Fast and Optimal Bayesian Approximations for Targeted Prediction

Daniel R. Kowal*

Abstract

Prediction is critical for decision-making under uncertainty and lends validity to statistical inference. With targeted prediction, the goal is to optimize predictions for specific decision tasks of interest, which we represent via functionals. Using tools for predictive decision analysis, we design a framework for constructing optimal, scalable, and simple approximations for targeted prediction under a Bayesian model. For a wide variety of approximations and (penalized) loss functions, we derive a convenient representation of the optimal targeted approximation that yields efficient and interpretable solutions. Customized out-of-sample predictive metrics are developed to evaluate and compare among targeted predictors. Through careful use of the posterior predictive distribution, we introduce a procedure that identifies a set of near-optimal predictors. These acceptable models can include different model forms or subsets of covariates and provide unique insights into the features and level of complexity needed for accurate targeted prediction. Simulations demonstrate excellent prediction, estimation, and variable selection capabilities. Targeted approximations are constructed for physical activity data from the National Health and Nutrition Examination Survey (NHANES) to better predict and understand the characteristics of intraday physical activity.

Keywords: decision theory; functional data; physical activity; variable selection

*Assistant Professor, Department of Statistics, Rice University (daniel.kowal@rice.edu).
1 Introduction

Prediction is a cornerstone of statistical analysis: it is essential for decision-making under uncertainty and provides validation for inference (Geisser, 1993). Predictive evaluations are crucial for model comparisons and selections (Gelfand et al., 1992) and offer diagnostic capabilities for detecting model misspecification (Gelman et al., 1996). More subtly, predictions provide an access point for model interpretability: namely, via identification of the model characteristics or variables which matter most for accuracy. However, the demands of many datasets—which can be high-dimensional, high-resolution, and multi-faceted—often necessitate sophisticated and complex models. Even when such models predict well, they can be cumbersome to deploy and difficult to summarize or interpret.

Our focus is targeted prediction, where predictions are customized for the decision tasks of interest. In practice, the translation of models into actionable decisions requires predictive quantities in the form of functionals of future or unobserved data. Predictions and predictive approximations should be optimized for these decision tasks—and targeted to the relevant functionals. Fundamentally, the target is essential for defining the correct (predictive) likelihood (Bjornstad, 1990; Bjørnstad, 1996). Absent specific functionals of interest, targeted prediction nonetheless offers a path for interpretable statistical learning (Murdoch et al., 2019): the functionals probe the data-generating process to uncover the predictability of distinct attributes.

To illustrate these points, we display wearable device data from the National Health and Nutrition Examination Survey (NHANES) in Figure 1. Physical activity (PA) trajectories are modeled as functional data and accompanied by subject-specific covariates; detailed descriptions of the data and the model are in Section 5. Scientific interest does not reside exclusively with these intraday profiles; we are also interested in functionals of these trajectories. Figure 1 denotes several such functionals: the average activity (avg), the peak...
activity level \( \text{max} \), and the time of peak activity \( \text{argmax} \). These functionals summarize daily PA and describe clear sources of variability in PA among the individuals. Other features are discernible as well, such as sedentary behavior and periods of absolute inactivity, and are investigated in Section 5. Our goal is construct simple, fast, and accurate predictive approximations targeted to these functionals. In doing so, we seek to improve predictive performance, streamline the decision-making process, and identify the model attributes and variables that matter most for accuracy—and determine how they differ among functionals.

![Figure 1: Intraday physical activity (gray line) and fitted values (blue line) for multiple subjects under model \( M \) in (17)-(19). The lines denote the empirical (solid gray) and predictive expected value (dashed blue) of \( \text{avg} \) (lower horizontal), \( \text{max} \) (upper horizontal), and \( \text{argmax} \) (vertical).](image)

To achieve these goals, we introduce new tools for predictive decision analysis. For any Bayesian model \( M \), we compute optimal predictive approximations targeted to each functional of interest. The optimal predictor minimizes a posterior predictive expected loss: the
predictive expectation inherits from \( \mathcal{M} \), while the loss is carefully tailored to each functional. For a variety of approximations and (penalized) loss functions, we derive a convenient representation of the optimal predictors that yields efficient and interpretable solutions. In particular, these solutions can be computed using existing software packages for penalized regression, which allows for widespread and immediate deployment of the proposed techniques. While intrinsically useful for prediction, the elicitation of these approximations is independently informative for understanding and summarizing the model \( \mathcal{M} \) posterior.

A key feature of our approach is the use of the model \( \mathcal{M} \) posterior predictive distribution. In particular, we leverage this predictive distribution to provide uncertainty quantification for out-of-sample predictive evaluations of the targeted approximations. Using these predictive tools, we introduce a procedure to identify not only the most accurate predictor, but also any predictor that performs nearly as well with some nonnegligible predictive probability. This strategy emerges as a Bayesian representation of the Rashomon effect, which observes that there often exists a multitude of acceptably accurate predictors (Breiman, 2001). The set of acceptable models is informative: it describes the shared characteristics and level of complexity needed for near-optimal targeted prediction. Importantly, we do not require any re-fitting of \( \mathcal{M} \)—neither for distinct functionals nor for out-of-sample evaluations—and instead design an efficient algorithm to approximate the relevant out-of-sample predictive quantities for each functional. The proposed methods are applied to both simulated and real data and demonstrate excellent prediction, estimation, and model selection capabilities.

There is a rich literature on the use of decision analysis to extract information from a Bayesian model. Bernardo and Smith (2009) provide foundational elements, while Vehtari and Ojanen (2012) give a prediction-centric survey. MacEachern (2001) and Gutiérrez-Peña and Walker (2006) use decision analysis to summarize Bayesian nonparametric models. Decision analysis is widely used for Bayesian variable selection, including for linear regression (Lindley, 1968; Bondell and Reich, 2012; Hahn and Carvalho, 2015), seeming unrelated.
regressions (Puelz et al., 2017), graphical models (Bashir et al., 2019), nonlinear regressions (Woody et al., 2019), functional regression (Kowal and Bourgeois, 2020), and time-varying parameter models (Huber et al., 2020). Predictive quantities are also useful for Bayesian model selection (Laud and Ibrahim, 1995; Meyer and Laud, 2002). Alternative approaches combine selection with Kullback-Leibler (KL) distributional approximations (Goutis and Robert, 1998; Nott and Leng, 2010; Tran et al., 2012; Piironen et al., 2018; Crawford et al., 2019). In general, these methods do not provide targeted prediction or summarization and nearly all rely on in-sample metrics to select a single model.

The remainder of the paper is outlined as follows. Section 2 introduces predictive decision analysis for optimal targeted prediction. Section 3 develops the methods and algorithms for predictive model evaluation, comparison, and selection. A simulation study is in Section 4. The PA data are analyzed in Section 5. Section 6 concludes. Online supplementary material includes computational details, supporting figures, proofs, and R code to reproduce the results from the simulated and real data analyses.

## 2 Targeted point prediction

Consider the paired data \( \{x_i, y_i\}_{i=1}^n \) with \( p \)-dimensional covariates \( x_i \) and \( m \)-dimensional response \( y_i \). The response variables \( y_i \) may be univariate (\( m = 1 \)), multivariate (\( m > 1 \)), or functional data with \( y_i = (y_i(\tau_1), \ldots, y_i(\tau_m))' \) observed on a domain \( T \subset \mathbb{R}^d \). Suppose we have a satisfactory Bayesian model \( \mathcal{M} \) parametrized by \( \theta \) with posterior \( p_{\mathcal{M}}(\theta|y) \). The requisite notion of “satisfactory” is made clear below, but fundamentally \( \mathcal{M} \) should (i) encapsulate the modeler’s beliefs about the data-generating process and (ii) demonstrate empirically the ability to capture the essential features of the data. While these criteria are standard for Bayesian modeling, they often demand highly complex and computationally intensive models. There is broad interest in developing simple yet accurate approximations
to $\mathcal{M}$, which further assist in producing fast predictions and interpretable model summaries.

We propose that approximations should target predictive variables $h(\tilde{y})$, where each $h$ is a pre-specified functional of interest and $\tilde{y} \sim p_M(\tilde{y}|y)$ is the predictive distribution of unobserved or future data conditional on observed data. The specification of $h$ reflects the prediction task: often the data $y$ are an input to a system or process $h$, which inherits predictive uncertainty when the data have not yet been observed. We focus on constructing optimal and scalable point prediction approximations to $h(\tilde{y})$. These targeted approximations are not burdened by the complexity required to capture the global distributional features of $p_M(\theta|y)$ or $p_M(\tilde{y}|y)$—which may be mostly irrelevant for predicting $h(\tilde{y})$. The approximations can be customized for each $h$ separately or constructed for multiple $h$ jointly, but in either case use the full posterior distribution under $\mathcal{M}$ to incorporate all the available data.

