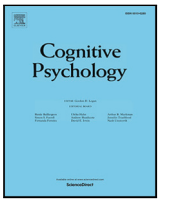
Cognitive Psychology 132 (2022) 101444

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych



Robust priors for regularized regression

Sebastian Bobadilla-Suarez [a,*], Matt Jones [b], Bradley C. Love [a,c]

[a] Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, UK
[b] Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO 80309-0345, USA
[c] The Alan Turing Institute, 96 Euston Road, London, NW1 2DB, UK

ARTICLE INFO

Dataset link: https://osf.io/fg4p5/, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), https://github.com/bobaseb/robustpriorsregression, https://github.com/bobaseb/rsatoolboxlss/tree/develop/LSSproject

Keywords:
Decision making
fMRI
Heuristics
Inductive bias
Inference
Robust priors

ABSTRACT

Induction benefits from useful priors. Penalized regression approaches, like ridge regression, shrink weights toward zero but zero association is usually not a sensible prior. Inspired by simple and robust decision heuristics humans use, we constructed non-zero priors for penalized regression models that provide robust and interpretable solutions across several tasks. Our approach enables estimates from a constrained model to serve as a prior for a more general model, yielding a principled way to interpolate between models of differing complexity. We successfully applied this approach to a number of decision and classification problems, as well as analyzing simulated brain imaging data. Models with robust priors had excellent worst-case performance. Solutions followed from the form of the heuristic that was used to derive the prior. These new algorithms can serve applications in data analysis and machine learning, as well as help in understanding how people transition from novice to expert performance.

1. Introduction

Inference from data is most successful when it involves a helpful inductive bias or prior belief. Regularized regression approaches, such as ridge regression, incorporate a penalty term that complements the fit term by providing a constraint on the solution, akin to how Occam's razor favors solutions that both fit the observed data and are simple. By incorporating such constraints or prior beliefs, the hope is that models will better predict future outcomes.

What makes a good prior belief or inductive bias? In the case of ridge regression, the norm of the regression coefficients is shrunk toward zero (Hoerl & Kennard, 1970; Tikhonov, 1943) to control model complexity and reduce overfitting. However, in many domains, zero is not a reasonable *a priori* guess for the true association between variables. For example, it would be strange to a priori predict the quality of a new home would be unaffected by the experience of the workers, quality of materials, reputation of the architect, etc. Because the world is somewhat predictable, a prior centered on the origin (Fig. 1) is inappropriate.

If not zero, where does one turn for a useful prior? One answer is to look to human behavior. Humans use an assortment of clever strategies for learning and decision-making that perform well even in conditions of low knowledge. Simple heuristics that are fast and frugal (Czerlinski, Gigerenzer, & Goldstein, 1999) excel when training examples are scarce (Parpart, Jones, & Love, 2018). People can also shift to more complex strategies when resources are available (Rieskamp & Otto, 2006). With increasing experience and expertise, humans often acquire a sophisticated understanding of domains.

Although heuristics are efficient and robust models in their own right, we propose they are a useful starting point or prior for more complete characterizations of domains. Advantages of heuristics include their ecological validity (Czerlinski et al., 1999;

* Corresponding author.
E-mail address: sebastian.suarez.12@ucl.ac.uk (S. Bobadilla-Suarez).

https://doi.org/10.1016/j.cogpsych.2021.101444
Received 27 April 2021; Received in revised form 26 October 2021; Accepted 29 October 2021
Available online 30 November 2021

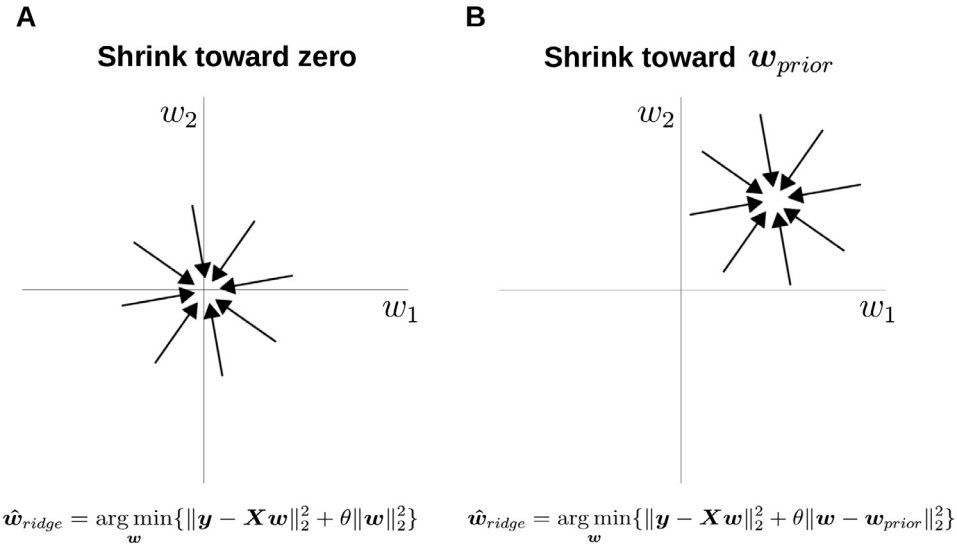


Fig. 1. Shrinking toward zero and non-zero priors. (A) Ridge regression shrinks weights toward $\vec{0}$ (Zero prior) in contrast to (B) a model using a prior based on the TAL heuristic where all weights are equal and non-zero (i.e., $w_{prior} \neq \vec{0}$). Using priors based on a heuristic (i.e., a constrained model) can increase robustness and interpretability. Eq. (1) is shown at the bottom of panel B. The other equation simply drops w_{prior}, equivalent to standard notation for ridge regression, here termed the Zero prior model.

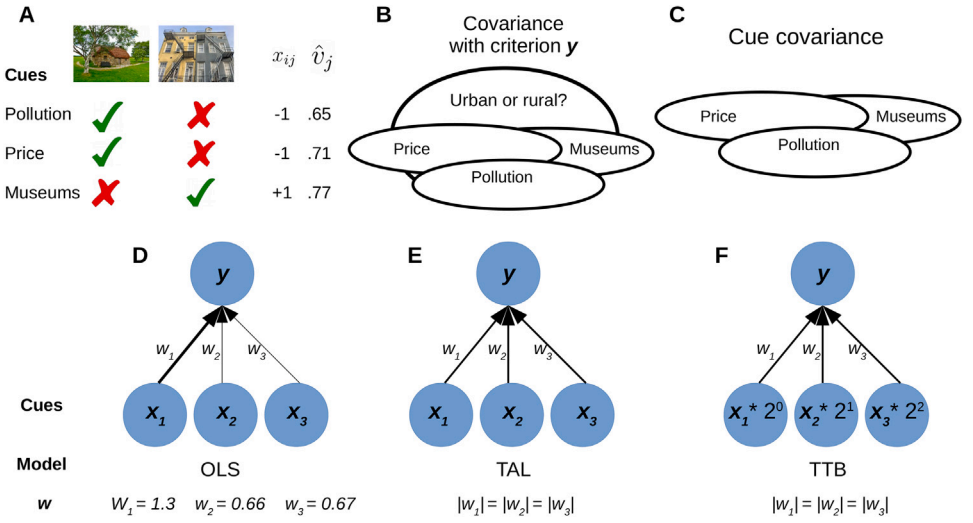


Fig. 2. TAL and TTB decision heuristics. (A) A hypothetical decision — choosing between a rural (−1) or urban (+1) home based on cues ordered by cue validity: low pollution, low price, and proximity to museums. Each cue is coded as −1 when favoring the left option (rural), +1 for the right option (urban), and 0 when the two options are equal on that cue. TAL sums cue values, choosing the rural home. TTB chooses based on the best cue (measured by $\hat{v}_j$, see Eq. (4)) that distinguishes the options. Here, TTB would choose the urban home based solely on proximity to museums. (B) The covariance of the cues with the criterion y (urban or rural home), which $\hat{v}_j$ measures. (C) The covariance of the cues with one another; TAL and TTB heuristics disregard this information. (D–F) Illustrations of OLS, TAL and TTB. Here, OLS strikes a balance for correlated cues; low price and proximity to museums are (negatively) correlated. Thus, low pollution presents a higher weight w_1. TAL and TTB equate the absolute value of all weights. TTB additionally ranks and scales predictors according to their predictive value $\hat{v}_j$ in a non-compensatory way (multiplying cues by powers of 2).

