# Building Trustworty Models (Draft)

Jiaxuan Wang

*<2017-03-07 Tuesday>*

The motivation is modified based on the motivation section from Lauren Naylor's original write-up.

## 1 motivation

Machine learning models in healthcare must achieve good predictive performance, but to be used in practice, they also must be interpretable. Interpretability can be defined in many ways depending on the context or setting. It can refer to how well a human can reproduce the calculations of a model, how intuitive the parameters and calculations are, or how well a human can understand how a model's algorithm works, even if they cannot reproduce it by hand [3]. In healthcare, we define an interpretable model as one that is able to provide reasons for its predictions. Past research has shown that decision trees are preferred among physicians because of their high level of interpretability [5, 2]. However, interpretability alone is not enough to completely gain physician's trust.

A model may provide reasons for its predictions, but if the reasons do not agree with what is known to be medically relevant, physicians will not trust it. For example, the lasso penalty, $\lambda \sum_{i=1}^{n} |\theta_i|$ is commonly used to create interpretable models. This penalty induces sparsity in the learned feature weights, so that predictions can be explained by a small number of relevant factors. While this improves interpretability, it does nothing to ensure that the selected features align with physicians' knowledge. If a feature that is known to be relevant is correlated with a feature that is not, the model may use the latter feature to make predictions and discard the former.

This example suggests that in addition to interpretability, we also want the model to be credible, that is to agree with prior knowledge in the field. Ideally, this increase in trust should not lower performance.

Apart from gaining trust, credible models promote actions. In the medical context, we often not only desire to predict the risk of a patient getting an infection, but also want to enact actionable plans to lower the risk. Admittedly, all models theoretically provide actions that will lower the risk by perturbing features to observe the effect on the predicted outcome. Then physicians can choose to focus on features that lower the risk the most. However, in a non-credible model, because of its lack of domain knowledge, it's very plausible to suggest an unorthodox action that is often costly compared to an already implemented procedure that solves the same issue (they solve the same issue because the known risk factor is correlated with the unknown feature). Alternatively, many features are simply not actionable, meaning that even though we know a feature is important, we cannot change it. For example, age is positively correlated to many disease, but we cannot change this feature. If we use a credible model and interpret risk factors as actionable factors denoting whether a feature can be changed, the suggested actions will make more sense.

It may sound too good to be true that there's no trade-off between performance and credibility: doesn't everything come at a cost? Given that we already have an accurate model, credibility filters features through domain expertise. The cost we pay is this outside knowledge, which in the medical context are known risk factors. To illustrate this point, consider an absurd example of trying to predict the number of people drowning in a month using season and number of ice-creams sold within that month as features. Anyone reasonable would agree that more people drowning in summer than in any other season because more people swim in summer. The season should be an obvious relevant feature for a learning algorithm. However, it is very likely that a model would choose the number of ice-creams sold as a predictive variable over season because more ice-creams sold implies summer and thus positively affects the number of people drowning. At the heart of this issue is that highly correlated features are indistinguishable to a model in their predictive power. In the extreme case of co-linear features, their optimal weights for minimizing prediction error can vary over a continuum as long as the sum of weights is fixed. There's no way for the model to know which one to pick in correlated features without outside guidance and that's what makes its reasoning cryptic. We can fix this issue by providing the model with domain knowledge.

The goal of this research is to create credible models in the sense that matches a clinician's medical knowledge. We aim to develop methods for combining the expert-based relevancy of features with a datadriven model. As a case study, we focus on the specific prediction task of predicting a

patient's risk of acquiring an infection with C. difficile.

## 2 objective

We want to incorporate known risk factors into models such that they favor known risk factors over unknown features when these features are correlated. We want the model to be end to end trainable so that users don't have to choose a hard threshold to cutoff correlation as is often the case in preprocessing.

## 3 TODO related work

While interpretability has been actively explored in the literature [3], credibility has never been formally studied. It is often assumed that an interpretable model implies trust, which is not true as we saw in the motivation. In fact, credibility is a super-set of interpretability. Credibility implies interpretability, but not the other way around.

Interpretability is usually approached through dimensionality reduction. We preprocess the data to eliminate correlation or encode the feature selection criterion into the model's objective function or both. Dimensionality reduction methods can be classified by whether they are continuous or discrete. Discrete methods include subset selection (examples includes best subset selection and its computationally efficient variant, stepwise selection), which explicitly selects subset of features so that the trade-off between model simplicity and loss is balanced, PCA, which minimizes noise in choosing orthogonal components, and ICA, which picks out independent components by assuming that they are not normally distributed. Continuous shrinkage methods refer to regularization, which is penalty added on the norm of model parameters. Regularization is the most oftenly used class of methods due to its non-intrusiveness on the training pipeline. In this work, we focus on regularization methods because of their better integration into the model so that no hard threshold are picked manually.

The most commonly used and analyzed regularizations are $L_1$ (lasso) and $L_2$ (ridge) norm due to their desirable statistical properties. Each of which can be interpreted as placing a prior distribution on feature weights `[8]`. The sparseness in feature weights induced by lasso's diamond shaped contour makes it more favorable in the context of eliminating irrelevant features, thus many extensions over it are proposed, including ordered weighted loss (OWL) [1], adaptive lasso [8], elastic net [9], and weighted lasso. While OWL, elastic

net, and weighted lasso are generalizations of lasso, adaptive lasso satisfies the oracle property in the sense that under mild regularity conditions, it identifies the right subset model and is consistent with true parameters (that is the learned features converge in distribution to the true underlying feature weights). However, adaptive lasso requires learning another model to set its weight, making it more cumbersome to use than others.

The most natural extension over the regular lasso is the weighted lasso, which introduces a weight $w_i$ for each feature: $\lambda \sum_{i=1}^{n} w_i |\theta_i|$. This penalty is used in [4] where the feature's weight is the inverse of its relevance (**TODO: Haven't read this paper yet, will do**). This approach causes the weights of less relevant features correlated with more relevant features to be driven to zero. However, we may not know the relevance of the features that have not been identified as risk factors: there may be undiscovered relationships not mentioned in the literature. If such a feature were correlated with a known risk factor, we would want to throw it out and use the known risk factor, but if it is not correlated with another feature and is predictive, we would like to keep it. Combining expert knowledge with a model is explored in [6]. The model is trained using features identified as relevant, along with the subset of other features from the data that give the most improvement to performance, while creating the least redundancy in the features. This work differs from ours because their list of relevant features is assumed to be known, and their motivation is to increase model performance, not credibility.

## 4 measuring success

Fixing the level of performance, the task of learning is to allocate weights to features so that desirable structures are kept. We want our model to be consistent with physician's knowledge. More concretely, we want the model to prefer more sparsity in the unknown set of features compared to the known set. This whole process should be data driven so that the known risk factors are merely suggestions for the model instead of forced constraints. We call a model credible if it satisfies the following properties:

1. within a group of dependent features, weights of known risk factors should be dense
2. within a group of dependent features of all unknown risk factors, the weights should be sparse
3. maintained model performance

Criteria 3) is acheived by grid searching over the validation set so that models in consideration have similar level of performance.

For 1) and 2), we measure the distance in distribution between each group of correlated features and the known risk factor indicator vector within that group. The metrics used is symmetric KL divergence, which is $\frac{KL(p\|q)+KL(q\|p)}{2}$ where $p$ and $q$ are distributions we are considering.

Here's an example of calculating KL divergence in a group of dependent features. It's trivial to extend this example for symmetric KL.

Assume $r = [1, 1, 0, 0]^T$ and $\theta = [0.1, 0.2, -0.01, 0.02]^T$ ($\theta$ excluding b term), we first normalize each vector so that their $\|\cdot\|_1$ is 1.

$r' = [0.5, 0.5, 0, 0]^T$, $\theta' = [0.32258065, 0.64516129, 0.03225806, 0.06451613]^T$

To avoid 0 appearing in log of KL divergence calculation, a small smooth factor of 1e-6 is added to any vector with 0, renormalizing giving

$r'' = [4.99999000e{-}01, 4.99999000e{-}01, 9.99996000e{-}07, 9.99996000e{-}07]^T$, $\theta'' = [0.32258065, 0.64516129, 0.03225806, 0.06451613]^T$

Then $KL(r''\|\theta'')$ is the reported result in each dependent group, where $KL(x\|y) = \sum_i p(x_i) \log \frac{p(x_i)}{p(y_i)}$

In the case where r is all 0 in relevant feature group, I give $\min_{v \in one\ hot\ vectors} KL(v\|\theta'')$ as a loss as to encourage sparse feature.

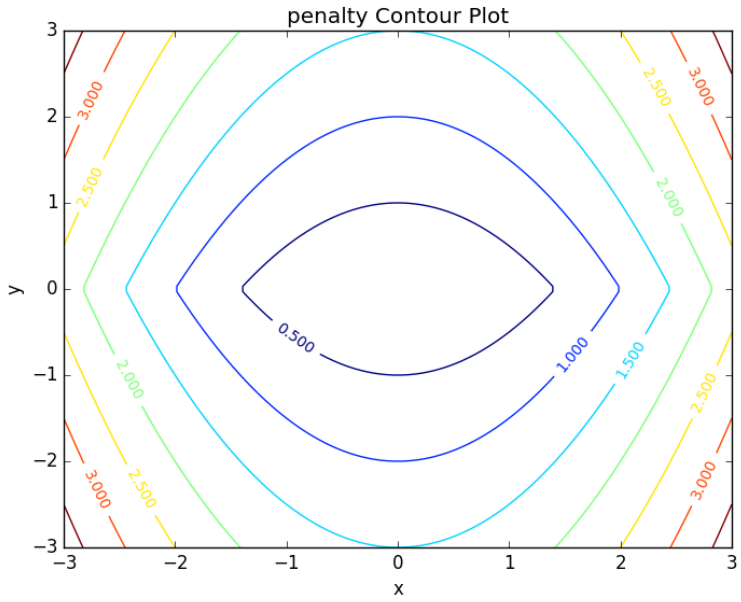
# 5 method

The most natural approach to encourage sparseness in unknown risk factors while maintaining dense weights in known risk factors is to constrain known risk factors using $l_2$ norm and unknown risk factors using $l_1$ norm. Formally, this penalty term can be written as

$pena(\theta) = \lambda(0.5(1-\beta)\|r \odot \theta\|_2^2 + \beta\|(1-r) \odot \theta\|_1)$

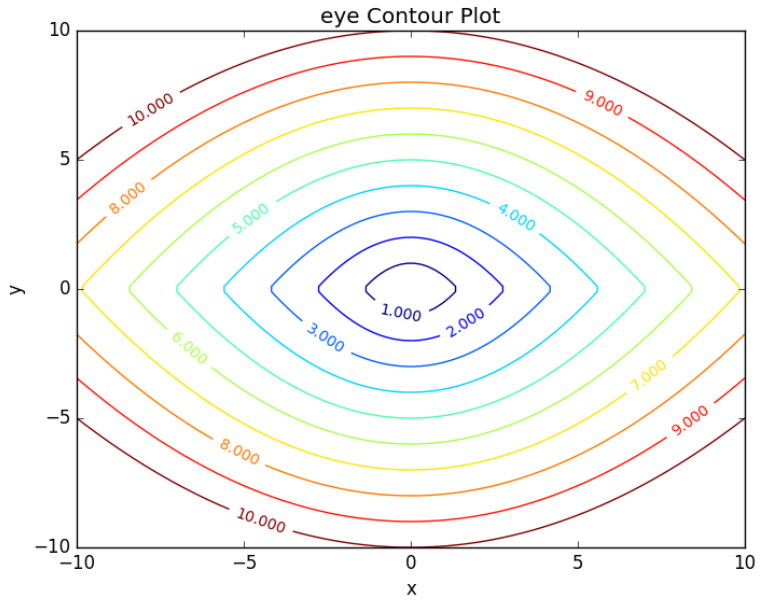
where r $\in \{0,1\}^{\mathrm{d}}$, $\theta \in \mathbb{R}^d$, $\lambda \in \mathbb{R}_+$, $\beta \in [0,1]$

Assuming x is the known risk factor and y is the unknown risk factor, we plot the contour of this penalty:




As the contour plot suggests, this penalty function is nonhomogeneous: that is $f(tx) \neq |t|f(x)$. In the case of perfectly correlated variables, this translate to model's sensitivity to $\lambda$: small $\lambda$ will let the model favor unknown risk factor y which is opposite to what we want.

