HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2023 February 27.

Published in final edited form as:

J Am Stat Assoc. 2022 ; 117(540): 1875–1886. doi:10.1080/01621459.2021.1891926.

Fast, Optimal, and Targeted Predictions using Parametrized Decision Analysis

Daniel R. Kowal* **[Dobelman Family Assistant Professor]**

*Department of Statistics, Rice University

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. Although classical decision analysis extracts predictions from a Bayesian model, these predictions are often difficult to interpret and slow to compute. Instead, we design a class of *parametrized actions* for Bayesian decision analysis that produce optimal, scalable, and simple targeted predictions. For a wide variety of action parametrizations and loss functions—including linear actions with sparsity constraints for targeted variable selection—we derive a convenient representation of the optimal targeted prediction 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, or *acceptable* targeted predictors, which 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 predictions 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

Bayesian statistics; functional data; physical activity; variable selection

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.

(daniel.kowal@rice.edu).

Our focus is *targeted prediction*, where predictions are customized for the decision tasks of interest. The translation of models into actionable decisions requires predictive quantities in the form of *functionals* of future or unobserved data. Predictions should be optimized for these decision tasks—and targeted to the relevant functionals. The target is fundamental for defining the correct (predictive) likelihood (Bjornstad, 1990). Absent specific functionals of interest, targeted prediction offers a path for interpretable statistical learning: 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; 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 the trajectories. Figure 1 shows several such functionals: the average activity (avg), the peak activity level (max), and the time of peak activity (argmax). These functionals summarize daily PA and describe clear sources of variability in PA among the individuals. Other features are discernible, such as sedentary behavior and periods of absolute inactivity, and are investigated in Section 5. However, Bayesian model-based point predictions alone do not explain what drives the variability among individuals and can be slow to compute out-of-sample.

Our goal is construct targeted predictions that improve accuracy, streamline decision making, and highlight the model attributes and covariates that matter most for prediction —which notably may differ among functionals. Building upon classical decision analysis, we introduce *parametrized actions* that extract optimal, simple, and fast predictions under any Bayesian model $\mathscr{M}$. The parameterizations exploit familiar model structures, such as linear, tree, and additive forms, while the actions minimize a posterior predictive expected loss that is customized for each functional. For a broad class of parametrized actions and loss functions, we derive a convenient representation of the optimal targeted prediction that yields efficient and interpretable solutions. These solutions can be computed using existing software packages for penalized regression, which allows for widespread and immediate deployment of the proposed techniques. The targeted predictions are constructed simultaneously for multiple functionals based on a single $\mathscr{M}$, which avoids the need to re-fit a Bayesian model for each functional. While intrinsically useful for prediction, the elicitation of multiple targeted predictors is also informative for understanding and summarizing the model $\mathscr{M}$ posterior.

A key feature of our approach is the use of the model $\mathscr{M}$ predictive distribution to provide uncertainty quantification for out-of-sample predictive evaluation. We design a procedure to identify not only the most accurate targeted 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* predictors is informative: it describes the shared characteristics and level of complexity needed for near-optimal targeted prediction. We do not require any re-fitting of $\mathscr{M}$ 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 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. The proposed methods expand upon a line of research for *posterior summarization*, most commonly for Bayesian variable selection, advocated by Lindley (1968) and rekindled by Hahn and Carvalho (2015). These techniques have been adapted for seemingly unrelated regressions (Puelz et al., 2017), graphical models (Bashir et al., 2019), nonlinear regressions (Woody et al., 2020), functional regression (Kowal and Bourgeois, 2020), and time-varying parameter models (Huber et al., 2020). Alternative approaches combine linear variable selection with Kullback-Leibler distributional approximations (Goutis and Robert, 1998; Nott and Leng, 2010; Tran et al., 2012; Crawford et al., 2019; Piironen et al., 2020). In general, these methods focus on global summarizations of a particular model $\mathscr{M}$ posterior distribution. By comparison, our emphasis on *predictive functionals* adds specificity and a direct link to the observables, which provides a solid foundation for (out-of-sample) predictive evaluations and broadens applicability among Bayesian models with different parameterizations.

The remainder of the paper is organized as follows. Section 2 introduces predictive decision analysis for optimal targeted prediction. Section 3 develops the methods and algorithms for predictive evaluations and comparisons. A simulation study is in Section 4. The PA data are analyzed in Section 5. Section 6 concludes. Online supplementary material includes methodological generalizations and further examples, computational details, additional results for the simulated and PA data, proofs, and R code to reproduce the analyses.

2 Targeted point prediction

Consider the paired data $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$ with p-dimensional covariates $\boldsymbol{x}_i$ and m-dimensional response $\boldsymbol{y}_i$. The response variables $\boldsymbol{y}_i$ may be univariate ($m = 1$), multivariate ($m > 1$), or functional data with $\boldsymbol{y}_i = (y_i(\tau_1), \ldots, y_i(\tau_m))'$ observed on a domain $\mathcal{T} \subset \mathbb{R}^d$. Suppose we have a satisfactory Bayesian model $\mathscr{M}$ parametrized by $\boldsymbol{\theta}$ with posterior $p_{\mathscr{M}}(\boldsymbol{\theta} \mid \boldsymbol{y})$. The requisite notion of "satisfactory" is made clear below, but fundamentally $\mathscr{M}$ should encapsulate the modeler's beliefs about the data-generating process and demonstrate empirically the ability to capture the essential features of the data. Although these criteria are standard for Bayesian modeling, they often demand highly complex and computationally intensive models. There is broad interest in extracting simple, accurate, and computationally efficient representations or summaries of $\mathscr{M}$, especially for prediction.

Our approach builds upon Bayesian decision analysis. First, we target the *predictive functionals* $h_1(\tilde{\boldsymbol{y}}), \ldots, h_J(\tilde{\boldsymbol{y}})$, where each h_j is a functional of interest and $\tilde{\boldsymbol{y}} \sim p_{\mathscr{M}}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y})$ is the predictive distribution of unobserved data at covariate value $\tilde{\boldsymbol{x}}$ and conditional on observed data. Each h_j reflects a prediction task: often the data $(\boldsymbol{x}, \boldsymbol{y})$ are an input to

a system h_j, which inherits predictive uncertainty when $\boldsymbol{y}$ has not yet been observed. Alternatively, the functionals $\{h_j\}$ can be selected to provide distinct summaries of the model $\mathscr{M}$. Next, we introduce a *parametrized action* $g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})$, which is a point prediction of $h(\tilde{\boldsymbol{y}})$ at $\tilde{\boldsymbol{x}}$ with unknown parameters $\boldsymbol{\delta}$. The role of g is to produce interpretable, fast, and accurate predictions targeted to h. Examples include linear, tree, and additive forms, but g is not required to match the structure of $\mathscr{M}$. The targeted predictions are not burdened by the complexity required to capture the global distributional features of $p_{\mathscr{M}}(\boldsymbol{\theta} \mid \boldsymbol{y})$ or $p_{\mathscr{M}}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y})$—which may be mostly irrelevant for predicting any particular $h_j(\tilde{\boldsymbol{y}})$—yet use the full posterior distribution under $\mathscr{M}$ to incorporate all available data. Lastly, we leverage the model $\mathscr{M}$ predictive distribution to quantify and compare the *out-of-sample* predictive accuracy of each parametrized action. Using this information, we assemble a collection of near-optimal, or *acceptable* targeted predictors, which offers unique insights into the predictability of $h_j(\tilde{\boldsymbol{y}})$.

For any functional $h_j = h$, predictive accuracy is measured by a loss function $\mathscr{L}_0\{h(\tilde{\boldsymbol{y}}), g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})\}$, which determines the loss from predicting $g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})$ when $h(\tilde{\boldsymbol{y}})$ is realized. To incorporate multiple covariate values $\tilde{\mathscr{X}} \coloneqq \{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$, we introduce an aggregate loss function

$$\overline{\mathscr{L}}_0\left[\{h(\tilde{\boldsymbol{y}}_i), g(\tilde{\boldsymbol{x}}_i; \boldsymbol{\delta})\}_{i=1}^{\tilde{n}}\right] \coloneqq \tilde{n}^{-1} \sum_{i=1}^{\tilde{n}} \mathscr{L}_0\{h(\tilde{\boldsymbol{y}}_i), g(\tilde{\boldsymbol{x}}_i; \boldsymbol{\delta})\},$$

where each $\tilde{\boldsymbol{y}}_i$ is the predictive variable at $\tilde{\boldsymbol{x}}_i$ under model $\mathscr{M}$. The choice of $\tilde{\mathscr{X}}$ can be distinct from the original covariates $\{\boldsymbol{x}_i\}_{i=1}^{n}$, for example to customize predictions for specific designs or subpopulations of interest, yet still leverages the full posterior distribution under model $\mathscr{M}$. We augment the aggregate loss with a complexity penalty $\mathscr{P}$ on the unknown parameters $\boldsymbol{\delta}$:

$$\overline{\mathscr{L}}_\lambda\left[\{h(\tilde{\boldsymbol{y}}_i), g(\tilde{\boldsymbol{x}}_i; \boldsymbol{\delta})\}_{i=1}^{\tilde{n}}\right] \coloneqq \overline{\mathscr{L}}_0\left[\{h(\tilde{\boldsymbol{y}}_i), g(\tilde{\boldsymbol{x}}_i; \boldsymbol{\delta})\}_{i=1}^{\tilde{n}}\right] + \lambda \mathscr{P}(\boldsymbol{\delta}),$$

where $\lambda \geq 0$ indexes a path of parameterized actions and determines the tradeoff between predictive accuracy ($\overline{\mathscr{L}}_0$) and complexity ($\mathscr{P}$).

