UNIVERSITY OF ALBERTA
SCHOOL OF PUBLIC HEALTH

Department of Public Health Sciences

Prediction Error Estimation With A Changing Covariates Distribution

Yan Yuan, PhD

June 4
2012 SSC, Guelph, ON

Prediction

Natural human desire; an important goal in many applied sciences and business sections.

Huge growth in predictive models and algorithms from Computer Science, Bioinformatics and Statistics.

An fundamental methodological issue is the quantification of models' prediction performance.

Measure of Prediction Performance

Prediction loss and prediction error (Korn and Simon, 1990; Graf et al., 1999; Hothorn et al., 2006; Gneiting et al., 2007; Lawless and Yuan, 2010**)**

Concordance measure (Pencina and D’Agostino, 2004**)**

ROC curve (Heagerty and Zheng, 2005, Cook 2007, Mann et al. 2010, Uno et al. 2011**)**

Statistical Prediction

- Predict random variable *Y* given covariates *Z* for some population
  - Denote true conditional distribution function of *Y* given *Z*

$$F_T(y \mid z) = Pr(Y \leq y \mid z)$$

- Training data

$$D = \{(y_i, z_i), i = 1, \ldots, n\}$$

- Modeling procedure (*M*)
  - The final model is given by $F_{\hat{\theta}}(y|z)$

- Point predictor

e.g. $\hat{Y}(Z) = G_{\hat{\theta}}(Z) = E_{\hat{\theta}}(Y|Z) = \int y dF_{\hat{\theta}}(y|z)$

Prediction Error

$$\pi(M;F_T,H_Z)=E_{Y,D,Z}[L(Y,G_{\hat{\theta}}(Z))]$$

- The prediction error depends on the model M
- The expectation is taken with respect to the *Y* values in the independent new data, the training data *D*, and **the distribution of covariates *Z***.

Estimation of Prediction Error

- Test data based estimator

$$\hat{\pi}^{test} = \frac{1}{m} \sum_{j=1}^{m} L(y_j, G_{\hat{\theta}}(z_j))$$

- Model based estimator

$$\hat{\pi}^{m} = \frac{1}{n} \sum_{i=1}^{n} E_{D^*} \int L(y, G_{\hat{\theta}}(z_i)) dF_{\hat{\theta}}(y|z_i).$$

- Apparent loss + adjustment term

$$\hat{\pi}^{A} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, G_{\hat{\theta}}(z_i)) + \hat{\Omega}$$