To address this issue, we propose eye penalty which is obtained by fixing a convex body in the contour of pena and scale it for different contour levels. We call the fixed contour the generating convex body. Consider the corners of the cross section between the known and unknown risk factors, we want the corners to have slope of magnitude 1 so that perfectly correlated features will favor known risk factors. The generating convex body is exactly determined via this criteria. The contour plot for the 2 dimensional case is again plotted.




The new contour plot demonstrates that eye penalty is indeed homogeneous.

While a derivation of this penalty and the proof of its properties can be found in the last section 7.0.1, I state the result:

### 5.1 formal definition of eye penalty

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

### 5.2 properties

1. eye is a norm
2. eye is $\beta$ free
3. eye is a generalization of lasso, ridge, and elastic net

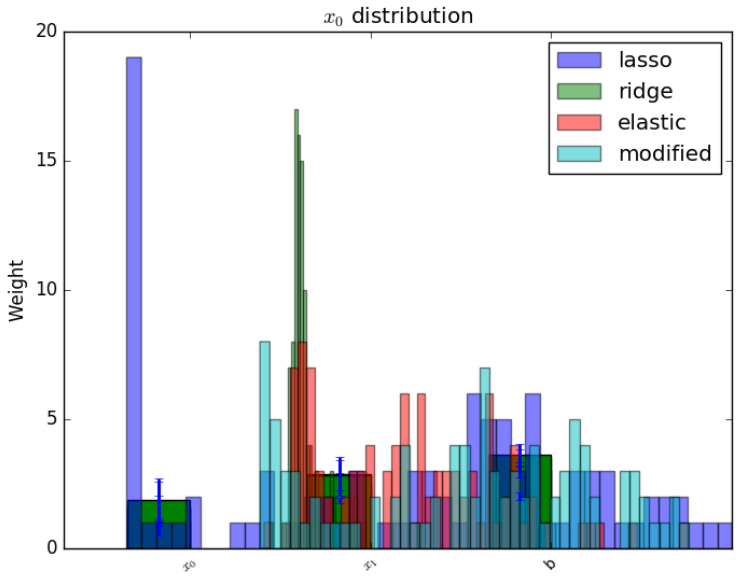
## 6 TODO experiments

Each experiment was ran with a different aim in mind. The first four experiments explore 2d data while the last four experiments explore high di-

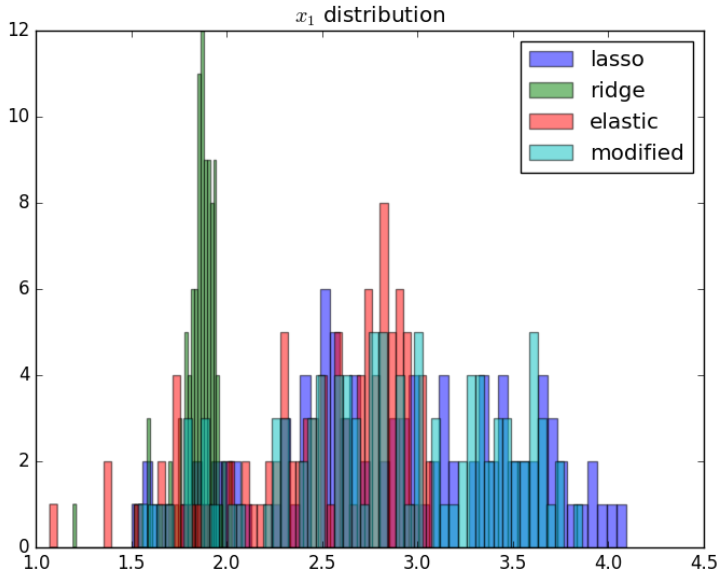
mensional data. The last experiments applies eye penalty to C. difficile prediction.

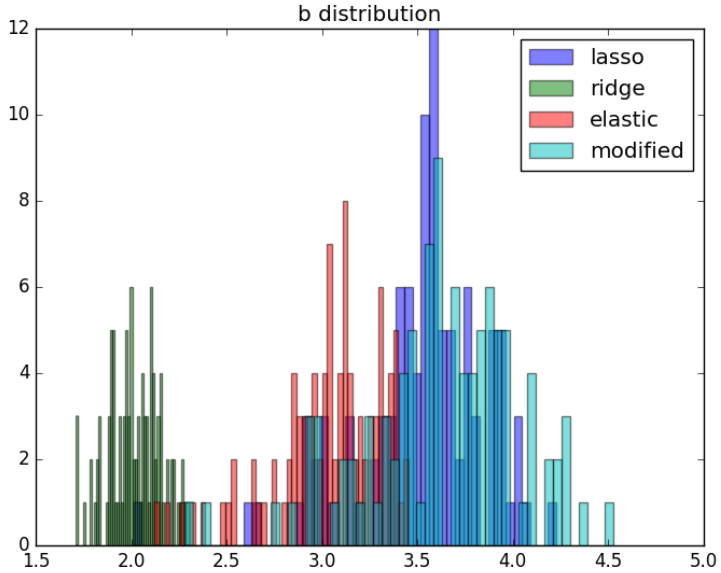
### 6.1 $1_{st}$ run (regularized b)

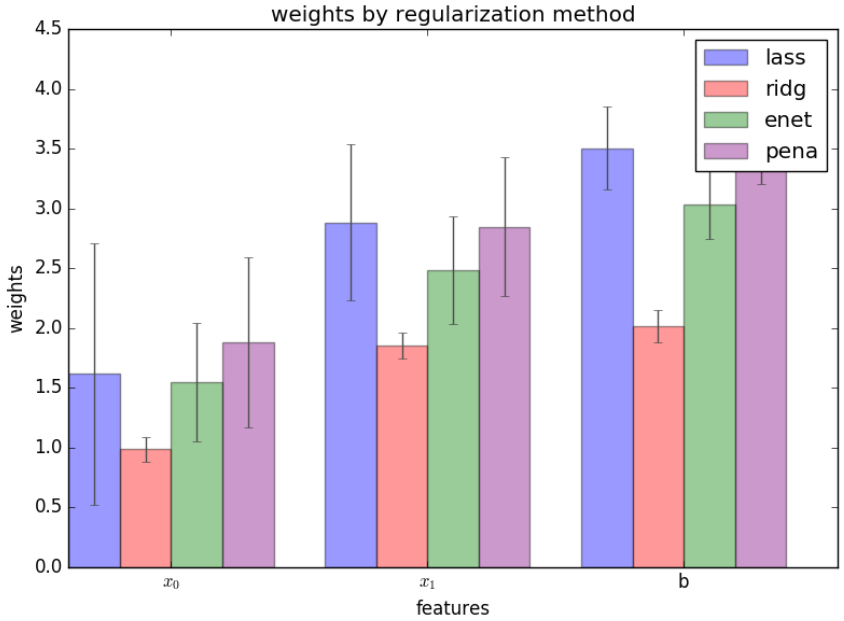
2 variables: $x_0$ known, $x_1$ unknown
b regularized
fix hyperparmeters to predefined value
repeat the following 100 times:
generate data, run the selected regularizers, record $\theta$




Note here the axes are wrongly labled. The y axis should be number count and x axis be weight.


$x_1$ distribution
lasso
ridge
elastic
modified
0
2
4
6
8
10
12
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5


b distribution
lasso
ridge
elastic
modified
0
2
4
6
8
10
12
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0




This experiment clearly shows that lasso is able to drive unknown factor to 0 in the unnormalized case (since $x_1 = 2\ x_0$, $x_1$ indeed get all the zero)

The flaw in this run is the lack of a validation set to set hyperparameters, which is addressed in second run 6.2.

### 6.1.1 data gen

Data n = 100:

h = linspace(-2.5, 1, n)

$x_0 \sim h$

$x_1 \sim 2h$

$y = h > 0.5$

r (known risk factors) = [1, 0]

Loss function is the negative loss likelihood of the logistic regression model.

Optimizer: AdaDelta

Number of Epoch: 1000

Regulizers: elastic net, lasso, ridge, penalty

## 6.2 $2_{nd}$ run (unregularized b, validation)

2 variables: $x_0$ known, $x_1$ unknown

b unregularized

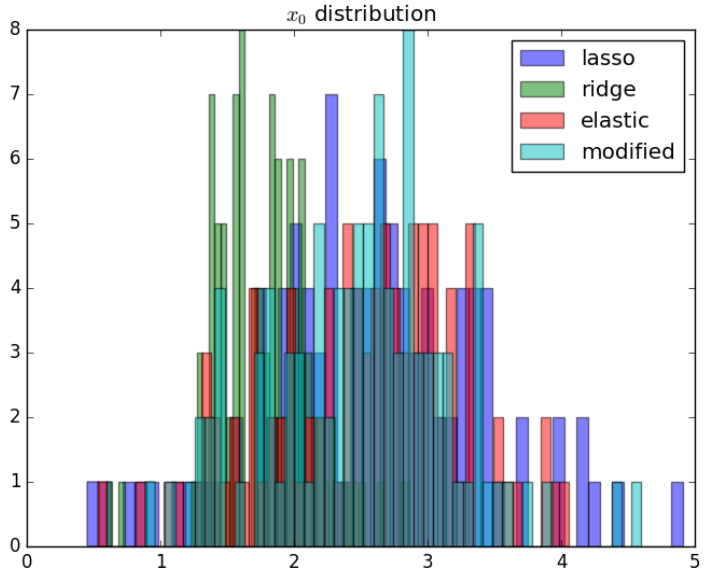
generate two datasets ($x_1 = 2x_0$), one for training, one for validation

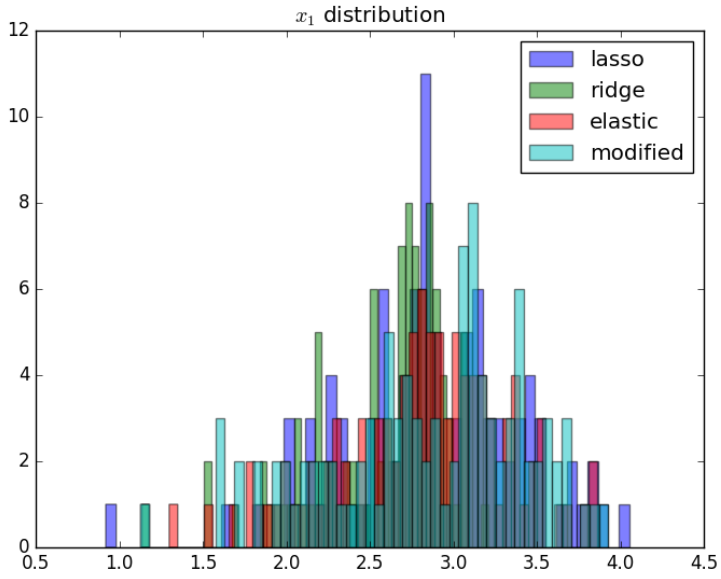
parameter search over the different hyperparams of the regularizers