A core attribute of the proposed approach is that only a single Bayesian model $\mathcal{M}$ is required. The model $\mathcal{M}$ is used to construct, evaluate, and compare all approximating models, and is the vessel for all subsequent uncertainty quantification. We operate in an $\mathcal{M}$-complete framework (Bernardo and Smith, 2009): we neither assume that the true model is among those under consideration ($\mathcal{M}$-closed) nor that all models under consideration are inadequate for prediction ($\mathcal{M}$-open). Rather, we only assume that $\mathcal{M}$ provides a sufficiently accurate predictive distribution for each target $h(\tilde{y})$. This requirement is empirically verifiable through standard posterior predictive diagnostics (Gelman et al., 1996), which assess the fit of a model $\mathcal{M}$ for predicting certain features of the data. While it is practically impossible for $\mathcal{M}$ to be adequate for every functional, many well-designed models are capable of describing multiple functionals. If needed, additional modeling flexibility is available through model aggregation such as averaging (Hoeting et al., 1999) or stacking (Yao et al., 2018). These and other techniques are compatible with the proposed framework.
2.1 Obtaining optimal point prediction parameters

Our first objective is to obtain optimal point predictions of $h(\tilde{y})$. These predictions will serve as the launching point for developing predictive evaluation metrics (Section 3.1) and identifying sufficiently accurate models (Section 3.2). We allow a general form for the predictor $g(\tilde{x}; \delta)$, which is a point prediction of $h(\tilde{y})$ at covariate value $\tilde{x}$ with unknown parameters $\delta$. Important examples include linear, additive, and tree models, but $g$ is not required to match the structure of $M$. While $h(\tilde{y})$ is a predictive quantity under model $M$, $g(\tilde{x}; \delta)$ is a deterministic function of its inputs, such as $g(\tilde{x}; \delta) = \tilde{x}'\delta$. For any predictor $g$, the goal is to identify values of the parameters $\delta$ that produce accurate point predictions of $h(\tilde{y})$.

Predictive accuracy is measured by a loss function $L_0\{h(\tilde{y}), g(\tilde{x}; \delta)\}$, which determines the loss from predicting $g(\tilde{x}; \delta)$ when $h(\tilde{y})$ is realized. Since $L_0$ depends on a random quantity $\tilde{y}$, Bayesian decision analysis proceeds by minimizing the posterior expected loss: $\hat{\delta}_0 := \underset{\delta}{\text{arg min}} E[\tilde{y}|y] L_0\{h(\tilde{y}), g(\tilde{x}; \delta)\}$. This operation averages the predictive loss over the distribution of future or unobserved values $h(\tilde{y})$ at $\tilde{x}$ under model $M$, and then selects parameters $\hat{\delta}_0$ that minimize this quantity.

To generate a path of approximations with varying complexity, we introduce a penalty $P$ on the unknown $\delta$:

$$L_\lambda\{h(\tilde{y}), g(\tilde{x}; \delta)\} := L_0\{h(\tilde{y}), g(\tilde{x}; \delta)\} + \lambda P(\delta) \quad (1)$$

where $\lambda \geq 0$ determines the tradeoff between predictive accuracy ($L_0$) and complexity ($P$).

The optimal point prediction parameters are

$$\hat{\delta}_A := \underset{\delta}{\text{arg min}} E[\tilde{y}|y] L_\lambda\{h(\tilde{y}), g(\tilde{x}; \delta)\}. \quad (2)$$

The subscript $A = (g, P, \lambda)$ denotes the essential components of the approximating model,
Figure 2: Given data \( \{x_i, y_i\}_{i=1}^n \), a Bayesian model \( M \) is constructed. For each functional \( h(\tilde{y}) \) and using model \( M \), multiple approximations \( A \) are optimized, evaluated, and compared. The optimal parameters \( \hat{\delta}_A \) are used to compute point predictions \( g(\tilde{x}; \hat{\delta}_A) \) of \( h(\tilde{y}) \) at \( \tilde{x} \).

which we define explicitly:

**Definition 1.** The approximation \( A := (g, P, \lambda) \) is a triple consisting of the predictor \( g \), the complexity penalty \( P \), and the complexity parameter \( \lambda \).

The challenge is to produce optimal point prediction parameters \( \hat{\delta}_A \) for distinct approximating models \( A \), and subsequently to evaluate and compare the resulting point predictions. A schematic is presented in Figure 2 given data \( \{x_i, y_i\}_{i=1}^n \), a Bayesian model \( M \) is constructed; for each functional \( h \), one or more approximations \( A \) are optimized for prediction; point predictions \( g(\tilde{x}; \hat{\delta}_A) \) are computed for \( h(\tilde{y}) \) at \( \tilde{x} \). The optimal parameters \( \hat{\delta}_A \) offer a summary of the posterior (predictive) distribution of model \( M \)—akin to posterior expectations, standard deviations, and credible intervals—but specifically targeted to \( h \).

From Figure 2, it is clear that \( \hat{\delta}_A \) depend on the approximation \( A \), the functional \( h \), the model \( M \), and the data \( \{x_i, y_i\}_{i=1}^n \). These optimal parameters \( \hat{\delta}_A \) also depend on the loss function \( L_0 \). Although generality of \( L_0 \) is desirable, tractability is essential for practical use.
A natural starting point is squared error loss

\[ \mathcal{L}_0\{h(\tilde{y}), g(\tilde{x}; \delta)\} = \|h(\tilde{y}) - g(\tilde{x}; \delta)\|_2^2 \]  \hspace{1cm} (3)

with generalizations considered below.

We identify a representation of the requisite optimization problem (2) that admits fast and interpretable solutions for a broad class of approximations:

**Theorem 1.** For design point \( \tilde{x} \) with \( \mathbb{E}[\tilde{y}|y] \|h(\tilde{y})\|_2^2 < \infty \), the optimal point prediction parameters in (2) under squared error loss are

\[ \hat{\delta}_A = \arg \min_\delta \left\{ \|\tilde{h} - g(\tilde{x}; \delta)\|_2^2 + \lambda P(\delta) \right\} \]  \hspace{1cm} (4)

where \( \tilde{h} := \mathbb{E}_{(\tilde{y}|y)} h(\tilde{y}) = \mathbb{E}_{(\theta|y)} \hat{h}(\theta) \) is the posterior expectation of \( \hat{h}(\theta) := \mathbb{E}_{\tilde{y}|\theta} h(\tilde{y}) \).

Theorem 1 establishes an equivalence between the solution to the posterior predictive expected loss (2) and a penalized least squares criterion. For any choice of functional \( h \) and approximation \( \mathcal{A} \), the predictive optimal \( \hat{\delta}_A \) only requires estimation of \( \tilde{h} \) and minimization of (4). The response variable \( \tilde{h} \) is the Bayes estimator of \( h(\tilde{y}) \) for model \( \mathcal{M} \) under squared error loss. Estimation of \( \tilde{h} \) is a standard Bayesian exercise, for example using posterior predictive samples: \( \tilde{h} \approx S^{-1} \sum_{s=1}^S h(\tilde{y}^s) \) for \( \tilde{y}^s \sim p_\mathcal{M}(\tilde{y}|y) \). Most commonly, posterior predictive samples are generated by iteratively drawing \( \theta^s \sim p_\mathcal{M}(\theta|y) \) from the posterior and \( \tilde{y}^s \sim p_\mathcal{M}(\tilde{y}|\theta^s) \) from the sampling distribution.

**Remark.** To shift the emphasis from prediction toward model summarization, we may replace the predictive functional \( h(\tilde{y}) \) with a posterior functional \( h(\theta) \), such as \( h(\theta) = h(\mathbb{E}_{\tilde{y}|\theta} \tilde{y}) \), with minimal changes to Theorem 1 and subsequent results.
2.2 Optimal point predictions over multiple design points

The solution \( \hat{\delta}_A \) in (4) depends on the design point \( \tilde{x} \). Naturally, we are often interested in optimizing parameters across \textit{multiple} covariate values, say \( \{ \tilde{x}_i \}_{i=1}^{n} \). These covariates do not need to correspond to the original covariates \( \{ x_i \}_{i=1}^{n} \) nor the variables used in \( \mathcal{M} \): the choice of \( \{ \tilde{x}_i \}_{i=1}^{n} \) can customize predictions for specific designs or subpopulations of interest—all while leveraging the full posterior distribution under model \( \mathcal{M} \). The aggregate loss function

\[
\bar{L}_\lambda[\{h(\tilde{y}_i), g(\tilde{x}_i; \delta)\}_{i=1}^{n}] := \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_\lambda\{h(\tilde{y}_i), g(\tilde{x}_i; \delta)\}
\]

suitably generalizes (1) to include the specified covariate values \( \{ \tilde{x}_i \}_{i=1}^{n} \) with corresponding predictive functionals \( \{h(\tilde{y}_i)\}_{i=1}^{n} \). Under the aggregate loss function, the decision parameters are optimized simultaneously over the joint posterior predictive distribution \( p_M(\tilde{y}_1, \ldots, \tilde{y}_n|y) \):

\[
\hat{\delta}_{A,\tilde{X}} := \arg \min_{\delta} \mathbb{E}_{\tilde{y}_1, \ldots, \tilde{y}_n|y} \left[ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_\lambda\{h(\tilde{y}_i), g(\tilde{x}_i; \delta)\} \right].
\]

For squared error loss, Theorem 1 is generalized accordingly:

**Corollary 1.** For design points \( \tilde{X} = \{ \tilde{x}_i \}_{i=1}^{n} \) with \( \mathbb{E}_{\tilde{y}_i|y}\|h(\tilde{y}_i)\|_2^2 < \infty \) for \( i = 1, \ldots, n \), the optimal point prediction parameters in (6) under the aggregate squared error loss are

\[
\hat{\delta}_{A,\tilde{X}} = \arg \min_{\delta} \left\{ \tilde{n}^{-1} \sum_{i=1}^{\tilde{n}} \|\tilde{h}_i - g(\tilde{x}_i; \delta)\|_2^2 + \lambda \mathcal{P}(\delta) \right\}
\]

where \( \tilde{h}_i := \mathbb{E}_{\tilde{y}_i|y} h(\tilde{y}_i) = \mathbb{E}_{\theta|y} \hat{h}_i(\theta) \) is the posterior expectation of \( \hat{h}_i(\theta) := \mathbb{E}_{\tilde{y}_i|y} h(\tilde{y}_i) \).