Since $\overline{\mathscr{L}}_\lambda$ depends on a random quantities $\{\tilde{\boldsymbol{y}}_i\}_{i=1}^{\tilde{n}}$, Bayesian decision analysis proceeds by optimizing for $\boldsymbol{\delta}$ over the joint posterior predictive distribution $p_{\mathscr{M}}(\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_{\tilde{n}} \mid \boldsymbol{y})$:

$$\hat{\boldsymbol{\delta}}_{\mathscr{A}} \coloneqq \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \mathbb{E}_{[\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_{\tilde{n}} \mid \boldsymbol{y}]} \overline{\mathscr{L}}_\lambda\left[\{h(\tilde{\boldsymbol{y}}_i), g(\tilde{\boldsymbol{x}}_i; \boldsymbol{\delta})\}_{i=1}^{\tilde{n}}\right]. \quad (1)$$

This operation averages the predictive loss over the joint distribution of future or unobserved values $\{h(\tilde{\boldsymbol{y}}_i)\}_{i=1}^{n}$ at $\tilde{\mathscr{X}}$ under model $\mathscr{M}$, and then selects parameters $\hat{\boldsymbol{\delta}}_{\mathscr{A}}$ that minimize this quantity. We define the *parametrized action* $\mathscr{A} \coloneqq (g, \mathscr{P}, \lambda)$ as a triple consisting of the

targeted predictor g, the complexity penalty $\mathscr{P}$, and the complexity parameter λ. Since we typically compare among parametrized actions for the same functional h, design points $\widetilde{\mathscr{X}}$, and Bayesian model $\mathscr{M}$, we suppress notational dependence on these terms.

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

By design, the optimal parameters $\hat{\boldsymbol{\delta}}_{\mathscr{A}}$ depend on the loss function $\mathscr{L}_0$. Generality of $\mathscr{L}_0$ is desirable, but tractability is essential for practical use. A natural starting point is squared error loss $\mathscr{L}_0\{h(\tilde{\boldsymbol{y}}), g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})\} = \| h(\tilde{\boldsymbol{y}}) - g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta}) \|_2^2$ with generalizations considered below. In this setting, we identify a representation of the requisite optimization problem (1) that admits fast and interpretable solutions for a broad class of parametrized actions:

Theorem 1. *When* $\mathbb{E}_{[\tilde{\boldsymbol{y}}_i \mid \boldsymbol{y}]} \| h(\tilde{\boldsymbol{y}}_i) \|_2^2 < \infty$ *at each* $\tilde{\boldsymbol{x}}_i \in \mathscr{X}, i = 1, \ldots, \tilde{n}$, *the optimal point prediction parameters in* (1) *under squared error loss are*

$$\hat{\boldsymbol{\delta}}_{\mathscr{A}} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \left\{ \tilde{n}^{-1} \sum_{i=1}^{\tilde{n}} \| \bar{h}_i - g(\tilde{\boldsymbol{x}}_i; \boldsymbol{\delta}) \|_2^2 + \lambda \mathscr{P}(\boldsymbol{\delta}) \right\} \tag{2}$$

where $\bar{h}_i := \mathbb{E}_{[\tilde{\boldsymbol{y}}_i \mid \boldsymbol{y}]} h(\tilde{\boldsymbol{y}}_i)$ *is the posterior predictive expectation of* $h(\tilde{\boldsymbol{y}}_i)$ *at* $\tilde{\boldsymbol{x}}_i$ *under model* $\mathscr{M}$.

Theorem 1 establishes an equivalence between the solution to the posterior predictive expected loss (1) and a penalized least squares criterion, with important computational implications. First, estimation of $\bar{h}_i$ is a standard Bayesian exercise, for example using posterior predictive samples: $\bar{h}_i \approx S^{-1} \sum_{s=1}^S h(\tilde{\boldsymbol{y}}_i^s)$ for $\tilde{\boldsymbol{y}}_i^s \sim p_{\mathscr{M}}(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{y})$ at $\tilde{\boldsymbol{x}}_i$. Most commonly, posterior predictive samples are generated by iteratively drawing $\boldsymbol{\theta}^s \sim p_{\mathscr{M}}(\boldsymbol{\theta} \mid \boldsymbol{y})$ from the posterior and $\tilde{\boldsymbol{y}}_i^s \sim p_{\mathscr{M}}(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{\theta}^s)$ from the sampling distribution. Second, the penalized least squares representation in (2) implies that the optimal point prediction parameters $\hat{\boldsymbol{\delta}}_{\mathscr{A}}$ can be computed easily and efficiently for many choices of $\mathscr{A}$ using existing algorithms and software. Third, the optimal parametrized actions produce fast out-of-sample targeted predictions: the prediction of $h(\tilde{\boldsymbol{y}})$ at *any* $\tilde{\boldsymbol{x}}$ is $g(\tilde{\boldsymbol{x}}; \hat{\boldsymbol{\delta}}_{\mathscr{A}})$, which is quick to compute for many choices of g. Lastly, the optimal parameters from (2) can be computed simultaneously for many parametrized actions $\mathscr{A}$ and distinct functionals h—all based on a single Bayesian model $\mathscr{M}$.

Remark. *Certain choices of h, such as binary functionals* $h(\tilde{\boldsymbol{y}}) \in \{0, 1\}$, *are incompatible with squared error loss. In the* supplementary material, *we discuss generalizations to*

deviance-based loss functions. Importantly, the core attributes of the proposed approach are maintained: computational speed, ease of implementation, and interpretability.

We illustrate the utility of this framework with the following examples; an additional example with $h(\tilde{\boldsymbol{y}}) \in \{0, 1\}$ is presented in the supplementary material.

Example 1 (Linear contrasts). Consider a (multivariate) regression model $\mathbb{E}_{[\boldsymbol{y}_i \mid \boldsymbol{\theta}]}\boldsymbol{y}_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ for $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,m})'$. The linear contrast $h(\tilde{\boldsymbol{y}}) = \boldsymbol{C}\tilde{\boldsymbol{y}}$ is often of interest: the matrix $\boldsymbol{C}$ can extract specific components of $\tilde{\boldsymbol{y}}$, evaluate differences between components of $\tilde{\boldsymbol{y}}$, and apply a linear weighting scheme to $\tilde{\boldsymbol{y}}$. For functional data with $y_{i,j} = y_i(\tau_j)$, the linear contrast can target subdomains $\boldsymbol{C}\tilde{\boldsymbol{y}} = \{\tilde{y}(\tau)\}_{\tau \in \mathcal{S}}$ for $\mathcal{S} \subset \mathcal{T}$ and evaluate derivatives of $\tilde{y}(\tau)$. In this setting, the predictive target simplifies to the posterior expectation $\bar{h} = \mathbb{E}_{[\boldsymbol{\theta} \mid \boldsymbol{y}]}\{\boldsymbol{C}f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}})\} = \boldsymbol{C}\mathbb{E}_{[\boldsymbol{\theta} \mid \boldsymbol{y}]}f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}})$. Given an estimate $\hat{f}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}})$ of the posterior expectation of the regression function at $\tilde{\boldsymbol{x}}$, the response variable $\bar{h}_i \approx \boldsymbol{C}\hat{f}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_i)$ needed for (2) is easily computable for many choices of $\boldsymbol{C}$. Notably, the predictive expected contrast $\boldsymbol{C}\hat{f}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ is distinct from the empirical contrast $h(\boldsymbol{y}_i) = \boldsymbol{C}\boldsymbol{y}_i$: the former can incorporate shrinkage, smoothness, and other regularization of the regression function $f_{\boldsymbol{\theta}}$ under $\mathcal{M}$. From a single Bayesian model $\mathcal{M}$, multiple parametrized actions $\mathcal{A}$ can be optimized for each contrast $\boldsymbol{C}$.

Example 2 (Functional data summaries). Suppose h is a scalar summary of a curve $\{y(\tau)\}_{\tau \in \mathcal{T}}$, such as the maximum $h(\tilde{\boldsymbol{y}}) = \max_\tau \tilde{y}(\tau)$ or the point at which the maximum occurs $h(\tilde{\boldsymbol{y}}) = \arg\max_\tau \tilde{y}(\tau)$, and let $\mathcal{M}$ be a Bayesian functional data model (Section 5 provides a detailed example). To select variables for optimal linear prediction of $h(\tilde{\boldsymbol{y}})$, we apply Theorem 1 with $g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta}) = \tilde{\boldsymbol{x}}'\boldsymbol{\delta}$ and an ℓ_1-penalty, $\mathcal{P}(\boldsymbol{\delta}) = \|\boldsymbol{\delta}\|_1 = \sum_{j=1}^{p} |\delta_j|$:

$$\hat{\boldsymbol{\delta}}_{\mathcal{A}} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \left\{ \tilde{n}^{-1} \sum_{i=1}^{\tilde{n}} \|\bar{h}_i - \tilde{\boldsymbol{x}}_i'\boldsymbol{\delta}\|_2^2 + \lambda \|\boldsymbol{\delta}\|_1 \right\}, \tag{3}$$

for example using the observed covariates $\tilde{\mathcal{X}} = \{\boldsymbol{x}_i\}_{i=1}^{n}$. The optimal parameters $\hat{\boldsymbol{\delta}}_{\mathcal{A}}$ are readily computable using existing software, such as glmnet in R (Friedman et al., 2010).

