

# Optimizing ROC Curves with a Sort-Based Surrogate Loss for Binary Classification and Changepoint Detection

Toby Dylan Hocking — [toby.hocking@nau.edu](mailto:toby.hocking@nau.edu)  
joint work with my student Jonathan Hillman  
Machine Learning Research Lab — <http://ml.nau.edu>  
School of Informatics, Computing and Cyber Systems  
Northern Arizona University, USA



## Problem Setting 1: ROC curves for evaluating supervised binary classification algorithms

Problem setting 2: ROC curves for evaluating supervised changepoint algorithms

Proposed surrogate loss for ROC curve optimization: Area Under  $\text{Min}\{\text{FP}, \text{FN}\}$  (AUM)

Empirical results: minimizing AUM results in optimized ROC curves

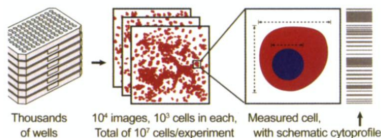
Discussion and Conclusions

# Problem: supervised binary classification

- ▶ Given pairs of inputs  $\mathbf{x} \in \mathbb{R}^p$  and outputs  $y \in \{0, 1\}$  can we learn a score  $f(\mathbf{x}) \in \mathbb{R}$ , predict  $y = 1$  when  $f(\mathbf{x}) > 0$ ?
- ▶ Example: email,  $\mathbf{x}$  = bag of words,  $y$  = spam or not.
- ▶ Example: images. Jones *et al.* PNAS 2009.

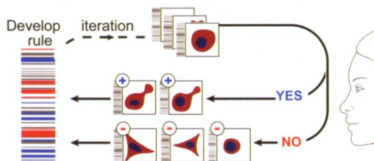
## A Automated Cell Image Processing

Cytoprofile of 500+ features measured for each cell



## B Iterative Machine Learning

System presents cells to biologist for scoring, in batches



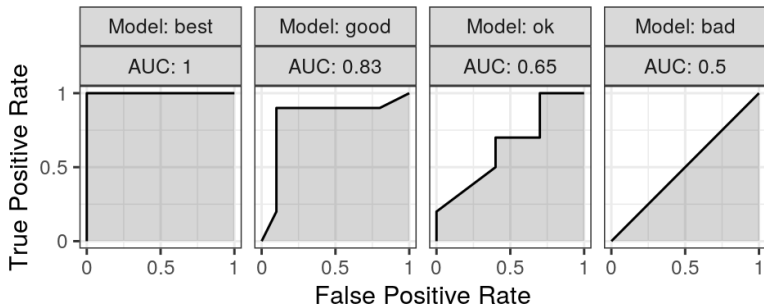
Most algorithms (SVM, Logistic regression, etc) minimize a differentiable surrogate of zero-one loss = sum of:

**False positives:**  $f(\mathbf{x}) > 0$  but  $y = 0$  (predict budding, but cell is not).

**False negatives:**  $f(\mathbf{x}) < 0$  but  $y = 1$  (predict not budding but cell is).

# Receiver Operating Characteristic (ROC) Curves

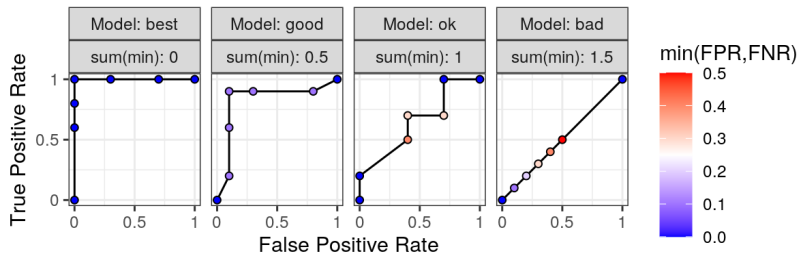
- ▶ Classic evaluation method from the signal processing literature (Egan and Egan, 1975).
- ▶ For a given set of predicted scores, plot True Positive Rate vs False Positive Rate, each point on the ROC curve is a different threshold of the predicted scores.
- ▶ Best classifier has a point near upper left ( $TPR=1$ ,  $FPR=0$ ), with large Area Under the Curve (AUC).



# Research question and new idea

Can we learn a binary classification function  $f$  which directly optimizes the ROC curve?

