Learning Credible Models

Jiaxuan Wang
University of Michigan
jiaxuan@umich.edu

Jeeheh Oh
University of Michigan
jeeheh@umich.edu

Haozhu Wang
University of Michigan
hzwang@umich.edu

Jenna Wiens
University of Michigan
wiensj@umich.edu

ABSTRACT

In many settings, it is important that a model be capable of providing reasons for its predictions (*i.e.*, the model must be interpretable). However, the model's reasoning may not conform with well-established knowledge. In such cases, while interpretable, the model lacks *credibility*. In this work, we formally define credibility in the linear setting and focus on techniques for learning models that are both accurate and credible. In particular, we propose a regularization penalty, expert yielded estimates (EYE), that incorporates expert knowledge about well-known relationships among covariates and the outcome of interest. We give both theoretical and empirical results comparing our proposed method to several other regularization techniques. Across a range of settings, experiments on both synthetic and real data show that models learned using the EYE penalty are significantly more credible than those learned using other penalties. Applied to two large-scale patient risk stratification task, our proposed technique results in a model whose top features overlap significantly with known clinical risk factors, while still achieving good predictive performance.

KEYWORDS

Model Interpretability, Regularization

ACM Reference Format:
Jiaxuan Wang, Jeeheh Oh, Haozhu Wang, and Jenna Wiens. 2018. Learning Credible Models. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3219819.3220070

1 INTRODUCTION

For adoption, predictive models must achieve good predictive performance. Often, however, good performance alone is not enough. In many settings, the model must also be interpretable or capable of providing reasons for its predictions. For example, in healthcare applications, research has shown that decision trees are preferred among physicians because of their high level of interpretability [15, 21]. Still, interpretability alone may not be enough to encourage adoption. If the reasons provided by the model do not agree, at least in part, with well-established domain knowledge, practitioners may be less likely to trust and adopt the model.

Often, one ends up trading off such credibility for interpretability, especially when it comes to learning sparse models. For example, regularization penalties, like the LASSO penalty, encourage sparsity in the learned feature weights, but in doing so may end up selecting features that are merely associated with the outcome rather than those that are known to affect the outcome. This can easily occur when there is a high-degree of collinearity present in one's data. In short, interpretability does not imply credibility.

Informally, a credible model is an interpretable model that i) provides reasons for its predictions that are, at least in part, inline with well-established domain knowledge, and ii) does no worse than other models in terms of predictive performance. While a user is more likely to adopt a model that agrees with well-established domain knowledge, one should not have to sacrifice accuracy to achieve such adoption. That is, the model should only agree with well-established knowledge, if it is consistent with the data. Relying on domain expertise alone would defeat the purpose of data-driven algorithms, and could result in worse performance. Admittedly, the definition of credibility is a subjective matter. In this work, we offer a first attempt to formalize the intuition behind a credible model.

Our main contributions include:

- formally defining credibility in the linear setting
- proposing a novel regularization term EYE (expert yielded estimates) to achieve this form of credibility.

Our proposed approach leverages domain expertise regarding known relationships between the set of covariates and the outcome. This domain expertise is used to guide the model in selecting among highly correlated features, while encouraging sparsity. Our proposed framework allows for a form of collaboration between the data-driven learning algorithm and the expert. We prove desirable properties of our approach in the least squares regression setting. Furthermore, we give empirical evidence of these properties on synthetic and real datasets. Applied to two large-scale patient risk stratification tasks, our proposed approach resulted in an accurate model and a feature ranking that, when compared to a set of well-established risk factors, yielded an average precision (AP) an order of magnitude greater than the second most credible model in one task, and twice as large in AP in the other task.

The rest of the paper is organized as follows. Section 2 reviews related work on variable selection and interpretability. Section 3 defines credibility and describes our proposed method in detail.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '18, August 19–23, 2018, London, United Kingdom
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3220070

Table 1: A comparison of relevant regularization penalties.

Method	Formulation	Sparsity	Grouping effect	Consistency
LASSO	$\|\boldsymbol{\theta}\|_1$	yes	no	conditioned [36]
ridge	$\frac{1}{2}\|\boldsymbol{\theta}\|_2^2$	no	yes	yes
elastic net	$\beta\|\boldsymbol{\theta}\|_1 + \frac{1}{2}(1-\beta)\|\boldsymbol{\theta}\|_2^2$	yes	yes	conditioned [14]
OWL	$\sum_{i=1}^{n} w_i \|\boldsymbol{\theta}\|_{[i]}$	yes	yes	unknown
weighted LASSO	$\|\boldsymbol{w} \odot \boldsymbol{\theta}\|_1$	yes	yes	no
weighted ridge	$\frac{1}{2}\|\boldsymbol{w} \odot \boldsymbol{\theta}\|_2^2$	no	no	no
adaptive LASSO	$\|\boldsymbol{w}^* \odot \boldsymbol{\theta}\|_1$	yes	no	conditioned [37]

Section 4 presents experiments and results. Section 5 summarizes the importance of our work and suggests potential extensions of our proposed method.

2 RELATED WORK

Credibility is closely related to interpretability, which has been actively explored in the literature [10, 17, 19, 24, 28, 30]. Yet, to the best of our knowledge, credibility has never been formally studied.

Interpretability is often achieved through dimensionality reduction. Common approaches include preprocessing the data to eliminate correlation, or embedding a feature selection criterion into the model's objective function. Embedding a regularization term in the objective function is often preferred over preprocessing techniques since it is nonintrusive in the training pipeline. Thus, while credible models could, in theory, be achieved by first preprocessing the data, we focus on a more general approach that relies on regularization.

The most common forms of regularization, l_1 (LASSO) and l_2 (ridge), can be interpreted as placing a prior distribution on feature weights [37] and can be solved analytically (LASSO in the orthogonal case, ridge in the general setting). The sparsity in feature weights induced by LASSO's diamond shaped contour is often desirable, thus many extensions of it have been proposed, including elastic net [38], ordered weighted LASSO (OWL) [7], adaptive LASSO [37], and weighted LASSO [3].

In Table 1, we summarize relevant properties for several common regularization terms. $\boldsymbol{\theta}$ represents the model parameters; $\beta \in [0, 1]$ is a hyperparameter that controls the tradeoff between the l_1 and l_2 norms; $\boldsymbol{w}$ is a set of non-negative weights for each feature; $\boldsymbol{w}^*$ is the optimal set of weights (according to a least squares solution) [37]; $|\boldsymbol{\theta}|_{[i]}$ is the ith largest parameter sorted by magnitude; and $\odot$ is the elementwise product. The grouping effect refers to the ability to group highly correlated covariates together [38], and consistency refers to the property that learned features converge in distribution to the true underlying feature weights [13]. Without the grouping effect, some relevant features identified as important by experts may end up not being selected because they are correlated with other relevant expert recommended features.