Mata et al., 2012) and robustness across decision problems. Their weakness is insensitivity to aspects of the data due to their rigid inductive bias (Geman, Bienenstock, & Doursat, 1992; Parpart et al., 2018). This weakness is ameliorated when heuristics function as priors within more complex models because priors can be overcome by additional data, much like how human experts develop more complex and nuanced knowledge with increasing experience in a domain. When data are abundant, the encompassing model would master the subtleties of the domain, whereas when data are scarce the heuristic prior would help guide predictions and increase robustness. Because the heuristics themselves are interpretable models, the solution of the encompassing model could be understood in terms of deviations from the heuristic prior.

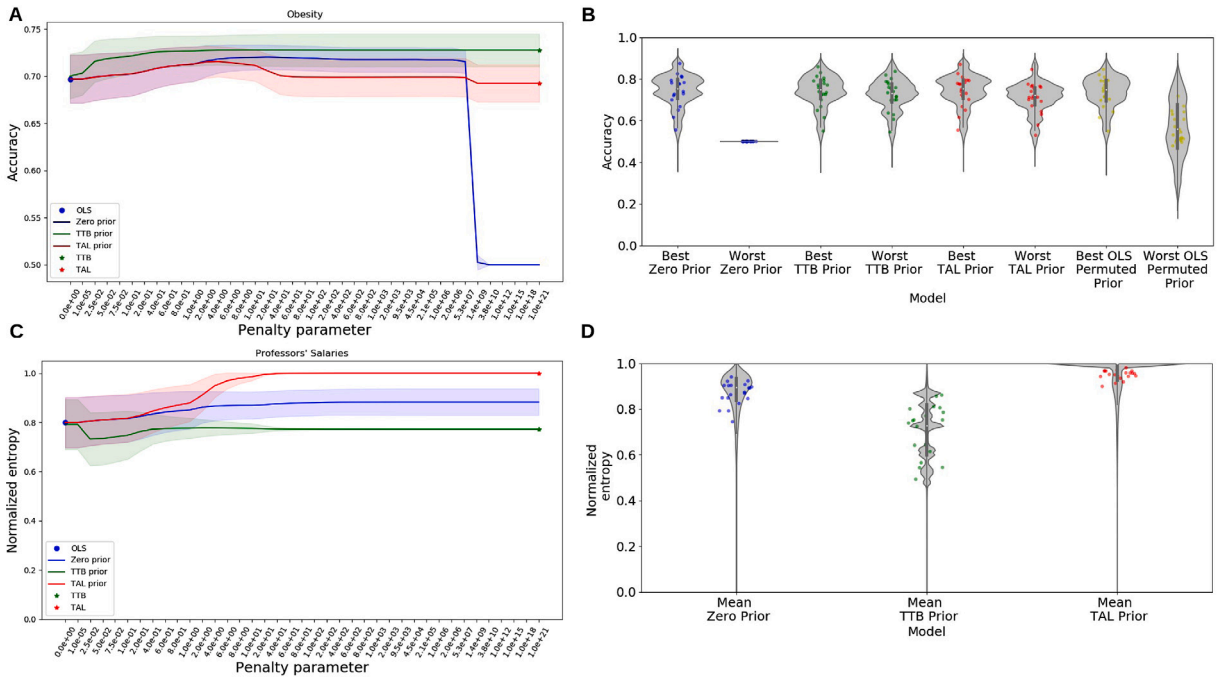


Fig. 3. Accuracy and normalized entropy for the 20 datasets in Application 1. Training set size was fixed at 50. (A) Test set accuracy across penalty values for the Obesity dataset. At low penalties, the models all agree with OLS (unpenalized). Under strong penalties, the Zero prior model (standard ridge regression) converges to chance performance as weights shrink toward $\vec{\mathbf{0}}$. TAL-prior and TTB-prior models converge toward their respective heuristics and are robust. (B) Test set accuracy for all 20 datasets for best and worst performing penalty values for each model (see SI). The OLS permuted prior model is a penalized regression model with a permuted OLS solution as prior. Heuristic prior models are most robust. (C) Normalized entropy (Eq. (17)) for the Professors' Salaries dataset, which measures how compensatory the weights are. The TAL-prior model becomes maximally compensatory as penalty increases, unlike the TTB-prior model. (D) Normalized entropy for all 20 datasets across all penalty values, which orders as TAL-prior, Zero prior, TTB-prior. For B and D, violins represent density estimates for the respective metric. Each dot is one of 20 datasets. For A and C, shaded areas represent 1 standard deviation.

2. Robust priors based on decision-making heuristics

We used two well-known heuristics, tallying (TAL) and take-the-best (TTB) (Czerlinski et al., 1999), as priors in regularized regression models. These heuristics predict which of two options is preferable. For example, TAL or TTB could predict whether a rural or urban home is preferable based on several cues (Fig. 2A). Each cue is ternary valued and indicates whether the left (−1) or right (+1) option is preferred on that dimension, with 0 for a tie. TAL is a simple majority voting rule whereas TTB bases its decision on the single most predictive cue that can discriminate between the alternatives (both heuristics explained below; also, see Fig. 2).

To use a heuristic as a prior, we use a two-step model-fitting procedure (cf. (Zou, 2006)). In step 1, we fit the heuristic to the training data. The resulting point estimate for the weight vector provides the penalty term (weighted by the penalty parameter θ) within a regularized regression model in step 2. The penalty term shrinks regression coefficients toward the heuristic solution, as opposed to $\vec{\mathbf{0}}$ as in ridge regression (Fig. 1). Increasing θ increases the strength of the prior, eventually pushing the regression solution to fully agree with the heuristic (cf. (Parpart et al., 2018)).

Our approach integrates heuristics with full-information (regression) models in a principled way that applies to a broad class of heuristics. The approach is to subtract a carefully constructed vector inside the penalty term of the well-known $L2$ cost function used in ridge regression. The cost function for standard ridge regression is

$$\hat{\boldsymbol{w}}_{ridge} = \arg\min_{\boldsymbol{w}} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \theta\|\boldsymbol{w} - \boldsymbol{w}_{prior}\|_2^2\}, \tag{1}$$

where $\|\cdot\|_2$ is the Euclidean (L_2) norm, $\boldsymbol{y}$ is the dependent variable $[y_1, \ldots, y_n]^T$, $\boldsymbol{X}$ is an $n \times m$ matrix with one column for each of the m predictor variables $\boldsymbol{x}_j$, $\boldsymbol{w}_{prior}$ is a column vector of zeros $\vec{\mathbf{0}} = [0_1, \ldots, 0_m]^T$, $\boldsymbol{w}$ is a vector of estimated regression coefficients $[w_1, \ldots, w_m]^T$, and $\theta \geq 0$ is a tunable penalty parameter.[1] Note that implicit priors are also found in regularized regression with all other norms (L_n) too, including LASSO regression (L_1). Thus, our insight generalizes to all other norms as well.

[1] Arrows over numerals and lowercase boldface are for vectors whereas uppercase boldface is for matrices or tensors. Carets are reserved for estimates, tildes for normalized scalars and overbars for arithmetic means.

The first term inside the arg min of Eq. (1) promotes goodness-of-fit in the model, whereas the second term – known as the penalty term – promotes smaller weights $\boldsymbol{w}$. As θ increases, the weights tend to $\boldsymbol{w}_{prior}$ in the limit (Fig. 1). (The derivation of the optimal weights $\boldsymbol{w}^*_{ridge}$ is included in the SI.) However, when $\theta = 0$, the model is equivalent to ordinary least squares (OLS) regression. OLS regression estimates coefficients without the penalty term:

$$\hat{\boldsymbol{w}}_{OLS} = \arg\min_{\boldsymbol{w}}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2\} \tag{2}$$

Normally, $\boldsymbol{w}_{prior}$ is not included in Eq. (1); it is implicit in more standard specifications of ridge regression, where the penalty term is written simply as $\theta\|\boldsymbol{w}\|_2^2$. Nevertheless, instead of a $\vec{\boldsymbol{0}}$ vector, one can generalize ridge regression with alternative constructions of $\boldsymbol{w}_{prior}$ (Fig. 1).

As argued above, choosing this vector intelligently might improve learning of $\hat{\boldsymbol{w}}_{ridge}$ by imposing a more sensible inductive bias (Geman et al., 1992). Although the decision-making literature has traditionally proposed that humans use certain classes of heuristics due to cognitive limitations (Bobadilla-Suarez & Love, 2018; Kahneman, Slovic, & Tversky, 1974; Simon, 1978), heuristics have also been justified from their ecological validity (Czerlinski et al., 1999; Dawes, 1979; Parpart et al., 2018). That is, the inductive biases they embody agree with the statistical structure of many natural environments, thus leading to better performance. Taking inspiration from the TAL and TTB heuristics, both in their success in describing human decision making (Bobadilla-Suarez & Love, 2018; Bröder, 2003; Newell & Shanks, 2003; Otworowska, Blokpoel, Sweers, Wareham, & van Rooij, 2018) and in application to real world statistical problems (Czerlinski et al., 1999; Parpart et al., 2018), we propose a construction of $\boldsymbol{w}_{prior}$ based on these heuristics.