For computing \( \hat{\delta}_{A,\tilde{X}} \), Corollary 1 only requires an estimate of the pointwise predictive expectation \( \tilde{h}_i \) at each \( \tilde{x}_i \in \tilde{X} \) and a solution to a penalized least squares objective. For many choices of \( \mathcal{A} \), the optimal point prediction parameters \( \hat{\delta}_{A,\tilde{X}} \) can be computed easily and efficiently using existing algorithms and software. Point predictions of \( h(\tilde{y}) \) at any \( \tilde{x} \)
are given by \(g(\tilde{x}; \hat{\delta}_{A_L})\), which is often quick to compute.

An interpretable special case of Corollary 1 is the following:

**Corollary 2.** Let \(A_L\) denote the unpenalized linear approximation with \(g(\tilde{x}; \delta) = \tilde{x}' \delta\) and \(\lambda = 0\). When \(\hat{h}_i(\theta) = \tilde{x}_i \theta\) and using the observed design points \(X := \{x_i\}_{i=1}^n\), the optimal point predictor parameters are given by the posterior expectation \(\hat{\delta}_{A_L,X} = E_{\theta|y}\theta\).

Corollary 2 is most familiar when \(M\) is a linear model and \(h\) is the identity functional. In this simplified setting, allowing \(\lambda > 0\) with a sparsity penalty for \(P\) recovers the decoupling shrinkage and selection approach for Bayesian variable selection (Hahn and Carvalho, 2015). However, the full benefits of Corollary 1 are realized by the generality of the model \(M\), the functionals \(h\), and the approximations \(A\).

While the flexibility in specifying \(h\) is undoubtedly a primary feature of this framework, certain choices of \(h\), such as binary functionals \(h(\tilde{y}) \in \{0, 1\}\), are incompatible with squared error loss. Modifications of \(L_0\) must be designed with care to maintain the core attributes of the proposed approach: computational speed, ease of implementation, and interpretability. We achieve these goals by replacing the squared error loss with the negative log-likelihood of an exponential family distribution, or the *deviance*:

\[
L_{0,EF}^* \{h(\tilde{y}), g(\tilde{x}; \delta)\} := F_0\{g(\tilde{x}; \delta)\} - T_0\{h(\tilde{y})\} - \sum_{j=1}^p F_j\{g(\tilde{x}; \delta)\} T_j\{h(\tilde{y})\}, \tag{8}
\]

where \(\{F_j\}_{j=1}^p\) are the natural parameters and \(\{T_j\}_{j=1}^p\) are the sufficient statistics, all of which have known forms for a given distribution in the exponential family.

Optimal point predictions are obtained again by minimizing the posterior expected loss, \(\hat{\delta}_A = \arg\min_\delta E_{\tilde{y}|y} L_{0,EF}^* \{h(\tilde{y}), g(\tilde{x}; \delta)\}\). As in Theorem 1, key simplifications are available:

**Theorem 2.** When \(E_{\tilde{y}|y}|T_0\{h(\tilde{y})\}| < \infty\), the optimal point prediction parameters under deviance loss (8) are \(\hat{\delta}_A = \arg\min_\delta \left[ F_0\{g(\tilde{x}; \delta)\} - \sum_{j=1}^p F_j\{g(\tilde{x}; \delta)\} T_j \right]\), where \(T_j := E_{\tilde{y}|y}|T_j\{h(\tilde{y})\}\) is the predictive expectation of the sufficient statistics \(j = 1, \ldots, p\) under \(M\).
Theorem 2 maintains the simplifications of Theorem 1 but under a more general family of loss functions. An optimal \( \hat{\delta}_A \) requires only estimation of \( T_j \), such as \( T_j \approx S^{-1} \sum_{s=1}^S T_j \{ h(\tilde{y}^s) \} \) for \( \tilde{y}^s \sim p_M(\tilde{y}|y) \), and minimization of the resulting deviance. Crucially, the requisite optimization problem retains the form of the exponential family log-likelihood, and therefore is efficiently solvable using existing software for many choices of \( g \). Extensions for multiple covariates \( \{ \tilde{x}_i \}_{i=1}^n \) and penalized loss functions \( \mathcal{L}^{EF}_{\lambda} := \mathcal{L}^{EF}_0 + \lambda \mathcal{P} \) proceed as in Corollary 1.

Despite the presence of the exponential family of distributions, we employ the loss function (8) only for point prediction of \( h(\tilde{y}) \). The loss function is chosen to reflect the nature of \( h(\tilde{y}) \), which may be binary, count-valued, nonnegative, or restricted to an interval—each of which features a distributional analog in the exponential family. However, the loss function (8) does not impose additional assumptions on the distribution of \( h(\tilde{y}) \): the predictive distribution is inherited from \( \mathcal{M} \), while Theorem 2 produces optimal point prediction parameters for an approximation \( A \) based on this loss function.

This approach is distinct from distributional approximations of \( \mathcal{M} \) based on KL divergence (Goutis and Robert, 1998; Nott and Leng, 2010; Tran et al., 2012; Piironen et al., 2018). These methods approximate \( p_M(y|\theta) \) with a distribution \( p_M(y|\hat{\delta}_{KL}) \) such that \( \hat{\delta}_{KL} = \arg\min_\delta D_{KL} \{ p_M(y|\theta), p_M(y|\delta) \} \), usually for variable selection. Those approximations are derived for the likelihood \( p_M(y|\theta) \) rather than the predictive distribution \( p_M(\tilde{y}|y) \) and do not target any particular functional \( h \). Consequently, the resulting global approximations may be unnecessarily complex or suboptimal locally for \( h(\tilde{y}) \). Indeed, Huggins et al. (2018) show that approximations deemed accurate by KL divergence can produce inaccurate point estimates of important posterior quantities—which may include \( h(\tilde{y}) \).

### 2.3 Examples of targeted prediction

We discuss three examples of predictive decision analysis for targeted approximation. In each case, many Bayesian models \( \mathcal{M} \) are compatible.
Example 1 (Linear contrasts). Consider a (multivariate) regression model $E_{y_i|\theta} y_i = f_\theta(x_i)$ for $y_i = (y_{i,1}, \ldots, y_{i,m})'$. The linear contrast $h(\tilde{y}) = C\tilde{y}$ is often of interest: the matrix $C$ can extract specific components of $\tilde{y}$, evaluate contrasts between components of $\tilde{y}$, and apply a linear weighting scheme to $\tilde{y}$. For functional data with $y_{i,j} = y_i(\tau_j)$, the linear operation $h(\tilde{y}) = C\tilde{y}$ can target specific subdomains of interest, e.g., $h(\tilde{y}) = C\tilde{y} = \{\tilde{y}(\tau) : \tau \in S\}$ for $S \subset T$, and evaluate derivatives of $\tilde{y}(\tau)$. For this class of functionals, the target of approximation simplifies to the posterior expectation $\bar{h}_i = E_{\theta|y_i} \{C f_\theta(\tilde{x})\} = C E_{\theta|y_i} f_\theta(\tilde{x})$.

Given an estimate $\hat{f}_\theta(\tilde{x})$ of the posterior expectation of the regression function $f_\theta$ at $\tilde{x}$, the response variable $\bar{h}_i = C \hat{f}_\theta(x_i)$ needed for (7) is easily computable for many choices of $C$. Notably, the predictive expected contrast $\bar{h}_i = C \hat{f}_\theta(x_i)$ is distinct from the empirical contrast $h(y_i) = Cy_i$: the former can incorporate shrinkage, smoothness, and other regularization of the regression function $f_\theta$ under $M$. As depicted in Figure 2, multiple approximating models $A$ can be optimized for each contrast $C$ under a single Bayesian model $M$.

