An Introduction to Utility-Maximizing Credit Scoring

Jiun-Hua Su

Institute of Economics
Academia Sinica

August 23rd, 2023

Outline:

- Credit Scoring and Its Common Statistical Methods
- Maximum Utility Estimation
  1. Complexity Penalized Maximum Utility Estimation
  2. Nonparametric Maximum Utility Estimation
  3. Profit-Maximizing Credit Scoring

Prerequisites:

- An introductory course in probability and statistics (Required)
- A course in mathematical statistics (Required/Suggested)
- An introductory course in statistical learning (Suggested/Optional)
- An introductory course in economics/finance (Optional)
- A course in mathematical analysis (Not Required)
- An introductory course in computer science (Not Required)

Literature on utility-maximizing binary prediction:

- Lieli and White (2010): "The Construction of Empirical Credit Scoring Rules Based on Maximization Principles"
- Elliott and Lieli (2013): "Predicting Binary Outcomes"
- Su (2021): "Model Selection in Utility-Maximizing Binary Prediction"
- Su (2023): "Utility-Maximizing Binary Prediction via the Nearest Neighbor Method and Its Application to Credit Scoring"

Part 0 Credit Scoring

Anderson (2007): "The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation"

Thomas et al. (2017): "Credit Scoring and Its Applications"

Credit Scoring

- In the current context 'credit' simply means, 'buy now, pay later', as indicated by Anderson (2007, p. 3).
- 板谷敏彥 (2022, p. 27):
  - 美索不達米亞的西帕爾出土之西元前一八二三年的泥板，上頭紀錄了一份借貸契約 *(*借據*)*，內容如下*:*

    「伊利 · 卡達里之子普茲魯姆，從沙瑪什 *(*太陽神*)* 處收到三十八又十六分之一謝克爾。普茲魯姆將按照沙瑪什神規定的利率支付利息。普茲魯姆將於收成之時，償還白銀本金和利息。」

    此處的借貸利率是依據神殿的規範，因此應該是百分之二十。還款時間爲「收成之時」。因爲小麥一年只收成一次，由此可知，借貸契約的期限爲一年以內，借貸目的是爲了耕種小麥。
  - 「如果商人違反規定，每古爾穀物收取超過六十卡利息、每謝克爾白銀收取超過六分之一謝克爾又六賽拉的利息，商人將喪失其所提供之物。」

    這就是《漢摩拉比法典》規範的利率上限。此依法條的存在供我們想像，當時美索不達米亞也存在高利貸，而且釀成了社會問題。

- Anderson (2007, p. 6):

  'What is credit scoring?' Simply stated, it is the use of statistical models to transform relevant data into numerical measures that guide credit decision.

A brief history of credit scoring:

- The arrival of credit cards in the late 1960s made the banks and other credit card issuers realize the usefulness of credit scoring. (Thomas et al., 2017, p. 4)
- For the most part, credit scoring was an American preserve prior to 1980, while most other countries relied upon traditional relationship lending, and risk-assessment procedures. (Anderson, 2007, p. 41)
- In the 1980s, the success of credit scoring in credit cards meant that bankds started using scoring for their other products, such as mortgages and personal loans, while in the last few years scoring has been used for home loans and small business loans. (Thomas et al., 2017, p. 5)
- However, the greatest impact on credit scoring since 2000 is the advent of the Basel Accords. (Thomas et al., 2017, p. 5)
  1. A bank should hold an amount of capital (regulatory capital);
  2. Under the internal ratings-based approach, a bank can provide its own estimates of PD, LGD, and EAD.

Two types of credit scoring:

1. Application scoring: to provide guidance on an 'accept/reject' decision on granting credit to a new applicant;
2. Behavioral scoring: to facilitate account management of existing customer, for example, adjustment of credit limit.

We will focus on the application scoring in what follows.

Statistical Methods in Credit Scoring

We will quickly go over the following methods in credit scoring:

- Linear Discriminant Analysis;
- Logistic Regression;
- Support Vector Machine.

Further methods can be found in Anderson (2007) and Thomas et al. (2017).

Fisher's Linear Discriminant Analysis (LDA)

- Setting: $Y \in \{0, 1\}$, $X \in \mathbb{R}^k$
- Assumption: $X|Y = y \sim \mathcal{N}(\mu_y, \Sigma)$
- Decision:

$$
\begin{aligned}
&\text{given } X = x, \text{ predict } Y = 1 \\
\Leftrightarrow\ & \mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y = 0|X = x) \\
\Leftrightarrow\ & 0 < \log\left\{\frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)}\right\} \\
&= \log\left\{\frac{\mathbb{P}(X = x|Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(X = x|Y = 0)\mathbb{P}(Y = 0)}\right\} \qquad \text{Bayes theorem} \\
&= x^\top \underbrace{\Sigma^{-1}(\mu_1 - \mu_0)}_{\equiv\beta} + \underbrace{\frac{1}{2}\left[\mu_0^\top \Sigma^{-1}\mu_0 - \mu_1^\top \Sigma^{-1}\mu_1\right] + \log\left\{\frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y = 0)}\right\}}_{\equiv\alpha}
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\mathbb{P}(Y=1|X=x)\\
&= \frac{\mathbb{P}(X=x|Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=x|Y=1)\mathbb{P}(Y=1)+\mathbb{P}(X=x|Y=0)\mathbb{P}(Y=0)}\\
&= \frac{1}{1+\left[\frac{\mathbb{P}(X=x|Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=x|Y=0)\mathbb{P}(Y=0)}\right]^{-1}}\\
&= \frac{1}{1+\exp\left\{-\log\left\{\frac{\mathbb{P}(X=x|Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=x|Y=0)\mathbb{P}(Y=0)}\right\}\right\}}.
\end{aligned}
$$

Thus, under the assumption that $X|Y=y\sim\mathcal{N}(\mu_y,\Sigma)$, we have

$$
\begin{aligned}
\mathbb{P}(Y=1|X=x) &= \frac{1}{1+\exp\{-(x^\top\beta+\alpha)\}}\\
&= \Lambda(x^\top\beta+\alpha),
\end{aligned}
$$

where $\Lambda(u)=(1+\exp\{-u\})^{-1}$ is the logistic function.

