Supplement to
"Fast, Optimal, and Targeted Predictions using Parametrized Decision Analysis"

Daniel R. Kowal*

February 11, 2021

This supplementary document contains the following: results for generalized loss functions (Section A); algorithms for out-of-sample approximations (Section B.1) and Gibbs sampling (Section B.2); additional results for the simulations (Section C) and physical activity data (Section D); and proofs of main results in the paper (Section E).

A Generalized loss functions

Although the flexibility in specifying h is undoubtedly a primary feature of the proposed framework, certain choices of h, such as binary functionals $h(\tilde{\boldsymbol{y}}) \in \{0, 1\}$, are incompatible with squared error loss. Generalized loss functions 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

*Dobelman Family Assistant Professor, Department of Statistics, Rice University, Houston, TX 77251-1892 (Daniel.Kowal@rice.edu).

negative log-likelihood of an exponential family distribution, or the *deviance loss*:

$$\mathcal{L}_0^{EF}\{h(\tilde{\boldsymbol{y}}), g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})\} := F_0\{g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})\} - T_0\{h(\tilde{\boldsymbol{y}})\} - \sum_{j=1}^{p} F_j\{g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})\} T_j\{h(\tilde{\boldsymbol{y}})\},$$

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 predictive expected loss under the Bayesian model $\mathcal{M}$:

$$\hat{\boldsymbol{\delta}}_{\mathcal{A}} := \arg\min_{\boldsymbol{\delta}} \mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]} \mathcal{L}_0^{EF}\{h(\tilde{\boldsymbol{y}}), g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})\}.$$

Key simplifications are available:

Theorem A.1. *When $\mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]} |T_0\{h(\tilde{\boldsymbol{y}})\}| < \infty$, the optimal point prediction parameters under the deviance loss $\mathcal{L}_0^{EF}$ are*

$$\hat{\boldsymbol{\delta}}_{\mathcal{A}} = \arg\min_{\boldsymbol{\delta}} \big[F_0\{g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})\} - \sum_{j=1}^{p} F_j\{g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})\}\overline{T_j}\big],$$

where $\overline{T_j} := \mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]} T_j\{h(\tilde{\boldsymbol{y}})\}$ is the predictive expectation of the sufficient statistics $j = 1, \ldots, p$ under $\mathcal{M}$.

An optimal $\hat{\boldsymbol{\delta}}_{\mathcal{A}}$ requires only estimation of $\overline{T_j}$, such as $\overline{T_j} \approx S^{-1} \sum_{s=1}^{S} T_j\{h(\tilde{\boldsymbol{y}}^s)\}$ for $\tilde{\boldsymbol{y}}^s \sim p_{\mathcal{M}}(\tilde{\boldsymbol{y}}|\boldsymbol{y})$, and minimization of the resulting deviance loss. 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 $\{\tilde{\boldsymbol{x}}_i\}_{i=1}^{\tilde{n}}$ and penalized loss functions $\mathcal{L}_\lambda^{EF} := \mathcal{L}_0^{EF} + \lambda \mathcal{P}$ are straightforward.

Despite the presence of the exponential family of *distributions*, we employ the deviance loss $\mathcal{L}_0^{EF}$ only for *point prediction* of $h(\tilde{\boldsymbol{y}})$. The loss function is chosen to reflect the nature of $h(\tilde{\boldsymbol{y}})$, 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 deviance loss does not impose additional assumptions on the distribution of $h(\tilde{\boldsymbol{y}})$: the predictive distribution is inherited from $\mathcal{M}$, while Theorem A.1 produces optimal point prediction parameters for a parametrized action $\mathcal{A}$ 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., 2020). These methods approximate $p_{\mathcal{M}}(\boldsymbol{y}|\boldsymbol{\theta})$ with a distribution $p_{\mathcal{M}}(\boldsymbol{y}|\hat{\boldsymbol{\delta}}_{KL})$ such that $\hat{\boldsymbol{\delta}}_{KL} = \arg\min_{\boldsymbol{\delta}} D_{KL}\left\{p_{\mathcal{M}}(\boldsymbol{y}|\boldsymbol{\theta}), p_{\mathcal{M}}(\boldsymbol{y}|\boldsymbol{\delta})\right\}$, usually for variable selection. Those approximations are derived for the *likelihood* $p_{\mathcal{M}}(\boldsymbol{y}|\boldsymbol{\theta})$ rather than the predictive distribution $p_{\mathcal{M}}(\tilde{\boldsymbol{y}}|\boldsymbol{y})$ and do not target any particular functional h. Consequently, the resulting global approximations may be unnecessarily complex or suboptimal locally for $h(\tilde{\boldsymbol{y}})$. 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 $h(\tilde{\boldsymbol{y}})$.

Example A.1 (Classification and cross-entropy)**.** Consider a binary functional $h(\tilde{\boldsymbol{y}}) \in \{0,1\}$, such as a discretized contrast for multivariate data, $h(\tilde{\boldsymbol{y}}) = \mathbb{I}\{\tilde{y}_1 > \tilde{y}_2\}$, or exceedance of a threshold t^* for functional data, $h(\tilde{\boldsymbol{y}}) = \mathbb{I}\{\exists \tau \in \mathcal{T} : \tilde{y}(\tau) > t^*\}$. The Bernoulli deviance for the canonical (logistic) link function is given by $\mathcal{L}_0^{EF}$ with $p = 1$, $F_0\{g(\tilde{\boldsymbol{x}};\boldsymbol{\delta})\} = \log[1 + \exp\{g(\tilde{\boldsymbol{x}};\boldsymbol{\delta})\}]$, $T_0 = 0$, $F_1\{g(\tilde{\boldsymbol{x}};\boldsymbol{\delta})\} = g(\tilde{\boldsymbol{x}};\boldsymbol{\delta})$, and $T_1\{h(\tilde{\boldsymbol{y}})\} = h(\tilde{\boldsymbol{y}})$. In this case, the deviance loss is the *cross-entropy*, which is a popular metric for classification. The predictive expectation $\overline{T_1} = \mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]} T_1\{h(\tilde{\boldsymbol{y}})\} = \mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]} h(\tilde{\boldsymbol{y}})$ required by Theorem A.1 is simply the posterior predictive probability of $\{h(\tilde{\boldsymbol{y}}) = 1\}$ under model $\mathcal{M}$. Interestingly, $\overline{T_1} \in [0,1]$ is on a continuous scale, and may contain more information than the binary empirical functional $h(\boldsymbol{y}) \in \{0,1\}$.

B Algorithms

B.1 SIR for out-of-sample predictive evaluation

The out-of-sample predictive metrics require computation of several quantities. For each data split $k = 1, \dots, K$, the out-of-sample *empirical* and *predictive* losses are

$$\mathbb{L}_{\mathcal{A}}^{out}(k) := \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} L\big\{h(\boldsymbol{y}_i), g(\boldsymbol{x}_i; \hat{\boldsymbol{\delta}}_{\mathcal{A}}^{-\mathcal{I}_k})\big\}, \quad \widetilde{\mathbb{L}}_{\mathcal{A}}^{out}(k) := \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} L\big\{h(\tilde{\boldsymbol{y}}_i^{-\mathcal{I}_k}), g(\boldsymbol{x}_i; \hat{\boldsymbol{\delta}}_{\mathcal{A}}^{-\mathcal{I}_k})\big\} \tag{B.1}$$

respectively, where $\hat{\boldsymbol{\delta}}_{\mathcal{A}}^{-\mathcal{I}_k}$ is estimated only using the training data $\boldsymbol{y}^{-\mathcal{I}_k} := \{\boldsymbol{y}_i\}_{i \notin \mathcal{I}_k}$,

$$\hat{\boldsymbol{\delta}}_{\mathcal{A}}^{-\mathcal{I}_k} := \arg\min_{\boldsymbol{\delta}} \mathbb{E}_{[\tilde{\boldsymbol{y}} | \boldsymbol{y}^{-\mathcal{I}_k}]} \bar{\mathcal{L}}_{\lambda}\big[\{h(\tilde{\boldsymbol{y}}_i), g(\tilde{\boldsymbol{x}}_i; \boldsymbol{\delta})\}_{i \notin \mathcal{I}_k}\big] \tag{B.2}$$

and similarly $\tilde{\boldsymbol{y}}_i^{-\mathcal{I}_k} \sim p_{\mathcal{M}}(\tilde{\boldsymbol{y}}_i | \boldsymbol{y}^{-\mathcal{I}_k})$ is the predictive variate at $\boldsymbol{x}_i$ conditional only on the training data. Evaluation of $\mathcal{A}$ is based on the averages of (B.1) across all data splits:

$$\mathbb{L}_{\mathcal{A}}^{out} := \frac{1}{K} \sum_{k=1}^{K} \mathbb{L}_{\mathcal{A}}^{out}(k), \quad \widetilde{\mathbb{L}}_{\mathcal{A}}^{out} := \frac{1}{K} \sum_{k=1}^{K} \widetilde{\mathbb{L}}_{\mathcal{A}}^{out}(k). \tag{B.3}$$

Under squared error loss, (B.2) simplifies as follows:

$$\hat{\boldsymbol{\delta}}_{\mathcal{A}}^{-\mathcal{I}_k} = \arg\min_{\boldsymbol{\delta}} \Big\{(n - |\mathcal{I}_k|)^{-1} \sum_{j \notin \mathcal{I}_k} \big\|\bar{h}_j^{-\mathcal{I}_k} - g(\boldsymbol{x}_j; \boldsymbol{\delta})\big\|_2^2 + \lambda \mathcal{P}(\boldsymbol{\delta})\Big\} \tag{B.4}$$

where $\bar{h}_j^{-\mathcal{I}_k} = \mathbb{E}_{[\tilde{\boldsymbol{y}}_j | \boldsymbol{y}^{-\mathcal{I}_k}]} h(\tilde{\boldsymbol{y}}_j)$ is the out-of-sample point prediction at $\boldsymbol{x}_j$. Similar simplifications are available for deviance loss.