In practice, we apply an adaptive variant of the ℓ_1-penalty. Motivated by the adaptive lasso (Zou, 2006), Kowal et al. (2020) introduce the penalty $\mathcal{P}(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_{j=1}^{p} \omega_j |\delta_j|$, where $\boldsymbol{\omega}_j = |\beta_j|^{-1}$ and β_j are the regression coefficients in a Gaussian linear model $\mathcal{M}$. For nonlinear or non-Gaussian models $\mathcal{M}$ and targeted predictions, we use the generalized weights $\boldsymbol{\omega} = |\tilde{\boldsymbol{\delta}}_0|^{-1}$, where $\tilde{\boldsymbol{\delta}}_0$ is the ℓ_2-projection of the predictive variables $h(\tilde{\boldsymbol{y}}_i)$ onto the predictor g. Bayesian decision analysis requires integration over the unknown $\boldsymbol{\theta}$, so the requisite penalty in (2) becomes the posterior expectation $\overline{\mathcal{P}(\boldsymbol{\delta})} := \mathbb{E}_{[\tilde{\boldsymbol{y}} \mid \boldsymbol{y}]}\mathcal{P}(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_{j=1}^{p} \hat{\omega}_j |\delta_j|$ for $\hat{\boldsymbol{\omega}} = \mathbb{E}_{[\tilde{\boldsymbol{y}} \mid \boldsymbol{y}]}(|\tilde{\boldsymbol{\delta}}_0|^{-1})$, which is estimable using posterior predictive samples.

The parameterized and targeted decision analysis from (1) features connections with classical decision analysis. Targeted prediction arises in classical decision analysis through the Bayes estimator $\bar{h}_i = \mathbb{E}_{[\tilde{\boldsymbol{y}}_i \mid \boldsymbol{y}]} h(\tilde{\boldsymbol{y}}_i)$, which is obtained from Theorem 1 as a special case:

Corollary 1. *Let* $\mathscr{A}_B = (g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta}) = \delta(\tilde{\boldsymbol{x}}), \lambda = 0)$ *denote an unrestricted and unpenalized action. The optimal point predictor parameters are* $\hat{\delta}(\tilde{\boldsymbol{x}}_i) = \bar{h}_i$.

However, action parametrization and penalization are valuable tools: they lend interpretability to the targeted prediction, highlight the balance between accuracy and simplicity, and often produce faster—and more accurate—out-of-sample predictions via $g(\tilde{\boldsymbol{x}}; \hat{\boldsymbol{\delta}}_{\mathscr{A}})$.

In some cases, the optimal actions $\hat{\boldsymbol{\delta}}_{\mathscr{A}}$ can be linked to the underlying model parameters $\boldsymbol{\theta}$, such as when the parameterization $\mathscr{A}$ matches the form of $\mathscr{M}$ and both are linear:

Corollary 2. *Let* $\mathscr{A}_L = (g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta}) = \tilde{\boldsymbol{x}}'\boldsymbol{\delta}, \lambda = 0)$ *denote a linear and unpenalized action. For a model* $\mathscr{M}$ *with* $\mathbb{E}_{[\tilde{\boldsymbol{y}}_i \mid \boldsymbol{\theta}]} h(\tilde{\boldsymbol{y}}_i) = \tilde{\boldsymbol{x}}_i'\boldsymbol{\theta}$ *and using the observed design points* $\tilde{\mathscr{X}} = \{\boldsymbol{x}_i\}_{i=1}^n$, *the optimal point predictor parameters are* $\hat{\boldsymbol{\delta}}_{\mathscr{A}_L} = \mathbb{E}_{\boldsymbol{\theta} \mid \boldsymbol{y}}\boldsymbol{\theta}$.

Corollary 2 is most familiar when $\mathscr{M}$ is a linear model and h is the identity. By further allowing $\lambda > 0$ with a sparsity penalty $\mathscr{P}$, we recover the *decoupling shrinkage and selection* approach for Bayesian linear variable selection (Hahn and Carvalho, 2015). Similar links to Woody et al. (2020) can be established for nonlinear regression.

Despite the potential connections to $\boldsymbol{\theta}$ in certain cases, the parametrized actions are not bound by the parametrization of model $\mathscr{M}$. The full benefits of Theorem 1 are realized by the simultaneous generality of the model $\mathscr{M}$, the functionals h, and the parametrized actions $\mathscr{A}$. Of course, we can shift the emphasis from prediction toward posterior summarization by replacing the predictive functional $h(\tilde{\boldsymbol{y}})$ with a posterior functional $h(\boldsymbol{\theta})$, such as $h(\boldsymbol{\theta}) = h\left(\mathbb{E}_{[\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}]}\tilde{\boldsymbol{y}}\right)$. However, we prefer the predictive functionals: they correspond to concrete observables that are comparable across Bayesian models (Geisser, 1993).

3 Predictive inference for model determination

Decision analysis extracts an optimal $\hat{\boldsymbol{\delta}}_{\mathscr{A}}$ by minimizing a posterior (predictive) expected loss function. However, this optimality is obtained only for a *given* parametrized action $\mathscr{A}$. The key implication of Theorem 1 is that optimal point predictions can be computed easily and efficiently for *many* $\mathscr{A}$ (see Figure 2). To fully exploit these benefits, additional tools are needed to evaluate, compare, and select among the parametrized actions.

We proceed to evaluate predictive performance out-of-sample, which best encapsulates the task of predicting new data. The Bayesian model $\mathscr{M}$ provides predictive uncertainty quantification for all evaluations and comparisons. These out-of-sample predictive comparisons serve to identify not only the best targeted predictor, but also those targeted

predictors that achieve an acceptable level of accuracy for out-of-sample prediction. The collection of acceptable targeted predictors illuminates the shared characteristics of near-optimal models, such as the important covariates, the forms of g and $\mathscr{P}$, and the level of complexity needed for accurate prediction of $h(\tilde{\boldsymbol{y}})$. This approach only requires a Bayesian model $\mathscr{M}$, 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 targeted predictor. We proceed nominally using $g(\tilde{\boldsymbol{x}}; \hat{\boldsymbol{\delta}}_{\mathscr{A}})$, but note that any point predictor of $h(\tilde{\boldsymbol{y}})$ at $\tilde{\boldsymbol{x}}$ can be used. Let $L(\boldsymbol{z}, \hat{\boldsymbol{z}})$ denote the loss associated with a prediction $\hat{\boldsymbol{z}}$ when $\boldsymbol{z}$ has occurred. We consider both *empirical* and *predictive* versions of the loss: the former uses empirical functionals $\boldsymbol{z} = h(\boldsymbol{y})$ and relies exclusively on the observed data, while the latter uses predictive functionals $\boldsymbol{z} = h(\tilde{\boldsymbol{y}})$ and inherits a predictive distribution under $\mathscr{M}$.

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

For each data split k, the out-of-sample *empirical* and *predictive* losses are

$$\begin{aligned} \mathbb{L}_{\mathscr{A}}^{out}(k) &:= \frac{1}{|\mathscr{I}_k|} \sum_{i \in \mathscr{I}_k} L\left\{ h(\boldsymbol{y}_i), g(\boldsymbol{x}_i; \hat{\boldsymbol{\delta}}_{\mathscr{A}}^{-\mathscr{I}_k}) \right\}, \;\; \mathbb{L}_{\mathscr{A}}^{out}(k) \\ &:= \frac{1}{|\mathscr{I}_k|} \sum_{i \in \mathscr{I}_k} L\left\{ h(\tilde{\boldsymbol{y}}_i^{-\mathscr{I}_k}), g(\boldsymbol{x}_i; \hat{\boldsymbol{\delta}}_{\mathscr{A}}^{-\mathscr{I}_k}) \right\} \end{aligned} \tag{4}$$

respectively, where $\hat{\boldsymbol{\delta}}_{\mathscr{A}}^{-\mathscr{I}_k} := \arg\min_{\boldsymbol{\delta}} \mathbb{E}_{[\tilde{\boldsymbol{y}} \mid \boldsymbol{y}^{-\mathscr{I}_k}]} \overline{\mathscr{L}}_{\lambda}\left[\{h(\tilde{\boldsymbol{y}}_i), g(\tilde{\boldsymbol{x}}_i; \boldsymbol{\delta})\}_{i \notin \mathscr{I}_k} \right]$ is optimized only using the training data $\boldsymbol{y}^{-\mathscr{I}_k} := \{\boldsymbol{y}_i\}_{i \notin \mathscr{I}_k}$, and similarly $\tilde{\boldsymbol{y}}_i^{-\mathscr{I}_k} \sim p_{\mathscr{M}}(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{y}^{-\mathscr{I}_k})$ is the predictive variate at $\boldsymbol{x}_i$ conditional only on the training data. Although in-sample versions are available, there is an important distinction between the *out-of-sample* predictive distribution, $p_{\mathscr{M}}(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{y}^{-\mathscr{I}_k})$, and the *in-sample* predictive distribution, $p_{\mathscr{M}}(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{y})$. The in-sample version conditions on both the training data $\boldsymbol{y}^{-\mathscr{I}_k}$ and the validation data $\boldsymbol{y}^{\mathscr{I}_k} := \{\boldsymbol{y}_i\}_{i \in \mathscr{I}_k}$, which overstates the accuracy and understates the uncertainty for a validation point $i \in \mathscr{I}_k$. The out-of-sample version avoids these issues and more closely resembles most practical prediction problems.