In terms of incorporating additional expert knowledge at training time, Sun *et al.* explore using features identified as relevant during training, along with a subset of other features that yield the greatest improvement in predictive performance [29]. This work differs from ours because they assume expert knowledge as ground truth, a potentially dangerous assumption when experts are wrong. Vapnik *et al.* explore the theory of learning with privileged information [31]. Though similar in setting, they use expert knowledge to accelerate the learning process, not to enforce credibility. Helleputte and Dupont use partially supervised approximation of zero-norm minimization (psAROM) to create a sparse set of relevant features. Much like weighted LASSO, psAROM does not exhibit the grouping effect, thus is unable to retain all known relevant features. Moreover, the non-convex objective function for psAROM makes exact optimization hard [11]. [5] looks at utilizing hierarchical expert information to learn embeddings that help model prediction of rare diseases. While it is an interesting approach, its model's interpretability is questionable. [25] constrains the input gradient of features that are believed not to be relevant in a neural network. In the linear setting, the method simplifies to l_2 regularization on unknown features, which is suboptimal for model interpretability because the learned weights are dense.

Perhaps closest to our proposed approach, and the concept of credibility, is related work in interpretability that focuses on enforcing monotonicity constraints between the covariates and the prediction [2, 16, 20, 23, 33]. The main idea behind this branch of work is to restrict classifiers to the set of monotone functions. This restriction could be probabilistic [16] or monotone in certain arguments identified by experts [2, 23, 33]. Though similar in aim (having models inline with domain expertise), previous work has focused on rule based systems. Other attempts to enforce monotonicity in nonlinear models [1, 26, 32] aim to increase performance. Again, relying too heavily on expert knowledge may result in a decrease in performance when experts are wrong. In contrast, we propose a general regularization technique that aims to increase credibility without decreasing performance. Moreover, in the linear setting, credible models satisfy monotonicity and sparsity constraints.

3 PROPOSED APPROACH

In this paper, we focus on linear models. Within this setting, we start by formally defining credibility in 3.1. Then, building off of a naïve approach in 3.2, we introduce our proposed approach in 3.3. In 3.4, we state important properties and theoretical results relevant to our proposed method.

3.1 Definition and Notation

Interpretability is a prerequisite for credibility. For linear models, interpretability is often defined as sparsity in the feature weights. Here, we define the set of features as $\mathcal{D}$. We assume that we have some domain expertise that identifies $\mathcal{K} \subseteq \mathcal{D}$, a subset of the features as known (or believed) to be important. Intuitively, among a group correlated features a credible model will select those in $\mathcal{K}$, if the relationship is consistent with the data.

Consider the following unconstrained empirical risk minimization problem, $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, X, \boldsymbol{y}) + n\lambda J(\boldsymbol{\theta}, \boldsymbol{r})$ that minimizes the sum of some loss function L and regularization term J. X is an n by d design matrix, where row $\boldsymbol{x}$ corresponds to one observation. The corresponding entry in $\boldsymbol{y} \in \mathbb{R}^n$ is the target value for $\boldsymbol{x}$. Let v_i denote the ith entry of a vector $\boldsymbol{v}$. $\lambda \in \mathbb{R}_{\geq 0}$ is the tradeoff between loss and regularization, and $\boldsymbol{r} \in \{0,1\}^d$ is the indicator array where $r_i = 1$ if $i \in \mathcal{K}$ and 0 otherwise. Note that our setting differs from the conventional setting only through the inclusion of $\boldsymbol{r}$ in the regularization term. For theoretical convenience, we prove theorems in the least squares regression setting and denote $\hat{\boldsymbol{\theta}}^{OLS}$ as the ordinary least squares solution. For experiments, we use logistic loss.

We denote $\boldsymbol{\theta}$ as the true underlying parameters. Then $\boldsymbol{\theta}_{\mathcal{K}}$ and $\boldsymbol{\theta}_{\mathcal{D}\setminus\mathcal{K}}$ are the true parameters associated with the subset of known and unknown features, respectively. Throughout the text, vectors are in bold, and estimates are denoted with a hat.

Definition A linear model is *credible* if

(i) Within a group of correlated *relevant* features $C \subseteq D$: $\hat{\boldsymbol{\theta}}_{\mathcal{K}\cap C}$ is dense, and $\hat{\boldsymbol{\theta}}_{C\setminus\mathcal{K}}$ is sparse (*structure constraint*).
(ii) Model performance is comparable with other regularization techniques (*performance constraint*)

Consider the following toy example where $|C| = 2$ and one of these features has been identified $\in \mathcal{K}$ by the expert, while the other has not. One could arbitrarily select among these two correlated features, including only one in the model. To increase credibility, we encourage the model to select the known feature (*i.e.*, the feature in $\mathcal{K}$)

We stress *relevant* in the definition because we do not care about the structure constraint if the group of variables does not contribute to the predictive performance. We assume expert knowledge is sparse compared to all features; thus a credible model is sparse due to the structure requirement. Credible models will result in dense weights among the known features, if the expert knowledge provided is indeed supported by the data. If experts are incorrect, *i.e.*, the set of features $\mathcal{K}$ are not relevant to the task at hand, then credible models will discard these variables, encouraging sparsity.

3.2 A Naïve Approach to Credibility

Intuitively, one may achieve credibility by constraining weights for known important factors with the l_2 norm and weights for other features with the l_1 norm. The l_2 norm will maintain a dense structure in known important factors and the l_1 norm will encourage sparsity on all remaining covariates. Formally, this penalty can be written as $q(\boldsymbol{\theta}) = (1-\beta)\|\boldsymbol{r} \odot \boldsymbol{\theta}\|_2^2 + 2\beta\|(\mathbf{1}-\boldsymbol{r}) \odot \boldsymbol{\theta}\|_1$ where $\boldsymbol{\theta} \in \mathbb{R}^d$, $\beta \in (0,1)$ controls the tradeoff between weights associated with the features in $\mathcal{K}$ and in $\mathcal{D} \setminus \mathcal{K}$.

Unfortunately, q does not encourage sparsity in $\hat{\boldsymbol{\theta}}_{\mathcal{D}\setminus\mathcal{K}}$. **Figure 1a** shows its contour plot. For a convex problem, each level set of the contour corresponds to a feasible region associated with a particular λ. A larger level value implies a smaller λ. It is clear from the figure that this penalty is non-homogeneous, that is $f(t\boldsymbol{x}) \neq |t| f(\boldsymbol{x})$. In a two-dimensional setting, when the covariates perfectly correlate with one another, the level curve for the loss function will have a slope of -1 corresponding to the violet dashed lines in **Figure 1**.

To understand why the slope must be -1, consider the classifier $y = \theta_{\mathcal{K}} x_1 + \theta_{\mathcal{D}\setminus\mathcal{K}} x_2$. Since x_1 and x_2 are perfectly correlated by assumption, we have $y = (\theta_{\mathcal{K}} + \theta_{\mathcal{D}\setminus\mathcal{K}})x_1$. Note that the loss value is fixed as long as $\theta_{\mathcal{K}} + \theta_{\mathcal{D}\setminus\mathcal{K}}$ is fixed, which means that each level curve of the loss function has the form $\theta_{\mathcal{K}} + \theta_{\mathcal{D}\setminus\mathcal{K}} = c$ for some scaler c, *i.e.*, $\theta_{\mathcal{D}\setminus\mathcal{K}} = -\theta_{\mathcal{K}} + c$. Thus, the slope of the violet lines must be -1 in **Figure 1**.