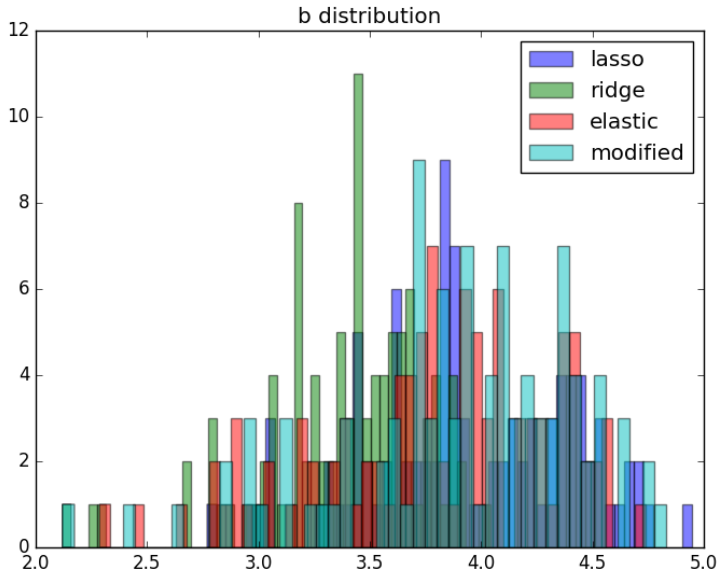
for each regularizer, use the hyperparmeters that acheives the minimal loss

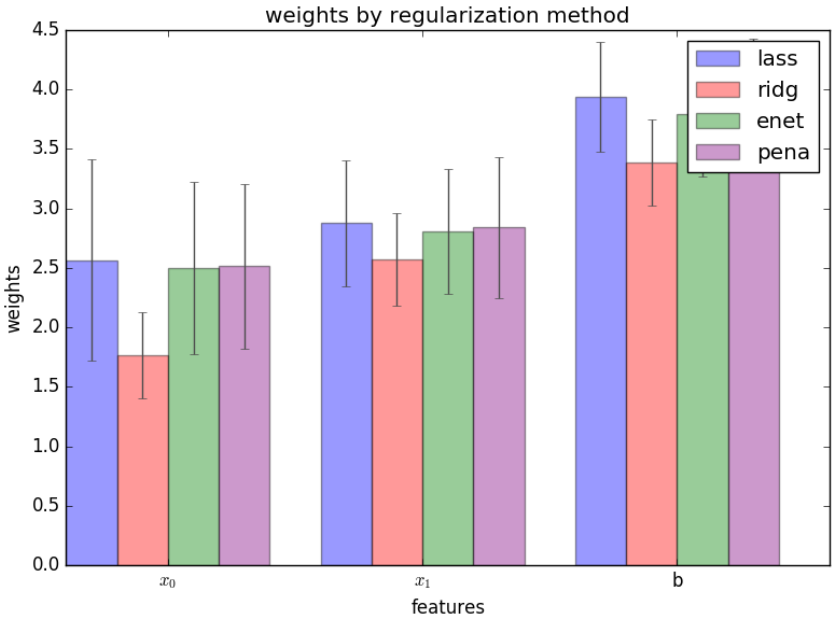
repeat the following 100 times:

generate data, run the selected regularizers, record $\theta$




$x_1$ distribution
lasso
ridge
elastic
modified
0
2
4
6
8
10
12
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5


b distribution
lasso
ridge
elastic
modified
0
2
4
6
8
10
12
2.0
2.5
3.0
3.5
4.0
4.5
5.0




No dicernable pattern in this run as data is unormalized. The addition of validation set makes the comparison fair between methods. The issue of normalization is addressed in 6.3

### 6.2.1 data gen

Data n = 100:

h = linspace(-2.5, 1, n)

$x_0 \sim h$

$x_1 \sim 2h$

$y = h > 0.5$

r (known risk factors) = [1, 0]

Loss function is the negative loss likelihood of the logistic regression model.

Optimizer: AdaDelta

Number of Epoch: 1000

Regulizers: elastic net, lasso, ridge, penalty

## 6.3 $3_{rd}$ run (data normalized, eye penalty)

2 variables: $x_0$ known, $x_1$ unknown

b unregularized

generate two datasets (x2 = 2x1), one for training, one for validation

normalize the data to 2 mean and 2 variance (validaton data is normalized using mean and variance for the training data)

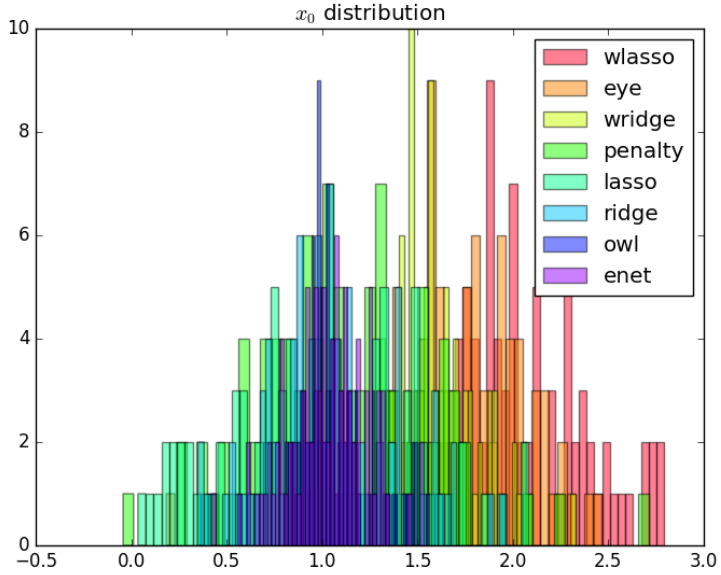
parameter search over the different hyperparams of the regularizers

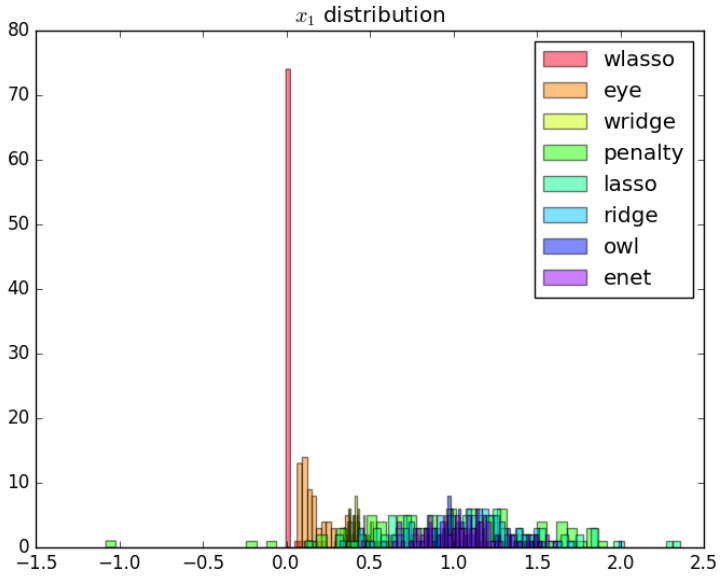
for each regularizer, use the hyperparmeters that acheives the minimal loss

repeat the following 100 times:

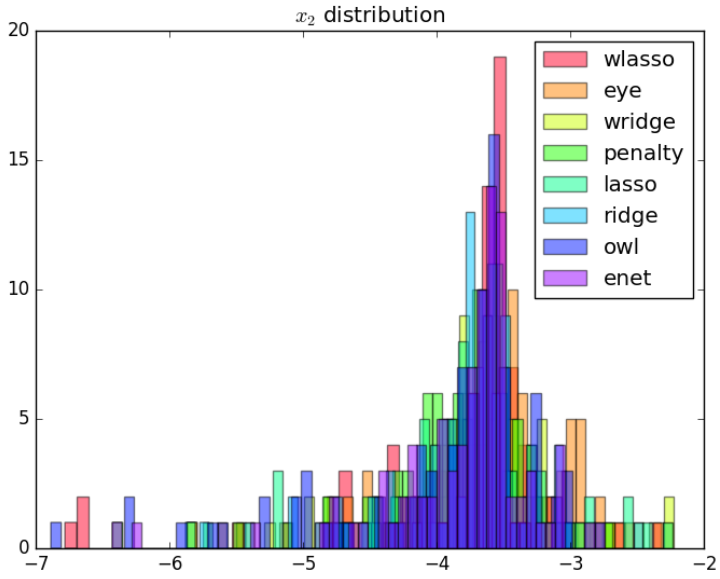
generate data, normalize data, run the selected regularizers, record $\theta$

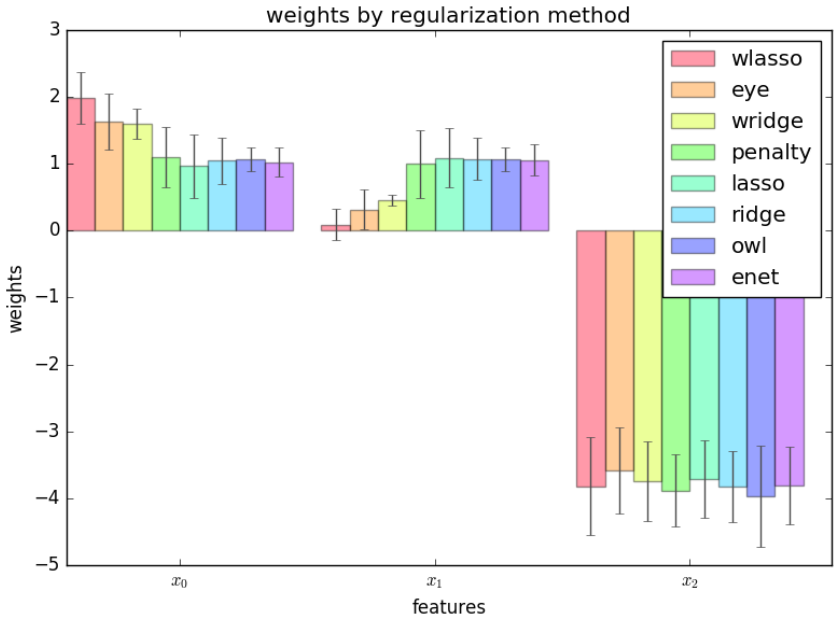
The choosing criteria is still loss b/c AUROC is always going to be 1 in the deterministic case:






Most weights of $x_1$ for weighted lasso and eye are pushed to 0, confirming our intuition.






In the next experiment 6.4, we explore the effect of noise on regularization.

### 6.3.1 data gen

Data n = 100:

h = linspace(-2.5, 1, n)

$x_0 \sim h$

$x_1 \sim 2h$

$y = h > 0.5$

r (known risk factors) = [1, 0]

Loss function is the negative loss likelihood of the logistic regression model.