- ▶ Most algorithms involve minimizing a differentiable surrogate of the zero-one loss, which is not the same.
- ▶ The Area Under the ROC Curve (AUC) is piecewise constant (gradient zero almost everywhere), so can not be used with gradient descent algorithms.
- ▶ We propose to encourage points to be in the upper left of ROC space, using a loss function which is a differentiable surrogate of the sum of  $\min(\text{FPR}, \text{FNR})$ .



Problem Setting 1: ROC curves for evaluating supervised binary classification algorithms

Problem setting 2: ROC curves for evaluating supervised changepoint algorithms

Proposed surrogate loss for ROC curve optimization: Area Under  $\text{Min}\{\text{FP}, \text{FN}\}$  (AUM)

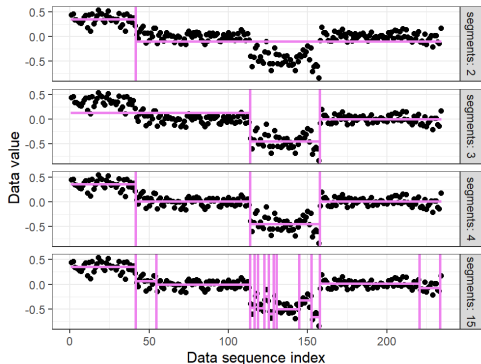
Empirical results: minimizing AUM results in optimized ROC curves

Discussion and Conclusions

# Problem: unsupervised changepoint detection

- ▶ Data sequence  $z_1, \dots, z_T$  at  $T$  points over time/space.
- ▶ Ex: DNA copy number data for cancer diagnosis,  $z_t \in \mathbb{R}$ .
- ▶ The penalized changepoint problem (Maidstone *et al.* 2017)

$$\arg \min_{u_1, \dots, u_T \in \mathbb{R}} \sum_{t=1}^T (u_t - z_t)^2 + \lambda \sum_{t=2}^T I[u_{t-1} \neq u_t].$$

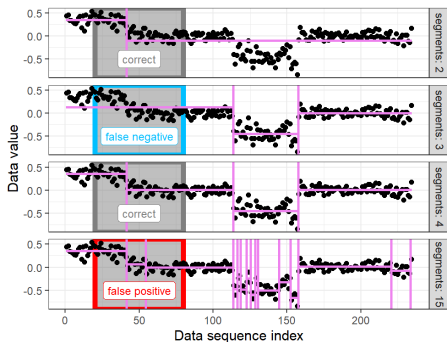


Larger penalty  $\lambda$  results in fewer changes/segments.

Smaller penalty  $\lambda$  results in more changes/segments.

# Problem: weakly supervised changepoint detection

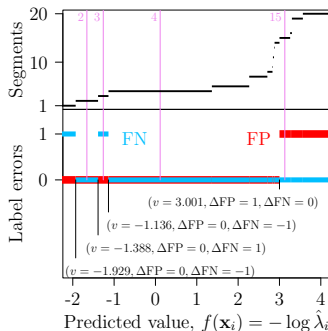
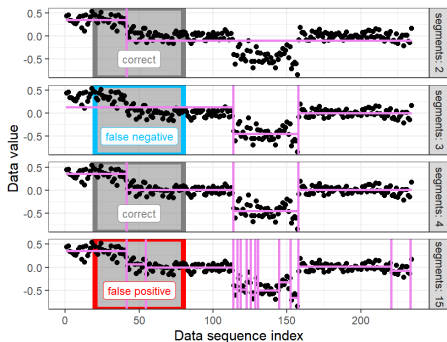
- ▶ First described by Hocking *et al.* ICML 2013.
- ▶ We are given a data sequence  $\mathbf{z}$  with labeled regions  $L$ .
- ▶ We compute features  $\mathbf{x} = \phi(\mathbf{z}) \in \mathbf{R}^p$  and want to learn a function  $f(\mathbf{x}) = -\log \lambda \in \mathbf{R}$  that minimizes label error (sum of false positives and false negatives), or maximizes AUC.