Logistic Regression

- Setting: $Y \in \{0,1\}$, $X \in \mathbb{R}^k$
- Assumption: $\mathbb{P}(Y = 1|X = x) = \Lambda(x^\top \beta + \alpha)$ for some $\beta \in \mathbb{R}^k$ and $\alpha \in \mathbb{R}$
- Motivation:
  1. Generalized linear model (with linear log odds)

     Specifying the link function Λ^{-1}, we have $\mathbb{E}[Y|X = x] = \Lambda(x^\top \beta + \alpha)$.
  2. Latent regression

     Under the assumptions that
     - i. $Y^* = X^\top \beta + \alpha - \varepsilon$, $\mathbb{P}(\varepsilon \le u|X = x) = \Lambda(u)$, and
     - ii. $Y = \begin{cases} 1, & \text{if } Y^* \ge 0; \\ 0, & \text{otherwise,} \end{cases}$

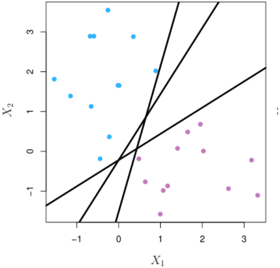
     we have $\mathbb{P}(Y = 1|X = x) = \Lambda(x^\top \beta + \alpha)$.
- Decision:

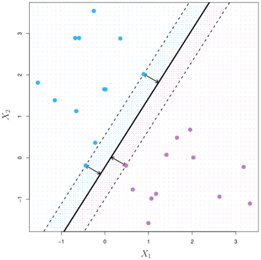
$$\text{given } X = x, \text{ predict } Y = 1 \Leftrightarrow \mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y = 0|X = x)$$

$$\Leftrightarrow 0 < \log\left\{\frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)}\right\} = x^\top \beta + \alpha$$

Vapnik's Support Vector Machine (SVM)

Figures 9.2 and 9.3 of James et al. (2021)





- Setting: $y \in \{-1, 1\}$, $x \in \mathbb{R}^k$
- We say the observations $\{(y_i, x_i)\}_{i=1}^n$ are linearly separable if there is a separating hyperplane with slope β and intercept α such that for all $i = 1, \ldots, n$,

$$\begin{cases} \beta^\top x_i + \alpha > 0, & \text{if } y_i = 1, \\ \beta^\top x_i + \alpha < 0, & \text{if } y_i = -1; \end{cases}$$

equivalently, $y_i(\beta^\top x_i + \alpha) > 0$.

- The hard support vector machine aims to find the separating hyperplane with the largest margin; specifically,

$$(\beta, \alpha) \in \arg\max_{(b,a):\|b\|=1} \min\{|b^\top x_i + a| : i = 1, 2, \ldots, n\}$$
$$\text{subject to } y_i(b^\top x_i + a) > 0 \text{ for all } i = 1, 2, \ldots, n.$$

Lemma 1

The following algorithm yields a solution to the hard-SVM.

1. *Input:* $\mathscr{D}_n = \{(y_i, x_i)\}_{i=1}^n$
2. *Solve:*

$$(\tilde{\beta}, \tilde{\alpha}) \in \arg\min_{(b,a)} \|b\|^2$$
$$\textit{subject to } y_i(b^\top x_i + a) \geq 1 \textit{ for all } i = 1, 2, \ldots, n.$$

3. *Output:* $\beta = \frac{\tilde{\beta}}{\|\tilde{\beta}\|}$ *and* $\alpha = \frac{\tilde{\alpha}}{\|\tilde{\beta}\|}$.

- If the observations $\{(y_i, x_i)\}_{i=1}^n$ are linearly non-separable, then we consider the soft support vector machine

$$(\beta, \alpha) \in \arg\min_{(b,a)} \frac{c}{n} \sum_{i=1}^{n} \xi_i + \|b\|^2$$

subject to $y_i(x_i^\top b + a) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for all $i = 1, \ldots, n$.

- Recap: the hard support vector machine can be solved by

$$(\tilde{\beta}, \tilde{\alpha}) \in \arg\min_{(b,a)} \|b\|^2$$

subject to $y_i(b^\top x_i + a) \geq 1$ for all $i = 1, 2, \ldots, n$.

- If the observations $\{(y_i, x_i)\}_{i=1}^n$ are linearly non-separable, then we consider the soft support vector machine

$$(\beta, \alpha) \in \arg\min_{(b,a)} \frac{c}{n} \sum_{i=1}^{n} \xi_i + \|b\|^2$$
$$\text{subject to } y_i(x_i^\top b + a) \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \text{ for all } i = 1, \ldots, n.$$

- The soft support vector machine can be equivalently recast as

$$(\beta, \alpha) \in \arg\min_{(b,a)} \frac{1}{n} \sum_{i=1}^{n} \max\left\{0, 1 - y_i(x_i^\top b + a)\right\} + \lambda\|b\|^2$$

  if we appropriately select λ.

- Setting: $Y \in \{-1, 1\}$, $X \in \mathbb{R}^k$
- Decision:

$$\text{given } X = x, \text{ predict } Y = 1 \Leftrightarrow x^\top \beta + \alpha > 0$$
$$\Leftrightarrow \text{sign}(\Lambda(x^\top \beta + \alpha) - 1/2) > 0$$

Stefan Banach:

> *A mathematician is a person who can find analogies between theorems; a better mathematician is one who can see analogies between proofs and the best mathematician can notice analogies between theories. One can imagine that the ultimate mathematician is one who can see analogies between analogies.*

Why 1/2?