Optimizer: AdaDelta

Number of Epoch: 1000

Regulizers: elastic net, lasso, ridge, penalty, eye, weighted lasso, weighted ridge, ordered weighted lasso

## 6.4 $4_{th}$ run (noise added)

2 variables: $x_0$ known, $x_1$ unknown

b unregularized

generate two datasets, one for training, one for validation

normalize the data to 2 mean and 2 variance (validaton data is normalized using mean and variance for the training data)

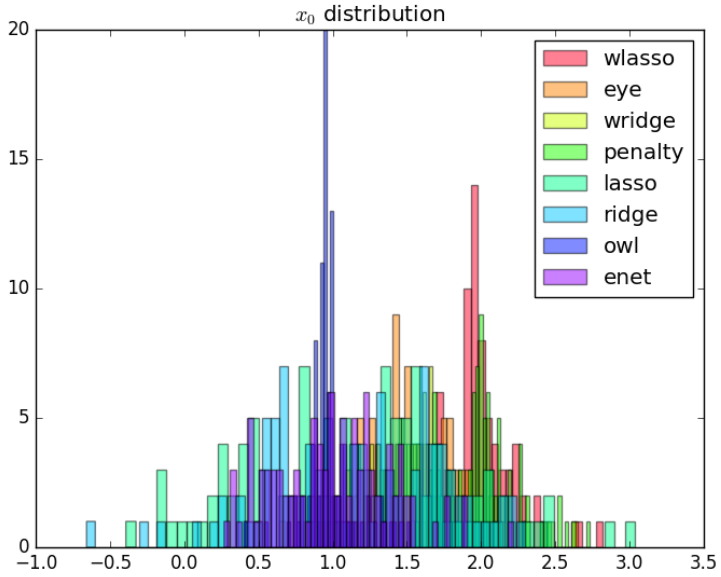
parameter search over the different hyperparams of the regularizers

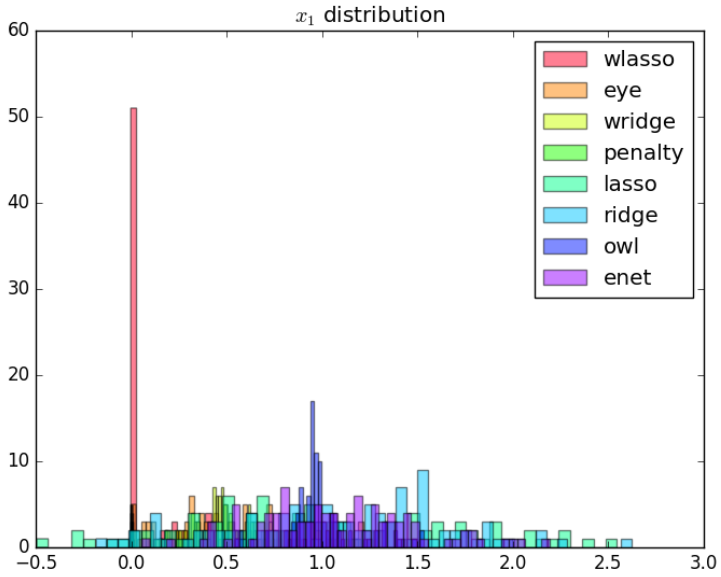
for each regularizer, use the hyperparmeters that acheives the minimal loss

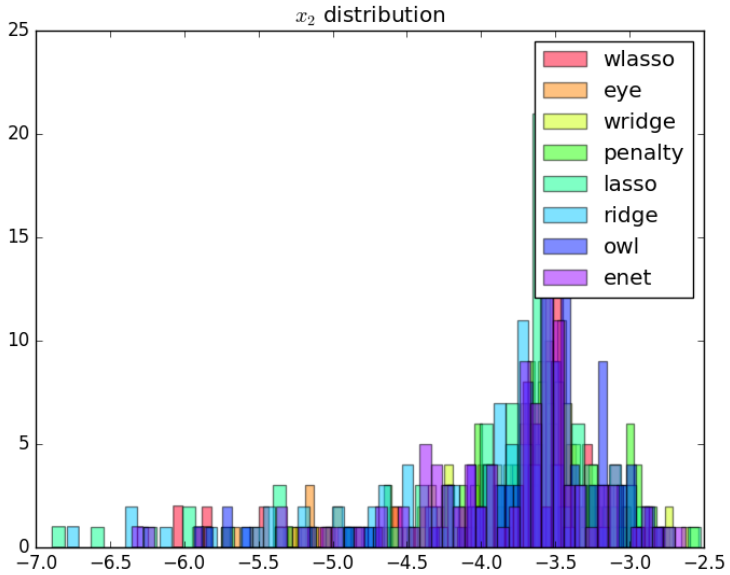
repeat the following 100 times:

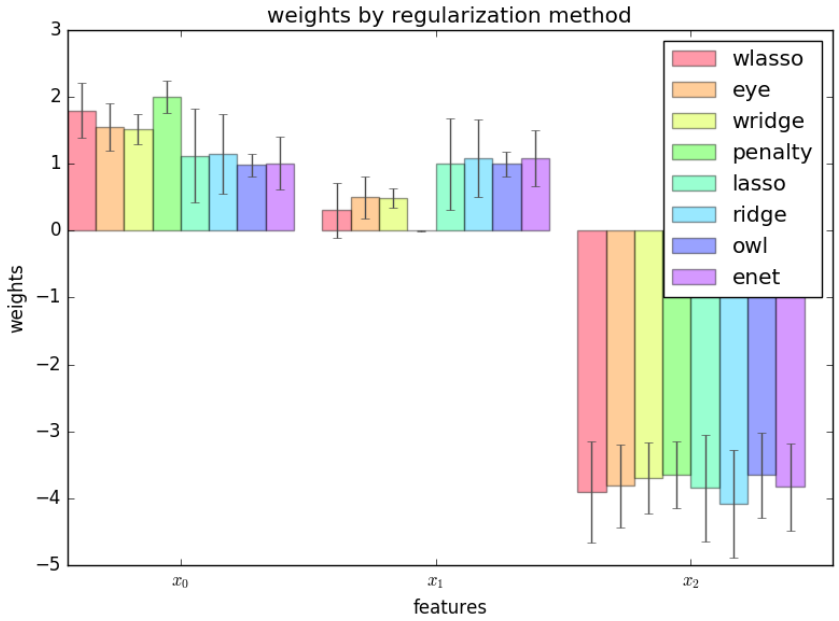
generate data ($x_i$ = Uniform(0..4) h + N(0,0.2)), normalize data, run the selected regularizers, record $\theta$

The choosing criteria is loss




$x_1$ distribution
wlasso
eye
wridge
penalty
lasso
ridge
owl
enet
60
50
40
30
20
10
0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0

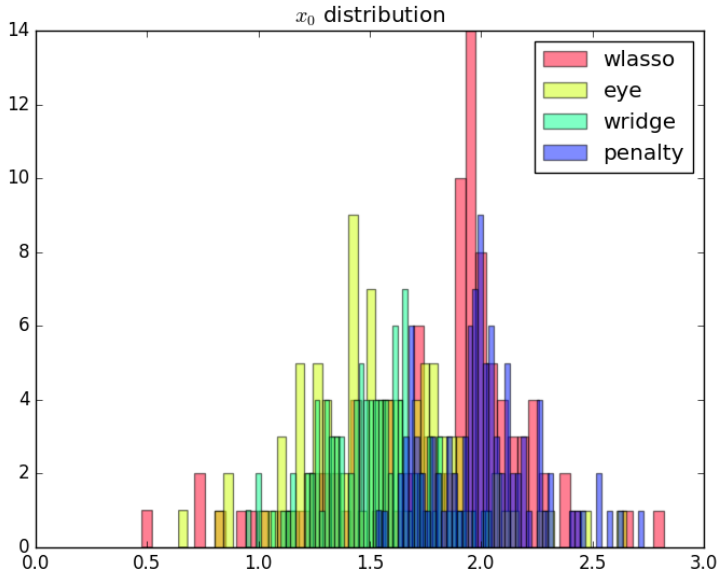

$x_2$ distribution
wlasso
eye
wridge
penalty
lasso
ridge
owl
enet
25
20
15
10
5
0
−7.0
−6.5
−6.0
−5.5
−5.0
−4.5
−4.0
−3.5
−3.0
−2.5

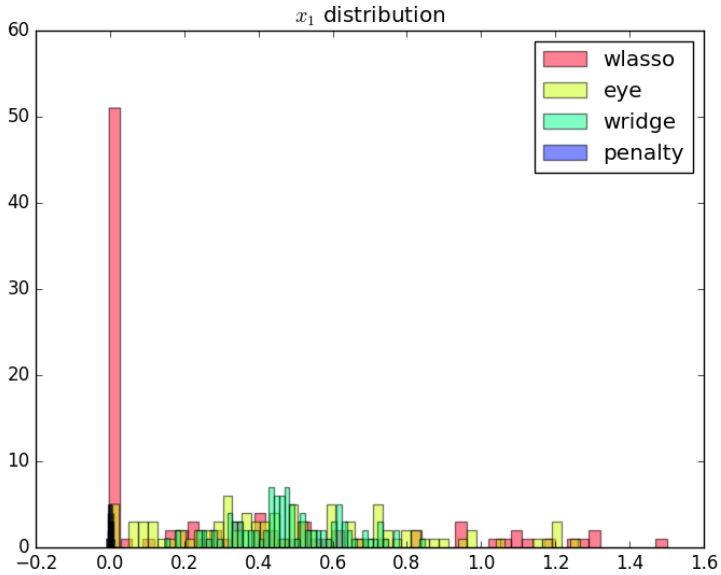



hyper parameter used:

- enet(0.01, 0.2)
- eye(array([ 1., 0.]), 0.01, 0.4)
- lasso(0.0001)
- OWL([2, 1], 0.01)
- penalty(array([ 1., 0.]), 0.1, 1.0)
- ridge(0.001)
- weightedLasso(array([ 1., 2.]), 0.01)
- weightedRidge(array([ 1., 2.]), 0.01)

The sparsity in penalty can be explained as I placed no constraint on the known risk factor (l1 ratio is 1), so it only regularizes $x_1$ not $x_0$



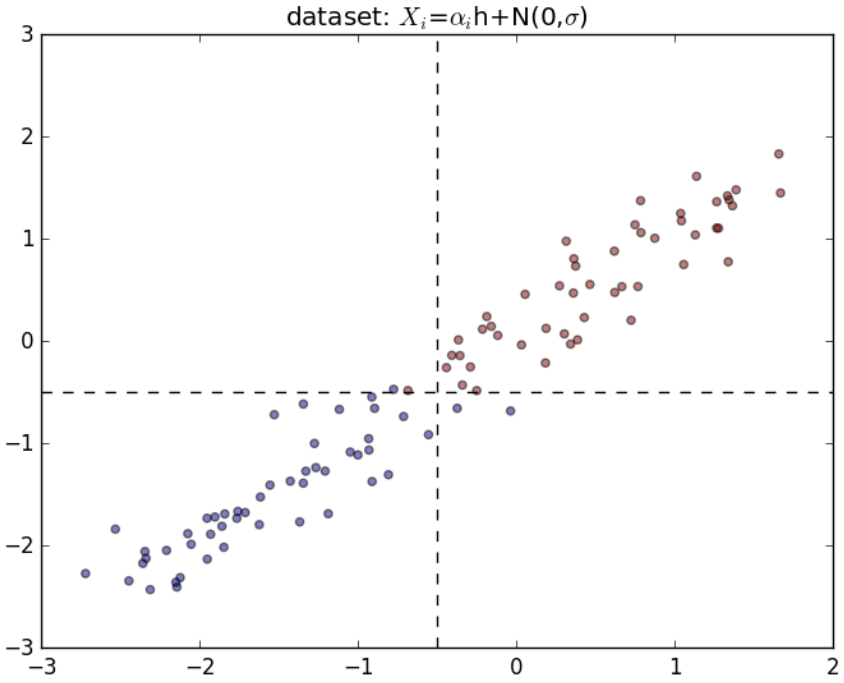

The noise in this experiment is chosen such that the model prefers regularization over unregularization (if is linearly separable, then grid search would favor unregularized case).

The next few experiments illustrate the performance of regularization on

high dimensional data.

### 6.4.1 data gen

Data n = 100:




h = linspace(-2.5, 1, n)
$x_0 \sim Uniform(1..4)h + N(0, 0.2)$
$x_1 \sim Uniform(1..4)h + N(0, 0.2)$
$y = h > 0.5$
r (known risk factors) = [1, 0]

Loss function is the negative loss likelihood of the logistic regression model.

