Carnegie Mellon University

Auton Lab

ICMLA 2013

Informative Projection Recovery for Classification, Clustering and Regression

MADALINA FITERAU
ARTUR DUBRAWSKI

Motivation

1. NEED COMPACT MODELS TO ENABLE ANALYSIS AND VISUALIZATION
2. LEVERAGING EXISTING STRUCTURE IN DATA → HIGH PERFORMANCE
3. COMPACT ENSEMBLES OF COMPLEMENTARY LOW-D SOLVERS

BORDER CONTROL

DIAGNOSTICS

VEHICLE CHECKS

Presentation Roadmap

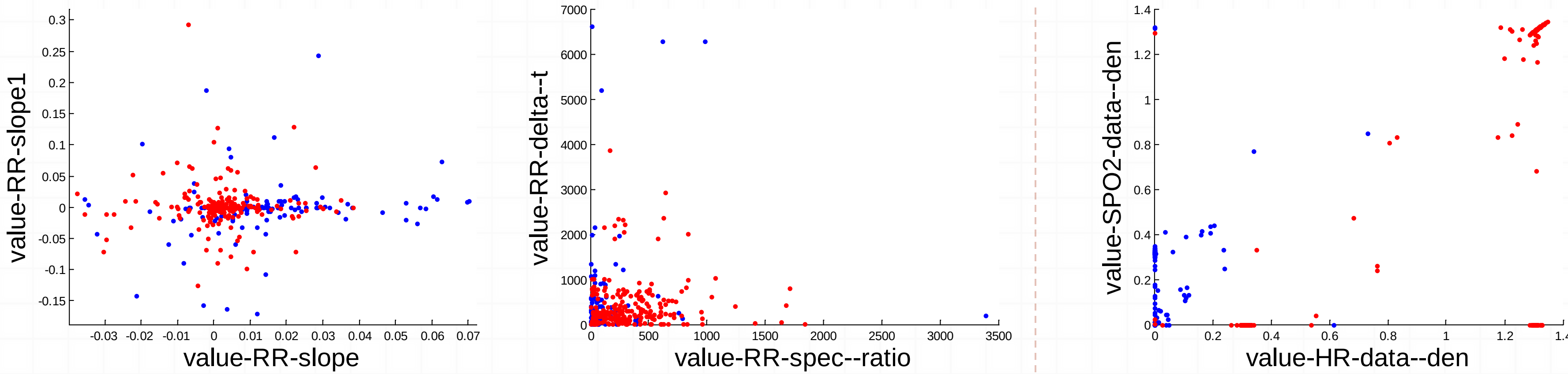
- Informative Projection Retrieval
- RIPR* Framework Overview
  * Regression-based Informative Projection Retrieval
- The Optimization Procedure
- Applicability to Learning Tasks
- Performance Evaluation
- Medical Application Case Study

Informative Projection Retrieval (IPR)

Projection Retrieval for a Learning Task

- problem of **selecting low-d (2D, 3D) subspaces**
- s.t. queries are resolved with **high-confidence**
- models perform the task with **low expected risk**

example: features represent vital signs and derived features; considering only the duty cycles of the signals might be sufficient



A multitude of projections where data is ‘noisy’

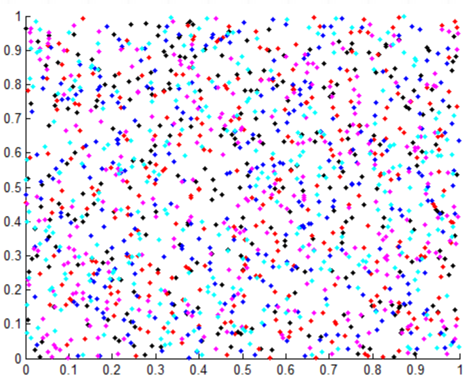
A small set where there is a clear separation

RIPR = Regression-based Informative Projection Retrieval*

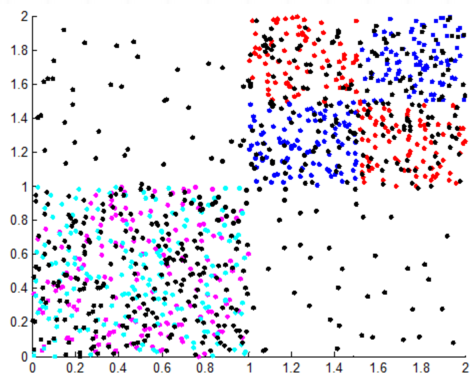
*A generalization of our prior work in“*Projection Retrieval for Classification*”, NIPS 2012

RIPR Target Datasets

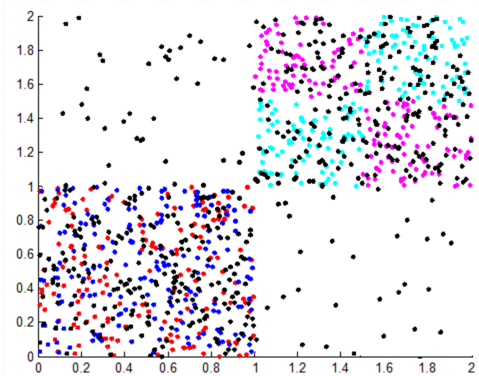
- Most of the features are redundant (non-informative)
- There exists one or several sets of features with structure
- The 'tidy' part of the set may span only part of the points
- Jointly, the sets of projections handle all data



Aspect of most projections



IP for blue/red group

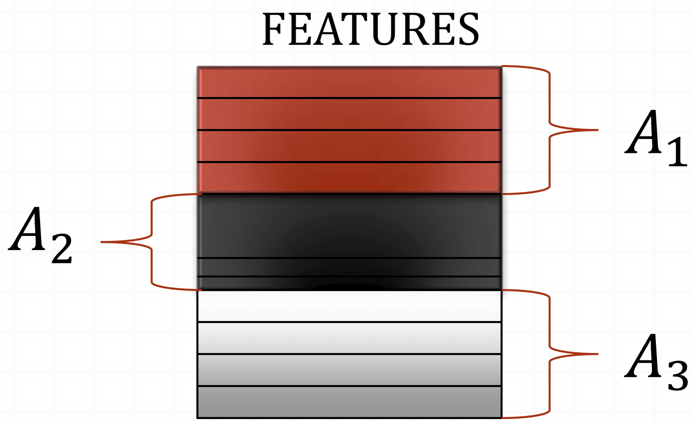


IP for light blue/purple group

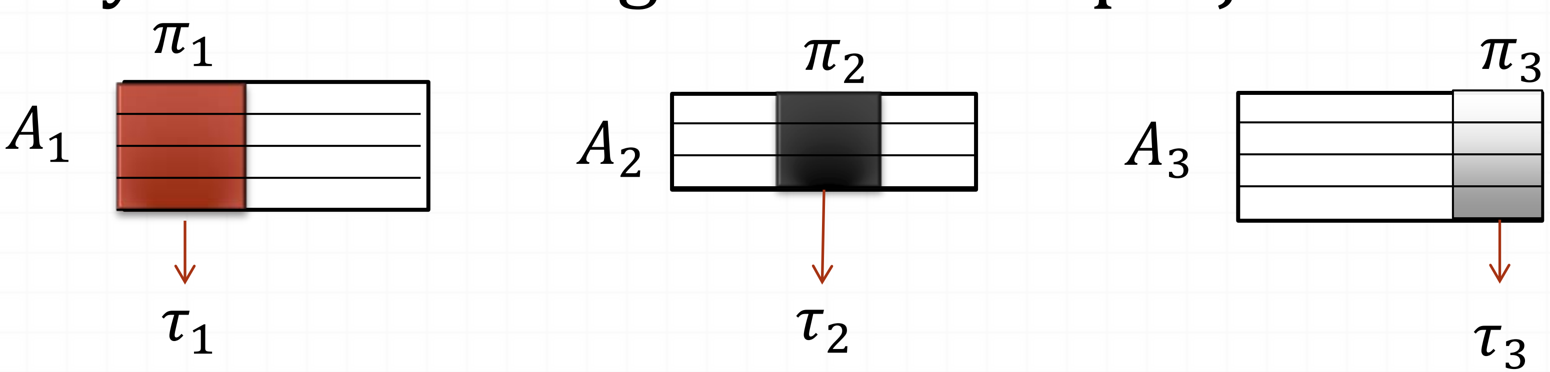
- Clinical Data - several sub-models, corresponding to underlying conditions and patient characteristics
- Human-engineered datasets - corrupted with artifacts which can be identified as low-dimensional patterns