Evaluation of $\mathscr{A}$ is based on the averages of (4) across all data splits,

$$\mathbb{L}_{\mathscr{A}}^{out} := K^{-1} \sum_{k=1}^{K} \mathbb{L}_{\mathscr{A}}^{out}(k), \;\; \mathbb{L}_{\mathscr{A}}^{out} := K^{-1} \sum_{k=1}^{K} \mathbb{L}_{\mathscr{A}}^{out}(k).$$

The K-fold aggregation averages over two sources of variability in (4): variability in the training sets $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i \notin \mathscr{I}_k}$, each of which results in a distinct estimate of the coefficients $\hat{\boldsymbol{\delta}}_{\mathscr{A}}^{-\mathscr{I}_k}$, and variability in the validation sets $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i \in \mathscr{I}_k}$, which evaluates predictions only at the validation design points $\{\boldsymbol{x}_i\}_{i \in \mathscr{I}_k}$. The contrast between $\mathbb{L}_{\mathscr{A}}^{out}$ and $\mathbb{L}_{\mathscr{A}}^{out}$ is important: $\mathbb{L}_{\mathscr{A}}^{out}$ is a point estimate of the risk under predictions from $\mathscr{A}$, while $\mathbb{L}_{\mathscr{A}}^{out}$ provides the distribution of out-of-sample loss under different *realizations* of the predictive variables $h(\tilde{\boldsymbol{y}}_i)$. Specifically, each $h(\boldsymbol{y}_i)$ for $i \in \mathscr{I}_k$ represents one possible realization of the out-of-sample target variable at $\boldsymbol{x}_i$; the predictive variable $h(\tilde{\boldsymbol{y}}_i^{-\mathscr{I}_k})$ for $\tilde{\boldsymbol{y}}_i^{-\mathscr{I}_k} \sim p_{\mathscr{M}}(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{y}^{-\mathscr{I}_k})$ expresses the distribution of possible realizations according to $\mathscr{M}$. The predictive loss $\mathbb{L}_{\mathscr{A}}^{out}$ incorporates this distributional information for out-of-sample predictive uncertainty quantification.

3.2 Predictive model selection

The out-of-sample empirical and predictive losses, $\mathbb{L}_{\mathscr{A}}^{out}$ and $\mathbb{L}_{\mathscr{A}}^{out}$, respectively, provide the ingredients needed to compare and select among targeted predictors. 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 the most accurate predictor, but also to gather those targeted predictors 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 targeted predictor, $\mathscr{A}_{min} := \arg\min_{\mathscr{A} \in \mathbb{A}} \mathbb{L}_{\mathscr{A}}^{out}$, which minimizes out-of-sample empirical loss as in classical K-fold cross-validation. The set $\mathbb{A}$ may include different forms for g and $\mathscr{P}$ and usually will include a path of λ values for each $(g, \mathscr{P})$ pair. We prefer relative rather than absolute accuracy because it directly references an empirically attainable accuracy level.

For any two actions $\mathscr{A}, \mathscr{A}' \in \mathbb{A}$, let $\mathbb{D}_{\mathscr{A}, \mathscr{A}'}^{out} := 100 \times (\mathbb{L}_{\mathscr{A}}^{out} - \mathbb{L}_{\mathscr{A}'}^{out}) \;/\; \mathbb{L}_{\mathscr{A}'}^{out}$ be the percent increase in out-of-sample predictive loss from $\mathscr{A}'$ to $\mathscr{A}$. We seek all parametrized actions $\mathscr{A}$ that perform within a margin $\eta \geq 0\%$ of the best model, $\mathbb{D}_{\mathscr{A}, \mathscr{A}_{min}}^{out} < \eta\%$, with probability at least $\varepsilon \in [0,1]$. The margin η acknowledges that near-optimal performance—especially for simple models—is often sufficient, while the probability level $\mathcal{E}$ incorporates predictive uncertainty. In concert, η and $\mathcal{E}$ provide domain-specific and model-informed leniency for admission

into a set of acceptable predictors. We formally define the set of *acceptable predictors* as follows:

Definition 1. *The set of acceptable predictors is*

$$\Lambda_{\eta,\varepsilon} := \left\{ \mathscr{A} \in \mathbb{A} : \mathbb{P}_{\mathscr{M}}(\mathbb{D}^{out}_{\mathscr{A},\mathscr{A}_{min}} < \eta) \geq \varepsilon \right\}, \text{ where } \mathbb{D}^{out}_{\mathscr{A},\mathscr{A}_{min}} := 100 \times (\mathbb{L}^{out}_{\mathscr{A}} - \mathbb{L}^{out}_{\mathscr{A}_{min}}) \; / \; \mathbb{L}^{out}_{\mathscr{A}_{min}}.$$

The probability $\mathbb{P}_{\mathscr{M}}$ is estimated using out-of-sample predictive draws under model $\mathscr{M}$ (see Section 3.3). Each set $\Lambda_{\eta,\varepsilon}$ is nonempty, since $\mathscr{A}_{min} \in \Lambda_{\eta,\varepsilon}$ for all η, ε, and nested: $\Lambda_{\eta,\varepsilon} \subseteq \Lambda_{\eta',\varepsilon'}$, for any $\eta' \geq \eta$ or $\varepsilon' \leq \varepsilon$, so increasing η or decreasing ε can expand the set of acceptable predictors. 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* of predictors for which the in-sample empirical loss is within a margin η of the best predictor. By comparison, $\Lambda_{\eta,\varepsilon}$ uses out-of-sample criteria for evaluation and incorporates predictive uncertainty via the Bayesian model $\mathscr{M}$.

The set of acceptable predictors also can be constructed using prediction intervals:

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

Viewed another way, $\mathscr{A}$ is *not* acceptable if the lower $1 - \varepsilon$ predictive interval for $\mathbb{D}^{out}_{\mathscr{A},\mathscr{A}_{min}}$ excludes η. From this perspective, *unacceptable* predictors are those $\mathscr{A}$ for which there is insufficient predictive probability (under $\mathscr{M}$) that the out-of-sample accuracy of $\mathscr{A}$ is within a certain margin of the best predictor. This definition is similar to the confidence sets of Lei (2019), which exclude any $\mathscr{A}$ for which the null hypothesis that $\mathscr{A}$ 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 predictor sets are derived entirely from the predictive distribution of $\mathscr{M}$ and accompanied by fast and accurate approximation algorithms (see Section 3.3).

Among acceptable predictors, we highlight the simplest one. For fixed $(g, \mathscr{P})$, the simplest predictor has the largest complexity penalty: $\lambda_{\eta,\varepsilon} := \max\{\lambda : (g, \mathscr{P}, \lambda) \in \Lambda_{\eta,\varepsilon}\}$. When $\mathscr{P}$ is a sparsity penalty such as (3), the simplest acceptable predictor contains the smallest set of covariates needed to (nearly) match the predictive accuracy of the best predictor—which may itself be $\mathscr{A}_{min}$. Selection based on $\lambda_{\eta,\varepsilon}$ resembles the *one-standard-error rule* (e.g., Hastie et al., 2009), which selects the simplest predictor for which the out-of-sample empirical loss is within one standard error of the best predictor. Instead, $\lambda_{\eta,\varepsilon}$ uses the out-of-sample predictive loss with posterior uncertainty quantification inherited from $\mathscr{M}$.

3.3 Fast approximations for out-of-sample predictive evaluation

The primary hurdle for out-of-sample predictive evaluations is computational: they require computing $\hat{\boldsymbol{\delta}}_{\mathscr{A}}^{-\mathscr{I}_k}$ and sampling $\tilde{\boldsymbol{y}}_i^{-\mathscr{I}_k} \sim p_{\mathscr{M}}(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{y}^{-\mathscr{I}_k})$ for each data split $k = 1,\ldots, K$. Re-fitting $\mathscr{M}$ on each training set $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i \notin \mathscr{I}_k}$ 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 $\mathscr{M}$ 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 squared error loss, but adaptations to other loss functions are straightforward.

To obtain $\hat{\boldsymbol{\delta}}_{\mathscr{A}}^{-\mathscr{I}_k}$, we equivalently represent the optimal action as in Theorem 1:

$$\hat{\boldsymbol{\delta}}_{\mathscr{A}}^{-\mathscr{I}_k} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \left\{ (n - \mid \mathscr{I}_k \mid)^{-1} \sum_{j \notin \mathscr{I}_k} \| \bar{h}_j^{-\mathscr{I}_k} - g(\boldsymbol{x}_j; \boldsymbol{\delta}) \|_2^2 + \lambda \mathscr{P}(\boldsymbol{\delta}) \right\} \tag{5}$$

where $\bar{h}_j^{-\mathscr{I}_k} = \mathbb{E}_{[\tilde{\boldsymbol{y}}_j \mid \boldsymbol{y}^{-\mathscr{I}_k}]} h(\tilde{\boldsymbol{y}}_j)$ is the out-of-sample point prediction at $\boldsymbol{x}_j$. As such, (5) is easily solvable for many choices of $\mathscr{A}$: all that is required is an estimate of $\bar{h}_j^{-\mathscr{I}_k}$ for each $j \notin \mathscr{I}_k$ in the training set. We estimate this quantity using importance sampling. Proposals $\{\tilde{\boldsymbol{y}}_j^s\}_{s=1}^S \sim p_{\mathscr{M}}(\tilde{\boldsymbol{y}}_j \mid \boldsymbol{y})$ are generated from the full predictive distribution by sampling $\{\boldsymbol{\theta}^s\}_{s=1}^S \sim p_{\mathscr{M}}(\boldsymbol{\theta} \mid \boldsymbol{y})$ from the full posterior and $\{\tilde{\boldsymbol{y}}_j^s\}_{s=1}^S \sim p_{\mathscr{M}}(\tilde{\boldsymbol{y}}_j \mid \boldsymbol{\theta}^s)$ from the likelihood. The full data posterior $p_{\mathscr{M}}(\boldsymbol{\theta} \mid \boldsymbol{y})$ serves as a proposal for the training data posterior $p_{\mathscr{M}}(\boldsymbol{\theta} \mid \boldsymbol{y}^{-\mathscr{I}_k})$ with importance weights $w_k^s \propto 1 \,/\, p(\boldsymbol{y}^{\mathscr{I}_k} \mid \boldsymbol{\theta}^s)$, with further factorization under conditional independence. The target can be estimated using $\bar{h}_j^{-\mathscr{I}_k} \approx \sum_{s=1}^S w_k^s h(\tilde{\boldsymbol{y}}_j^s)$ or based on SIR sampling. In some cases, it is beneficial to regularize the importance weights (Ionides, 2008; Vehtari et al., 2015), but our empirical results remain unchanged with or without regularization. Successful variants of this approach exist for Bayesian model selection (Gelfand et al., 1992) and evaluating prediction distributions (Vehtari and Ojanen, 2012).