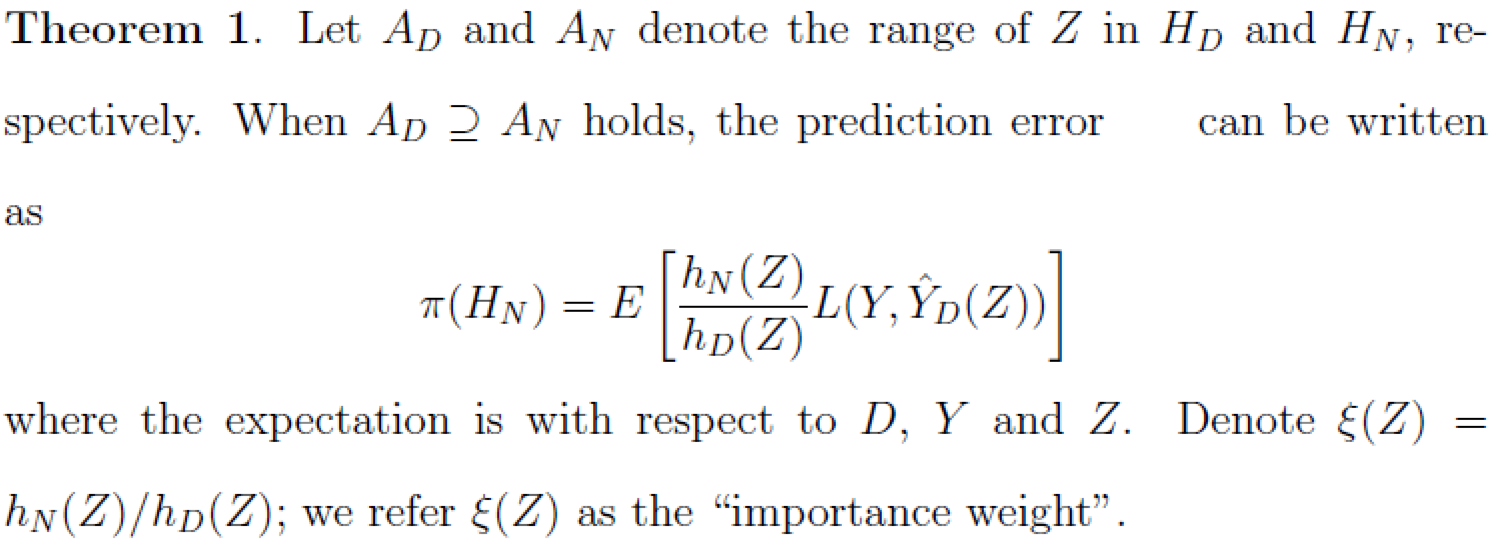
- Cross-validation loss

$$\hat{\pi}^{cv} = \frac{1}{n} \sum_{v=1}^{V} \sum_{i \in S_v} L(y_i, G_{\hat{\theta}(-v)}(z_i))$$

Changing Covariates Distribution

- **Assumption**
  - Covariate Z is typically assumed to be uniformly distributed on the values $(z_1,\ldots, z_n)$ observed in D.
  - The distribution of the covariates is the same in the new/future data as in the observed “training data” D that were used to derive the predictive model.
- **Problem**
  - In practice, the distributions of covariates in the training data ($H_D(Z)$) and new data ($H_N(Z)$) are often different.

Our Idea

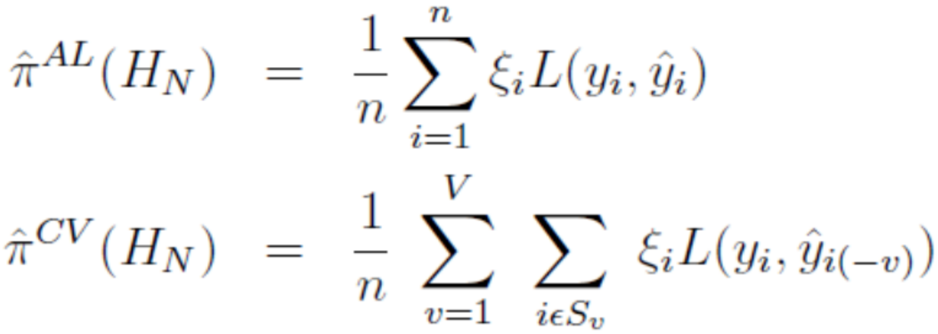


Theorem 1. Let A_D and A_N denote the range of Z in H_D and H_N, respectively. When $A_D \supseteq A_N$ holds, the prediction error can be written as

$$\pi(H_N) = E\left[\frac{h_N(Z)}{h_D(Z)} L(Y, \hat{Y}_D(Z))\right]$$

where the expectation is with respect to D, Y and Z. Denote $\xi(Z) = h_N(Z)/h_D(Z)$; we refer $\xi(Z)$ as the "importance weight".

Weighted estimator



$$\hat{\pi}^{AL}(H_N) = \frac{1}{n}\sum_{i=1}^{n} \xi_i L(y_i, \hat{y}_i)$$

$$\hat{\pi}^{CV}(H_N) = \frac{1}{n}\sum_{v=1}^{V} \sum_{i \epsilon S_v} \xi_i L(y_i, \hat{y}_{i(-v)})$$

Simple Case

Z_i is discrete;

Support of H_N is contained in H_D

$$\hat{\xi}_i = \frac{\hat{h}_N(z_i)}{\hat{h}_D(z_i)} = \frac{n}{m}\frac{c'(a_k)}{c(a_k)}$$

where

$$c(a_k) = \sum_{i=1}^{n} I(z_i = a_k) \text{ and } c'(a_k) = \sum_{j=1}^{m} I(z'_j = a_k)$$

A Simple Example

Z	0	1	Sample size
Training data	20	10	*n*=30
New data	10	15	*m*=25

$$\xi_{(z_i=0)} = \frac{n}{m}\frac{c'(0)}{c(0)} = \frac{30}{25}\frac{10}{20} = \frac{3}{5}$$

$$\xi_{(z_i=1)} = \frac{n}{m}\frac{c'(1)}{c(1)} = \frac{30}{25}\frac{15}{10} = \frac{9}{5}$$