The out-of-sample predictive metrics require estimates of the out-of-sample point prediction $\bar{h}_j^{-\mathcal{I}_k}$, a solution to the penalized least squares problem (B.4), and out-of-sample predictive draws $\tilde{\boldsymbol{y}}_i^{-\mathcal{I}_k} \sim p_{\mathcal{M}}(\tilde{\boldsymbol{y}}_i | \boldsymbol{y}^{-\mathcal{I}_k})$. We obtain these quantities without Bayesian model re-fitting using the sampling-importance resampling (SIR) algorithm in Algorithm 1. The samples $\{\tilde{\boldsymbol{y}}_i^{\tilde{s}}\}_{\tilde{s}=\tilde{s}_1}^{\tilde{S}}$ constitute draws from $p_{\mathcal{M}}(\tilde{\boldsymbol{y}}_i|\boldsymbol{y}^{-\mathcal{I}_k})$ for each $i \in \mathcal{I}_k$; we use $\tilde{S} = \lfloor S/2 \rfloor$ SIR samples. These samples, along with $\hat{\boldsymbol{\delta}}_{\mathcal{A}}^{-\mathcal{I}_k}$, are sufficient for computing the key terms in predictive evaluation: the out-of-sample predictive loss $\widetilde{\mathbb{L}}_{\mathcal{A}}^{out}$, the predictive discrepancies $\widetilde{\mathbb{D}}_{\mathcal{A},\mathcal{A}'}^{out}$, and the set of acceptable models $\Lambda_{\eta,\varepsilon}$.

Algorithm 1: Out-of-sample predictive loss.

```
1. Obtain posterior samples $\{\boldsymbol{\theta}^s\}_{s=1}^S \sim p_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{y})$;
2. For each training set $k = 1, \ldots, K$:
    (a) Compute $\log w_k^s \overset{c}{=} -\log p_{\mathcal{M}}(\boldsymbol{y}^{\mathcal{I}_k}|\boldsymbol{\theta}^s) = -\sum_{i \in \mathcal{I}_k} \log p_{\mathcal{M}}(\boldsymbol{y}_i|\boldsymbol{\theta}^s)$ (up to a constant);
    (b) Sample $\{\tilde{s}_1, \ldots, \tilde{s}_{\tilde{S}}\}$ without replacement from $\{1, \ldots, S\}$ with probability
        weights $\{w_k^1, \ldots, w_k^S\}$;
    (c) Sample $\tilde{\boldsymbol{y}}_i^{\tilde{s}} \sim p_{\mathcal{M}}(\tilde{\boldsymbol{y}}_i|\boldsymbol{\theta}^{\tilde{s}})$ for $\tilde{s} = \tilde{s}_1, \ldots, \tilde{s}_{\tilde{S}}$ and $i = 1, \ldots, n$;
    (d) Compute $\bar{h}_j^{-\mathcal{I}_k} \approx \tilde{S}^{-1} \sum_{\tilde{s}=\tilde{s}_1}^{\tilde{S}} h(\tilde{\boldsymbol{y}}_j^{\tilde{s}})$ for $j \notin \mathcal{I}_k$;
    (e) Compute $\hat{\boldsymbol{\delta}}_{\mathcal{A}}^{-\mathcal{I}_k}$ by solving (B.4) for each $\mathcal{A} \in \mathbb{A}$;
    (f) Compute $\mathbb{L}_{\mathcal{A}}^{out}(k)$ and $\{\widetilde{\mathbb{L}}_{\mathcal{A}}^{out,\tilde{s}}(k)\}_{\tilde{s}=\tilde{s}_1}^{\tilde{S}}$ in (B.1) using $\hat{\boldsymbol{\delta}}_{\mathcal{A}}^{-\mathcal{I}_k}$ and $\{h(\tilde{\boldsymbol{y}}_i^{\tilde{s}})\}_{\tilde{s}=\tilde{s}_1}^{\tilde{S}}$;
3. Compute $\mathbb{L}_{\mathcal{A}}^{out}$ and $\{\widetilde{\mathbb{L}}_{\mathcal{A}}^{out,\tilde{s}}\}_{\tilde{s}=\tilde{s}_1}^{\tilde{S}}$ in (B.3) using $\mathbb{L}_{\mathcal{A}}^{out}(k)$ and $\{\widetilde{\mathbb{L}}_{\mathcal{A}}^{out,\tilde{s}}(k)\}_{\tilde{s}=\tilde{s}_1}^{\tilde{S}}$.
```

Algorithm 1 allows for recycled computations: steps 1–2(c) are shared across all functionals h, parametrized actions $\mathcal{A}$, and loss functions L. As a result, these computations are a one-time cost for all predictive evaluations under $\mathcal{M}$. Step 2(d) depends only on the functional h, while steps 2(e)–3 depend on h, $\mathcal{A}$, and L. The solutions $\hat{\boldsymbol{\delta}}_{\mathcal{A}}$ and the Bayes estimators $\bar{h}_i$ from the *full* dataset are not needed for predictive evaluation.

B.2 Gibbs sampling for the STAR functional regression model

We design a Gibbs sampling algorithm for the proposed count-valued functional regression model based on the simultaneous transformation and rounding (STAR) framework of Kowal and Canale (2020). For each individual, we aggregate physical activity (PA) data 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 $\mathcal{M}$ is the following:

$$y_{i,j}^{tot} = \texttt{round}(y_{i,j}^*), \quad z_{i,j}^* = \texttt{transform}(y_{i,j}^*) \tag{B.5}$$

$$z_{i,j}^* = \boldsymbol{b}'(\tau_j)\boldsymbol{\theta}_i + \sigma_\epsilon \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} t_\nu(0,1) \tag{B.6}$$

$$\theta_{i,\ell} = \boldsymbol{x}_i' \boldsymbol{\alpha}_\ell + \sigma_{\gamma_i}\gamma_{i,\ell}, \quad \gamma_{i,\ell} \overset{iid}{\sim} N(0,1) \tag{B.7}$$

with the priors $\alpha_{\ell,j} \overset{indep}{\sim} N(0, \sigma_{\alpha_j}^2)$ and $\sigma_\epsilon^{-2}, \sigma_{\gamma_i}^{-2}, \sigma_{\alpha_j}^{-2} \overset{iid}{\sim} \text{Gamma}(0.01, 0.01)$. In (B.5), `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 $\texttt{round}(t) = \lfloor t \rfloor$ for $t > 0$ and $\texttt{round}(t) = 0$ for $t \leq 0$, so $y_{i,j}^{tot} = 0$ whenever $y_{i,j}^* < 0$. Within the Box-Cox family, we find that $\texttt{transform}(t) = 2(\sqrt{t} - 1)$ is adequate for the predictive functionals of interest. The functional regression model is given in (B.6) and (B.7): $\boldsymbol{b}$ is a vector of 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. We use a spline basis with the reparametrization of Scheipl et al. (2012), which simultaneously orthogonalizes $\boldsymbol{b}$ and diagonalizes the prior variance of the basis coefficients. This diagonalization justifies the assumption of independence across basis coefficients in (B.7). Heavy-tailed innovations ($\nu = 3$) are introduced to model large spikes in PA. Within the Gibbs sampler, we use the parameter expansion $\epsilon_i | \xi_{\epsilon_i} \sim N(0, \xi_{\epsilon_i}^{-1})$ with $\xi_{\epsilon_i} \overset{iid}{\sim} \text{Gamma}(\nu/2, \nu/2)$.

We introduce the following notation. Let $\boldsymbol{B}$ denote the $m \times L$ basis matrix with columns $\boldsymbol{B}_\ell' = (b_\ell(\tau_1), \ldots, b_\ell(\tau_m))$ for $\ell = 1, \ldots, L$. Since $\boldsymbol{B}'\boldsymbol{B}$ is diagonal by orthogonalization, let $\texttt{diag}_\ell(\boldsymbol{B}'\boldsymbol{B})$ denote the ℓth diagonal element. Let $\boldsymbol{z}_i^* = (z_{i,1}^*, \ldots, \boldsymbol{z}_{i,m}^*)'$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, $\boldsymbol{\theta}_\ell = (\theta_{1,\ell}, \ldots, \theta_{n,\ell})'$, and n_i the number of binned observations for subject i (i.e., the number of days of data for subject i times 5 minutes per bin).