A Dual-Objective Training Process

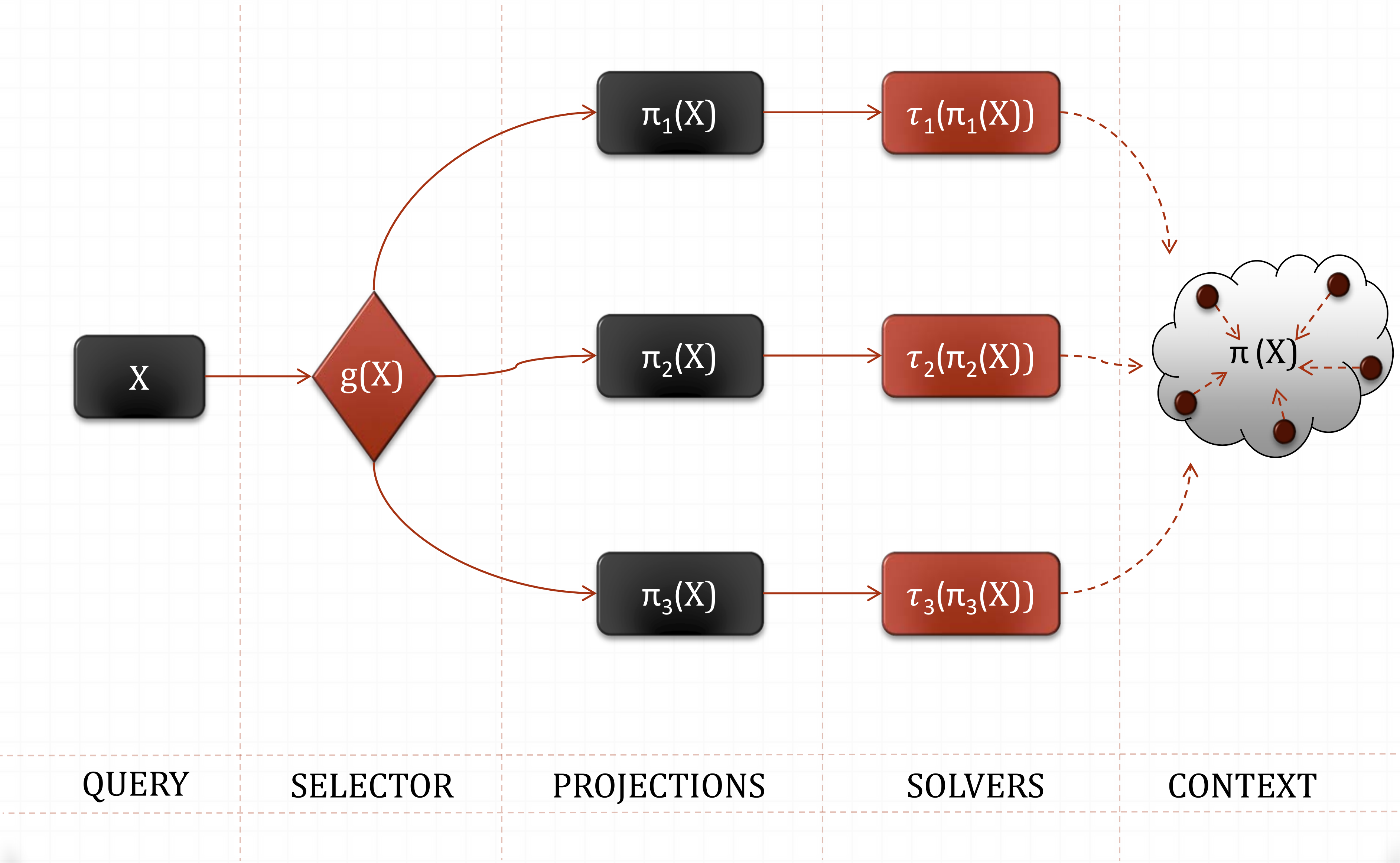
1. Data is split across informative projections



2. Each projection has a solver trained using only the data assigned to that projection



RIPR Framework



RIPR Model

Model components:

- Set of d-dimensional, axis-aligned sub-spaces of the original feature space $\text{P} \in \Pi$
- Each projection has an assigned solver of the task T; the solvers are selected from some solver class $\mathcal{T}$
- A selection function g, which yields, for a query point x, the projection/solver pair $(\pi_{g(x)}, \tau_{g(x)})$ for the point;
- $\ell\left(\tau_{g(x)}\left(\pi_{g(x)}\right), y\right)$ represents the model loss at point x

Dataset → $\{(x_1, y_1) \dots (x_n, y_n) \in \mathcal{X}^n \times \mathcal{Y}^n\}$

$$\mathcal{M}_d = \{ \; P = \{\pi : \pi \in \Pi, dim(\pi) \le d\},$$
$$T = \{\tau_i : \tau_i \in \mathcal{T}, \tau_i : \mathcal{X} \rightarrow, \mathcal{Y} \; \forall i = 1 \dots |P|\},$$
$$g \in \{f : \mathcal{X} \rightarrow \{1 \dots |P|\}\}$$

Target model

Small set of projections

Solvers

Selection function

RIPR Objective Function

Model components:

- Set of d-dimensional, axis-aligned sub-spaces of the original feature space P ϵ Π
- Each projection has an assigned solver of the task T; the solvers are selected from some solver class $\mathcal{T}$
- A selection function g, which yields, for a query point x, the projection/solver pair $(\pi_{g(x)}, \tau_{g(x)})$ for the point;
- $\ell\left(\tau_{g(x)}\left(\pi_{g(x)}\right), y\right)$ represents the model loss at point x

Minimization:

$$M^* = argmin_{M \epsilon \mathcal{M}_d} \mathbb{E}_{\mathcal{X}}\, \ell\left(\tau_{g(x)}\left(\pi_{g(x)}\right), y\right)$$

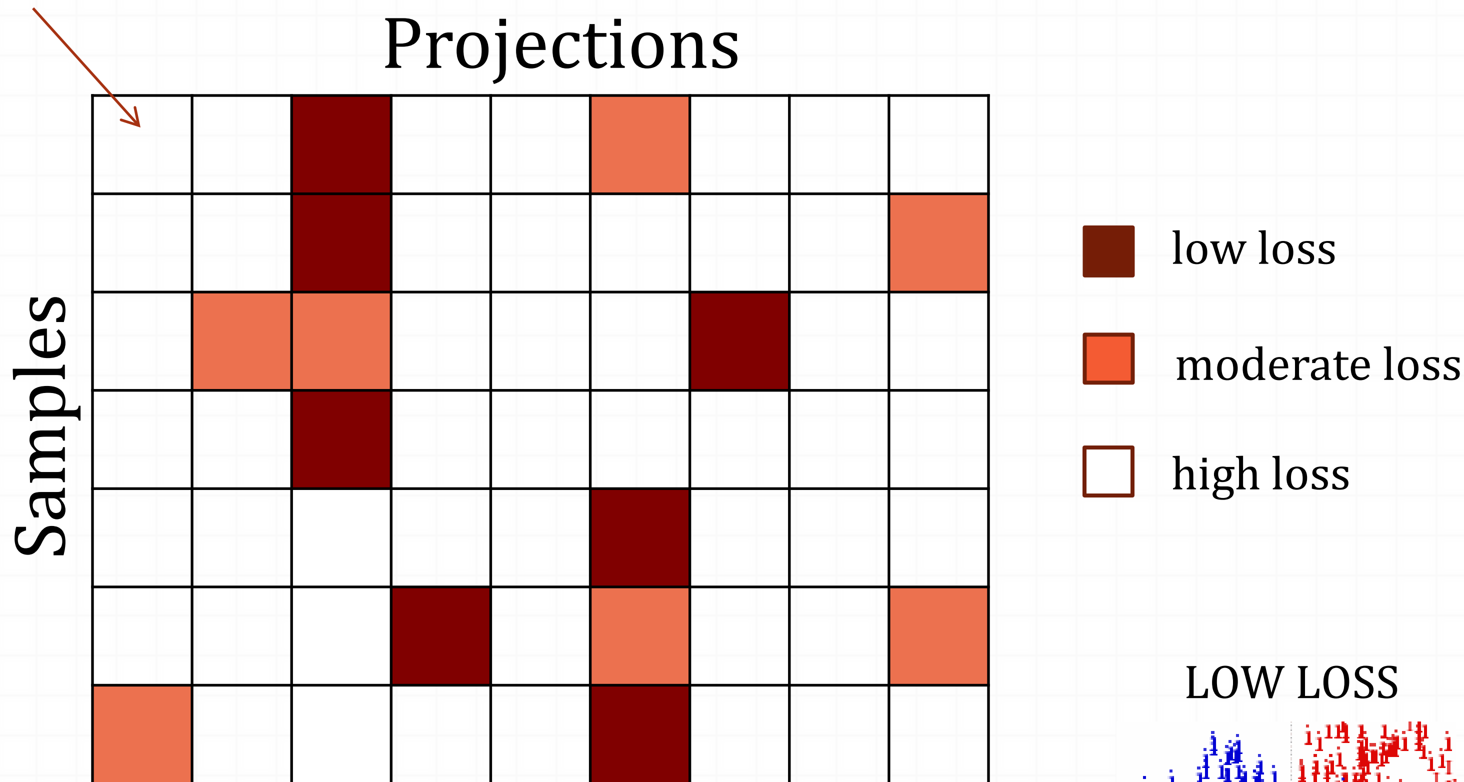
Expected loss for task solver trained on projection assigned to point

Presentation Roadmap

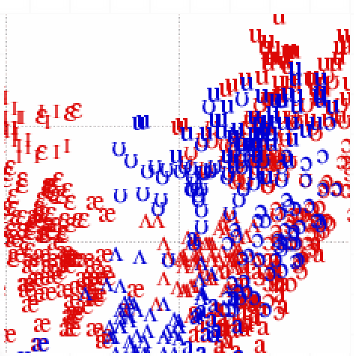
- Informative Projection Retrieval
- RIPR Framework Overview
- The Optimization Procedure
- Applicability to Learning Tasks
- Performance Evaluation
- Medical Application Case Study

