$ and growing $p$

Figure 4 summarizes the results with $n = 101$ and having $p$ take values in \{20, 40, 60, 80, 100\}. In broad terms, the behavior of both the mean tpr and that for the HPM are as expected.

As $p$ grows the distribution gradually shifts towards lower values, with the BB$(1, p^2)$ prior concentrating around lower tpr values, followed by the MD$(\theta = 1)$, with the BB$(1, p)$ taking the highest values. Nevertheless, the distributions under the MD$(\theta = 1)$ and the BB$(1, p)$ are remarkably close. It is worth highlighting that the distributions for the tpr metrics when $\zeta = 3/4$ under the BB$(1, p^2)$ prior become more and more diluted as $p$ grows, providing evidence for how this prior’s unduly harsh penalization leads to the exclusion of multiple true signals. Conversely, the other two priors are able to remain concentrated around tpr values close to one.

In terms of the false positive terms, the mean fdr is tightly concentrated around zero for the BB$(1, p^2)$, with the distribution from the MD$(\theta = 1)$ prior being relatively close having most of its mass close to 0.05, and with the BB$(1, p)$ performing the worst, taking mostly fdr values between 0.1 and 0.2. Regarding the HPM’s fdr distribution, these all concentrate around 0 for all priors, but more tightly when $\zeta = 3/4$, indicating that the modal model mostly includes true terms irrespective of the model space prior used in the settings we considered.
6.2 Posterior Concentration

To evaluate the posterior concentration we assess four metrics: posterior probability of the modal model (i.e., the HPM), posterior probability of the true model, the number of models needed to accumulate 95% of the total probability mass of the models in the space, and the rank for the true model. These are shown in Figure 5, with panel (a) displaying the results for fixed $p$ and growing $n$, and panel (b) showing the results for growing $p$ and fixed $n$. Taken together with the results described for the $tpr$ and $fdr$, these metrics provide a holistic view of how the priors behave. The two cases for $\zeta$ reflect different expectations in posterior behavior.

First, we discuss the case when $\zeta = 1/2$. Again, we find evidence that BB$(1, p^2)$ prior assigns an unwarranted amount of probability mass to the HPM in all scenarios, even when $p = 20$ and $n = 41$ the distribution of the probability assigned to the HPM peaks around 0.9. The excessively high modal posterior model probability together with the low probabilities assigned to the true model under most scenarios result from this prior’s severe aversion to false positives. At the other end of the spectrum is the BB$(1, p)$ model, which spreads out posterior probabilities across many more models due to its insufficient control of false positives. While this also results in a higher average true positive rate, it also implies that many more models are needed to accumulate 95% of the model space posterior mass. This is particularly evident in most scenarios with growing $p$ (see Figure 5 panel b), where the number of models needed to accumulate 95% of the distribution is orders of magnitude greater for the BB$(1, p^2)$ than for the other two priors. As expected, the matryoshka doll can be found between these two extremes. It finds the most important predictors and concentrates posterior probability mass on only hundreds of models. This much needed posterior model space uncertainty is strongly contrasted with the misplaced certainty obtained by the BB$(1, p^2)$, which concentrates on less than 10 models 95% of the probability mass across all scenarios, assigning very large probabilities to the HPM, but often excluding the true model from the set of these top models (see scenarios with varying $p$).

Second, when $\zeta = 3/4$ (i.e., the signal is better distributed across predictors) we see that when $n$ is small or $p$ is large, the BB$(1, p^2)$ provides an unjustified amount of concentration on a small number of models, and as $p$ grows, the lower the probability assigned to the true model. In contrast, the BB$(1, p)$ spreads the probability mass across far too many models, thus assigning relatively low posterior probabilities, even to the modal model. As expected, the matryoshka doll provides a data driven compromise between these two extremes. For example, when $n = 101$ and $p = 80$, the matryoshka doll concentrates 95% of its probability mass on hundreds of models and assigns probabilities to the true model, such that the dominant mode of its distribution is about 0.8. These results for the matryoshka doll mediate a suitable compromise between the handful of models (in the low tens) with very high posterior probabilities that the BB$(1, p^2)$ allocates probability mass to, and the thousands of models with relatively low probabilities found with the BB$(1, p)$.