The Gibbs sampling algorithm is in Algorithm 2. Following the sampling of $z_{i,j}^*$, the remaining steps are mostly standard for functional regression. The primary differences are (i) the sampler for the basis coefficients $\theta_{i,\ell}$ uses the orthogonality of the basis matrix $\boldsymbol{B}$ to improve computational efficiency and (ii) the posterior predictive draws of $\tilde{y}_{i,j}$ are generated according to STAR. Predictive functionals are computed by evaluating $h(\tilde{\boldsymbol{y}}_i)$ for each draw $\tilde{\boldsymbol{y}}_i = (\tilde{y}_{i,1}, \ldots, \tilde{y}_{i,m})'$ from the posterior predictive distribution of the (average) intraday PA.

Algorithm 2: Gibbs sampler for count-valued functional regression.

```
1. Sample $[z_{i,j}^* | -] \overset{indep}{\sim} N\big(\boldsymbol{b}'(\tau_j)\boldsymbol{\theta}_i, \sigma_\epsilon^2/\xi_{\epsilon_i}\big)$ truncated to $[\texttt{transform}(y_{i,j}^{tot}), \texttt{transform}(y_{i,j}^{tot} + 1))$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$;
2. Sample $[\theta_{i,\ell} | -] \overset{indep}{\sim} N\big(Q_{\theta_{i,\ell}}^{-1}\ell_{\theta_{i,\ell}}, Q_{\theta_{i,\ell}}^{-1}\big)$ where $Q_{\theta_{i,\ell}} = \texttt{diag}_\ell(\boldsymbol{B}'\boldsymbol{B})\xi_{\epsilon_i}/\sigma_\epsilon^2 + 1/\sigma_{\gamma_i}^2$ and $\ell_{\theta_{i,\ell}} = \boldsymbol{B}_\ell'\boldsymbol{z}_i^*\xi_{\epsilon_i}/\sigma_\epsilon^2 + \boldsymbol{x}_i'\boldsymbol{\alpha}_\ell/\sigma_{\gamma_i}^2$ for $i = 1, \ldots, n$ and $\ell = 1, \ldots, L$;
3. Sample $[\boldsymbol{\alpha}_\ell | -] \overset{indep}{\sim} N\big(\boldsymbol{Q}_{\alpha_\ell}^{-1}\boldsymbol{\ell}_{\alpha_\ell}, \boldsymbol{Q}_{\alpha_\ell}^{-1}\big)$ where $\boldsymbol{Q}_{\alpha_\ell} = \boldsymbol{X}'\texttt{diag}(\{\sigma_{\gamma_i}^{-2}\}_{i=1}^n)\boldsymbol{X} + \texttt{diag}(\{\sigma_{\alpha_j}^{-2}\}_{j=1}^p)$ and $\boldsymbol{\ell}_{\alpha_\ell} = \boldsymbol{X}'\texttt{diag}(\{\sigma_{\gamma_i}^{-2}\}_{i=1}^n)\boldsymbol{\theta}_\ell$ for $\ell = 1, \ldots, L$;
4. Sample $[\sigma_\epsilon^{-2} | -] \sim \text{Gamma}\big(nm/2, \sum_{i=1}^n \sum_{j=1}^m \xi_{\epsilon_i}\{z_{i,j}^* - \boldsymbol{b}'(\tau_j)\boldsymbol{\theta}_i\}^2/2\big)$;
5. Sample $[\xi_{\epsilon_i} | -] \overset{indep}{\sim} \text{Gamma}\big(m/2 + \nu/2, \sigma_\epsilon^{-2} \sum_{j=1}^m \{z_{i,j}^* - \boldsymbol{b}'(\tau_j)\boldsymbol{\theta}_i\}^2/2 + \nu/2\big)$ for $i = 1, \ldots, n$;
6. Sample $[\sigma_{\alpha_j}^{-2} | -] \overset{indep}{\sim} \text{Gamma}\big(L/2 + 0.01, \sum_{\ell=1}^L \alpha_{\ell,j}^2/2 + 0.01\big)$ for $j = 1, \ldots, p$;
7. Sample $[\sigma_{\gamma_i}^{-2} | -] \overset{indep}{\sim} \text{Gamma}\big(L/2 + 0.01, \sum_{\ell=1}^L (\theta_{i,\ell} - \boldsymbol{x}_i'\boldsymbol{\alpha}_\ell)^2/2 + 0.01\big)$ for $i = 1, \ldots, n$;
8. Compute $\tilde{y}_{i,j} = \tilde{y}_{i,j}^{tot}/n_i$ where $\tilde{y}_{i,j}^{tot} = \texttt{round}\{\texttt{transform}^{-1}(\tilde{z}_{i,j})\}$ and $[\tilde{z}_{i,j} | -] \overset{indep}{\sim} N\big(\boldsymbol{b}'(\tau_j)\boldsymbol{\theta}_i, \sigma_\epsilon^2/\xi_{\epsilon_i}\big)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$;
```

C Simulation study

C.1 Additional simulation results

We evaluate targeted point predictions of h, estimates of the linear coefficients, and variable selection properties using simulated data. The main paper reports results for $m = 200$, $p = 50$, and $n \in \{100, 500\}$. Here, we consider four variations: a smaller sample size ($n = 75$), more predictors than functional observations ($p > n$), sensitivity to $\varepsilon \in \{0.05, 0.10, 0.20, 0.50\}$, and a validation dataset where the distribution of the design points $\tilde{\boldsymbol{x}}$ differs from that of the training data.

An example of the simulated data is presented in Figure C.1. The functions are piecewise linear and concave with a single breakpoint—the argmax—which is determined by a sparse linear model. Gaussian noise was added to produce the functional data $\boldsymbol{y}_i$. Note that we restrict the functional $h(Y_i^*) = \boldsymbol{x}_i'\boldsymbol{\beta}^*$ to the interval $[0.2, 0.8]$ by shifting and scaling the (nonzero) coefficients as follows: set $\beta_0^* \leftarrow \min\{\boldsymbol{x}_i'\boldsymbol{\beta}^*\}$; shift $\beta_j^* \leftarrow \beta_j^* \times 0.6/\text{range}\{\boldsymbol{x}_i'\boldsymbol{\beta}^*\}$ for $j = 0, 1, \ldots, p$; and reset $\beta_0^* \leftarrow \beta_0^* + 0.2$. Figure C.1 demonstrates that the empirical argmax values $h(\boldsymbol{y}_i)$ are sensitive to outlying data points, while the Bayesian model $\mathcal{M}$ estimates $\bar{h} = \mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]} h(\tilde{\boldsymbol{y}})$ are more robust. Most interesting, the proposed point predictions further improve upon $\mathcal{M}$ and more closely match the true argmax values. These results are confirmed more rigorously in the simulation study in the main paper.

C.2 Marginal variable selection

We present the marginal variable selection results for $m = 200$, $p = 50$, and $n \in \{75, 100, 500\}$ in Table C.1. Specifically, we report true positive and true negative rates for each selection method averaged across the p predictors and the 100 simulations. Notably, `proposed(out)` offers the best marginal variable selection properties, with large TPRs and TNRs in all settings. By comparison, `projpred` reports lower TPRs while `proposed(in)` suffers from

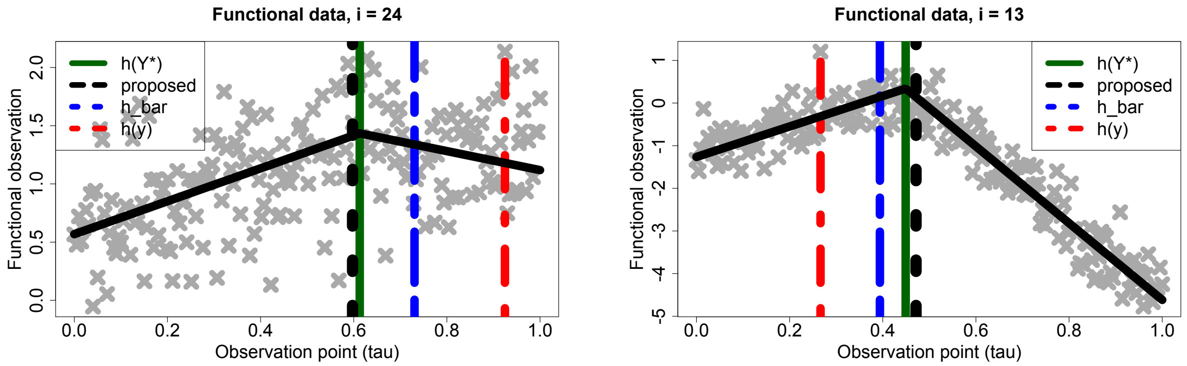


Figure C.1: True functions (solid black line) and functional data observations $\boldsymbol{y}_i$ (gray x-marks) for two curves. Vertical lines denote the true argmax functional (dark green), the proposed point prediction (dotted black), the model $\mathcal{M}$ prediction (dashed blue), and the empirical value $h(\boldsymbol{y}_i)$ (dot-dashed red). The proposed predictions more closely match the true values.

lower TNRs. These results illustrate the gains from the parametrized decision analysis under a Bayesian model $\mathcal{M}$—rather than variable selection based on the empirical functionals $\{\boldsymbol{x}_i, h(\boldsymbol{y}_i)\}_{i=1}^n$—and highlight the crucial distinction between in-sample and out-of-sample variable selection.

		adaptive lasso	projpred	proposed(in)	proposed(out)
$n = 75$	TPR	0.89	0.73	0.99	0.96
	TNR	0.90	0.94	0.46	0.94
$n = 100$	TPR	0.95	0.82	1.00	0.99
	TNR	0.97	0.96	0.48	0.98
$n = 500$	TPR	1.00	1.00	1.00	1.00
	TNR	1.00	0.99	0.46	0.99

Table C.1: True positive rates (TPR) and true negative rates (TNR) for the synthetic data ($p = 50$, $m = 200$). The parametric action with out-of-sample selection offers the best marginal variable selection properties, especially for smaller sample sizes.