Starting point: the loss matrix

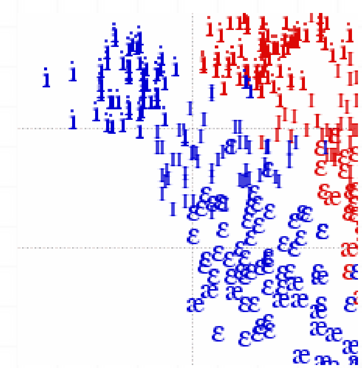
Loss estimators



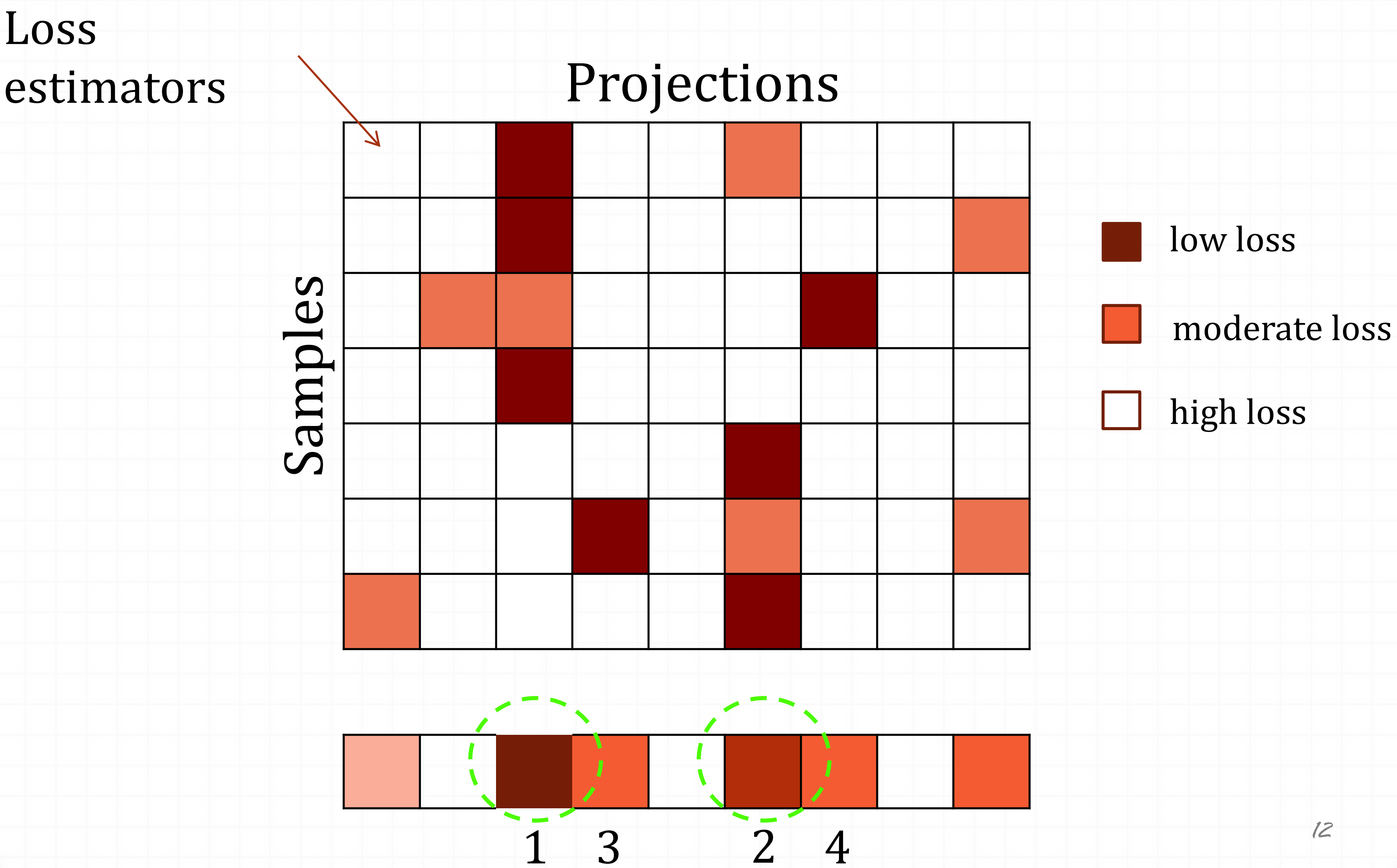
HIGH LOSS



LOW LOSS

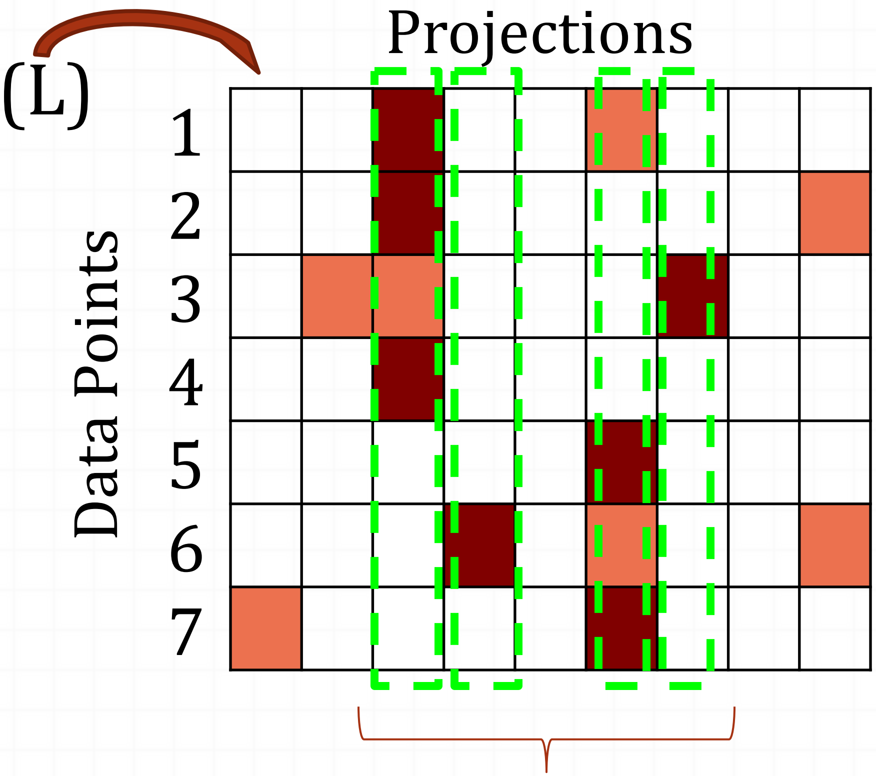


Starting point: the loss matrix



The Optimization Procedure

Matrix of Loss Estimators (L)



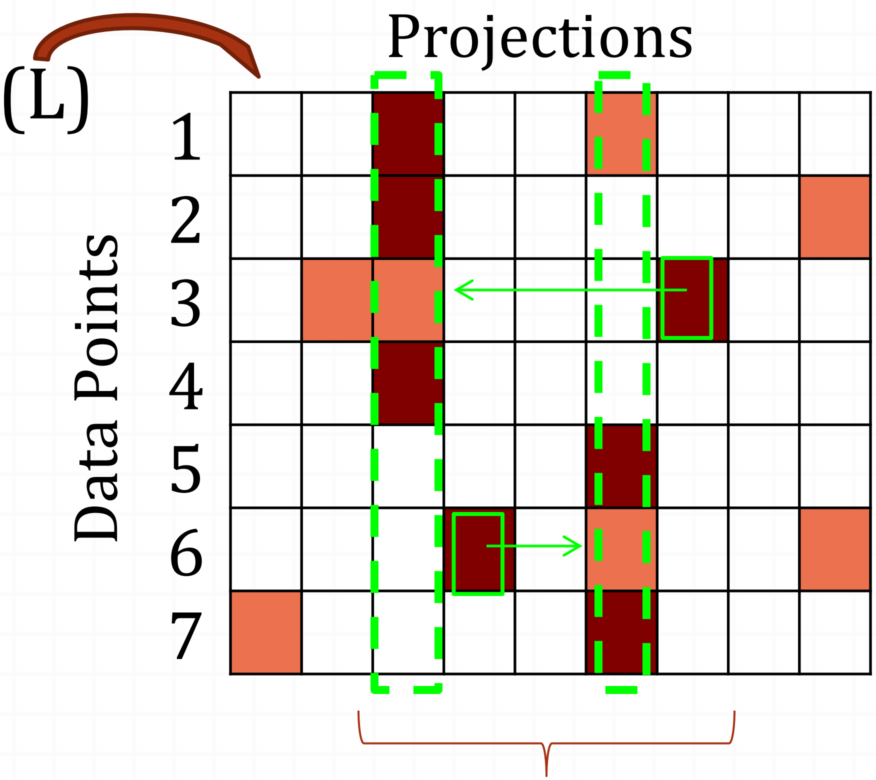
optimal

nearly optimal

We introduce a penalty over # of columns to limit the # of projections in the model

The Optimization Procedure

Matrix of Loss Estimators (L)

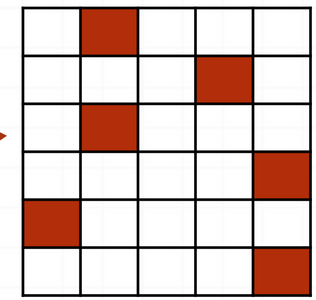


- optimal
- nearly optimal

Suboptimal projections will be used for some of the points

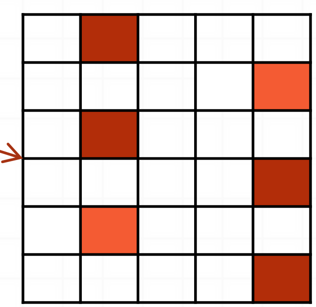
Regression for Informative Projection Recovery (RIPR)

- RIPR learns a binary selection matrix B in a manner resembling the adaptive lasso



- Iterative procedure
  - Initialize selection matrix B
  - Compute multiplier δ inversely proportional with projection popularity

  - Use penalty $|B\delta|_1 \rightarrow$ new B



The RIPR Algorithm

1. Compute loss matrix L, target T

2. Estimate selection matrix B

$$min_B \parallel \mathrm{T} - L \otimes \mathrm{B}\mathcal{J}_{|\Pi|,1} \parallel_2{}^2 + \lambda \sum_{k=1}^{|\Pi|} |B_k|_1$$

ITERATE UNTIL CONVERGENCE

3. Compute multiplier δ inversely proportional with utility

$$\delta_k = |B_k|_1, \qquad \delta = 1 - \delta/|\delta|_1$$

4. Obtain new selection matrix B penalizing Bδ

$$min_B \parallel \mathrm{T} - L \otimes \mathrm{B}\mathcal{J}_{|\Pi|,1} \parallel_2{}^2 + \lambda |B\delta|_1$$

where $L_{ij} \otimes \mathrm{B}_{ij} = L_{ij} B_{ij}$

Applicability to Learning Tasks

We show how RIPR can solve the following tasks:

- Classification
- Semi-supervised classification
- Clustering
- Regression

The matrix of loss estimators is computed differently for each of these tasks.