Optimizer: AdaDelta

Number of Epoch: 1000

Regulizers: elastic net, lasso, ridge, OWL, weighted lasso, weighted ridge, penalty, eye penalty

## 6.5 $5_{th}$ run (nd data, sweep r, fix correlation of 0.04, fix theta to 1)

b unregularized

generate two datasets, one for training, one for validation

normalize the data to 2 mean and 2 variance (validaton data is normalized using mean and variance for the training data)

parameter search over the different hyperparams of the regularizers (each of the final candidate has loss around 0.083)

for each regularizer, use the hyperparmeters that acheives the minimal loss

repeat the following 10 times:

generate data (detailed in nd data generation section), normalize data, run the selected regularizers, record $\theta$

The choosing criteria is loss

KL divergence metric filtering for relevant features:

eye: 2.5722261048

wlasso: 5.18104309657

wridge: 6.8364694347

lasso: 18.9613782735

ridge: 12.7547711529

owl: 13.5265637342

enet: 17.7231341012

KL divergence metric including irrelevant features:

eye: 13.1307145901

wlasso: 7.55507729218

wridge: 11.5881850514

lasso: 31.1710069808

ridge: 16.9635832109

owl: 17.5479982613

enet: 30.2439873411

kl/emd$_{\text{metricvisual}}$ (generated using gen$_{\text{result}}$.py:gen$_{\text{ndlosscsv}}$, is in .pages format so assumes mac, included in attachment)

### 6.5.1 data gen (genPartitionData)

Data n = 5000

n relevant groups (nrgroups) = 11

n irrelevant group (nirgroups) = 11

correlated variables pergroup (npergroup) = 10

$h_i \sim Uniform(-3, 1, n)$

$\theta_{\text{i}} = 1 \; \forall$ i

$x_{i,j} \sim Uniform(1..2)h_i + N(0, 0.2)$ for i $\in$ [n] for j $\in$ [npergrop]

$y = \frac{\sum_{i=1}^{nrgroups} h_i \theta_i}{\sum_{i=1}^{nrgroups} |\theta_i|} > -1$

r (known risk factors): for each correlated variable group, putting in one more known risk factor than the previous group

Loss function is the negative loss likelihood of the logistic regression model.

Optimizer: AdaDelta

Number of Epoch: 1000

Regulizers: elastic net, lasso, ridge, OWL, weighted lasso, weighted ridge, eye penalty

### 6.6 TODO $6_{\text{th}}$ run (sweep corelation, fix r, fix theta to 1) (to be fused with weekly report)

b unregularized

generate two datasets, one for training, one for validation

normalize the data to 2 mean and 2 variance (validaton data is normalized using mean and variance for the training data)

parameter search over the different hyperparams of the regularizers (each of the final candidate has loss around 0.083)

for each regularizer, use the hyperparmeters that acheives the minimal loss

repeat the following 10 times:

generate data (detailed in nd data generation section), normalize data, run the selected regularizers, record $\theta$

The choosing criteria is loss

construct a covariance matrix with 10 different blocks on diagnal with variables in each block having a different covariance value. This experiment is to discover the relationship between noise level and credibility.

- run(eye(r, 0.05), outdir="result$_{\text{eye}}$")
- run(enet(0.01, 0.1), outdir="result$_{\text{enet}}$")
- run(lasso(0.0005), outdir="result$_{\text{lasso}}$")
- run(ridge(0.01), outdir="result$_{\text{ridge}}$")
- run(weightedLasso(w1, 0.005), outdir="result$_{\text{wlasso}}$")
- run(weightedRidge(w1, 0.01), outdir="result$_{\text{wridge}}$")
- run(OWL(owl1, 0.001), outdir="result$_{\text{owl}}$")

### 6.6.1 general nd data generation

Data n = 2000

n relevant groups (nrgroups) = 11

n irrelevant group (nirgroups) = 0

correlated variables pergroup (npergroup) = 4

Given a covariance matrix C

Do cholesky decomposition: C = A $A^T$

$h \sim N(0, 1, shape = (n, d))$

x = h $A^T$

$\theta_i = 1 \; \forall$ i

$y = X\theta + N(0, 5, n) > 0$

note that the noise added to y makes the problem linearly inseparable so that regularization makes sense (otherwise validation will always choose the least regularized classifier).

r (known risk factors): for each dependent group, set half as known, half as unknown

Loss function is the negative loss likelihood of the logistic regression model.

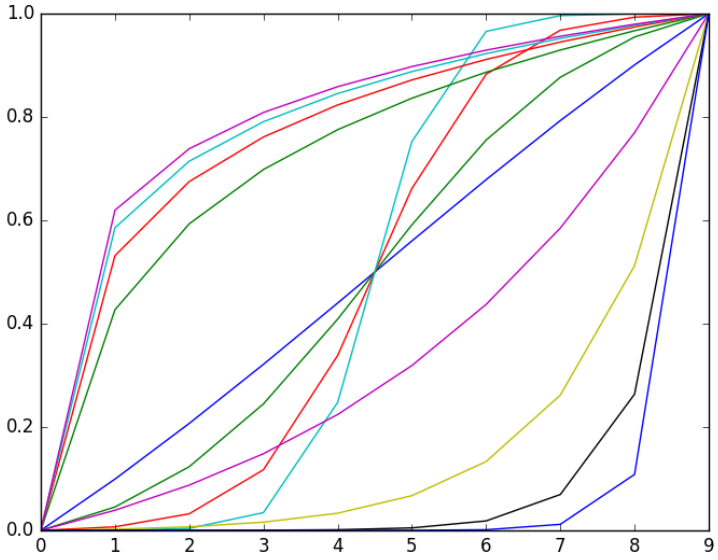
Optimizer: AdaDelta

Number of Epoch: 1000

Regulizers: elastic net, lasso, ridge, OWL, weighted lasso, weighted ridge, eye penalty

## 6.7 TODO $7_{th}$ run (sweep fractional r, fix correlation, fix theta) (to be fused with weekly report)

To extend r to be fractional, we consider setting r according to parametrized functions: log, exp, sigmoid, and linear.




### 6.7.1 data generation

fix correlation of 0.99, generate groups of 10 data via a block diagonal covariance matrix. The rest is the same as run 6

## 6.8 TODO $8_{\text{th}}$ run (sweep theta, fix r, fix correlation)

Try different theta in data generation. I expect this will not make a difference in dependent groups compared to run 5, 6, and 7.

What I mean is that different theta in the same dependent group will have the same effect for all regularizations as long as the sum of theta is the same. So it is questionable whether or not to run this experiment.

## 6.9 TODO real data

After graduating from simulated data, we will apply eye penalty to C. difficile prediction.

# 7 summary of regularizations used in this work

### 7.0.1 eye penalty

$q(\theta) = 2\beta\|(1-r)\odot\theta\|_1 + (1-\beta)\|r\odot\theta\|_2^2$

$pena(\theta) = \lambda q(\theta)$

where r $\in \{0,1\}^d$, $\theta \in$l $\mathbb{R}^d$, $\lambda \in \mathbb{R}_+$, $\beta \in$ (0,1) ($\beta$ is also called l1 ratio in this text)

For any constant c

$pena(\theta) = c$

is convex because pena is convex (addition of positively weighted norms)

similarly, $q(\theta) = c$ is also convex

c can be chosen so that slope in the first quadrant between known risk factor x and unknown risk factor is -1

we define eye norm as a an atomic norm $\|\cdot\|_A$ as introduced in cite:chandrasekaran2012convex

$\|x\|_A := \inf\{t > 0 \mid x \in tconv(A)\}$

Let $A = \{x \mid q(x) = \frac{\beta^2}{1-\beta}\}$, we get the eye penalty

Note that A is already a convex set, adding in scaling factor $\lambda$, we have

$$eye(x) = \lambda \inf\{t > 0 \mid x \in \{tx \mid q(x) = \frac{\beta^2}{1-\beta}\}\} \tag{2}$$

We will show that (2) is equivalent to (1)

1. derivation of (2)

   The main intuition is to set c so that the slope in the first quadrant between known risk factor x and unknown risk factor is -1. Since we only care about this interaction between known and unknown risk factors and that $x \mid pena(x) = c$ is symmetric about origin, WLOG, we let y be the unknown feature and x be the known risk factor with constraint y $\geq 0$, x $\geq 0$.

$$\lambda[2\beta y + (1-\beta)x^2] = c \tag{3}$$

$$\rightarrow 2\beta y + (1-\beta)x^2 = \frac{c}{\lambda} \tag{4}$$

$$\rightarrow y = \frac{c}{2\lambda\beta} - \frac{(1-\beta)x^2}{2\beta} \tag{5}$$

$$\rightarrow y = 0 \Rightarrow x = \sqrt{\frac{c}{\lambda(1-\beta)}} \tag{6}$$

$$\rightarrow f'(x) = -\frac{(1-\beta)}{\beta}x \tag{7}$$

$$\rightarrow f'(\sqrt{\frac{c}{\lambda(1-\beta)}}) = -\frac{1-\beta}{\beta}\sqrt{\frac{c}{\lambda(1-\beta)}} = -1 \tag{8}$$

$$\rightarrow c = \frac{\lambda\beta^2}{1-\beta} \tag{9}$$

$$\rightarrow 2\beta y + (1-\beta)x^2 = \frac{\beta^2}{1-\beta} \tag{10}$$

Thus, we just need q(x) = $\frac{\beta^2}{1-\beta}$

2. properties:

   - A is symmetric about origin (x ∈ A then -x ∈ A), so this is a norm
     (a) eye(t $\theta$) = |t| eye($\theta$)
     (b) eye($\theta$ + $\beta$) ≤ eye($\theta$) + eye($\beta$)
     (c) eye($\theta$) = 0 iff $\theta$ = 0
   - $\beta$ conserve the shape of contour

     *Proof.* consider the contour $B_1$ = {x: $eye_{\beta_1}$}(x) = t} and$B_2$ = {x: $eye_{\beta_2}$}(x) = t}
     We want to show $B_1$ is similar to $B_2$
     case1: t = 0, then $B_1 = B_2 = \{0\}$ by property a3
     case2: t ≠ 0
     we can equivalently write $B_1$ and $B_2$ as: (by definition and a1 and q convex)
     $B_1$ = t {x: x ∈ {x | $q_{\beta_1}$(x) = $\frac{\beta_1^2}{1-\beta_1}$ }}
     $B_2$ = t {x: x ∈ {x | $q_{\beta_2}$(x) = $\frac{\beta_2^2}{1-\beta_2}$ }}

let $B_{1'} = \{x: x \in \{x \mid q_{\beta_1}(x) = \frac{\beta_1^2}{1-\beta_1}\}\}$ and $B_{2'} = \{x: x \in t \{x \mid q_{\beta_2}(x) = \frac{\beta_1^2}{1-\beta_2}\}\}$

Claim: $B_{2'} = \frac{\beta_2(1-\beta_1)}{\beta_1(1-\beta_2)} B_{1'}$

It should be clear that if this claim is true then $B_1$ is similar to $B_2$ and we are done

take $x \in B_{1'}$

then $q_{\beta_1}(x) = 2\beta_1\|(1-r) \odot x\|_1 + (1-\beta_1)\|r \odot x\|_2^2 = \frac{\beta_1^2}{1-\beta_1}$

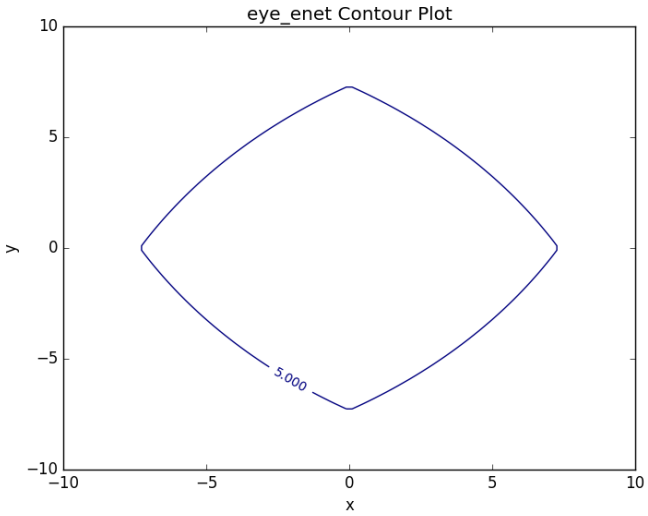
let x' = $\frac{\beta_2(1-\beta_1)}{\beta_1(1-\beta_2)}$ x