Example 2 (Functional data summaries). Suppose $h$ is a scalar summary of a curve $\{y(\tau)\}_{\tau \in T}$, such as the maximum $h(\tilde{y}) = \max_\tau \tilde{y}(\tau)$ or the point at which the maximum occurs $h(\tilde{y}) = \arg \max_\tau \tilde{y}(\tau)$, and let $M$ be a Bayesian functional data model (Section 5 provides a detailed example). To select variables for optimal linear prediction of $h(\tilde{y})$, we can apply Corollary 1 with $g(\tilde{x}; \delta) = \tilde{x}'\delta$ and an $\ell_1$-penalty, $P(\delta) = ||\delta||_1 = \sum_{j=1}^p |\delta_j|$: 

$$
\delta_{\hat{A},X} = \arg \min_\delta \left\{ n^{-1} \sum_{i=1}^n ||\bar{h}_i - x_i'\delta||_2^2 + \lambda ||\delta||_1 \right\}
$$

where $X = \{x_i\}_{i=1}^n$ are the observed covariates. The optimal parameters $\delta_{\hat{A},X}$ are readily computable using existing software, such as glmnet in R (Friedman et al. 2010).

In practice, we apply an adaptive variant of the $\ell_1$-penalty. Motivated by the adaptive lasso (Zou 2006), Kowal et al. (2020) introduce the penalty $P(\delta, \theta) = \sum_{j=1}^p \omega_j |\delta_j|$, where $\omega_j = |\beta_j|^{-1}$ and $\beta_j$ are the regression coefficients in a Gaussian linear model $M$. For nonlinear
or non-Gaussian models $\mathcal{M}$ and targeted approximations to $h$, we use the generalized weights $\omega = |\tilde{d}_0|^{-1}$, where $\tilde{d}_0$ is the $\ell_2$-projection of the predictive variables $h(\tilde{y})$ onto the predictor $g$. Bayesian decision analysis requires integration over the unknown $\theta$, so the requisite penalty in (4) or (7) becomes the posterior expectation $P(\delta) := E_{\tilde{y}|y}[P(\delta, \theta) = \sum_{j=1}^{p} \hat{\omega}_j |\delta_j|$ for $\hat{\omega} = E_{\tilde{y}|y}[|\tilde{d}_0|^{-1}]$, which is estimable using posterior predictive samples.

Example 3 (Classification and cross-entropy). Consider a binary functional $h(\tilde{y}) \in \{0, 1\}$, such as a discretized contrast for multivariate data, $h(\tilde{y}) = I\{\tilde{y}_1 > \tilde{y}_2\}$, or exceedance of a threshold $t^*$ for functional data, $h(\tilde{y}) = I(\exists \tau \in \mathcal{T} : \tilde{y}(\tau) > t^*)$. The Bernoulli deviance for the canonical (logistic) link function is given by (8) with $p = 1$, $F_0\{g(\tilde{x}; \delta)\} = \log[1 + \exp\{g(\tilde{x}; \delta)\}]$, $T_0 = 0$, $F_1\{g(\tilde{x}; \delta)\} = g(\tilde{x}; \delta)$, and $T_1\{h(\tilde{y})\} = h(\tilde{y})$. In this case, (8) is the cross-entropy, which is a popular metric for classification. The predictive expectation $T_1 = E_{\tilde{y}|y}T_1\{h(\tilde{y})\} = E_{\tilde{y}|y}h(\tilde{y})$ required by Theorem 2 is simply the posterior predictive probability of $\{h(\tilde{y}) = 1\}$ under model $\mathcal{M}$. Interestingly, $T_1 \in [0, 1]$ is on a continuous scale, and may contain more information than the binary empirical functional $h(y) \in \{0, 1\}$.

3 Predictive inference for model determination

Decision analysis extracts an optimal $\hat{\delta}_A$ by minimizing a posterior (predictive) expected loss function. However, this optimality is obtained only for a given approximating model $A$ and loss function $\mathcal{L}$. The key implication of Theorems 1 and 2 is that optimal point predictions can be computed easily and efficiently for many approximating models $\mathcal{A}$, loss functions $\mathcal{L}$, and functionals $h$. To exploit fully these benefits, additional tools are needed to evaluate, compare, and select among approximating models.

Our strategy features several important attributes. First, predictive performance is evaluated out-of-sample, which best encapsulates the task of predicting new data. Second, we use the Bayesian model $\mathcal{M}$ to provide predictive uncertainty quantification for all evaluations.
and comparisons. Third, we leverage these out-of-sample predictive comparisons to identify not only the best approximating model, but also those approximating models that achieve an acceptable level of accuracy for out-of-sample prediction. The collection of acceptable models can provide more information than the best predictor alone: it allows us to study the shared characteristics of near-optimal models, such as the important covariates, the forms of \( g \) and \( \mathcal{P} \), and the level of complexity required for sufficiently accurate predictions. Lastly, the approach is general: all that is required is a Bayesian model \( \mathcal{M} \), an evaluative loss function \( L \), and the design points at which to evaluate the predictions under some \( g \).

### 3.1 Predictive model evaluation

The path toward model comparisons and selection begins with evaluation of a single approximating model. We proceed nominally using the predictor \( g(\tilde{x}; \hat{\delta}_A) \) for an approximation \( \mathcal{A} \), but note that any point predictor of \( h(\tilde{y}) \) at \( \tilde{x} \) can be used. Let \( L(z, \hat{z}) \) denote the loss associated with a prediction \( \hat{z} \) when \( z \) has occurred. In general, the evaluative loss \( L \) can be different from the loss \( \mathcal{L} \) used for obtaining \( \hat{\delta}_A \), although this is not recommended. We consider both empirical and predictive versions of the loss: the former uses empirical functionals \( z = h(y) \) and relies exclusively on the observed data, while the latter uses predictive functionals \( z = h(\tilde{y}) \) and inherits a predictive distribution under \( \mathcal{M} \).

Out-of-sample evaluation necessitates a division of the data into training and testing sets: all model-fitting and optimization is restricted to the training data, while all predictive evaluations are conducted on the testing data. Dependence on any particular training/testing split is reduced by repeating this procedure for \( K \) randomly-selected splits akin to \( K \)-fold cross-validation; in practice we use \( K = 10 \). Let \( \mathcal{I}_k \subset \{1, \ldots, n\} \) denote the \( k \)th testing set, where each data point appears in (at least) one testing set, \( \bigcup_{k=1}^{K} \mathcal{I}_k = \{1, \ldots, n\} \). We prefer testing sets that are equally-sized, mutually exclusive, and selected randomly from \( \{1, \ldots, n\} \), although other designs are compatible. Importantly, we do not require re-fitting.
of the Bayesian model $\mathcal{M}$ on each training set, and instead use computationally efficient approximation techniques based on a single fit of $\mathcal{M}$ to the full data (see Section 3.3).

For each training/testing split $k$, the out-of-sample empirical and predictive losses are

$$
\mathbb{L}^\text{out}_A(k) := \frac{1}{|I_k|} \sum_{i \in I_k} L\{h(y_i), g(x_i; \hat{\delta}^{-I_k})\}, \quad \mathbb{\tilde{L}}^\text{out}_A(k) := \frac{1}{|I_k|} \sum_{i \in I_k} L\{h(\tilde{y}_i^{-I_k}), g(x_i; \hat{\delta}^{-I_k})\}
$$

(10)

respectively, where $\hat{\delta}^{-I_k}$ is estimated only using the training data $y^{-I_k} := \{y_i\}_{i \not\in I_k}$,

$$
\hat{\delta}^{-I_k} := \arg \min_\delta \mathbb{E}_{[\tilde{y}|y^{-I_k}]\tilde{\mathcal{L}}_\lambda}\{h(\tilde{y}_i), g(\tilde{x}_i; \delta)\}_{i \not\in I_k}
$$

(11)

and similarly $\tilde{y}_i^{-I_k} \sim p_M(\tilde{y}_i|y^{-I_k})$ is the predictive variate at $x_i$ conditional only on the training data. Although in-sample versions are available, there is an important distinction between the out-of-sample predictive distribution, $p_M(\tilde{y}_i|y^{-I_k})$, and the in-sample predictive distribution, $p_M(\tilde{y}_i|y)$. The in-sample version $p_M(\tilde{y}_i|y) = p_M(\tilde{y}_i|y^{-I_k}, y^{I_k})$ conditions on both the training data $y^{-I_k}$ and the testing data $y^{I_k} := \{y_i\}_{i \in I_k}$, which overstates the accuracy and understates the uncertainty for a testing point $\tilde{y}_i$, $i \in I_k$. The out-of-sample version avoids these issues and more closely resembles most practical prediction problems.