The generality of the method does not stop here: RIPR can solve any learning task for which the risk can be decomposed using consistent loss estimators.

Loss Estimators: Classification

Neighbor-based estimator for conditional entropy*:

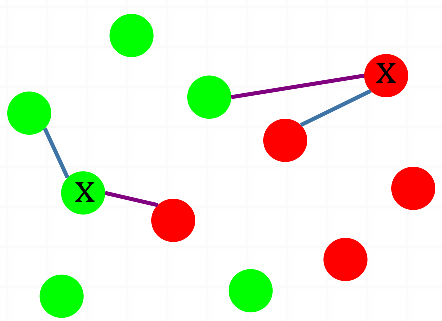
$$\widehat{H}(Y|X \in \mathcal{A}) \propto \frac{1}{n}\sum_{i=1}^{n} I[x_i \in \mathcal{A}]\left(\frac{n-1}{n}\left(\frac{\boldsymbol{dist_{k+1}(x_i, X_{y_i})}}{\boldsymbol{dist_k(x_i, X_{\neg y_i})}}\right)^{\dim(X)}\right)^{1-\alpha}$$

For a projection π, the estimator is $\widehat{H}(Y|\pi(X); g(X) \to \pi)$.
The optimal model can be computed through the minimization:

$$\widehat{M} = argmin_{M \epsilon \mathcal{M}_d} \sum_{\pi_j \in \Pi} \sum_{i=1}^{n} I[g(x_i) \to \pi_j]\left(\frac{\boldsymbol{dist_{k+1}(\pi_j(x_i), \pi_j(X_{y_i}))}}{\boldsymbol{dist_k(\pi_j(x_i), \pi_j(X_{\neg y_i}))}}\right)^{\dim(\pi_j)(1-\alpha)}$$

B_{ij}- selection matrix

L_{ij}-local entropy contributions



$$T_i = min_j L_{ij}$$

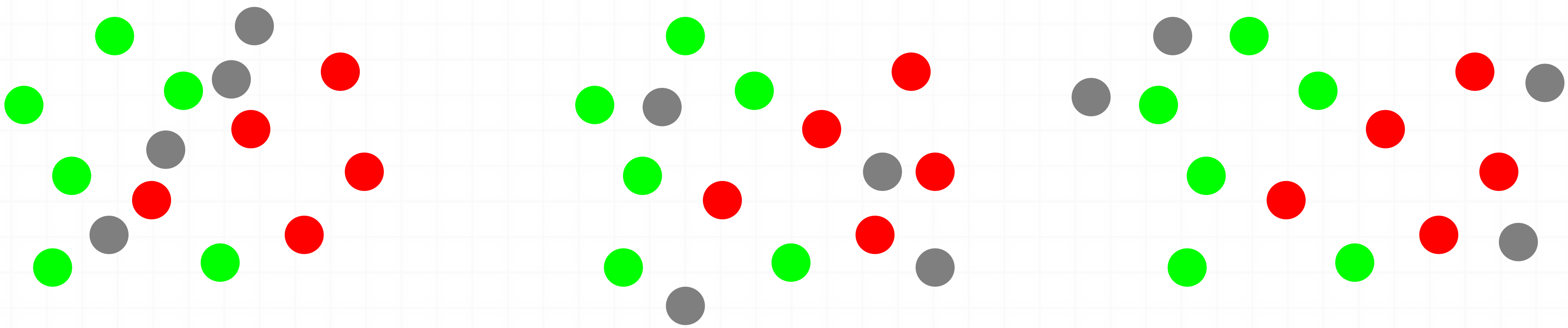
Based on the divergence estimator by Poczos and Schneider, *"On the estimation of alpha-divergences"* (AISTATS 2011)

Loss Estimators: Semi-supervised Classification

- For labeled samples: same as for classification
- For unlabeled samples:
  - Consider all possible label assignments
  - Assume the most 'confident' label (with smallest loss)

  Equivalent to

  - Penalizing unlabeled samples proportional to how ambivalent they are to the label assigned



POOR DECENT GOOD

Loss Estimators: Semi-supervised Classification

- For labeled samples: same as for classification
- For unlabeled samples:
  - Consider all possible label assignments
  - Assume the most 'confident' label (with smallest loss)

  Equivalent to

  - Penalizing unlabeled samples proportional to how ambivalent they are to the label assigned

$$R_{ssc}\left(X_{\in\mathcal{A}(\pi_j)}\right) = \sum_{x_i \in labeled} \left(\frac{\boldsymbol{dist_{k+1}(\pi_j(x_i), \pi_j(X_{y_i}))}}{\boldsymbol{dist_k(\pi_j(x_i), \pi_j(X_{\neg y_i}))}}\right)^{\dim(\pi_j)(1-\alpha)} +$$

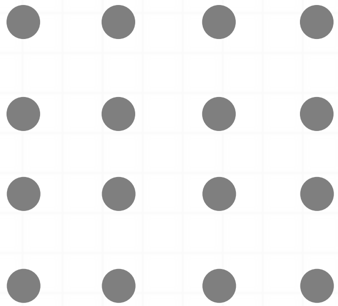
$$\sum_{x_i \in unlabeled} min_{\gamma \epsilon \mathcal{Y}} \left(\frac{\boldsymbol{dist_{k+1}(\pi_j(x_i), \pi_j(X_\gamma))}}{\boldsymbol{dist_k(\pi_j(x_i), \pi_j(X_{\neg\gamma}))}}\right)^{\dim(\pi_j)(1-\alpha)}$$

Entropy Estimators for Clustering

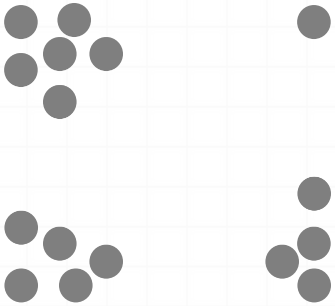
- Point-wise estimators are problematic for clustering
- An ensemble view of the data is typically required
- It is unknown which data should be assigned to which projection prior to clustering

Entropy Estimators for Clustering

- Point-wise estimators are problematic for clustering
- An ensemble view of the data is typically required
- It is unknown which data should be assigned to which projection prior to clustering
- We focus on density-based clustering
- The loss is lower for densely packed regions
- We eliminate dimensionality issues by considering negative KL divergence to uniform on the same space



POOR



GOOD

Entropy Estimators for Clustering

- Point-wise estimators are problematic for clustering
- An ensemble view of the data is typically required
- It is unknown which data should be assigned to which projection prior to clustering
- We focus on density-based clustering
- The loss is lower for densely packed regions
- We eliminate dimensionality issues by considering negative KL divergence to uniform on the same space*

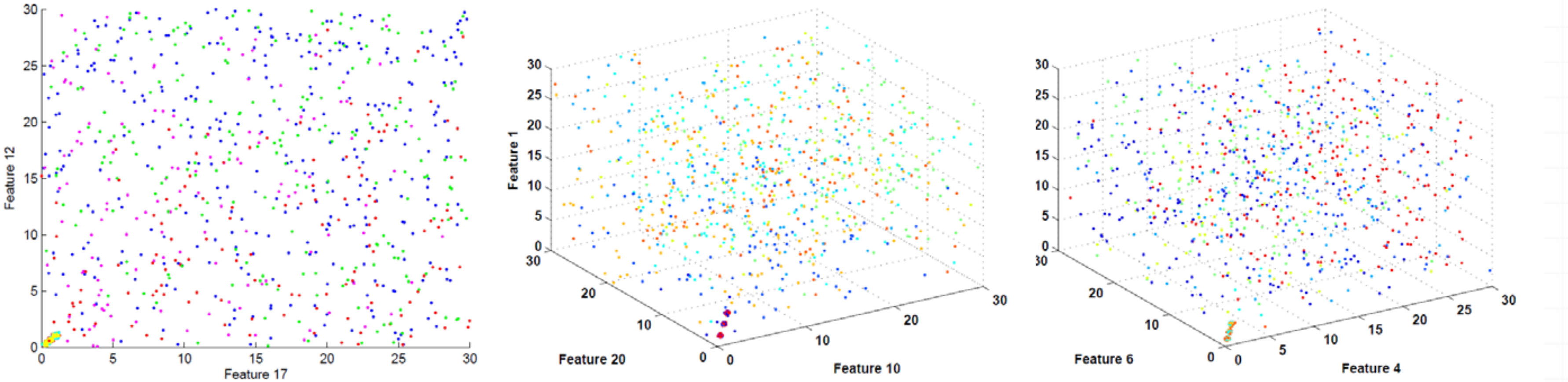
$$R_{clustering}\left(X_{\in \mathcal{A}(\pi_j)}\right) = \rightarrow -KL\left(\pi_j(X), \pi_j(Unif)\right)$$

$$\ell\left(\tau_j\left(\pi_j(x)\right)\right) = \left(\frac{dist\left(\pi_j(x), \pi_j(X)\right)}{dist\left(\pi_j(x), \pi_j(U)\right)}\right)^{\dim(\pi_j)(1-\alpha)}$$