C.3 Smaller sample size

In Figure C.2, we present the prediction and estimation comparisons for the case of $n = 75$, $m = 200$, and $p = 50$. The results are similar to those for $n = 100$: clear improvements in targeted prediction are obtained by (i) fitting to $h(\tilde{\boldsymbol{y}}_i)$ (via $\bar{h}_i$) rather than $h(\boldsymbol{y}_i)$, (ii) including covariate information, (iii) incorporating penalization or variable selection, and (iv) selecting the complexity based on out-of-sample evaluations.

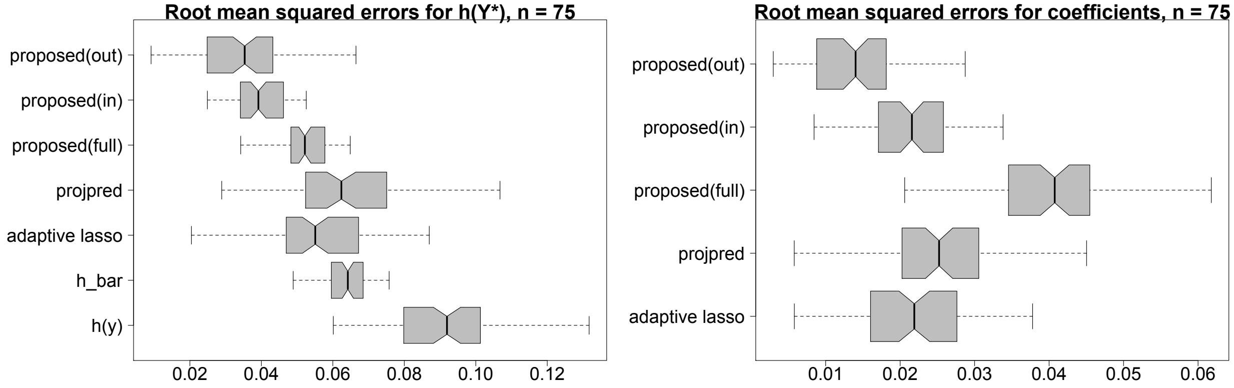


Figure C.2: RMSEs for the true functionals $h(Y^*)$ (**left**) and the true regression coefficients $\boldsymbol{\beta}^*$ (**right**) for $n = 75$ across 100 simulated datasets. Non-overlapping notches indicate significant differences between medians. The proposed point predictions and estimates using out-of-sample selection of λ are most accurate.

C.4 More covariates than observations: $p > n$

For the case of $p > n$, we consider two scenarios, both with $m = 200$: $(n = 200, p = 500)$ and $(n = 100, p = 200)$. The remaining simulation specifications are unchanged.

The adaptive penalization in the proposed approach requires modification due to $p > n$. Recall that the penalty $\mathcal{P}(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_{j=1}^{p} \omega_j |\delta_j|$, where the weights ω_j derive from model $\mathcal{M}$. Previously, we used $\boldsymbol{\omega} = |\tilde{\boldsymbol{\delta}}_0|^{-1}$, where $\tilde{\boldsymbol{\delta}}_0$ is the ℓ_2-projection of the predictive variables $\{h(\tilde{\boldsymbol{y}}_i)\}_{i=1}^{n}$ onto $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, which inherits a predictive distribution under $\mathcal{M}$. To adjust for $p > n$, we instead use the weights $\boldsymbol{\omega} = |\tilde{\boldsymbol{\delta}}_R|^{-1}$, where $\tilde{\boldsymbol{\delta}}_R$ is the ridge regression estimator on $\{\boldsymbol{x}_i, h(\tilde{\boldsymbol{y}}_i)\}_{i=1}^{n}$. The ridge tuning parameter is obtained by fitting a ridge regression to $\{\boldsymbol{x}_i, \bar{h}_i\}_{i=1}^{n}$ and selecting the parameter that minimizes cross-validated mean squared error. As a result, the same tuning parameter is used across all posterior predictive samples of $\{\tilde{\boldsymbol{y}}_i\}_{i=1}^{n}$. As before, we integrate the penalty over the posterior predictive distribution,

$\overline{\mathcal{P}(\boldsymbol{\delta})} := \mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]}\mathcal{P}(\boldsymbol{\delta}, \boldsymbol{\theta}) = \sum_{j=1}^{p} \hat{\omega}_j |\delta_j|$ for $\hat{\boldsymbol{\omega}} = \mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]}(|\tilde{\boldsymbol{\delta}}_R|^{-1})$, which is estimable using posterior predictive samples. The weights for the adaptive lasso fit to the empirical functionals $\{\boldsymbol{x}_i, h(\boldsymbol{y}_i)\}_{i=1}^n$ are similarly modified using a ridge-based adjustment.

Point predictions of $h(Y_i^*)$ and estimates of $\boldsymbol{\beta}^*$ are evaluated using root mean squared errors (RMSEs) in Figure C.3. Most notably, the parametrized linear action with out-of-sample variable selection (`proposed(out)`) is consistently the most accurate for prediction and estimation. Perhaps most interesting, the (unparametrized) Bayes estimator $\bar{h}$ outperforms the empirical functional $h(\boldsymbol{y})$ as well as the adaptive lasso and the parametrized linear action without selection (`proposed(full)`) for point prediction. Clearly, the parametrized action offers the greatest advantages when the complexity penalty is included. Although in-sample and out-of-sample selection provide equally accurate point prediction, the out-of-sample selection improves the estimation of the linear coefficients. Note that we have omitted `projpred` due to extremely slow computation times for the Bayesian linear model sampling algorithm (`stan_glm`) in the `rstanarm` package under these choices of n and p.

C.5 Sensitivity to ε

To assess sensitivity to the probability level ε, which defines the acceptable predictor set and therefore the smallest acceptable predictor, we evaluate the prediction and estimation accuracy for sparse linear actions with $\varepsilon \in \{0.05, 0.10, 0.20, 0.50\}$. We consider the case of $n = 100$, $m = 200$, and $p = 50$ and present the results in Figure C.4. As expected, the prediction and estimation accuracy improves as ε increases, which pulls the simplest acceptable predictor toward the best predictor $\mathcal{A}_{min}$. The smallest value of $\varepsilon = 0.01$ sacrifices predictive ability for sparsity. Notably, the TPRs are comparable in all cases—0.98 for $\varepsilon = 0.01$ and 0.99 otherwise—but the TNRs decline from 0.99 ($\varepsilon = 0.01$) to 0.93 ($\varepsilon = 0.50$). These results affirm our choice of $\varepsilon = 0.10$, which simultaneously produces excellent predictions, estimations, and variable selection.