Evaluation of $\mathcal{A}$ is based on the averages of (10) across all training/testing splits:

$$
\mathbb{L}_A := \frac{1}{K} \sum_{k=1}^K \mathbb{L}^\text{out}_A(k), \quad \mathbb{\tilde{L}}_A := \frac{1}{K} \sum_{k=1}^K \mathbb{\tilde{L}}^\text{out}_A(k).
$$

(12)

The $K$-fold aggregation averages over two sources of variability in (10): variability in the training sets $\{x_i, y_i\}_{i \not\in I_k}$, each of which results in a distinct estimate of the coefficients $\hat{\delta}^{-I_k}$, and variability in the testing sets $\{x_i, y_i\}_{i \in I_k}$, which evaluates predictions only at the testing design points $\{x_i\}_{i \in I_k}$. The contrast between $\mathbb{L}_A$ and $\mathbb{\tilde{L}}_A$ is important: $\mathbb{L}_A$ is a point estimate of the risk under predictions from $\mathcal{A}$, while $\mathbb{\tilde{L}}_A$ provides the distribution of
out-of-sample loss under different realizations of the predictive variables $h(\tilde{y}_i)$. Specifically, each $h(y_i)$ for $i \in \mathcal{I}_k$ represents one possible realization of the out-of-sample target variable at $x_i$; the predictive variable $h(\tilde{y}_i^{-T_k})$ for $\tilde{y}_i^{-T_k} \sim p_M(\tilde{y}_i | y^{-T_k})$ expresses the distribution of all possible realizations according to $\mathcal{M}$. The predictive loss $\tilde{\mathbb{L}}^{{\text{out}}}_A$ incorporates this distributional information for out-of-sample predictive uncertainty quantification.

3.2 Predictive model selection

The out-of-sample empirical and predictive losses in (12) provide the necessary ingredients for comparison and selection among approximating models. Predictive quantities have proven useful for Bayesian model selection; see Vehtari and Ojanen (2012) for a thorough review. Our goal is not only to identify or select the most accurate model, but also to gather those models that achieve an acceptable level of accuracy. In doing so, we introduce a Bayesian representation of the Rashomon effect, which observes that for many practical applications, many approaches can achieve adequate predictive accuracy (Breiman, 2001).

The proposed notion of “acceptable” accuracy is defined relative to the most accurate model among all approximations $A \in \mathcal{A}$ under consideration. We prefer relative rather than absolute accuracy because it directly references an empirically attainable accuracy level. Following standard practice, we define the best model to be the minimizer of the out-of-sample empirical loss

$$\mathcal{A}_{\text{min}} := \arg \min_{A \in \mathcal{A}} \tilde{\mathbb{L}}^{{\text{out}}}_A \quad (13)$$

exactly as in classical $K$-fold cross-validation. The set $\mathcal{A}$ may include different forms for $g$ and $\mathcal{P}$ and usually will include a path of $\lambda$ values for each $(g, \mathcal{P})$ pair.

For any two models $A, A' \in \mathcal{A}$, let $\tilde{\mathbb{D}}^{{\text{out}}}_{A, A'} := \tilde{\mathbb{L}}^{{\text{out}}}_A - \tilde{\mathbb{L}}^{{\text{out}}}_{A'}$ be the difference in out-of-sample predictive loss. We seek approximations $A$ that perform within a margin $\eta \geq 0$ of the best model, $\tilde{\mathbb{D}}^{{\text{out}}}_{A, \mathcal{A}_{\text{min}}} < \eta$, with probability at least $\varepsilon \in [0, 1]$. The margin $\eta$ acknowledges
that near-optimal performance—especially for simple models—is often sufficient, while the probability level $\varepsilon$ incorporates predictive uncertainty. In concert, $\eta$ and $\varepsilon$ provide domain-specific and model-informed leniency for admission into a set of acceptable models. We formally define the set of acceptable models as follows:

**Definition 2.** The set of acceptable models is $\Lambda_{\eta,\varepsilon} := \{ A \in \mathcal{A} : \mathbb{P}_M(\mathbb{D}_{\mathcal{A},\mathcal{A}_{\text{min}}}^{\text{out}} < \eta) \geq \varepsilon \}$, where $\mathbb{D}_{\mathcal{A},\mathcal{A}_{\text{min}}}^{\text{out}} = \mathbb{I}_A^{\text{out}} - \mathbb{I}_{\mathcal{A}_{\text{min}}}^{\text{out}}$ and $\mathcal{A}_{\text{min}}$ is defined in (13).

The probability $\mathbb{P}_M$ is estimated using out-of-sample predictive draws under model $\mathcal{M}$ (see Section 3.3). The set of acceptable models is nonempty, since $\mathcal{A}_{\text{min}} \in \Lambda_{\eta,\varepsilon}$ for all $\eta, \varepsilon$. In addition, these sets are nested: $\Lambda_{\eta',\varepsilon'} \subseteq \Lambda_{\eta,\varepsilon}$ for any $\eta' \geq \eta$ or $\varepsilon' \leq \varepsilon$, so increasing $\eta$ or decreasing $\varepsilon$ can expand the set of acceptable models. The special case of sparse Bayesian linear regression was considered in Kowal et al. (2020). With similar intentions, Tulabandhula and Rudin (2013) and Semenova and Rudin (2019) define a Rashomon set as the set of models for which the in-sample loss is within a margin $\eta$ of the best model. The acceptable model set has two distinguishing features: it uses out-of-sample criteria for evaluation and incorporates predictive uncertainty via the Bayesian model $\mathcal{M}$.

The acceptable model set also can be constructed using prediction intervals:

**Lemma 1.** A model is acceptable $\mathcal{A} \in \Lambda_{\eta,\varepsilon}$ if and only if there exists a lower $(1-\varepsilon)$ posterior prediction interval for $\mathbb{D}_{\mathcal{A},\mathcal{A}_{\text{min}}}^{\text{out}}$ that includes $\eta$.

Viewed another way, $\mathcal{A}$ is not acceptable if the lower $1 - \varepsilon$ predictive interval for $\mathbb{D}_{\mathcal{A},\mathcal{A}_{\text{min}}}^{\text{out}}$ excludes $\eta$. From this perspective, unacceptable models are those $\mathcal{A}$ for which there is insufficient predictive probability (under $\mathcal{M}$) that the out-of-sample accuracy of $\mathcal{A}$ is within a certain margin of the best model. This definition is similar to the confidence sets of Lei (2019), which exclude any $\mathcal{A}$ for which the null hypothesis that $\mathcal{A}$ produces best predictive risk is rejected. Lei (2019) relies on a customized bootstrap procedure, which adds substantial computational burden to the model-fitting and cross-validation procedures. By
comparison, acceptable model sets are derived entirely from the predictive distribution of \( \mathcal{M} \) and accompanied by fast and accurate approximation algorithms (see Section 3.3).

Among acceptable models, we highlight the simplest model. For fixed \((g, \mathcal{P})\), the simplest model has the largest complexity penalty:

\[
\mathcal{A}_{\eta, \varepsilon}(g, \mathcal{P}) := (g, \mathcal{P}, \lambda_{\eta, \varepsilon}), \quad \lambda_{\eta, \varepsilon} := \max \{ \lambda : (g, \mathcal{P}, \lambda) \in \Lambda_{\eta, \varepsilon} \}. \tag{14}
\]

When \( \mathcal{P} \) is a sparsity penalty such as (9), the simplest acceptable model contains the smallest set of covariates needed to (nearly) match the predictive accuracy of the best model \( \mathcal{A}_{\min} \). In some cases, the best model is also the simplest acceptable model, \( \mathcal{A}_{\min} = \mathcal{A}_{\eta, \varepsilon}(g, \mathcal{P}) \). Selection based on \( \lambda_{\eta, \varepsilon} \) resembles the one-standard-error rule (e.g., Hastie et al., 2009), which selects the simplest model for which the out-of-sample empirical loss is within one standard error of the best model. Instead, \( \lambda_{\eta, \varepsilon} \) uses the out-of-sample predictive loss, with uncertainty quantification inherited from the out-of-sample predictive distribution of \( \mathcal{M} \).