By the KKT conditions, with $\lambda > 0$, the optimal solution (red dots for each level curve in **Figure 1**) occurs at the boundary of the contour with the same slope ($\lambda = 0$ means the problem is unconstrained, then all methods are equal). We observe that with small λ, the large constraint region forces the model to favor features not in $\mathcal{K}$ because the point on the boundary with slope of -1 occurs near $\theta_{\mathcal{D}\setminus\mathcal{K}}$ axis, leading to a model that is not credible.

3.3 The Expert Yielded Estimates (EYE) Penalty

To address this sensitivity to the choice of hyperparameter, we propose the EYE penalty, obtained by fixing a level curve of q and scaling it for different contour levels. The trick is to force the slope of level curve in the positive quadrant to approach -1 as $\theta_{\mathcal{D}\setminus\mathcal{K}}$ approaches 0. Note that since q is symmetric around both axes, we can just focus on one "corner." That is, we want the "corner" on the right of the level curve to have a slope of -1, so that $\hat{\boldsymbol{\theta}}$ hits it in the perfectly correlated case. In fact, as long as $-1 \leq$ the "corner" slope ≤ 0, we achieve the desired feature selection. In the extreme case of slope 0 ($\beta = 1$), we do not penalize $\boldsymbol{\theta}_{\mathcal{K}}$ at all. Using a slope with a magnitude smaller than 1 assumes that features in $\mathcal{K}$ are much more relevant than other features, thus biasing $\hat{\boldsymbol{\theta}}_{\mathcal{K}}$. Since we do not wish to bias $\hat{\boldsymbol{\theta}}_{\mathcal{K}}$ towards larger values, if the solution is inconsistent with the data, we keep the slope as -1. This minimizes the effect of our potential prejudices, while maintaining the desirable feature selection properties. Casting our intuition mathematically yields the EYE penalty:

$$eye(\boldsymbol{x}) = \inf\left\{ t > 0 \mid \boldsymbol{x} \in \left\{ t\boldsymbol{x} \mid q(\boldsymbol{x}) \leq \frac{\beta^2}{1-\beta} \right\} \right\} \tag{1}$$

where t is a scaling factor to make EYE homogeneous and the inner set defines the level curve to fix. Note that β only scales the EYE penalty, thus can rewrite the penalty as:

$$eye(\boldsymbol{\theta}) = \|(\mathbf{1}-\boldsymbol{r}) \odot \boldsymbol{\theta}\|_1 + \sqrt{\|(\mathbf{1}-\boldsymbol{r}) \odot \boldsymbol{\theta}\|_1^2 + \|\boldsymbol{r} \odot \boldsymbol{\theta}\|_2^2} \tag{2}$$

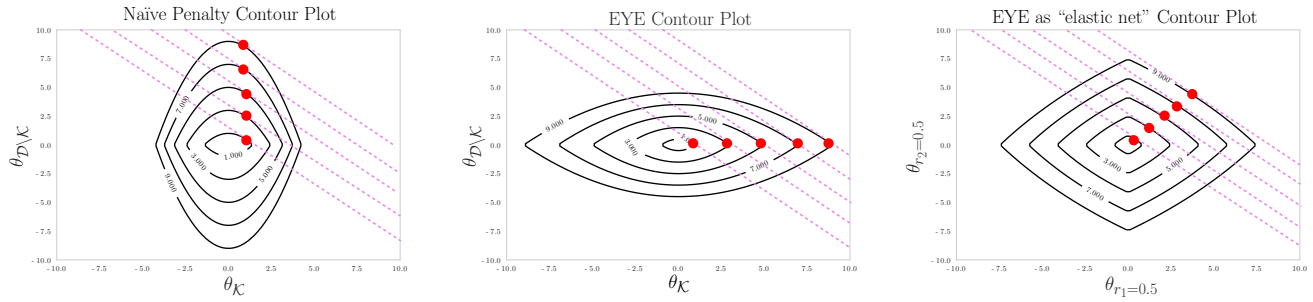
Derivations of (1) and (2) are included in the Appendix[1]. **Figure 1b** shows the contour plot of EYE penalty (note that the optimal solution for each level set occurs at the "corner" as desired).

3.4 EYE Properties

In this section, we give theoretical results for the proposed EYE penalty. We include detailed proofs in the Appendix[1]. While the first three properties are general, the last three properties are valid in the least squares regression setting, *i.e.*, $Loss(\boldsymbol{\theta}, X, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2$. We focus on the least square regression setting because a closed form solution exists, though our method is applicable to the classification setting as well (demonstrated in section 4).

EYE is a norm: This comes for free as **Equation (1)** is an atomic norm [4], thus, convex.

[1]please refer to section 6 of https://arxiv.org/pdf/1711.03190.pdf due to page limitation



(a) (b) (c)

Figure 1: Visualization of selected regularization penalties. Dashed violet lines denote level sets for the loss function when features are perfectly correlated; red dots are the optimal points for each feasible region. A large feasible region (level sets with large labeled values) corresponds to a small λ. (a) The naïve penalty ($\beta = 0.5$) favors $\theta_{\mathcal{D}\setminus\mathcal{K}}$ as the feasible region grows. (b) EYE consistently favors $\theta_{\mathcal{K}}$. (c) When $\boldsymbol{r} = \mathbf{0.5}$, EYE produces a contour plot similar to elastic net. Setting $\boldsymbol{r} = \mathbf{0.5}$ represents a situation in which two features i and j are equally "known" and perfectly correlated. In this setting, $\hat{\theta}_i = \hat{\theta}_j$ (*i.e.*, highly correlated known factors have similar weights)

EYE is β free: Similar to elastic net and the naïve penalty q, EYE is a combination of the l_1 and l_2 norms, but it omits the extra parameter β. This leads to a quadratic reduction in the hyperparameter search space for EYE compared to elastic net and q.

EYE is a generalization of LASSO, l_2 norm, and "elastic net": Setting $\boldsymbol{r} = \mathbf{1}$ and $\mathbf{0}$, we recover the l_2 norm and LASSO penalties, respectively. Relaxing $\boldsymbol{r}$ from a binary valued vector to a float valued vector, so that $\boldsymbol{r} = \mathbf{0.5}$, we get the elastic net shaped contour (**Figure 1c**). Elastic net is in quotes because the contour represents one particular level set, and elastic net is non-homogeneous.

EYE promotes sparse models: Assuming $X^\top X = I$, the solution to EYE penalized least squares regression is sparse. **Figure 3** illustrates this effect in the context with other regularization penalties.

EYE favors a solution that is sparse in $\hat{\boldsymbol{\theta}}_{\mathcal{D}\setminus\mathcal{K}}$ and dense in $\hat{\boldsymbol{\theta}}_{\mathcal{K}}$: In a setting in which covariates are perfectly correlated, $\hat{\boldsymbol{\theta}}_{\mathcal{D}\setminus\mathcal{K}}$ will be set to exactly zero. Conversely, $\hat{\boldsymbol{\theta}}_{\mathcal{K}}$ has nonzero entries. Moreover, the learned weights will be the same for every entry of $\hat{\boldsymbol{\theta}}_{\mathcal{K}}$ (*e.g.*, **Figure 1c**). This verifies the first part of the structure constraint. We also note that when the group of correlated features are all in $\mathcal{D} \setminus \mathcal{K}$, the objective function reverts back to LASSO, so that the weights are sparse, substantiating the second part of the structure constraint.