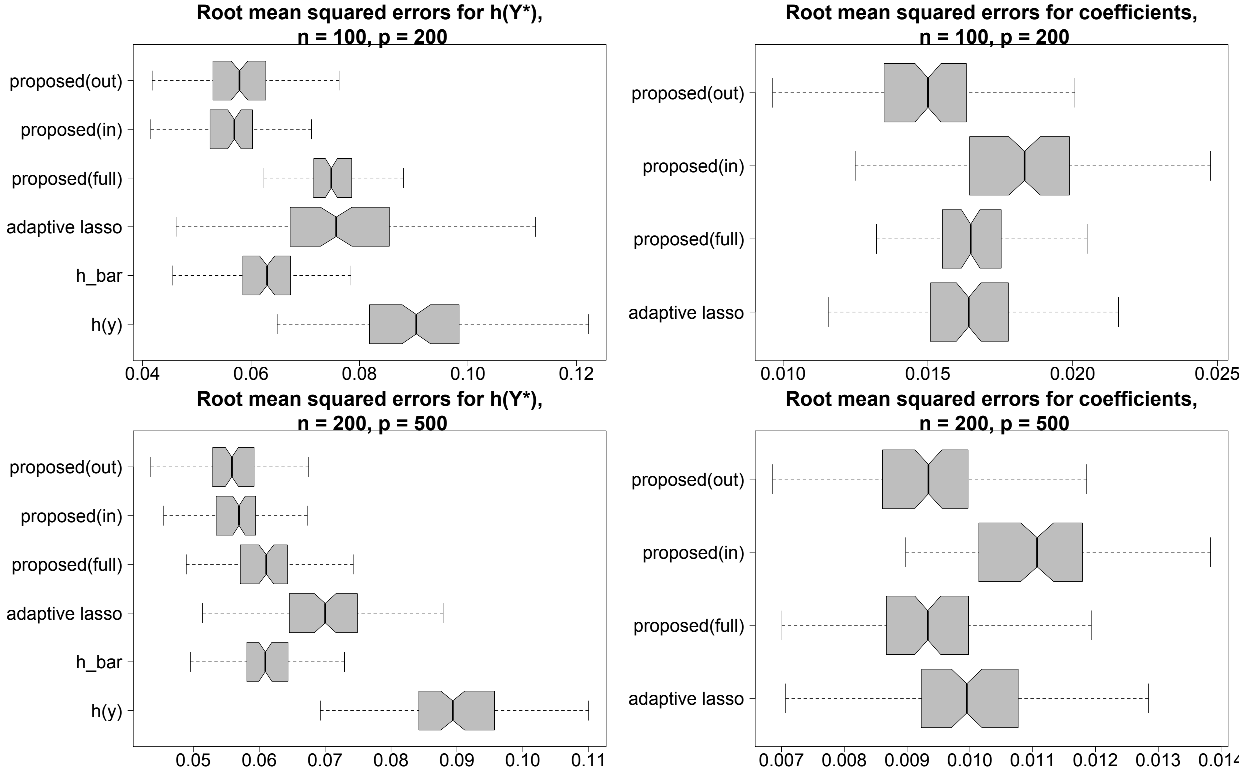
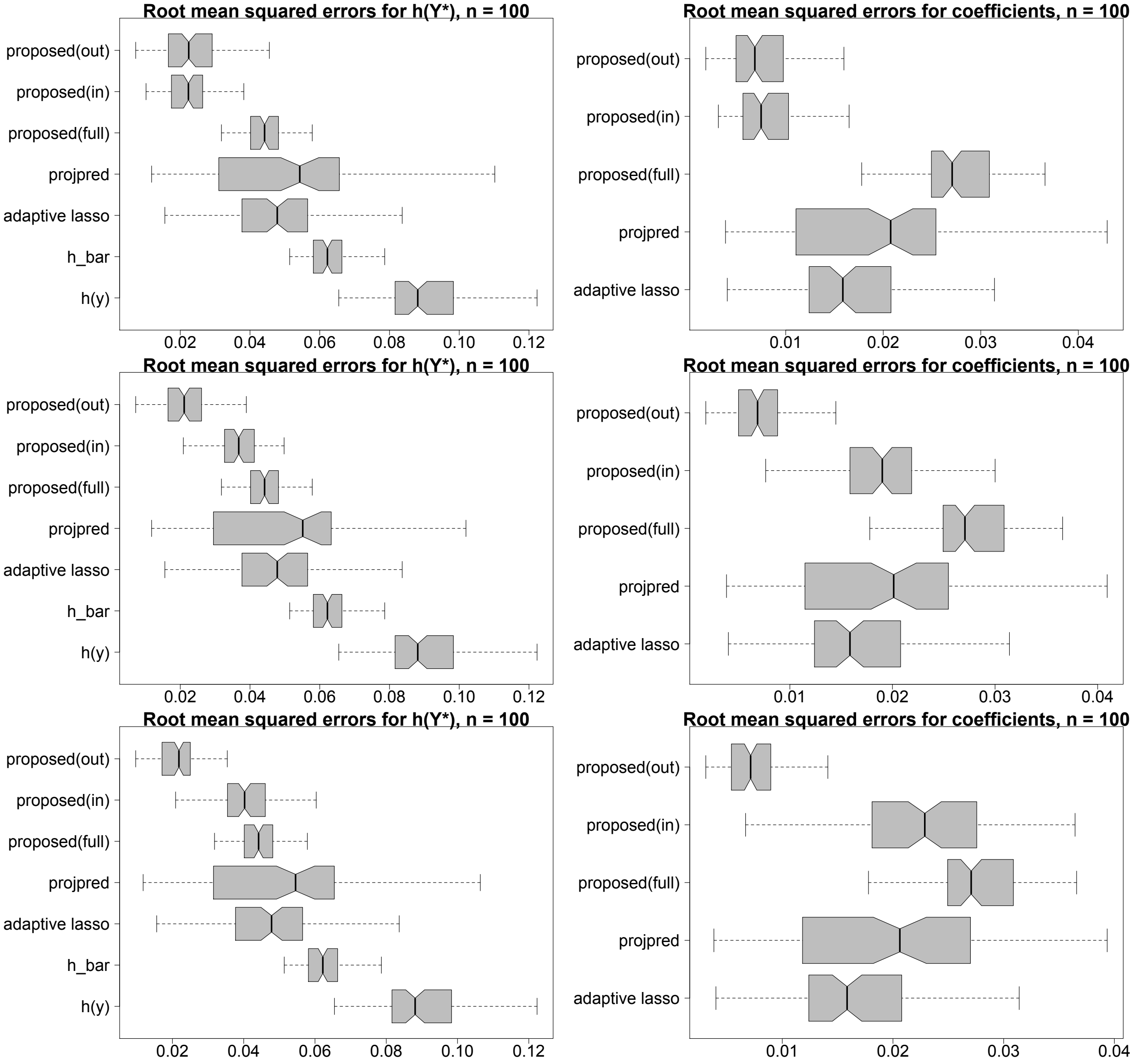


Figure C.3: RMSEs for the true functionals $h(Y^*)$ (**left**) and the true regression coefficients $\boldsymbol{\beta}^*$ (**right**) for $n = 100, p = 200$ (**top**) and $n = 200, p = 500$ (**bottom**) across 100 simulated datasets. Non-overlapping notches indicate significant differences between medians. The parametrized actions with out-of-sample selection are most accurate for prediction and estimation.

C.6 Differential training and testing covariates

We evaluate parameterized actions for targeted prediction for the scenario in which the distribution of the covariates used for training differs substantially from the distribution of covariates used for testing. For the training data, the covariates $\{x_{i,j}\}$ are drawn from marginal standard normal distributions with $\text{Cor}(x_{i,j}, x_{i,j'}) = (0.75)^{|j-j'|}$, and half of these covariates are binarized: $x_{i,j} \leftarrow \mathbb{I}\{x_{i,j} \geq 0\}$. For the testing data, we simulate covariates $\{x_{i,j}^*\}$ from marginal standard t_3 distributions with the same correlation structure, $\text{Cor}(x_{i,j}^*, x_{i,j'}^*) = (0.75)^{|j-j'|}$, but a different binarized threshold for the discrete variables: $x_{i,j}^* \leftarrow \mathbb{I}\{x_{i,j}^* > 0.5\}$. Compared to the training covariates, the testing covariates feature


Root mean squared errors for h(Y*), n = 100

Root mean squared errors for coefficients, n = 100

Root mean squared errors for h(Y*), n = 100

Root mean squared errors for coefficients, n = 100

Root mean squared errors for h(Y*), n = 100

Root mean squared errors for coefficients, n = 100

Figure C.4: RMSEs for the true functionals $h(Y^*)$ (**left**) and the true regression coefficients $\boldsymbol{\beta}^*$ (**right**) for $\varepsilon = 0.01$ (**top**), $\varepsilon = 0.20$ (**middle**), and $\varepsilon = 0.50$ (**bottom**) with $n = 100$, $m = 200$, and $p = 50$ 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, with better results for larger ε.

continuous variables with heavier-tailed marginal distributions and discrete variables with a distorted allocation of zeros and ones. In both the training and testing datasets, the continuous covariates are centered and scaled to sample standard deviation 0.5. Given the testing covariates $\{x^*_{i,j}\}$, we simulate the latent testing functions $\{Y^*_{test,i}(\tau) : \tau \in [0,1]\}$ as in the original simulation study.

Point predictions of $h(Y^*_{test,i})$ are evaluated using root mean squared errors in Figure C.5. We present results for $m = 200$ and $p = 50$ with $n \in \{75, 500\}$; the results for $n = 100$ are similar to those for $n = 75$ and are omitted. The parametrized linear actions with out-of-sample variable selection (`proposed(out)`) produce the most accurate point predictions, with substantial gains arriving for the smaller sample size. Notably, the (unparametrized) Bayes estimator $\bar{h}$—which is nominally optimal under squared error loss in classical decision analysis— underperforms relative to the parameterized actions. In addition, this estimator is the slowest to compute: we estimate $\bar{h}_i$ by sampling from the posterior predictive distribution at each testing point $\boldsymbol{x}^*_i$, applying h to each simulated function, and computing the MCMC sample mean across draws. By comparison, point predictions under the linear actions are simply computed as $\hat{\boldsymbol{\delta}}'_{\mathcal{A}} \boldsymbol{x}^*_i$.