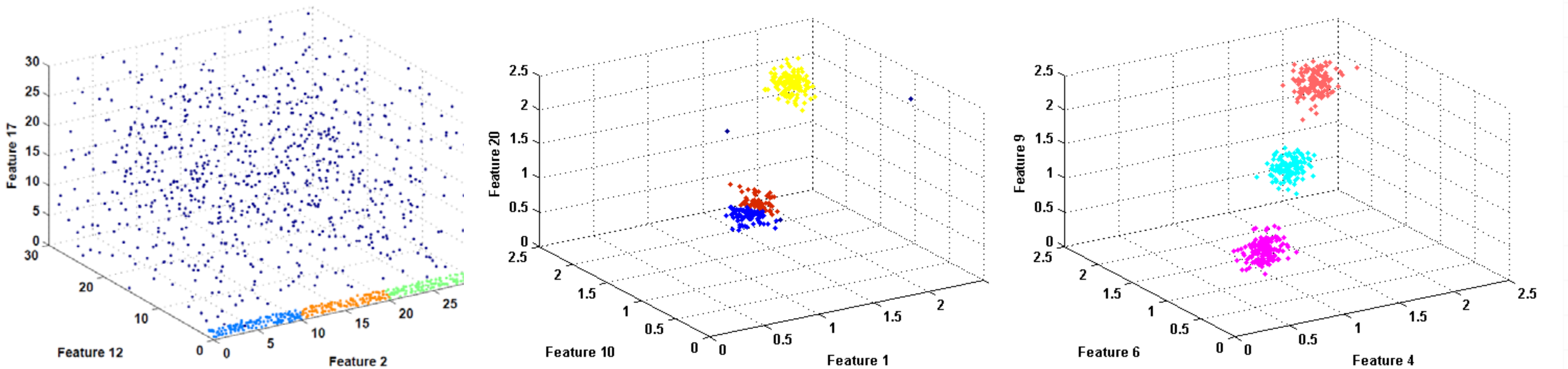
* some scaling issues remain

Low-d Clustering: Why it Works

K-Means model projected on (known) informative features



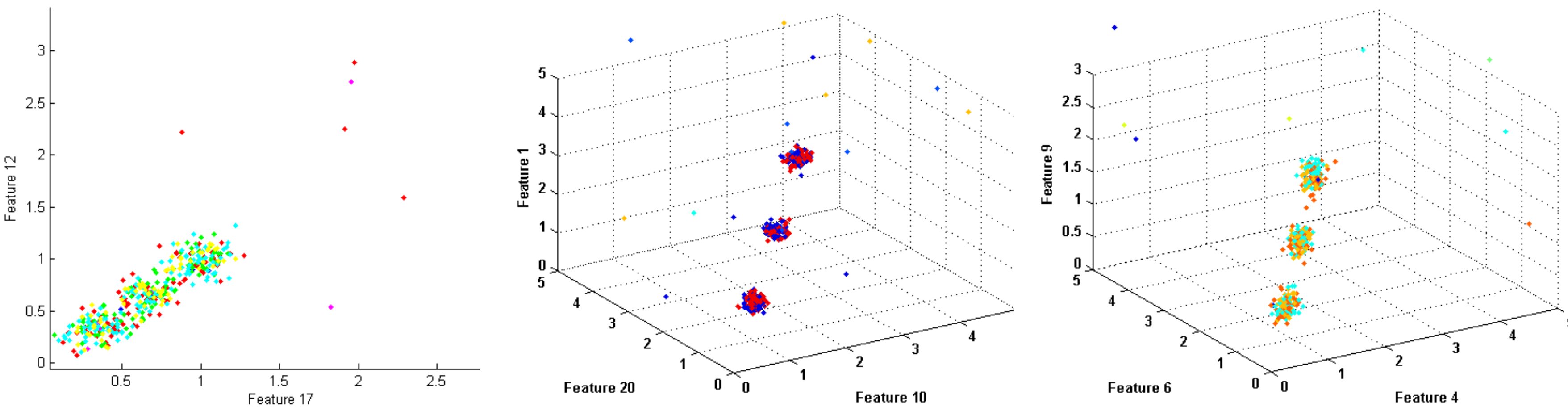
Representation of RIPR model – recovered projections and assigned data



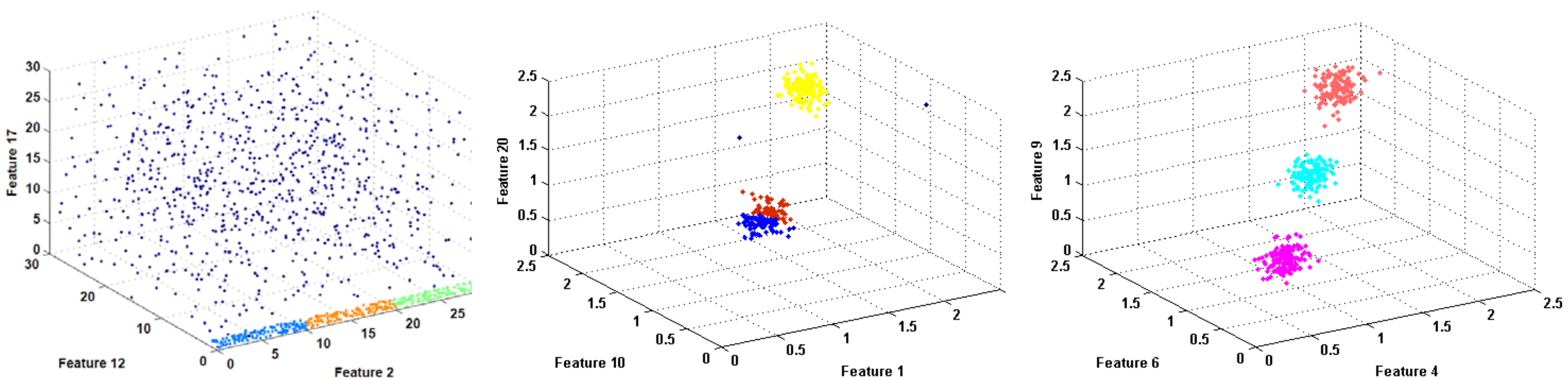
The hidden structure in data is clearly revealed by the RIPR model.

Low-d Clustering: Why it Works

K-Means model projected on (known) informative features



Representation of RIPR model – recovered projections and assigned data



The hidden structure in data is clearly revealed by the RIPR model.

Loss/Risk for common Learning Tasks

Learning Task	Loss/Risk
Classification *	Classification error approximated by conditional entropy $R_{cls}(\mathcal{X}) = \mathbb{E}_{\mathcal{X}}[y \neq h_{g(x)}(\pi_{g(x)}(x))] \approx \mathrm{H}(\mathrm{y}\vert\pi_{g(x)}(\mathrm{x}))$
Semi-supervised classification	Conditional entropy for labeled samples plus best case entropy over label assignments for unlabeled samples $R_{ssc}(\mathcal{X}) = R_{cls}(\mathcal{X}) + min_{\gamma\epsilon\mathcal{Y}} H_{x\ unlabeled}(\gamma\vert\pi_{g(x)}(\mathrm{x}))$
Clustering	Negative divergence between distribution of data and a uniform distribution on the same sample space $R_{clustering} = -KL(\pi_{g(x)}(\mathrm{x})\vert\vert uniform(\pi_{g(x)}(\mathcal{X}))$
Regression	Mean squared error $R_{reg}(\mathcal{X}) = \mathbb{E}_{\mathcal{X}}[(y - h_{g(x)}(\pi_{g(x)}(x)))^2]$

* The object of prior work: "*Projection Retrieval for Classification*", NIPS 2012

Assigning a Projection to a Query

Problem: how to select the appropriate projection for a specific query q?