EYE groups highly correlated known factors together: If $\hat{\theta}_i\hat{\theta}_j > 0$ and the design matrix is standardized, then

$$\frac{|r_i^2\hat{\theta}_i - r_j^2\hat{\theta}_j|}{Z} \leq \frac{\sqrt{2(1-\rho)}\|\boldsymbol{y}\|_2}{n\lambda} + |r_i - r_j|\left(1 + \frac{\|(\mathbf{1}-\boldsymbol{r})\odot\hat{\boldsymbol{\theta}}\|_1}{Z}\right)$$

where ρ is the sample covariance between x_i and x_j, and $Z = \sqrt{\|(\mathbf{1}-\boldsymbol{r})\odot\hat{\boldsymbol{\theta}}\|_1^2 + \|\boldsymbol{r}\odot\hat{\boldsymbol{\theta}}\|_2^2}$.

This implies that when $r_i = r_j \neq 0$

$$\frac{|\hat{\theta}_i - \hat{\theta}_j|}{Z} \leq \frac{\sqrt{2(1-\rho)}\|\boldsymbol{y}\|_2}{r_i^2 n\lambda}$$

I.e., the more correlated known important factors are, the more similar their weights will be. This is analogous to the grouping effect.

4 EXPERIMENTS

In this section, we empirically verify EYE's ability to yield credible models through a series of experiments. We compare EYE to a number of other regularization penalties across a range of settings using both synthetic and real data.

4.1 Measuring Credibility

Criterion (i): density in the set of known relevant features and sparsity in the set of unknown. In a two dimensional setting, we measure $\log|\frac{\theta_{\mathcal{K}}}{\theta_{\mathcal{D}\setminus\mathcal{K}}}|$ as a proxy for desirable weight structure (the higher the better). In a high-dimensional setting, highly correlated covariates form groups. For each group of correlated features, if known factors exist and are indeed important, then the shape of the learned weights should match $\boldsymbol{r}$ in the corresponding groups. *E.g.*, given two correlated features x_1 and x_2 that are associated with the outcome, if $r_1 = 0$ and $r_2 = 1$, then $\theta_1 = 0$ and $\theta_2 \neq 0$. Thus, to measure credibility, we use the symmetric KL divergence, $symKL(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}', \boldsymbol{r}') = \frac{1}{2}\left(KL(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}'\|\boldsymbol{r}') + KL(\boldsymbol{r}'\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}')\right)$, between the normalized absolute value of learned weights and the normalized $\boldsymbol{r}$ for each group g. For groups of relevant features that do not contain known factors, the learned weights should be sparse (*i.e.*, all weight should be placed on a single feature within the group). Thus, we report $\min_{\boldsymbol{x}\in \textit{one hot vectors}} symKL(\boldsymbol{x}, \hat{\boldsymbol{\theta}}')$ for such groups. As $symKL$ decreases, the credibility of a model increases. Note that $symKL$ only measures the shape of weights within *each group* of correlated features and does not assume expert knowledge is correct (*e.g.*, all weights within a group could be near zero).

In our experiments on real data, we do not know the true underlying $\boldsymbol{\theta}$ and the partition of groups. In this case, we measure credibility by computing the fraction of known important factors in the top n features sorted by the absolute feature weights learned by the model. We sweep n from 1 to d and report the average precision (AP) between $|\hat{\boldsymbol{\theta}}|$ and $\boldsymbol{r}$.

Criterion (ii): maintained classification performance. Recall that we want to learn a credible model without sacrificing

model performance. That is, there should be no statistically significant difference in performance between a credible model and the best performing one (in this case, we focus on best linear models learned using other regularization techniques). We measure model performance in terms of the area under the receiver operating characteristic curve (AUC). In our experiments, we split our data into train, validation, and test sets. We train a model for each hyperparameter and bootstrap the validation set 100 times and record performance on each bootstrap sample. We want a model that is both accurate and sparse (measured using the Gini coefficient due to its desirable properties [12]). To ensure accuracy, for each regularization method, we remove models that are significantly worse than the best model in that regularization class using the validation set bootstrapped 100 times (p value set at .05). From this filtered set, we choose the sparsest model and report criteria (i) and (ii) on the held-out test set.

4.2 Experimental Setup and Benchmarks

We compare EYE to the regularization penalties in **Table 1** across various settings. We exclude ridge from our comparisons, because it produces a dense model (**Figure 3**). In addition, we exclude adaptive LASSO because it requires an additional stage of processing.

We set the weights, $\boldsymbol{w}$, in **Table 1**, to mimic the effect of the $\boldsymbol{r}$. This gives a subset of the regularization techniques according to the same kind of expert knowledge that our proposed approach uses. In weighted LASSO and weighted ridge, the values in $\boldsymbol{w}_{\mathcal{D}\setminus\mathcal{K}}$ were swept from 1 to 3 times the magnitude of the values in $\boldsymbol{w}_{\mathcal{K}}$ to penalize unknown factors more heavily. For OWL, we set the weights in two ways. In the first case, we only penalize $|\hat{\theta}|_{[1]}$, effectively recovering the l_∞ norm. In the second case, weights for the m largest entries in $\hat{\theta}$ are set to be twice the magnitude of the rest, where m is the number of known important factors. Note that a direct translation from known factors to weights is not possible in OWL, since the weights are determined based on the learned ordering. We implemented all models as a single layer perceptron with a softmax trained using the ADADELTA algorithm [35] minimizing the logistic loss.

4.3 Validation on Synthetic Datasets

To test EYE under a range of settings, we construct several synthetic datasets [2]. In all experiments, we generate the data and run logistic regression with EYE and each regularization benchmark. In all of our experiments on synthetic data, we found no statistically significant differences in AUC, thus satisfying the performance constraint. These experiments expose the limitations of the naïve penalty, measure sensitivity to noise and to correlation in covariates, explore different shapes of $\boldsymbol{r}$, and examine the effect of the accuracy of expert knowledge on credibility. In all cases, the EYE penalty leads to the most credible model, validating our theoretical results.

4.3.1 Limitations of the Naïve Penalty: Sensitivity to Hyperparameters. The naïve penalty q appears to be a natural solution for building credible linear models. However, since q is non-homogeneous, as the constraint region grows, the models begin to prefer features *not* in $\mathcal{K}$. Since small λ corresponds to a large constraint region, we vary λ to expose this undesirable behavior.

We sample 100 data points uniformly at random from -2.5 to 1.5 to create $\boldsymbol{v}$. We set $X = [\boldsymbol{v}, \boldsymbol{v}]$ to produce two perfectly correlated features with one known factor. We set $\boldsymbol{\theta} = [1, 1]$ (note that since the two features are perfectly correlated, it doesn't matter how $\boldsymbol{\theta}$ is assigned), and assign the label $\boldsymbol{y}$ as $\mathbb{1}_{\boldsymbol{\theta}^\top \boldsymbol{x} > 0}(\boldsymbol{x})$ for each data point $\boldsymbol{x}$.