Below we discuss how to construct priors from TAL and TTB. We then report three applications. In the first application, we compared generalization performance (test set accuracy) and interpretability of model solutions on 20 classical datasets previously used in the decision-making literature (Czerlinski et al., 1999; Parpart et al., 2018). There, the decision-making problem was to choose the better item within a pair (see Fig. 2A and below). In the second application, we evaluated our approach within a classification paradigm in which a single item is assigned to one of two classes (e.g., friend or foe?). In the final application, we demonstrated the generality and benefits of our approach by analyzing simulated brain imaging data where the prior is derived from a technique (Mumford, Turner, Ashby, & Poldrack, 2012) that seeks to minimize collinearity amongst predictors in a manner that parallels how we derive heuristic priors.

2.1. TAL and TTB heuristics

TAL and TTB do not adapt their form or complexity in light of the data. For example, TAL is an equal-weights algorithm that uses only the signs of the coefficients (Czerlinski et al., 1999; Dawes, 1979): the estimated weights $\hat{w}_j$ are constrained to be either -1 or 1 (Fig. 2E).

The Tallying decision rule (TAL) is defined as

$$\hat{y}_i = \text{sign}\left(\sum_j \text{sign}(\hat{v}_j x_{ij})\right) \tag{3}$$

$$\text{where } \hat{v}_j = \frac{R - W}{R + W} \in [-1, 1] \tag{4}$$

A cue's estimated cue validity $\hat{v}_j$ is defined as the difference between the numbers of correct predictions R and incorrect predictions W, divided by the total number of predictions across all observations $R + W$ (Martignon, Hoffrage, et al., 1999).[2] Observations that do not present a prediction (i.e., $x_{ij} = 0$) are ignored. Notably, cue validities depend only on the relationship between each cue and the outcome, and not on covariance between cue. Thus, the definition of $\hat{v}_j$ is what makes heuristics insensitive to cue covariance. When assessing model performance, the validities $\hat{v}_j$ are estimated for each training set.

The Take-the-Best (TTB) decision rule is

$$\hat{y}_i = \text{sign}(x_{ij^*}) \tag{5}$$

$$\text{where } j^* = \arg\max_j \left\{|\ \hat{v}_j\ |\colon x_{ij} \neq 0\right\} \tag{6}$$

Whereas TAL sums the signs of the predictors to determine its response, TTB relies on the top predictor that differentiates the two options. When there is no evidence for either option, both TAL and TTB choose randomly (with probability $p = 0.5$ for each option). This occurs for TAL when Eq. (3) yields 0 and for TTB when every x_{ij} equals 0.

We now define $\boldsymbol{w}_{prior}$ based on the TAL heuristic, referred to as $\boldsymbol{w}_{TAL\ prior}$. First, we determine a scalar coefficient $\hat{w}_{TAL\ scale}$ shared across all predictor variables $\boldsymbol{x}_j$:

$$\hat{w}_{TAL\ scale} = \arg\min_w\{\|\boldsymbol{y} - \boldsymbol{X}(\hat{\boldsymbol{q}}w)\|_2^2\}. \tag{7}$$

[2] Previous literature has defined cue validities as $R/(R+W)$. The definition used here is a linear transformation that simplifies description of the models.

This equation is the same as Eq. (2) except that the vector $\boldsymbol{w}$ has been replaced by the product of a scalar w and a column vector $\hat{\boldsymbol{q}}$, which has cue directionalities $\hat{q}_j = \text{sign}(\hat{v}_j)$. Using this scalar, we define:

$$\boldsymbol{w}_{TAL\ prior} = \hat{\boldsymbol{q}}\hat{w}_{TAL\ scale}. \tag{8}$$

To construct $\boldsymbol{w}_{TTB\ prior}$, we build from the intuition that TTB is equivalent to a noncompensatory weight vector such as $2^{\hat{\boldsymbol{r}}}$, where $\hat{\boldsymbol{r}}$ is a vector of ascending ranks for the absolute cue validities, $\hat{r}_j = |\{j' : |\hat{v}_{j'}| < |\hat{v}_j|\}|$. Paralleling the definition of the $\boldsymbol{w}_{TAL\ prior}$, for the $\boldsymbol{w}_{TTB\ prior}$, we also determine a shared scalar:

$$\hat{w}_{TTB\ scale} = \arg\min_{\boldsymbol{w}}\{\|\boldsymbol{y} - \boldsymbol{X}_{TTB}(\hat{\boldsymbol{q}}w)\|_2^2\} \tag{9}$$

$$\boldsymbol{w}_{TTB\ prior} = \hat{\boldsymbol{q}}\hat{w}_{TTB\ scale}. \tag{10}$$

However, we have a new design matrix $\boldsymbol{X}_{TTB}$ in Eq. (9), defined by

$$\boldsymbol{X}_{TTB} = \boldsymbol{X}\hat{\boldsymbol{P}} \tag{11}$$

with $\hat{\boldsymbol{P}}$ being a diagonal matrix

$$\hat{\boldsymbol{P}} = diag(\phi^{\hat{\boldsymbol{r}}}), \quad \text{where } \phi := 2. \tag{12}$$

This transformation has the effect of encoding cue validity directly into the design matrix by scaling each regressor according to a geometric progression. In order for $\boldsymbol{w}_{TTB\ prior}$ to function appropriately as a $\boldsymbol{w}_{prior}$, the original design matrix $\boldsymbol{X}$ is replaced with $\boldsymbol{X}_{TTB}$ in Eq. (1). When instead $\phi := 1$, we recover the TAL prior. Note also that this entire procedure is nearly equivalent to working with the original design matrix $\boldsymbol{X}$ and taking $\boldsymbol{w}_{TTB\ prior}$ to be proportional to $\hat{\boldsymbol{P}}\hat{\boldsymbol{q}}$ (the vector of signed exponentiated ranks, $\hat{q}_j\phi^{\hat{r}_j}$) rather than to $\hat{\boldsymbol{q}}$, except that the weights are differentially penalized according to their ranks.

With these priors defined, we can now formally specify two regularized regression models. The TAL-prior model is defined by Eq. (1) with $\boldsymbol{w}_{prior} = \boldsymbol{w}_{TAL\ prior}$. The TTB-prior model is defined by Eq. (1) with $\boldsymbol{w}_{prior} = \boldsymbol{w}_{TTB\ prior}$ and with $\boldsymbol{X}$ replaced by $\boldsymbol{X}_{TTB}$.

In contrast to OLS, the use of the common scalar for all cues in the prior for both TAL-prior and TTB-prior highlights that both heuristics are insensitive to cue covariance information (see Fig. 2E,F). For TAL-prior, the common scalar reflects the fact that TAL is a (fully) compensatory strategy, whereas the design matrix in TTB-prior, $\boldsymbol{X}_{TTB}$, reflects the fact that TTB is a non-compensatory strategy. Later, we will evaluate how these differing priors affect the nature of penalized regression solutions.

2.1.1. Logistic ridge regression

The first two applications reported here use logistic ridge regression (Le Cessie & Van Houwelingen, 1992; Schaefer, Roi, & Wolfe, 1984; van Wieringen, 2015). To estimate weights for penalized logistic regression, $\hat{\boldsymbol{w}}_{log(ridge)}$, we first obtain a scale parameter for an unpenalized logistic regression via maximum likelihood, where as above the weight vector is constrained to be proportional to the cue directionalities:

$$\hat{w}_{log\ scale} = \arg\max_{w} \Pr[\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w} = w\hat{\boldsymbol{q}}]. \tag{13}$$

The likelihood for logistic regression is as usual:

$$\Pr[y_i|\boldsymbol{X}_{i\cdot}, \boldsymbol{w}] = \begin{cases} \frac{1}{1+\exp(\boldsymbol{X}_{i\cdot}\boldsymbol{w})}, & y_i = 0 \\ \frac{1}{1+\exp(-\boldsymbol{X}_{i\cdot}\boldsymbol{w})}, & y_i = 1, \end{cases} \tag{14}$$

where $\boldsymbol{X}_{i\cdot}$ denotes the ith row of $\boldsymbol{X}$. We then define $\boldsymbol{w}_{log\ prior}$ as

$$\boldsymbol{w}_{log\ prior} = \hat{\boldsymbol{q}} \mid \hat{w}_{log\ scale} \mid . \tag{15}$$

We then insert $\boldsymbol{w}_{log\ prior}$ into our final objective function for regularized logistic regression:

$$\hat{\boldsymbol{w}}_{log(ridge)} = \arg\max_{\boldsymbol{w}} \left\{ \Pr[\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}] - \tfrac{1}{2}\theta\|\boldsymbol{w} - \boldsymbol{w}_{log\ prior}\|_2^2 \right\}. \tag{16}$$

See Supplementary Information (SI) for an approximation of $\hat{\boldsymbol{w}}_{log(ridge)}$ for regularized logistic regression via the Newton–Raphson iterative algorithm.