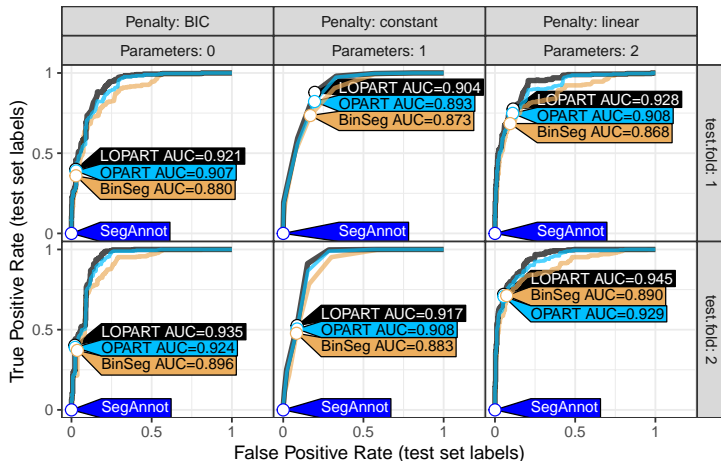
# Problem: weakly supervised changepoint detection

- ▶ First described by Hocking *et al.* ICML 2013.
- ▶ We are given a data sequence  $\mathbf{z}$  with labeled regions  $L$ .
- ▶ We compute features  $\mathbf{x} = \phi(\mathbf{z}) \in \mathbf{R}^p$  and want to learn a function  $f(\mathbf{x}) = -\log \lambda \in \mathbf{R}$  that minimizes label error (sum of false positives and false negatives), or maximizes AUC.



# Comparing changepoint algorithms using ROC curves

Hocking TD, Srivastava A. Labeled Optimal Partitioning. Accepted in Computational Statistics, arXiv:2006.13967.



LOPART algorithm (R package LOPART) has consistently larger test AUC than previous algorithms.

Problem Setting 1: ROC curves for evaluating supervised binary classification algorithms

Problem setting 2: ROC curves for evaluating supervised changepoint algorithms

Proposed surrogate loss for ROC curve optimization: Area Under  $\text{Min}\{\text{FP}, \text{FN}\}$  (AUM)

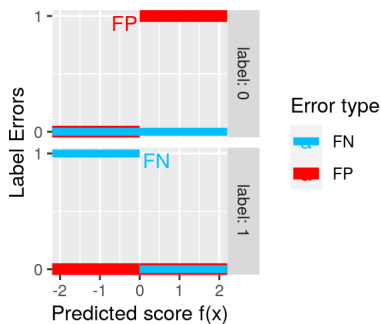
Empirical results: minimizing AUM results in optimized ROC curves

Discussion and Conclusions

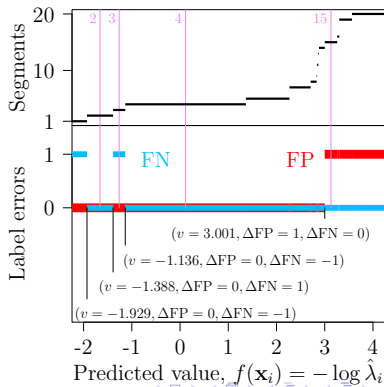
# Algorithm inputs: predictions and label error functions

- ▶ Each observation  $i \in \{1, \dots, n\}$  has a predicted value  $\hat{y}_i \in \mathbb{R}$ .
- ▶ Breakpoints  $b \in \{1, \dots, B\}$  used to represent label error via tuple  $(v_b, \Delta FP_b, \Delta FN_b, \mathcal{I}_b)$ .
- ▶ There are changes  $\Delta FP_b, \Delta FN_b$  at predicted value  $v_b \in \mathbb{R}$  in error function  $\mathcal{I}_b \in \{1, \dots, n\}$ .

## Binary classification



## Changepoint detection

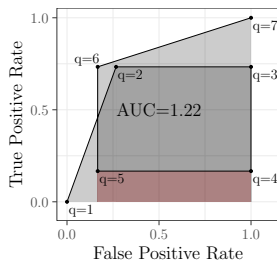
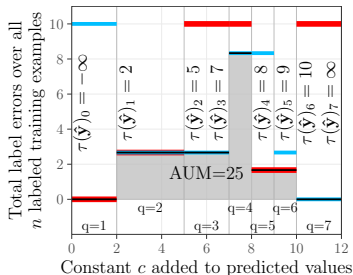