In the online supplement we provide additional results for these simulations, introduce the algorithm we use for both model space enumeration as well exploration when
Figure 5: Posterior concentration metrics for simulation experiments over finite model spaces as either the sample size \((n)\) or the total number of predictors in the full model \((p)\) grow.
larger spaces are considered, test its performance, and provide additional simulations to evaluate priori performance as both \( p \) and the size of the true model grow.

7 Discussion

In this paper we propose a new construction for model space priors with a specific focus on regression models. The key motivation behind this work was designing a prior over the space of models with both a meaningful interpretation and a construction from first principles instilled with desirable properties. The matryoshka doll prior meets these conditions, striking a well-balanced penalization profile. The matryoshka doll prior relies on a reformulation of the way in which the model space is viewed. Instead of considering indicator variables for covariate inclusion, the model space is viewed as a series of local null and alternative hypotheses. This induces the isomorphism theorems and the limiting Poisson prior on model complexity presented in Section 3. Furthermore, we provide a generalization of the matryoshka doll class of priors, which arises from any representation of a model space prior satisfying the isomorphism condition with \( P(M_0|M_k, I) = \frac{\eta_k}{1+\eta_k} \), which also converges to a Poisson distribution so long as \( \lim_{k \to \infty} \eta_k \) converges to a positive real number.

We argue that the limiting Poisson prior identified provides a natural complexity penalization that strikes an attractive balance that eludes beta binomial constructions, with the BB\((a, p^n)\) excessively controlling false positives thus having a tendency to omit some true signals, and conversely with the BB\((a, b)\) or BB\((a, \lambda p)\) displaying weak control over the inclusion of false positive terms. The adaptation of the penalty to the number of covariates in the matryoshka doll provides both low penalization for adding the first few covariates to a model and strong control of false positives when adding covariates that explain a small fraction of the variation in the response. In our opinion, this balanced behavior is “just right” for penalizing model complexity in regression models.

As a future direction, we are investigating the expansion of the model space prior construction (using children or nesting models) to more structured posets. In particular, we are interested in essentially non-parametric Bayesian models such as polynomial surface regression or multi-resolution analysis. The model spaces for these models can be forced to adhere to either weak or strong heredity rules [14]. For a fixed maximal degree (or subspace depth), the model space is thus a subset of \( M_p \), but the challenge lies in accommodating heredity requirements in asymptotic and self-similarity arguments analogous to those made in Section 4 and enumerating model complexity classes.

Funding

Womack was partially supported by NIH award 5R01DC018813-04.

Taylor-Rodríguez was partially supported by NSF RTG DMS award 2136228, NIH award 5R01DC018813-04, and by Portland State University’s FDG and Sabbatical programs.
References

[1] Altomare, D., Consonni, G., and La Rocca, L. (2013). “Objective Bayesian Search of Gaussian Directed Acyclic Graphical Models for Ordered Variables with Non-Local Priors.” *Biometrics*, 69(2): 478–487.

[2] Berger, J. O. and Pericchi, L. R. (1996). “The intrinsic Bayes factor for model selection and prediction.” *Journal of the American Statistical Association*, 91(433): 109–122.

[3] Casella, G., Girón, F. J., Martínez, M. L., and Moreno, E. (2009). “Consistency of Bayesian procedures for variable selection.” *Annals of Statistics*, 1207–1228.

[4] Casella, G. and Moreno, E. (2012). “Objective Bayesian variable selection.” *Journal of the American Statistical Association*.

[5] Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). “Bayesian linear regression with sparse priors.” *Annals of Statistics*, 43(5): 1986–2018. URL http://dx.doi.org/10.1214/15-AOS1334

[6] Girón, F. J., Moreno, E., Casella, G., and Martínez, M. (2010). “Consistency of objective Bayes factors for nonnested linear models and increasing model dimension.” *RACSAM-Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matematicas*, 104(1): 57–67.

