Classification model and survival analysis in biomedicine

Xiaoqian Jiang, PhD

Biomedical Informatics
University of California San Diego

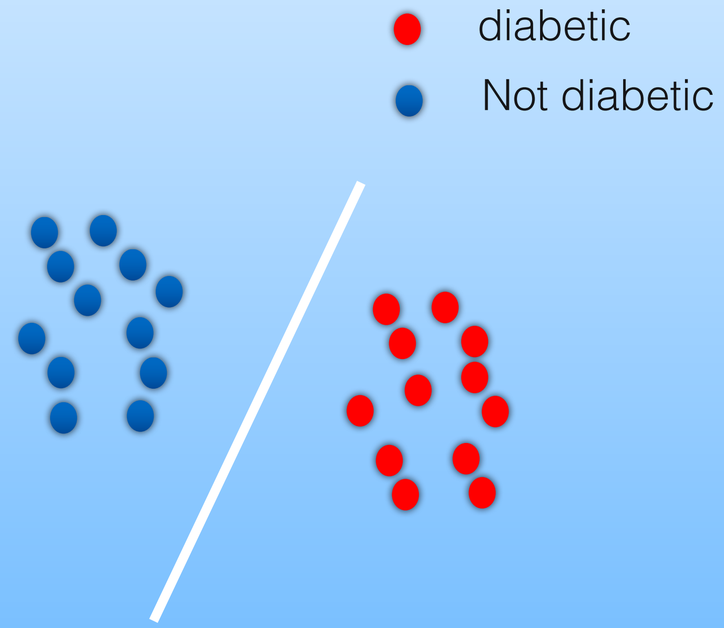
Predictive models in biomedicine

- Predictive models can facilitate risk assessment, diagnosis, prognosis in clinical care
- Typical biomedical predictive models include classification models (logistic regression) and time-to-event models (Cox proportional hazards)

Classification models

Binary classification

Age	Gender	BMI	Diabetes
12	f	18	n
45	m	36	y
…	…	…	…



A high level overview

- Inputs
  - Training examples {<Xi, yi>}, i stands for the i-th example.
- Output
  - Approximation of the unknown target function *f* that map X->y

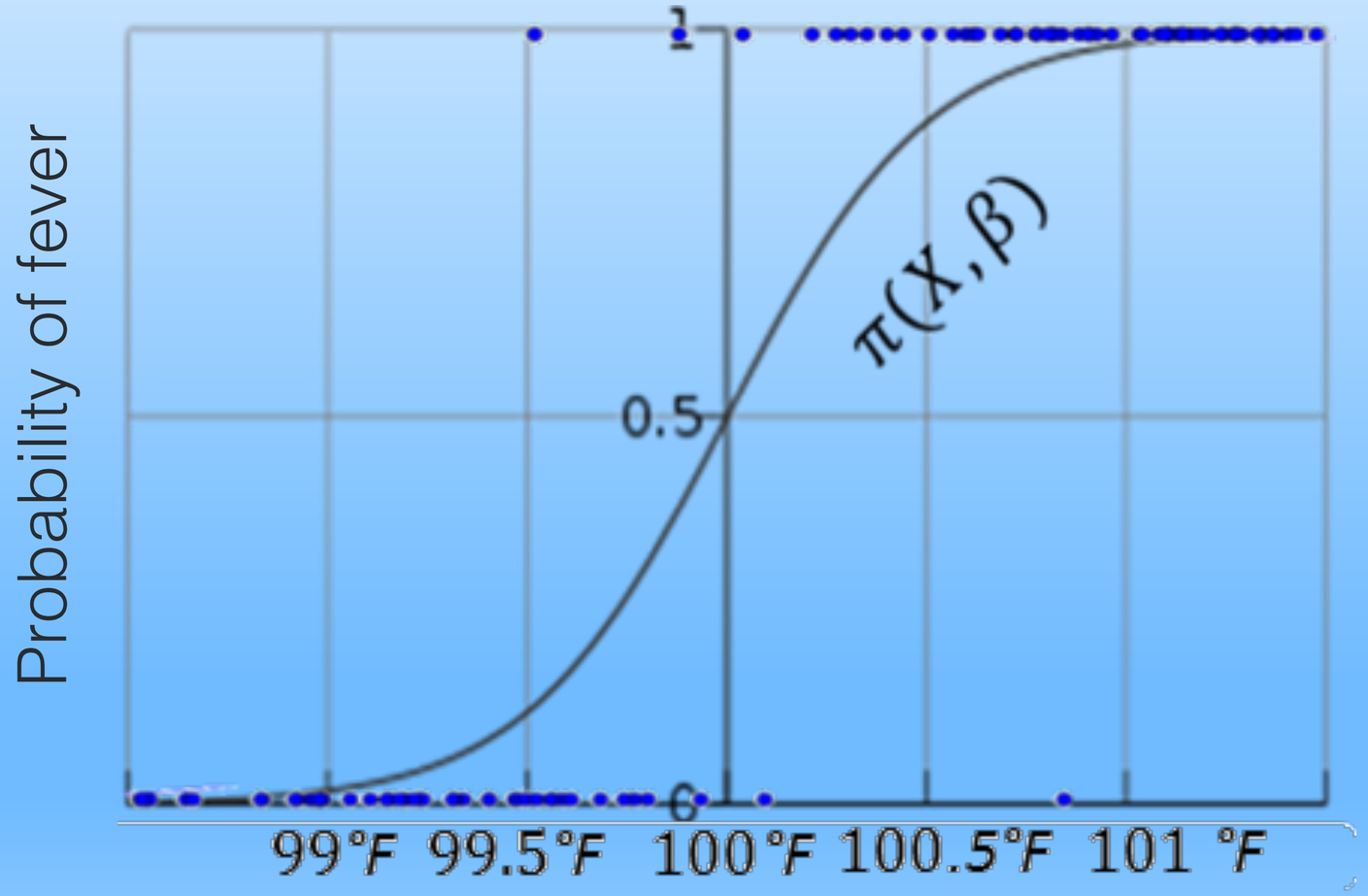
There are many classification models

- Decision tree
- Naïve Bayes
- Ada-boost
- Support vector machine
- Deep learning

In biomedicine, classification power is critical but there are other important concerns like calibration and interpretation

Models allow explicit interpretation of the results are more favorable in practice

Logistic Regression



Maximum Likelihood Estimation

- Estimated probability based on observations of a binary response Y and covariates X

$$P(Y = 1|X) = \pi(X, \beta) = \frac{1}{1 + e^{-X\beta}}$$

Binary response

Covariates

Logit function

Model parameter

- Likelihood function based on observed data

Number of records

$$l(\beta) = \sum_{1}^{n} [y_i \log \pi(x_i, \beta) + (1 - y_i) \log(1 - \pi(x_i, \beta))]$$

Maximum Likelihood Estimation

- Gradient descent algorithm for calculation

$$P(Y = 1|X) = \pi(X, \beta) = \frac{1}{1 + e^{-X\beta}}$$

$l(\beta)$ is a concave function

$$l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

$$\beta^{(k+1)} = \beta^{(k)} - \lambda \frac{\partial l(\beta^{(k)})}{\partial \beta^{(k)}}$$

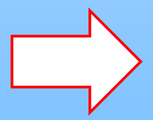
λ is a learning rate parameter, which controls the convergence speed

Maximum Likelihood Estimation

- Newton-Raphson algorithm for calculation

$$P(Y = 1|X) = \pi(X, \beta) = \frac{1}{1 + e^{-X\beta}}$$

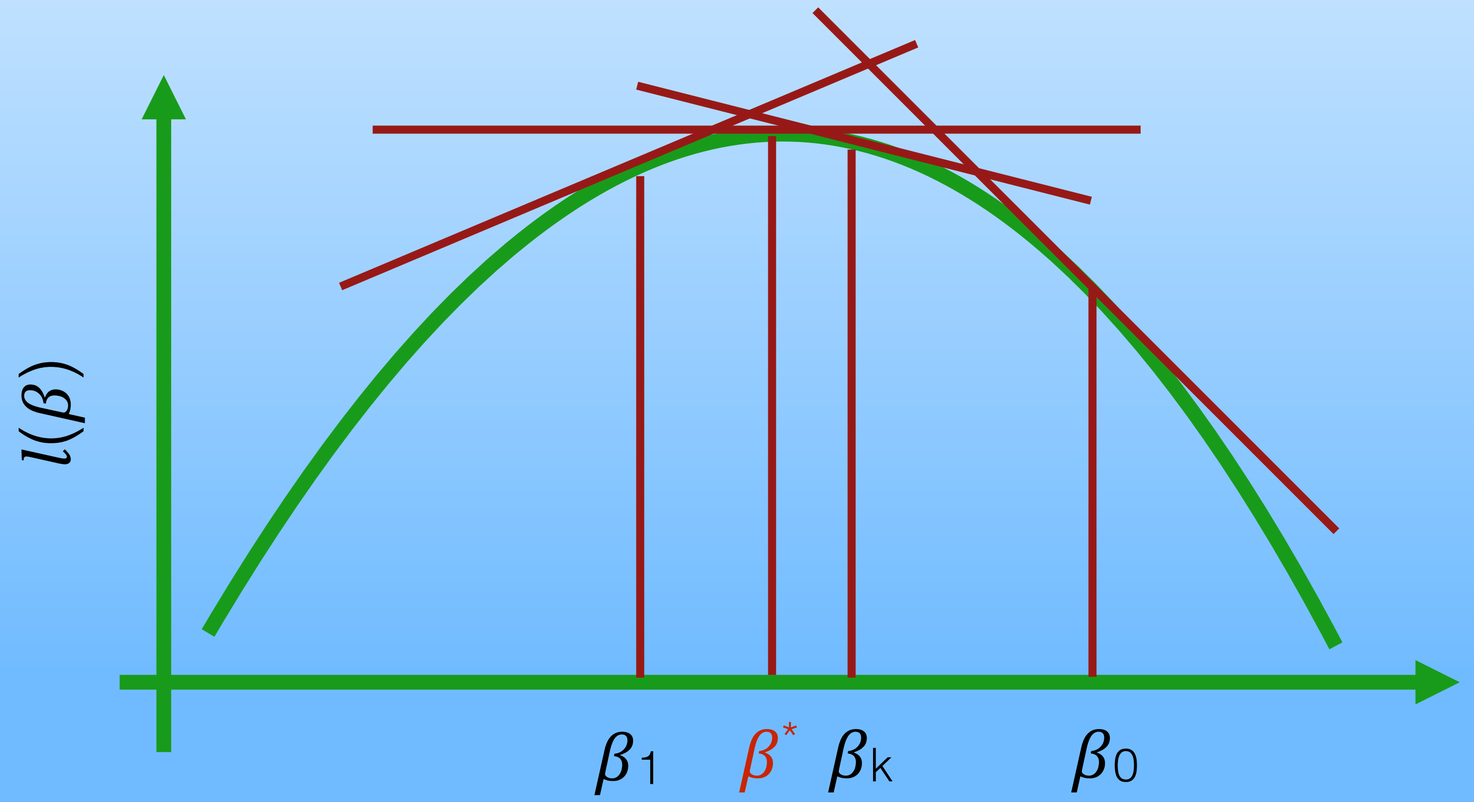
$l(\beta)$ is a concave function



$$l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

$$\beta^{(k+1)} = \beta^{(k)} - \left[\frac{\partial^2 l(\beta^{(k)})}{\partial \beta^{(k)} \partial \beta^{(k)^T}} \right]^{-1} \frac{\partial l(\beta^{(k)})}{\partial \beta^{(k)}}$$

Newton-Raphson (NR) Algorithm



$$l(\boldsymbol{\beta}) = \sum_{i=1}^{D} \{\boldsymbol{\beta}^T \sum_{l \in \mathcal{D}_i} \mathbf{z}^l - d_i \log[\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)]\}$$

$$\beta^{(k+1)} = \beta^{(k)} - \left[\frac{\partial^2 l(\beta^{(k)})}{\partial \beta^{(k)} \partial \beta^{(k)^T}}\right]^{-1} \frac{\partial l(\beta^{(k)})}{\partial \beta^{(k)}}$$

Other regression models

- Logistic regression is one type of the Generalized Linear Model, if we change the plugin function, it can handle response variable with different distributional assumption
  - If $\pi(X, \beta) = \beta X$, the model become the simple linear regression, which has a closed-form solution for the parameter β given a continuous outcome variable
  - If $\pi(X, \beta) = e^{\beta X}$, the model becomes Poisson regression and can handle outcome variable exhibiting a Poisson distribution
  - If $\pi(X, \beta)$ is log-linear, the model can handle categorical data

$$P(Y = 1|X) = \pi(X, \beta) = \frac{1}{1 + e^{-X\beta}}$$

Binary response

Covariates

Plugin function

Model parameter

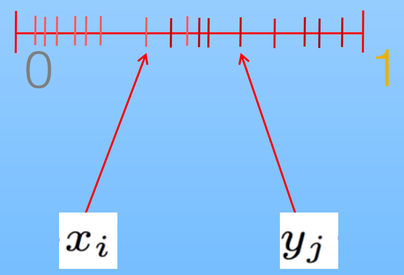
Outputs of the logistic regression

- Estimated probability based on observations of a binary response variable and covariates
- Estimated parameters that maximize the likelihood function
- Estimated standard deviation for each attribute
- Odds ratio and *z*-test statistics (indicating significant attribute)

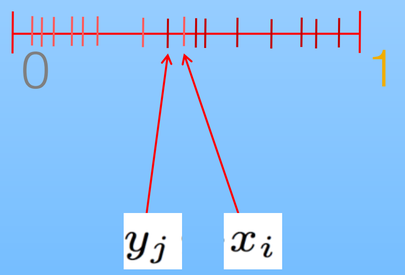
We need to evaluate the trained model first to measure its discrimination and goodness-of-fit to confirm that it is a good model to use

Calculation of AUC

$$\hat{\theta} = \frac{1}{m \cdot n} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{1}_{x_i < y_j}$$

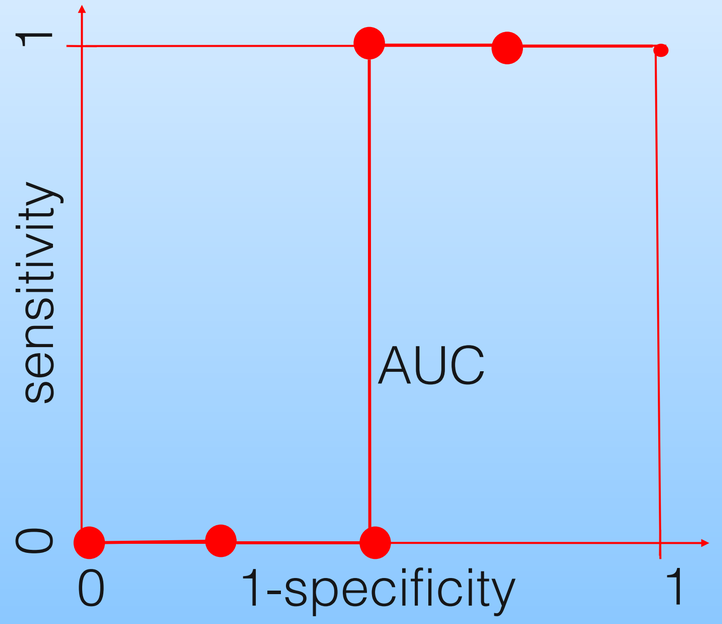


$\mathbf{1}_{x_i < y_j}$ = 1



$\mathbf{1}_{x_i < y_j}$ = 0

Calculation of AUC



Prediction	Label
0.1	0
0.2	0
0.32	1
0.33	0
0.35	0

	P	N	
"P"	0	4	4
"N"	1	0	1
	1	4	

[0,1]

	P	N	
"P"	0	3	4
"N"	1	1	1
	1	4	

[0,0.75]

	P	N	
"P"	0	2	4
"N"	1	2	1
	1	4	

[0,0.5]

	P	N	
"P"	1	2	4
"N"	0	2	1
	1	4	

[1,0.5]

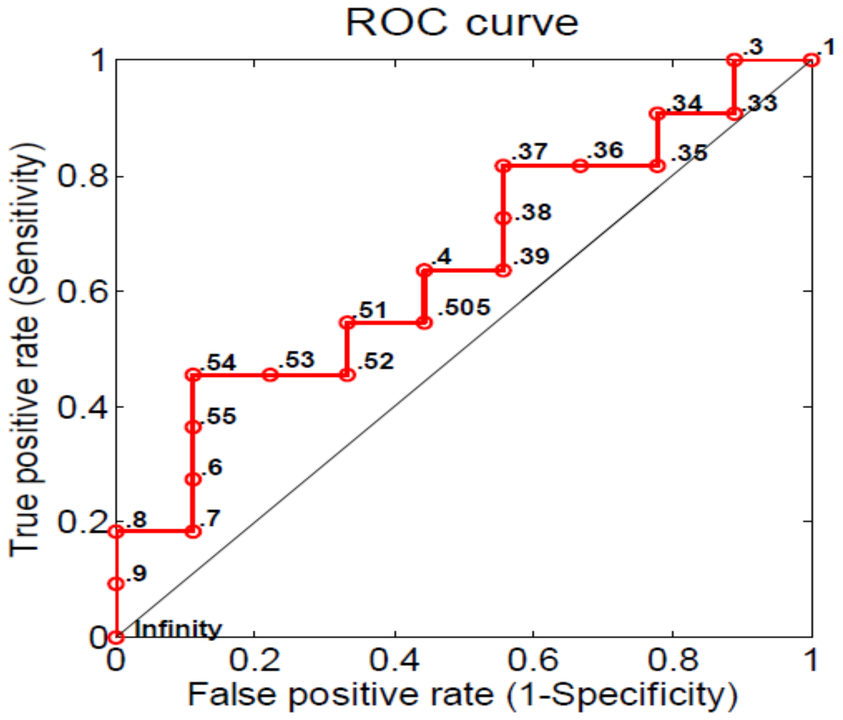
	P	N	
"P"	1	1	4
"N"	1	3	1
	1	4	

[1,0.25]