Figure 2a shows the log ratio for credibility for different settings of λ and β. First note that as λ approaches zero, the log ratio approaches 0 for all methods because the models are effectively unconstrained. With nontrivial λ and large β, both EYE and the naïve penalty result in high credibility. This is expected as a large β will constrain known important factors less, thus placing more weight on them. For β in the lower range, the log ratio is negative because the naïve penalty penalizes known features more. For β in the middle range, the log ratio varies from credible to non-credible, exhibiting the artifact of non-homogeneity (the penalty contour is elongated along $\theta_{\mathcal{K}}$ as λ decreases, thus again favoring $X_{\mathcal{D}\setminus\mathcal{K}}$). Since we want the log ratio> 0 for all nontrivial λ, the naïve penalty with $\beta < 0.8$ fails.

The naïve penalty with large β also fails to produce credible models because the resulting models have worse classification performance. In particular, when $\beta > 0.8$, the naïve penalty overemphasizes the relevancy of known important factors. As shown in **Figure 2b**, the naïve penalty with large β performs considerably worse in terms of accuracy than EYE for large λ. On small λ, their performance are comparable. This is expected because EYE introduces less bias towards known important factors.

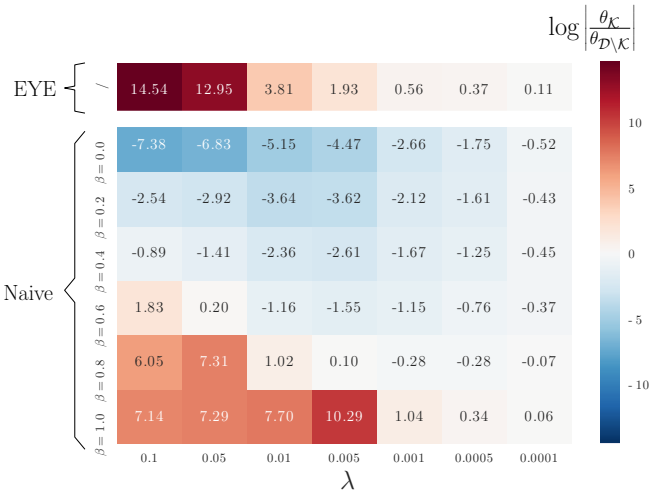
4.3.2 Varying the Degree of Collinearity. We can show theoretically that EYE results in a credible model when features are highly correlated. However, the robustness of EYE in the presence of noise is unknown. To explore how EYE responds to changes in correlation between features, we conduct an experiment in a high-dimensional setting.

We generate 10 groups of data, each having 30 features, with 15 in $\mathcal{K}$. We assigned each group a correlation score from 0 to 0.9 (here, we exclude the perfectly correlated case as it will be examined in detail in the next experiment). Intra-group feature correlations are fixed to the group's correlation score, while inter-group feature correlations are 0.

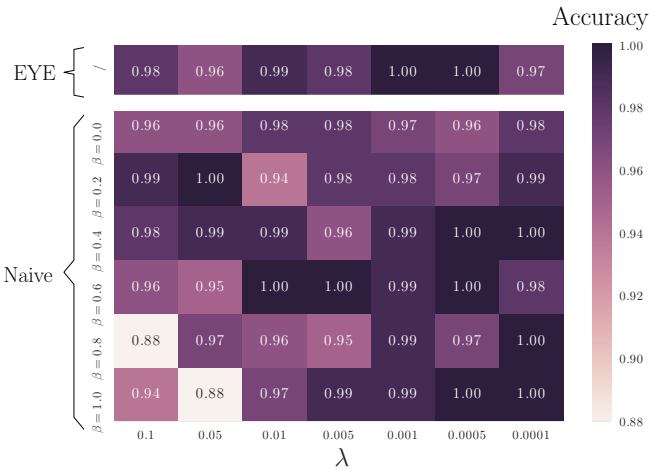
Figure 4a plots the $symKL$ for each group. Moving from left to right, the correlation increases in step size of 0.1 from 0 to 0.9. As correlation increases, the EYE regularized model achieves the smallest $symKL$, and becomes the most credible model. In comparison, the other approaches do not achieve the same degree of credibility though, weighted LASSO and weighted ridge do exhibit a similar trend. However, since weighted LASSO fails to capture denseness in known important factors and weighted ridge fails to capture sparseness in unknown features, EYE leads to a more credible model. As correlation increases, LASSO actually produces a less credible model (as expected).

4.3.3 Varying Percentage of Known Important Factors. Besides varying correlation, we also vary the percentage of known important factors within a group of correlated features. We observe that EYE is consistently better than other methods.

[2] code available at https://github.com/nathanwang000/credible_learning



(a)



(b)

Figure 2: A comparison of the naïve penalty and EYE. (a) EYE meets the structural constraint better than naïve penalty with small and mid-ranged β (b) EYE has better performance than naïve Penalty with large β.

In this experiment, we generate groups of data C_i where $i = 0, \ldots, 10$, each having 10 features. Features in each group are perfectly correlated, and features across groups are independent. Each group has a different number of features in $\mathcal{K}$, *e.g.*, group 0 has 0 known relevant factors and group 10 has 10 known important factors.

Figure 4b plots the $symKL$ for each group of features. The groups are sorted by $|C_i \cap \mathcal{K}|$. When $|C_i \cap \mathcal{K}| = 0$, the model should be sparse. Indeed, for group 0, we observe that EYE, LASSO, and weighted LASSO do equally well (EYE in fact degenerates to LASSO in this case), closely followed by elastic net. Weighted ridge and OWL, on the other hand, do poorly since they encourage dense models. For other groups, EYE penalty achieves the best result (lowest $symKL$). This can be explained by property 3.4 as EYE sets the weights the same for correlated features in $\mathcal{K}$ while zeroing

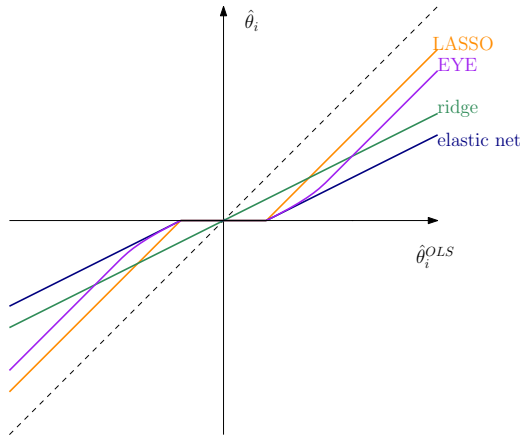
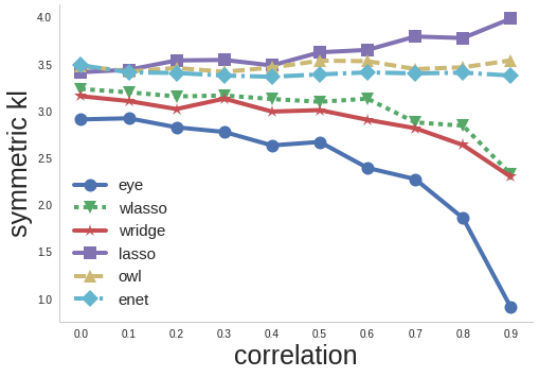
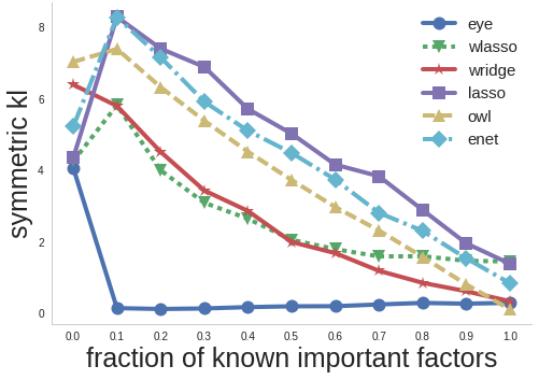