# Proposed surrogate loss, Area Under Min (AUM)

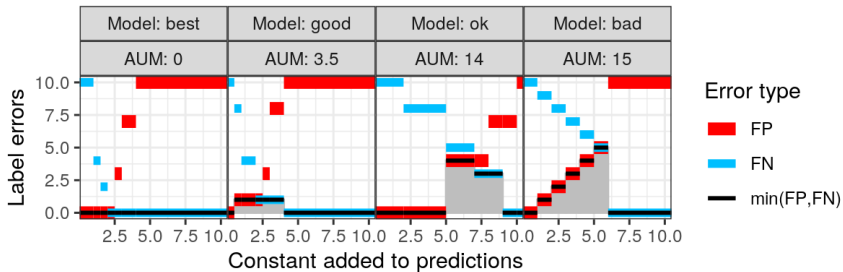
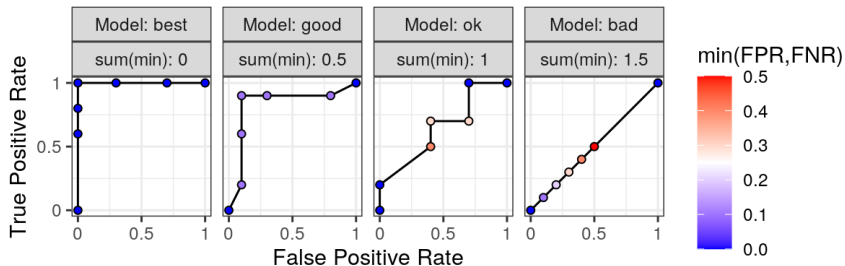
- ▶ Threshold  $t_b = v_b - \hat{y}_{\mathcal{I}_b} = \tau(\hat{\mathbf{y}})_q$  is largest constant you can add to predictions and still be on ROC point  $q$ .
- ▶ Proposed surrogate loss, Area Under Min (AUM) of total FP/FN, computed via sort and modified cumsum:

$$\underline{\text{FP}}_b = \sum_{j: t_j < t_b} \Delta \text{FP}_j, \quad \overline{\text{FP}}_b = \sum_{j: t_j \leq t_b} \Delta \text{FP}_j,$$

$$\underline{\text{FN}}_b = \sum_{j: t_j \geq t_b} -\Delta \text{FN}_j, \quad \overline{\text{FN}}_b = \sum_{j: t_j > t_b} -\Delta \text{FN}_j.$$



# Small AUM is correlated with large AUC

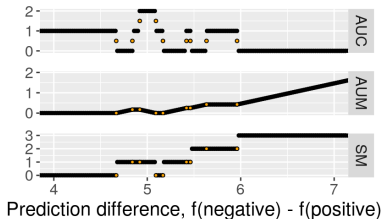


# Proposed algorithm computes two directional derivatives

- ▶ Gradient only defined when function is differentiable, but AUM is not differentiable everywhere (see below).
- ▶ Directional derivatives always computable (R package aum),

$$\nabla_{\mathbf{v}(-1,i)}\text{AUM}(\hat{\mathbf{y}}) = \sum_{b:\mathcal{I}_b=i} \min\{\overline{\text{FP}}_b, \overline{\text{FN}}_b\} - \min\{\overline{\text{FP}}_b - \Delta\text{FP}_b, \overline{\text{FN}}_b - \Delta\text{FN}_b\},$$

$$\nabla_{\mathbf{v}(1,i)}\text{AUM}(\hat{\mathbf{y}}) = \sum_{b:\mathcal{I}_b=i} \min\{\underline{\text{FP}}_b + \Delta\text{FP}_b, \underline{\text{FN}}_b + \Delta\text{FN}_b\} - \min\{\underline{\text{FP}}_b, \underline{\text{FN}}_b\}.$$



Proposed learning algo uses mean of these two directional derivatives as “gradient.”

Problem Setting 1: ROC curves for evaluating supervised binary classification algorithms

Problem setting 2: ROC curves for evaluating supervised changepoint algorithms

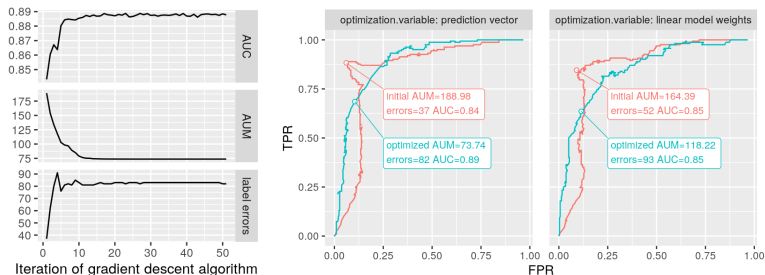
Proposed surrogate loss for ROC curve optimization: Area Under  $\text{Min}\{\text{FP}, \text{FN}\}$  (AUM)