SIR provides a mechanism for sampling $\tilde{\boldsymbol{y}}_i^{-\mathscr{I}_k} \sim p_{\mathscr{M}}(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{y}^{-\mathscr{I}_k})$ using the importance weights $\{w_k^s\}_{s=1}^S$, which in turn provides out-of-sample predictive draws of $\mathbb{L}_{\mathscr{A}}^{out}$ and $\mathbb{D}_{\mathscr{A},\mathscr{A}'}^{out}$, for any actions $\mathscr{A}, \mathscr{A}' \in \mathbb{A}$. The idea is to obtain the proposal samples $\{\tilde{\boldsymbol{y}}_j^s\}_{s=1}^S \sim p_{\mathscr{M}}(\tilde{\boldsymbol{y}}_j \mid \boldsymbol{y})$ from the full posterior distribution and then subsample from $\{\tilde{\boldsymbol{y}}_j^s\}_{s=1}^S$ without replacement based on the corresponding importance weights $\{w_k^s\}_{s=1}^S$. 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. For targeted prediction, these evaluations must be directed toward a particular *functional* of the response variable. Specifically, we generate functional data $\{Y_i^*(\tau): \tau \in [0,1]\}$ such that the argmax of each function, $\tau_i^* := \arg\max_\tau Y_i^*(\tau) = h(Y_i^*)$, is linearly associated with a subset of covariates, $\tau_i^* = \boldsymbol{x}_i'\boldsymbol{\beta}^*$. The covariates are correlated and mixed continuous and discrete: we draw $x_{i,j}$ from marginal standard normal distributions with $\mathrm{Cor}(x_{i,j}, x_{i,j'}) = (0.75)^{|j-j'|}$ 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 standard deviation 0.5. For the true coefficients $\{\beta_j^*\}_{j=1}^p$, we randomly select 5% for $\beta_j^* = 1$, 5% for $\beta_j^* = -1$, and leave the remaining values at zero with the exception of the intercept, $\beta_0^* = 1$. The coefficients $\{\beta_j^*\}_{j=0}^p$ are rescaled such that $\tau_i^* = \boldsymbol{x}_i'\boldsymbol{\beta}^* \in [0.2, 0.8]$ to ensure that the argmax occurs away from the boundary; see the supplementary material. 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 $a_{0,i} \overset{iid}{\sim} N(0,1), a_{1,i}, a_{2,i} \overset{iid}{\sim} \chi_5^2$, and $(x)_+ := x\mathbb{I}\{x \geq 0\}$. By construction, Y_i^* is piecewise linear and concave with a single breakpoint, $\tau_i^* = \arg\max_\tau Y_i^*(\tau)$, and therefore $h(Y_i^*) = \boldsymbol{x}_i'\boldsymbol{\beta}^*$. Finally, the observed data $\boldsymbol{y}_i$ 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, $m = 200$ observation points, and varying sample sizes $n \in \{75, 100, 500\}$. For each simulated dataset $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$, 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 $\mathscr{M}$ does *not* reflect the true data-generating process, yet our targeted predictions are derived from the predictive distribution under $\mathscr{M}$. We consider linear actions $g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta}) = \tilde{\boldsymbol{x}}'\boldsymbol{\delta}$ with the adaptive ℓ_1-penalty from Example 2 and computed using glmnet in R (Friedman et al., 2010). In this case, the set of parametrized actions $\mathbb{A}$ is determined by the path of λ values, which control the sparsity of the linear action $\boldsymbol{\delta}$. For benchmark comparisons, we use the adaptive lasso (Zou, 2006) and projection predictive feature selection (Piironen et al., 2020) on the empirical functionals $\{\boldsymbol{x}_i, h(\boldsymbol{y}_i)\}_{i=1}^n$. Model sizes were selected using 10-fold cross-validation. Implementation of Piironen et al. (2020) uses the projpred package in R; for the requisite Bayesian linear model, we assume double exponential priors for the linear coefficients, but results are unchanged for Gaussian and t-priors.

To validate the proposed definition of acceptable predictor sets, we investigate a simple yet important question: does the true model belong to $\Lambda_{\eta,e}$? Specifically, we determine whether the true set of active variables $\{j: \beta_j^* \neq 0\}$ matches the set of active variables for *any*

acceptable predictor $\mathscr{A} \in \Lambda_{\eta,\varepsilon}$. 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 λ path and correctly evaluated by $\mathbb{D}^{out}_{\mathscr{A},\mathscr{A}'}$. Correct identification is only satisfied when *all* and *only* the true active variables $\{j: \beta^*_j \neq 0\}$ are nonzero according to $\mathscr{A}$.

For this task, we compute $\varepsilon_{max}(\eta) := \mathbb{P}_{\mathscr{M}}\left(\mathbb{D}^{out}_{\mathscr{A}^*,\mathscr{A}_{min}} < \eta\right)$, which is the maximum probability level for which the true model $\mathscr{A}^*$ is acceptable. The margin η corresponds to the percent increase in loss relative to $\mathscr{A}_{min}$. By design, $\mathscr{A}^* \in \Lambda_{\eta,\varepsilon'}$, remains acceptable for any smaller probability level $\varepsilon' \leq \varepsilon_{max}(\eta)$. Most important, we set $\varepsilon_{max}(\eta) = 0$ if $\mathscr{A}^*$ is not on the λ path. For each simulated dataset, we compute $\varepsilon_{max}(\eta)$ for a grid of η% values. The results averaged across 100 simulations are in Figure 3. Naturally, $\varepsilon_{max}(\eta)$ uniformly increases with the sample size for each value of η. When $\eta = 0$, the average maximum probability levels are $\varepsilon_{max}(0) \in \{0.21, 0.39, 0.54\}$ for $n \in \{75, 100, 500\}$, respectively, which suggests that a cutoff of $\varepsilon = 0.1$ is capable of capturing the true model even when zero margin is allowed. Notably, $\varepsilon_{max}(\eta)$ does *not* converge to one as η increases for the smaller sample sizes $n \in \{75, 100\}$. The reason is simple: if $\mathscr{A}^*$ is not discovered along the λ path, then $\varepsilon_{max}(\eta) = 0$ by definition—regardless of the choice of η. This result demonstrates the importance of the set of predictors *under consideration* $\mathbb{A}$, which here is determined entirely by the selected variables in the glmnet solution path.

Next, we evaluate point predictions of $h(Y^*_i)$ and estimates of $\boldsymbol{\beta}^*$ using root mean squared errors (RMSEs). The parametrized actions $\hat{\boldsymbol{\delta}}_\lambda$ and point predictions $g(\tilde{\boldsymbol{x}}; \hat{\boldsymbol{\delta}}_\lambda) = \tilde{\boldsymbol{x}}'\hat{\boldsymbol{\delta}}_\lambda$ are computed for multiple choices of λ: the simplest acceptable predictor $\lambda = \lambda_{\eta,\varepsilon}$ with $\eta = 0$ and $\varepsilon = 0.1$ (proposed(out)); the analogous choice of λ based on *in-sample* evaluations (proposed(in)); and the unpenalized linear action with $\lambda = 0$ (proposed(full)). For comparisons, we include the aforementioned adaptive lasso and projpred, the point predictions $\bar{h}_i$ under model $\mathscr{M}$ (h_bar; see Corollary 1), and the empirical functionals $h(\mathbf{y}_i)$ (h(y)). The results are in Figure 4. In summary, clear improvements in targeted prediction are obtained by (i) fitting to $h(\tilde{\mathbf{y}}_i)$ (via $\bar{h}_i$) rather than $h(\mathbf{y}_i)$, (ii) including covariate information, (iii) incorporating penalization or variable selection, and (iv) selecting the complexity λ based on out-of-sample criteria. The targeted actions $\mathscr{A}$ vastly outperform the model $\mathscr{M}$ predictions—even though each $\mathscr{A}$ is based entirely on the predictive distribution from $\mathscr{M}$. Lastly, the accurate estimation of the linear coefficients is important: the estimates $\hat{\boldsymbol{\delta}}_\lambda$ describe the partial linear effects of each x_j on targeted prediction of $h(\tilde{\mathbf{y}})$.

The supplementary material includes additional comparisons. Marginal variable selection is evaluated based on true positive and negative rates, with proposed(out) offering the best performance among these methods. Results for high dimensional data with $p > n$ ($n = 200$, $p = 500$ and $n = 100$, $p = 200$) confirm the prediction and estimation advantages of the proposed approach. Sensitivity to $\varepsilon \in \{0.05, 0.10, 0.20, 0.50\}$ is also studied for prediction, estimation, and selection. Lastly, we evaluate the robustness in predictive accuracy among these methods. Specifically, we consider the setting in which

the distribution of the covariates differs significantly between the training and validation datasets. The parametrized actions offer superior targeted predictions, especially for small to moderate sample sizes.