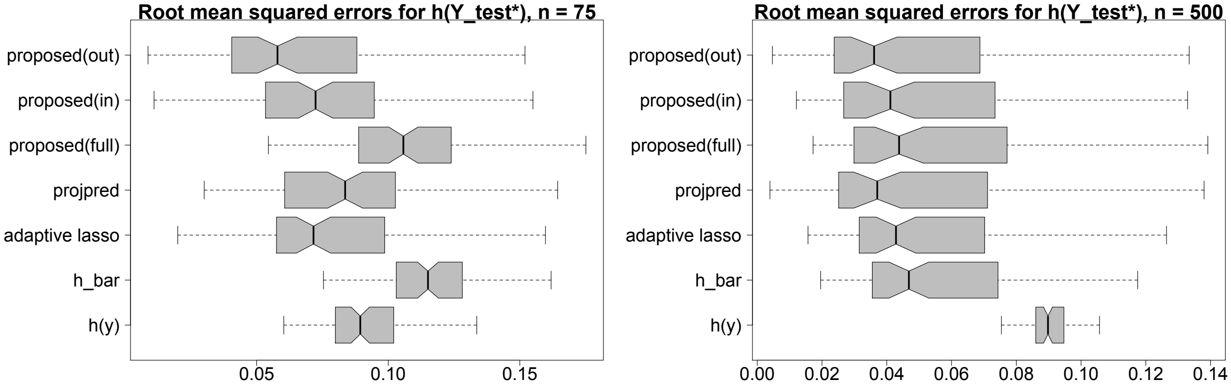


Figure C.5: RMSEs for the true functionals $h(Y^*_{test})$ on the testing data with distinctly distributed covariates $\{x^*_{i,j}\}$ for $n = 75$ (**left**) and $n = 500$ (**right**) across 100 simulated datasets. Non-overlapping notches indicate significant differences between medians. The proposed point predictions and estimates using out-of-sample selection are most accurate.

D Physical activity data analysis

D.1 Additional results: physical activity data

We expand upon our analysis of the National Health and Nutrition Examination Survey (NHANES) physical activity (PA) data. Targeted predictions for each functional (see the main paper) were constructed using a linear action $g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta}) = \tilde{\boldsymbol{x}}'\boldsymbol{\delta}$ with an adaptive ℓ_1-penalty. 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 model $\lambda = \lambda_{0,0.1}$ (`proposed(out)`) and the full model $\lambda = 0$ (`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. Here, we present posterior predictive diagnostics, additional results for vigorous PA, and an expanded set of covariates based on quadratic and interaction effects.

D.2 Posterior predictive diagnostics

Posterior predictive diagnostics for the functionals of interest are provided in Figure D.1, which plots the sample (kernel) density estimates for the empirical functionals $\{h(\boldsymbol{y}_i)\}_{i=1}^n$ and the predictive functionals $\{h(\tilde{\boldsymbol{y}}_i)\}_{i=1}^n$ for 500 draws from the posterior predictive distribution under model $\mathcal{M}$. 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 ν, but alternative choices of `transform` or $\boldsymbol{b}$ (such as wavelets) produce inferior results.

The posterior predictive diagnostics for the binary functional `zeros(1am-5am)` are in Figure D.2. Model $\mathcal{M}$ is successful in distinguishing between the classes of `zeros(1am-5am)`.

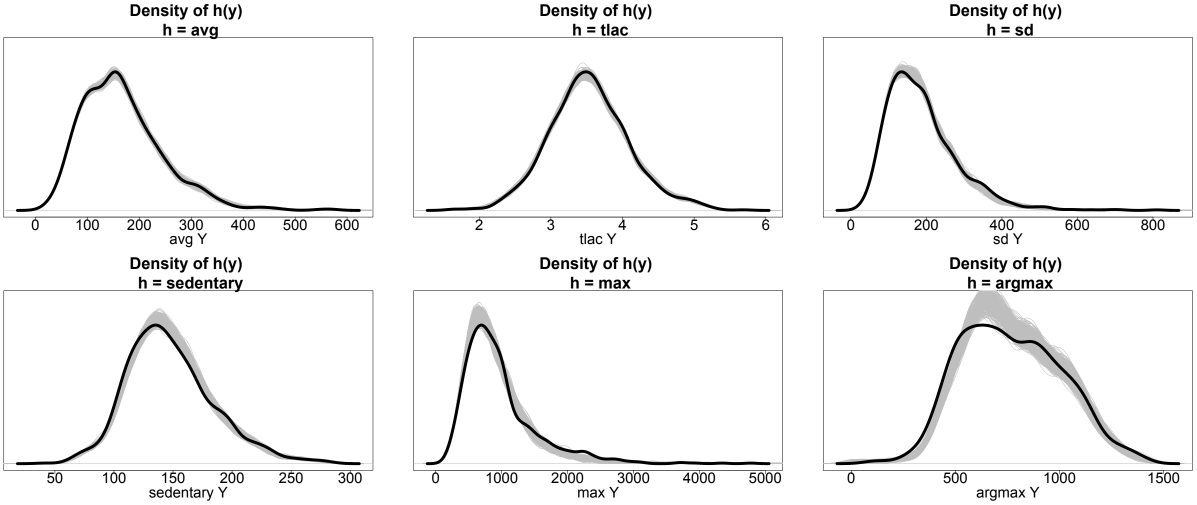


Figure D.1: Kernel density estimates for the empirical functionals $\{h(\boldsymbol{y}_i)\}_{i=1}^n$ (black line) and the predictive functionals $\{h(\tilde{\boldsymbol{y}}_i)\}_{i=1}^n$ (gray lines) for 500 predictive samples. The model $\mathcal{M}$ in (B.5)-(B.7) appears to be adequate for these functionals, with some difficulty for `argmax` around the mode.

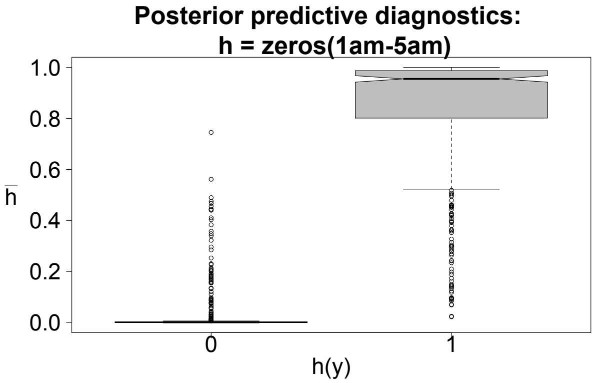


Figure D.2: Posterior predictive diagnostics: the binary empirical functionals $\{h(\boldsymbol{y}_i)\}_{i=1}^n$ and the predictive expected functionals $\{\bar{h}_i\}_{i=1}^n$. The Bayesian model $\mathcal{M}$ appears adequate.

D.3 Out-of-sample evaluations for vigorous PA

Due to the similarity among the vigorous PA measures—`avg`, `sd`, and `max`—the main paper only presented results for `max`. In Figure D.3, we present the predictive and empirical loss relative to the best predictor $\mathcal{A}_{min}$ for all of the vigorous PA measures. Clearly, the results are similar, and indeed the selected variables are identical up to the smallest acceptable linear predictor of size 10.

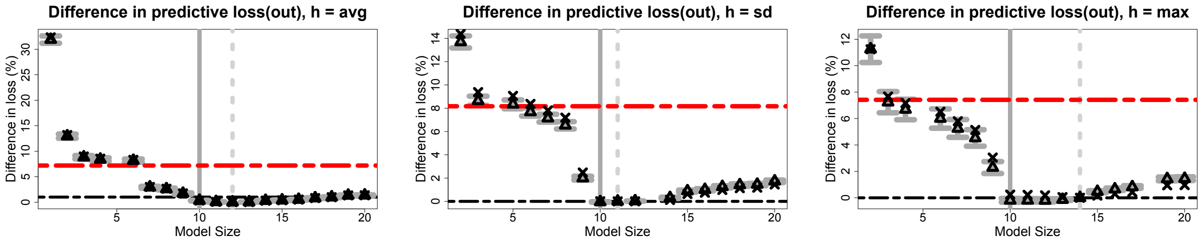


Figure D.3: Approximate out-of-sample squared error loss for sparse linear actions targeted to vigorous PA. Results are presented for each size as a percent increase in loss relative to $\mathcal{A}_{min}$. The predictive expectations (triangles) and 80% intervals (gray bars) are included with the empirical 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).

D.4 Targeted prediction with quadratic and interaction effects

We expand the set of covariates by including quadratic effects for age and BMI as well as pairwise interactions for each of age and BMI with race, gender, the behavioral attributes, and the self-reported comorbidity factors. The main effects (age and BMI) are included in all parametrized actions.

The targeted predictions are evaluated using the proposed (approximate) out-of-sample metrics. For each functional h and linear action size indexed by λ, Figure D.4 presents the predictive and empirical loss relative to the best model $\mathcal{A}_{min}$. The predictive expectations align closely with the empirical values, which suggests that model $\mathcal{M}$ is adequate for these predictive metrics. The parameterized actions including interactions now prefer including more variables, which suggests that many of these interactions are important for prediction. In all cases, the smallest acceptable predictor outperforms the adaptive lasso.