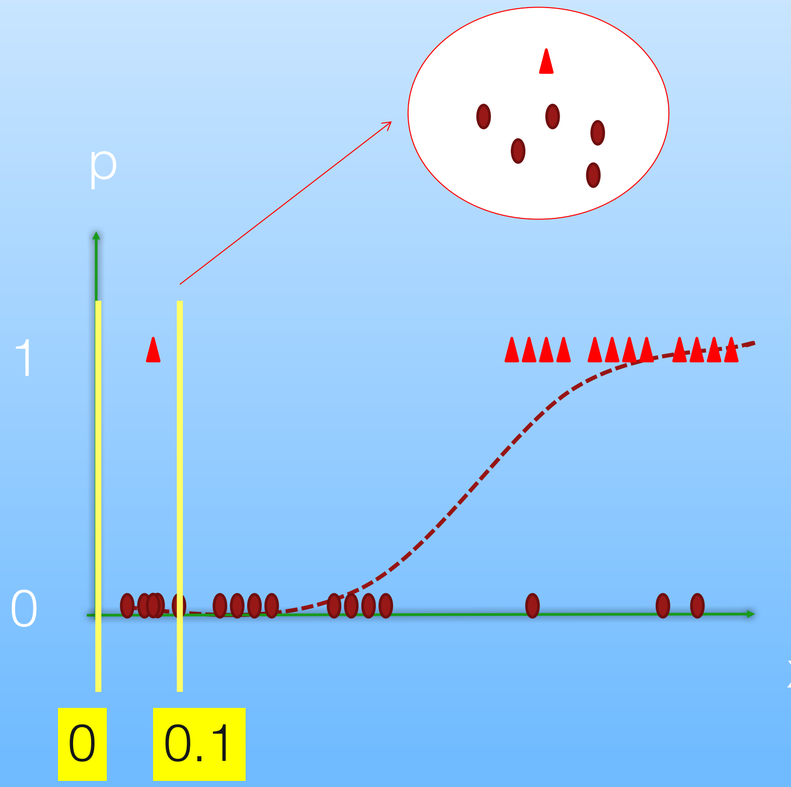
Discriminative models are not necessarily calibrated

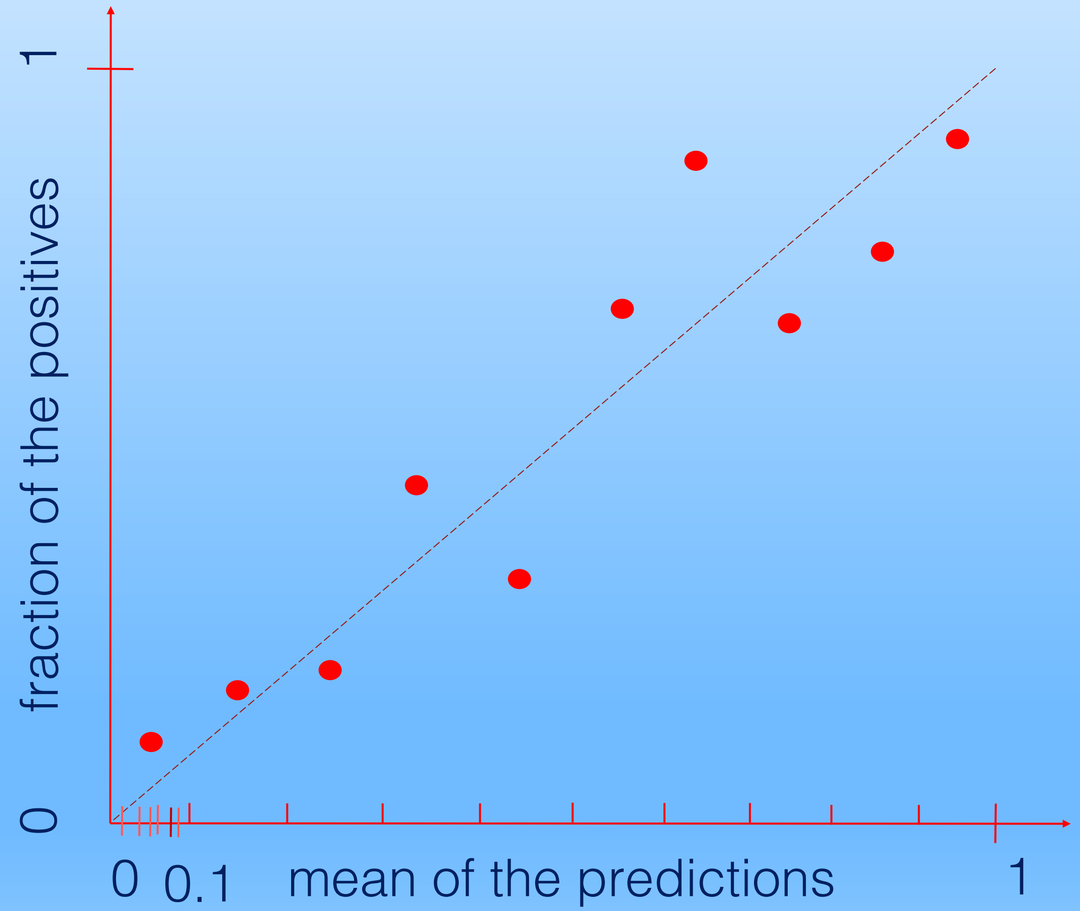
#	1	2	3	4	5	6	7	8	9	10
C	p	p	n	p	p	p	n	n	p	n
P	.9	.8	.7	.6	.55	.54	.53	.52	.51	.505
#	11	12	13	14	15	16	17	18	19	20
C	p	n	p	p	n	n	p	n	p	n
P	.4	.39	.38	.37	.36	.35	.34	.33	.3	.1

(1) Probabilistic Classifier A



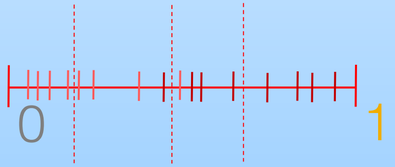
Reliability diagram

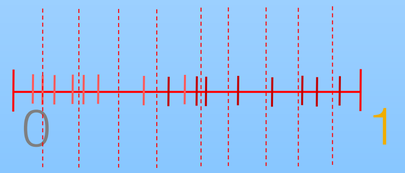




Goodness-of-fit test

- Most famous test is Hosmer and Lemeshow
  - C statistic
  - H statistic



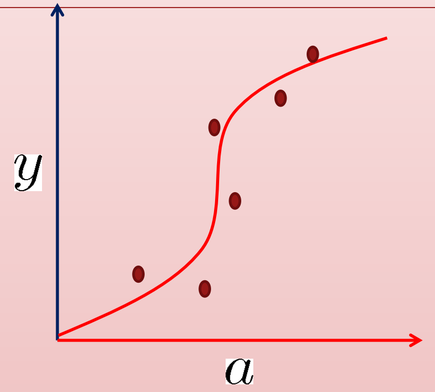
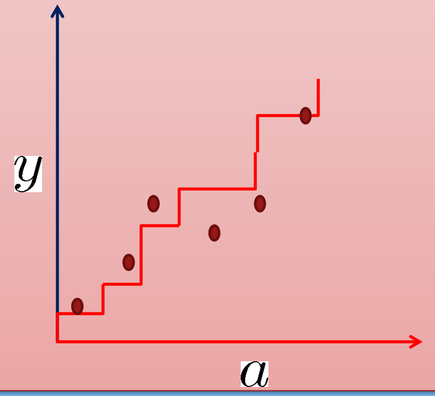


  - X^2 distribution with g-2 degrees of freedom

$$\sum_{k=1}^{g} \frac{(o_k - n_k' \bar{\pi}_k)^2}{n_k' \bar{\pi}_k (1 - \bar{\pi}_k)}$$

Can we improve calibration without hurting classification?

Yes, use functions that preserve the ranking order

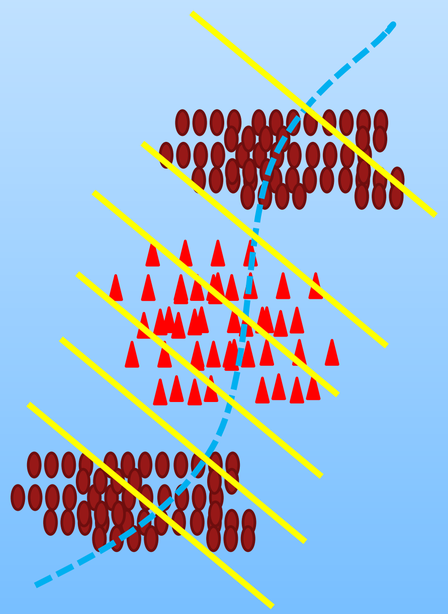
Function	Formulations
	$y_i = \frac{1}{1 + e^{-Aa_i + B}}$
	$\min \sum_{i=1}^{n} w_i \|\|y_i - a_i\|\|^2$ subject to $y_i \geq y_{i+1}\ \forall i.$

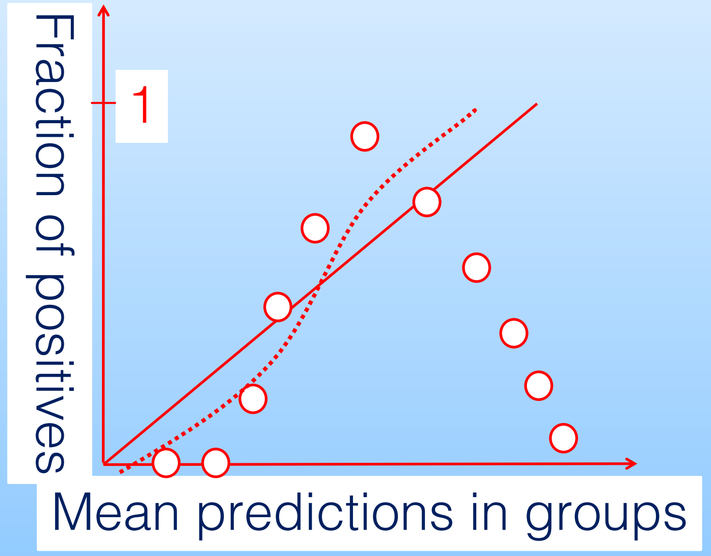
Platt scaling

Isotonic Regression

Monotonic increasing functions “preserves” the ordering, thus the discrimination power.

Limitation of Platt Scaling



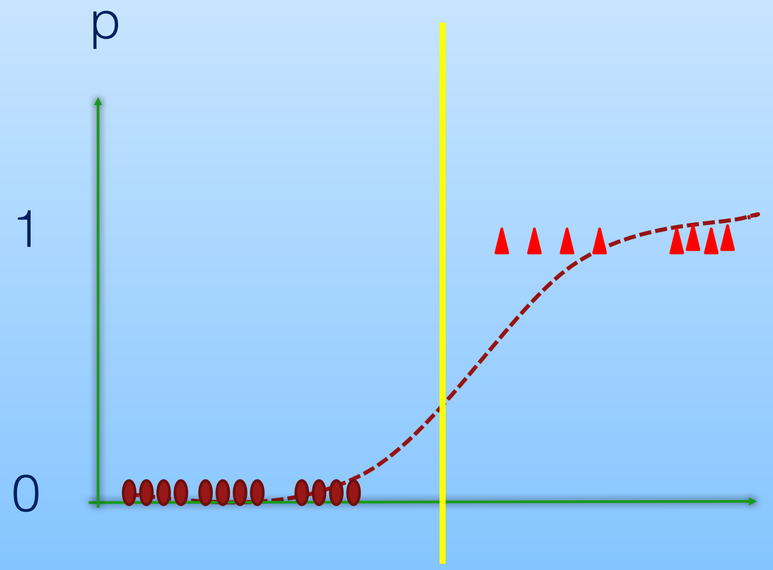


$$y_i = \frac{1}{1 + e^{-Aa_i + B}}$$

Platt Scaling

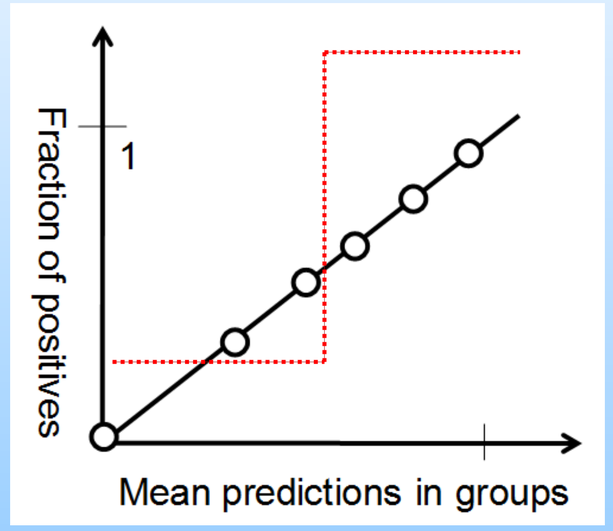
Does not fit!

Limitation of Isotonic Regression



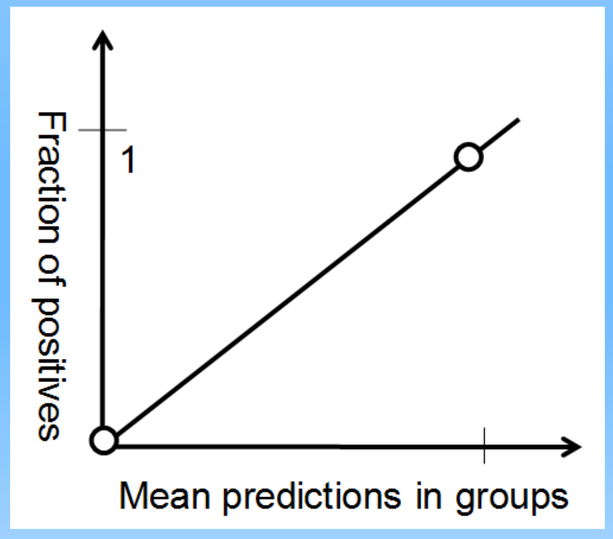
▲ Class “1”

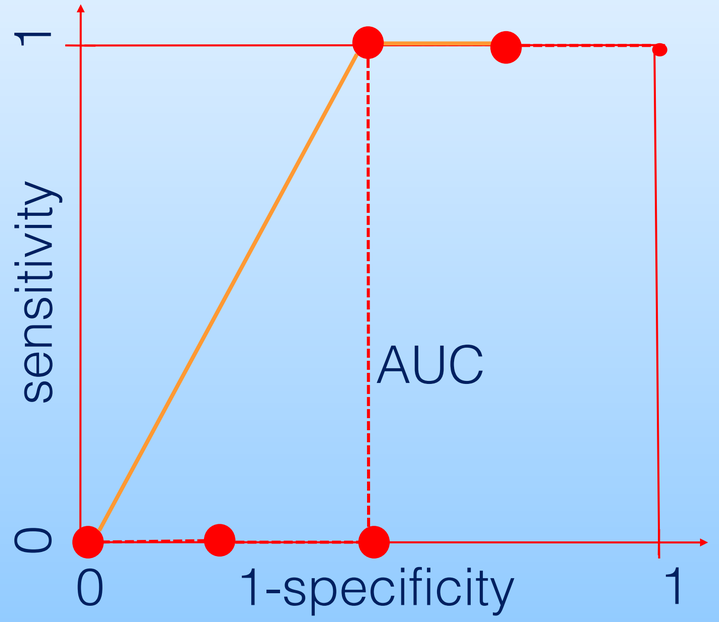
● Class “0”



Isotonic Regression

$$\min \sum_{i=1}^{n} w_i ||y_i - a_i||^2$$
$$\text{subject to } y_i \geq y_{i+1} \ \forall i.$$





Prediction	Label
0.1	0
0.2	0
0.32	1
0.33	0
0.35	0

Prediction	Label
0.1	0
0.2	0
0.33	1
0.33	0
0.33	0

Isotonic Regression

	P	N
"P"	0	4
"N"	1	0

[0,1]

	P	N
"P"	1	2
"N"	0	2

[1,0.5]

	P	N
"P"	1	1
"N"	1	3

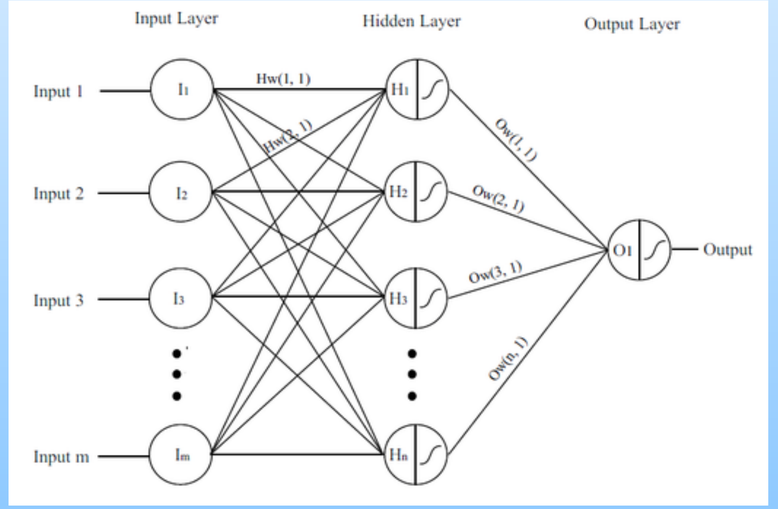
[1,0.25]