[7] Hans, C. (2010). “Model uncertainty and variable selection in Bayesian lasso regression.” *Statistics and Computing*, 20(2): 221–229.

[8] Johnson, V. E. and Rossell, D. (2010). “On the use of non-local prior densities in Bayesian hypothesis tests.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2): 143–170.

[9] — (2012). “Bayesian model selection in high-dimensional settings.” *Journal of the American Statistical Association*, 107(498): 649–660.

[10] Ley, E. and Steel, M. F. (2009). “On the effect of prior assumptions in Bayesian model averaging with applications to growth regression.” *Journal of Applied Econometrics*, 24(4): 651–674.

[11] Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). “Mixtures of g priors for Bayesian variable selection.” *Journal of the American Statistical Association*, 103(481).

[12] Moreno, E., Girón, F. J., and Casella, G. (2010). “Consistency of objective Bayes factors as the model dimension grows.” *Annals of Statistics*, 1937–1952.

[13] Scott, J. G. and Berger, J. O. (2010). “Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem.” *Annals of Statistics*, 38(5): 2587–2619.

[14] Taylor-Rodriguez, D., Womack, A., and Bliznyuk, N. (2016). “Bayesian Variable selection on model spaces constrained by heredity conditions.” *Journal of Computational and Graphical Statistics*, 25(2): 515–535.
The matryoshka doll model space prior

[15] Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C., and Schildkraut, J. M. (2010). “Bayesian model search and multilevel inference for SNP association studies.” *Annals of Applied Statistics*, 4(3): 1342. 2, 9, 11, 15

[16] Womack, A. J., León-Novelo, L., and Casella, G. (2014). “Inference From Intrinsic Bayes’ Procedures Under Model Selection and Uncertainty.” *Journal of the American Statistical Association*, 109(507): 1040–1053. 2

[17] Zellner, A. and Siow, A. (1980). “Posterior odds ratios for selected regression hypotheses.” *Trabajos de Estadística y de Investigación Operativa*, 31(1): 585–603. 2
### Appendix A: Summary tables Section 6 simulations

| Fixed | Varying | Prior       | Prob Mode | Prob $M_T$ | Rank $M_T$ | mods. needed |
|-------|---------|-------------|-----------|------------|------------|--------------|
|       | $n = 41$| BB(1,$p$)   | 0.26      | 0          | 55         | 943.5        |
|       |         | MD($\theta = 1$) | 0.38      | 0          | 89         | 106          |
|       |         | BB(1,$p^2$) | 0.66      | 0          | 311.5      | 4            |
|       | $n = 81$| BB(1,$p$)   | 0.29      | 0.03       | 5          | 379.5        |
|       |         | MD($\theta = 1$) | 0.45      | 0.01       | 7          | 44           |
|       |         | BB(1,$p^2$) | 0.68      | 0          | 21         | 5            |
| $\zeta = 1/2$, $p = 20$ | $n = 121$| BB(1,$p$)   | 0.34      | 0.11       | 2          | 209.5        |
|       |         | MD($\theta = 1$) | 0.52      | 0.08       | 3          | 28           |
|       |         | BB(1,$p^2$) | 0.71      | 0.01       | 5          | 4            |
|       | $n = 161$| BB(1,$p$)   | 0.40      | 0.25       | 1          | 133          |
|       |         | MD($\theta = 1$) | 0.61      | 0.26       | 2          | 21           |
|       |         | BB(1,$p^2$) | 0.74      | 0.07       | 2          | 3            |
|       | $n = 201$| BB(1,$p$)   | 0.45      | 0.35       | 1          | 95           |
|       |         | MD($\theta = 1$) | 0.66      | 0.44       | 1          | 17           |
|       |         | BB(1,$p^2$) | 0.80      | 0.22       | 2          | 3            |