Solution: select the projection in P for which the estimated loss* at q is smallest.

$$(\hat{k}, \hat{y}) = argmin_{(k,y)} \hat{\ell}(\tau_k(\pi_k), y)$$

$$\text{where } k \in \{1 \ldots |P|\}$$

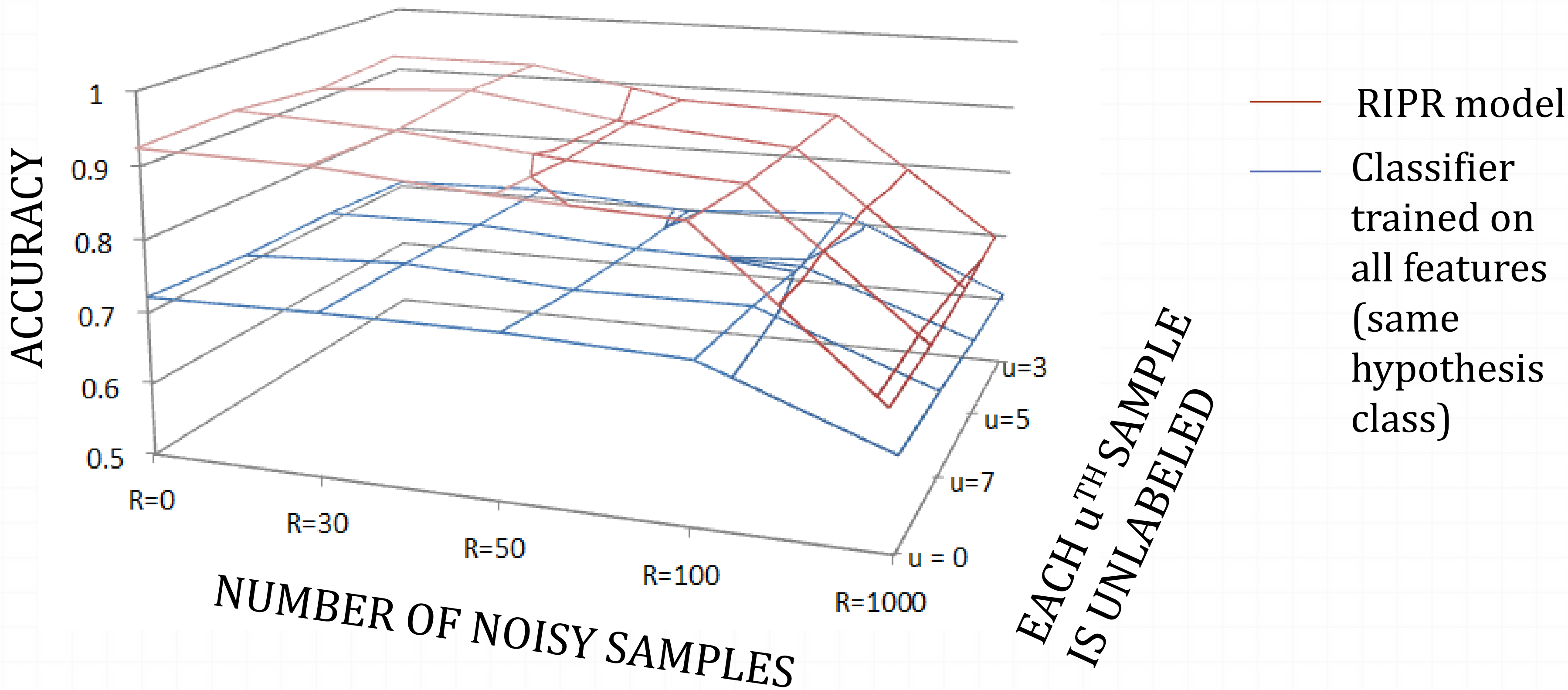
*For clustering, the loss estimator is computed considering the cluster assignments determined during learning.

Presentation Roadmap

- Informative Projection Retrieval
- RIPR Framework Overview
- The Optimization Procedure
- Applicability to Learning Tasks
- Performance Evaluation
- Medical Application Case Study

Semi-supervised classification - artificial data -

DATASET CONTAINS 3 INFORMATIVE PROJECTIONS, 3000 LABELED POINTS.

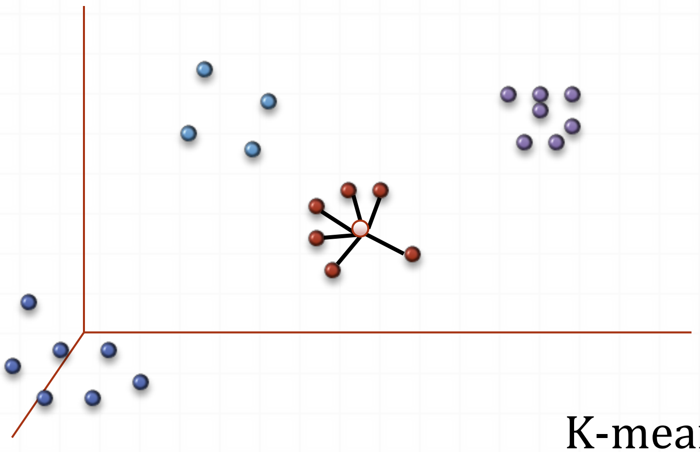


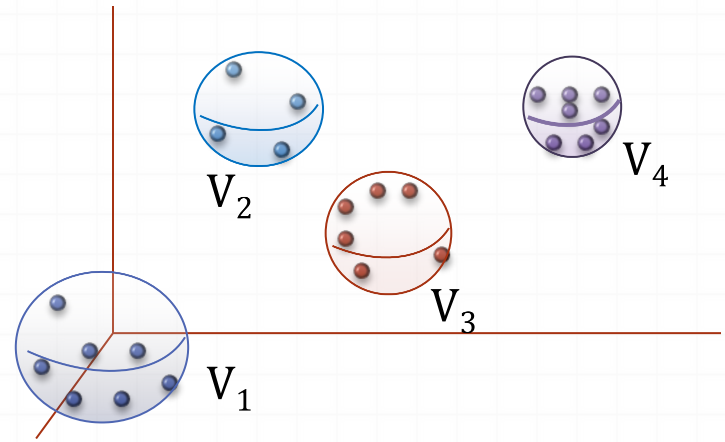
RIPR CORRECTLY RECOVERS THE PROJECTIONS FOR ALL SETINGS TESTED.
LEVERAGING THIS STRUCTURE, RIPR ACHIEVES HIGHER ACCURACY.

Clustering
- evaluation metrics -

DISTORTION – mean distance to cluster centers

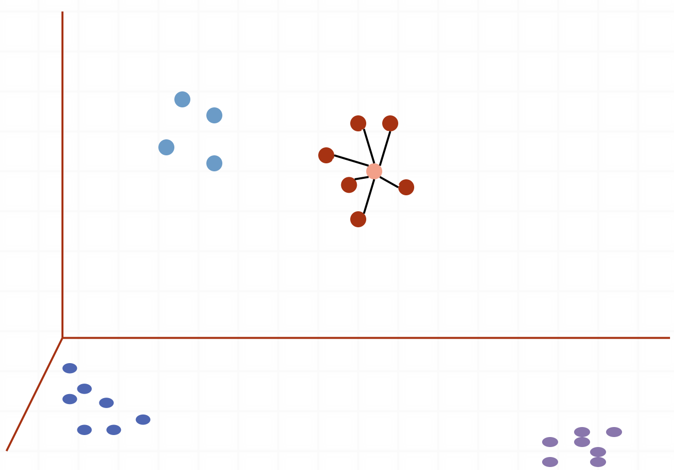
LOG CLUSTER VOLUME

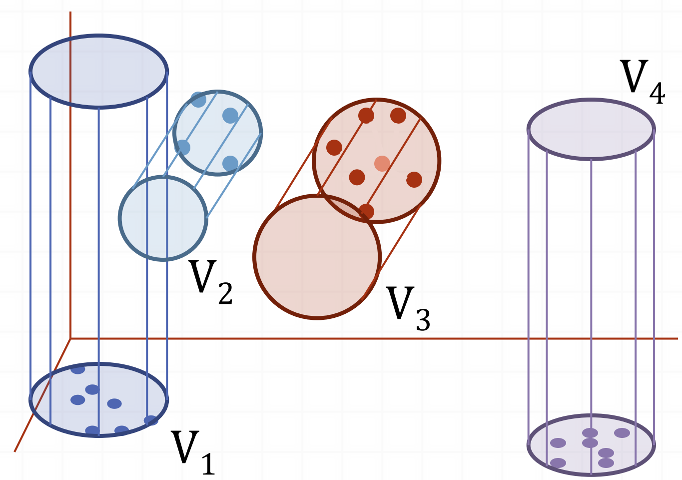




K-means Model

Ripped K-means Model





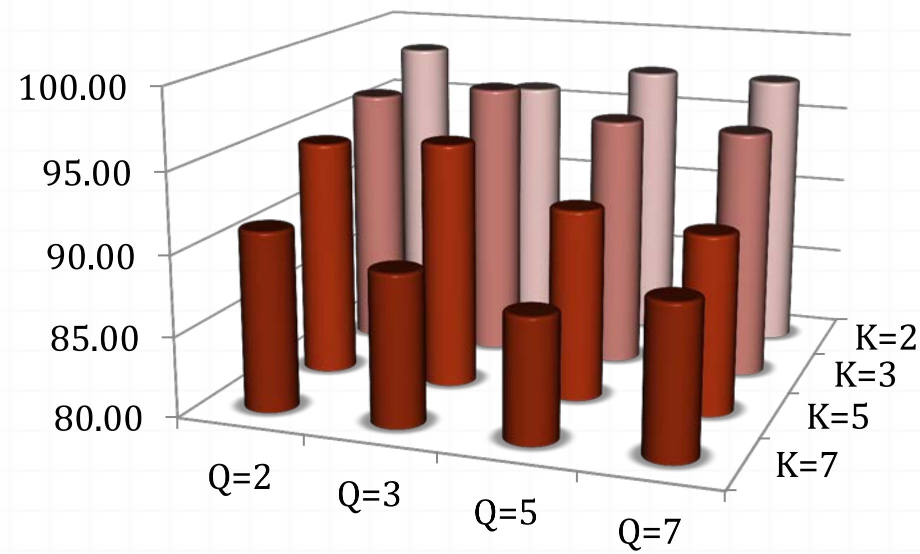
Clustering
- artificial data -

Settings		Distortion		Log Volume	
Q	**K**	**RIPR**	**Kmeans**	**RIPR**	**Kmeans**
2	2	865	12,318	27.41	29.17
2	3	622	12,203	27.56	29.01
2	5	440	12,060	27.78	29.06
2	7	375	11,909	27.92	28.97
3	2	1,344	25,704	31.08	32.47
3	3	872	25,472	31.20	32.77
3	5	648	25,247	31.45	32.78
3	7	530	24,979	31.57	32.55
5	2	2,683	66,801	35.65	37.26
5	3	1,484	66,352	35.79	37.16
5	5	1,065	65,419	36.00	37.09
5	7	842	64,946	36.17	37.08
7	2	4,621	127,558	38.66	40.25
7	3	2,174	126,309	38.86	40.21
7	5	1,480	124,436	39.05	40.10
7	7	1,238	123,151	39.13	40.11

RIPR MODELS ARE MORE COMPACT

Q = NUMBER OF INFORMATIVE PROJECTIONS
K = NUMBER OF CLUSTERS ON EACH PROJECTION

PERCENTAGE REDUCTION IN SUM OF CLUSTER VOLUME



COMPRESSION IS REDUCED AS MORE CLUSTERS/PROJECTIONS ARE ADDED

NOTE: THE K-MEANS AND RIPR MODELS HAVE THE NUMBER OF CLUSTERS.