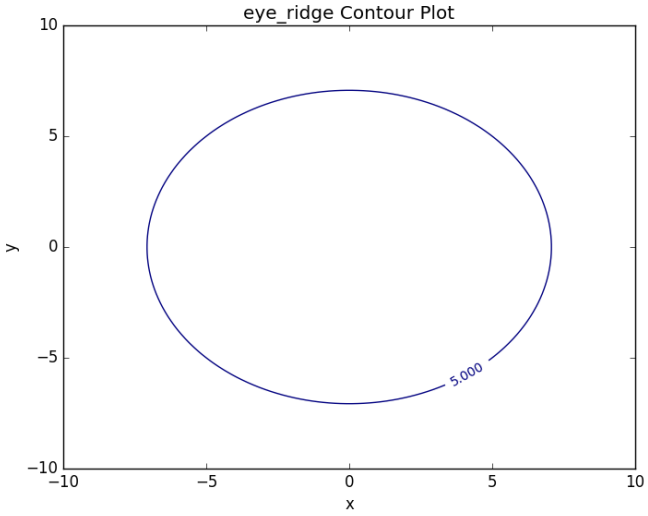
$$
\begin{aligned}
q_{\beta_2}(x') &= 2\beta_2\|(1-r) \odot x'\|_1 + (1-\beta_2)\|r \odot x'\|_2^2 \\
&= \frac{2\beta_2^2(1-\beta_1)}{\beta_1(1-\beta_2)}\|(1-r) \odot x\|_1 + \frac{\beta_2^2(1-\beta_1)^2}{\beta_1^2(1-\beta_2)}\|r \odot x\|_2^2 \\
&= \frac{\beta_2^2(1-\beta_1)}{\beta_1^2(1-\beta_2)}(2\beta_1\|(1-r) \odot x\|_1 + (1-\beta_1)\|r \odot x\|_2^2) \\
&= \frac{\beta_2^2(1-\beta_1)}{\beta_1^2(1-\beta_2)}\frac{\beta_1^2}{1-\beta_1} \\
&= \frac{\beta_2^2}{1-\beta_2}
\end{aligned}
$$

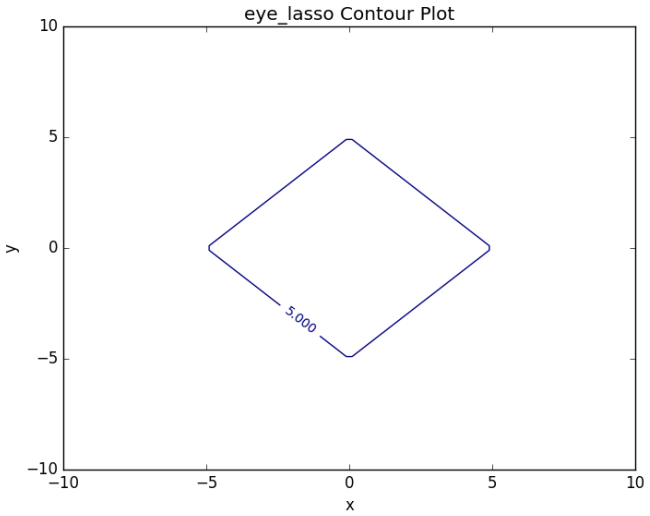
so x' $\in B_{2'}$. Thus $\frac{\beta_2(1-\beta_1)}{\beta_1(1-beta_2)} B_{1'} \subset B_{2'}$. The other direction is similarly proven. □

- eye as a generalization of elastic net, lasso, and ridge

  By relaxing the constraint of r from binary to float, we can recover elastic net(setting r=0.5 * $\mathbf{1}$). Even without extending r, we can recover ridge (r= $\mathbf{1}$) and lasso (r= $\mathbf{0}$)


eye_enet Contour Plot
10
5
0
−5
−10
y
−10
−5
0
5
10
x
5.000

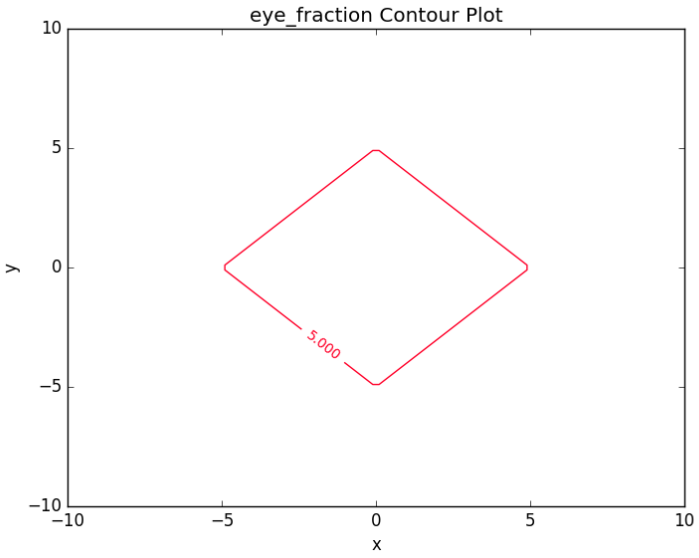

eye_ridge Contour Plot
10
5
0
−5
−10
y
−10
−5
0
5
10
x
5.000



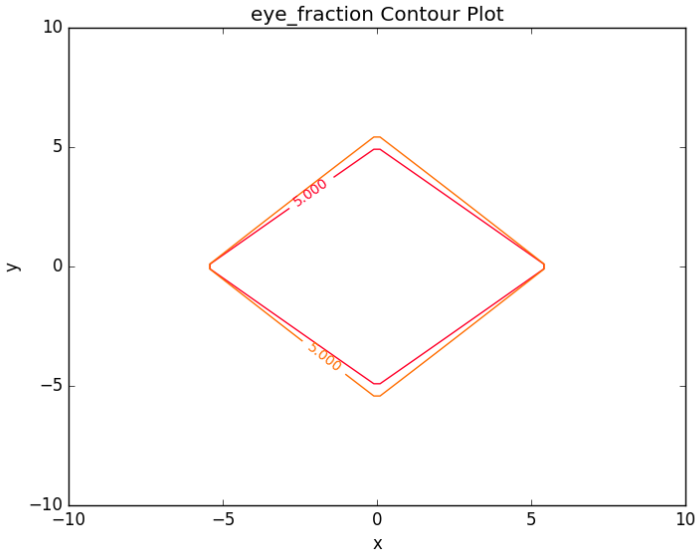

3. extending r to $[0,1]^d$ At times, it makes sense for risk factor to be fractionally weighted (eg. odds ratio in medical documents).

   varying $r_1$ and $r_2$ (in the following plot, $r_2$ are sweep from 0 up to $r_1$ with stepsize of 0.1)

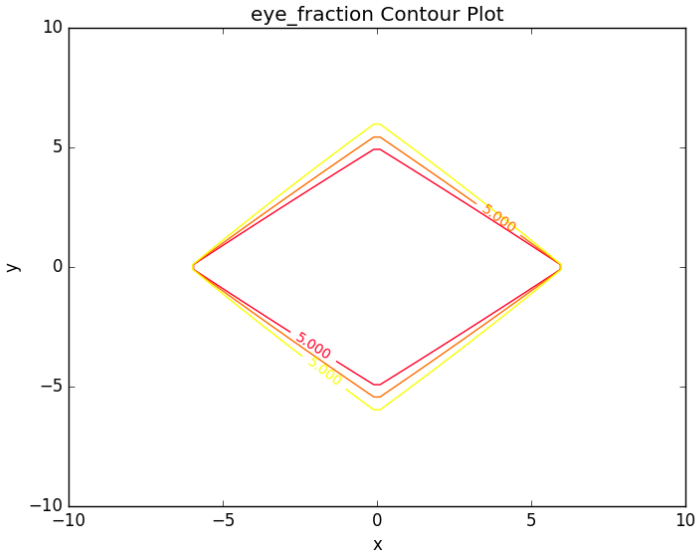
   $r_1 = 0.0$



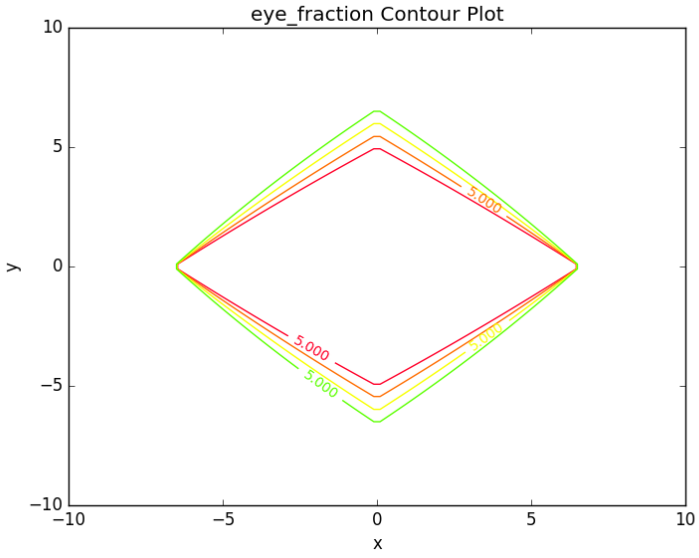

   $r_1 = 0.1$


eye_fraction Contour Plot
10
5
0
−5
−10
y
−10
−5
0
5
10
x
5.000
5.000

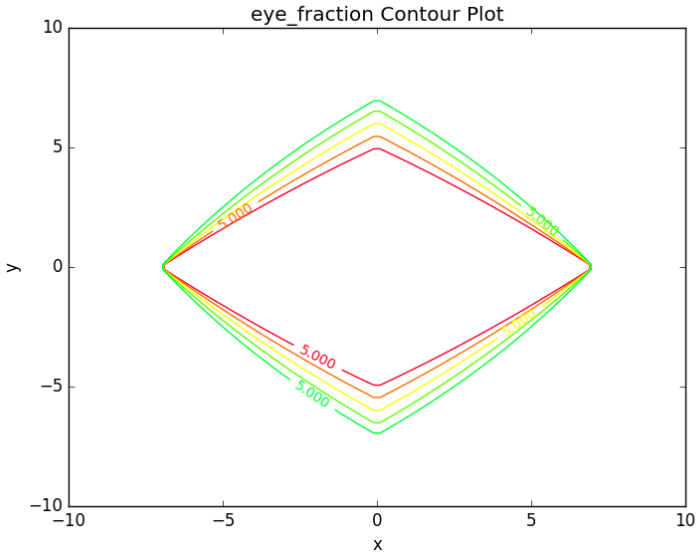
$r_1 = 0.2$


eye_fraction Contour Plot
10
5
0
−5
−10
y
−10
−5
0
5
10
x
5.000
5.000
5.000

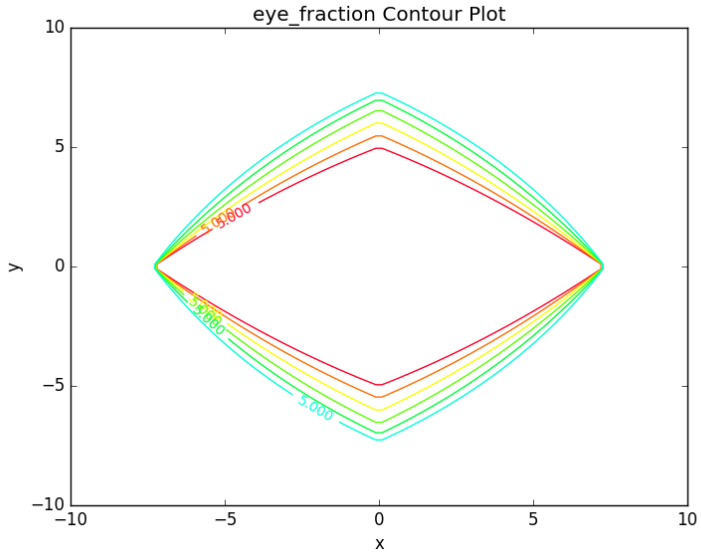
$r_1 = 0.3$


eye_fraction Contour Plot
10
5
0
−5
−10
−10
−5
0
5
10
x
y
5.000
5.000
5.000
5.000
5.000