The results for the binary functional `zeros(1am-5am)` are presented in Figure D.5. For illustration, we include both (approximate) out-of-sample and in-sample versions of the predictive metrics, and include the linear actions with and without interactions. The distinction

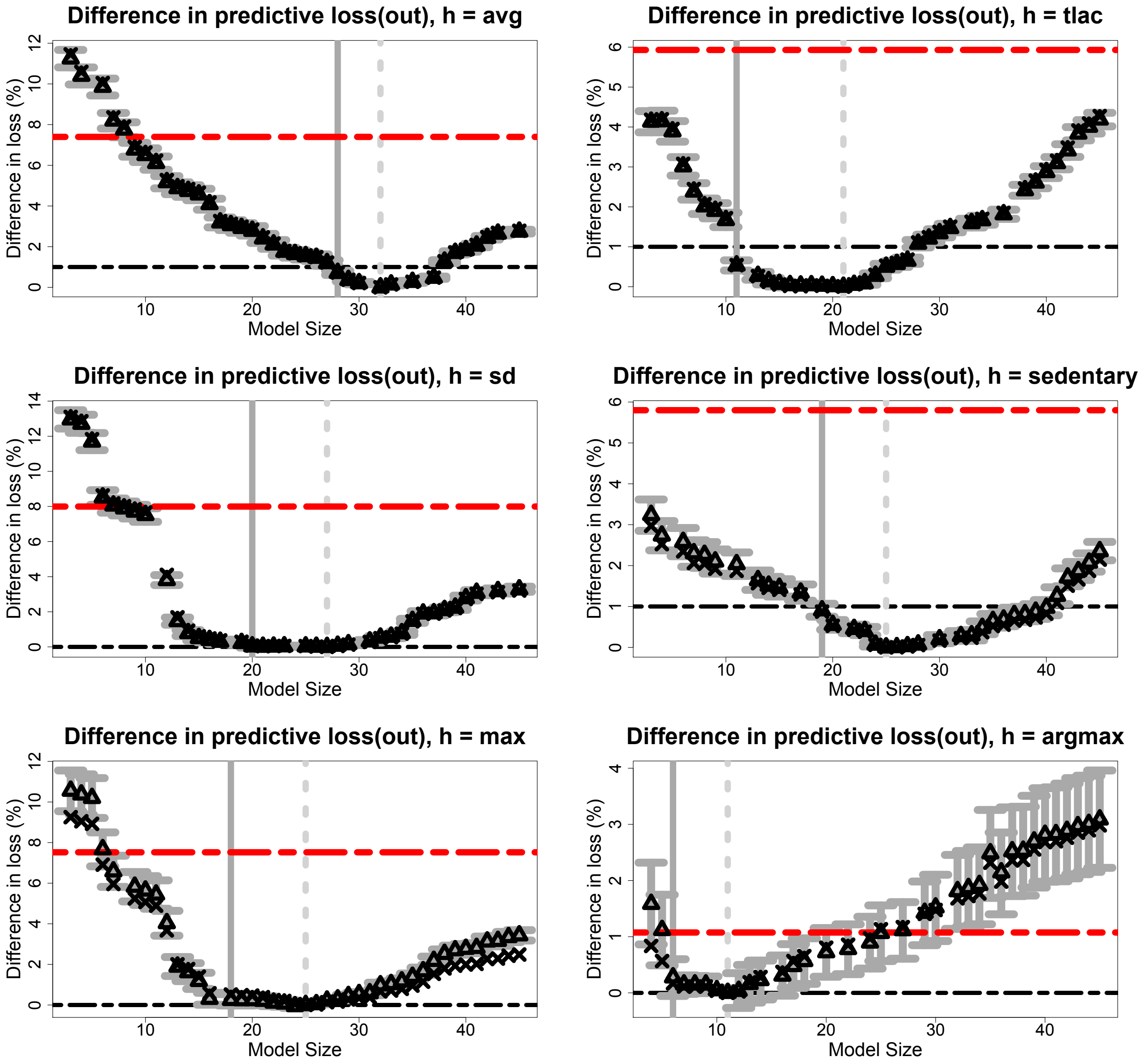


Figure D.4: Approximate out-of-sample squared error loss for sparse linear models (with interactions) targeted to each functional. Results are presented for each size as a percent increase in loss relative to $\mathcal{A}_{min}$. The predictive expectations (triangles) and 80% intervals (gray bars) are 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).

is stark: the in-sample version favors much larger sets of covariates, which are not supported by the out-of-sample metrics. For the out-of-sample version with interactions, the selected variables are race, gender, smoking status, a quadratic age effect, and the interactions age × cancer and BMI × race (BMI and age are included automatically).

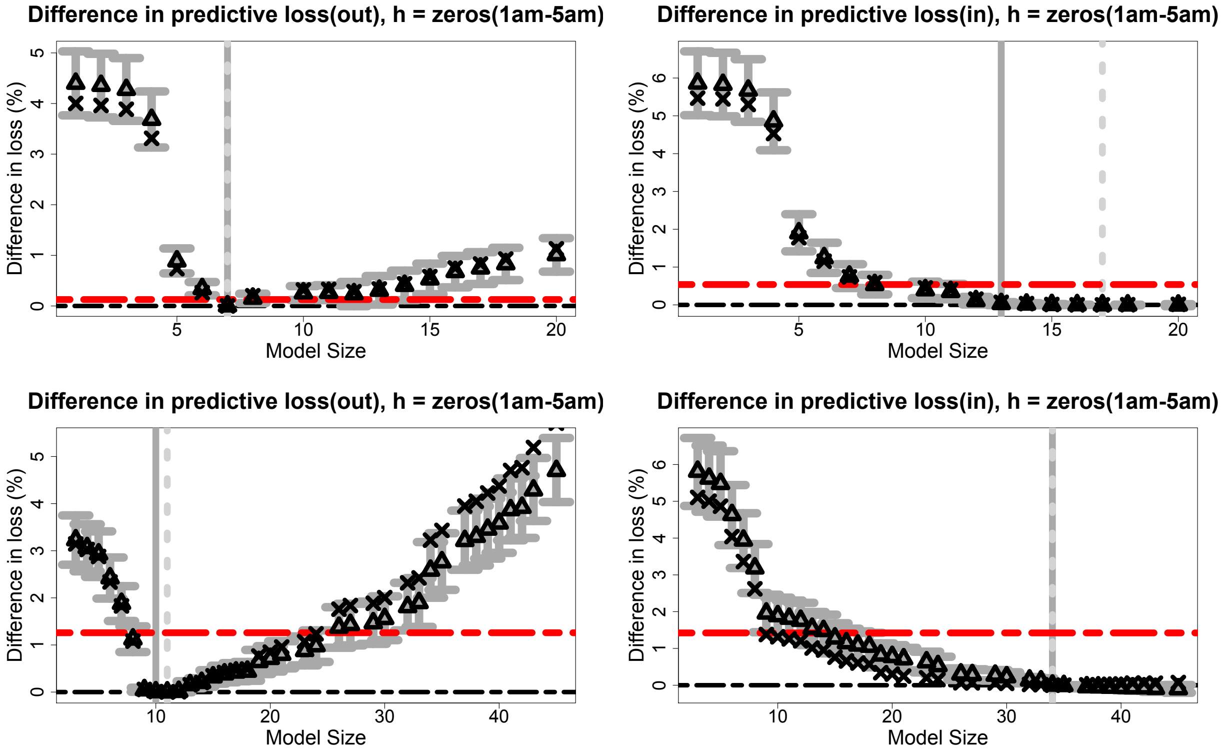


Figure D.5: Approximate out-of-sample (**left**) and in-sample (**right**) cross-entropy loss for sparse linear models without (**top**) and with (**bottom**) interactions targeted to `zeros(1am-5am)`. Results are presented for each size as a percent increase in loss relative to $\mathcal{A}_{min}$. 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).

D.5 Selected covariates and direction of estimated effects

The selected covariates and direction of estimates effects for each functional are provided in Table D.1. These results are based on the simplest acceptable predictor described in the main paper and using the original set of covariates, i.e., excluding interactions and quadratic effects. There is strong consensus for the selected covariates and the directions among `avg`, `tlac`, `sd`, and `max`, while `sedentary` consistently features the opposite sign. Focusing on `sedentary` for conciseness, we note that BMI, age, smoking, diabetes, coronary heart disease, and total cholesterol are positively associated with sedentary behavior (i.e., less physical activity). These results appear to be reasonable. Since `argmax` is not accurately predicted by any of the candidate predictors, the simplest acceptable predictor only selects one variable.

	`avg`	`tlac`	`sd`	`sedentary`	`max`	`argmax`
BMI	-	-	-	+	-	0
Age	-	-	-	+	-	0
Gender: female	-	-	-	+	-	0
Race: Black	-	0	-	+	-	0
Race: Hispanic	+	+	+	-	+	-
Race: Other	0	+	0	-	0	0
Education: HS	0	0	0	-	0	0
Education: more than HS	0	+	0	0	0	0
Smoker: current	-	0	-	+	-	0
Smoker: former	0	-	0	0	0	0
Diabetes	-	-	-	+	-	0
Congestive heart failure	+	0	0	0	0	0
Coronary heart disease	-	-	-	+	-	0
HDL Cholesterol	+	+	+	-	+	0
Total Cholesterol	-	-	0	+	0	0

Table D.1: Signs of the estimated effects for the selected variables for each functional. Continuous variables are italicized; the remaining variables are binary. Baselines for the categorical variables are Race: White, Education: less than high school (HS), and Smoker: never.