| $p = 20$ | BB(1,$p$) | 0.3      | 0.1      | 3          | 253        |
|          | MD($\theta = 1$) | 0.5      | 0        | 4          | 33          |
|          | BB(1,$p^2$) | 0.7      | 0        | 7          | 4           |
| $p = 40$ | BB(1,$p$) | 0.3      | 0        | 5          | 1623.5     |
|          | MD($\theta = 1$) | 0.5      | 0        | 6          | 78          |
|          | BB(1,$p^2$) | 0.7      | 0        | 29         | 3           |

| $\zeta = 1/2$, $n = 101$ | $p = 60$ | BB(1,$p$) | 0.3 | 0 | 8 | 3481 |
|                          |          | MD($\theta = 1$) | 0.5 | 0 | 12 | 131 |
|                          |          | BB(1,$p^2$) | 0.8 | 0 | 74 | 2 |
| $p = 80$ | BB(1,$p$) | 0.3 | 0 | 10.5 | 5560.5 |
|          | MD($\theta = 1$) | 0.5 | 0 | 15 | 174.5 |
|          | BB(1,$p^2$) | 0.8 | 0 | 140 | 2 |
| $p = 100$ | BB(1,$p$) | 0.3 | 0 | 13 | 8453 |
|           | MD($\theta = 1$) | 0.5 | 0 | 19 | 227 |
|           | BB(1,$p^2$) | 0.8 | 0 | 201 | 2 |

Table 2: Medians of posterior concentration measures over 1000 simulated datasets from simulations in Section 6, having signal decay rate $\zeta = 1/2$ as either $n$ or $p$ grow. $\eta = 1/(e - 1)$ so that $\theta = 1$ in matryoshka doll.
Table 3: Medians of posterior concentration measures over 1000 simulated datasets from simulations in Section 6, having signal decay rate $\zeta = 3/4$ as either $n$ or $p$ grow. $\eta = 1/(e - 1)$ so that $\theta = 1$ in matryoshka doll.
Appendix B: Model space exploration algorithm

In practice, when considering variable selection search algorithms are often preferred to exhaustive enumeration of the model space due to the computationally prohibitive cost the latter approach. Here we put forth a fast and effective strategy to search the space of models by building a binary tree with each model as a possible node.

Algorithm 1 performs model enumeration using a bifurcation and recursion algorithm. This algorithm provides a computationally efficient way to perform model space enumeration. To do so, it creates a total ordering out of the model space by starting at the least element, the base model (nested in all other models in the space), \( M_0 \), and taking the greediest path to the greatest element, the full model, \( M_F \). It then goes back to the last place in the path, say \( M \), where there are children models available that have not yet been chosen and sets the next model to be the best child of \( M \) that has not yet been visited. This is repeated until the model space is exhausted.

To set some notation, let \( A \subseteq \{1, \ldots, p\} \), and define \( Q(A) \) to be the unnormalized posterior probability of \( M_A \), which is given by \( Q(A) = f(y|X, M_A) \pi(M_A) \). This can be computed for any \( A \) whenever needed. Let \( \text{Info}(A) \) represent all of the information about model \( M_A \) that can be computed without enumerating the model space and that we want to write to file or memory. Furthermore, let \( U \) denote a set of terms that are not in \( M_A \), which is initialized at \( U = A^c \) whenever \( A \) is updated in the algorithm.

**Algorithm 1: Greedy Binary Tree via Bifurcation and Recursion**

```plaintext
Define: B&R (A ⊆ \{1, \ldots, p\}, U ⊆ A^c)
Set: u^* = \arg\max_{u \in U} Q(A \cup \{u\})
Write: \text{Info}(A \cup \{u^*\})
if |U| > 1 then
    Call: B&R (A \cup \{u^*\}, U \setminus \{u^*\})
    Call: B&R (A, U \setminus \{u^*\})
end if
end B&R

Run:
Write: \text{Info}(\emptyset)
Call: B&R (\emptyset, \{1, \ldots, p\})
end Run
```

In Algorithm 2, we perform the same bifurcation and recursion algorithm with the addition of a pruning step, limiting wasteful exploration of regions of the model space where little probability mass can be found. To implement this last step, we assume \( \epsilon \in (0, 1) \), and the algorithm tests if the probability for a proposed model relative to its path is greater than \( \epsilon \) to continue with the recursion, if it is not then the branch is pruned at that last model.