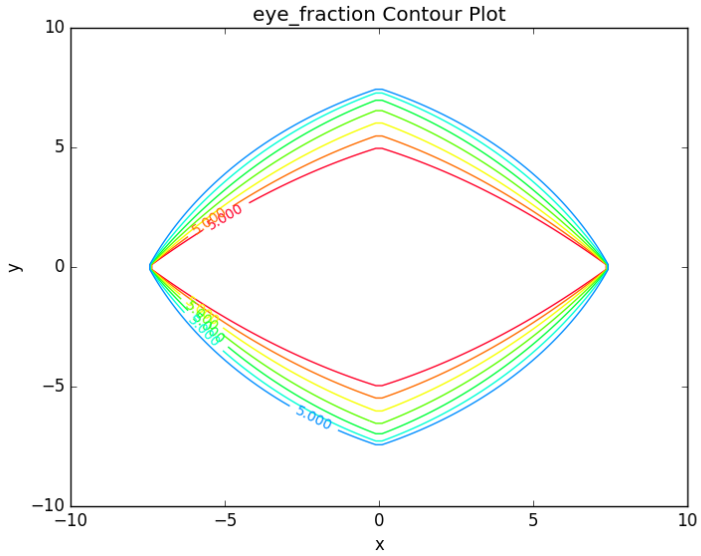
$r_1 = 0.4$


eye_fraction Contour Plot
10
5
0
−5
−10
−10
−5
0
5
10
x
y
5.000
5.000
5.000
5.000
5.000

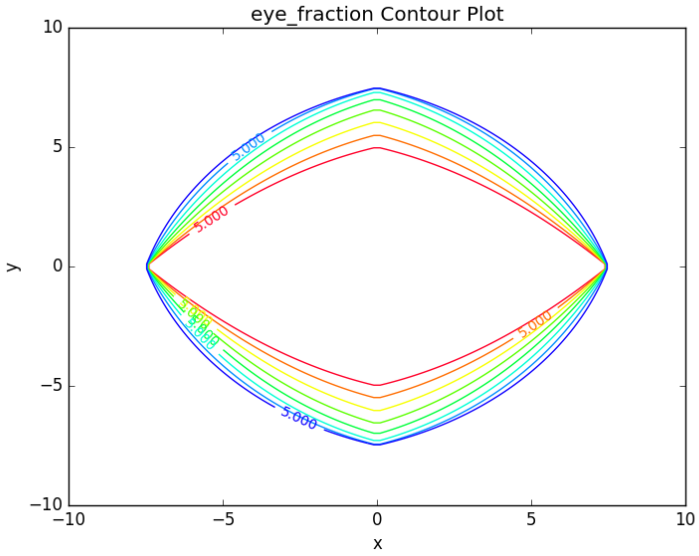
$r_1 = 0.5$


eye_fraction Contour Plot
10
5
0
−5
−10
−10
−5
0
5
10
x
y
5.000

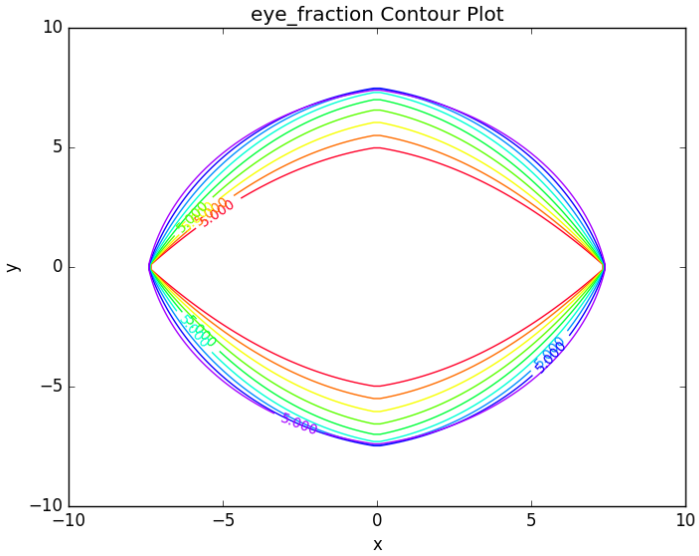
$r_1 = 0.6$


eye_fraction Contour Plot
10
5
0
−5
−10
−10
−5
0
5
10
x
y
5.000

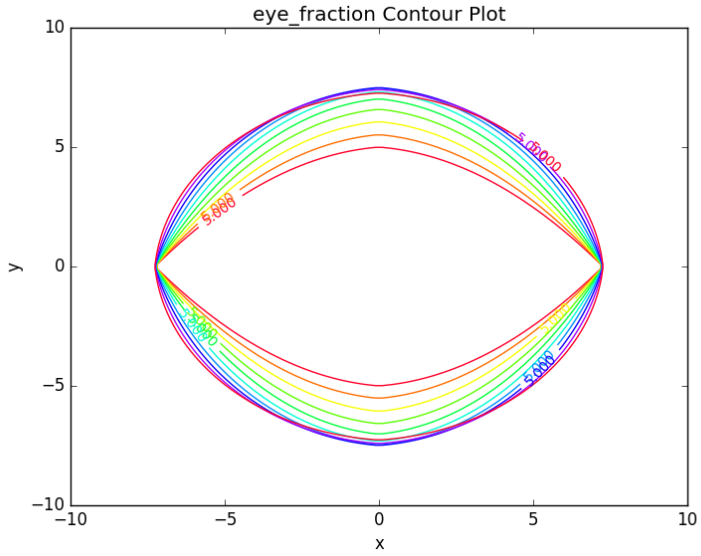
$r_1 = 0.7$


eye_fraction Contour Plot
10
5
0
−5
−10
y
−10
−5
0
5
10
x
5.000
5.000
5.000
5.000
5.000
5.000

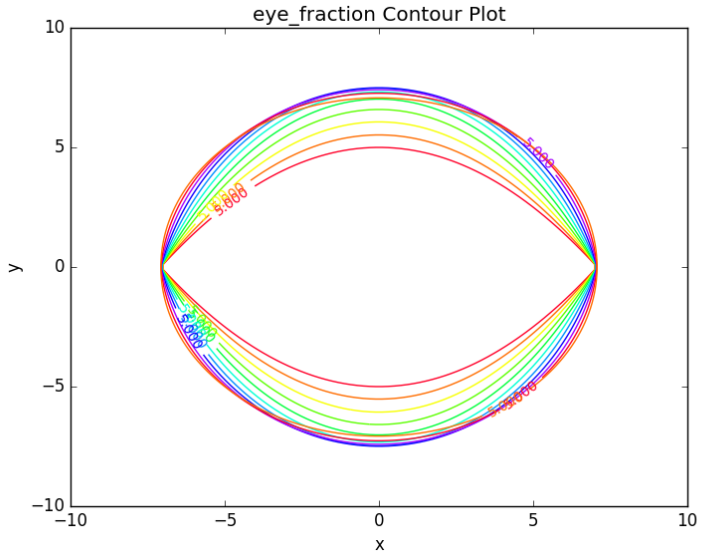
$r_1 = 0.8$


eye_fraction Contour Plot
10
5
0
−5
−10
y
−10
−5
0
5
10
x
5.000
5.000
5.000
5.000
5.000
5.000

$r_1 = 0.9$

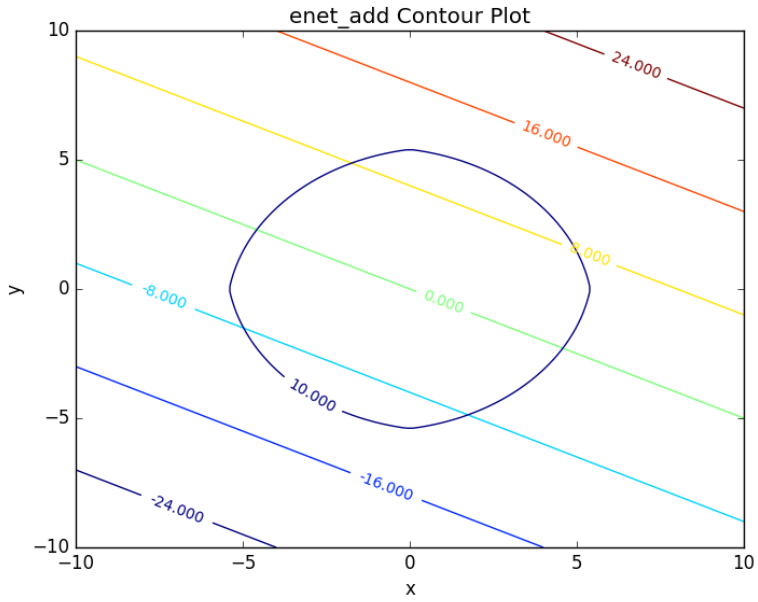



$r_1 = 1.0$



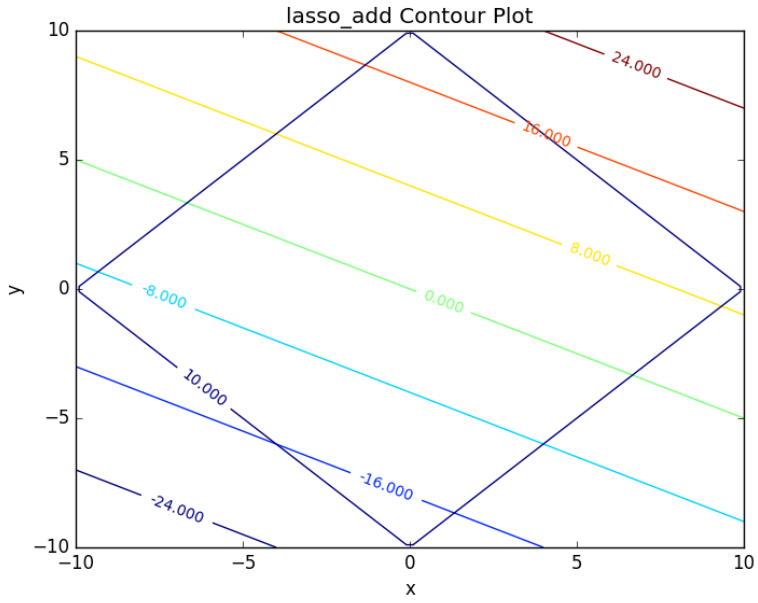

### 7.0.2 elastic net

$\lambda(\beta\|\theta\|_1 + 0.5(1-\beta)\|\theta\|_2^2)$ where $\beta \in [0,1]$




### 7.0.3 lasso [7]

$\lambda\|\theta\|_1$



In the orthogonal case: $X^T X = I$

we have

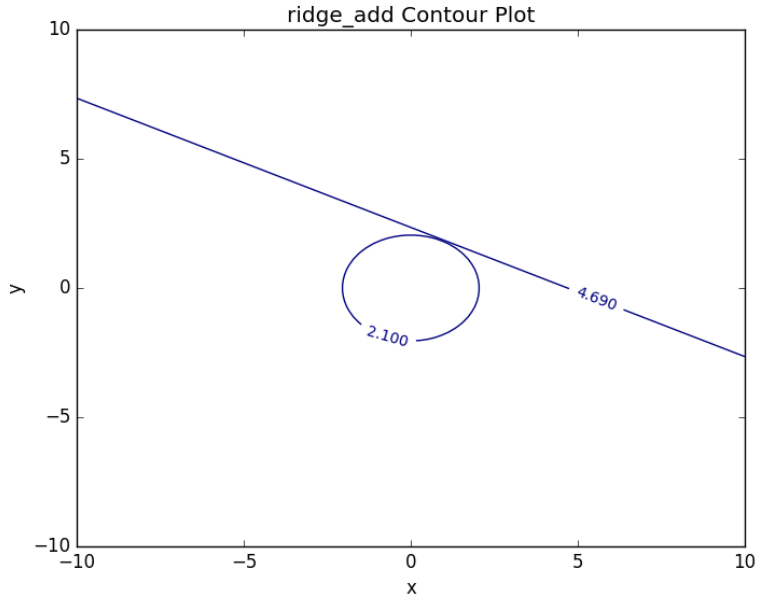
$\theta = S_{n\lambda}(\theta^{OLS})$

where $\theta^{OLS}$ is the value of ordinary least square

and $S_\lambda(x) = xmax(0, 1 - \frac{\lambda}{|x|})$ (the soft threshold function, note that in the limiting case converge to $y = x$)