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 predictions that more accurately predict and explain the defining characteristics of PA. Specifically, we consider the following functionals $h(\tilde{\boldsymbol{y}})$ for intraday PA $\tilde{\boldsymbol{y}} = (\tilde{y}(\tau_1), \ldots, \tilde{y}(\tau_m))'$ at times-of-day $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_m$:

$$\int \tilde{y}(\tau) d\tau \quad \int \log\{\tilde{y}(\tau)+1\} d\tau \quad \| \tilde{y} - \int \tilde{y}(\tau) d\tau \|_{L^2} \quad \int \mathbb{I}\{\tilde{y}(\tau) \leq 100\} d\tau \quad \max_\tau \tilde{y}(\tau) \quad \arg\max_\tau \tilde{y}(\tau)$$
avg tlac sd sedentary max argmax

where avg captures average daily activity, tlac is the total log activity count and measures moderate activity (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: $\text{zeros(1am}-\text{5am)} := \mathbb{I}\{\tilde{y}(\tau) = 0 \text{ for all } \boldsymbol{\tau} \in [\text{1am,5am}]\}$. Individuals with zeros(1am-5am) = 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 zeros(1am-5am) = 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, cancer, and stroke), and lab measurements (total cholesterol, HDL cholesterol, systolic blood pressure). Data pre-processing generally follows 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 sample standard deviation 0.5.

In accordance with the schematic in Figure 2, targeted predictive decision analysis begins with a Bayesian model $\mathscr{M}$. 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 Kowal and Canale (2020). STAR formalizes the popular approach of *transforming* count data prior to applying 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 τ_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 $\mathscr{M}$ is the following:

$$y_{i,j}^{tot} = \text{round}(y_{i,j}^*), \quad z_{i,j}^* = \text{transform}(y_{i,j}^*) \tag{6}$$

$$z_{i,j}^* = \boldsymbol{b}'(\tau_j)\boldsymbol{\theta}_i + \sigma_\epsilon \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} t_v(0,1) \tag{7}$$

$$\theta_{i,\ell} = \boldsymbol{x}_i' \boldsymbol{\alpha}_\ell + \sigma_{\gamma_i} \gamma_{i,\ell}, \quad \gamma_{i,\ell} \overset{iid}{\sim} N(0,1) \tag{8}$$

with $\alpha_{\ell,j} \overset{indep}{\sim} N(0, \sigma_{\alpha_j}^2)$ and $\sigma_\epsilon^{-2}, \sigma_{\gamma_i}^{-2}, \sigma_{\alpha_j}^{-2} \overset{iid}{\sim} \text{Gamma}(0.01, 0.01)$. In (6), 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 $\text{round}(t) = \lfloor t \rfloor$ for $t > 0$ and $\text{round}(t) = 0$ for $t \leq 0$, so $y_{i,j}^{tot} = 0$ whenever $y_{i,j}^* < 0$, and set $\text{transform}(t) = 2(\sqrt{t} - 1)$ in the Box-Cox family. In the functional regression levels (7)-(8), $\boldsymbol{b}$ is a vector of spline basis functions with basis coefficients $\boldsymbol{\theta}_i$ for subject i and $\boldsymbol{\alpha}_\ell$ is the vector of regression coefficients for each basis coefficient ℓ. The spline basis is reparametrized to orthogonalize $\boldsymbol{b}$ and diagonalize the prior variance of the basis coefficients, which justifies the assumption of independence across basis coefficients in (8). Heavy-tailed innovations ($v = 3$) 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 (see the supplementary material) demonstrate adequacy of $\mathscr{M}$ for the functionals of interest. These results are insensitive to v, but alternative choices of transform (e.g., $\log t$) or $\boldsymbol{b}$ (e.g., wavelets) produce inferior results.

Targeted predictions for each functional were constructed using a linear action $g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta}) = \tilde{\boldsymbol{x}}'\boldsymbol{\delta}$ with an adaptive ℓ_1-penalty (see Example 2). Trees were also considered but were not competitive. The set of parametrized actions $\mathbb{A}$ is given by the path of λ values computed using glmnet in R (Friedman et al., 2010): we highlight the simplest acceptable action $\lambda = \lambda_{0,0.1}$ (proposed(out)) and the unpenalized linear action $\lambda = 0$ (proposed(full)). For comparison, we fit an adaptive lasso to $\{\boldsymbol{x}_i, h(\boldsymbol{y}_i)\}_{i=1}^n$ for each h. Squared error loss is used

for all but zeros(1am-5am) which uses cross-entropy. In the supplementary material, we consider quadratic effects for age and BMI and pairwise interactions for each of age and BMI with race, 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 complexity λ—which indexes the number of nonzero elements in $\hat{\boldsymbol{\delta}}_{\mathscr{A}}$ —Figure 5 presents the percent increase in predictive and empirical loss relative to the best predictor $\mathscr{A}_{min}$. The measures of vigorous PA (avg, sd, and max) produce nearly identical results, so we only include max here; avg and sd are in the supplement. The predictive expectations align closely with the empirical values, which suggests that model $\mathscr{M}$ is adequate for these predictive metrics.

For each functional, we obtain optimal or near-optimal predictions with only about 10 covariates with better accuracy than the adaptive lasso. Many of the selected covariates are shared among functionals: age, BMI, gender, race, HDL cholesterol, and CHD are selected for all but argmax, while smoking status (avg, sd, max), diabetes (avg, sd, sedentary, max), and total cholesterol (tlac, sedentary) appear as well. The functionals measuring vigorous PA agree on the selected variables, including negative effects for diabetes and smoking. Most distinct is argmax: while $\mathscr{A}_{min}$ includes 11 covariates, the predictive uncertainty quantification from $\mathbb{D}^{out}_{\mathscr{A},\mathscr{A}_{min}}$ indicates that linear predictors with as few as one covariate (race) are acceptable. These covariates are simply not linearly predictive of argmax: the difference between $\mathscr{A}_{min}$ and any other $\mathscr{A} \in \mathbb{A}$ is less than 1%. Details on the selected covariates and the direction of the estimated effects are provided in the supplement.

Robustness to the choice of η is also illustrated in Figure 5. We select $\eta = 0\%$ for max and argmax and $\eta = 1\%$ for tlac and sedentary, which highlights the purpose of η: by allowing $\eta > 0$, we can obtain targeted predictors with fewer covariates. By comparison, increasing the margin to $\eta = 1\%$ for max and argmax does not change the smallest acceptable predictor.

To validate the approximations in Figure 5, we augment the analysis with a truly out-of-sample prediction evaluation. For each of 20 training/validation splits, model $\mathscr{M}$ and the adaptive lasso are fit to the training data and sparse linear actions are targeted to each h. We emphasize that this exercise is computationally intensive: the MCMC for model $\mathscr{M}$ 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 5 compute in under two seconds.

Point predictions were generated for the validation data using $\bar{h}$ under $\mathscr{M}$ (h_bar), the adaptive lasso, and sparse linear actions with $\lambda = \lambda_{0,0.1}$ (proposed(out)), $\lambda = 0$ (proposed(full)), and $\mathscr{A}_{min}$. Since $\mathscr{A}_{min}$ is also the unique acceptable predictor when $\varepsilon = 1$, $\eta = 0$, it provides information about robustness to $\mathcal{E}$ and η. The point predictions under $\mathscr{M}$ are highly inaccurate—and so excluded from Figure 6—and slow to compute: we draw $\tilde{\boldsymbol{y}} \sim p_{\mathscr{M}}(\tilde{\boldsymbol{y}} \mid \boldsymbol{y})$ at each validation point $\tilde{\boldsymbol{x}}$ and then average $h(\tilde{\boldsymbol{y}})$ over these draws. The targeted actions simply evaluate $g(\tilde{\boldsymbol{x}}; \hat{\boldsymbol{\delta}}_{\mathscr{A}}) = \tilde{\boldsymbol{x}}' \hat{\boldsymbol{\delta}}_{\mathscr{A}}$, which is faster, simpler, less susceptible

to Monte Carlo error, and empirically more accurate. Predictions were evaluated on the empirical functionals $h(\mathbf{y}_i)$ in the validation data using mean squared prediction error.

The results from the out-of-sample prediction exercise are in Figure 6. The smallest acceptable predictor proposed(out) performs almost identical to the best predictor $\mathcal{A}_{min}$ despite using fewer covariates—which is precisely the goal of the acceptable predictor sets and the out-of-sample approximations in Figure 5. 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* $h(\mathbf{y}_i)$, which are used for training the adaptive lasso but *not* for the proposed methods. Instead, the parametrized actions are trained using $\bar{h}_i$—which is itself a poor out-of-sample predictor. However, the targeted actions only rely on the in-sample adequacy of $\bar{h}_i$ and, unlike models trained to the empirical functionals, leverage both the model-based regularization and the uncertainty quantification provided by $\mathcal{M}$. 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 the performance comparisons in Figure 6 confirm those in Figure 5, 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 an aggregate loss function 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 and provide 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 prediction, estimation, and 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. Using targeted prediction with sparse linear actions, 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.

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 among targeted predictors for each functional h, and is the vessel for all subsequent uncertainty quantification. Although it is practically impossible for $\mathcal{M}$ to be adequate for every functional, many well-designed models are capable of describing multiple functionals. We

only require that $\mathscr{M}$ provides a sufficiently accurate predictive distribution for each $h(\tilde{\boldsymbol{y}})$, which is empirically verifiable through standard posterior predictive diagnostics (Gelman et al., 1996). When the predictive distribution of $\mathscr{M}$ is intractable or computationally prohibitive, the proposed methods remain compatible with any approximation algorithm for $p_{\mathscr{M}}\{h(\tilde{\boldsymbol{y}}) \mid \boldsymbol{y}\}$.