In addition to the definitions provided above, in Algorithm 2 the pruning step depends on the quantity \( q \), which measures the probability that has been accumulated
from $M_0$ along the branch leading up to the model being considered. Note that $q$ is initialized at $Q(\emptyset)$. As before, the algorithm proposes at each step the most advantageous available step from a model $M_i$, say $M_i^\star$. If there is enough evidence for $M_i^\star$, then it is added to the path as $M_{i+1}$ and we continue the construction. Starting from $M_0$ and going along the branch, once the evidence in favor of some $M_i^\star$ is too small relative to the branch, the move is rejected and the descendants of $M_i$ are pruned from the model space and removed from consideration. Once such pruning occurs, we move back down the branch to the most recently visited model with a children set that has not been completed visited and pruned. We then propose a best move from that model and evaluate its merit relative to its path from $M_0$.

Algorithm 2: Pruned Greedy Binary Tree via Bifurcation and Recursion

Define: PB&R ($A \subseteq \{1, \ldots, p\}$, $U \subseteq A^c$, $q > 0$)
Set: $u^\star = \arg\max_{u \in U} Q(A \cup \{u\})$
Write: Info($A \cup \{u^\star\}$)
if $|U| > 1$
\hspace{1em}if $\frac{Q(A \cup \{u^\star\}]}{(q + Q(A \cup \{u^\star\}))} > \epsilon$
\hspace{2em}Call: PB&R ($A \cup \{u^\star\}, U \setminus \{u^\star\}, q + Q(A \cup \{u^\star\})$)
\hspace{1em}end if
\hspace{1em}Call: PB&R ($A, U \setminus \{u^\star\}, q$)
end if
end PB&R
Run:
\hspace{1em}Write: Info($\emptyset$)
\hspace{1em}Call: PB&R ($\emptyset, \{1, \ldots, p\}, Q(\emptyset)$)
end Run

Appendix C: Algorithm performance evaluation

To validate the proposed algorithm, we make use of the same set of simulations presented in Section 6 of the main body of paper and the compare results from enumeration (generated using Algorithm 1) to those obtained using Algorithm 2, and derive three metrics. First, the total variation norm ($\text{tv norm}$) between the probabilities of models found by the search algorithm and the probabilities for those same models obtained from enumeration. The second metric considered is the total probability mass discovered by the search ($\text{prob. found search}$), which was obtained by adding together the true model posterior probabilities (i.e., those from enumeration) of models found by the search algorithm. Finally, we extracted the maximum posterior probability among models not discovered by the algorithm ($\text{max prob. missed}$).

The two panels in Figure 6 summarize the results obtained for simulations with growing $n$ and growing $p$, respectively. We find that the model space prior considered influences the effectiveness of the algorithm, which is to be expected given that the
Figure 6: Algorithm efficacy metrics for simulation experiments over finite model spaces as either the same size \( n \) or the total number of predictors in the full model \( p \) grow.
posterior concentration strongly depends on the model space prior. The general patterns we find are:

1. In terms of the \(\text{tv norm}\), the BB(1, \(p^2\)) is always closest to zero, meaning that the renormalized probabilities obtained with the models found by the algorithm closely match their true values, and that the bulk of the probability mass is found. The MD(\(\theta = 1\)) follows closely, also being close to zero over most settings while being shifted slightly to the right with respect to the BB(1, \(p^2\)). Finally, under the BB(1, \(p\)) the \(\text{tv norm}\) lie slightly further away from zero. This pattern based on the prior makes sense since it is impacted by how each prior distributes the probability mass over models, with the BB(1, \(p^2\)) concentrating the mass over only a few models while the BB(1, \(p\)) comparatively dilutes it over the space. Differences in the behavior of the \(\text{tv norm}\) across distributions are most evident in the scenarios with growing \(p\).