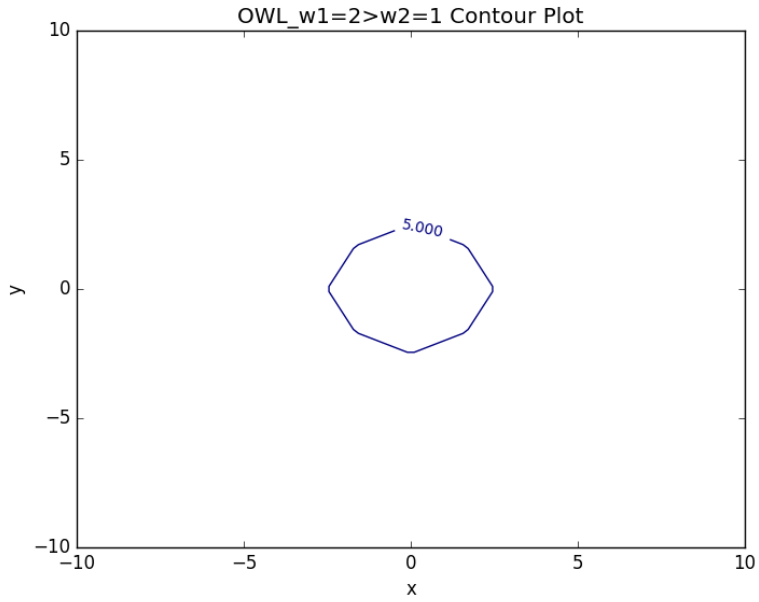
### 7.0.4 ridge

$0.5\lambda\|\theta\|_2^2$

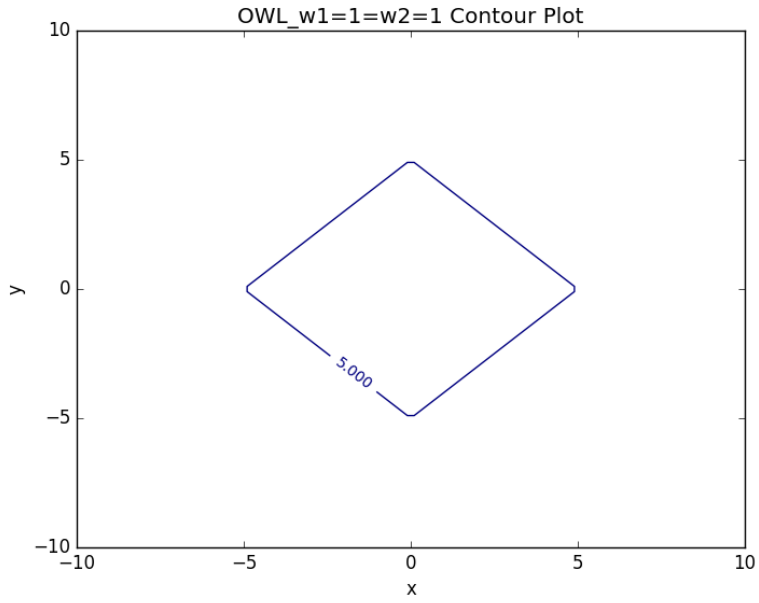



### 7.0.5 OWL

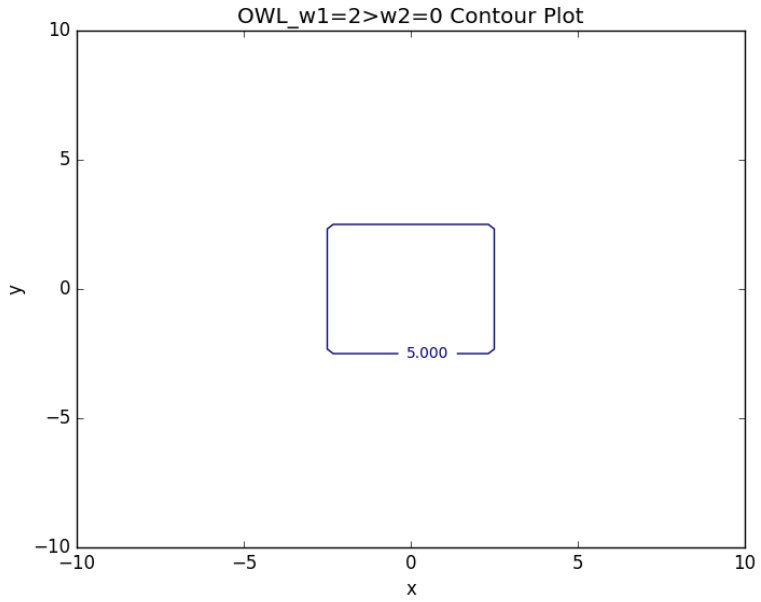
$\lambda \sum_{i=1}^{n} w_i |x|_{[i]}$ where $w \in K_{m+}$ (monotone nonnegative cone)




degenerated case: back to lasso




degenerated case: back to $l_\infty$




some properties:

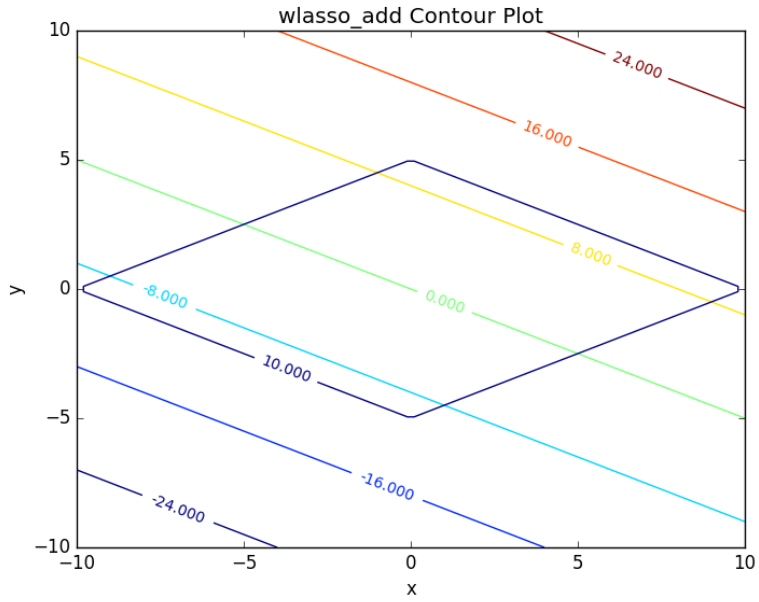
generalization of OSCAR norm

symmetry with respect to signed permutations

in the regular case, the minimal atomic set for this norm is known (the corners are easily calculated)

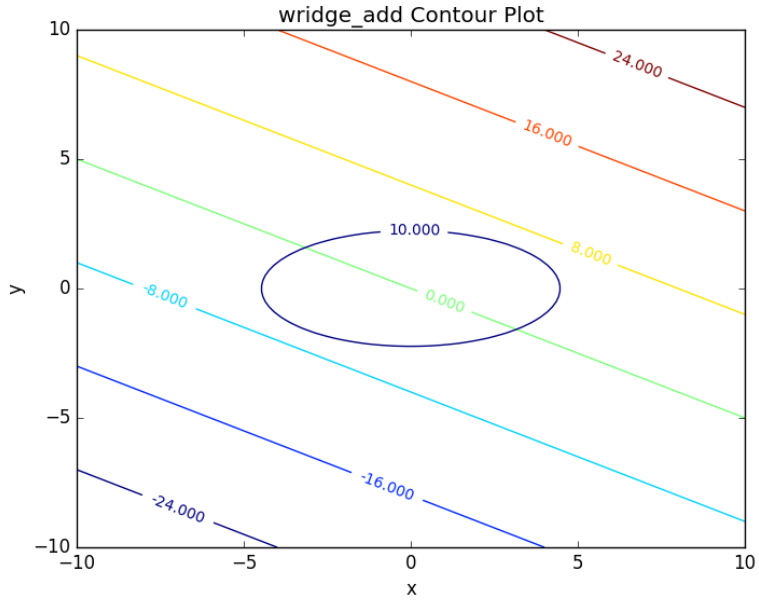
### 7.0.6 weighted lasso

$\lambda \|w \odot \theta\|_1$ where w $\in \mathbb{R}_+^d$



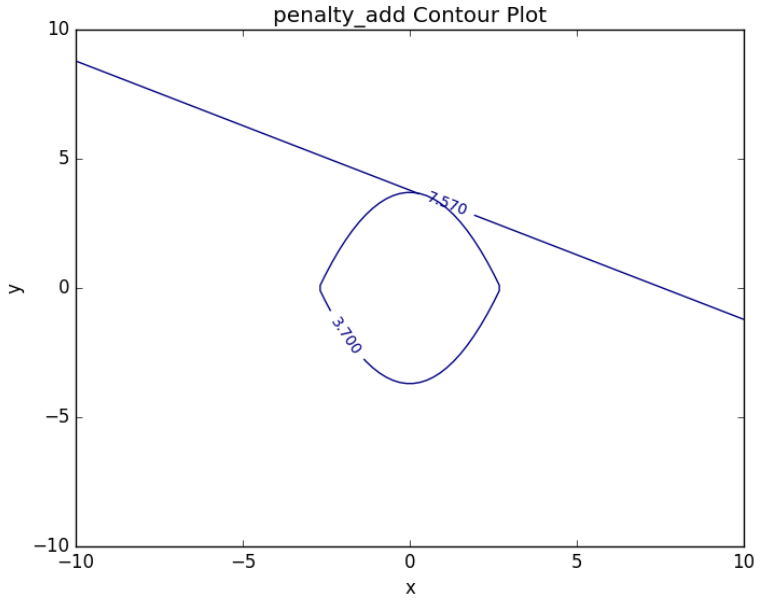

### 7.0.7 weighted ridge

$0.5\lambda\|w \odot \theta\|_2^2$ where w $\in \mathbb{R}_+^d$



### 7.0.8 old penalty

$\lambda(0.5(1-\beta)\|r \odot \theta\|_2^2 + \beta\|(1-r) \odot \theta\|_1)$ where r $\in 0, 1^d$, $\theta \in \mathbb{R}^d$, $\lambda \in \mathbb{R}$, $\beta \in$ [0,1]




## References


[1] Mario AT Figueiredo and Robert D Nowak. Sparse estimation with strongly correlated variables using ordered weighted l1 regularization. *arXiv preprint arXiv:1409.4005*, 2014.

[2] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

[3] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[4] Yun Liu, Collin M Stultz, John V Guttag, Kun-Ta Chuang, Fu-Wen Liang, and Huey-Jen Su. Transferring knowledge from text to predict disease onset. *arXiv preprint arXiv:1608.02071*, 2016.

[5] Geert Meyfroidt, Fabian Güiza, Jan Ramon, and Maurice Bruynooghe. Machine learning techniques to examine large patient databases. *Best Practice & Research Clinical Anaesthesiology*, 23(1):127–143, 2009.

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

[7] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

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

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

# 8 for discussion:

1. validate method for reporting p-value

   $H_0$: mean loss or auroc across method

   $H_1$: mean loss or auroc for each method

   test if $H_0$ different from $H_1$?

# 9 TODO with real data

- minimal implementation with example to jeeheh
- run through sparse matrix

(800000,50000) scipy sparse matrix
format: csr