Empirical results: minimizing AUM results in optimized ROC curves

Discussion and Conclusions

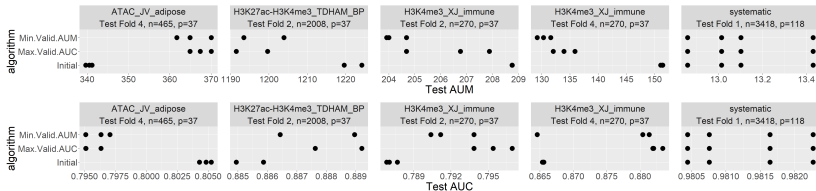


# AUM gradient descent results in increased train AUC for a real changepoint problem



- ▶ Left/middle: changepoint problem initialized to prediction vector with min label errors, gradient descent on prediction vector.
- ▶ Right: linear model initialized by minimizing regularized convex loss (surrogate for label error, Hocking *et al.* ICML 2013), gradient descent on weight vector.

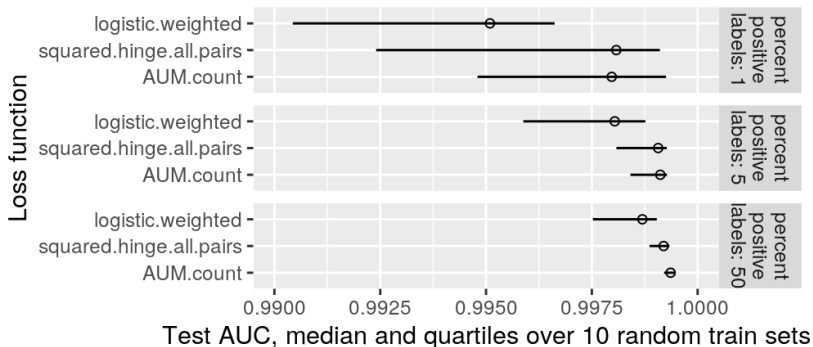
# Learning algorithm results in better test AUC/AUM for changepoint problems



- ▶ Five changepoint problems (panels from left to right).
- ▶ Two evaluation metrics (AUM=top, AUC=bottom).
- ▶ Three algorithms (Y axis), Initial=Min regularized convex loss (surrogate for label error, Hocking *et al.* ICML 2013), Min.Valid.AUM/Max.Valid.AUC=AUM gradient descent with early stopping regularization.
- ▶ Four points = Four random initializations.

# Learning algorithm competitive for unbalanced binary classification

(b) AUM compared to baselines



- ▶ Squared hinge all pairs is a classic/popular surrogate loss function for AUC optimization. (Yan *et al.* ICML 2003)
- ▶ All linear models with early stopping regularization.

Problem Setting 1: ROC curves for evaluating supervised binary classification algorithms

Problem setting 2: ROC curves for evaluating supervised changepoint algorithms

Proposed surrogate loss for ROC curve optimization: Area Under  $\text{Min}\{\text{FP}, \text{FN}\}$  (AUM)

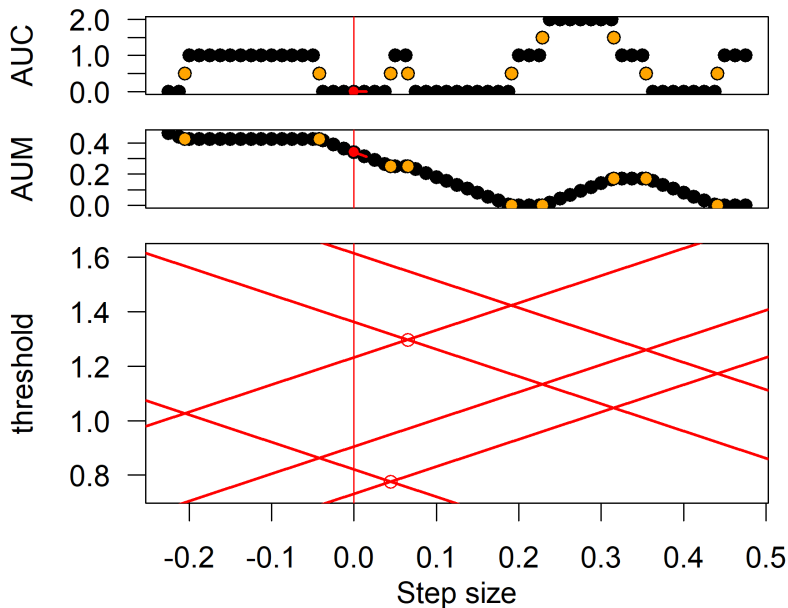
Empirical results: minimizing AUM results in optimized ROC curves