- Setting: $Y \in \{0, 1\}$, $X \in \mathbb{R}^k$
  (Without loss of generality, 0 can be replaced with -1 here.)
- Two outcomes and $\mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y = 0|X = x)$
- The implicit rationale is to find a classifier such that the misclassification rate is minimized.
- The Bayes decision rule

$$g^*(x) = \begin{cases} 1, & \text{if } \mathbb{P}(Y = 1|X = x) > 1/2, \\ 0, & \text{otherwise,} \end{cases}$$

  has the property that for all $g : \mathbb{R}^d \to \{0, 1\}$,

$$\mathbb{P}(g^*(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y).$$

  See the discussion in Devroye et al. (1996) and Hastie et al. (2009).
- We make a decision in pursuit of statistical accuracy.
- Should we do so in credit scoring?

Part 1 Maximum Utility Estimation

Elliott and Lieli (2013): “Predicting Binary Outcomes”

Binary Decision and Binary Prediction

Making a binary decision based on an uncertain binary outcome is common in modern economic activities.

Granger and Machina (2006) suggest that decision making based on the prediction should be driven by a decision maker's preference.

- Lieli and White (2010) study how a utility-maximizing lender's approval or rejection depends on his or her binary prediction about a borrower's default.

  Two scenarios

  - The lender rejects the loan & the borrower complies fully with the terms of the contract 🙁
  - The lender approves the loan & the borrower defaults 😭

- Further examples can be found in Elliott and Timmermann (2016).

Decision-Based Binary Prediction

Elliott and Lieli (2013):

A decision maker chooses a binary decision $a \in \{-1, 1\}$ to maximize his or her expected utility

$$\max_{a \in \{-1,1\}} \mathbb{E}\left[U(a, Y, X) | X = x\right], \quad (1)$$

where $X = x$ is a d-dimensional vector of observed covariates, and $Y \in \{-1, 1\}$ is not observable at the time of the decision.

Application: Profit-Maximizing Credit Scoring

- $U = \pi$, lender's profit function

	Good $(Y = 1)$	Bad $(Y = -1)$
Approve (A)	$\pi_{A,1}(x) > 0$	$\pi_{A,-1}(x) < 0$
Reject (R)	$\pi_{R,1}(x) = 0$	$\pi_{R,-1}(x) = 0$

- X: loan characteristics (e.g. interest rate and duration)

Assumptions

$$\max_{a \in \{-1,1\}} \mathbb{E}\left[U(a,Y,X)|X=x\right]$$

Assumptions imposed by Elliott and Lieli:

A1 The conditional probability $\mathbb{P}(Y=1 \mid X=x)$ does not depend on the binary decision a.

A2 For all x in the support $\mathcal{X} \subseteq \mathbb{R}^d$ of X, $U(1,1,x) > U(-1,1,x)$ and $U(-1,-1,x) > U(1,-1,x)$.

A3 For any $a, y \in \{1,-1\}$, $U(a,y,\cdot)$ is Borel measurable; in addition, there is some $M > 0$ such that $|U(a,y,x)| \leq M$ for all $x \in \mathcal{X}$ and $a, y \in \{1,-1\}$.

Optimal Decision Rule

$$\max_{a \in \{-1,1\}} \mathbb{E}\left[U(a, Y, X) | X = x\right]$$

Elliott and Lieli (2013) show that under Assumptions A1 and A2, we can obtain an optimal decision rule (after observing $X = x$)

$$\begin{aligned} a^*(x) &\equiv \begin{cases} 1 & \text{if } p^*(x) \geq c(x), \\ -1 & \text{otherwise,} \end{cases} \\ &= \mathsf{sign}(p^*(X) - c(X)) \end{aligned}$$

where $p^*(x) \equiv \mathbb{P}(Y = 1 \mid X = x)$ and

$$c(x) \equiv \frac{U(-1,-1,x) - U(1,-1,x)}{U(1,1,x) - U(-1,1,x) + U(-1,-1,x) - U(1,-1,x)} \in (0,1)$$

is a cutoff function derived from the utility function, which is known in principle to the decision maker.

To solve $\max_{a\in\{-1,1\}} \mathbb{E}\left[U(a,Y,X)|X=x\right]$, we let $u_{a,Y}(X) \equiv U(a,Y,X)$ for ease of notation.

$a=1$:

$$\begin{aligned}
&\mathbb{E}[u_{1,Y}(X)|X=x] \\
&= p(Y=1|X=x)u_{1,1}(x) + p(Y=-1|X=x)u_{1,-1}(x) \\
&= p^*(x)[u_{1,1}(x) - u_{1,-1}(x)] + u_{1,-1}(x)
\end{aligned}$$

$a=-1$:

$$\begin{aligned}
&\mathbb{E}[u_{-1,Y}(X)|X=x] \\
&= p(Y=1|X=x)u_{-1,1}(x) + p(Y=-1|X=x)u_{-1,-1}(x) \\
&= p^*(x)[u_{-1,1}(x) - u_{-1,-1}(x)] + u_{-1,-1}(x)
\end{aligned}$$

We have

$$\begin{aligned}
a^*(x) = 1 \ \text{ if and only if } \ & p^*(x)[u_{1,1}(x) - u_{1,-1}(x)] + u_{1,-1}(x) \\
& \geq p^*(x)[u_{-1,1}(x) - u_{-1,-1}(x)] + u_{-1,-1}(x);
\end{aligned}$$

i.e., $p^*(x) \geq c(x) \equiv \frac{u_{-1,-1}(x) - u_{1,-1}(x)}{u_{1,1}(x) - u_{-1,1}(x) + u_{-1,-1}(x) - u_{1,-1}(x)}$.

- To achieve maximal expected utility in (1), we only need the correct specification of sign$(p^*(x) - c(x))$.

Elliott and Lieli (2013)

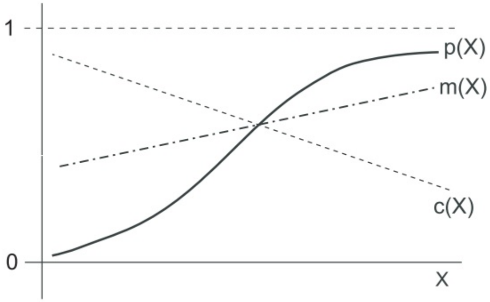


Fig. 1. Here $p(X)$ gives the probability that $Y = 1$ given scalar X, $c(X)$ gives the cutoff for the decision rule, $m(X)$ gives a function that differs from $p(X)$ everywhere but at the cutoff and so delivers the same decisions.