3. Application I: Heuristic decision making

Regularized regression models with heuristic priors were evaluated on the 20 datasets that have been previously used to compare heuristics with ordinary least squares (OLS) regression (Czerlinski et al., 1999; Katsikopoulos, Schooler, & Hertwig, 2010; Parpart et al., 2018). For each of the 20 problems, the cues for the two options on each trial were binary valued (see Methods below for more details), which leads to ternary-valued inputs according to our coding scheme (see Fig. 2A).

3.1. Methods

The preprocessed data were retrieved from an Open Science Foundation (OSF) repository (Parpart, Jones, & Love, 2019), used to evaluate the half-ridge and COR models by Parpart et al. (2018). In accord with previous research, cue attributes were dichotomized by median split (Czerlinski et al., 1999; Parpart et al., 2018).

The data were transformed into a format appropriate for decision-making problems where all pairwise comparisons between observations were encoded as the signed differences in (binary) attributes (possible values: -1, 0, and $+1$). The decision in our coding scheme is -1 for the left choice and $+1$ for the right choice, which was mapped to 0 and 1 for logistic regression: e.g., in the homestead example (Fig. 2A), rural is coded as -1 (or 0 for logistic models) and urban is coded as $+1$. This is common procedure in the decision-making literature (Czerlinski et al., 1999; Katsikopoulos et al., 2010; Parpart et al., 2018). Formally, this consists of training pairs $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ with $\boldsymbol{x}_i \in \{-1, 0, 1\}^m$. Training sets consisted of 50 training pairs from which the priors were learned from. All results were computed for 1000 iterations (i.e., different partitions into training and test sets) for all penalty values.

As the penalty parameter θ increases, the penalized regression models with the $\boldsymbol{w}_{TAL\ prior}$ and $\boldsymbol{w}_{TTB\ prior}$ converge to their corresponding heuristics (Fig. 3A). As a sanity check, the TAL-prior and TTB-prior models were validated on simulated data in the SI (Figures S1 and S2, respectively) by tracking their agreement with OLS predictions. Effectively, agreement with OLS is higher for low penalty values and agreement with TAL-prior or TTB-prior is higher for high penalty values. Individual plots for each of the twenty datasets are also included in the SI (Figures S3 and S4).

Although regression models are interpretable in that each feature's importance follows from its weight, the heuristic penalty terms make clear how the prior shapes the solution and how the solution differs from the prior, which itself is an interpretable solution. To evaluate how the form of the solution changes as a function of the prior, we calculated normalized Shannon entropy defined as:

$$\tilde{H} = \frac{-\sum_j \iota\tilde{w}_j log_2 \iota\tilde{w}_j}{-\sum_j \frac{1}{m} log_2 \frac{1}{m}} \tag{17}$$

where $\iota\tilde{w}_j$ is

$$\iota\tilde{w}_j = \frac{|\iota\hat{w}_j|}{\|\hat{\boldsymbol{w}}\|_1} \phi^{\hat{r}_j}, \tag{18}$$

$\phi := 1$ for TAL-prior and $\phi := 2$ for TTB-prior, and $\|\cdot\|_1$ is the L_1 norm, such that $\tilde{H} \in [0, 1]$ for any number (m) of predictors. Eq. (17) provides an intuitive measure of how compensatory a solution is. The measure will peak at 1 when the predictive force of the weights is uniform, as in the TAL heuristic.

3.2. Results

As predicted, these models are robust across the range of θ values because they converge to a reasonable estimate (i.e., a sensible heuristic). In contrast, while ridge regression performs well overall, its performance suffers at higher penalty values as its weights are pulled toward $\vec{\mathbf{0}}$. The robustness of the penalized regression models with heuristic priors held across the 20 datasets (Fig. 3B). Notice that regularization using any nonzero prior is not sufficient for robustness — an ad hoc nonzero prior (OLS Permuted Prior) was not robust. The OLS permuted prior model is a penalized regression model with a permuted OLS solution as prior (i.e., where the weights from the OLS solution have been permuted).

We confirmed that $\tilde{H}$ for the TAL-prior model would converge to 1 with increasing penalty θ, in contrast to the TTB-prior model. We also predicted ridge regression's $\tilde{H}$ would be somewhat lower than TAL-prior's. That is, convergence to a $\vec{\mathbf{0}}$ weights vector for standard ridge regression is nonsensical and effectively resisted in the optimization, providing more heterogeneous weights than otherwise expected. These predictions held (Figs. 3C,D).

The results presented here hold under an alternative training scheme where we also evaluate OLS as a prior itself (*Splitting training data* in the SI). OLS performs worse as a prior on the majority of datasets (Figure S7 in the SI) and also shows higher variance overall (Figure S8 in the SI).

In these 20 decision problems, models using priors based on TAL and TTB were robust across the entire range of prior strengths. These penalized regression models shrunk to a reasonable prior based on a simple heuristic that discards covariance information amongst predictive cues. The forms of the solutions were interpretable and followed from the priors.

4. Application II: Breast Cancer classification

In this application, we conducted the same analyses as in Application I, but for a classification problem as opposed to a forced choice between two options. We applied the models to the Breast Cancer Wisconsin (Diagnostic) Data Set from the UCI data repository (Blake, Keogh, & Merz, 1995). In this task, models predicted whether an item was cancerous or not based on binary features (see Methods for more details). The predictors were discrete as in Application I, though the identical approach would apply to continuous predictors or to a mixture of discrete and continuous predictors.[3]

[3] Our approach easily generalizes to the continuous setting by placing predictors on the same scale through some form of normalization (e.g., z-scoring).

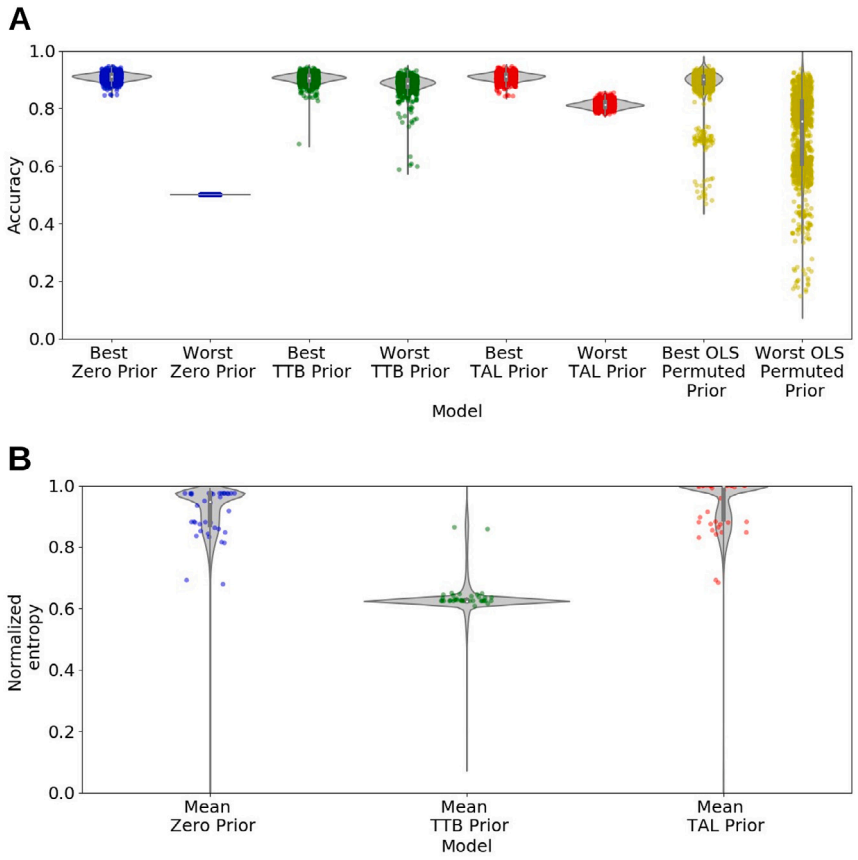


Fig. 4. Generalization performance and normalized entropy for Breast Cancer Wisconsin (Diagnostic) Data Set after training on 100 items (Application II). The same models are considered here as in Application I (see Fig. 3). (A) Test set accuracy for the best and worst performing penalty value for each model (see SI). The key finding is that models with heuristic priors are most robust. (B) Normalized entropy (Eq. (17)) averaged across the range of penalty values reflects how compensatory a model's predictions are, led by TAL-prior, followed by the Zero prior, and finally the TTB-prior model. Each dot represents one of the tested penalty values averaged over 1000 train–test splits. The gray violins represent the respective density estimates in both panels.