2. For \text{prob. found search} we find that the BB(1, \(p^2\)) consistently finds most of the probability mass (i.e., it is always close to one), followed slightly lower values from the MD(\(\theta = 1\)), which is then followed by lower values (or drastically lower values depending of the scenario) by the BB(1, \(p\)). This result follows the same intuition as with the \(\text{tv norm}\), that is, this result is a consequence of the different posterior concentration regimes provided by the different priors.

3. Over all scenarios considered there is barely any difference across priors in the \text{max prob. missed}. This metric is invariably close to zero, meaning that the algorithm is able to find the vast majority or all high probability models regardless of the prior.

Appendix D: Prior behavior on growing model spaces

In this section we explore the behavior of the matryoshka doll prior through simulations and provide empirical support for the claims made in previous sections. We compare the matryoshka doll prior (MD(\(\theta = 1\)), to the BB(1, \(b\)) prior with \(b \in \{1, p, p^2\}\). We evaluate the posterior behavior in terms of signal reconstruction and posterior concentration. Both the number of possible predictors as well as the number of predictors in the true signal increase with sample size. The predictors are generated as independent draws from a standard normal distribution and the response is normal with the appropriate mean structure and variance \(\sigma^2 = 1\). We evaluate: (i) posterior behavior, both with finite samples and asymptotically as the number of predictors grows with \(n\); and (ii) the influence of having the signal in the predictors decay rapidly (i.e., a large proportion of signal in the true model concentrates in a few of the predictors) or slowly (i.e., signal in true model is more evenly spread out across all predictors).

The sample sizes considered are \(n = 200, 2000\). The size of the full model \(M_F\) is \(p = 2n\) and the size of the true model \(M_T\) is \(|M_T| \approx \sqrt{n}\). We assume that \(M_T \subset M_F\). To control how the signal is distributed across all predictors in \(M_T\), we assume its \(j\)th largest regression coefficient \(\beta_j\) satisfies \(\beta_j^2 \propto \zeta^j\) with \(\zeta \in \{1/2, 9/10\}\). We assume that
the asymptotic mean structure is generated from the full geometric series with signal to noise ratio $\sum \beta_j^2 / \sigma^2 = 2$. The mean structure for $M_T$ for finite $n$ is generated by the series truncated to the appropriate number of terms. With each combination of $n$ and $\zeta$ we simulated 100 datasets.

Setting the decay rate $\zeta = 1/2$ makes the magnitude of the coefficients decrease rapidly and concentrate 99% of the true signal in seven regression coefficients. In contrast, setting $\zeta = 9/10$ promotes a slow decay, spreading out the signal more evenly across all predictors and requiring forty four predictors to account for 99% of the total signal.

The model space exploration algorithm described in Section 6 was driven by model posterior probabilities obtained with Zellner-Siow priors for the model parameters and the BB$(1,1)$ prior for models. The branch pruning cutoff was set to $\varepsilon = 0.01$. These choices enabled the algorithm to visit a large portion of the model space because the BB$(1,1)$ prior provides the weakest complexity penalization among all the priors considered. The posterior probabilities resulting for the other model space priors were calculated using the Bayes’ factors obtained from the search and renormalization.

We evaluate the model space priors in terms of false positive inclusion, true positive inclusion, model complexity, and posterior concentration. The specific metrics used are described in the corresponding section of the analysis of the simulation study.

D.1 False and true positive rates

The false and true positive rate metrics are both discrete and continuous.