References

- Zadrozny B, Elkan C. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining New York, NY, USA: ACM; 2002. p. 694–699.
- Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. J Am Med Inform Assoc 2012;19(2):263–274. PMID: 21984587
- Wu Y, Jiang X, Kim J, Ohno-Machado L, Jolla L. I-spline Smoothing for Calibrating Predictive Models. AMIA Summits Transl Sci Proc 2012. p. 39–46.
- Jiang X, Menon A, Wang S, Kim J, Ohno-Machado L. Doubly Optimized Calibrated Support Vector Machine (DOC-SVM): An Algorithm for Joint Optimization of Discrimination and Calibration. PLoS One 2012 Jan;7(11):e48823. PMID: 23139819
- Menon AK, Edu AU, Edu MU, Jiang X, Vambu S, Elkan C, et al. Predicting accurate probabilities with a ranking loss. International Conference on Machine Learning Edinburgh, Scotland, UK; 2012. p. CoRR abs/1206.4661.
- Jiang X, Osl M, Kim J, Ohno-Machado L, Wu Y, Jolla L. Smooth Isotonic Regression: A New Method to Calibrate Predictive Models. AMIA Summits Transl Sci Proc San Francisco, CA; 2011. p. 16–20. PMID: 22211175
- Naeini MP, Cooper GF, Hauskrecht M. Obtaining Well Calibrated Probabilities Using Bayesian Binning. Proc Conf AAAI Artif Intell 2015 Jan;2015:2901–2907. PMID: 25927013
- Naeini MP, Cooper GF. Binary Classifier Calibration using an Ensemble of Near Isotonic Regression Models [Internet]. arXiv [csLG]. 2015. Available from: http://arxiv.org/abs/1511.05191

Logistic regression and deep learning

- Artificial Neural Network (and some deep learning models) is essentially a stack of logistic regression units that repeatedly use the sigmoidal function for optimization



Function train(X, Y)

1. $W_0 = random(K, H)$ // weights between input and hidden layers of size $K \times H$
2. $W_1 = random(H)$ // weights between hidden and output layers of size $H \times 1$
3. For $j = 1 : max\_Iteration$
4. $layer_1 = sigmoid(X * W_0)$ // forward outputs of size $N \times H$
5. If($do\_dropout$) // if the dropout flag is true
6. $layer_1 = layer_1 * random.binomial(H, 1 - p/(1 - p))$ // randomly dropout
7. $layer_2 = sigmoid(layer_1 * W_1)$ // forward outputs of size $N \times 1$
8. $layer_{2delta} = (layer_2 - Y) .* (layer_2 .* (1 - layer_2))$ // backpropagation errors $N \times 1$
9. $layer_{1delta} = (layer_{2delta} * {W_1}^T) .* (layer_1 .* (1 - layer_1))$ // backpropagation errors $N \times H$
10. $W_1 = W_1 - (\gamma * {layer_1}^T * layer_{2delta})$ //adjust the weights of size $H \times 1$
11. $W_0 = W_0 - (\gamma * X^T * layer_{1delta})$ //adjust the weights of size $K \times H$

Survival analysis

Survival Data

Patient ID	Age	DrugDose	Race	Treatment	Time	Censored
1	37	7	0	1	3	1
2	37	6	0	0	4	1
3	39	0	0	0	4	0
4	36	2	0	1	4	1
5	35	12	1	1	5	0
6	33	2	1	1	5	1
7	29	3	0	0	5	0
8	37	0	0	1	5	1
9	35	1	0	0	6	1
10	30	3	1	0	6	0
11	43	0	1	1	6	1
12	42	20	0	0	7	0

Survival Data

Covariates Z

Patient ID	Age	DrugDose	Race	Treatment	Time	Censored
1	37	7	0	1	3	1
2	37	6	0	0	4	1
3	39	0	0	0	4	0
4	36	2	0	1	4	1
5	35	12	1	1	5	0
6	33	2	1	1	5	1
7	29	3	0	0	5	0
8	37	0	0	1	5	1
9	35	1	0	0	6	1
10	30	3	1	0	6	0
11	43	0	1	1	6	1
12	42	20	0	0	7	0

Survival Data

Time-to-event: The time from entry into a study until a subject has a particular outcome t

Patient ID	Age	DrugDose	Race	Treatment	Time	Censored
1	37	7	0	1	3	1
2	37	6	0	0	4	1
3	39	0	0	0	4	0
4	36	2	0	1	4	1
5	35	12	1	1	5	0
6	33	2	1	1	5	1
7	29	3	0	0	5	0
8	37	0	0	1	5	1
9	35	1	0	0	6	1
10	30	3	1	0	6	0
11	43	0	1	1	6	1
12	42	20	0	0	7	0

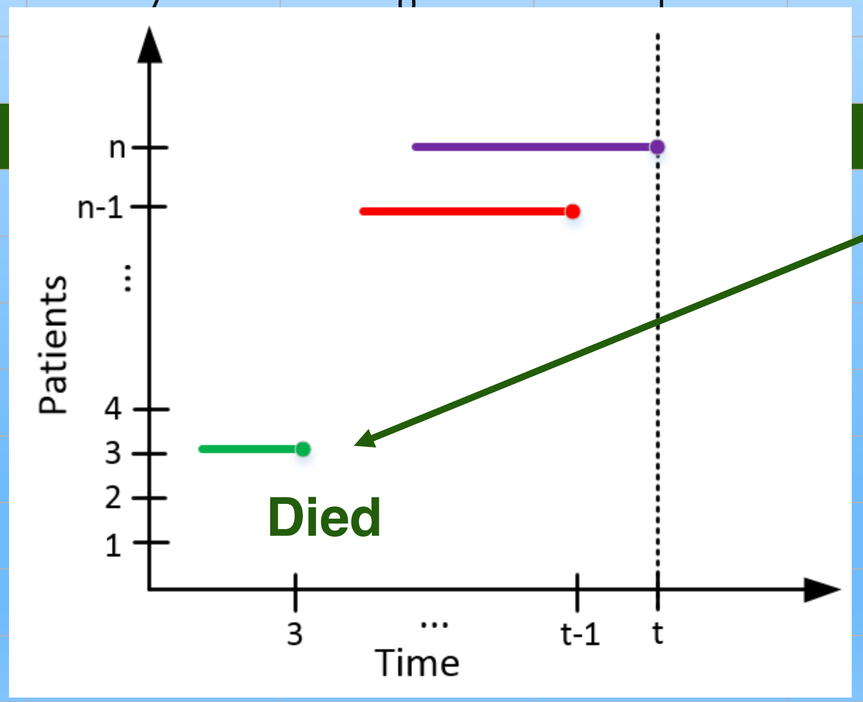
Tied events

Survival Data

Patient ID	Age	DrugDose	Race	Treatment	Time	Censored
1	37	7	0	1	3	1
2	37	6	0	0	4	1
3	39	0	0	0	4	0
4	36	2	0	1	4	1
5	35	12	1	1	5	0
6	33	2	1	1	5	1
7	29	3	0	0	5	0
8	37	0	0	1	5	1
9	35	1	0	0	6	1
10	30	3	1	0	6	0
11	43	0	1	1	6	1
12	42	20	0	0	7	1

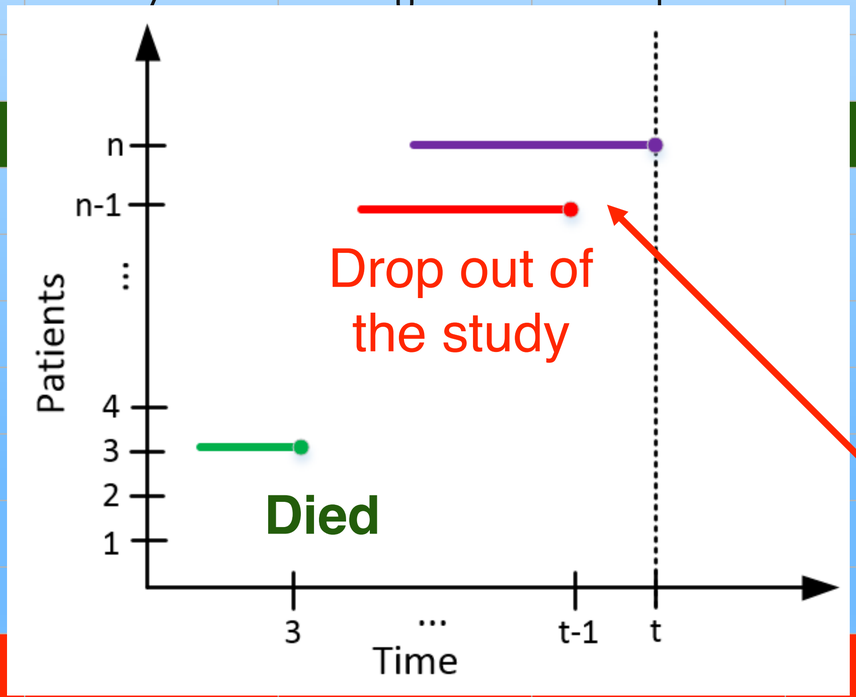
Survival Data

Patient ID	Age	DrugDose	Race	Treatment	Time	Censored
1	37	7	0	1	3	1
2	37				4	1
3	39				4	0
4	36				4	1
5	35				5	0
6	33				5	1
7	29				5	0
8	37				5	1
9	35				6	1
10	30				6	0
11	43				6	1
12	42	20	0	0	7	1



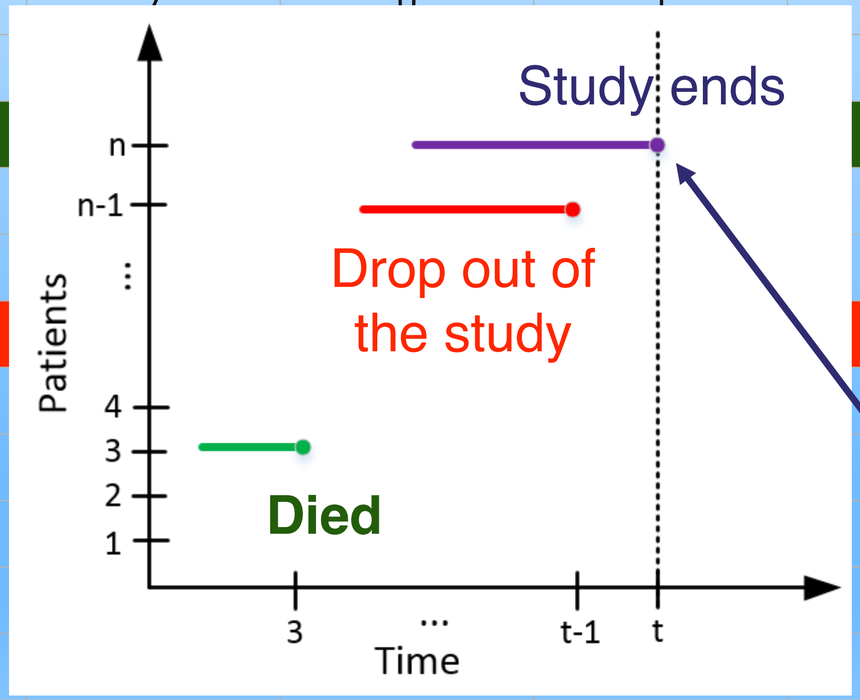
Survival Data

Patient ID	Age	DrugDose	Race	Treatment	Time	Censored
1	37	7	0	1	3	1
2	37				4	1
3	39				4	0
4	36				4	1
5	35				5	0
6	33				5	1
7	29				5	0
8	37				5	1
9	35				6	1
10	30				6	0
11	43				6	1
12	42	20	0	0	7	1



Survival Data

Patient ID	Age	DrugDose	Race	Treatment	Time	Censored
1	37	7	0	1	3	1
2	37				4	1
3	39				4	0
4	36				4	1
5	35				5	0
6	33				5	1
7	29				5	0
8	37				5	1
9	35				6	1
10	30				6	0
11	43				6	1
12	42	20	0	0	7	1



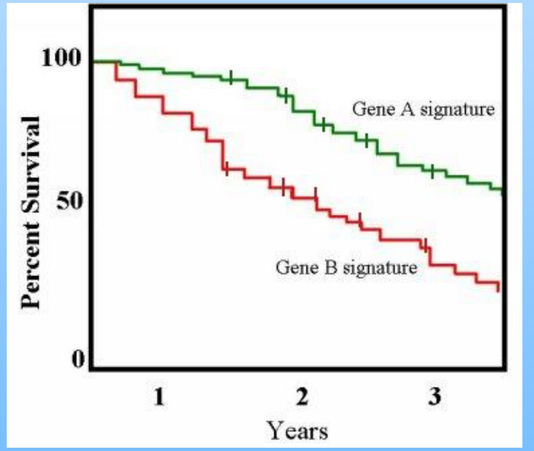
Kaplan-Meier survival estimate

The estimator is given by

$$\hat{S}(t) = \prod_{i:t_i<t}(1 - \frac{d_i}{n_i})$$

Here d_i is the number of events and n_i is the total individuals at risk at time i, this is a straightforward model to represent the percent of survival at a given time t

But this model does not give individual risk at a given time



https://en.wikipedia.org/wiki/Kaplan–Meier_estimator

Cox Proportional Hazards

Breslow's partial likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp(\boldsymbol{\beta}^T \sum_{l \in \mathcal{D}_i} \mathbf{z}^l)}{[\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)]^{d_i}}$$

Covariates of the *l*-th subject in the given set

Number of subjects has time-to-event at t=t_i

D: the number of distinct event time

A set of subjects has risk at t=t_i

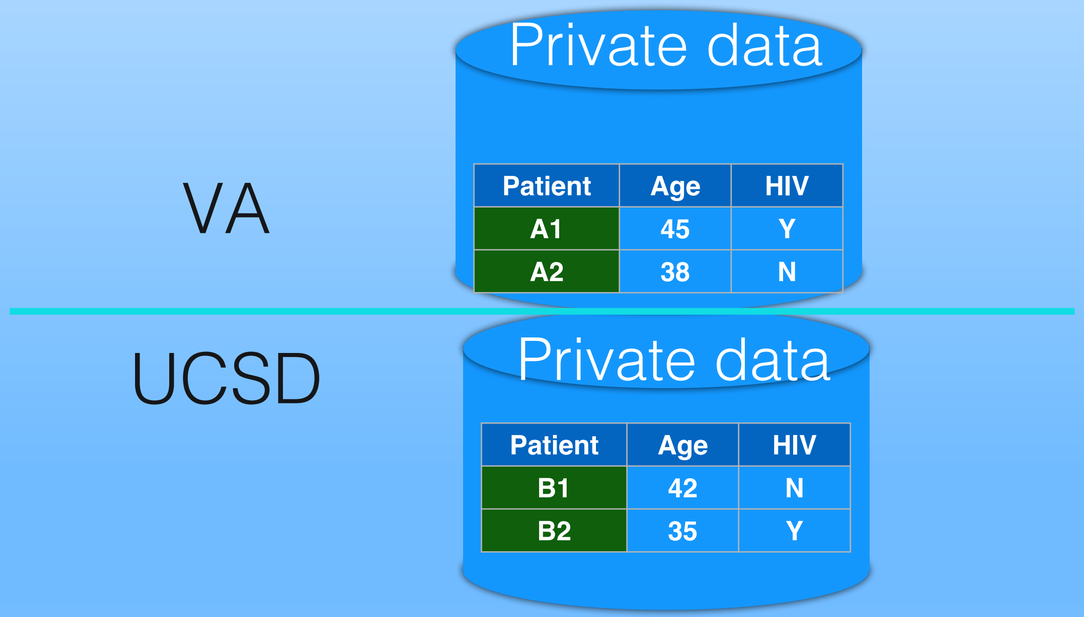
A set of subjects has time-to-event at t=t_i

Id	Covariates z		Time	Censored	
1	…	i=1	3	1	
2	…		4	1	
3	…	i=2	4	1	
4	…		4	0	
5	…		5	1	$\mathcal{R}_3$
6	…	i=3	5	1	
7	..		5	0	$\mathcal{D}_3$
8	..		5	0	