Maximum Utility Estimation

Elliott and Lieli (2013) also show that the decision-making problem in (1) can be equivalently written as

$$\max_f \mathbb{E}\left[b(X)[Y+1-2c(X)]\text{sign}(f(X)-c(X))\right],$$

where $b(x) \equiv U(1,1,x) - U(-1,1,x) + U(-1,-1,x) - U(1,-1,x)$ is the denominator of $c(x)$ and the maximum is taken over all measurable functions from $\mathcal{X}$ to $\mathbb{R}$.

Decomposition:

$$\underbrace{b(X)[Y+1-2c(X)]}_{=\begin{cases} 2[U(1,1,X)-U(-1,1,X)]>0, & \text{if } Y=1 \\ 2[U(1,-1,X)-U(-1,-1,X)]<0, & \text{if } Y=-1. \end{cases}} \qquad \overbrace{\text{sign}(f(X)-c(X))}^{=A}$$

- If $AY = 1$, then $b(X)[Y+1-2c(X)] > 0$;
- If $AY = -1$, then $b(X)[Y+1-2c(X)] < 0$.

Given a sample of observations $\{(Y_i, X_i)\}_{i=1}^n$ and a pre-specified class $\mathcal{F}$ of functions, a maximum utility estimator is defined as

$$\hat{f}_{\mathsf{mu}} \in \arg\max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} b(X_i)[Y_i + 1 - 2c(X_i)]\mathsf{sign}(f(X_i) - c(X_i)).$$

The associated prediction rule is $x \mapsto \mathsf{sign}(\hat{f}_{\mathsf{mu}}(x) - c(x))$.

Manski's (1975, 1985) maximum score estimator is a special case of this maximum utility estimator. (Note that $Y_i\mathsf{sign}(f(X_i) - c(X_i))$ is the score for observation i.)

Overfitting

According to the simulation results, Elliott and Lieli make the following comments:

> *"Both ML and MU have a strong tendency to overfit in sample, however the problem seems more severe for the MU method. This creates challenges for model selection."*

> *"There are a large number of methods for model selection for classification schemes, although none have been shown to extend to the general methods of this paper."*

To alleviate the in-sample overfitting in the maximum utility estimation, Su (2021) further studies the complexity-penalized utility-maximizing prediction rule.

Part 2 Complexity Penalized Maximum Utility Estimation

Su (2021): "Model Selection in Utility-Maximizing Binary Prediction"

Cost-Sensitive Binary Classification

Cause of overfitting?

The maximum utility estimation can be viewed as binary classification in which the cost of misclassification for each in sample observation may be different.

$$
\begin{aligned}
\hat{f}_{\mathsf{mu}} &\in \arg\max_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} b(X_i)[Y_i + 1 - 2c(X_i)]\mathsf{sign}(f(X_i) - c(X_i)) \\
&= \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \underbrace{b(X_i)[Y_i(1 - 2c(X_i)) + 1]}_{\text{cost of mismatch } \geq 0} \mathbb{1}_{[Y_i \neq \mathsf{sign}(f(X_i) - c(X_i))]}.
\end{aligned}
$$

Derivation

Nature of the Overfitting in MU Estimation

- If the cost of mismatch is a constant, then the maximum utility estimation reduces to the traditional binary classification in machine learning

$$\begin{aligned}\hat{f}_{\mathsf{mu}} &\in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \underbrace{b(X_i)[Y_i(1-2c(X_i))+1]}_{\text{cost of mismatch}} \mathbb{1}_{[Y_i\neq \mathsf{sign}(f(X_i)-c(X_i))]} \\ &= \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[Y_i\neq \mathsf{sign}(f(X_i)-c(X_i))]}.\end{aligned}$$

- Moreover, if the in-sample observations can be perfectly separated by $\mathcal{F}$, then

$$\begin{aligned}0 &= \min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[Y_i\neq \mathsf{sign}(f(X_i)-c(X_i))]} \\ &= \min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} b(X_i)[Y_i(1-2c(X_i))+1]\mathbb{1}_{[Y_i\neq \mathsf{sign}(f(X_i)-c(X_i))]}.\end{aligned}$$

$\boxed{\text{Cause of overfitting: Complicated } \mathcal{F}}$

Structural Risk Minimization

How to alleviate overfitting? Vapnik's (1982) structural risk minimization

$$\max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} b(X_i)[Y_i + 1 - 2c(X_i)]\mathsf{sign}(f(X_i) - c(X_i))$$

- The utility of a predictor f evaluated at the observation (y, x) is denoted by

$$s(y, x, f) \equiv b(x)[y + 1 - 2c(x)]\mathsf{sign}(f(x) - c(x))$$

- Given a predictor f constructed based on a sample $\mathscr{D}_n \equiv \{(Y_i, X_i)\}_{i=1}^n$ of observations with sample size n, its expected utility is

$$S(f) \equiv \mathbb{E}[s(Y, X, f) | \mathscr{D}_n]$$

and its empirical utility is

$$S_n(f) \equiv \frac{1}{n} \sum_{i=1}^{n} s(Y_i, X_i, f).$$

Utility-Maximizing Prediction Rule

- Consider nondecreasing sieve $\{\mathcal{F}_k\}_{k=1}^{\infty}$; i.e.,

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_k \subset \cdots \text{ and } \mathcal{F} \equiv \bigcup_{k=1}^{\infty} \mathcal{F}_k.$$

  For example, $\mathcal{F}_k = \mathcal{P}_k$ is the class of polynomial transformations on $\mathcal{X}$ of order at most k.