Discussion and Conclusions

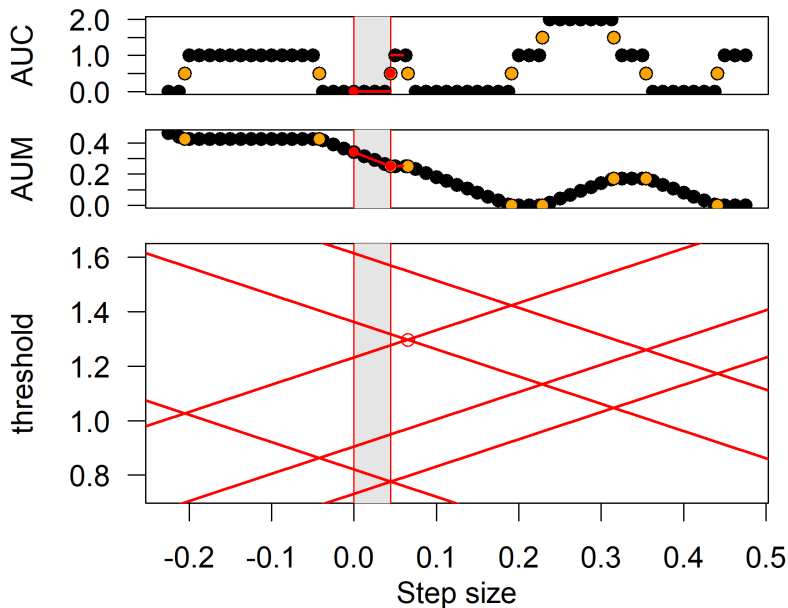
# Discussion and Conclusions, Pre-print arXiv:2107.01285

- ▶ ROC curves are used to evaluate binary classification and changepoint detection algorithms.
- ▶ We propose a new loss function,  $AUM = \text{Area Under Min}(FP, FN)$ , which is a differentiable surrogate of the sum of  $\text{Min}(FP, FN)$  over all points on the ROC curve.
- ▶ We propose new algorithm for efficient AUM and directional derivative computation.
- ▶ Implementations available in R and python/torch:  
<https://cloud.r-project.org/web/packages/aum/>  
<https://tdhock.github.io/blog/2022/aum-learning/>
- ▶ Empirical results provide evidence that learning using AUM minimization results in ROC curve optimization (encourages monotonic/regular curves with large AUC).
- ▶ Future work: exploiting piecewise linear structure of the AUM loss, other model classes, other problems/objectives.