Approaches

1. **Matching weights**

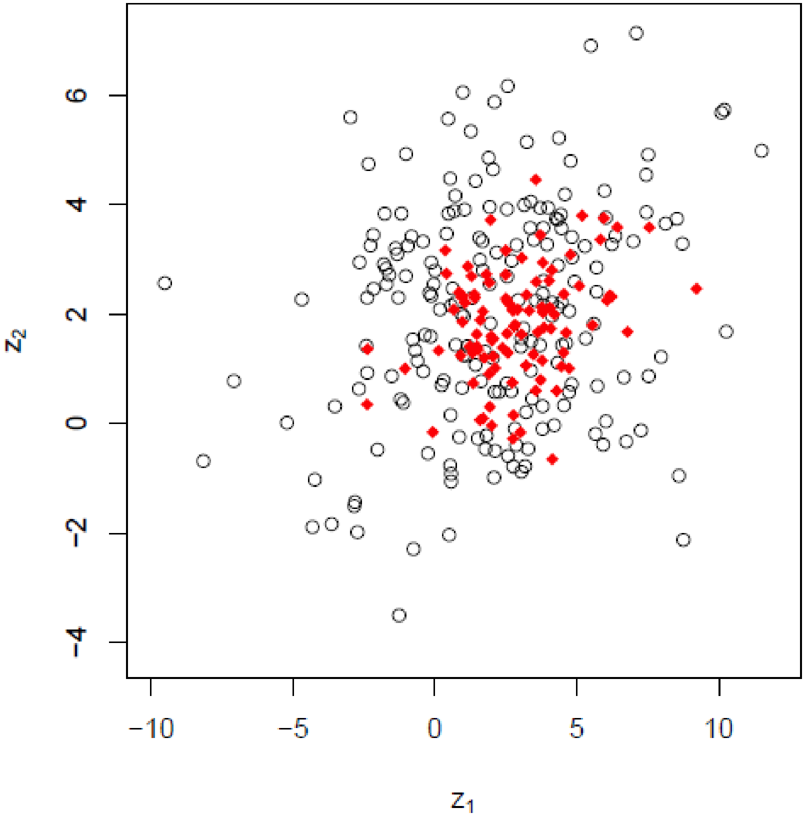
   Nearest neighbor matching (Mahalanobis distance)

   Genetic matching (Generalized Mahalanobis distance)