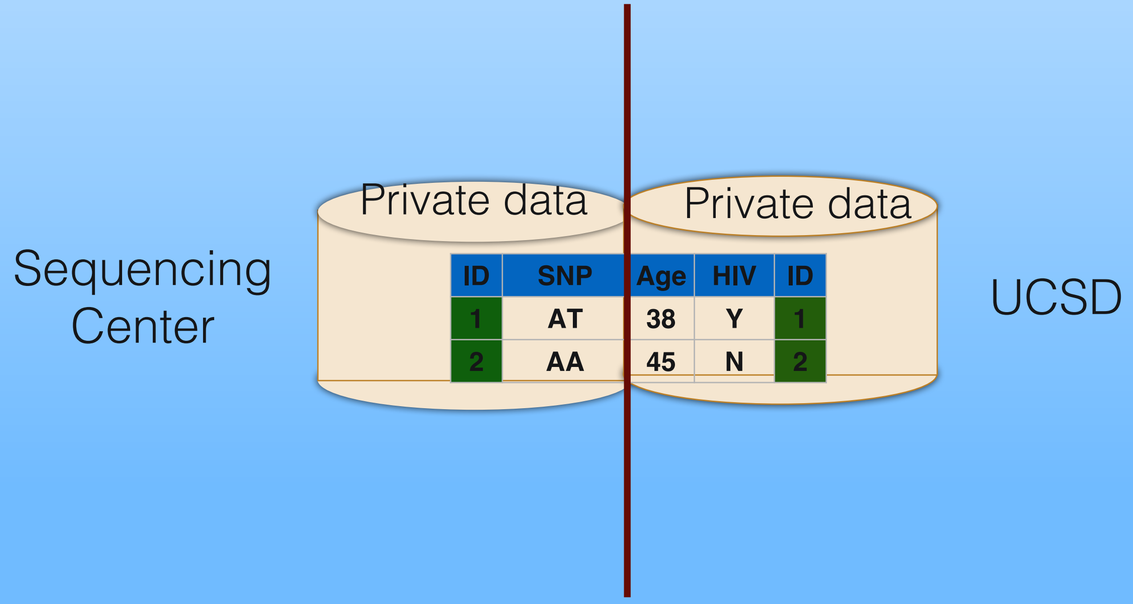
$d_3=2$

When data are distributed

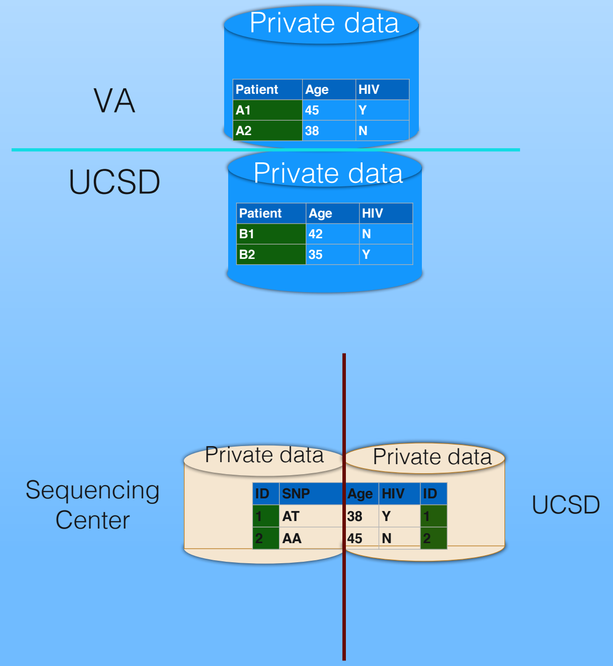
Horizontally distributed data



Vertically distributed data



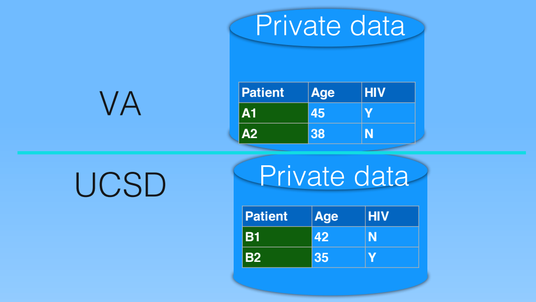
Common Distributed Data Analysis



	Logistic Regression	Cox Model (Survival analysis)
Frequentist	GLORE[1,2,3]	WebDISCO[5]
Bayesian	EXPLORER[4]	To be investigated
Frequentist	VERTIGO[6]	VERTICOX[WIP]
Bayesian	To be investigated	To be investigated

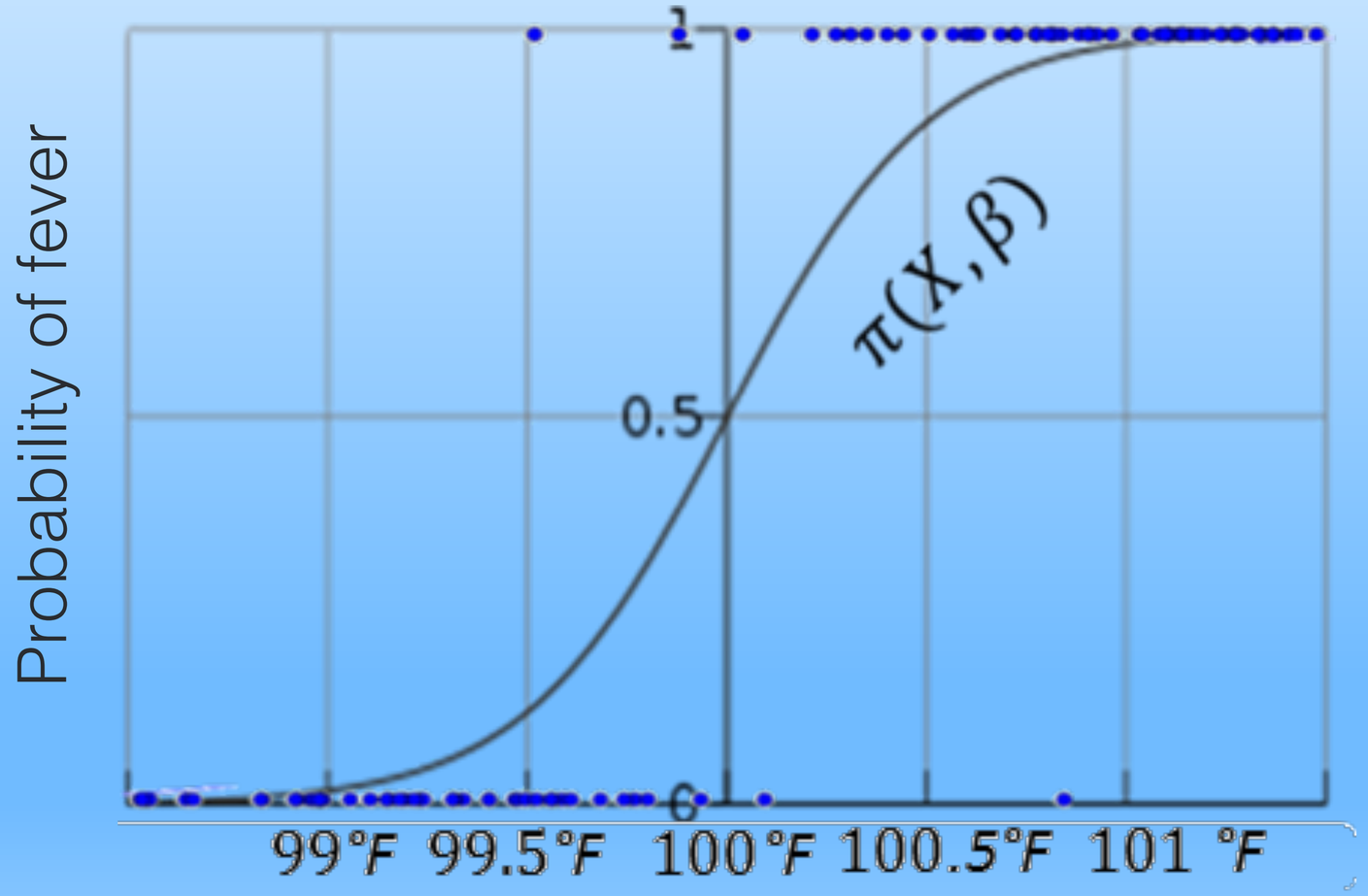
1. Jiang X *et al.* WebGLORE: a web service for Grid LOgistic REgression. *Bioinformatics* 2013
2. Wu Y, Jiang X, *et al.* Grid multi-category response logistic models. *BMC Med Inform Decis Mak* 2015
3. Jiang X, Wu Y, *et al.,* Development of a web service for analysis in a distributed network. *EGEMS (Wash DC) Academy Health*; 2014
4. Wang, S., Jiang, X, *et al.,* Expectation propagation logistic regression (explorer): distributed privacy-preserving online model learning. *Journal of biomedical informatics*, 2013
5. Lu C-L, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *J Am Med Inform Assoc* 2015
6. Jiang X, Li Y, et al. VERTIcal Grid lOgistic regression (VERTIGO). *J Am Med Inform Assoc* 2016

GLORE: Grid LOgistic Regression (extends to generalized linear models)



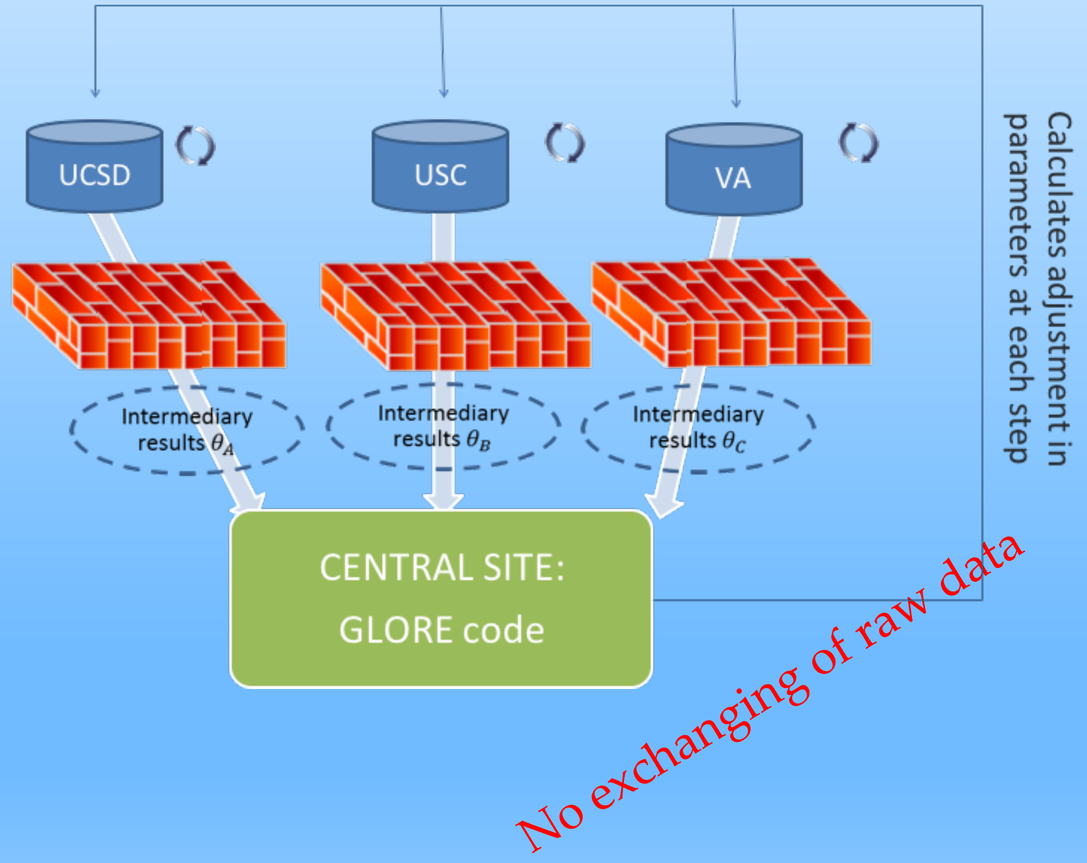
Jiang X, *et al.* WebGLORE: a web service for Grid LOgistic REgression. *Bioinformatics* 2013

Logistic Regression



Foundation of GLORE

- Support *p-1* features are consistent over *k* sites
- In each iteration, intermediary result of a *pxp* matrix and a *p*-dimensional vector are transmitted to the central site for optimization



Maximum Likelihood Estimation

- Estimated probability based on observations of a binary response Y and covariates X

$$P(Y = 1|X) = \pi(X, \beta) = \frac{1}{1 + e^{-X\beta}}$$

Binary response

Covariates

Logit function

Model parameter

- Likelihood function based on observed data (centralized)

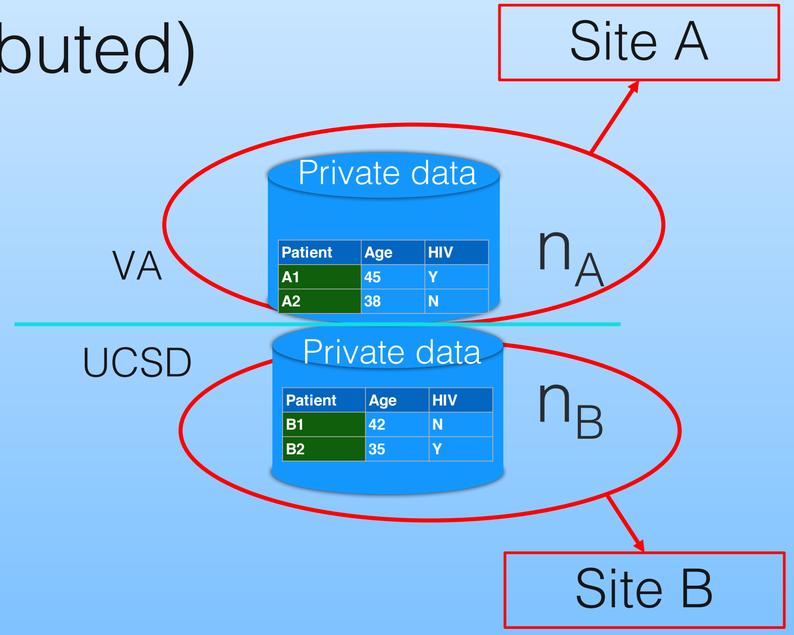
Number of records

$$l(\beta) = \sum_{1}^{n} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

Maximum Likelihood Estimation

- Likelihood function based on observed data (distributed)

$$P(Y = 1|X) = \pi(X, \beta) = \frac{1}{1 + e^{-X\beta}}$$



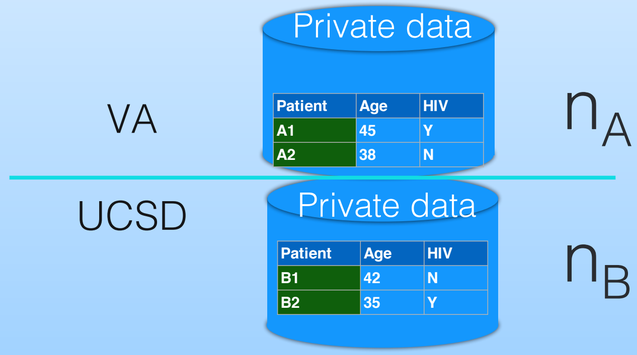
Number of records held by site A

Number of records held by site B

$$l(\beta) = \sum_{1}^{n_A + n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

$$l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

$$l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$



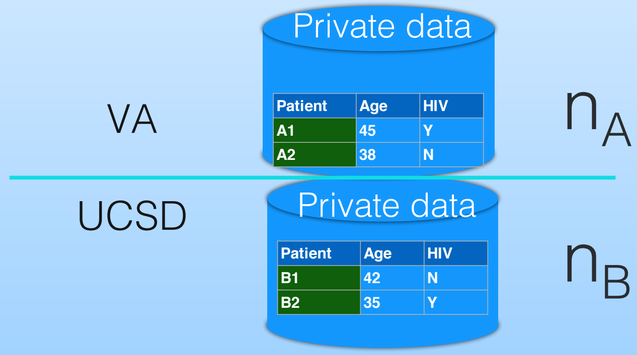
$$l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

$$\beta^{(k+1)} = \beta^{(k)} - \left[\frac{\partial^2 l(\beta^{(k)})}{\partial \beta^{(k)} \partial \beta^{(k)^T}} \right]^{-1} \frac{\partial l(\beta^{(k)})}{\partial \beta^{(k)}}$$

$$= \beta^{(k)} + \left[\bar{X}^T W\left(\bar{X}, \beta^{(k)}\right) \bar{X} \right]^{-1} \bar{X}^T \left[\bar{Y} - \Pi\left(\bar{X}, \beta^{(k)}\right) \right]$$

Global variance-covariance matrix

Global prediction outcomes

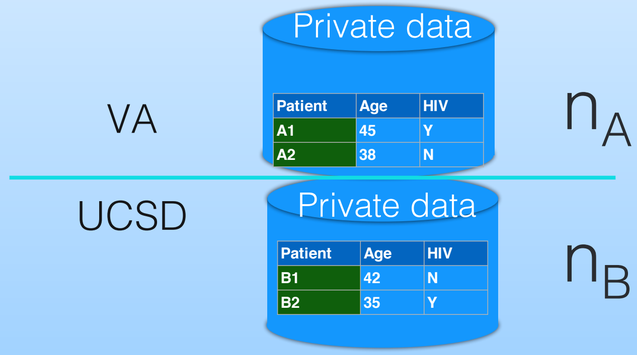


$$l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

$$\begin{aligned}
\beta^{(k+1)} &= \beta^{(k)} - \left[\frac{\partial^2 l(\beta^{(k)})}{\partial \beta^{(k)} \partial \beta^{(k)T}} \right]^{-1} \frac{\partial l(\beta^{(k)})}{\partial \beta^{(k)}} \\
&= \beta^{(k)} + \left[\bar{X}^T W\left(\bar{X}, \beta^{(k)}\right) \bar{X} \right]^{-1} \bar{X}^T \left[\bar{Y} - \Pi\left(\bar{X}, \beta^{(k)}\right) \right] \\
&= \beta^{(k)} + \left[\bar{X}_A^T W_A\left(\bar{X}_A, \beta^{(k)}\right) \bar{X}_A + \bar{X}_B^T W_B\left(\bar{X}_B, \beta^{(k)}\right) \bar{X}_B \right]^{-1} \\
&\quad \cdot \{ \bar{X}_A^T [\bar{Y}_A - \Pi_A(\bar{X}_A, \beta)] + \bar{X}_B^T [\bar{Y}_B - \Pi_B(\bar{X}_B, \beta)] \}.
\end{aligned}$$

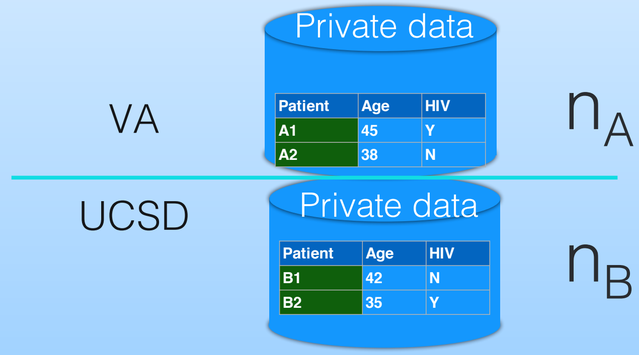
Local variance-covariance matrix

Local prediction outcomes



$$l(\beta) = \sum_{1}^{n_A+n_B} \left[y_i \log \pi(x_i, \beta) + (1 - y_i) \log\left(1 - \pi(x_i, \beta)\right) \right]$$

$$\begin{aligned}
\beta^{(k+1)} &= \beta^{(k)} - \left[\frac{\partial^2 l(\beta^{(k)})}{\partial \beta^{(k)} \partial \beta^{(k)^T}} \right]^{-1} \frac{\partial l(\beta^{(k)})}{\partial \beta^{(k)}} \\
&= \beta^{(k)} + \left[\bar{X}^T W\left(\bar{X}, \beta^{(k)}\right) \bar{X} \right]^{-1} \bar{X}^T \left[\bar{Y} - \Pi\left(\bar{X}, \beta^{(k)}\right) \right] \\
&= \beta^{(k)} + \left[\bar{X}_A^T W_A\left(\bar{X}_A, \beta^{(k)}\right) \bar{X}_A + \bar{X}_B^T W_B\left(\bar{X}_B, \beta^{(k)}\right) \bar{X}_B \right]^{-1} \\
&\quad \cdot \{ \bar{X}_A^T [\bar{Y}_A - \Pi_A(\bar{X}_A, \beta)] + \bar{X}_B^T [\bar{Y}_B - \Pi_B(\bar{X}_B, \beta)] \}.
\end{aligned}$$