4.1. Methods

The data comprises nine cues ($m = 9$) that describe characteristics of the cell nuclei present in digitized images of fine needle aspirate (FNA) of breast masses (Blake et al., 1995). Data points with missing cue values were removed, resulting in a total of $n = 478$ observations. All variables were binarized by median split.

In an analogous fashion to how we constructed $\boldsymbol{X}$ in Application I, here we transformed the original data by median splits. For each cue, if the value was equal to the median, it received a value of 0, if it was above the median it was equal to $+1$ and if it was below the median it was equal to -1. Formally, this also produces training pairs $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ with $\boldsymbol{x}_i \in \{-1, 0, 1\}^m$. Training sets consisted of 100 training pairs from which the priors were learned from. However, we did not construct a matrix of pairwise comparisons of observations as before. The dependent variable was binary, $\boldsymbol{y} \in \{-1, +1\}$, coding for malignant and benign tumors, respectively. This preprocessing of the data $\boldsymbol{X}$ is closer to the way regression models are calculated for everyday applications. Both the mean test accuracy (Fig. 4A and Figure S5 in the SI) and the mean normalized entropy (Fig. 4B and Figure S6 in the SI) were averaged over 1000 iterations for each penalty value.

4.2. Results

The results were in accord with Application I. The models with a heuristic prior were robust across the range of θ values (Fig. 4A). As in Application I, the priors shaped the form of the solution in the predicted manner (Fig. 4B) with the TAL-prior model having the most compensatory solutions.

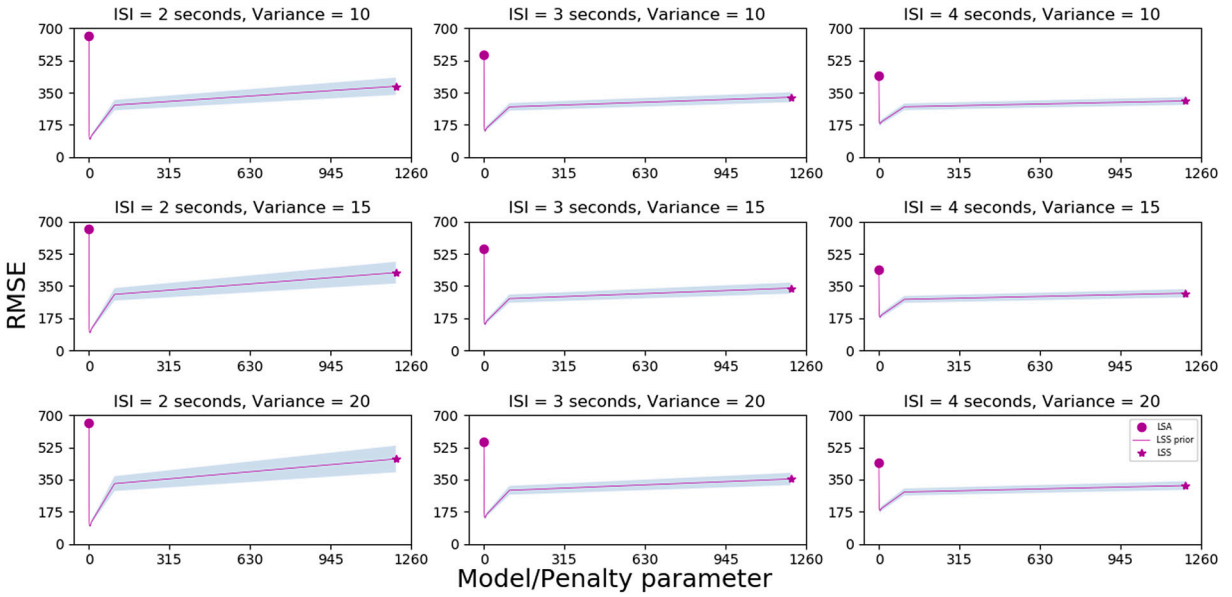


Fig. 5. The fMRI simulation results. Root mean squared error averaged over voxels and simulations ($\overline{\text{RMSE}}$) between estimated regression weights $\hat{\boldsymbol{w}}$ of each model and the true data-generating weights $\boldsymbol{\Psi}$ (see Methods) for different interstimulus intervals (ISI) and variances (signal-to-noise ratio, SNR: σ^2_{Ψ}, see Methods). The 3×3 design is presented such that each row displays a different level of variance while each column displays a different ISI. Our LSS-prior model is equivalent to the LSA model (purple circle) when $\theta = 0$ and the LSS model (purple star) when the θ penalty parameter is large. Shaded areas represent three standard deviations above and below the estimate. Lower $\overline{\text{RMSE}}$ values convey superior performance. The key finding is our penalized regression model is superior to the standard LSA model and the LSS model (which serves as the prior) for moderate prior strength θ. For all panels, the results for all penalty values (in each model and dataset) are averaged over 1000 train–test splits. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Application III: Estimation in brain imaging analyses

In Applications I and II, the task was to generalize from training items to make decisions about test items. In Application III, the objective was to estimate the weights themselves. We considered simulated functional magnetic resonance imaging (fMRI) time series that allowed for comparing estimates to ground truth.

Brain imaging datasets are challenging to analyze because they measure the brain's hemodynamic response, which is a temporal and spatially autocorrelated, high-dimensional, noisy, and time-lagged signal. The signal is composed of thousands of voxels (voluminous pixels) with coordinates in space (x, y, z) and time (n time-points). Correlations across space and time due to psychological (e.g., Visscher, Kahana, and Sekuler (2009)), neurovascular (Boynton, Engel, & Heeger, 2012) and physical (Smith et al., 1999) effects complicate the independence and linearity assumptions used to model the signal in each voxel. Furthermore, the observed blood-oxygen-level dependent (BOLD) signal is only indirectly related to the outcome variable of interest (neural activity), via the hemodynamic response function (HRF), which is normally modeled as a double gamma function.

In task fMRI, the BOLD time series for a voxel is modeled by weighting events, such as a sequence of pictures (e.g., dog, truck, face, etc.) presented to a study participant. In addition to nuisance regressors, one typically estimates a beta weight for each event (convolved with the HRF). We refer to this standard method as least squares all (LSA), which is unpenalized and plays a role analogous to OLS in Applications I and II.

However, for the reasons discussed above, collinearity in the time series can compromise parameter estimation (Mumford, Poline, & Poldrack, 2015), particularly in rapid event designs (e.g., trial duration of one or two seconds). One proposed solution, which we refer to as least squares separate (LSS), is to estimate a separate model for each event rather than a single model for all events (Rissman, Gazzaley, & D'esposito, 2004). Each model estimates one beta weight for the target event (i.e., trial) and a second shared beta weight for all other events (Turner, 2010). In practice, LSS produces better (less variable) estimates by being less sensitive to collinearity in the time series (Mumford et al., 2012).

We view LSS as analogous to the heuristics considered in Applications I and II. The TAL and TTB heuristics are insensitive to cue covariance. Specifically, cue validity and cue direction are estimated individually for each predictor. Moreover, we implemented these heuristics in a regression framework with a single beta weight (e.g., $\hat{w}_{TAL\ scale}$) to derive a prior. In both models, simplification is achieved by forcing multiple predictors to share a single regression weight. Analogously, each LSS model forces all but the target event (out of potentially hundreds of events) to share a common beta weight.

Like TAL and TTB, we predicted that LSS would provide an effective prior for a penalized regression model because it provides a reasonable and robust starting point to move from when the data warrant. We predicted that a penalized regression model with an LSS prior would outperform both LSS (high θ) and the LSA approach ($\theta = 0$).