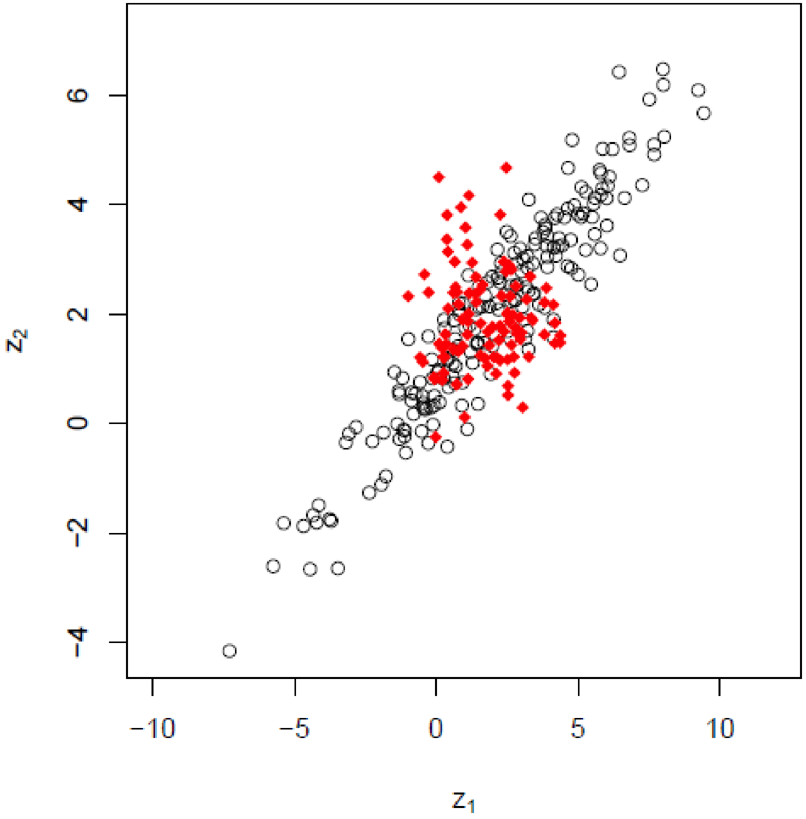
2. **Kernel weights**

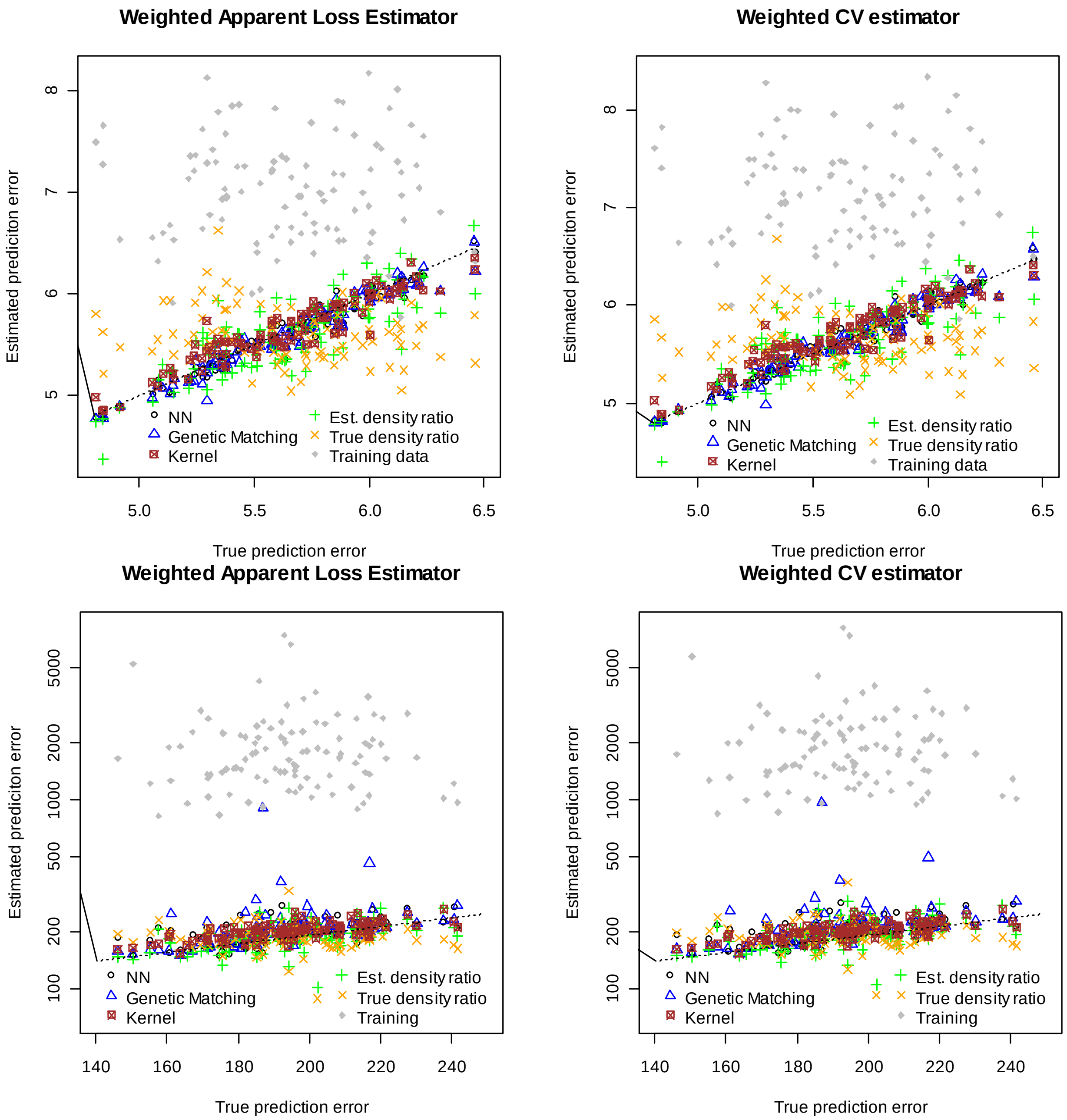
   Base the estimation of $h_N(z)$ on kernel smoothing