Local variance-covariance matrix

$$W_A(\bar{X}_A, \beta) = \begin{bmatrix} \pi(x_1, \beta)(1 - \pi(x_1, \beta)) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \pi(x_{n_A}, \beta)(1 - \pi(x_{n_A}, \beta)) \end{bmatrix},$$

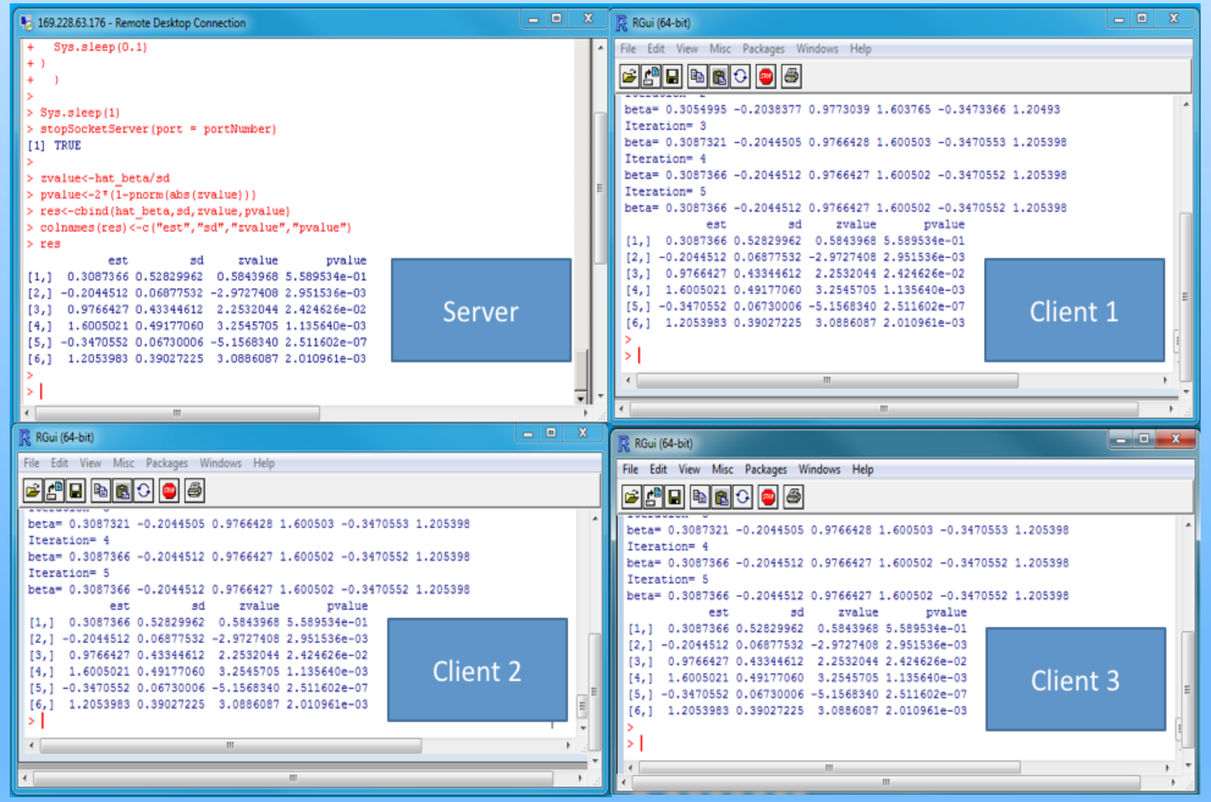
$$W_B(\bar{X}_B, \beta) = \begin{bmatrix} \pi(x_{n_A+1}, \beta)(1 - \pi(x_{n_A+1}, \beta) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \pi(x_{n_A+n_B}, \beta)(1 - \pi(x_{n_A+n_B}, \beta)) \end{bmatrix},$$

$$\Pi_A(\bar{X}_A, \beta) = \begin{bmatrix} \pi(x_1, \beta) \\ \vdots \\ \pi(x_{n_A}, \beta) \end{bmatrix}, \text{ and } \Pi_B(\bar{X}_B, \beta) = \begin{bmatrix} \pi(x_{n_A+1}, \beta) \\ \vdots \\ \pi(x_{n_A+n_B}, \beta) \end{bmatrix}.$$

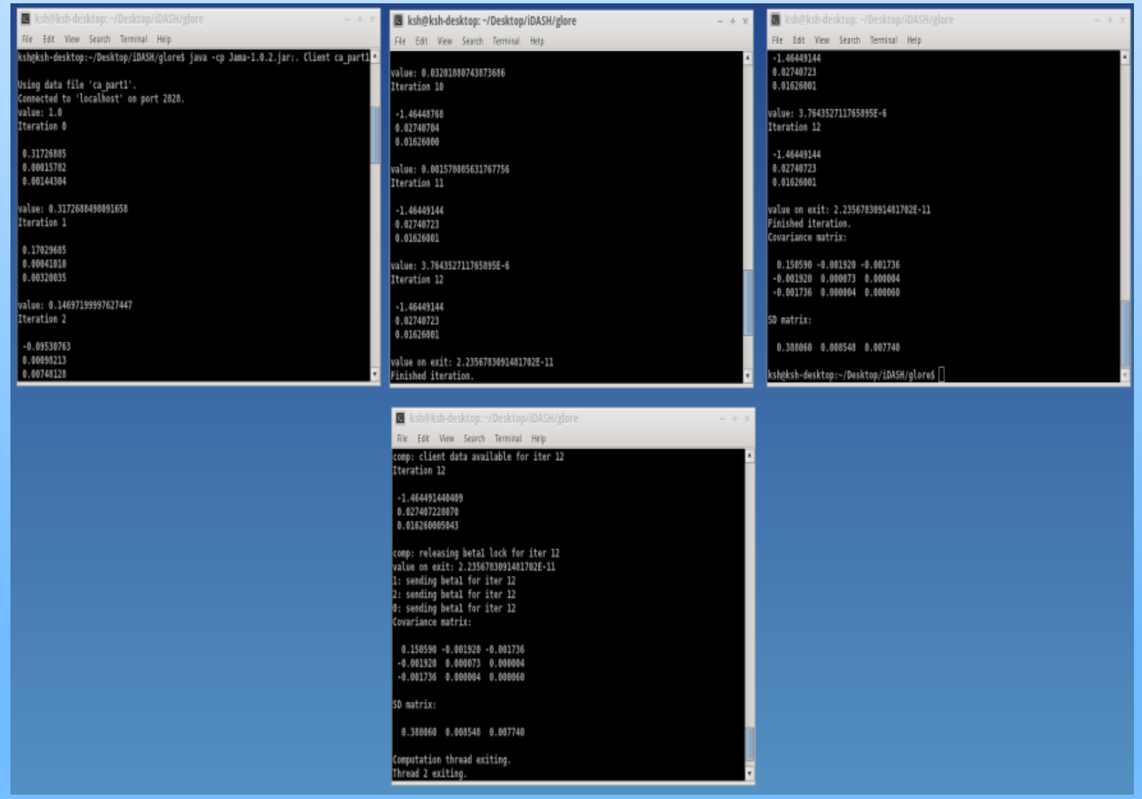
Local prediction outcomes

Backbone Implementation

- R backbone



- JAVA backbone

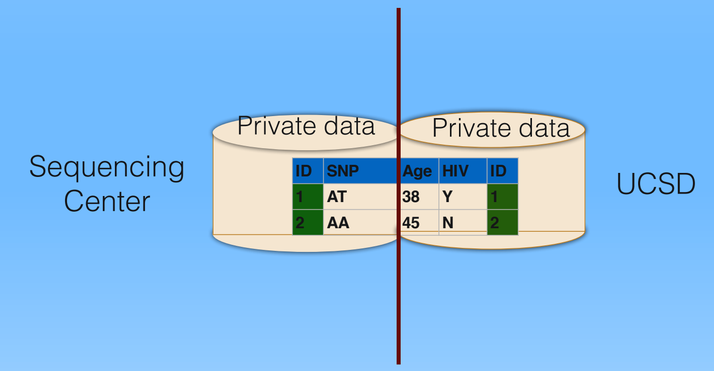


https://github.com/x1jiang/glore

Validation

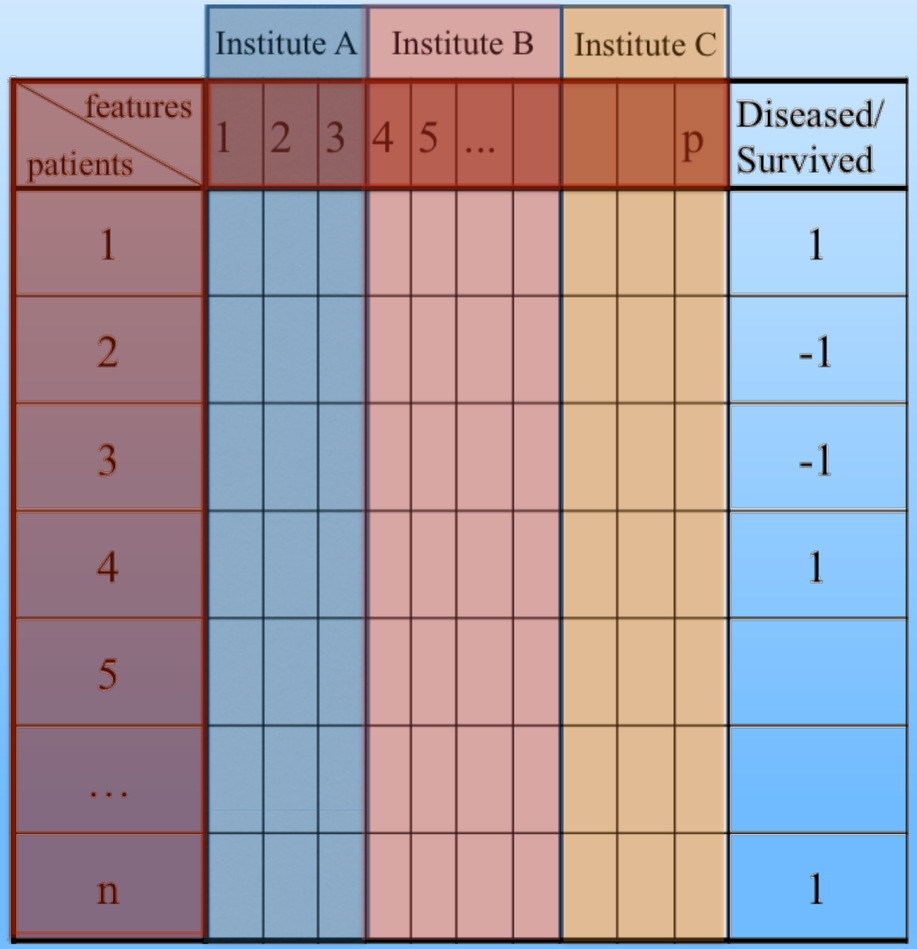
- Public datasets from dbGap
- Cincinnati data (*ImproveCareNow*! CDRN)
- No difference in the estimation from the centralized model

VERTIGO: VERTIcal Grid lOgistic regression



Jiang X, *et al*. VERTIcal Grid lOgistic regression (VERTIGO). *J Am Med Inform Assoc* 2016

VERTIcal Grid lOgistic regression (VERTIGO)



	Institute A			Institute B				Institute C			
features / patients	1	2	3	4	5	...				p	Diseased/ Survived
1											1
2											-1
3											-1
4											1
5											
…											
n											1

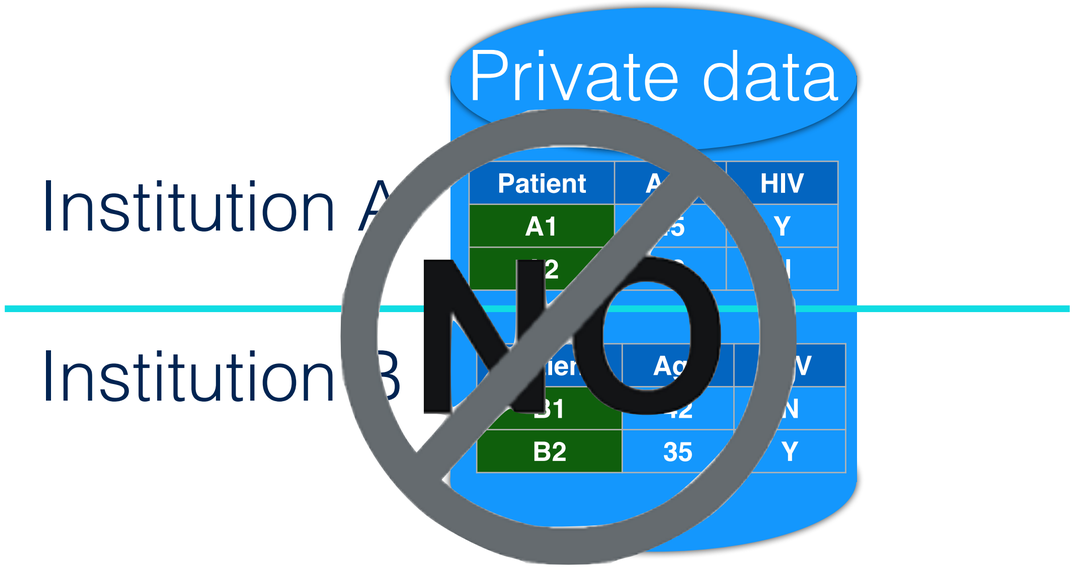
VERTIcal Grid lOgistic regression (VERTIGO)

	Institute A			Institute B				Institute C			
features / patients	1	2	3	4	5	...				p	Diseased/ Survived
1											1
2											-1
3											-1
4											1
5											
...											
n											1

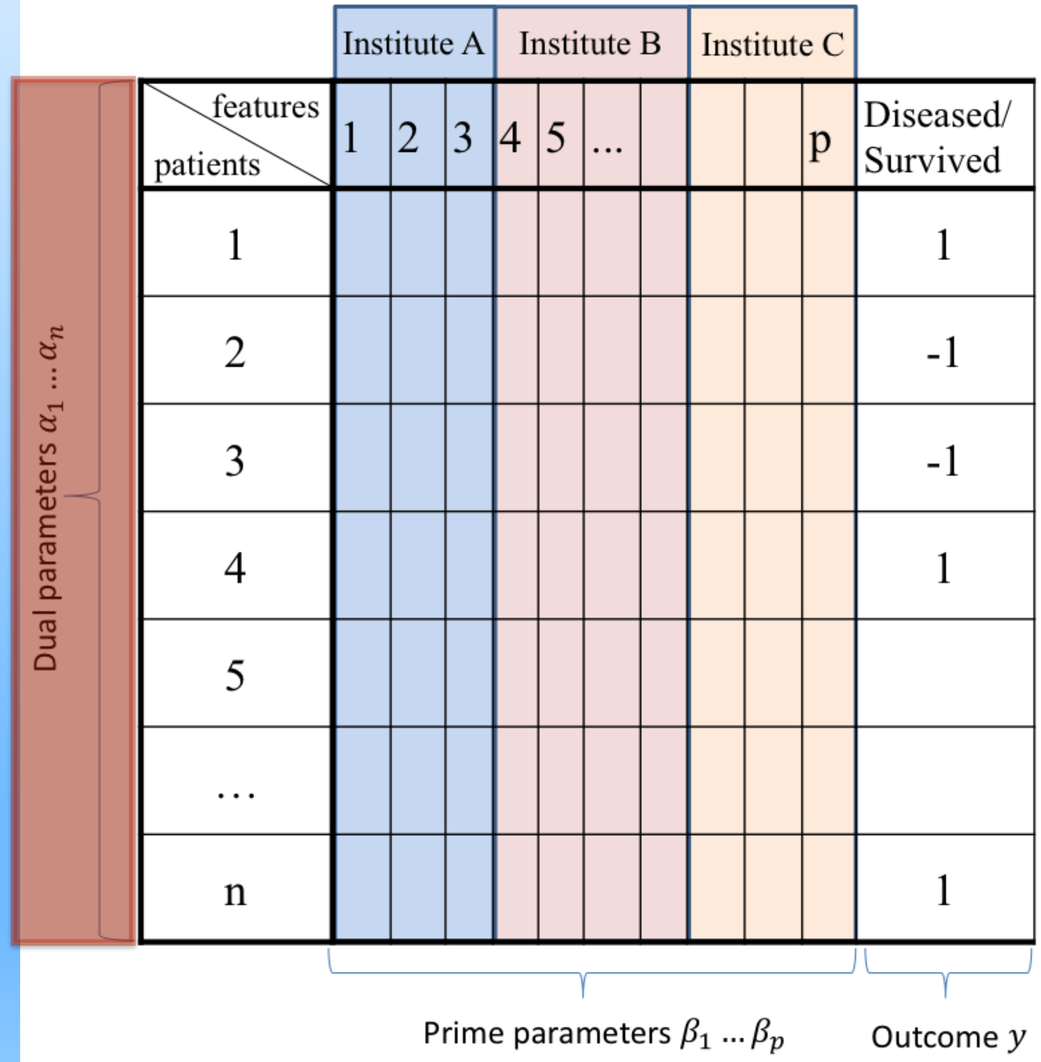
Prime parameters $\beta_1 \dots \beta_p$ Outcome y