### 3.3 Fast approximations for out-of-sample predictive evaluation

The primary hurdle for out-of-sample predictive evaluations is computational: they require solving (11) for \( \hat{\delta}_{\mathcal{A}}^{\mathcal{T}_k} \) and sampling \( \tilde{y}_i^{\mathcal{T}_k} \sim p_{\mathcal{M}}(\tilde{y}_i | y^{\mathcal{T}_k}) \) for each training/testing split \( k = 1, \ldots, K \). Re-fitting \( \mathcal{M} \) on each training set \( \{x_i, y_i\}_{i \notin \mathcal{T}_k} \) is impractical and in many cases computationally infeasible. To address these challenges, we develop efficient approximations that require only a single fit of the Bayesian model \( \mathcal{M} \) to the data—which is already necessary for standard posterior inference. Specifically, we use a sampling-importance resampling (SIR) algorithm with the full posterior predictive distribution as a proposal for the relevant out-of-sample predictive distributions. The subsequent results focus on the squared error loss (3), but adaptations to other loss functions such as (8) are straightforward.
To obtain \( \delta_{-I_k}^{\mathcal{A}} \), we reformulate the optimization (11) in a more convenient representation:

\[
\delta_{-I_k}^{\mathcal{A}} = \arg \min_{\delta} \left\{ (n - |I_k|)^{-1} \sum_{j \notin I_k} \| \bar{h}_j - I_k \|_2^2 + \lambda P(\delta) \right\} \tag{15}
\]

where \( \bar{h}_j = \mathbb{E}[\tilde{y}_j | y_{-I_k}] h(\tilde{y}_j) \) is the out-of-sample point prediction at \( x_j \). The equivalence between (11) and (15) is a simple consequence of Corollary 1. As such, (15) is easily solvable for many choices of \( \mathcal{A} \): all that is required is estimation of the out-of-sample point prediction \( \bar{h}_j \) for each \( j \notin I_k \) in the training set. A preliminary approximation of \( \bar{h}_j \) is available using importance sampling:

\[
\bar{h}_j = \int \int h(\tilde{y}_j) p_M(\tilde{y}_j | \theta) p_M(\theta | y_{-I_k}) d\tilde{y}_j d\theta \approx \sum_{s=1}^S w_s h(\tilde{y}_s) \tag{16}
\]

where \( \{w_s\}_{s=1}^S \) are the importance weights for training dataset \( k \) and \( \{\tilde{y}_s\}_{s=1}^S \) are samples from a proposal distribution. It some cases, it is beneficial to regularize the importance weights (Ionides, 2008; Vehtari et al., 2015).

The integral in (16) suggests a sampling procedure: we generate proposals \( \{\tilde{y}_s\}_{s=1}^S \sim p_M(\tilde{y} | y) \) from the full predictive distribution by sampling \( \{\theta_s\}_{s=1}^S \sim p_M(\theta | y) \) from the full posterior and \( \{\tilde{y}_s\}_{s=1}^S \sim p_M(\tilde{y}_j | \theta_s) \) from the likelihood. The full data posterior \( p_M(\theta | y) \) serves as a proposal for the training data posterior \( p_M(\theta | y_{-I_k}) \) with importance weights \( w_s \propto 1/p(y_{I_k} | \theta_s) = \prod_{i \in I_k} 1/p(y_i | \theta_s) \) where factorization follows under conditional independence. Variants of this approach have been used successfully for Bayesian model selection (Gelfand et al., 1992) and evaluating prediction distributions (Vehtari and Ojanen, 2012).

SIR provides a mechanism for sampling \( \tilde{y}_i^{-I_k} \sim p_M(\tilde{y}_i | y_{-I_k}) \) using the importance weights \( \{w_s\}_{s=1}^S \), which in turn provides out-of-sample predictive draws of \( \tilde{I}_A^{out} \) and \( \tilde{D}_A^{out} \) for any approximations \( \mathcal{A}, \mathcal{A}' \in \mathcal{A} \). The idea is to obtain the proposal samples \( \{\tilde{y}_s\}_{s=1}^S \sim p_M(\tilde{y}_j | y) \) from the full posterior distribution and then subsample from \( \{\tilde{y}_s\}_{s=1}^S \) without re-
placement based on the corresponding importance weights \( \{w^s_k\}_{s=1}^S \). The full SIR algorithm details are provided in the supplementary material.

4 Simulation study

We evaluate the selection capabilities and predictive accuracy of the proposed techniques using synthetic data. The synthetic data must be constructed carefully: for targeted prediction, the covariates \( x \) are associated with a particular functional \( h \) of the response variable.

We generate functional data such that the argmax of each function is linearly associated with a subset of the covariates. The true functions \( \{Y^*_i(\tau) : \tau \in [0, 1]\} \) are piecewise linear and concave with a single breakpoint, \( \tau^*_i := \arg \max_{\tau} Y^*_i(\tau) = h(Y^*_i) \), which is determined by covariates \( \tau^*_i = x'_i \beta^* \). The covariates are correlated and mixed continuous and discrete: we draw \( x_{i,j} \) from marginal standard normal distributions with \( \text{Cor}(x_{i,j}, x_{i,j}') = 0.75 \) and binarize half of these \( p \) variables, \( x_{i,j} \leftarrow \mathbb{I}\{x_{i,j} \geq 0\} \). The continuous covariates are centered and scaled to sample mean zero and sample standard deviation 0.5. For the true coefficients \( \{\beta^*_j\}_{j=1}^p \), we initially select 5% for \( \beta^*_j = 1 \), 5% for \( \beta^*_j = -1 \), and leave the remaining values at zero with the exception of the intercept, \( \beta^*_0 = 1 \). We restrict \( h(Y^*_i) = x'_i \beta^* \) to the interval \([0.2, 0.8]\) by shifting and scaling the (nonzero) coefficients as follows: set \( \beta^*_0 \leftarrow \min\{x'_i \beta^*\} \); shift \( \beta^*_j \leftarrow \beta^*_j \times 0.6/\text{range}\{x'_i \beta^*\} \) for \( j = 0, 1, \ldots, p \); and reset \( \beta^*_0 \leftarrow \beta^*_0 + 0.2 \). Proceeding with \( \tau^*_i := x'_i \beta^* \), the true functions are computed as \( Y^*_i(\tau) = a_{0,i} + a_{1,i}\tau - (a_{1,i} + a_{2,i})(\tau - \tau^*_i)_+ \), where \( a_{0,i} \overset{iid}{\sim} N(0, 1) \), \( a_{1,i}, a_{2,i} \overset{iid}{\sim} \chi^2_5 \), and \( (x)_+ := x\mathbb{I}\{x \geq 0\} \). Finally, the observed data \( y_i \) are generated by adding Gaussian noise to \( Y^*_i(\tau) \) at \( m \) equally-spaced points with a root signal-to-noise ratio of 5. Example figures are provided in the supplementary material.

The synthetic data-generating process is repeated 100 times for \( p = 50 \) covariates, \( m = 50 \) observation points, and varying sample sizes \( n \in \{75, 100, 500\} \). For each simulated dataset \( \{x_i, y_i\}_{i=1}^n \), we compute the posterior and predictive distributions under the Bayesian
function-on-scalars regression model of Kowal and Bourgeois (2020), which models a linear association between the functional data response and the scalar covariates. We emphasize that this model $\mathcal{M}$ does not reflect the synthetic data-generating process, yet our targeted predictions will be based entirely on the posterior and predictive distributions under $\mathcal{M}$. We consider linear approximating models $g(\tilde{x}; \delta) = \tilde{x}'\delta$ with the adaptive $\ell_1$-penalty from Example 2 and computed using glmnet in R (Friedman et al., 2010). In this case, model complexity is determined by sparsity, and the set of approximating models $\mathcal{A}$ is entirely determined by the path of $\lambda$ values. For benchmark comparisons, we fit the adaptive lasso (Zou, 2006) to the empirical functionals $\{x_i, h(y_i)\}_{i=1}^n$ and select the tuning parameter using the one-standard-error rule from 10-fold cross-validation.

To validate the proposed definition of acceptable model sets, we investigate a simple yet important question: does the true model belong to $\Lambda_{\eta,\varepsilon}$? Specifically, we determine whether the true set of active variables $\{j: \beta_j^* \neq 0\}$ matches the set of active variables for any acceptable model $A \in \Lambda_{\eta,\varepsilon}$. This task is challenging: we do not assume knowledge of the active variables, so the true model only belongs to $\Lambda_{\eta,\varepsilon}$ when it is both correctly identified along the $\lambda$ path and correctly evaluated by $\tilde{D}_{\mathcal{A}, \mathcal{A}'}^{\text{out}}$. Correct identification is only satisfied when all and only the true active variables $\{j: \beta_j^* \neq 0\}$ are nonzero according to $\mathcal{A}$.

For this task, we compute $\varepsilon_{\max}(A^*) := P_{\mathcal{M}}(\tilde{D}_{A^*, \mathcal{A}_{\text{min}}}^{\text{out}} < \eta)$, which is the maximum probability level for which the true model $A^*$ is acceptable. By design, $A^* \in \Lambda_{\eta,\varepsilon'}$ remains acceptable for any smaller probability level $\varepsilon' \leq \varepsilon_{\max}(A^*)$. Most important, we set $\varepsilon_{\max}(A^*) = 0$ if $A^*$ is not on the $\lambda$ path. For each simulated dataset, we compute $\varepsilon_{\max}(A^*)$ for a grid of $\eta$ values, which are expressed as a percentage of the true observation error variance. The results averaged across 100 simulations are in Figure 3. Naturally, $\varepsilon_{\max}(A^*)$ uniformly increases with the sample size for all values of $\eta$. When $\eta = 0$, the average maximum probability levels are $\varepsilon_{\max}(A^*) \in \{0.08, 0.21, 0.50\}$ for $n \in \{75, 100, 500\}$, respectively, which suggests that a cutoff of $\varepsilon = 0.05$ or $\varepsilon = 0.1$ is capable of capturing the true model even when zero
margin is allowed. Notably, $\varepsilon_{max}(A^*)$ does not converge to one as $\eta$ increases for the smaller sample sizes $n \in \{75, 100\}$. The reason is simple: if $A^*$ is not discovered along the $\lambda$ path, then $\varepsilon_{max}(A^*) = 0$ by definition—regardless of the choice of $\eta$. This result demonstrates the importance of the set of approximating models under consideration, $A$, which here is determined entirely by the selected variables in the glmnet solution path.