- For each $\mathcal{F}_k$, we select a maximum utility estimator

$$\hat{f}_k \in \arg\max_{f \in \mathcal{F}_k} S_n(f).$$

  We define a *utility-maximizing prediction rule* (UMPR) as a maximum utility estimator $\hat{f}_k$ that maximizes the complexity penalized empirical utility; specifically,

$$\tilde{f}_n \equiv \hat{f}_{\hat{k}_n}\ , \text{ where } \hat{k}_n = \arg\max_{k \in \mathbb{N}} S_n(\hat{f}_k) - C_n(k).$$

Heuristic Idea of Structural Risk Minimization

Utility-Maximizing Prediction Rule (UMPR):

$$\tilde{f}_n \equiv \hat{f}_{\hat{k}_n}\ , \text{ where } \hat{k}_n = \arg\max_{k \in \mathbb{N}} S_n(\hat{f}_k) - C_n(k).$$

Heuristic Idea:

If $C_n(k) \simeq \underbrace{S_n(\hat{f}_k) - S(\hat{f}_k)}_{\text{magnitude of overfitting}}$,

then $S(\hat{f}_k) \simeq S_n(\hat{f}_k) - C_n(k)$,

$$\begin{aligned} \hat{k}_n &= \arg\max_{k \in \mathbb{N}} \overbrace{S_n(\hat{f}_k) - C_n(k)}^{\text{penalized empirical utility}} \\ &\simeq \arg\max_{k \in \mathbb{N}} S(\hat{f}_k) \end{aligned}$$

and thus

$$S(\tilde{f}_n) \gtrsim S(\hat{f}_k) \text{ for all } k.$$

Resemblance between UMPR and AIC

Utility-Maximizing Prediction Rule (UMPR):

$$\tilde{f}_n \equiv \hat{f}_{\hat{k}_n}\ , \text{ where } \hat{k}_n = \arg\max_{k\in\mathbb{N}} S_n(\hat{f}_k) - C_n(k).$$

Akaike Information Criterion (AIC):

$$\tilde{f}_n^{\mathsf{IC}} \equiv \hat{f}_{\check{k}_n}^{\mathsf{ML}}\ , \text{ where } \check{k}_n = \arg\max_{k\in\mathbb{N}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\hat{f}_k^{\mathsf{ML}}|Y_i, X_i) - C_n^{\mathsf{IC}}(k).$$

- $\mathcal{L}(\hat{f}_k^{\mathsf{ML}}|Y_i, X_i) = \left(\frac{1+Y_i}{2}\right)\log \hat{f}_k^{\mathsf{ML}}(X_i) + \left(\frac{1-Y_i}{2}\right)\log[1-\hat{f}_k^{\mathsf{ML}}(X_i)]$
- $\mathcal{L}$: the log-likelihood function of a single observation (Y, X)
- $\hat{f}_k^{\mathsf{ML}}$: the maximum likelihood estimator in $\mathcal{F}_k$
- $C_n^{\mathsf{IC}}(k)$: $\frac{1}{n}$ $\times$ the number of free parameters in $\mathcal{F}_k$

	UMPR	AIC
Fitting of p^*	local	global
Validity of penalty	non-asymptotic	asymptotic
Methodology	discriminative	generative

$$\begin{tikzcd} \{(Y_i, X_i)\}_{i=1}^n \arrow[rr] \arrow[dr] & & \max\limits_{a \in \{-1,1\}} \mathbb{E}\left[u_{a,Y}(X) | X = x\right] \\ & p^*(x) \equiv \mathbb{P}(Y = 1 \mid X = x) \arrow[ur, dashed] & \end{tikzcd}$$

Vapnik (in the 1990s): "When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one."

Computability-Bounded Rationality

- Heuristic Idea: $C_n(k) \simeq \underbrace{S_n(\hat{f}_k) - S(\hat{f}_k)}_{\text{magnitude of overfitting}}$
- $S_n(\hat{f}_k) - S(\hat{f}_k) \leq \underbrace{\sup_{f \in \mathcal{F}_k} (S_n(f) - S(f))}_{\text{maximal magnitude of overfitting}}$

We avoid measurability complications by imposing the following assumption:

A4 For each $k \in \mathbb{N}$, the class $\mathcal{F}_k$ of functions is countable.

In a computer program, there are only countably many computable real numbers.

This assumption could be interpreted as a decision maker's computability-bounded rationality, as in Richter and Wong (1999).

Concentration Inequality

Theorem (McDiarmid, 1989)

Suppose that $g : \mathcal{Z}^n \to \mathbb{R}$ *satisfies*

$$\sup_{\substack{z_1,\ldots,z_n, \\ z_i' \in \mathcal{Z}}} |g(z_1, \ldots, z_n) - g(z_1, \ldots, z_{(i-1)}, z_i', z_{(i+1)}, \ldots, z_n)| \leq c_i$$

for $1 \leq i \leq n$*. If* $Z_1, \ldots, Z_n$ *are independent random variables taking values in a set* $\mathcal{Z}$*, then for any* $t > 0$*,*

$$\mathbb{P}\left(g(Z_1, \ldots, Z_n) - \mathbb{E}[g(Z_1, \ldots, Z_n)] > t\right) \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right\}.$$

- Talagrand (1996):

  A random variable that depends (in a "smooth" way) on the influence of many independent variables (but not too much on any of them) is essentially constant.

- Boucheron et al. (2013): "Concentration Inequalities: A Nonasymptotic Theory of Independence"

Application

- Taking $g = \sup_{f \in \mathcal{F}_k}(S_n(f) - S(f))$ in McDiarmid's (1989) inequality, we obtain

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}_k}(S_n(f) - S(f)) - \mathbb{E}\left[\sup_{f \in \mathcal{F}_k}(S_n(f) - S(f))\right] > \varepsilon\right) \leq \exp\left\{-\frac{n\varepsilon^2}{32M^2}\right\}.$$

- This inequality implies that given i.i.d. observations, $|S_n(\hat{f}_k) - S(\hat{f}_k)|$ converges almost surely to zero whenever $\mathcal{F}_k$ is a VC-subgraph class.
- See the discussion immediately after Corollary 1 of Su (2021).