GLORE (horizontally distributed data)

Horizontally distributed data



VERTIcal Grid lOgistic regression (VERTIGO)



Primal form GLORE (horizontally distributed data)

Log-likelihood: $l(\beta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \beta^T z_i)) - \frac{\lambda}{2} \beta^T \beta$

Prime parameter

Objective: $\max_{\beta} l(\beta)$

Dual form VERTIGO (Vertically distributed data)

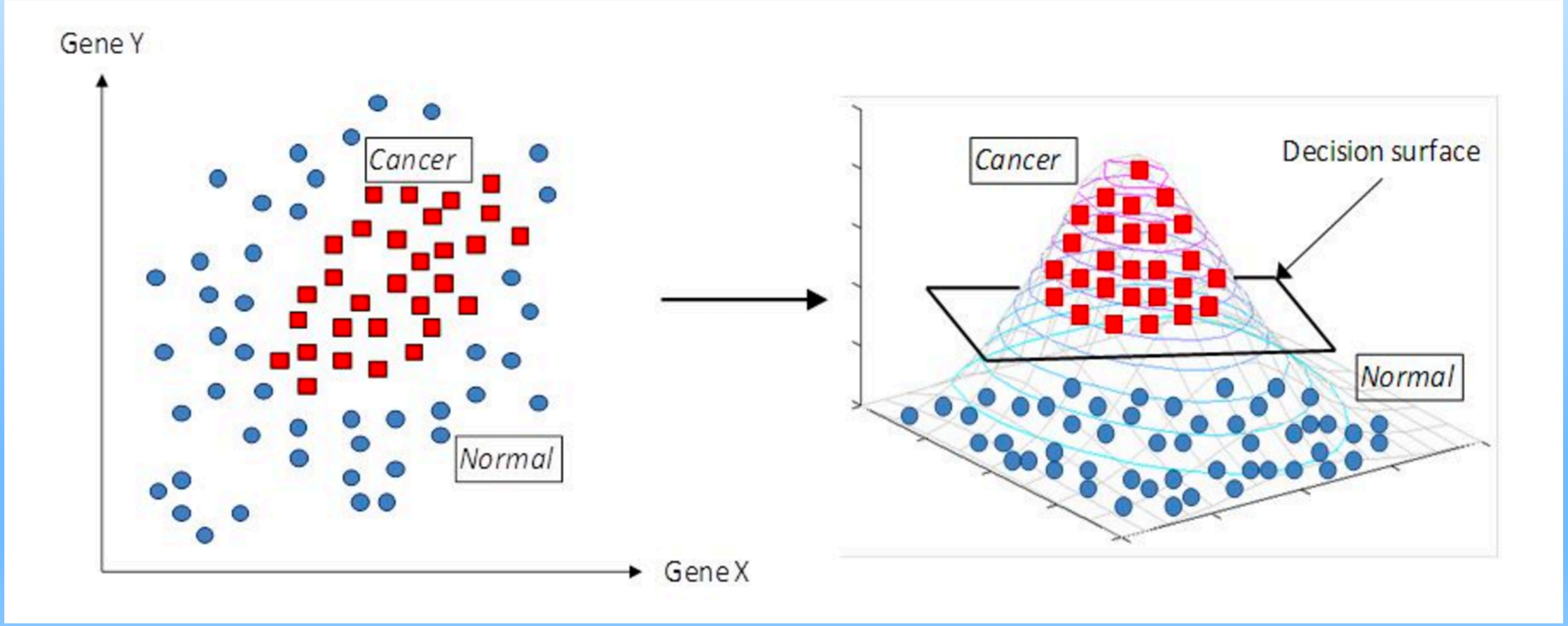
Log-likelihood: $J(\alpha) = \frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j z_i^T z_j - \sum_{i=1}^{n} H(\alpha_i)$

$H(\alpha_i) = -\alpha_i \log \alpha_i - (1 - \alpha_i) \log(1 - \alpha_i)$

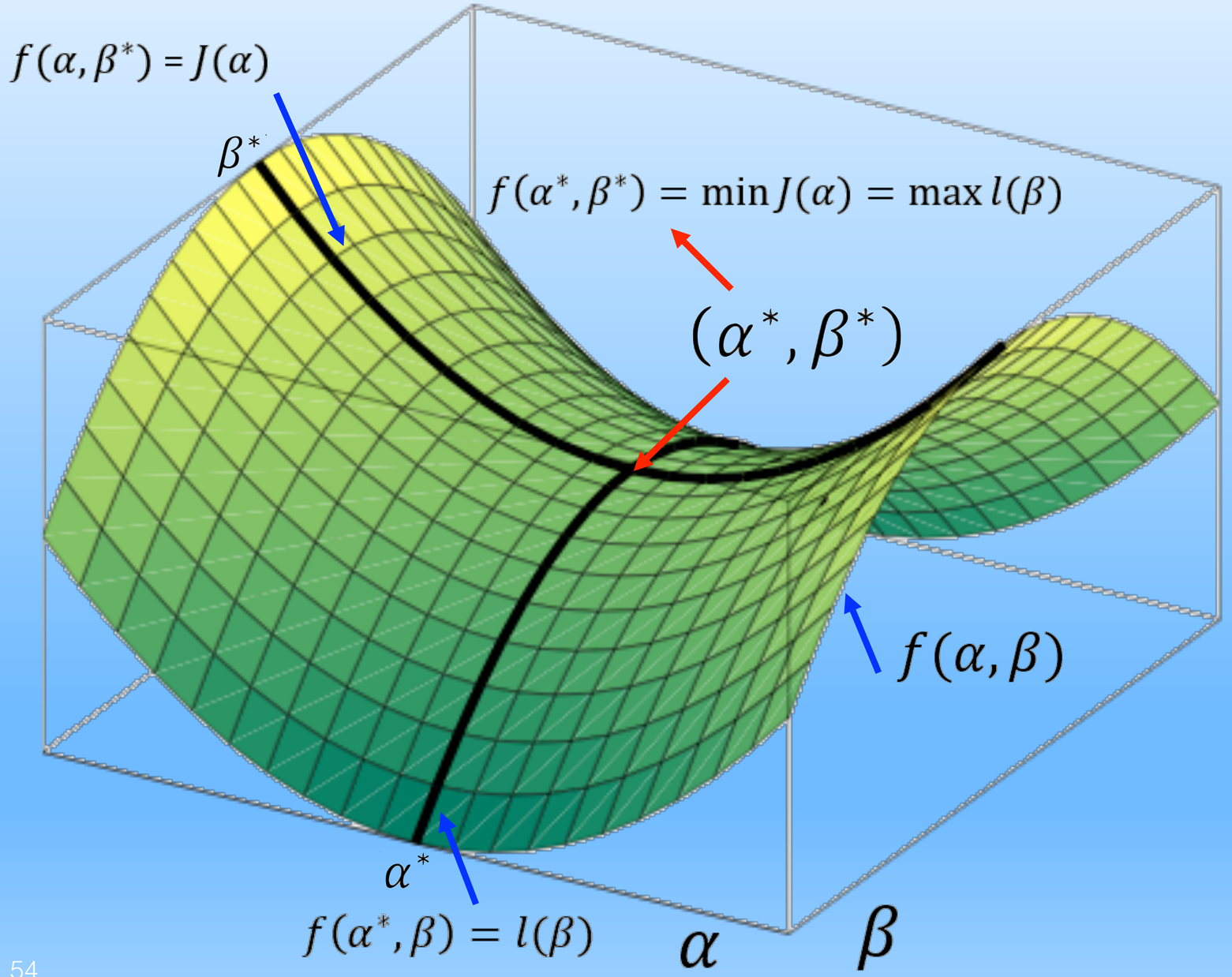
Dual parameter

Objective: $\min_{\alpha} J(\alpha)$

VERTIcal Grid lOgistic regression (VERTIGO)



VERTIcal Grid lOgistic regression (VERTIGO)



Primal form

Log-likelihood: $l(\beta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \beta^T z_i)) - \frac{\lambda}{2} \beta^T \beta$

Prime parameter

Objective: $\max_{\beta} l(\beta)$

Dual form

Log-likelihood: $J(\alpha) = \frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j z_i^T z_j - \sum_{i=1}^{n} H(\alpha_i)$

$H(\alpha_i) = -\alpha_i \log \alpha_i - (1 - \alpha_i) \log(1 - \alpha_i)$

Dual parameter

Objective: $\min_{\alpha} J(\alpha)$

Distributed NR Algorithm for VERTIGO in Dual Form

	First derivative:	**Hessian matrix:**
Centralized	$J_i'(\alpha) = \lambda^{-1} y_i \sum_j \alpha_j y_j z_j^T z_i + log\frac{\alpha_i}{1-\alpha_i}$	$J_{i,j}''(\alpha) = \lambda^{-1}\mathrm{diag}(y) Z Z^T \mathrm{diag}(y) + \mathrm{diag}(\frac{\alpha_i}{1-\alpha_i})$
Distributed	$J_i'(\alpha) = \lambda^{-1} y_i \sum_j \alpha_j y_j \left(\sum_k \left(z_j^{\text{site}_k}\right)^T \left(z_i^{\text{site}_k}\right) \right) + \log\frac{\alpha_i}{1-\alpha_i}$	$J_{i,j}''(\alpha) = \lambda^{-1}\mathrm{diag}(y) \left(\sum_k \left(Z^{\text{site}_k}\right)\left(Z^{\text{site}_k}\right)^T \right) \mathrm{diag}(y) + \mathrm{diag}(\frac{\alpha_i}{1-\alpha_i})$

Local statistics from site *k*

Local statistics from site *k*

Experiments (Average Response Time)

# of total records	VERTIGO				Primal optimization
	Iterative Hessian on CPU	Fixed Hessian on CPU	Iterative Hessian on GPU	Fixed Hessian on GPU	
2,000	2.84	7.03	**1.69**	7.01	0.72
4,000	12.51	8.68	6.55	8.22	1.71
8,000	66.38	15.30	28.57	11.05	4.10
20,000	815.6	101.7	-	-	26.5

Average computing time（seconds）for training LR models (party=2).
The red numbers are the best performers in each row

Cox proportional Hazards model

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{Z}) = \lambda_0(t)\exp(\beta_1 Z_1 + \cdots + \beta_p Z_p)$$

Time to event

Covariates

Baseline Hazard function

Coefficients

Non-parametric term

Parametric term

Breslow's partial likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp(\boldsymbol{\beta}^T \sum_{l\in\mathcal{D}_i} \mathbf{z}^l)}{[\sum_{l\in\mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)]^{d_i}}$$

Cox proportional Hazards model

Breslow's partial likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp(\boldsymbol{\beta}^T \sum_{l \in \mathcal{D}_i} \mathbf{z}^l)}{[\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)]^{d_i}}$$

Covariates of the *l*-th subject in the given set

Number of subjects has time-to-event at t=t_i

D: the number of distinct event time

A set of subjects has risk at t=t_i

A set of subjects has time-to-event at t=t_i

Id	Covariates z		Time	Censored	
1	…	i=1	3	1	
2	…		4	1	
3	…	i=2	4	1	
4	…		4	0	
5	…		5	1	$\mathcal{R}_3$
6	…	i=3	5	1	
7	..		5	0	$\mathcal{D}_3$
8	..		5	0	

$d_3=2$

Cox proportional Hazards model

Breslow's partial likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp(\boldsymbol{\beta}^T \sum_{l \in \mathcal{D}_i} \mathbf{z}^l)}{[\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)]^{d_i}}$$

Covariates of the *l*-th subject in the given set

Number of subjects has time-to-event at t=t_i

D: the number of distinct event time

A set of subjects has risk at t=t_i

A set of subjects has time-to-event at t=t_i

Log likelihood function

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{D} \{\boldsymbol{\beta}^T \sum_{l \in \mathcal{D}_i} \mathbf{z}^l - d_i \log[\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)]\}$$

$l(\beta)$ is a concave function

Newton-Raphson (NR) Algorithm

$$\beta^{(k+1)} = \beta^{(k)} - \left[\frac{\partial^2 l(\beta^{(k)})}{\partial \beta^{(k)} \partial \beta^{(k)^T}}\right]^{-1} \frac{\partial l(\beta^{(k)})}{\partial \beta^{(k)}}$$

Horizontally Distributed Survival Data

	Patient ID	Age	DrugDose	Race	Treatment	Time	Censored
Site1	1	37	7	0	1	3	1
	2	37	6	0	0	4	1
	3	39	0	0	0	4	0
	4	36	2	0	1	4	1
	…	…	…	…	…	…	…
Site k	i+1	35	12	1	1	5	0
	i+2	33	2	1	1	5	1
	i+3	29	3	0	0	5	0
	i+4	37	0	0	1	5	1
	…	…	…	…	…	…	…
Site M	n-3	35	1	0	0	6	1
	n-2	30	3	1	0	6	0
	n-1	43	0	1	1	6	0
	n	42	20	0	0	7	0

Distributed Cox Model

$$\beta^{new} = \beta^{old} - [l''(\beta^{old})]^{-1} l'(\beta^{old})$$

First derivative

Hessian Matrix

Centralized

$$l'_r(\boldsymbol{\beta}) = \sum_{i=1}^{D} \{\sum_{l \in \mathcal{D}_i} z_r^l - d_i \frac{\sum_{l \in \mathcal{R}_i} z_r^l \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}{\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}\}$$

$$l''_{r,q}(\boldsymbol{\beta}) = -\sum_{i=1}^{D} d_i \left\{\frac{\sum_{l \in \mathcal{R}_i} z_r^l z_q^l \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}{\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)} - \frac{\sum_{l \in \mathcal{R}_i} z_r^l \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}{\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)} \frac{\sum_{l \in \mathcal{R}_i} z_q^l \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}{\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}\right\}.$$

Distributed

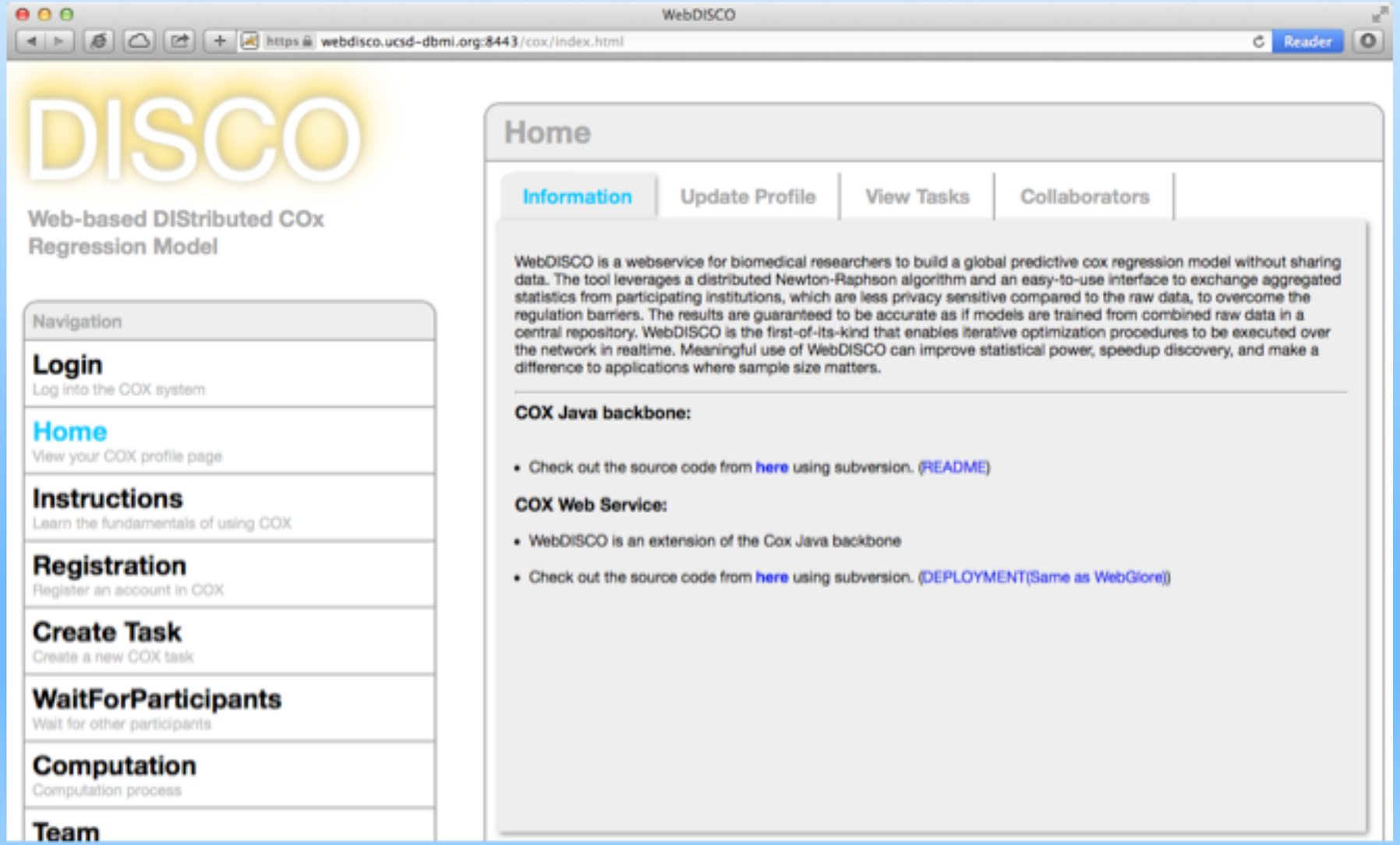
$$l'_r(\boldsymbol{\beta}) = \sum_{k=1}^{M} \sum_{i=1}^{D} \sum_{l \in \mathcal{D}_i^k} z_r^l - \sum_{i=1}^{D} \left(\sum_{k=1}^{M} |\mathcal{D}_i^k|\right) \frac{\sum_{k=1}^{M} \sum_{l \in \mathcal{R}_i^k} z_r^l \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}{\sum_{k=1}^{M} \sum_{l \in \mathcal{R}_i^k} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}$$