Future work will establish uncertainty quantification for the optimal point prediction parameters $\hat{\boldsymbol{\delta}}_{\mathscr{A}}$. This task is nontrivial: frequentist uncertainty estimates for penalized regression are generally *not* valid, since the data have already been used to obtain the posterior (predictive) distribution under model $\mathscr{M}$. A promising alternative is to project the predictive targets $h(\tilde{\boldsymbol{y}})$ onto $g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})$, which induces a predictive distribution for the resulting parameter $\boldsymbol{\delta}$. Similar posterior projections have proven useful for linear variable selection (Woody et al., 2020) with growing theoretical justification (Patra and Dunson, 2018).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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

Bashir A, Carvalho CM, Hahn PR, and Jones MB (2019). Post-processing posteriors over precision matrices to produce sparse graph estimates. Bayesian Analysis, 14(4): 1075–1090.

Bernardo JM and Smith AFM (2009). Bayesian theory, volume 405. John Wiley & Sons.

Bjornstad JF (1990). Predictive likelihood: A review. Statistical Science, pages 242–254.

Breiman L (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science, 16(3):199–231.

Crawford L, Flaxman SR, Runcie DE, and West M (2019). Variable prioritization in nonlinear black box methods: A genetic association case study. Ann. Appl. Stat, 13(2):958–989. [PubMed: 32542104]

Fishman EI, Steeves JA, Zipunnikov V, Koster A, Berrigan D, Harris TA, and Murphy R (2016). Association between Objectively Measured Physical Activity and Mortality in Nhanes. Medicine and Science in Sports and Exercise, 48(7):1303–1311. [PubMed: 26848889]

Friedman J, Hastie T, and Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1–22. [PubMed: 20808728]

Geisser S (1993). Predictive inference, volume 55. CRC press.

Gelfand AE, Dey DK, and Chang H (1992). Model determination using predictive distributions, with implementation via sampling-based methods (with discussion). Bayesian Statistics 4, 4:147–167.

Gelman A, Meng XL, and Stern H (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6(4):733–807.

Goutis C and Robert CP (1998). Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. Biometrika, 85(1):29–37.

Gutiérrez-Peña E and Walker SG (2006). Statistical Decision Problems and Bayesian Nonparametric Methods. International Statistical Review, 73(3) :309–330.

Hahn PR and Carvalho CM (2015). Decoupling shrinkage and selection in bayesian linear models: A posterior summary perspective. Journal of the American Statistical Association, 110(509):435–448.

Hastie T, Tibshirani R, and Friedman J (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Huber F, Koop G, and Onorante L (2020). Inducing Sparsity and Shrinkage in Time-Varying Parameter Models. Journal of Business and Economic Statistics, pages 1–48.

Ionides EL (2008). Truncated importance sampling. Journal of Computational and Graphical Statistics, 17(2):295–311.

Kowal DR and Bourgeois DC (2020). Bayesian Function-on-Scalars Regression for High-Dimensional Data. Journal of Computational and Graphical Statistics, pages 1–10. [PubMed: 33013150]

Kowal DR, Bravo M, Leong H, Griffin RJ, Ensor KB, and Miranda ML (2020). Bayesian Variable Selection for Understanding Mixtures in Environmental Exposures.

Kowal DR and Canale A (2020). Simultaneous Transformation and Rounding (STAR) Models for Integer-Valued Data. Electronic Journal of Statistics, 14(1):1744–1772.

Lei J (2019). Cross-Validation With Confidence. Journal of the American Statistical Association, 0(0):1–53.

Leroux A, Di J, Smirnova E, Mcguffey EJ, Cao Q, Bayatmokhtari E, Tabacu L, Zipunnikov V, Urbanek JK, and Crainiceanu C (2019). Organizing and Analyzing the Activity Data in NHANES. Statistics in Biosciences, 11(2):262–287. [PubMed: 32047572]

Lindley DV (1968). The Choice of Variables in Multiple Regression. Journal of the Royal Statistical Society: Series B (Methodological), 30(1):31–53.

MacEachern SN (2001). Decision theoretic aspects of dependent nonparametric processes. In Bayesian Methods with Applications to Science, Policy, and Official Statistics, pages 551–560.

Nott DJ and Leng C (2010). Bayesian projection approaches to variable selection in generalized linear models. Computational Statistics and Data Analysis, 54(12):3227–3241.

Patra S and Dunson DB (2018). Constrained Bayesian Inference through Posterior Projections. arXiv preprint arXiv:1812.05741

Piironen J, Paasiniemi M, and Vehtari A (2020). Projective inference in high-dimensional problems: Prediction and feature selection. Electronic Journal of Statistics, 14(1):2155–2197.

Puelz D, Hahn PR, and Carvalho CM (2017). Variable selection in seemingly unrelated regressions with random predictors. Bayesian Analysis, 12(4):969–989.

Rao JNK (2011). Impact of frequentist and Bayesian methods on survey sampling practice: a selective appraisal. Statistical Science, 26(2):240–256.

Schmid D, Ricci C, and Leitzmann MF (2015). Associations of objectively assessed physical activity and sedentary time with all-cause mortality in US adults: The NHANES study. PLoS ONE, 10(3).

Semenova L and Rudin C (2019). A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. arXiv preprint arXiv: 1908.01755

Smirnova E, Leroux A, Cao Q, Tabacu L, Zipunnikov V, Crainiceanu C, and Urbanek JK (2019). The Predictive Performance of Objective Measures of Physical Activity Derived From Accelerometry Data for 5-Year All-Cause Mortality in Older Adults: National Health and Nutritional Examination Survey 2003–2006. The Journals of Gerontology: Series A.

Tran MN, Nott DJ, and Leng C (2012). The predictive Lasso. Statistics and Computing, 22(5):1069–1084.

Tulabandhula T and Rudin C (2013). Machine learning with operational costs. Journal of Machine Learning Research, 14(1):1989–2028.

Vehtari A and Ojanen J (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. Statistics Surveys, 6(1):142–228.

Vehtari A, Simpson D, Gelman A, Yao Y, and Gabry J (2015). Pareto Smoothed Importance Sampling. arXiv preprint arXiv:1507.02646

Wolff-Hughes DL, Bassett DR, and White T (2018). In response to: Re-evaluating the effect of age on physical activity over the lifespan. Preventive Medicine, 106:231–232. [PubMed: 29169827]

Woody S, Carvalho CM, and Murray JS (2020). Model interpretation through lower-dimensional posterior summarization. Journal of Computational and Graphical Statistics, pages 1–9. [PubMed: 33013150]

Zou H (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101 (476):1418–1429.

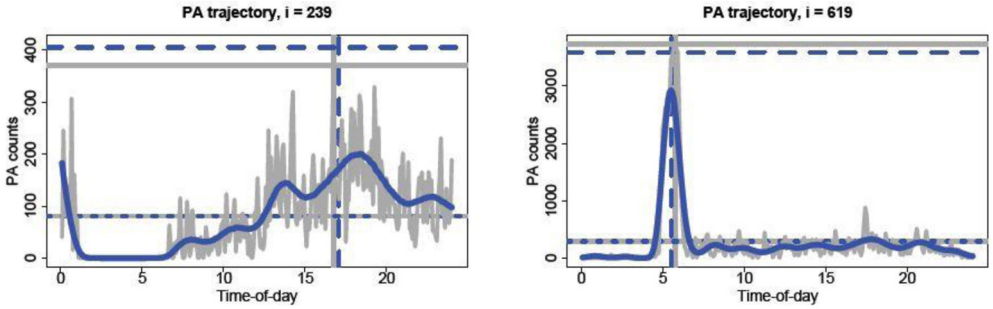

Fig. 1.

Intraday physical activity (gray line) and fitted values (blue line) for two subjects under model $\mathscr{M}$ in (6)-(8). The lines denote the empirical (solid gray) and predictive expected value (dashed blue) of avg (lower horizontal), max (upper horizontal), and argmax (vertical).

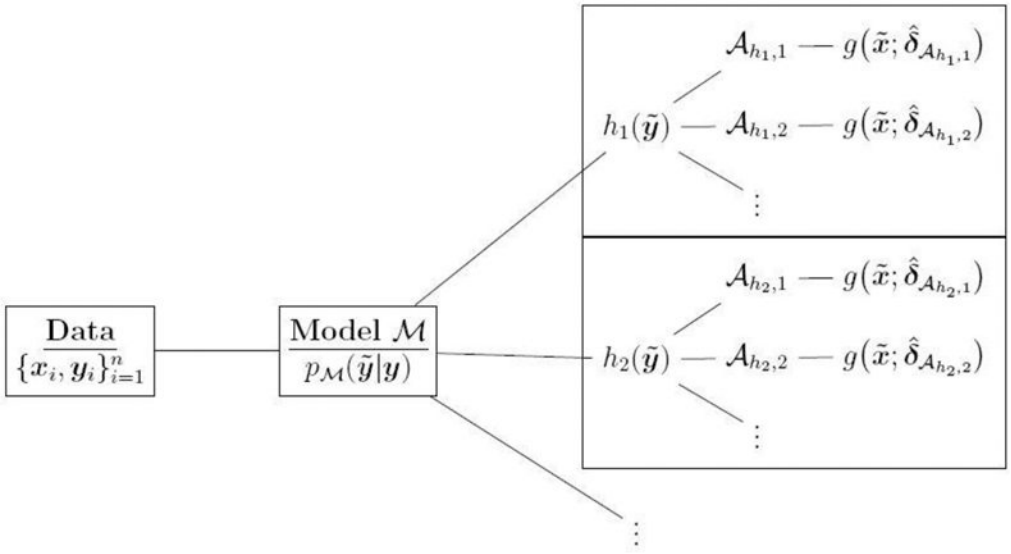

Fig. 2.