Data-Dependent Penalty

- Suppose that we have the ghost sample $\{(Y_i', X_i')\}_{i=1}^n$.

  (That is, the observations $(Y_1', X_1'), \ldots, (Y_n', X_n')$ are distributed as $(Y_1, X_1), \ldots, (Y_n, X_n)$ and independent of them.)

  $S_n'(f)$: empirical utility of f constructed based on the ghost sample

- The common symmetrization argument implies that

$$
\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}_k}(S_n(f) - S(f))\right] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}_k}\Big(S_n(f) - \mathbb{E}[S_n'(f)|\mathscr{D}_n]\Big)\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}_k}\mathbb{E}\left[\Big(S_n(f) - S_n'(f)\Big)\Big|\mathscr{D}_n\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\max_{f\in\mathcal{F}_k}\Big(S_n(f) - S_n'(f)\Big)\Big|\mathscr{D}_n\right]\right] \\
&= \mathbb{E}\left[\max_{f\in\mathcal{F}_k}(S_n(f) - S_n'(f))\right].
\end{aligned}
$$

- We have $\mathbb{E}\left[\sup_{f\in\mathcal{F}_k}\left(S_n(f)-S(f)\right)\right] \le \mathbb{E}\left[\max_{f\in\mathcal{F}_k}\left(S_n(f)-S'_n(f)\right)\right]$.
- It follows from McDiarmid's (1989) inequality that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}_k}\left(S_n(f)-S(f)\right)-\max_{f\in\mathcal{F}_k}\left(S_n(f)-S'_n(f)\right)\ge\varepsilon\right)\\
&\le \mathbb{P}\left(\sup_{f\in\mathcal{F}_k}\left(S_n(f)-S(f)\right)-\max_{f\in\mathcal{F}_k}\left(S_n(f)-S'_n(f)\right)\right.\\
&\qquad\left.-\,\mathbb{E}\left[\sup_{f\in\mathcal{F}_k}\left(S_n(f)-S(f)\right)-\max_{f\in\mathcal{F}_k}\left(S_n(f)-S'_n(f)\right)\right]\ge\varepsilon\right)\\
&\le \exp\left\{-\frac{n\varepsilon^2}{c_0M^2}\right\}
\end{aligned}$$

  for some constant $c_0>0$.
- Therefore, we obtain

$$\sup_{f\in\mathcal{F}_k}\left(S_n(f)-S(f)\right)\le\max_{f\in\mathcal{F}_k}\left(S_n(f)-S'_n(f)\right)+\mathrm{O}\left(\frac{1}{\sqrt{n}}\right)$$

  with high probability.

Maximal Discrepancy (MD)

- In practice, the lack of the ghost sample invalidates the direct estimation of $\max_{f\in\mathcal{F}_k}(S_n(f) - S'_n(f))$.
- We partition the sample into two nonoverlapping and roughly equal-sized subsamples; for example, the sample $\mathscr{D}_n$ is partitioned into two subsamples $\mathscr{D}_{n/2}^{(1)} = \{(Y_{2i-1}, X_{2i-1})\}_{i=1}^{n/2}$ and $\mathscr{D}_{n/2}^{(2)} = \{(Y_{2i}, X_{2i})\}_{i=1}^{n/2}$.
- We define the maximal discrepancy complexity penalty as

$$C_n^{\mathsf{MD}}(k;\alpha) \equiv \max_{f\in\mathcal{F}_k}\left(\frac{2}{n}\sum_{i=1}^{n/2} s(Y_{2i-1}, X_{2i-1}, f) - \frac{2}{n}\sum_{i=1}^{n/2} s(Y_{2i}, X_{2i}, f)\right) + 24M\chi_n(k;\alpha).$$

- We define the maximal discrepancy complexity penalty as

$$C_n^{\mathsf{MD}}(k;\alpha) \equiv \max_{f\in\mathcal{F}_k}\left(\frac{2}{n}\sum_{i=1}^{n/2} s(Y_{2i-1}, X_{2i-1}, f) - \frac{2}{n}\sum_{i=1}^{n/2} s(Y_{2i}, X_{2i}, f)\right) + 24M\chi_n(k;\alpha).$$

- Recap:
  1. Heuristic Idea: $C_n(k) \simeq \underbrace{S_n(\hat{f}_k) - S(\hat{f}_k)}_{\text{magnitude of overfitting}}$
  2. $S_n(\hat{f}_k) - S(\hat{f}_k) \leq \underbrace{\sup_{f\in\mathcal{F}_k} (S_n(f) - S(f))}_{\text{maximal magnitude of overfitting}}$
  3. With high probability,

$$\underbrace{S_n(\hat{f}_k) - S(\hat{f}_k)}_{\text{magnitude of overfitting}} \leq \sup_{f\in\mathcal{F}_k} (S_n(f) - S(f)) \leq \max_{f\in\mathcal{F}_k}\big(S_n(f) - S_n'(f)\big) + \mathrm{O}\left(\frac{1}{\sqrt{n}}\right).$$

- We define the maximal discrepancy complexity penalty as

$$C_n^{\mathsf{MD}}(k;\alpha) \equiv \max_{f \in \mathcal{F}_k} \left(\frac{2}{n} \sum_{i=1}^{n/2} s(Y_{2i-1}, X_{2i-1}, f) - \frac{2}{n} \sum_{i=1}^{n/2} s(Y_{2i}, X_{2i}, f) \right) + 24M\chi_n(k;\alpha).$$

- Let V_k be the Vapnik-Chervonenkis (VC) dimension of the class $\{x \mapsto \mathsf{sign}(f(x) - c(x)) : f \in \mathcal{F}_k\}$. The technical term

$$\chi_n(k;\alpha) \equiv \sqrt{\frac{(1+\alpha)\log\{V_k\}}{2n}}$$

is included in the penalty to guarantee that $\zeta(\alpha) \equiv \sum_{k=1}^{\infty} V_k^{-(1+\alpha)}$ is summable for some α_0. The tuning parameter $\alpha > 0$ can be selected by the tenfold cross-validation method.