Figure 3: When the design matrix is orthonormal, EYE, elastic net, and LASSO will set features with small ordinary least squares solution to exactly 0. In contrast, ridge is dense.



(a)



(b)

Figure 4: Comparisons of EYE with other methods under various settings (a) EYE leads to the most credible models in all correlations. (b) EYE leads to the most credible model for all shapes of r.

Table 2: EYE leads to the most credible model on a synthetic dataset (mean ± stdev)

Method	$\sum_{g=1}^{n} symKL_g$	AUC
EYE	**0.442** ± 0.128	0.900 ± 0.044
wLASSO	0.929 ± 0.147	0.898 ± 0.044
wridge	1.441 ± 0.241	0.899 ± 0.045
LASSO	2.483 ± 0.440	0.898 ± 0.044
elastic net	2.673 ± 0.399	0.893 ± 0.044
OWL	3.125 ± 0.329	0.900 ± 0.044

out weights in $\mathcal{D} \setminus \mathcal{K}$. Again, LASSO performed the worst overall because it ignores $\boldsymbol{r}$ and is sparse even when $\boldsymbol{r}$ is dense.

4.3.4 Varying Accuracy of Expert Knowledge. The experiments above only test cases where $\boldsymbol{\theta}$ is elementwise positive and where expert knowledge is correct (*i.e.*, the features identified by the expert were indeed relevant). To simulate a more general scenario in which the expert may be wrong, we use the following generative process:

(1) Select the number of independent groups, $n \sim$ Poisson(10)
(2) For each group i in n groups
   (a) Sample a group weight, $w^{(i)} \sim$ Normal(0,1)
   (b) Sample the number of features, $m^{(i)} \sim$ Poisson(20)
   (c) Sample known important factor indicator array, $\boldsymbol{r}^{(i)} \sim$ Bernoulli(0.5)$^{m^{(i)}}$
   (d) Assign true relevance $\boldsymbol{\theta}^{(i)} \in \mathbb{R}^{m^{(i)}}$ by distributing $w^{(i)}$ according to $\boldsymbol{r}^{(i)}$ (*e.g.*, if $w^{(i)} = 3$ and $\boldsymbol{r}^{(i)} = [0, 1, 1]$, then $\boldsymbol{\theta}^{(i)} = [0, 1.5, 1.5]$)
(3) Generate covariance matrix C such that intra-group feature correlation=0.95 and inter-group feature correlation=0
(4) Generate 5000 i.i.d. samples $\boldsymbol{x}_i \in \mathbb{R}^{\sum_{i=1}^{n} m^{(i)}} \sim$ Normal($\boldsymbol{0}$, C)
(5) Choose label $y_i \sim$ Bernoulli($sigmoid(\boldsymbol{\theta}^\top \boldsymbol{x}_i)$) where $\boldsymbol{\theta}$ is the concatenated array from $\boldsymbol{\theta}^{(i)}$

Generating data this way covers cases where expert knowledge is wrong as feature group relevance and $\boldsymbol{r}$ are independently assigned. It also allows the number of features and weights for each group to be different. **Table 2** summarizes performance and credibility for each method averaged across 100 runs. EYE achieves the lowest sum of $symKL$ for each group of correlated features. In terms of AUC, the best models for each penalty are comparable, confirming that EYE is able to recover from the expert's mistakes.

4.4 Application to a Real Clinical Prediction Task

After verifying desirable properties in synthetic datasets, we apply EYE to a large-scale clinical classification task. In particular, we consider the task of identifying patients at greatest risk of acquiring an infection during their hospital stay. We selected a task from healthcare since credibility and interpretability are critical to ensuring the safe adoption of such models. We focus on predicting which patients will acquire a *Clostridium difficile* infection (CDI), a particularly nasty healthcare-associated infection. Using electronic health record (EHR) data from a large academic US hospital, we aim to learn a credible model that produces accurate *daily* estimates of patient risk for CDI.

4.4.1 The Dataset. We consider all adult hospitalizations between 2010 and 2015. We exclude hospitalizations in which the patient is discharged or diagnosed with CDI before the 3rd calendar day, since we are interested in healthcare-acquired infections (as opposed to community-acquired). Our final study population consists of 143, 602 adult hospitalizations. Cases of CDI are clinically diagnosed by positive laboratory test. We label a hospitalization with a positive laboratory test for CDI as +1, and 0 otherwise. 1.09% of the study population is labeled positive.

4.4.2 The Task. We frame the problem as a prediction task: the goal is to predict whether or not the patient will be clinically diagnosed with CDI at some point in the future during their visit. In lieu of a single prediction at 24 hours, we make predictions every 24 hours. To generate a single AUC given multiple predictions per patient, we classify patients as high-risk if their risk ever exceeds the decision threshold, and low-risk otherwise. By sweeping the decision threshold, we generate a single receiver operating characteristic curve and a single AUC in which each hospitalization is represented exactly once.

4.4.3 Feature Extraction. We use the same feature extraction pipeline as described in [22]. In particular, we extract high-dimensional feature vectors for each day of a patient's admission from the structured contents of the EHR (*e.g.*, medication, procedures, in-hospital locations etc.). Most variables are categorical and are mapped to binary features. Continuous features are either binned by quintiles or well-established reference ranges (*e.g.*, a normal heart rate is 60-100 beats per minute). If a feature is not measured (*e.g.*, missing vital), then we explicitly encode this missingness. Finally, we discard rare features that are not present in more than .05% of the observations. This feature processing resulted in 4,739 binary variables. Of these variables, 264 corresponded to known risk factors. We identified these variables working with experts in infectious disease who identified key factors based on the literature [6, 8, 34].

4.4.4 Analysis. We train and validate the models on data from the first five years (n=444, 184 days), and test on the held-out most recent year (n=217, 793 days). Using the training data, we select hyperparameters using a grid search for λ and β from 10^{-10} to 10^{10} and 0 to 1 respectively. The final hyperparameters are selected based on model performance and sparsity as detailed in section 4.1.