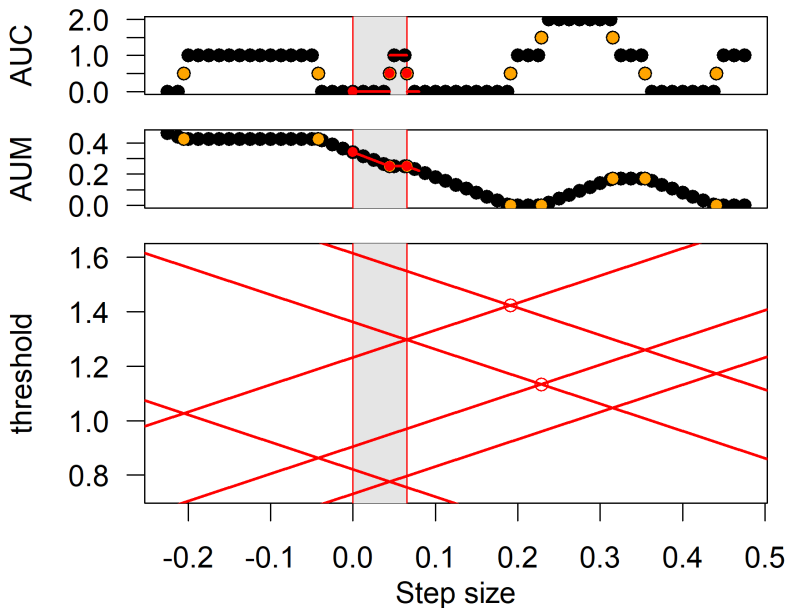
## AUM/AUC line search, iteration 1



## AUM/AUC line search, iteration 2

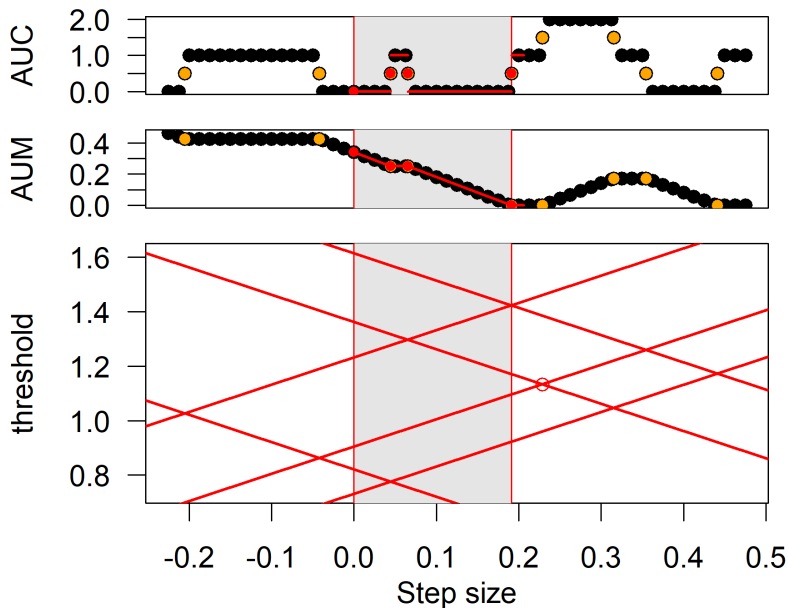


## AUM/AUC line search, iteration 3

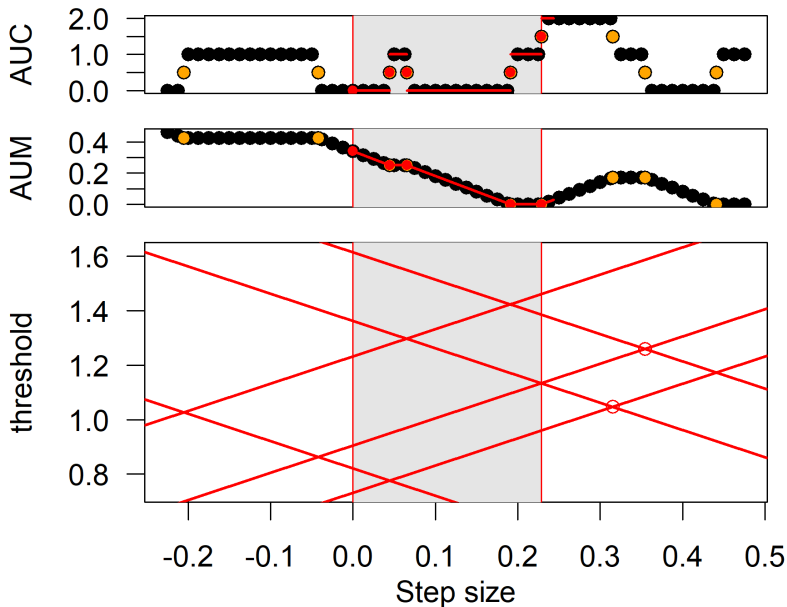




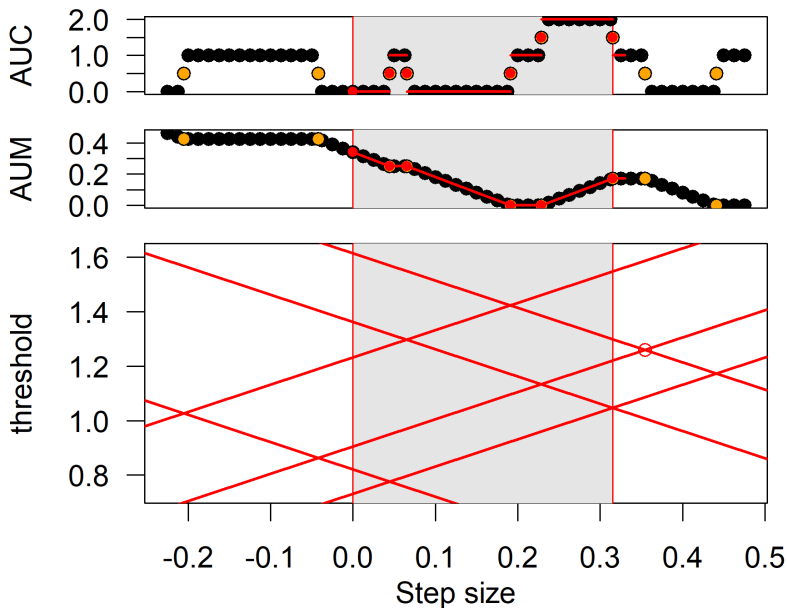
## AUM/AUC line search, iteration 4



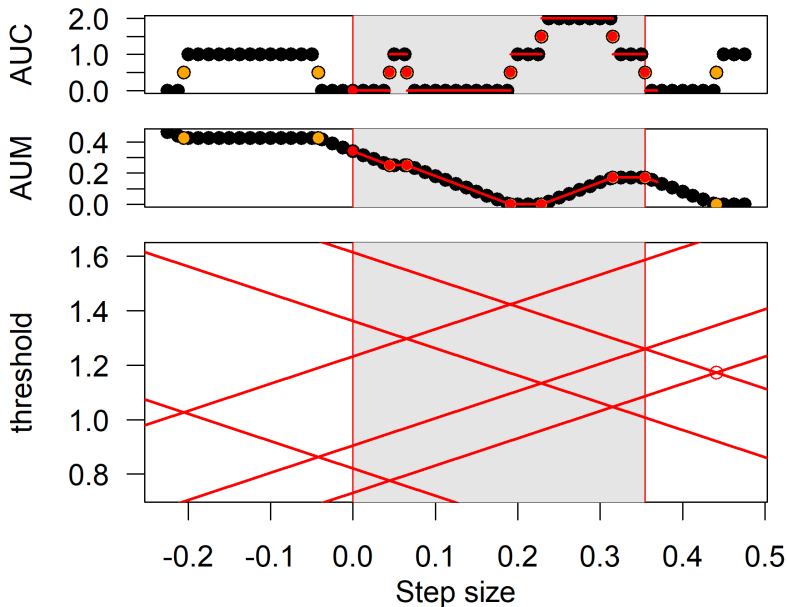