Pseudo-Random Maximal Discrepancy

- We draw a sequence $(\sigma_1, \sigma_2, \ldots, \sigma_{n/2})$ of i.i.d. Rademacher random variables that are independent of $\mathscr{D}_n$; that is, $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$.
- We consider the pseudo-random maximal discrepancy complexity penalty (without a technical term)

$$\max_{f \in \mathcal{F}_k} \frac{2}{n} \sum_{i=1}^{n/2} \sigma_i \Big(s(Y_{2i-1}, X_{2i-1}, f) - s(Y_{2i}, X_{2i}, f) \Big).$$

- The previous maximal discrepancy complexity penalty is a special case.

Rademacher Complexity (RC)

Rademacher complexity (Koltchinskii (2001) and Bartlett et al. (2002)) is commonly used to construct a data-dependent penalty in the traditional binary classification.

Let $\{\sigma_i\}_{i=1}^n$ be a sequence of i.i.d. Rademacher random variables that are independent of $\mathscr{D}_n$.

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}_k}\left(S_n(f)-S(f)\right)\right] &\leq \mathbb{E}\left[\max_{f\in\mathcal{F}_k}\Big(S_n(f)-S_n'(f)\Big)\right] \\
&= \mathbb{E}\left[\max_{f\in\mathcal{F}_k}\frac{1}{n}\sum_{i=1}^n\Big(s(Y_i,X_i,f)-s(Y_i',X_i',f)\Big)\right] \\
&= \mathbb{E}\left[\max_{f\in\mathcal{F}_k}\frac{1}{n}\sum_{i=1}^n\sigma_i\Big(s(Y_i,X_i,f)-s(Y_i',X_i',f)\Big)\right] \\
&\leq \mathbb{E}\left[\max_{f\in\mathcal{F}_k}\frac{1}{n}\sum_{i=1}^n\sigma_i s(Y_i,X_i,f)\right]+\mathbb{E}\left[\max_{f\in\mathcal{F}_k}\frac{1}{n}\sum_{i=1}^n(-\sigma_i)s(Y_i',X_i',f)\right] \\
&= \mathbb{E}\left[\max_{f\in\mathcal{F}_k}\frac{2}{n}\sum_{i=1}^n\sigma_i s(Y_i,X_i,f)\right].
\end{aligned}$$

- $\mathbb{E}\left[\sup_{f\in\mathcal{F}_k}(S_n(f)-S(f))\right] \le \mathbb{E}\left[\max_{f\in\mathcal{F}_k}\frac{2}{n}\sum_{i=1}^n \sigma_i s(Y_i,X_i,f)\right]$
- Applying McDiarmid's (1989) inequality, we have

$$\sup_{f\in\mathcal{F}_k}(S_n(f)-S(f)) \le \underbrace{\mathbb{E}\left[\max_{f\in\mathcal{F}_k}\frac{2}{n}\sum_{i=1}^n \sigma_i s(Y_i,X_i,f)\middle|\mathscr{D}_n\right]}_{\text{empirical Rademacher complexity}} + \mathrm{O}\left(\frac{1}{\sqrt{n}}\right)$$

  with high probability.

We define the simulated Rademacher complexity penalty as

$$C_n^{\mathsf{RC}}(k;\alpha,m) \equiv \frac{1}{m}\sum_{j=1}^m\left(\max_{f\in\mathcal{F}_k}\frac{2}{n}\sum_{i=1}^n \sigma_i^{(j)} s(Y_i,X_i,f)\right) + \gamma_{m,n}(M)\chi_n(k;\alpha),$$

where $\{\sigma^{(j)}\}_{j=1}^m = \{(\sigma_1^{(j)},\sigma_2^{(j)},\ldots,\sigma_n^{(j)})\}_{j=1}^m$ is the collection of i.i.d. Rademacher random vectors that are independent of $\mathscr{D}_n$, and $\gamma_{m,n}$ is a deterministic function that satisfies

$$\gamma_{m,n}(M) = \begin{cases} 40M, & \text{if } n \le m < \infty, \\ (16\ell+40)M, & \text{if } n/(\ell+1)^2 \le m < n/\ell^2 \text{ and } \ell\in\mathbb{N}. \end{cases}$$

- $\gamma_{m,n}$ is designed to control the extra randomness introduced by $\{\sigma^{(j)}\}_{j=1}^m$.

Bootstrap Complexity (BC)

- Note that $\sigma \overset{\mathsf{d}}{=} 2B - 1$, where $B \sim \text{Ber}(1/2)$.
- The simulated Rademacher complexity penalty (without a technical term) satisfies

$$
\begin{aligned}
&\frac{1}{m}\sum_{j=1}^{m}\left(\max_{f\in\mathcal{F}_k}\frac{2}{n}\sum_{i=1}^{n}\sigma_i^{(j)}s(Y_i,X_i,f)\right)\\
\overset{\mathsf{d}}{=}&\frac{2}{m}\sum_{j=1}^{m}\left(\max_{f\in\mathcal{F}_k}\frac{1}{n}\sum_{i=1}^{n}(2B_i^{(j)}-1)s(Y_i,X_i,f)\right).
\end{aligned}
$$

- Fromont (2007) suggests using bootstrap to construct a complexity penalty.