5.1. Methods

To build a continuum of models between LSA and LSS, we include the weights derived from LSS as a target (i.e., prior) in the penalty term within a regularized LSA model. Thus, the weights from the LSS-prior model are estimated with the following objective:

$$\arg\min_{\boldsymbol{w}}\{\|\boldsymbol{y}-\boldsymbol{X}_{LSA}\boldsymbol{w}\|_2^2+\theta\|\boldsymbol{w}-\boldsymbol{w}_{LSS\ prior}\|_2^2\} \tag{19}$$

Paralleling our treatment of the decision heuristics as priors, Eq. (19) specifies a continuum of models ranging from LSA ($\theta = 0$) to LSS ($\theta \to \infty$). For all models, $\boldsymbol{y}$ is the activation time series for a single voxel; with spatial indices (i.e., coordinates in brain space) its notation is $\boldsymbol{y}_{\text{xyz}}$.

Both LSA and LSS are known as massive univariate GLMs, since they model each voxel independently. For a given voxel, LSA estimates weights as

$$\hat{\boldsymbol{w}}_{LSA}=(\boldsymbol{X}_{LSA}^T\boldsymbol{X}_{LSA})^{-1}\boldsymbol{X}_{LSA}^T\boldsymbol{y} \tag{20}$$

where $\boldsymbol{y}$ is the BOLD response time series for a voxel and $\boldsymbol{X}_{LSA}$ is the $n \times m$ design matrix with number of columns m equal to the number of trials ℓ, with only one event per trial. (Each $\boldsymbol{x}_j$ is an event for LSA, but this changes for LSS.) This means that a column in $\boldsymbol{X}_{LSA}$ models a single event in the experiment. The number of brain scans or time-points n is usually larger than the number of trials (events) ($n > \ell$) because more than one brain scan is acquired per trial. Quite commonly, a regressor models an event (such as stimulus presentation) with a boxcar function, that models the duration of the stimulus, convolved with a double gamma HRF (Boynton et al., 2012). We will not focus here on how the regressors that model the BOLD signal are constructed. Instead, we focus on the GLMs that receive those regressors as input.

The LSS model differs from the LSA model in that the $\boldsymbol{X}_{LSA}$ matrix is replaced with a set of matrices $\boldsymbol{X}_{LSS_1},\ldots,\boldsymbol{X}_{LSS_\ell}$ which results in one GLM per trial:

$$\begin{aligned}\hat{w}_{LSS_1}&=\boldsymbol{s}[(\boldsymbol{X}_{LSS_1}^T\boldsymbol{X}_{LSS_1})^{-1}\boldsymbol{X}_{LSS_1}^T\boldsymbol{y}]\\&\vdots\\\hat{w}_{LSS_\ell}&=\boldsymbol{s}[(\boldsymbol{X}_{LSS_\ell}^T\boldsymbol{X}_{LSS_\ell})^{-1}\boldsymbol{X}_{LSS_\ell}^T\boldsymbol{y}]\end{aligned} \tag{21}$$

Each $\boldsymbol{X}_{LSS_k}$ has dimensions $n \times 2$, where n is the same as before. Each weight $\hat{w}_{LSS_k}$ is selected as the first coefficient from its respective GLM, via multiplication by $\boldsymbol{s} = [1 \quad 0]$. Each $\boldsymbol{X}_{LSS_k}$ is constructed as mentioned above, with the first predictor variable $\boldsymbol{X}_{LSS_k}\boldsymbol{s}^T$ modeling a single experimental trial of interest (i.e., the kth trial) and the second predictor being a nuisance variable $\boldsymbol{X}_{LSS_k}[0 \quad 1]^T$ modeling all other trials in the experiment (i.e., all $\ell - 1$ trials excluding trial k). The LSS-prior model in the main text uses $\hat{\boldsymbol{w}}_{LSS}$ as $\hat{\boldsymbol{w}}_{LSS\ prior}$.

5.1.1. Simulated fMRI data

There were 1000 simulations performed for each of 9 different designs (see below) with varying levels of signal-to-noise ratio (SNR) and interstimulus intervals (ISI; time between events). The simulations were performed on modified code from the rsatoolbox (Nili et al., 2014), which can be consulted at:
https://github.com/bobaseb/rsa_toolbox_lss/tree/develop/LSS_project

Each simulation consisted of a cluster (i.e., region of interest) of task-sensitive signal voxels with observed data generated for all ℓ trials by weights $\boldsymbol{\Psi} \in \mathbb{R}^{m\times d\times d\times d}$, where $m = \ell$ and each spatial dimension $d = 7$. The weights $\boldsymbol{\Psi}$ were embedded in an array tripled along each spatial dimension $\boldsymbol{\Omega} \in \mathbb{R}^{m\times 3d\times 3d\times 3d}$ (i.e., the simulated brain). The weights for non-task-sensitive voxels in $\boldsymbol{\Omega}$ (i.e., those not in $\boldsymbol{\Psi}$) were set to zero. Scanner noise $\boldsymbol{E} \in \mathbb{R}^{n\times 3d\times 3d\times 3d}$ had entries $\epsilon_{i\text{xyz}}$ drawn i.i.d. from a centered normal distribution $\mathcal{N}(0,\sigma^2_{scanner})$, where $\sigma^2_{scanner} = 10\,000$, and was added to $\boldsymbol{X}_{LSA}\boldsymbol{\Omega}$ to generate the observed signal:

$$\boldsymbol{Y}=\boldsymbol{X}_{LSA}\boldsymbol{\Omega}+\boldsymbol{E}. \tag{22}$$

Thus, for a single voxel described by a set of spatial coordinates x, y, z, we have data across time in $\boldsymbol{Y} \in \mathbb{R}^{n\times 3d\times 3d\times 3d}$, represented as $\boldsymbol{y}_{\text{xyz}}$. For observations $\boldsymbol{Y}$, the subset corresponding to voxels that are task-sensitive is denoted as $\boldsymbol{Y}_{\boldsymbol{\Psi}}$. Notice the use of $\boldsymbol{X}_{LSA}$ to generate simulated data instead of $\boldsymbol{X}_{LSS}$. In fact, there is no straightforward way to construct weights $\boldsymbol{\Psi}$ (embedded in $\boldsymbol{\Omega}$) to multiply with the set of matrices $\boldsymbol{X}_{LSS_k}$.

To simulate spatiotemporal correlations in the data, the scanner noise $\boldsymbol{E}$ was smoothed along its four axes for each run (two runs total, see below), using a Gaussian spatiotemporal smoothing kernel with full width at half maximum (FWHM) equal to 4 mm for the three spatial dimensions and 4.5 s for the temporal dimension. (Voxel size was set in millimeters at the default value of $3 \times 3 \times 3.75$ in the rsatoolbox.)

For each simulation, each coordinate of the effect center (x, y, z, as defined in the rsatoolbox) — where the signal voxels were placed inside the simulated brain $\boldsymbol{\Omega}$ — was uniformly sampled between 1 and 11 inclusive. Two separate runs ($r = 2$) were simulated on each of the 1000 iterations and each run had 20 repetitions of each of two stimulus types ($\ell = 20$ s and s = 2). Simulating more than one run and stimulus type contributes to the ecological validity of the simulation, especially for studies that focus on classification (MVPA) where one run is used for training and another for testing. Repetition time (TR; duration for obtaining one full brain scan) was set to 1 s and event duration (ED, the duration of a stimulus on the screen in the MRI scanner) was set to 1.5 s.

A trial's duration is given by $ED + ISI$. There are also $\lceil \ell/3 \rceil$ null epochs, randomly interspersed with the trials, where no stimulus is shown, each with a duration of $ED + ISI$ seconds. This kind of experimental design is common because it further helps reduce collinearity between trials and aid in the estimation of $\boldsymbol{w}$. Thus, for each run $n \gtrapprox \frac{\frac{4}{3}\ell \times (ED+ISI)+t_{end}}{TR}$, where t_{end} is a temporal slack after the last trial that allows the BOLD signal enough time to decay. The exact number of time-points n depends on the HRF model that was used (Boynton et al., 2012). This information is encoded in the design matrix $\boldsymbol{X}_{LSA}$.

To sample the data-generating weights $\boldsymbol{\Psi}$ with correlations between (task-sensitive) voxels, we did the following: For each of the $\ell/s = 20$ trials per stimulus, we sampled from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Each entry $\mu_1, \ldots, \mu_{d^3}$ in mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ (for each stimulus, respectively) was i.i.d., drawn from a normal distribution $\mathcal{N}(0, \sigma^2_{\boldsymbol{\Psi}})$ for three levels of SNR ($\sigma^2_{\boldsymbol{\Psi}} \in \{10, 15, 20\}$). These were sampled for each iteration (a thousand iterations total) in each of the nine designs (Fig. 5) but kept constant across runs. The covariance matrix $\boldsymbol{\Sigma}$, with dimensions $d^3 \times d^3$, induces the correlations between task-sensitive voxels and was kept constant across runs but resampled on different iterations. It was drawn from a scaled Wishart distribution $W(\boldsymbol{V}, df)/df$ with degrees of freedom $df = d^3$. The symmetric positive definite matrix $\boldsymbol{V}$ was constructed with ones on the diagonal and 0.7 for all off-diagonal values, representing a high degree of correlation between task-sensitive voxels. As presented in Fig. 5, the 3 × 3 design of the simulations had three levels of ISI ∈ {2, 3, 4} (in seconds) and three levels of SNR (as mentioned above).