Given data $\{x_i, y_i\}_{i=1}^n$, a Bayesian model $\mathcal{M}$ is constructed. For each functional $h(\tilde{y})$ and using model $\mathcal{M}$, multiple parametrized actions $\mathcal{A}$ are optimized, evaluated, and compared. The optimal parameters $\hat{\delta}_{\mathcal{A}}$ are used to compute point predictions $g(\tilde{x}; \hat{\delta}_{\mathcal{A}})$ of $h(\tilde{y})$ at $\tilde{x}$.

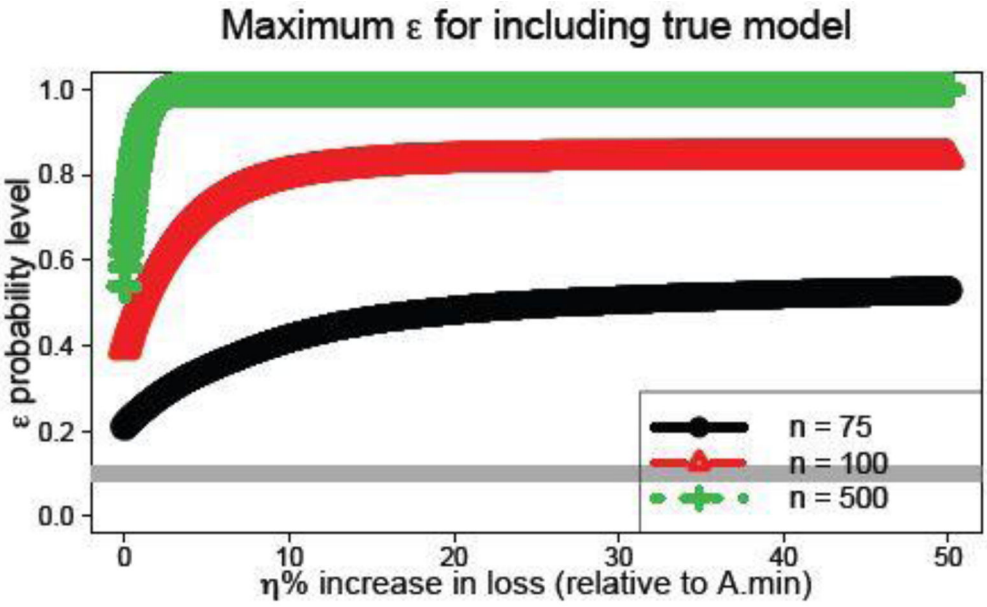

Fig. 3.

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

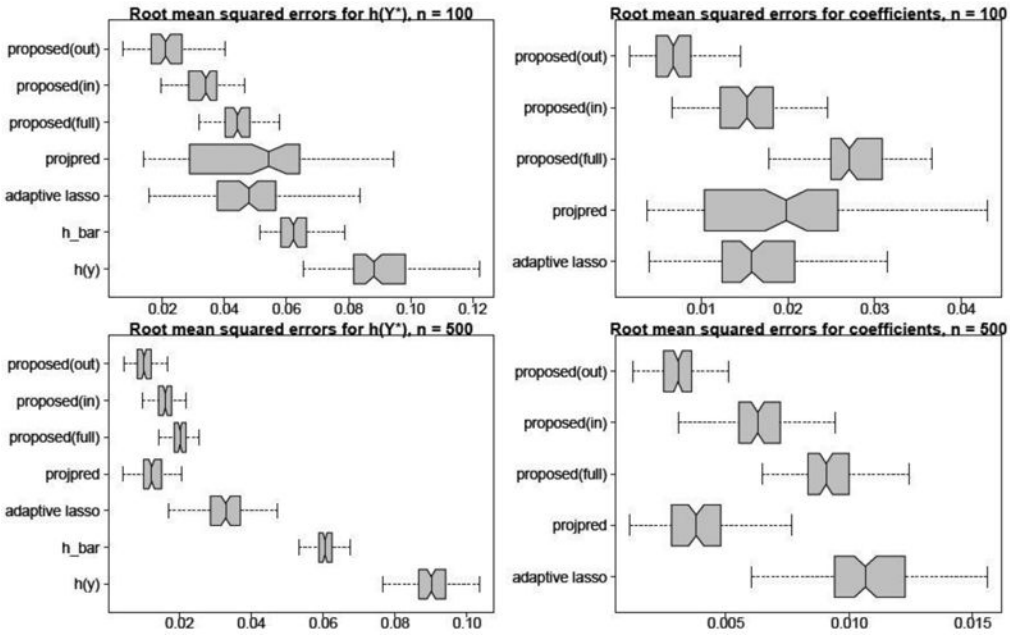

Fig. 4.

RMSEs for the true functionals $h(Y^*)$ (**left**) and the true regression coefficients β^* (**right**) for $n = 100$ (**top**) and $n = 500$ (**bottom**) across 100 simulated datasets. Non-overlapping notches indicate significant differences between medians. The parametrized actions with out-of-sample selection are most accurate for prediction and estimation.

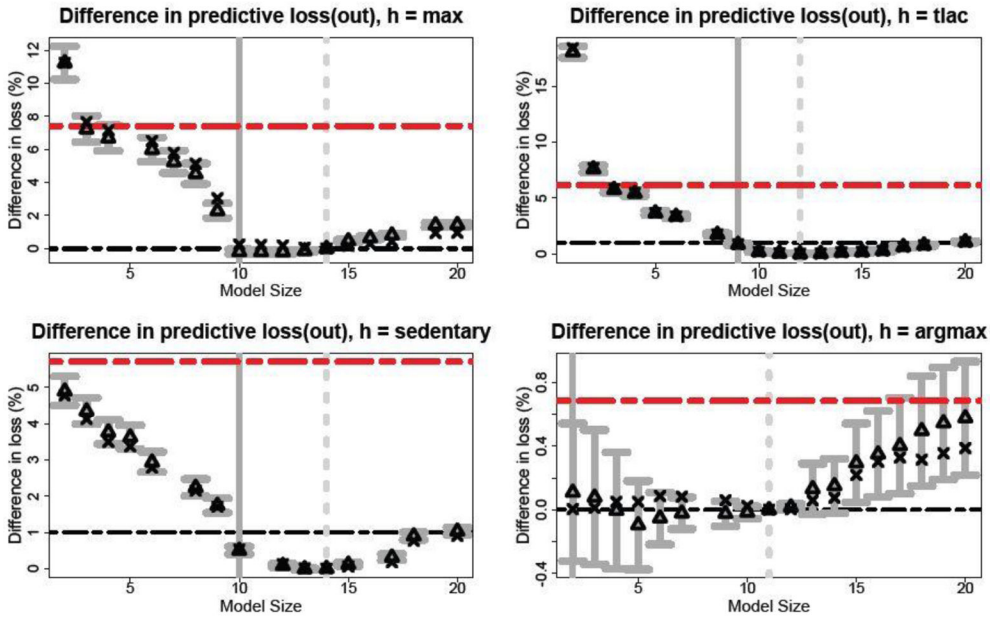

Fig. 5.

Approximate out-of-sample squared error loss for sparse linear actions targeted to each functional. Results are presented for each size as a percent increase in loss relative to $\mathcal{A}_{min}$. 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 η and the vertical lines denote $\lambda_{\eta,0.1}$ (solid) and $\mathcal{A}_{min}$ (dashed).

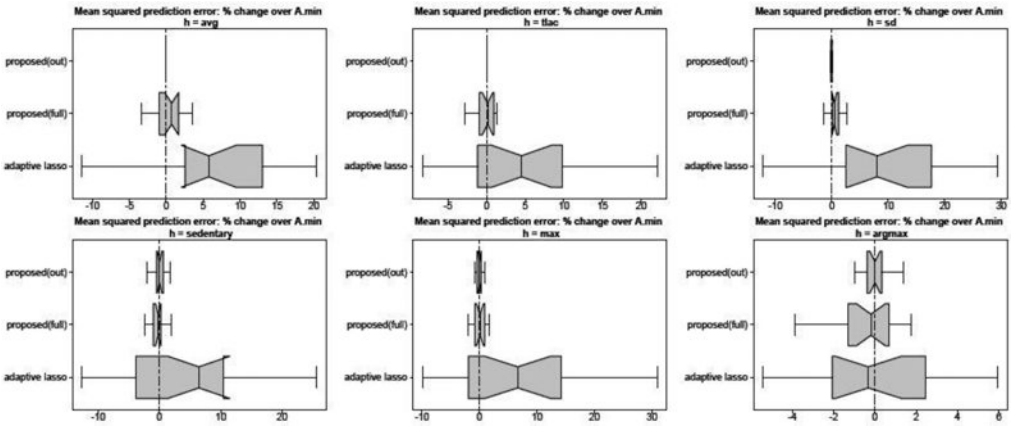

Fig. 6.

Mean squared prediction error for each functional across 20 training/validation splits. Results are presented as a percent increase relative to $\mathscr{A}_{min}$; values below zero (vertical line) indicate improvement over $\mathscr{A}_{min}$. Non-overlapping notches indicate significant differences between medians. Point predictions from $\mathscr{M}$ (h_bar) are noncompetitive and omitted. Both proposed(out) proposed(full) improve upon adaptive lasso and h_bar, while proposed(out) is most accurate and performs almost identical to $\mathscr{A}_{min}$ despite using fewer covariates.