For each regularization method, we report the AUC on the held-out test set, and the average precision (AP) between $|\hat{\boldsymbol{\theta}}|$ and $\boldsymbol{r}$ (see Section 4.1). **Table 3** summarizes the results on the test set with various regularizations.

Relative to the other common regularization techniques, EYE achieves an AP that is an order of magnitude higher, while maintaining good predictive performance. Moreover, EYE leads to one of the sparsest models, increasing model interpretability.

For comparison, we include a model based on only the 264 expert features (trained using l_2 regularized logistic regression) "expert-features-only." This baseline trivially achieves AP of 1, since it only uses expert features, but performs poorly relative to the other tasks. This confirms that simply retaining expert features is not enough to solve this task.

In addition, we include a baseline, "EYE-random-r", in which we randomly permuted $\boldsymbol{r}$. This corresponds to the setting where the expert is incorrect and is providing information about features that may be irrelevant. In this setting, EYE achieves a high AUC and

Table 3: EYE leads to the most credible model on both the *C. diff* and *PhysioNet Challenge* datasets; it keeps more of the factors identified in the clinical literature, while performing on par with other regularization techniques; it also has very sparse weights, second only to the model that just uses features in the risk factors

	C. diff			*PhysioNet Challenge*		
Method	AP	AUC	sparsity$^+$	AP	AUC	sparsity$^+$
expert-features-only	1*	0.598	**0.998**	1*	0.754	**0.877**
EYE	**0.204**	0.753	0.980	**0.671**	0.815	0.794
wLASSO	0.033	0.764	0.884	0.300	0.810	0.824
LASSO	0.032	0.760	0.856	0.131	**0.823**	0.779
wridge	0.031	**0.768**	0.755	0.209	0.810	0.069
elastic net	0.031	0.754	0.880	0.153	0.818	0.649
EYE-random-r	0.031	0.748	0.936	0.589	0.792	0.779
OWL	0.028	0.548	0.544	0.108	0.794	0.046

$^+$ percentage of near-zero feature weights, where near-zero is defined as < 0.01 of the largest absolute feature weight
* expert-features-only logistic regression trivially achieves AP of 1 simply because it only uses expert features

low AP. This confirms that EYE is not severely biased by incorrect expert knowledge. Moreover, we believe this to be a feature of the approach, since it can highlight settings in which the data and expert disagree.

4.5 Application to PhysioNet Challenge Dataset

To further validate our approach, we turn to a publicly available benchmark dataset from PhysioNet [9]. In this task, the goal is to predict in-hospital mortality using EHR data collected in intensive care units (ICUs). Similar to above using the EYE penalty we trained a model and evaluated it in terms of predictive performance, average precision (AP), and model sparsity.

4.5.1 The Dataset. We use the ICU data provided in the PhysioNet Challenge 2012 [27] to train our model. This challenge utilizes a subset of the MIMIC-III dataset. We focus on this subset rather than using the entire dataset, since the goal is not to achieve state-of-the-art in in-hospital mortality prediction, but simply to evaluate the performance of the EYE penalty. The challenge data consist of three sets, each set containing data for 4000 patients. In our experiments, we use set A, since it is the only publicly labeled subset. We split the data randomly, reserving 25% as the held-out test set.

4.5.2 The Task. Using data collected during the first two days of an ICU stay, we aim to predict which patients survive their hospitalizations, and which patients do not. In contrast to the *C. diff* task, here, we make a single prediction per patient at 48 hours.

4.5.3 Feature Extraction. The PhysioNet challenge dataset has considerably fewer features relative to the earlier task. In total, for each patient the data contain four general descriptors (*e.g.*, age) and 37 time-varying variables (*e.g.*, glucose, pH, etc.) measured possibly multiple times during the first 48 hours of the patient's ICU stay. We describe our feature extraction process below. Since again the goal was not state-of-the-art prediction on this particular task, we performed standard preprocessing without iteration/optimization.

We represent each patient by a vector containing 130 features. More specifically, for each time-varying variable we compute the maximum, mean, and minimum over the 48 hour window, yielding 111 features. In addition, for each of the 15 time-varying variables used in the Simplified Acute Physiology Score (SAPS-I) [18] we extract the most abnormal value observed within the first 24 hours,based on the SAPS scoring system. We concatenate these 126 features along with the 4 general descriptors producing a final vector of length 130. Out of the 130 variables, we consider the 15 SAPS-I variables along with age as expert knowledge. SAPS-I is a scoring system used to predict ICU mortality in patients greater than the age of 15 and thus corresponds to factors believed to increase patient risk.

4.5.4 Analysis. Using the training data, we select hyperparameters in the same way we did earlier. As with the previous experiment on the *C. diff* dataset, for each regularization method, we report both AUC and AP on the held-out test set for this task. Again, we compared the model learned using the EYE penalty to the other baselines. **Table 3** summarizes our results on the held-out test set.

Overall, we observed a similar trend as to what we observed for the *C. diff* dataset. Compared to the other common regularization techniques, EYE achieves significantly higher AP and results in a sparse model. In terms of discriminative performance it performs on par with the other techniques. Again, we see that a model based on the expert features alone (i.e., *expert-features-only*) performs worse than the other regularization techniques. However, the difference in performance is not as striking as it was earlier. This suggests that perhaps the additional features (beyond the 16 SAPS-I features) do not provide much complementary information. Interestingly, the model using randomly permuted $\boldsymbol{r}$ ("EYE-random-r") achieves high AUC and AP. We suspect this may be due to the amount of collinearity present in the data. The non-expert and expert features are highly correlated with one another and thus both subsets are predictive (*i.e.*, supported by the data).

5 DISCUSSION & CONCLUSION

In this work, we extended the notion of interpretability to credibility and presented a formal definition of credibility in a linear setting. We proposed a regularization penalty, EYE, that encourages such credibility. Our proposed approach incorporates domain knowledge about which factors are known (or believed) to be important. Our incorporation of expert knowledge results in increased credibility, encouraging model adoption, while maintaining model performance. Through a series of experiments on synthetic data, we showed that sparsity inducing regularization such as LASSO, weighted LASSO, elastic net, and OWL do not always produce credible models. In contrast, EYE produces a model that is provably credible in the least squares regression setting, and one that is consistently credible across a variety of settings.

Applied to two large-scale patient risk stratification tasks, EYE produced a model that was significantly better at highlighting known important factors, while being comparable in terms of predictive performance with other regularization techniques. Moreover, we demonstrated how the proposed approach does not lead to worse performance when the expert is wrong. This is especially important in a clinical setting, where some relationships between variables and the outcome of interest may be less well-established.

There are several important limitations of the proposed approach. We focused on a linear setting and one form of expert knowledge. In the future, we plan to extend the notion of credibility to other settings. Furthermore, we do not claim that EYE is the optimal approach to yield credibility (we give no proof on that). Compared to other regularization penalties considered in this paper, EYE introduces the least amount of bias, while striving to attain credibility.

While interpretable models have garnered attention in recent years, increased interpretability should not have to come at the expense of decreased credibility. Predictive performance and sparsity being equal, a data-driven model that reflects what is known or well-accepted in one's domain (in addition to what is unknown, but reflected in the data) is preferred over a purely data-driven model that highlights unusual features due to collinearity in the data. Moreover, correlations can be fragile and break over time; thus, credible models that select those features that are known to be associated with the outcome of interest may also be more robust to such changes over time.