Clustering
- UCI data -

SUM OF MEAN DISTANCES TO CLUSTER CENTERS AND LOG CLUSTER VOLUME

UCI Dataset	Mean Distortion		*% Distortion Reduction*	Log Volume of Clusters on All Dimensions		*% Volume Reduction*
	RIPR	Kmeans		RIPR	Kmeans	
Seeds	16	107	*90.73*	3.33	4.21	86.83
Libras	9	265	*98.54*	-2.52	3.15	100.00
MiniBOONE	125	1,154,704	*99.99*	104.23	107.77	99.97
Cell	40,877	8,181,327	*99.78*	23.75	29.39	100.00
Concrete	1,370	55,594	*98.01*	21.39	22.91	97.01

LOWER IS BETTER. RIPR MODELS ALWAYS HAVE A SMALLER TOTAL VOLUME.

Regression
- artificial data -

ACCURACY OF RIPPED SVM COMPARED TO ACCURACY OF STANDARD SVM
- THE NUMBER OF INFORMATIVE PROJECTIONS : 2-10
- PERCENTAGE OF NOISY SAMPLES: 0-50% (OUT OF 1600)

	IP #	2	3	5	7	10		2	3	5	7	10
		MSE RIPPED-SVM					MSE SVM					
NOISY SAMPLES	0%	0.05	0.27	0.05	0.02	0.23		0.27	1.16	0.11	0.1	0.43
	6.25%	0.42	1.26	0.34	1.45	0.52		0.8	1.02	0.6	2.99	0.94
	12.5%	0.5	0.86	0.8	0.33	0.99		0.97	1.27	0.29	0.68	1.44
	25%	0.63	1.47	1.34	1.61	0.11		0.4	1.26	1.64	1.71	0.08
	50%	0.69	0.38	1.12	0.68	1.1		0.52	0.06	0.91	0.9	1.16

PRECISION AND RECALL OF THE RECOVERED PROJECTIONS

		RIPR Precision					RIPR Recall					
NOISY SAMPLES	0%	1	1	0.4	0.43	0.3		0.67	1	0.67	1	1
	6.25%	1	0.67	0.6	0.43	0.2		0.67	0.67	1	1	0.67
	12.5%	1	1	0.6	0.43	0.3		0.67	1	1	1	1
	25%	1	1	0.6	0.43	0.1		0.67	1	1	1	0.33
	50%	1	0.67	0.4	0.29	0.3		0.67	0.67	0.67	0.67	1

Case Study – Alert Classification
- importance of artifact adjudication -

- Intensive Care Unit vital sign monitoring system
- Alerts are raised when patient health status deteriorates
- One alert is issued every 90s

- A significant amount of alerts are artifacts
- Frequent alerts cause alarm fatigue in medical staff
- Quality of care diminished unless artifacts are identified

Case Study – Alert Classification - vital sign data processing -

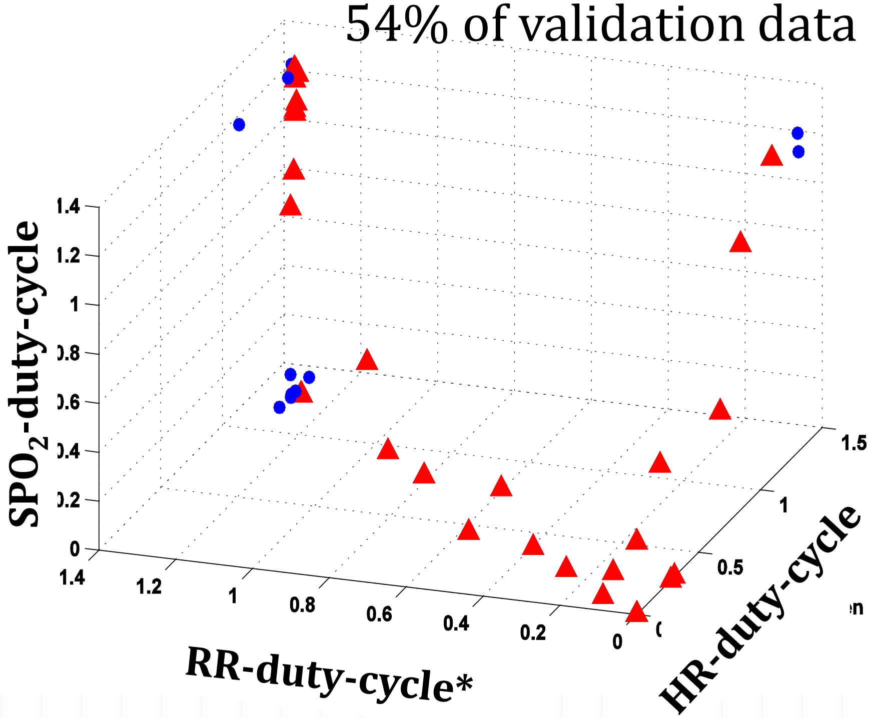
- Each alert is associated with the first abnormal vital sign
  - Heart Rate (HR), Respiratory Rate (RR)
  - Systolic (SBP) and Diastolic (DBP) Blood Pressure
  - Peripheral arterial oxygen saturation (SpO2)
- 812 of the samples were labeled by clinicians (~10%)
- Extracted temporal features and derived metrics
  - Vitals collected during the alert event
  - Data starting 4 minutes before alert onset
  - Moving window statistics
  - Metrics such as duty cycle
  - Data collected for each vital independently

Case Study – Alert Classification - performance -

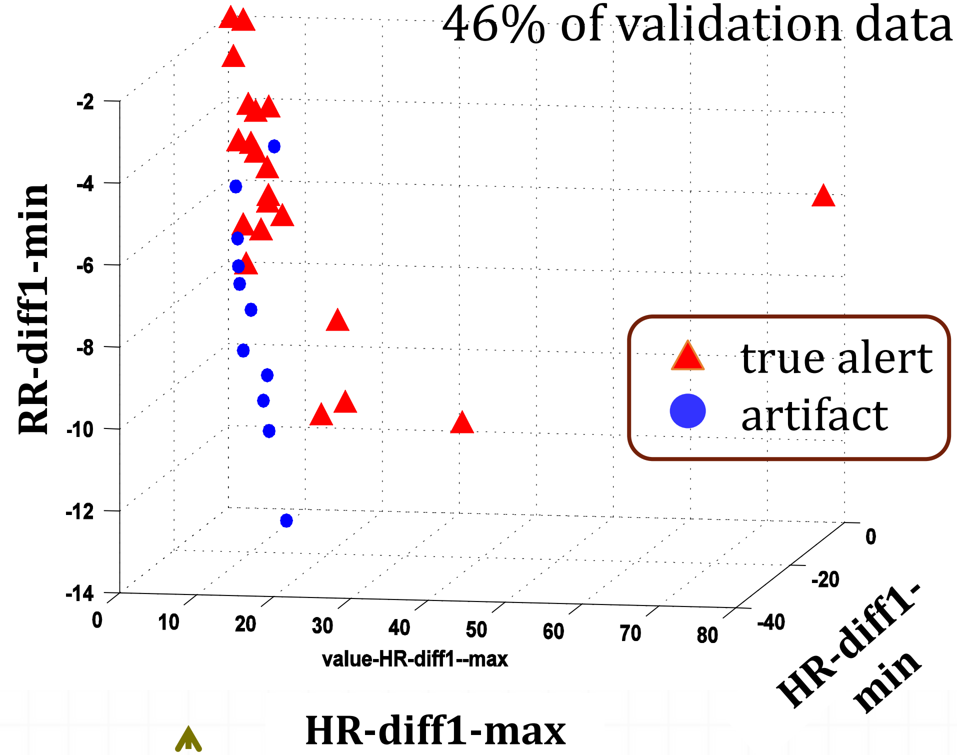
Alarm Type	RR	BP		SPO_2	
	2D	2D	3D	2D	3D
Accuracy	0.98	0.833	0.885	0.911	0.9151
Precision	0.979	0.858	0.896	0.929	0.9176
Recall	0.991	0.93	0.958	0.945	0.9957