## AUM/AUC line search, iteration 5



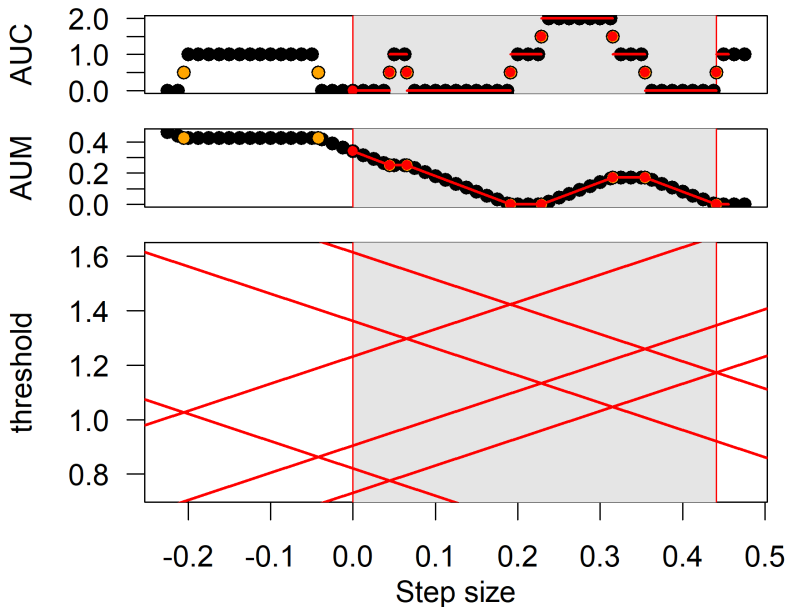
## AUM/AUC line search, iteration 6



## AUM/AUC line search, iteration 7



## AUM/AUC line search, iteration 8



Thanks to co-author Jonathan Hillman! (second from left)



Contact: [toby.hocking@nau.edu](mailto:toby.hocking@nau.edu)