Finally, though we focused on credibility, our proposed regularization technique could be extended to other settings in which the user would like to guide variable selection. For example, instead of encoding knowledge pertaining to which variables are known risk factors, $\mathbf{r}$ could encode information about which variables are actionable. This in turn could lead to more *actionable* models.

6 ACKNOWLEDGEMENT

This work was supported by the National Science Foundation (NSF award no. IIS-1553146); the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (grant no. U01AI124255). The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the National Science Foundation nor the National Institute of Allergy and Infectious Diseases of the National Institutes of Health.

REFERENCES

[1] Eric E Altendorf, Angelo C Restificar, and Thomas G Dietterich. 2012. Learning from sparse data by exploiting monotonicity constraints. *arXiv preprint arXiv:1207.1364* (2012).
[2] Arie Ben-David. 1995. Monotonicity maintenance in information-theoretic machine learning algorithms. *Machine Learning* 19, 1 (1995), 29–43.
[3] Linn Cecilie Bergersen, Ingrid K Glad, and Heidi Lyng. 2011. Weighted lasso with data integration. *Statistical applications in genetics and molecular biology* 10, 1 (2011).
[4] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. 2012. The convex geometry of linear inverse problems. *Foundations of Computational mathematics* 12, 6 (2012), 805–849.
[5] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 787–795.
[6] Erik R Dubberke, Yan Yan, Kimberly A Reske, Anne M Butler, Joshua Doherty, Victor Pham, and Victoria J Fraser. 2011. Development and validation of a Clostridium difficile infection risk prediction model. *Infection Control & Hospital Epidemiology* 32, 04 (2011), 360–366.
[7] Mario AT Figueiredo and Robert D Nowak. 2014. Sparse estimation with strongly correlated variables using ordered weighted l1 regularization. *arXiv preprint arXiv:1409.4005* (2014).
[8] KW Garey, TK Dao-Tran, ZD Jiang, MP Price, LO Gentry, and HL Dupont. 2008. A clinical risk index for Clostridium difficile infection in hospitalised patients receiving broad-spectrum antibiotics. *Journal of Hospital Infection* 70, 2 (2008), 142–147.
[9] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. 2000 (June 13). PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* 101, 23 (2000 (June 13)), e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.
[10] Satoshi Hara and Takanori Maehara. 2016. Finding Alternate Features in Lasso. *arXiv preprint arXiv:1611.05940* (2016).
[11] Thibault Helleputte and Pierre Dupont. 2009. Partially supervised feature selection with regularized linear models. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 409–416.
[12] Niall Hurley and Scott Rickard. 2009. Comparing measures of sparsity. *IEEE Transactions on Information Theory* 55, 10 (2009), 4723–4741.
[13] Il'dar Abdulovič Ibragimov and Rafail Z Has' minskii. 2013. *Statistical estimation: asymptotic theory*. Vol. 16. Springer Science & Business Media. 30 pages.
[14] Jinzhu Jia and Bin Yu. 2010. ON MODEL SELECTION CONSISTENCY OF THE ELASTIC NET WHEN p » n. *Statistica Sinica* (2010), 595–611.
[15] Igor Kononenko. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine* 23, 1 (2001), 89–109.
[16] Wojciech Kotłowski and Roman Słowiński. 2009. Rule learning with monotonicity constraints. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 537–544.
[17] Himabindu Lakkaraju and Cynthia Rudin. 2017. Learning Cost-Effective and Interpretable Treatment Regimes. In *Artificial Intelligence and Statistics*. 166–175.
[18] Jean-Roger Le Gall, Philippe Loirat, Annick Alperovitch, Paul Glaser, Claude Granthil, Daniel Mathieu, Philippe Mercier, Remi Thomas, and Daniel Villers. 1984. A simplified acute physiology score for ICU patients. *Critical care medicine* 12, 11 (1984), 975–977.
[19] Zachary C Lipton. 2016. The mythos of model interpretability. *ICML Workshop on Human Interpretability in Machine Learning* (2016).
[20] David Martens, Jan Vanthienen, Wouter Verbeke, and Bart Baesens. 2011. Performance of classification models from a user perspective. *Decision Support Systems* 51, 4 (2011), 782–793.
[21] Geert Meyfroidt, Fabian Güiza, Jan Ramon, and Maurice Bruynooghe. 2009. Machine learning techniques to examine large patient databases. *Best Practice & Research Clinical Anaesthesiology* 23, 1 (2009), 127–143.
[22] Jeeheh Oh, Maggie Makar, Christopher Fusco, Robert McCaffrey, Krishna Rao, Erin Ryan, Laraine Washer, Lauren West, Vincent Young, John Guttag, David Hooper, Erica Shenoy, and Jenna Wiens. 2018. A Generalizable, Data-Driven Approach to Predict Daily Risk of Clostridium difficile Infection at Two Large Academic Health Centers. *Infection Control and Hospital Epidemiology* (2018).
[23] Michael J Pazzani, S Mani, William R Shankle, et al. 2001. Acceptance of rules generated by machine learning among medical experts. *Methods of information in medicine* 40, 5 (2001), 380–385.
[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144.

[25] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717* (2017).

[26] Joseph Sill. 1998. Monotonic networks. *Advances in neural information processing systems* (1998), 661–667.

[27] Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. 2012. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology, 2012.* IEEE, 245–248.

[28] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* 41, 3 (2014), 647–665.

[29] Jimeng Sun, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Ebadollahi, Zahra Daar, and Walter F Stewart. 2012. Combining knowledge and data driven insights for identifying risk factors using electronic health records.. In *AMIA*, Vol. 2012. 901–10.

[30] Berk Ustun and Cynthia Rudin. 2014. Methods and models for interpretable linear classification. *arXiv preprint arXiv:1405.4047* (2014).

[31] Vladimir Vapnik and Rauf Izmailov. 2015. Learning using privileged information: similarity control and knowledge transfer. *Journal of Machine Learning Research* 16 (2015), 2023–2049.

[32] Marina Velikova, Hennie Daniels, and Ad Feelders. 2006. Solving partially monotone problems with neural networks. In *Proceedings of the International Conference on Neural Networks, Vienna, Austria.*

[33] Wouter Verbeke, David Martens, Christophe Mues, and Bart Baesens. 2011. Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications* 38, 3 (2011), 2354–2364.

[34] Jenna Wiens, Wayne N Campbell, Ella S Franklin, John V Guttag, and Eric Horvitz. 2014. Learning Data-Driven Patient Risk Str. jpegication Models for Clostridium difficile. In *Open forum infectious diseases*, Vol. 1. Oxford University Press, ofu045.

[35] Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* (2012).

[36] Peng Zhao and Bin Yu. 2006. On model selection consistency of lasso. *Journal of Machine learning research* 7, Nov (2006), 2541–2563.

[37] Hui Zou. 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101, 476 (2006), 1418–1429.

[38] Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society* 67, 2 (2005), 301–320.