1. False positive rates:

\[
\text{FPR}_{\text{discr}} = \sum_{M \in M} \text{fpr}_{\text{discr}}(M) \cdot p(M | y),
\]

\[
\text{FPR}_{\text{cont}} = \sum_{M \in M} \text{fpr}_{\text{cont}}(M) \cdot p(M | y), \quad \text{with}
\]

\[
\text{fpr}_{\text{discr}}(M) = \frac{|M \setminus M_T|}{|M_T|}, \quad \text{and} \quad \text{fpr}_{\text{cont}}(M) = 1 - \frac{R^2_{M \cap M_T}}{R^2_M},
\]

where $M_T$ represents the true model, $M \setminus M_T$ corresponds to the set of terms in $M$ but not in $M_T$, $M \cap M_T$ denotes the set of terms in both $M$ and $M_T$, and finally, $R^2_A$ represents the unadjusted $R$-squared for $M_A$.

2. True positive rates:

\[
\text{TPR}_{\text{discr}} = \sum_{M \in M} \text{tpr}_{\text{discr}}(M) \cdot p(M | y),
\]
The matryoshka doll model space prior

\[
\text{TPR}_{\text{cont}} = \sum_{M \in \mathcal{M}} \text{tpr}_{\text{cont}}(M) \cdot p(M|y), \quad \text{with}
\]

\[
\text{tpr}_{\text{discr}}(M) = \frac{|M \cap M_T|}{|M_T|} \quad \text{and} \quad \text{tpr}_{\text{cont}}(M) = \frac{R^2_{M \cap M_T}}{R^2_{M_T}}.
\]

The discrete and a continuous versions for the FPR and TPR convey different information about signal reconstruction. The discrete versions inform about the number of (incorrect or correct) predictors selected, while the continuous version provides insights about the amount of signal explained (by the false or true predictors). A small continuous FPR need not correspond to a small discrete FPR. Similarly, a large continuous TPR need not correspond to a large discrete TPR.

Overall, the results observed in terms of FPR and TPR follow the expected patterns (see Figures 7 and 8), implying a tradeoff between the FPR and TPR. The resulting metrics are neatly ordered, with those for the BB(1, p²) sitting at one extreme, followed by the matryoshka doll, then the BB(1, p) and the BB(1, 1) at the opposite end of the spectrum.

Figure 7: Expected false discovery rates. Sample sizes \( n = 200, 400, \ldots, 2000 \) and decay rates of true coefficient signal strength \( \zeta = 1/2 \) and \( \zeta = 9/10 \).

In terms of the FPR (Figure 7), both for the continuous and discrete versions, the BB(1, p²) is consistently near zero regardless of the sample size or the signal decay rate, which is expected as this prior yields the strongest penalization. The matryoshka doll prior takes only slightly higher values than the BB(1, p²) in both FPR_{discr} and FPR_{cont}. The FPR_{discr} values for the matryoshka doll prior are significantly smaller than those for the BB(1, p) and BB(1, 1) priors. Finally, regardless of the prior, the values observed for
When $\zeta = 9/10$ and the signal is more evenly spread out across predictors, the choice of model space prior impacts the inference, with the strength of its influence mediated by the sample size. With a sample size of $n = 200$, the BB$(1, p^2)$ prior severely penalizes the addition of covariates into the model and the BB$(1, 1)$ and BB$(1, p)$ allow in the most true predictors. As expected, the matryoshka doll produces a TPR that sits between these two extremes. With $n = 2000$, the same ordering remains but the differences across all priors become considerably less pronounced, especially in terms of the signal (see panel (b) in Figure 8). In terms of the number of true predictors, the simulation density of discrete TPR for the matryoshka doll overlaps with those of both the BB$(1, p^2)$ and the BB$(1, p)$; however, the latter distributions display little overlap.
among themselves (Figure 8 panel (a)). In terms of the continuous TPR, the same ordering appears. Notably, the BB(1, \(p^2\)) prior’s extreme control of false positives is driving down the continuous TPR and meaningful variation in the data from covariates in the true model is being missed.