Figure 3: The maximum probability level $\varepsilon_{max}(A^*)$ for which the true model is acceptable, $A^* \in \Lambda_{\eta,\varepsilon}$, across values of $\eta$. For any smaller probability level $\varepsilon' \leq \varepsilon_{max}(A^*)$, the true model remains acceptable: $A^* \in \Lambda_{\eta,\varepsilon'}$. The horizontal gray line is $\varepsilon = 0.1$.

Next, we evaluate point predictions of $h(Y_i^*)$ and estimates of $\beta^*$ using root mean squared errors (RMSEs). For the proposed approach, the coefficients $\hat{\delta}_\lambda$ and point predictions $g(\bar{x}; \hat{\delta}_\lambda) = \bar{x}'\hat{\delta}_\lambda$ are computed for multiple choices of $\lambda$: the simplest acceptable model $\lambda = \lambda_{\eta,\varepsilon}$ with $\eta = 0$ and $\varepsilon = 0.1$ (proposed(out)); the analogous choice of $\lambda$ based on in-sample evaluations (proposed(in)); and the full linear approximation with $\lambda = 0$ (proposed(full)). For comparisons, we include the aforementioned adaptive lasso, the point predictions $\bar{h}_i$ under model $M$ (h_bar), and the empirical functionals $h(y_i)$ ($h(y)$). The results are in Figure 4. In summary, clear improvements in targeted prediction are obtained by (i) fitting to $h(\tilde{y}_i)$ (via $\tilde{h}_i$) rather than $h(y_i)$, (ii) including covariate information, (iii) incorporating penalization or variable selection, and (iv) selecting the complexity $\lambda$ based on out-of-sample evaluations. Perhaps most importantly, the targeted approxima-
tions $\mathcal{A}$ vastly outperform the model $\mathcal{M}$ predictions—even though $\mathcal{A}$ is based entirely on the posterior and predictive distributions from $\mathcal{M}$. Similar results are confirmed for $n = 100$ in the supplementary material.

Figure 4: RMSEs for the true functionals $h(Y^*)$ (left) and the true regression coefficients $\beta^*$ (right) across 100 simulated datasets. Non-overlapping notches indicate significant differences between medians. The targeted approximations are most accurate for prediction and estimation.

5 Physical activity data analysis

We apply targeted prediction to study physical activity (PA) data from the National Health and Nutrition Examination Survey (NHANES). NHANES is a large survey conducted by the Centers for Disease Control to study the health and wellness of the U.S. population. We analyze data from the 2005-2006 cohort, which features minute-by-minute PA data measured by hip-worn accelerometers (see Figure 1). To date, the 2005-2006 cohort is the most recent publicly available NHANES PA data. These data are high-resolution and empirical measurements of PA, and offer an opportunity to study intraday activity profiles.

PA has been linked to all-cause mortality not only in total daily activity (Schmid et al., 2015) but also via other functionals that describe activity behaviors (Fishman et al., 2016; Smirnova et al., 2019). Our goal is to construct targeted approximations that better predict
and illuminate the defining characteristics of PA. Specifically, we consider the following functionals $h(\tilde{y})$ for intraday PA $\tilde{y} = (\tilde{y}(\tau_1), \ldots, \tilde{y}(\tau_m))^t$ at times-of-day $\tau_1, \ldots, \tau_m$:

<table>
<thead>
<tr>
<th>avg $\int \tilde{y}(\tau),d\tau$</th>
<th>t lac $\int \log {\tilde{y}(\tau) + 1},d\tau$</th>
<th>sd $|\tilde{y} - \int \tilde{y}(\tau),d\tau|_{L^2}$</th>
<th>sedentary $\int \mathbb{I}{\tilde{y}(\tau) \leq 100},d\tau$</th>
<th>max $\max_{\tau} \tilde{y}(\tau)$</th>
<th>argmax $\arg \max_{\tau} \tilde{y}(\tau)$</th>
</tr>
</thead>
</table>

where avg captures average daily activity, t lac is the total log activity count and measures moderate activity \cite{Wolff-Hughes et al. 2018}, sd targets the intraday variability in PA, sedentary computes the amount of time below a low activity threshold, max is the peak activity level, and argmax is the time of peak activity. In addition, we include a binary indicator of absolute inactivity during sleeping hours: $\text{zeros}(1\text{am}-5\text{am}) := \mathbb{I}\{\tilde{y}(\tau) = 0\}$ for all $\tau \in [1\text{am}, 5\text{am}]$. Individuals with $\text{zeros}(1\text{am}-5\text{am}) = 1$ likely removed the accelerometer during sleep in accordance with the NHANES instructions. Since we omit subjects with insufficient accelerometer wear time ($< 10$ hours), individuals with $\text{zeros}(1\text{am}-5\text{am}) = 1$ are active at other times of the day.

The PA data are accompanied by demographic variables (age, gender, body mass index (BMI), race, and education level), behavioral attributes (smoking status and alcohol consumption), self-reported comorbidity factors (diabetes, coronary heart disease (CHD), congestive heart failure (CHF), cancer, and stroke), and lab measurements (total cholesterol, HDL Cholesterol, systolic blood pressure). Data pre-processing generally follows \cite{Leroux et al. 2019} using the R package rnhanesdata. We consider individuals aged 50-85 without mobility problems and without missing covariates. The continuous covariates are centered and scaled to have sample mean zero and sample standard deviation 0.5 \cite{Gelman 2008}.

In accordance with the schematic in Figure 2, targeted predictive decision analysis begins with a Bayesian model $M$. Since the PA data are intraday activity counts, we use a count-valued functional regression model based on the simultaneous transformation and rounding (STAR) framework of \cite{Kowal and Canale 2020}. STAR formalizes the popular approach of transforming count data prior to apply Gaussian models, but includes a latent rounding
layer to produce a valid count-valued data-generating process. STAR models can capture zero-inflation, over- and under-dispersion, and boundedness or censoring, and provide a path for adapting continuous data models and algorithms to the count data setting.

For each individual, we aggregate PA across all available days (at least three and at most seven days per subject) in five-minute bins. Let $y_{i,j}$ and $y_{i,j}^{tot}$ and denote the average and total PA, respectively, for subject $i$ at time $\tau_j$, where $i = 1, \ldots, n = 1012$ and $j = 1, \ldots, m = 288$. Total PA is count-valued and will serve as the input for the STAR model, while all subsequent functionals and predictive distributions use average PA. Model $\mathcal{M}$ is the following:

$$y_{i,j}^{tot} = \text{round}(y_{i,j}^*), \quad z_{i,j}^* = \text{transform}(y_{i,j}^*)$$

$$z_{i,j}^* = b^j(\tau_j)\theta_i + \sigma_\epsilon \epsilon_i, \quad \epsilon_i \sim t_\nu(0, 1)$$

$$\theta_{i,\ell} = x'_{i} \alpha_\ell + \sigma_\gamma \gamma_{i,\ell}, \quad \gamma_{i,\ell} \sim \text{Gamma}(0.01, 0.01)$$

with the priors $\alpha_{\ell,j} \sim \text{N}(0, \sigma_\alpha^2)$ and $\sigma_\epsilon^{-2}, \sigma_\gamma^{-2}, \sigma_\alpha^{-2} \sim \text{Gamma}(0.01, 0.01)$. In \[17\], round maps the latent continuous data $y_{i,j}^*$ to $\{0, 1, \ldots, \infty\}$, while transform maps $y_{i,j}^*$ to $\mathbb{R}$ for continuous data modeling. We use round$(t) = \lfloor t \rfloor$ for $t > 0$ and round$(t) = 0$ for $t \leq 0$, so $y_{i,j}^{tot} = 0$ whenever $y_{i,j}^* < 0$. Within the Box-Cox family, we find that transform$(t) = 2(\sqrt{t} - 1)$ is adequate for the predictive functionals of interest (see Figure 5). In the functional regression levels \[18\]-\[19\], $b$ is a vector of basis functions with basis coefficients $\theta_i$ for subject $i$ and $\alpha_\ell$ is the vector of regression coefficients for each basis coefficient. We use a spline basis with the reparametrization of Scheipl et al. (2012), which simultaneously orthogonalizes $b$ and diagonalizes the prior variance of the basis coefficients. This diagonalization justifies the assumption of independence across basis coefficients in \[19\]. Heavy-tailed innovations ($\nu = 3$) are introduced to model large spikes in PA.