Local statistics from site k

$$l''_{r,q}(\boldsymbol{\beta}) = -\sum_{i=1}^{D} \left(\sum_{k=1}^{M} |\mathcal{D}_i^k|\right) \{\frac{\sum_{k=1}^{M} \sum_{l \in \mathcal{R}_i^k} z_r^l z_q^l \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}{\sum_{k=1}^{M} \sum_{l \in \mathcal{R}_i^k} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)} - \frac{\sum_{k=1}^{M} \sum_{l \in \mathcal{R}_i^k} z_r^l \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}{\sum_{k=1}^{M} \sum_{l \in \mathcal{R}_i^k} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)} \frac{\sum_{k=1}^{M} \sum_{l \in \mathcal{R}_i^k} z_q^l \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}{\sum_{k=1}^{M} \sum_{l \in \mathcal{R}_i^k} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)}\}$$

Local statistics from site k

WebDISCO: Web-based DIStributed COx Regression Model



Experiments (estimation accuracy)

UMASS Aids Research Unit (UARU) IMPACT study

Coefficients*	β learned in WebDISCO		Differences	Other feature statistics		
	1 site	2 sites		*se*	*z*	*p*
β_1	-0.035664309	-0.035664309	5.06E-13	0.0118	-3.023	0.0025
β_2	0.017800253	0.017800253	8.30E-15	0.0068	2.617	0.0089
β_3	0.053507037	0.053507037	9.21E-12	0.1704	0.314	0.75
β_4	-0.051884396	-0.051884396	6.68E-13	0.1349	-0.385	0.7
β_5	0.299188008	0.299188008	2.78E-12	0.2249	1.33	0.18
β_6	0.251226316	0.251226316	5.09E-12	0.2047	1.227	0.22
β_7	0.025211755	0.025211755	4.47E-12	0.0102	2.46	0.014
β_8	-0.455673981	-0.455673981	2.87E-12	0.1678	-2.716	0.0066
β_9	-0.275508156	-0.275508156	2.16E-12	0.1366	-2.017	0.044
β_{10}	-0.621506977	-0.621506977	8.81E-12	0.2111	-2.944	0.0032

Experiments (Average Response Time)

Average Response Time (seconds)		
UCSD and Emory	**UCSD and Duke**	**Emory and Duke**
5.50	**9.25**	**9.33**

Comparison of response time between different participant sites. WebDISCO provides robust real-time service given all average response time less than 10 seconds.

Vertical Cox Proportional Hazards

Breslow's partial likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp(\boldsymbol{\beta}^T \sum_{l \in \mathcal{D}_i} \mathbf{z}^l)}{[\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)]^{d_i}}$$

- Covariates of the *l*-th subject in the given set
- D: the number of distinct event time
- A set of subjects has risk at t=t_i
- A set of subjects has time-to-event at t=t_i
- Number of subjects has time-to-event at t=t_i

Log likelihood function

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{D} \{\boldsymbol{\beta}^T \sum_{l \in \mathcal{D}_i} \mathbf{z}^l - d_i \log[\sum_{l \in \mathcal{R}_i} \exp(\boldsymbol{\beta}^T \mathbf{z}^l)]\}$$

$f(z)$ is not decomposable and cannot be obtained based on intermediary statistics

$g(\beta)$ is decomposable over K institutions

$$\min_{\eta_{lk}, \beta_k} \sum_{i=1}^{T} \left[d_i \cdot \log \sum_{l \in R_i} \exp\left(\sum_{k=1}^{K} \eta_{lk} \right) \right] - \sum_{i=1}^{T} \sum_{n \in D_i} \sum_{k=1}^{K} \beta_k^T x_{nk}$$

$$\text{s.t. } \beta_k^T x_{nk} - \eta_{nk} = 0, \quad \forall n = 1, \cdots, N, \quad \forall k = 1, \cdots, K$$

Constraints for β and η

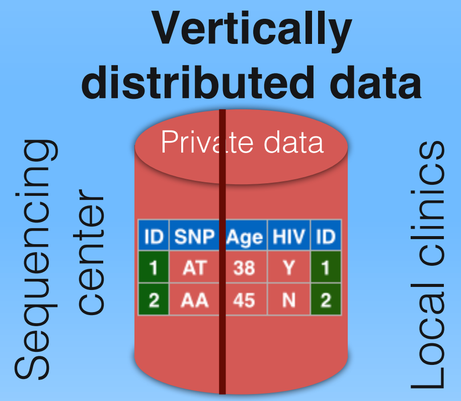
Alternating Direction Methods of Multipliers (ADMM)

ADMM problem form (with f, g convex)

$$\begin{array}{ll} \text{minimize} & f(x) + g(z) \\ \text{subject to} & Ax + Bz = c \end{array}$$

– two sets of variables, with separable objective

DIADEM: DIstributed Cox model based on Alternating Direction mEthod of Multipliers



Background

- Cox proportional hazard model (centralized)
  - Hazard function $\lambda(t|x)$ provides hazard for the patient with covariate x at time t

$$\lambda(t|x) = \lambda_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x})$$

Baseline hazard function over time t

Model parameters

  - Maximizing likelihood function for model parameter estimation

$$l(\beta) = \prod_{t=1}^{T} \frac{\exp\left(\sum_{n\in D_t}\beta^T\boldsymbol{x}_n\right)}{\left[\sum_{j\in R_t}\exp\left(\beta^T\boldsymbol{x}_j\right)\right]^{d_t}}$$

Index set of samples with observed events

Index set of samples at risk for the event

Number of events at time t

Background

- Distributed optimization
  - M Covariates distributed across K institutions
  - Training set: N samples $(x_{n1}, x_{n2}, \cdots, x_{nK})$, $n = 1, \cdots, N$, where x_{nk} is the M_k-tuple covariate held by the k-th institution
  - Parameter estimation by maximizing log-likelihood

$$l(\beta) = \sum_{t=1}^{T} \sum_{n \in D_t} \sum_{k=1}^{K} \beta_k^T x_{nk} - \sum_{t=1}^{T} \left[d_t \cdot \log \sum_{j \in R_t} \exp\left(\sum_{k=1}^{K} \beta_k^T x_{nk} \right) \right]$$

Number of institutions

M_k covariates held by the k-th institution

Model parameter for the k-th institution

Motivation

- **Dual optimization**
  - Legendre transform

$$l(\beta, \alpha) = \sum_i \left[\alpha_i \beta^T x_i - H(\alpha_i) - \alpha_i \, log \sum_{\substack{t_j \geq t_i, \\ j \neq i}} exp(\beta^T x_i) \right] - \frac{\lambda}{2} \beta^T \beta$$

Dual parameters

$$H(\alpha_i) = -\alpha_i \log \alpha_i - (1 - \alpha_i) \log(1 - \alpha_i)$$

- Cannot obtain analytic solution β to maximize $l(\beta, \alpha)$ as VERTIGO
- Approximate solution leads to loss in accuracy

Motivation

- Distributed optimization based on ADMM
  - Robust methods for vertically partitioned data
    - Log-likelihood function $l(\beta)$ is not decomposable over multiple institutions
    - Introduce auxiliary variables z for equivalent problem with decomposable $g(\beta)$
  - Support federated data analysis
    - Do not share patient-level data
    - No communications between institutions

$f(z)$ is not decomposable and solved in the server based on aggregated statistics

$g(\beta)$ is decomposable over K institutions

$$\min_{z_{jk},\beta_k} \sum_{t=1}^{T} \left[d_t \cdot \log \sum_{j \in R_t} \exp\left(\sum_{k=1}^{K} z_{jk} \right) \right] - \sum_{t=1}^{T} \sum_{n \in D_t} \sum_{k=1}^{K} \beta_k^T x_{nk}$$

$$\text{s.t. } \beta_k^T x_{nk} - z_{nk} = 0, \quad \forall n = 1, \cdots, N, \quad \forall k = 1, \cdots, K$$

Constraints for β and z

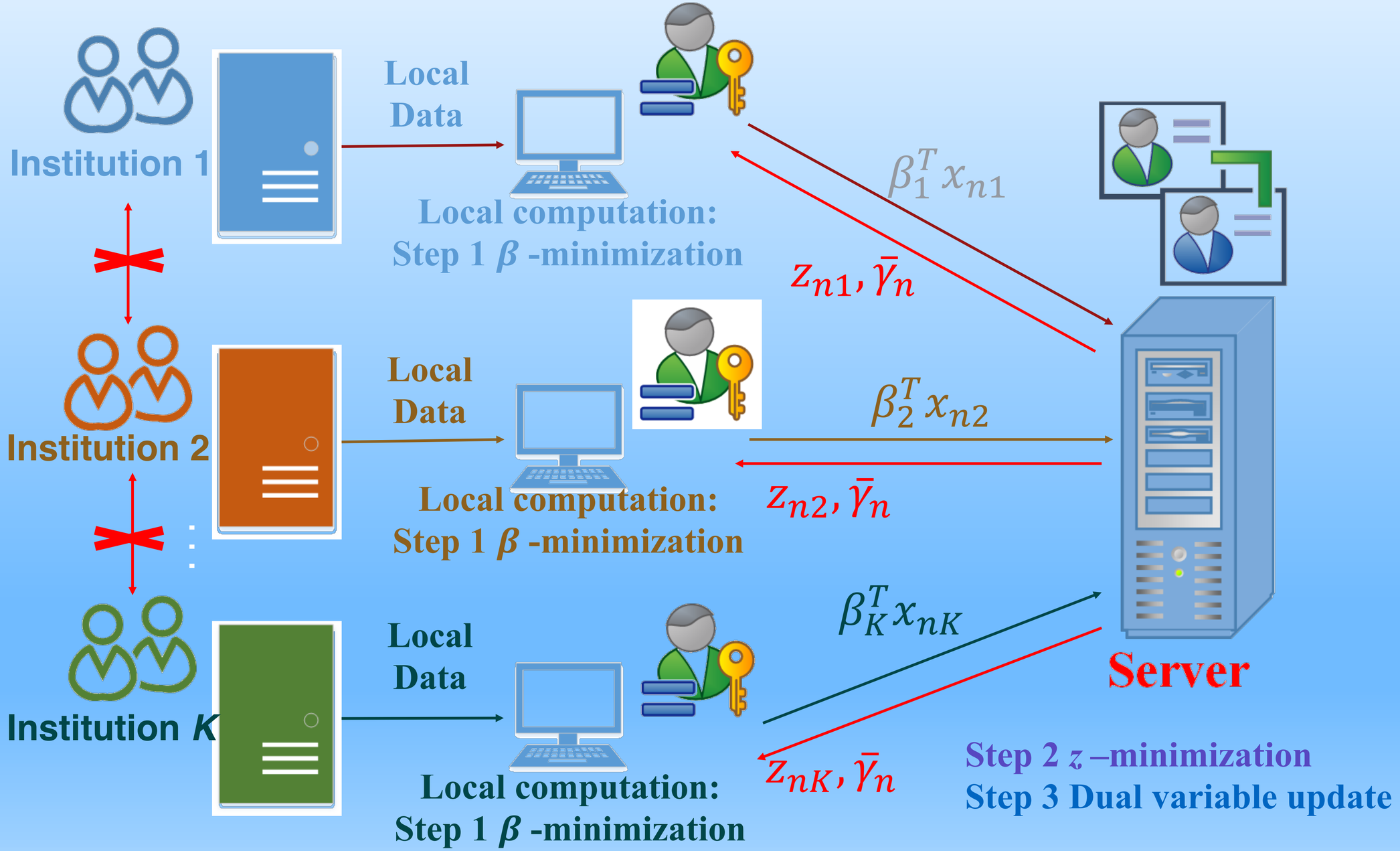
Formulation

- Distributed optimization based on ADMM
  - Introduce dual variable γ_{nk} and regularization parameter ρ for unconstrained optimization problem

$$
\begin{aligned}
&L(\beta, \gamma, z) \\
&= \sum_{t=1}^{T} \left[d_t \cdot \log \sum_{j \in R_t} \exp\left(\sum_{k=1}^{K} z_{jk} \right) \right] - \sum_{t=1}^{T} \sum_{n \in D_t} \sum_{k=1}^{K} \beta_k^T x_{nk} \\
&+ \sum_{n=1}^{N} \sum_{k=1}^{K} \left[\gamma_{nk} \left(\beta_k^T x_{nk} - z_{nk} \right) + \frac{\rho}{2} \left(\beta_k^T x_{nk} - z_{nk} \right)^2 \right]
\end{aligned}
$$

  - Iteratively obtain $\{\beta^*, \gamma^*, z^*\} = \underset{\beta, \gamma, z}{\mathbf{argmin}} L(\beta, \gamma, z)$ with three steps
    - Step 1: β-minimization with fixed z and γ separately at the K institutions
    - Step 2: z-minimization with fixed γ and updated β at the server
    - Step 3: update dual variable γ with updated β and z at the server

Formulation



Formulation

- Step 1: β-minimization
  - Fixing z and γ for β-minimization. At the p-th iteration, the k-th institution minimizes

$$L_\beta(\beta_k) = -\sum_{t=1}^{T}\sum_{n\in D_t}\beta_k^T x_{nk} + \sum_{n=1}^{N}\left[\frac{\rho}{2}\left(\beta_k^T x_{nk}\right)^2 + \left(\gamma_{nk}^{(p-1)} - \rho z_{nk}^{(p-1)}\right)\beta_k^T x_{nk}\right]$$

  - Separately solve β_k over local data for the k-th institution

$$\beta_k^{(p)} = \left[\rho\sum_{n=1}^{N} x_{nk}x_{nk}^T\right]^{-1}\left[\sum_{n=1}^{N}\left(\rho z_{nk}^{(p-1)} - \gamma_{nk}^{(p-1)}\right)x_{nk} + \sum_{t=1}^{T}\sum_{n\in D_t} x_{nk}\right]$$

  - Send aggregated results to the server

$$\sigma_{nk}^{(p)} = \left[\beta_k^{(p)}\right]^T x_{nk}, \qquad \bar{\sigma}_n^{(p)} = \frac{1}{K}\sum_{k=1}^{K}\sigma_{nk}^{(p)}$$

Formulation

- Step 2: z-minimization
  - Fixing β and γ for z-minimization. At the p-th iteration, the server minimizes

$$L_z(z) = \sum_{t=1}^{T}\left[d_t \cdot \log \sum_{j \in R_t} \exp\left(\sum_{k=1}^{K} z_{jk}\right)\right] + \sum_{n=1}^{N}\sum_{k=1}^{K}\left[\frac{\rho}{2} z_{nk}^2 - (\rho\sigma_{nk} + \gamma_{nk}) z_{nk}\right]$$

  - Since $L_z(z)$ cannot be analytically solved, we introduce equivalent optimization using surrogate variables $\bar{z}_n$ for $L_z(\bar{z})$.

$$\min \sum_{t=1}^{T}\left[d_t \cdot \log \sum_{j \in R_t} \exp\left(K\bar{z}_j\right)\right] + \sum_{n=1}^{N}\sum_{k=1}^{K}\left[\frac{\rho}{2} z_{nk}^2 - (\rho\sigma_{nk} + \gamma_{nk}) z_{nk}\right]$$

$$\text{s.t. } \bar{z}_n = \frac{1}{K}\sum_{k=1}^{K} z_{nk}, \quad n = 1, \cdots, N$$

Surrogate variables with constraints

Formulation

- Step 2: z-minimization (cont'd)
  - z_{nk} can be represented by $\bar{z}_n$

$$z_{nk}^{(p)} = \bar{z}_n^{(p)} + \sigma_{nk}^{(p)} + \frac{\gamma_{nk}^{(p-1)}}{\rho} - \frac{1}{K}\sum_{k=1}^{K}\left[\sigma_{nk}^{(p)} + \frac{\gamma_{nk}^{(p-1)}}{\rho}\right]$$

  - The equivalent unconstrained optimization problem $L_{\bar{z}}(\bar{z})$ is

$$\min_{\bar{z}} L_{\bar{z}}(\bar{z}) = \sum_{t=1}^{T}\left[d_t \cdot \log \sum_{j \in R_t} \exp\left(K\bar{z}_j\right)\right] + K\rho \sum_{n=1}^{N}\left[\frac{\bar{z}_n^2}{2} - \left(\bar{\sigma}_n^{(p)} + \frac{\bar{\gamma}_n^{(p-1)}}{\rho}\right)\bar{z}_n\right]$$

  - $L_{\bar{z}}(\bar{z})$ can be solved using the Newton-Raphson method

Formulation

- Step 3: Update dual variable γ
  - Update dual variable γ_{nk} corresponding to the n-th record held by the k-th institution

$$\gamma_{nk}^{(p)} = \frac{1}{K}\sum_{k=1}^{K}\gamma_{nk}^{(p-1)} + \rho\left(\bar{\sigma}_n^{(p)} - \bar{z}_n^{(p)}\right)$$

  - All the K institutions share the same dual variables $\bar{\gamma} = (\bar{\gamma}_1, \cdots, \bar{\gamma}_N)$, where $\bar{\gamma}_n = \sum_{k=1}^{K}\gamma_{nk}/K$

$$\gamma_{nk}^{(p)} = \bar{\gamma}_n^{(p)} = \bar{\gamma}_n^{(p-1)} + \rho\left(\bar{\sigma}_n^{(p)} - \bar{z}_n^{(p)}\right)$$

  - Send z_{nk}, $\bar{\gamma}_n$ and $\bar{\sigma}_n$ back to the k-th institution