E Proofs

Proof (Theorem 1). The posterior predictive expected loss is

$$\begin{aligned}
\mathbb{E}_{[\tilde{\boldsymbol{y}}_1,\ldots,\tilde{\boldsymbol{y}}_{\tilde{n}}|\boldsymbol{y}]}\bar{\mathcal{L}}_\lambda\left[\{h(\tilde{\boldsymbol{y}}_i), g(\tilde{\boldsymbol{x}}_i;\boldsymbol{\delta})\}_{i=1}^{\tilde{n}}\right] &= \mathbb{E}_{[\tilde{\boldsymbol{y}}_1,\ldots,\tilde{\boldsymbol{y}}_{\tilde{n}}|\boldsymbol{y}]}\left[\tilde{n}^{-1}\sum_{i=1}^{\tilde{n}}\mathcal{L}_0\{h(\tilde{\boldsymbol{y}}_i), g(\tilde{\boldsymbol{x}}_i;\boldsymbol{\delta})\} + \lambda\mathcal{P}(\boldsymbol{\delta})\right] \\
&= \tilde{n}^{-1}\sum_{i=1}^{\tilde{n}}\mathbb{E}_{[\tilde{\boldsymbol{y}}_i|\boldsymbol{y}]}\mathcal{L}_0\{h(\tilde{\boldsymbol{y}}_i), g(\tilde{\boldsymbol{x}}_i;\boldsymbol{\delta})\} + \lambda\mathcal{P}(\boldsymbol{\delta}) \\
&= \tilde{n}^{-1}\sum_{i=1}^{\tilde{n}}\mathbb{E}_{[\tilde{\boldsymbol{y}}_i|\boldsymbol{y}]}\left\|h(\tilde{\boldsymbol{y}}_i) - g(\tilde{\boldsymbol{x}}_i;\boldsymbol{\delta})\right\|_2^2 + \lambda\mathcal{P}(\boldsymbol{\delta}).
\end{aligned}$$

Focusing on one expectation term in the summand, we simplify:

$$\begin{aligned}
\mathbb{E}_{[\tilde{\boldsymbol{y}}_i|\boldsymbol{y}]}\left\|h(\tilde{\boldsymbol{y}}_i) - g(\tilde{\boldsymbol{x}}_i;\boldsymbol{\delta})\right\|_2^2 &= \mathbb{E}_{[\tilde{\boldsymbol{y}}_i|\boldsymbol{y}]}\left\|\{h(\tilde{\boldsymbol{y}}_i) - \bar{h}_i\} + \{\bar{h}_i - g(\tilde{\boldsymbol{x}}_i;\boldsymbol{\delta})\}\right\|_2^2 \\
&= \hat{v}_i + \left\|\bar{h}_i - g(\tilde{\boldsymbol{x}}_i;\boldsymbol{\delta})\right\|_2^2
\end{aligned}$$

where $\bar{h}_i \coloneqq \mathbb{E}_{[\tilde{\boldsymbol{y}}_i|\boldsymbol{y}]}h(\tilde{\boldsymbol{y}}_i)$ and $\hat{v}_i \coloneqq \mathbb{E}_{[\tilde{\boldsymbol{y}}_i|\boldsymbol{y}]}\left\|h(\tilde{\boldsymbol{y}}_i) - \bar{h}_i\right\|_2^2$. Since $\hat{v}_i < \infty$ is a constant that does not depend on $\boldsymbol{\delta}$, the result follows immediately. □

Proof (Corollary 1). From Theorem 1, the optimal action for $\mathcal{A}_B = (g(\tilde{\boldsymbol{x}};\boldsymbol{\delta}) = \delta(\tilde{\boldsymbol{x}}), \lambda = 0)$ is

$$\begin{aligned}
\hat{\boldsymbol{\delta}}_{\mathcal{A}_B} &= \arg\min_{\boldsymbol{\delta}}\left\{\tilde{n}^{-1}\sum_{i=1}^{\tilde{n}}\left\|\bar{h}_i - g(\tilde{\boldsymbol{x}}_i;\boldsymbol{\delta})\right\|_2^2 + \lambda\mathcal{P}(\boldsymbol{\delta})\right\} \\
&= \arg\min_{\boldsymbol{\delta}}\left\{\tilde{n}^{-1}\sum_{i=1}^{\tilde{n}}\left\|\bar{h}_i - \delta(\tilde{\boldsymbol{x}}_i)\right\|_2^2\right\}
\end{aligned}$$

which can be minimized (to zero) by setting $\delta(\tilde{\boldsymbol{x}}_i) = \bar{h}_i$ for each $i = 1, \ldots, \tilde{n}$. □

Proof (Corollary 2). When $\mathbb{E}_{[\tilde{\boldsymbol{y}}_i|\boldsymbol{\theta}]}h(\tilde{\boldsymbol{y}}_i) = \tilde{\boldsymbol{x}}_i'\boldsymbol{\theta}$, the Bayes action is

$$\begin{aligned}\bar{h}_i &= \mathbb{E}_{[\tilde{\boldsymbol{y}}_i|\boldsymbol{y}]}h(\tilde{\boldsymbol{y}}_i) \\ &= \mathbb{E}_{[\boldsymbol{\theta}|\boldsymbol{y}]}\mathbb{E}_{[\tilde{\boldsymbol{y}}_i|\boldsymbol{\theta}]}h(\tilde{\boldsymbol{y}}_i) \\ &= \tilde{\boldsymbol{x}}_i'\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\boldsymbol{\theta}.\end{aligned}$$

From Theorem 1, the optimal action for $\mathcal{A}_L = (g(\tilde{\boldsymbol{x}};\boldsymbol{\delta}) = \tilde{\boldsymbol{x}}'\boldsymbol{\delta}, \lambda = 0)$ at the observed design points $\widetilde{\mathcal{X}} = \{\boldsymbol{x}_i\}_{i=1}^n$ is

$$\begin{aligned}\hat{\boldsymbol{\delta}}_{\mathcal{A}_L} &= \arg\min_{\boldsymbol{\delta}}\left\{n^{-1}\sum_{i=1}^n \left\|\bar{h}_i - g(\boldsymbol{x}_i;\boldsymbol{\delta})\right\|_2^2 + \lambda\mathcal{P}(\boldsymbol{\delta})\right\} \\ &= \arg\min_{\boldsymbol{\delta}}\left\{n^{-1}\sum_{i=1}^n \left\|\boldsymbol{x}_i'\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\boldsymbol{\theta} - \boldsymbol{x}_i'\boldsymbol{\delta}\right\|_2^2\right\}\end{aligned}$$

which can be minimized (to zero) by setting $\hat{\boldsymbol{\delta}}_{\mathcal{A}_L} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\boldsymbol{\theta}$. □

Proof (Lemma 1). First, suppose $\mathcal{A} \in \Lambda_{\eta,\varepsilon}$. Then (ℓ,∞) is a $(1-\varepsilon)$ lower interval for $\widetilde{\mathbb{D}}^{out}_{\mathcal{A},\mathcal{A}_{min}}$. Next, let (ℓ,∞) be a lower $(1-\varepsilon)$ interval for $\widetilde{\mathbb{D}}^{out}_{\mathcal{A},\mathcal{A}_{min}}$. If η belongs to the interval, then $\eta > \ell$ and $\mathbb{P}_{\mathcal{M}}\big(\widetilde{\mathbb{D}}^{out}_{\mathcal{A},\mathcal{A}_{min}} < \eta\big) \geq \mathbb{P}_{\mathcal{M}}\big(\widetilde{\mathbb{D}}^{out}_{\mathcal{A},\mathcal{A}_{min}} < \ell\big) = \varepsilon$, which satisfies the criteria of the acceptable model set $\Lambda_{\eta,\varepsilon}$. □

Proof (Theorem A.1). The posterior predictive expected loss simplifies to

$$\mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]}\mathcal{L}_0^{EF}\{h(\tilde{\boldsymbol{y}}), g(\tilde{\boldsymbol{x}};\boldsymbol{\delta})\} = F_0\{g(\tilde{\boldsymbol{x}};\boldsymbol{\delta})\} - \mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]}T_0\{h(\tilde{\boldsymbol{y}})\} - \sum_{j=1}^p F_j\{g(\tilde{\boldsymbol{x}};\boldsymbol{\delta})\}\mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]}T_j\{h(\tilde{\boldsymbol{y}})\}$$

by linearity of expectation. Since $\mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]}T_0\{h(\tilde{\boldsymbol{y}})\}$ does not depend on $\boldsymbol{\delta}$ and $\mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]}|T_0\{h(\tilde{\boldsymbol{y}})\}| < \infty$, the minimizer of the posterior predictive expected loss is invariant to this term. It follows that

$$
\begin{aligned}
\hat{\boldsymbol{\delta}}_{\mathcal{A}} &\coloneqq \arg\min_{\boldsymbol{\delta}} \mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]} \mathcal{L}_0^{EF}\{h(\tilde{\boldsymbol{y}}), g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})\} \\
&= \arg\min_{\boldsymbol{\delta}} \left[F_0\{g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})\} - \sum_{j=1}^{p} F_j\{g(\tilde{\boldsymbol{x}}; \boldsymbol{\delta})\}\overline{T_j}\right]
\end{aligned}
$$

with $\overline{T_j} \coloneqq \mathbb{E}_{[\tilde{\boldsymbol{y}}|\boldsymbol{y}]} T_j\{h(\tilde{\boldsymbol{y}})\}$. □

References

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.

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

Huggins, J. H., Campbell, T., Kasprzak, M., and Broderick, T. (2018). Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach. *arXiv preprint arXiv:1809.09505*.

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

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

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.

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500):1518–1532.

Tran, M. N., Nott, D. J., and Leng, C. (2012). The predictive Lasso. *Statistics and Computing*, 22(5):1069–1084.