After sampling all $\ell \times d^3$ weights for a run, we have the object $\boldsymbol{M} \in \mathbb{R}^{\ell \times d^3}$. This matrix of weights (trials by voxels) $\boldsymbol{M}$ was permuted along the temporal dimension ℓ and was arbitrarily mapped to the spatial coordinates of $\boldsymbol{\Psi}$ – and by implication, of $\boldsymbol{\Omega}$ too – such that $\boldsymbol{M} \rightarrow \boldsymbol{\Psi} \rightarrow \boldsymbol{\Omega}$, before applying Eq. (22).

5.1.2. Model scoring

Our evaluations of the models were done with the root mean squared error (RMSE) of each $\hat{\boldsymbol{w}}$ for each model (i.e., LSA, LSS, LSA-prior and LSS-prior models) with respect to the ground truth of each vector ψ_{xyz} in $\boldsymbol{\Psi}$:

$$RMSE(\hat{\boldsymbol{w}}_{xyz}, \boldsymbol{\psi}_{xyz}) = \sqrt{\sum_{j=1}^{\ell} \frac{(\hat{w}_j - \psi_j)^2}{\ell}} \tag{23}$$

averaged across all the weights for task-sensitive voxels in $\boldsymbol{\Psi}$ and the 1000 iterations of simulated data:

$$\overline{\text{RMSE}} = \sum_{1}^{1000} \sum_{xyz} \frac{RMSE(\hat{\boldsymbol{w}}_{xyz}, \psi_{xyz})}{1000 d^3} \tag{24}$$

5.2. Results

Our main prediction held (see Fig. 5). Across a range of task conditions, our penalized regression approach with $\boldsymbol{w}_{LSS\ prior}$ outperformed both LSS (equivalent to large θ) and LSA (equivalent to $\theta = 0$) for intermediate penalty values of θ (Fig. 5). LSS provided an effective prior for our penalized regression. Replicating previous work, $\overline{\text{RMSE}}$ was lower for LSS than LSA, akin to less-is-more effects in which decision heuristics can best OLS (e.g., TTB in Fig. 3A).

6. General discussion

We looked toward human decision making to identify an effective prior for regularized regression and found that decision heuristics which disregard cue covariance information offer a number of advantages, such as robustness and interpretability. These heuristics offered a sensible starting point compared to the usual way of defining $\boldsymbol{w}_{prior}$ for most ridge regression applications (i.e., as the zero vector).

Here we have presented three different types of applications in over twenty different datasets, germane to the fields of decision making, fMRI analysis, and statistical modeling. We have validated the utility of using heuristics like TAL and TTB to construct $\boldsymbol{w}_{prior}$, as well as using other algorithms which lack a normative foundation and parallel the operation of heuristics — like LSS in the case of fMRI time series modeling.

Three main benefits of no-covariance priors are worth highlighting. First, predictions using a no-covariance prior are likely to provide at least as good, if not better, predictions than a vector of zeros as coefficients. Examples of the TTB-prior model outperforming other models are seen in Fig. 3A and in the lower $\overline{\text{RMSE}}$ values obtained in Application III.

Second, catastrophic failure of the model is avoided for extremely high values of θ, whereas in normal ridge regression, convergence to the zero vector for high penalty values results in essentially random guessing for the comparison and classification tasks presented here. Convergence toward very small weights may also create implementation issues on digital computers which have limited precision. For example, differences in how floating point numbers are represented in supporting software libraries could reduce the reproducibility of results.

Third, this class of priors has theoretical significance. On the one hand, the model class introduced here further integrates the notions of heuristic decision making and full information algorithms along one continuum of models (as in (Parpart et al., 2018)). Choosing heuristic priors that contrast compensatoriness of the environment, like TAL and TTB do, helps us interpret both the solutions of our models and the environment itself in an easier way than is possible with OLS or the Zero prior model. Likewise, the solution of the encompassing model could be understood in terms of deviations from the heuristic prior. Other informative comparisons could be made to the OLS solution, including how it diverges from the heuristic prior. Finally, our framework provides

a way to simulate fMRI data with LSS weights, previously not possible due to the arbitrariness of defining weights for the LSS nuisance variables (see Application III).

The theoretical contribution of this model class is worth emphasizing since it also provides a lens on why heuristics are useful in the first place. The priors offered by heuristics confer robustness; unlike the Zero prior, they embody a sensible inductive bias. This dovetails with why heuristics can operate defeasibly. Speculatively, humans and other cognitive agents may have evolved to implement these priors as a rule. Like Occam's razor, humans also show bias toward simple solutions for many decision-making tasks (Gigerenzer, Todd, & TAR, 1999; Kahneman et al., 1974). With expertise (i.e., acquiring more data), the solutions can change (Hornsby & Love, 2014), but initially, very general strategies like assuming independence among covariates have been documented (Bröder, 2000; Gigerenzer & Goldstein, 1996). Of course, this is only one notion of expertise. Other notions could include less effort during inference or rule application, finding appropriate features of a domain and ease of searching for new strategies or creating new ones. Furthermore, experts are not even guaranteed to perform better than statistical techniques (cf. (Meehl, 1954)).

Instead of being all-or-none, heuristic use may move along a continuum (Newell, 2005) as a function of prior strength and experience. Indeed, heuristic use in human decision making is not without its caveats (Newell, Weston, & Shanks, 2003), as is their supposed frugality (Bobadilla-Suarez & Love, 2018; Dougherty, Franco-Watkins, & Thomas, 2008). What is clear is that no heuristic will be best in all environments (cf. no free lunch theorem). Instead, each heuristic is best suited to certain environments and can be seen as embodying a prior that reflects beliefs about the environment.

Of course, this non-universality raises the critical question of how does one choose which heuristic to use? This question closely mirrors the inductive challenge of choosing a prior for a Bayesian model. A general solution to choosing the best heuristic is computationally intractable (Rich et al., 2021), though effective solutions have been offered (Rieskamp & Otto, 2006; Scheibehenne, Rieskamp, & Wagenmakers, 2013). Intuitively, if one believed, for whatever reason, that an environment was governed by numerous additive factors, then a heuristic like TAL would be a good strategy to adopt. The problem of strategy selection closely relates to the problem of meta-learning, or learning to learn, in which one determines how to choose hyperparameters, architectures, general strategies, etc. that will perform well in a task (Schweighofer & Doya, 2003). With enough data one can test which heuristic performs best on a sub-sample. However, in the low data regime this might not be possible. Our results show the differences between TTB and TAL used as a prior may not be too significant but future work should explore this angle.

With reference to models of human decision making, this class of algorithms has further potential. Referring back to the roots of regularized regression, Tikhonov (1943) initially constructed this type of regularization in a more general form:

$$\hat{\boldsymbol{w}}_{ridge} = \arg\min_{\boldsymbol{w}} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \|\boldsymbol{\Gamma}(\boldsymbol{w} - \boldsymbol{w}_{prior})\|_2^2\} \tag{25}$$

where θ has been replaced with a matrix $\boldsymbol{\Gamma}$. This enables the implementation of different penalty values for different directions in weight space. Admittedly, choosing $\boldsymbol{\Gamma}$ would require knowledge of the data. Our results suggest there might be some advantage in this kind of stepwise approach, where one model's output provides another model's prior. From a psychological point of view, this would enable modeling attention through the scaling of dimensions (Nosofsky, 1986). Although empirical studies show humans usually employ attention solely along individual dimensions (i.e., the diagonal of $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}$ (Jones, Love, & Maddox, 2005; Kruschke, 1993)), other applications (like our fMRI example) could benefit from this generality (Bobadilla-Suarez, Ahlheim, Mehrotra, Panos, & Love, 2020). Generalizations of such regression algorithms include adding a matrix that puts weights on observations themselves (van Wieringen, 2015) or even using heuristic regularizers for more complex models like neural networks. As in all modeling endeavors, the researcher should make clear how the model is intended (Jones & Love, 2011). For instance, a penalized regression approach could be proposed and evaluated as a normative account of what should be done, a high-level description of what people actually do, or an algorithmic account of the processes people engage in. We suggest further work on expertise (e.g., transitioning from novice to expert) could engage with any of the mentioned modeling strategies.