Explanatory Example

1st Iteration

$MSE = 1.18$

Institution 1

	Age	BDI
x_{11}	37	11
x_{21}	36	19
x_{31}	42	23
x_{41}	24	24
x_{51}	37	11
x_{61}	35	17
x_{71}	38	15

Step 1: β-minimization

$\beta_1^{(1)}$	(0.0202, 0.0158)
$\sigma_{11}^{(1)}$	0.9207
$\sigma_{21}^{(1)}$	1.0270
$\sigma_{31}^{(1)}$	1.2113
$\sigma_{41}^{(1)}$	0.8638
$\sigma_{51}^{(1)}$	0.9207
$\sigma_{61}^{(1)}$	0.9752
$\sigma_{71}^{(1)}$	1.0041

n	1	2	3	4	5	6	7
t	4	4	7	7	7	10	10
observed	1	1	1	1	1	1	1

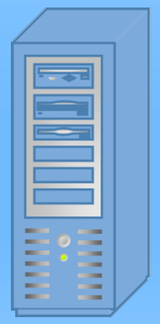
t	4	7	10
D_t	{1,2}	{3,4,5}	{6,7}
R_t	{1,2}	{3,4,5}	{6,7}
d_t	2	3	2

$\beta_2^{(1)}$	(0.8571, 0.5714)
$\sigma_{12}^{(1)}$	0.8571
$\sigma_{22}^{(1)}$	0
$\sigma_{32}^{(1)}$	0.8571
$\sigma_{42}^{(1)}$	0.8571
$\sigma_{52}^{(1)}$	0.5714
$\sigma_{62}^{(1)}$	1.4286
$\sigma_{72}^{(1)}$	0

Institution 2

	Heroin use	Cocaine use
x_{12}	1	0
x_{22}	0	0
x_{32}	1	0
x_{42}	1	0
x_{52}	0	1
x_{62}	1	1
x_{72}	0	0

Step 1: β-minimization



Server

$\sigma_{n1}^{(1)}$

$\bar{\sigma}_1^{(1)}$	0.9207
$\bar{\sigma}_2^{(1)}$	1.0270
$\bar{\sigma}_3^{(1)}$	1.2113
$\bar{\sigma}_4^{(1)}$	0.8638
$\bar{\sigma}_5^{(1)}$	0.9207
$\bar{\sigma}_6^{(1)}$	0.9752
$\bar{\sigma}_7^{(1)}$	1.0041

$z_{n1}^{(1)}, \bar{\gamma}_n^{(1)}$

$\bar{z}_1^{(1)}$	0.2853	$z_{11}^{(1)}$	0.3171	$z_{12}^{(1)}$	0.2535
$\bar{z}_2^{(1)}$	0.0982	$z_{21}^{(1)}$	0.6117	$z_{22}^{(1)}$	-0.4152
$\bar{z}_3^{(1)}$	-0.1783	$z_{31}^{(1)}$	-0.0012	$z_{32}^{(1)}$	0.3554
$\bar{z}_4^{(1)}$	-0.2309	$z_{41}^{(1)}$	-0.2276	$z_{42}^{(1)}$	-0.2343
$\bar{z}_5^{(1)}$	-0.2678	$z_{51}^{(1)}$	-0.0931	$z_{52}^{(1)}$	-0.4424
$\bar{z}_6^{(1)}$	-0.1310	$z_{61}^{(1)}$	-0.3577	$z_{62}^{(1)}$	0.0957
$\bar{z}_7^{(1)}$	-0.8284	$z_{71}^{(1)}$	-0.3263	$z_{72}^{(1)}$	-1.3304

$z_{n2}^{(1)}, \bar{\gamma}_n^{(1)}$

$\sigma_{n2}^{(1)}$

Step 2: z-minimization

$\bar{\gamma}_1^{(1)}$	0.6036
$\bar{\gamma}_2^{(1)}$	0.4152
$\bar{\gamma}_3^{(1)}$	1.2125
$\bar{\gamma}_4^{(1)}$	1.0914
$\bar{\gamma}_5^{(1)}$	1.0139
$\bar{\gamma}_6^{(1)}$	1.3329
$\bar{\gamma}_7^{(1)}$	1.3304

Step 3: update γ

Explanatory Example

Institution 1

	Age	BDI
x_{11}	37	11
x_{21}	36	19
x_{31}	42	23
x_{41}	24	24
x_{51}	37	11
x_{61}	35	17
x_{71}	38	15

Step 1: β-minimization

$z_{n1}^{(1)}$, $\bar{\gamma}_n^{(1)}$

$\beta_1^{(2)}$	(0.0081, −0.0174)
$\sigma_{11}^{(2)}$	0.1070
$\sigma_{21}^{(2)}$	-0.0403
$\sigma_{31}^{(2)}$	-0.0616
$\sigma_{41}^{(2)}$	-0.2243
$\sigma_{51}^{(2)}$	0.1070
$\sigma_{61}^{(2)}$	-0.136
$\sigma_{71}^{(2)}$	0.0455

2nd Iteration $MSE = 0.74$

n	1	2	3	4	5	6	7
t	4	4	7	7	7	10	10
observe	1	1	1	1	1	1	1

t	4	7	10
D_t	{1,2}	{3,4,5}	{6,7}
R_t	{1,2}	{3,4,5}	{6,7}
d_t	2	3	2

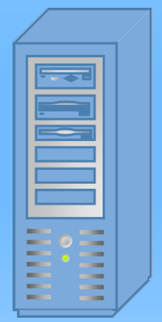
$z_{n2}^{(1)}$, $\bar{\gamma}_n^{(1)}$

$\beta_2^{(2)}$	(−0.0383, −0.3276)
$\sigma_{12}^{(2)}$	-0.0383
$\sigma_{22}^{(2)}$	0
$\sigma_{32}^{(2)}$	-0.0383
$\sigma_{42}^{(2)}$	-0.0383
$\sigma_{52}^{(2)}$	-0.3276
$\sigma_{62}^{(2)}$	-0.3659
$\sigma_{72}^{(2)}$	0

Institution 2

	Heroin use	Cocaine use
x_{12}	1	0
x_{22}	0	0
x_{32}	1	0
x_{42}	1	0
x_{52}	0	1
x_{62}	1	1
x_{72}	0	0

Step 1: β-minimization



Server

$\sigma_{n1}^{(2)}$

$\bar{\sigma}_1^{(2)}$	0.0344
$\bar{\sigma}_2^{(2)}$	-0.0202
$\bar{\sigma}_3^{(2)}$	-0.0500
$\bar{\sigma}_4^{(2)}$	-0.1313
$\bar{\sigma}_5^{(2)}$	-0.1103
$\bar{\sigma}_6^{(2)}$	-0.1897
$\bar{\sigma}_7^{(2)}$	0.0227

$z_{n1}^{(2)}, \bar{\gamma}_n^{(2)}$

$\bar{z}_n^{(1)}$, $\bar{\gamma}_n^{(1)}$

$\bar{z}_1^{(2)}$	0.1869	$z_{11}^{(2)}$	0.2596	$z_{12}^{(2)}$	0.1142
$\bar{z}_2^{(2)}$	0.0512	$z_{21}^{(2)}$	0.0310	$z_{22}^{(2)}$	0.0713
$\bar{z}_3^{(2)}$	-0.0502	$z_{31}^{(2)}$	-0.0619	$z_{32}^{(2)}$	-0.0386
$\bar{z}_4^{(2)}$	-0.1119	$z_{41}^{(2)}$	-0.2049	$z_{42}^{(2)}$	-0.0190
$\bar{z}_5^{(2)}$	-0.1301	$z_{51}^{(2)}$	0.0872	$z_{52}^{(2)}$	-0.3474
$\bar{z}_6^{(2)}$	-0.0559	$z_{61}^{(2)}$	0.1202	$z_{62}^{(2)}$	-0.2321
$\bar{z}_7^{(2)}$	-0.3342	$z_{71}^{(2)}$	-0.3114	$z_{72}^{(2)}$	-0.3569

Step 2: z-minimization

$z_{n2}^{(2)}, \bar{\gamma}_n^{(2)}$

$\sigma_{n2}^{(2)}$

$\bar{\gamma}_1^{(2)}$	0.4511
$\bar{\gamma}_2^{(2)}$	0.3439
$\bar{\gamma}_3^{(2)}$	1.2128
$\bar{\gamma}_4^{(2)}$	1.0720
$\bar{\gamma}_5^{(2)}$	1.0337
$\bar{\gamma}_6^{(2)}$	1.1991
$\bar{\gamma}_7^{(2)}$	1.6873

Step 3: update γ

Explanatory Example

2000th Iteration $MSE = 1.51\times10^{-17}$

Institution 1

	Age	BDI
x_{11}	37	11
x_{21}	36	19
x_{31}	42	23
x_{41}	24	24
x_{51}	37	11
x_{61}	35	17
x_{71}	38	15

Step 1: β-minimization

$\boldsymbol{\beta_1^*}$ $(-0.0104, -0.0902)$

n	1	2	3	4	5	6	7
t	4	4	7	7	7	10	10
observed	1	1	1	1	1	1	1

t	4	7	10
D_t	{1,2}	{3,4,5}	{6,7}
R_t	{1,2}	{3,4,5}	{6,7}
d_t	2	3	2

$\boldsymbol{\beta_2^*}$ $(0.8026, -0.5069$

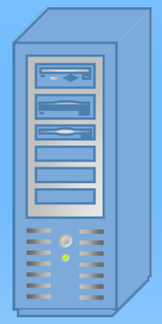
Institution 2

	Heroin use	Cocaine use
x_{12}	1	0
x_{22}	0	0
x_{32}	1	0
x_{42}	1	0
x_{52}	0	1
x_{62}	1	1
x_{72}	0	0

Step 1: β-minimization

$\boldsymbol{z_{n1}^{(1999)}}$ $\boldsymbol{\bar{\gamma}_n^{(1999)}}$

$\boldsymbol{z_{n2}^{(1999)}}$ $\boldsymbol{\bar{\gamma}_n^{(1999)}}$



Server

$z_{11}^{(1999)}$	-1.3751	$\bar{\gamma}_1^{(1999)}$	0.7058	$z_{12}^{(1999)}$	0.8026
$z_{21}^{(1999)}$	-2.0859	$\bar{\gamma}_2^{(1999)}$	0.1554	$z_{22}^{(1999)}$	0
$z_{31}^{(1999)}$	-2.5087	$\bar{\gamma}_3^{(1999)}$	1.0251	$z_{32}^{(1999)}$	0.8026
$z_{41}^{(1999)}$	-2.4124	$\bar{\gamma}_4^{(1999)}$	1.1288	$z_{42}^{(1999)}$	0.8026
$z_{51}^{(1999)}$	-1.3751	$\bar{\gamma}_5^{(1999)}$	0.8597	$z_{52}^{(1999)}$	-0.5069
$z_{61}^{(1999)}$	-1.8953	$\bar{\gamma}_6^{(1999)}$	1.1403	$z_{62}^{(1999)}$	0.2957
$z_{71}^{(1999)}$	-1.7461	$\bar{\gamma}_7^{(1999)}$	1.9849	$z_{72}^{(1999)}$	0

Convergence Analysis

- The distributed Cox model is linearly convergent to the optimum under the ADMM framework
  - Primal residual $\left\|X^T\beta^{(p)} - \bar{z}^{(p)}\right\|_2^2 \to 0,\ p \to \infty$
  - Dual residual $\left\|\gamma^{(p)} - \gamma^*\right\|_2^2 \to 0,\ p \to \infty$
  - Objective function $f(\bar{z}) + g(\beta) \to f(\bar{z}^*) + g(\beta^*),\ \ p \to \infty$
  - Convergence rate is related to ρ

Performance Analysis

- Computational complexity
  - Server side: $O(N^3)$ for N records
  - Client side: $O((N+M)M_k^2)$ for N records with M_k covariates
- Communication cost
  - Each institution send $O(N)$ for each iteration
  - Each institution receives $O(2N)$ for each iteration

Implementation

- Developed in JULIA, a high-level, high performance programming language. (Matlab style, C speed)
- Support asynchronized computation, no single point of failure
- Plan to support webservice in the future

Experimental results

- Model parameter estimation
  - Surveillance Epidemiology and End Results (SEER) Dataset

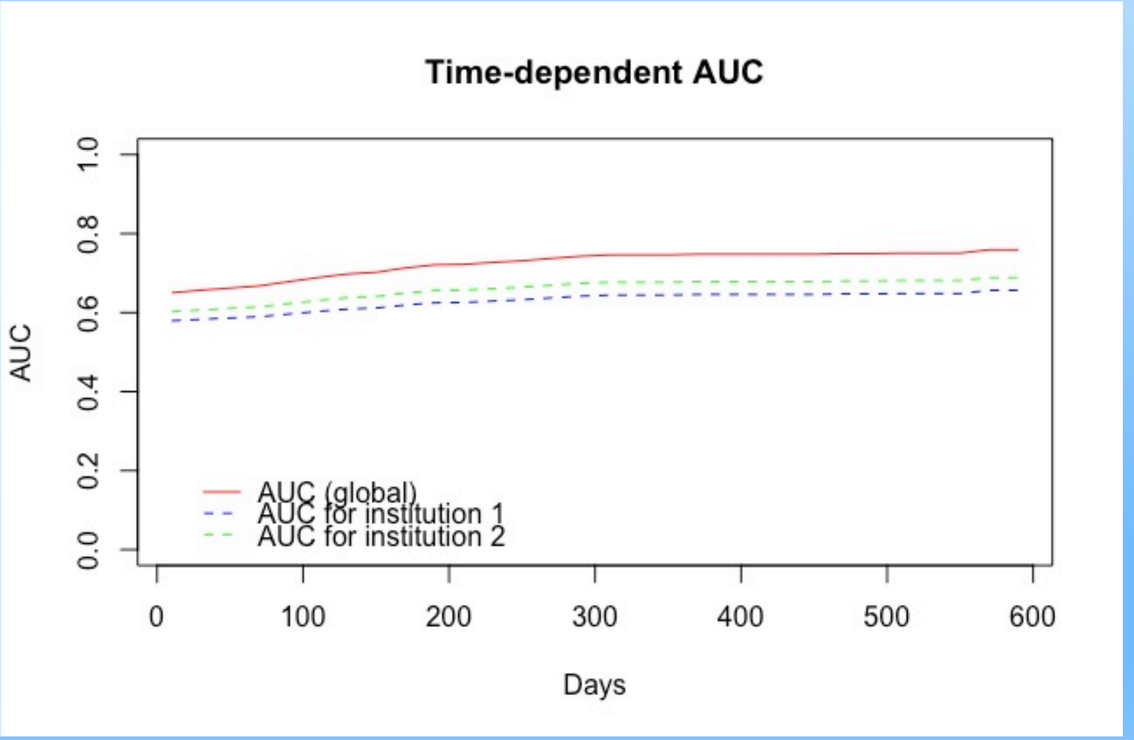
# of institutions	# of iterations	Estimation error (MSE)	Time cost (second)	
			Server	Client
2	2000	1.46×10^{-12}	6017.85	0.23
3	2000	3.18×10^{-12}	6946.83	0.22

  - UMASS Aids Research Unit (UARU) IMPACT Study Dataset

# of institutions	# of iterations	Estimation error (MSE)	Time cost (second)	
			Server	Client
2	1500	1.39×10^{-23}	45.22	0.08
3	1500	1.50×10^{-23}	45.35	0.10

Experimental results

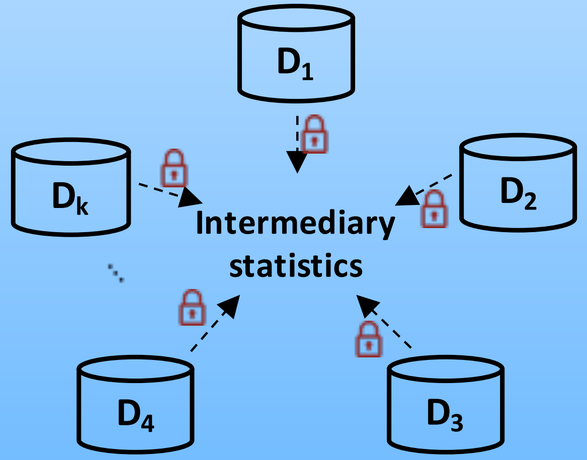
- Time-dependent AUC Score
  - Global AUC varies in [0.64,0.75]
  - Global AUC outperforms those obtained from local data held by the two institutions



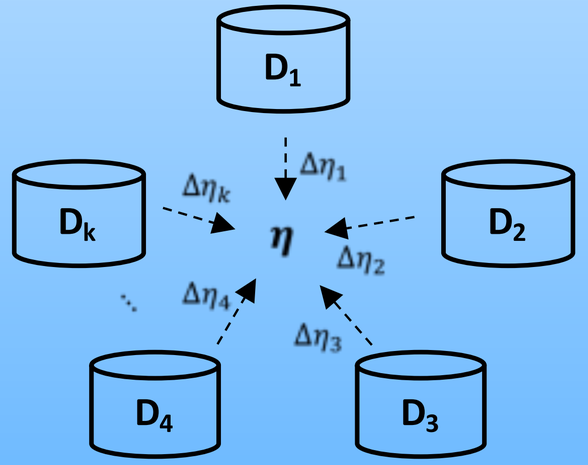
Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Stat Med* 2006; 25: 3474–86.

Additional Protection to Statistics

- Using secure primitive to safeguard the communication



- Masking the pattern before transmitting



- Using reference (public) data to improve the privacy protection

Jiang X, et al. Differential-Private Data Publishing Through Component Analysis. Trans Data Priv 2013
Jiang X. *et al*. Privacy Preserving RBF Kernel Support Vector Machine. *Biomed Res Int,* 2014;
Ji Z, Jiang X, *et al*. Differentially private distributed logistic regression using private and public data. *BMC Med Genomics BioMed Central*, 2014

Summary

- Predictive modeling forms a critical building block in biomedical informatics

Questions?

Thanks for listening