Distribution of Covariates Simulation Setting 1



Distribution of Covariates Simulation Setting 2

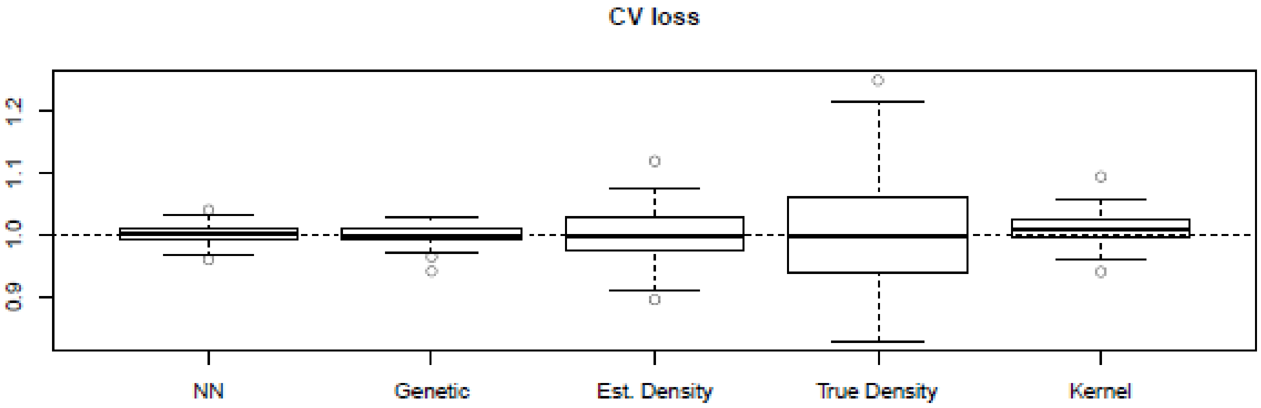




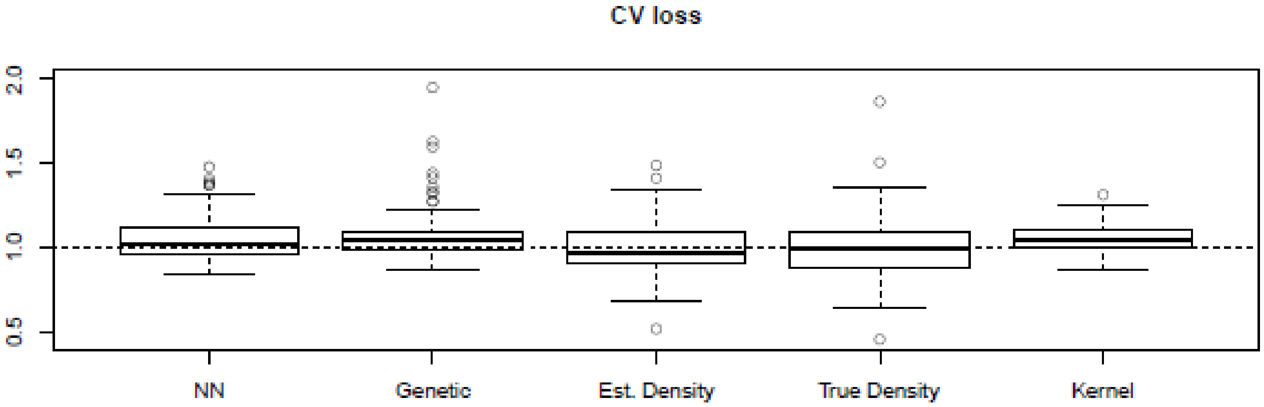
Weighted Apparent Loss Estimator
Weighted CV estimator
Estimated prediciton error
True prediction error
5
6
7
8
5.0
5.5
6.0
6.5
NN
Genetic Matching
Kernel
Est. density ratio
True density ratio
Training data
Weighted Apparent Loss Estimator
Weighted CV estimator
100
200
500
1000
2000
5000
140
160
180
200
220
240
Training

Ratio

Setting 1



Setting 2



Future work

- Distance measure
- Categorical variable
- Extension to higher dimension
- Real world example

Thank you

UNIVERSITY OF ALBERTA
SCHOOL OF PUBLIC HEALTH