Furthermore, the models presented here provide only point estimates of $\hat{\boldsymbol{w}}_{ridge}$, but there is also no obstacle in expanding them to the Bayesian setting to obtain the full posterior distribution as well. In fact, it is well known that ridge regression, like LASSO regression, has a Bayesian interpretation (Friedman, Hastie, & Tibshirani, 2001; Parpart et al., 2018; Tibshirani, 1996). Our two-step approach engages in a double counting of the data (cf. (Zou, 2006)) which could suffer from bias and undue confidence in predictions. A Bayesian formulation could address this potential issue, providing new insights on why our two-step approach works and in which environments. This exciting future direction could expand the reach of our approach by placing it on a normative footing, enabling inquiry into the models' confidence.

In conclusion, we find that priors motivated by decision heuristics are valuable both methodologically and theoretically. Assuming independence among predictor variables offers a reasonable prior or starting point in most situations. These priors are themselves data-informed models that perform robustly when the penalty value (i.e., prior) is overly strong. Although ridge regression may not routinely suffer from extreme penalty values in practice, use of the TAL and TTB priors do not appear to have any significant downside and may be judged a more sensible choice and perhaps more akin to how people learn than ridge regression's null vector prior. Linking insights across fields as disparate as decision making and advanced methodologies for fMRI data analysis, we are confident that these robust priors for regularized regression will find even further utility in other fields, surpassing the theoretical contributions that we have hinted at here.

CRediT authorship contribution statement

Sebastian Bobadilla-Suarez: Coded all the analyses and simulations, derived the objective function for Equation 16 along with its Newton-Raphson estimate, Wrote the initial draft, Interpreted the results and provided critical comments on the manuscript. **Matt Jones:** Interpreted the results and provided critical comments on the manuscript. **Bradley C. Love:** Developed the study concept and derived the objective function for Equation 1, Interpreted the results and provided critical comments on the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and materials availability

1. The 20 datasets from Application I can be retrieved from: https://osf.io/fg4p5/
2. The Breast Cancer Wisconsin (Diagnostic) Data Set from Application II can be retrieved from: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
3. Code and data for Applications I and II can be found at: https://github.com/bobaseb/robustpriorsregression
4. Code for the simulations of the brain imaging data can be found at a fork of the RSA Toolbox (Nili et al. 2014) (develop branch): https://github.com/bobaseb/rsatoolboxlss/tree/develop/LSSproject

Acknowledgments

We thank Franziska Bröker, Brett Roads and the rest of the Love Lab for comments on an earlier version of this manuscript.

Funding

This research was supported by the National Institutes of Health (NIH), United States Grant 1P01HD080679, Royal Society Wolfson Fellowship, United Kingdom 183029, and Wellcome Trust Senior Investigator Award, United Kingdom WT106931MA from BCL, and National Science Foundation (NSF), United States Grant 2020906 to MJ.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cogpsych.2021.101444.

References

Blake, C., Keogh, E., & Merz, C. (1995). Uci repository of machine learning databases. (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)) (Accessed: 2019-06-26).

Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A., & Love, BC. (2020). Measures of neural similarity. *Computational Brain & Behavior*, *3*(4), 369–383.

Bobadilla-Suarez, S., & Love, BC. (2018). Fast or frugal, but not both: Decision heuristics under time pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(1), 24.

Boynton, G., Engel, S., & Heeger, D. (2012). Linear systems analysis of the fMRI signal. *NeuroImage*, *62*(2), 975–984.

Bröder, A. (2000). Assessing the empirical validity of the take-the-best heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning Memory, and Cognition*, *26*(5), 1332.

Bröder, A. (2003). Decision making with the adaptive toolbox: influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology, Learning, Memory, and Cognition*, *29*(4), 611–625.

Czerlinski, J., Gigerenzer, G., & Goldstein, DG. (1999). How good are simple heuristics? In *Evolution and cognition*, *Simple heuristics that make us smart* (pp. 97–118). New York, NY, US: Oxford University Press.

Dawes, RM. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571.

Dougherty, MR., Franco-Watkins, AM., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological Review*, *115*(1), 199–213.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *Springer series in statistics*, *The elements of statistical learning* (p. 1). Berlin: Springer.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

Gigerenzer, G., & Goldstein, DG. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Gigerenzer, G., Todd, PM., & TAR, Group (Eds.), (1999). *Simple heuristics that make us smart*. USA: Oxford University Press.

Hoerl, AE., & Kennard, RW (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hornsby, AN., & Love, BC. (2014). Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*, *3*(2), 72–76.

Jones, M., & Love, BC. (2011). BayesIan fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(04), 169–188.

Jones, M., Love, BC., & Maddox, WT. (2005). Stimulus generalization in category learning In *Proceedings of the annual meeting of the cognitive science society,Vol. 27*, No.27.

Kahneman, D., Slovic, P., & Tversky, A. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *1854157*, 1124–1131.

Katsikopoulos, KV., Schooler, LJ., & Hertwig, R. (2010). The robust beauty of ordinary information. *Psychological Review*, *117*(4), 1259–1266.

Kruschke, JK. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*(1), 3–36.

Le Cessie, S., & Van Houwelingen, JC. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *41*(1), 191–201.

Martignon, L., Hoffrage, U., et al. (1999). Why does one-reason decision making work. *Simple Heuristics that Make Us Smart*, 119–140.

Mata, R., et al. (2012). Ecological rationality: a framework for understanding and aiding the aging decision maker. *Frontiers in Neuroscience*, *6*(19).

Meehl, PE. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Mumford, JA., Poline, JB., & Poldrack, RA. (2015). Orthogonalization of regressors in fMRI models. *PLoS One*, *10*(4), Article e0126255.

Mumford, JA., Turner, BO., Ashby, FG., & Poldrack, RA. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, *59*(3), 2636–2643.

Newell, BR. (2005). Re-visions of rationality? *Trends in Cognitive Sciences*, *9*(1), 11–15.

Newell, BR., & Shanks, DR. (2003). Take the best or look at the rest? Factors influencing one-reason decision making. *Journal of Experimental Psychology, Learning, Memory, and Cognition*, *29*(1), 53–65.
Newell, BR., Weston, NJ., & Shanks, DR (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone takes-the-best. *Organizational Behavior and Human Decision Processes*, *91*(1), 82–96.
Nili, H., et al. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*(4).
Nosofsky, RM. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.
Otworowska, M., Blokpoel, M., Sweers, M., Wareham, T., & van Rooij, I. (2018). Demons of ecological rationality. *Cognitive Science*, *42*(3), 1057–1066.
Parpart, P., Jones, M., & Love, BC. (2018). Heuristics as Bayesian inference under extreme priors. *Cognitive Psychology*, *102*, 127–144.
Parpart, Paula, Jones, Matt, & Love, Bradley C. (2019). 2018 Osf repository. heuristics as bayesian inference under extreme priors. (https://osf.io/pb9yt/) (Accessed: 2019-06-26).
Rich, P., et al. (2021). Naturalism, tractability and the adaptive toolbox. *Synthese*, *198*(6), 5749–5784.
Rieskamp, J., & Otto, PE. (2006). Ssl: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207.
Rissman, J., Gazzaley, A., & D'esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage*, *23*(2), 752–763.
Schaefer, RL., Roi, LD., & Wolfe, RA (1984). A ridge logistic estimator. *Communications in Statistics. Theory and Methods*, *13*(1), 99–113.
Scheibehenne, B., Rieskamp, J., & Wagenmakers, EJ. (2013). Testing adaptive toolbox models: a Bayesian hierarchical approach. *Psychological Review*, *120*(1), 39–64.
Schweighofer, N., & Doya, K. (2003). Meta-learning in reinforcement learning. *Neural Networks*, *16*(1), 5–9.
Simon, HA. (1978). Rationality as process and as product of a thought. *American Economic Review*, *68*(2), 1.
Smith, AM., et al. (1999). Investigation of low frequency drift in fMRI signal. *Neuroimage*, *9*(5), 526–533.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.
Tikhonov, AN. (1943). On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR. Vol. 39* (pp. 195–198).
Turner, B. (2010). A comparison of methods for the use of pattern classification on rapid event-related fMRI data. In *Poster session presented at the annual meeting of the society for neuroscience*, San Diego, CA. (p. 0).
Visscher, KM., Kahana, MJ., & Sekuler, R. (2009). Trial-to-trial carry-over of item- and relational-information in auditory short-term memory.
van Wieringen, WN. (2015). Lecture notes on ridge regression. arXiv preprint arXiv:1509.09169.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.