Case Study – Alert Classification - RIPR model for blood pressure -

54% of validation data



46% of validation data



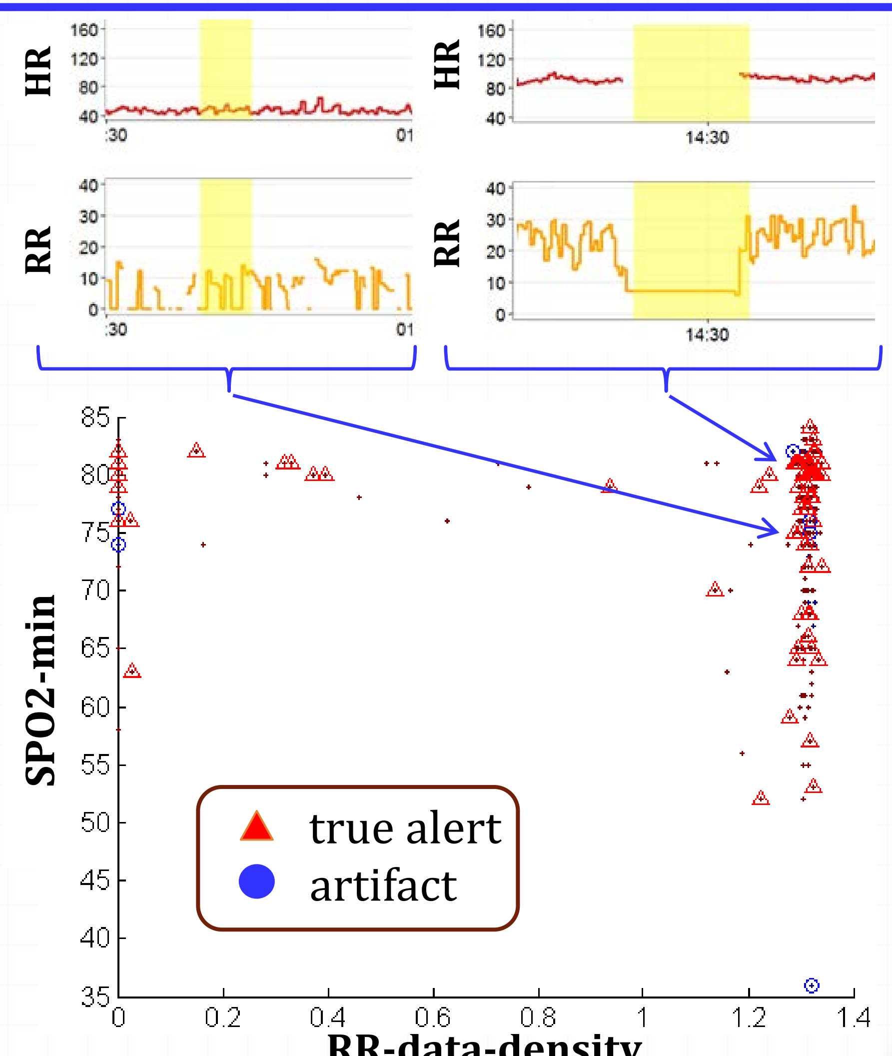
Alarm Type	RR	BP		SPO_2	
	2D	2D	3D	2D	3D
Accuracy	0.98	0.833	0.885	0.911	0.9151
Precision	0.979	0.858	0.896	0.929	0.9176
Recall	0.991	0.93	0.958	0.945	0.9957

RIPR identifies interpretable projections which adjudicate alerts.

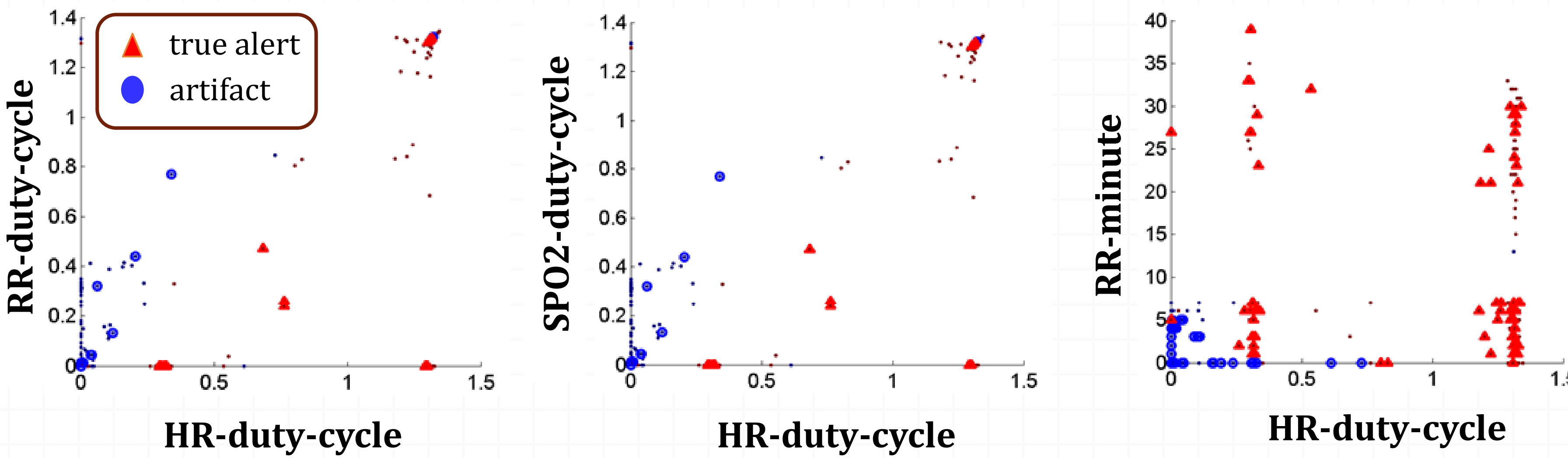
*duty cycle = number of readings over time units: a low value indicates high sparseness

Case Study – Alert Classification
- utility of RIPR models -

- The model selects HR duty cycle as the most important dimension in RR artifact classification, validating expect intuition
- Uncommon RR artifacts are classified as true alerts
- The RR signals are irregular
- Such cases can be identified through using variance of signal (new features added)
- RIPR model pointed out some mislabeled alerts



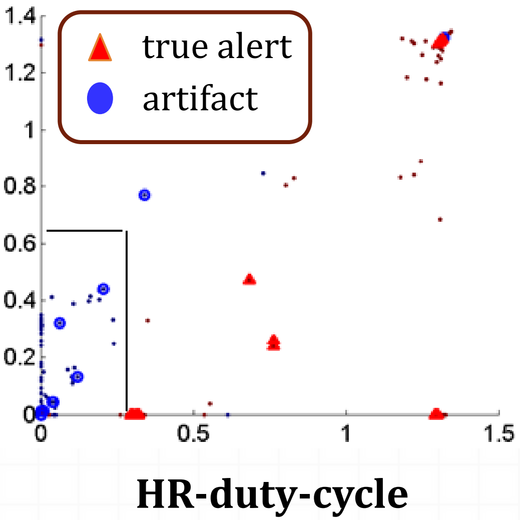
Case Study – Alert Classification - deriving rules -

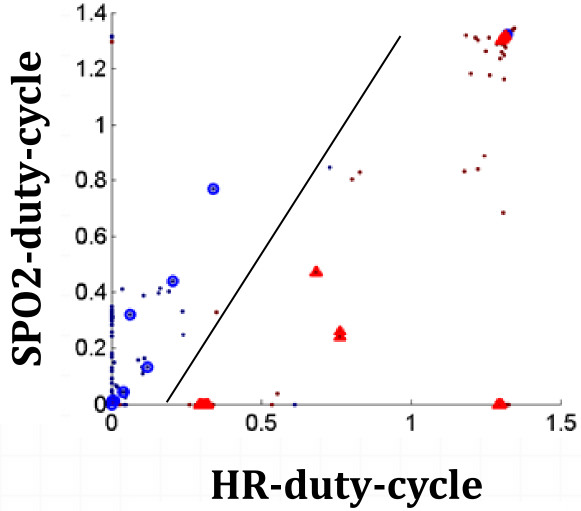


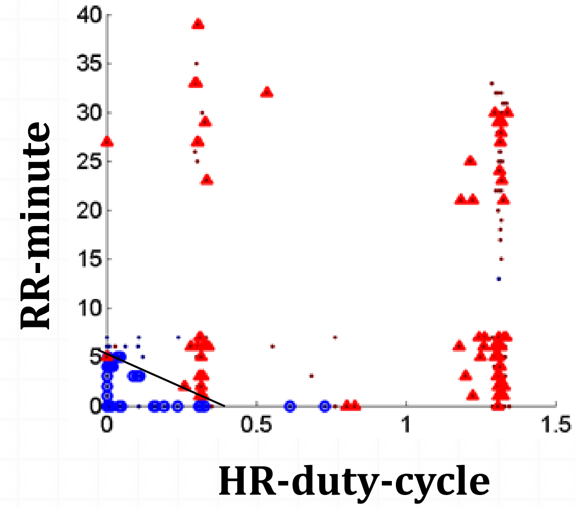
Alarm Type	RR	BP		SPO_2	
	2D	2D	3D	2D	3D
Accuracy	0.98	0.833	0.885	0.911	0.9151
Precision	0.979	0.858	0.896	0.929	0.9176
Recall	0.991	0.93	0.958	0.945	0.9957

*data density = number of readings over time units: a low value indicates high sparseness

Case Study – Alert Classification - deriving rules -







RR-duty-cycle* <= 0.6
and
HR-duty-cycle <= 0.25

HR-data-density –
SPO_2-data-density <= 0.2

HR-data-density/0.3
+ RR-min/5 <= 1

Summary

- Informative Projection Retrieval is relevant to many applications requiring interaction with human users
- We generalized RIPR, our solution to the IPR problem, to a wide range of learning tasks (classification, regression, clustering)
- RIPR expresses loss though divergence estimators
  - Semi-supervised models: penalize unlabeled data that cannot be confidently assigned to a class
  - Clustering models: favor high data density
- RIPR models are compact and well-performing in practice
  - IPs accurately recovered
  - Often more accurate than classifiers trained on all features
- Overall, RIPR contributes to the improvement of the quality of care for ICU