Posterior inference is conducted based on 5000 samples from a Gibbs sampler after discarding a burn-in of 5000 iterations; the algorithm is detailed in the supplementary material.
Posterior predictive diagnostics for the functionals of interest are provided in Figure 5, which plots the sample (kernel) density estimates for the empirical functionals \( \{ h(y_i) \}_{i=1}^{n} \) and the predictive functionals \( \{ h(\tilde{y}_i) \}_{i=1}^{n} \) for 500 draws from the posterior predictive distribution under model \( \mathcal{M} \) (see the supplement for evaluation of \texttt{zeros(1am-5am)}). There is substantial overlap between the densities of the empirical and predictive functionals, which suggests adequacy of \( \mathcal{M} \) for these functionals. These encouraging results are insensitive to \( \nu \), but alternative choices of \texttt{transform} or \texttt{b} (such as wavelets) produce inferior results.

Figure 5: Kernel density estimates for the empirical functionals \( \{ h(y_i) \}_{i=1}^{n} \) (black line) and the predictive functionals \( \{ h(\tilde{y}_i) \}_{i=1}^{n} \) (gray lines) for 500 predictive samples. The model \( \mathcal{M} \) in (17)-(19) appears to be adequate for these functionals, with some difficulty for \texttt{argmax} around the mode.

Targeted approximations for each functional were constructed using a linear model \( g(\tilde{x}; \delta) = \tilde{x}'\delta \) with an adaptive \( \ell_1 \)-penalty (see Example 2). Regression trees were also considered but were not competitive. The set of approximations \( \mathbb{A} \) is given by the path of \( \lambda \) values computed using \texttt{glmnet} in \texttt{R} \cite{glmnet}: we highlight the simplest acceptable model \( \lambda = \lambda_{0,0.1} \) (\texttt{proposed(out)}) and the full model \( \lambda = 0 \) (\texttt{proposed(full)}). For comparison, we fit an adaptive lasso to \( \{ x_i, h(y_i) \}_{i=1}^{n} \) for each \( h \). Squared error loss is used for all but \texttt{zeros(1am-5am)} which uses cross-entropy. In the supplementary material, we also consider
quadratic effects for age and BMI as well as pairwise interactions for each of age and BMI with ethnicity, gender, the behavioral attributes, and the self-reported comorbidity factors.

The targeted predictions are evaluated out-of-sample using the approximations from Section 3.3. For each functional $h$ and model size indexed by $\lambda$, Figure 6 presents the predictive and empirical loss relative to the best model $A_{\text{min}}$. The measures of vigorous PA—$\text{avg}$, $\text{sd}$, and $\text{max}$—produce nearly identical results, so we include only $\text{max}$. The predictive expectations align closely with the empirical values, which suggests that model $M$ is adequate for these predictive metrics. For each functional, we obtain optimal or near-optimal approximations with only about 10 covariates, with better accuracy than the adaptive lasso. Many of the selected covariates are shared among functionals: age, BMI, gender, ethnicity, HDL cholesterol, and CHD are selected for all but $\text{argmax}$, while smoking status ($\text{avg}$, $\text{sd}$, $\text{max}$), diabetes ($\text{avg}$, $\text{sd}$, $\text{sedentary}$, $\text{max}$), and total cholesterol ($\text{tlac}$, $\text{sedentary}$) appear as well. The functionals measuring vigorous PA agree on the selected variables, including negative effects for diabetes and smoking. Most distinct is $\text{argmax}$: while $A_{\text{min}}$ includes 11 covariates, the predictive uncertainty quantification from $\hat{\eta}_{A_{\text{min}}}^{\text{out}}$ indicates that approximations with as few as one covariate (ethnicity) are acceptable. These covariates are simply not linearly predictive of $\text{argmax}$: the difference between $A_{\text{min}}$ and any other $A$ is less than 1%.

To validate the approximations in Figure 6, we augment the analysis with a truly out-of-sample prediction evaluation. For each of 20 training/testing splits, model $M$ and the adaptive lasso are fit to the training data and sparse linear approximations are targeted to each $h$. We emphasize that this exercise is computationally intensive: the MCMC for model $M$ requires about 30 minutes per 10000 iterations (using R on a MacBook Pro, 2.8 GHz Intel Core i7), so repeating the model-fitting process 20 times is extremely slow. Comparatively, the approximations used for Figure 6 compute in under two seconds.

Point predictions were generated for the testing data using model $M$ ($h_{\text{bar}}$), the adaptive lasso, and the sparse linear approximations with $\lambda = \lambda_{0.01}$ (proposed(out)), $\lambda = 0$
Figure 6: Approximate out-of-sample squared error loss for sparse linear models targeted to each functional. Results are presented for each model size as a percent difference relative to $A_{min}$. The predictive expectations (triangles) and 80% intervals (gray bars) are included with the empirical relative loss for each model size (x-marks) and the adaptive lasso (red lines). The horizontal black lines denote the choices of $\eta$ and the vertical lines denote $\lambda_{\eta,0.1}$ (solid) and $A_{min}$ (dashed).

(proposed(full)) and $A_{min}$. The point predictions under $\mathcal{M}$ are not only highly inaccurate—and so excluded from Figure 7—but also slow to compute: we draw $\tilde{y} \sim p_\mathcal{M}(\tilde{y}|y)$ at each out-of-sample testing point $\tilde{x}$ and then average $h(\tilde{y})$ over these draws. The targeted approximations simply evaluate $g(\tilde{x}; \hat{\delta}_A) = \tilde{x}'\hat{\delta}_A$, which is faster, simpler, less susceptible to Monte Carlo error, and empirically more accurate. Predictions were evaluated on the empirical functionals $h(y_i)$ in the testing data using mean squared prediction error.

The results from the out-of-sample prediction exercise are in Figure 7. The smallest acceptable model proposed(out) performs almost identical to the best model $A_{min}$ despite using fewer covariates. Both proposed(out) and proposed(full) outperform the adaptive
lasso, in some cases by a large margin. The strength of this result is remarkable: the predictions are evaluated on the empirical functionals $h(y_i)$, which are used for training the adaptive lasso but not the proposed methods. Instead, proposed(out), proposed(full), and $A_{min}$ are trained using $\bar{h}_i$ (see Corollary 1)—which is itself a poor out-of-sample predictor. However, the targeted approximations only rely on the in-sample adequacy of $\bar{h}_i$ (see Figure 5) and, unlike models trained to the empirical functionals, leverage both the model-based regularization and the uncertainty quantification provided by $M$. In summary, the targeted predictors improve upon both the empirical predictor and the model-based predictor from which they were derived. Lastly, we note that Figure 7 confirms Figure 6 which validates the accuracy of the out-of-sample approximations from Section 3.3.

Since NHANES data are collected using a stratified multistage probability sampling design, it is natural to question the absence of survey weights from this analysis. Although it is straightforward to incorporate the survey weights into the aggregate loss function (5) to mimic a design-based approach (e.g., [Rao, 2011]), the unweighted approach has its merits. By design, NHANES oversamples certain subpopulations to ensure representation in the dataset. So although our out-of-sample predictions are not evaluated on a representative sample of the U.S. population, they are evaluated on a carefully-curated sample that includes key demographic, income, and age groups within the U.S. population.

6 Discussion

Using predictive decision analysis, we constructed optimal, simple, and efficient predictions from Bayesian models. These predictions were targeted to specific functionals to provide customization for key decision tasks and new avenues for model summarization. Out-of-sample predictive evaluations were computed using fast approximation algorithms and accompanied by predictive uncertainty quantification. Simulation studies demonstrated the predictive and
Figure 7: Mean squared prediction error for each functional across 20 training/testing splits. Results are presented as percent difference relative to $A_{\text{min}}$; values below zero (vertical line) indicate improvement relative to $A_{\text{min}}$. Non-overlapping notches indicate significant differences between medians. Point predictions from $M$ ($h_{\text{bar}}$) are noncompetitive and omitted. Both proposed(out) proposed(full) improve upon both adaptive lasso and $h_{\text{bar}}$, while proposed(out) is most accurate and performs almost identical to $A_{\text{min}}$ despite using fewer covariates.

model selection capabilities of the proposed approach. The methods were applied to a large physical activity dataset, for which we built a count-valued functional regression model. Us-
ing targeted prediction with sparse linear approximations, we identified 10 covariates that provide near-optimal out-of-sample predictions for important and descriptive PA functionals, with substantial gains in accuracy over both Bayesian and non-Bayesian predictors.

The proposed framework leverages—and requires—the predictive distribution under a Bayesian model $M$. When the predictive distribution of $M$ is intractable or computationally prohibitive, these methods remain compatible with any approximation algorithm for the predictive distribution of $h(\tilde{y})$. Customized modifications of existing algorithms, such as variational inference (Blei et al., 2017) or loss-calibrated inference (Lacoste-Julien et al., 2011), can be designed for this purpose. Such approximations could bypass unnecessary posterior computations and provide efficient updates for the predictive distribution of $h(\tilde{y})$—as well as the accompanying targeted approximations—as new data arrive.

Acknowledgements

Research was sponsored by the Army Research Office (W911NF-20-1-0184) and the National Institute of Environmental Health Sciences of the National Institutes of Health (R01ES028819). The content, views, and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, the National Institutes of Health, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References