### D.2 Posterior concentration

Finally, to evaluate the posterior concentration we assess three metrics: posterior expected model complexity, posterior probability of the modal model, and the number of models needed to accumulate 95% of the total probability mass of the models found by the search algorithm. These are shown in Figure 9 and medians are presented in Table 4. Together with true and false positive rates, these metrics help to provide a more complete understanding of the behaviors of the priors. The two cases for \(\zeta\) reflect different expectations in posterior behavior.

| Rate | Prior | \(n = 200\) | \(n = 2000\) |
|------|-------|------------|-------------|
|      | Comp  | Conc       | Max         | Comp  | Conc       | Max         |
| \(\zeta = \frac{1}{2}\) | BB(1, 1) | 5.4  | 3364 | 0.179 | 8.09 | 9466 | 0.292 |
|      | BB(1, \(p\)) | 4.7  | 1698 | 0.332 | 7.64 | 5562 | 0.47 |
|      | MD(\(\eta\)) | 4.26 | 422  | 0.565 | 7.15 | 651  | 0.746 |
|      | BB(1, \(p^2\)) | 3.02 | 2    | 0.887 | 6.16 | 2    | 0.93 |
| \(\zeta = \frac{9}{10}\) | BB(1, 1) | 12.13 | 30830 | 0.016 | 34.4 | 390398 | 0.006 |
|      | BB(1, \(p\)) | 9.54 | 29168 | 0.053 | 32.66 | 389586 | 0.017 |
|      | MD(\(\eta\)) | 5.49 | 3529 | 0.188 | 27.91 | 14738 | 0.183 |
|      | BB(1, \(p^2\)) | 0.16 | 2 | 0.871 | 23.28 | 10 | 0.414 |

Table 4: Medians of measures over 100 simulated datasets. \(\eta = 1/(e - 1)\) so that \(\theta = 1\) in matryoshka doll. Comp is posterior expected model complexity. Conc is the number of models needed to comprise 95% of the model space posterior (rounded to whole numbers). Max is the posterior probability of the modal model.

First, we discuss the case when \(\zeta = 1/2\) and the true model is well approximated by seven predictors. The BB(1, \(p^2\)) prior provides an unwarranted amount of certainty (or lack of uncertainty) in regards to what the “true” model is and what constitutes the set of good models. The excessively high modal posterior model probability and the posterior concentration on a few models reflects this prior’s unnecessarily severe penalization of false positives. On the other end of the spectrum is the BB(1, 1) and BB(1, \(p\)) model, which have a higher false positive rate and thus concentrate less in the posterior. This causes a higher average true positive rate, but also implies that thousands of models are needed to accumulate 95% of the model space posterior mass.

As expected, the matryoshka doll sits between these two extremes. It finds the most important predictors and concentrates on hundreds and not thousands of models. This posterior model space uncertainty is strongly contrasted with the certainty obtained by the BB(1, \(p^2\)), which concentrates on one or two models.
Figure 9: Posterior average model complexity, modal posterior model probability, and number of models needed to accumulate 95% of the total probability mass. Sample sizes $n = 200$ and $n = 2000$ and decay rates of true coefficient signal strength $\zeta = 1/2$ and $\zeta = 9/10$. 

(a) Model Complexity

(b) Maximum probability

(c) Models needed for 95% probability mass
Second, we discuss the case when $\zeta = 9/10$ and the model needs a large number of predictors to capture the mean structure. Especially when $n$ is small, the BB$(1, p^2)$ provides an unjustified amount of concentration on a small number of models. In contrast, the BB$(1, 1)$ and BB$(1, p)$ spread the probability mass across too many models and assign low posterior probability to their modal models. As expected, the matryoshka doll provides a data driven compromise between these two extremes. For example, when $n = 2000$, its concentration on tens of thousands of models, as opposed to tens of models for BB$(1, p^2)$ and hundreds of thousands for BB$(1, 1)$ and BB$(1, p)$, is reasonable given the complexity of its modal model and the number of covariates left that explain meaningful variation in the response variable. The average posterior complexity for the matryoshka doll prior is about 28 and we need 44 covariates to explain 99% of the variation in the response. This leads to tens of thousands of models that are nested in the true model and that nest the modal model. It is completely reasonable to have the posterior express this amount of uncertainty in what the true model is when $\zeta = 9/10$. 