We define the bootstrap complexity penalty as

$$C_n^{\mathsf{BC}}(k;\alpha,m) \equiv \left(\frac{n}{n-1}\right)^n \frac{1}{m}\sum_{j=1}^{m}\left(\max_{f\in\mathcal{F}_k}\frac{1}{n}\sum_{i=1}^{n}\left(W_{n,i}^{(j)}-1\right)s(Y_i,X_i,f)\right) + \gamma'_{m,n}(M)\chi_n(k;\alpha),$$

where $\{W_n^{(j)}\}_{j=1}^m = \{(W_{n,1}^{(j)}, W_{n,2}^{(j)}, \ldots, W_{n,n}^{(j)})\}_{j=1}^m$ is the collection of i.i.d. multinomial vectors with parameters n and $(1/n, 1/n, \ldots, 1/n)$ such that $\{W_n^{(j)}\}_{j=1}^m$ is independent of $\mathscr{D}_n$, and

$$\gamma'_{m,n}(M) = \begin{cases} 56M, & \text{if } n \le m < \infty, \\ (32\ell+56)M, & \text{if } n/(\ell+1)^2 \le m < n/\ell^2 \text{ and } \ell \in \mathbb{N}. \end{cases}$$

- $\gamma'_{m,n}$ is designed to control the extra randomness introduced by $\{W_n^{(j)}\}_{j=1}^m$.

Theoretical Properties of the UMPR

- Recap: $S(f) = \mathbb{E}[b(X)[Y + 1 - 2c(X)]\text{sign}(f(X) - c(X))|\mathscr{D}_n]$
- Let $S^* \equiv S(p^*)$ be the maximal expected utility and $S_k^* \equiv \sup_{f \in \mathcal{F}_k} S(f)$ for each k.
- $S^* - S_k^*$: approximation error for $\mathcal{F}_k$
- $\mathbb{E}[S(\tilde{f}_n)]$: generalized expected utility of the UMPR

Theorem 1

Suppose that (i) the data $\mathscr{D}_n = \{(Y_i, X_i)\}_{i=1}^n$ are i.i.d., (ii) $\mathcal{F}_k$ is a VC-subgraph class with VC index V_k for each k, (iii) $\zeta(\alpha_0) < \infty$ for some α_0, and (iv) Assumptions A1-A4 hold.

If the UMPR $\tilde{f}_n$ is constructed based on the penalty C_n^{RC} with tuning parameter α_0, then we have for any $n \in \mathbb{N}$,

$$
\begin{aligned}
& S^* - \mathbb{E}[S(\tilde{f}_n)] \\
& \leq \min_k \left\{ (S^* - S_k^*) + \mathbb{E}\left[C_n^{RC}(k; \alpha_0, m) \right] \right\} + \gamma_{m,n}(M) \sqrt{\frac{1 + \log\{2\zeta(\alpha_0)\}}{2n}}.
\end{aligned}
$$

Trade-off:

$$
k \uparrow \quad \Rightarrow \quad \begin{array}{l} \text{approximation error } S^* - S_k^* \downarrow \\ \text{expected complexity penalty } \mathbb{E}\left[C_n^{\mathrm{RC}}(k; \alpha_0, m) \right] \uparrow \end{array}
$$

Corollary 1

Suppose that the assumptions of Theorem 1 hold. If in addition $m/n \geq 1/\bar{\ell}^2$ *for some positive integer* $\bar{\ell}$*, then there are positive constants* κ_1 *and* κ_2 *only depending on* M*, and* κ_3 *depending on* $(M, \bar{\ell})$ *such that for each* $k \in \mathbb{N}$ *and* $n \geq 8$*,*

$$\mathbb{E}\left[C_n^{RC}(k; \alpha_0, m)\right] \leq \kappa_1 \sqrt{\frac{V_k}{n}} + \kappa_2 V_k \frac{(\log\{n\})^2}{n} + \kappa_3 \sqrt{1+\alpha_0} \sqrt{\frac{\log\{V_k\}}{n}}.$$

Moreover, the UMPR $\tilde{f}_n$ *constructed based on the penalty* C_n^{RC} *with tuning parameter* α_0 *satisfies*

$$\lim_{n\to\infty} S(\tilde{f}_n) = S^* \text{ with probability one}$$

for any distribution of (Y, X) *such that* $\lim_{k\to\infty} S_k^* = S^*$*.*

- Using the other penalties to construct the UMPR, we obtain similar results.

Proposition 1

Suppose Assumptions A1 and A2 hold. For any (measurable) deterministic function $f : \mathcal{X} \mapsto \mathbb{R}$, we have

$$S^* - S(f) = 4\,\mathbb{E}\left[b(X)[p^*(X) - c(X)](\mathbb{1}_{[p^*(X)\geq c(X)]} - \mathbb{1}_{[f(X)\geq c(X)]})\right] \geq 0$$

and

$$S^* - S(f) \leq 4\,\mathbb{E}\left[b(X)|p^*(X) - f(X)|\right] \leq 16M \sup_{x\in\mathcal{X}} |p^*(x) - f(x)|.$$

- For each $k \in \mathbb{N}$,

$$0 \leq \underbrace{S^* - S_k^*}_{\text{approximation error}} \leq 16M \underbrace{\inf_{f\in\mathcal{F}_k} \sup_{x\in\mathcal{X}} |p^*(x) - f(x)|}_{\text{uniform distance between } p^* \text{ and } \mathcal{F}_k}.$$

- If we specify $\mathcal{F}_k = \mathcal{P}_k$, then

$$\inf_{f\in\mathcal{F}_k} \sup_{x\in\mathcal{X}} |f(x) - p^*(x)| \to 0 \ \text{ as } k \to \infty$$

  whenever p^* is continuous on the compact support $\mathcal{X} \subseteq \mathbb{R}^d$.
  (Stone-Weierstrass approximation theorem)

Monte Carlo Experiments

We consider the simulation designs in Elliott and Lieli (2013).

DGP1 The covariate X follows the distribution $5 \cdot \text{beta}(1, 1.3) - 2.5$ and $p^*(X) = \Lambda(-0.5X + 0.2X^3)$ where Λ is the standard logistic function; i.e., $\Lambda(u) = (1 + \exp\{-u\})^{-1}$ for all $u \in \mathbb{R}$;

Pref.1 $b(X) = 20$ and $c(X) = 0.5$;

Pref.2 $b(X) = 20$ and $c(X) = 0.5 + 0.025X$;

DGP2 Both covariates X_1 and X_2 are independent and uniformly distributed on $[-3.5, 3.5]$ and $p^*(X_1, X_2) = \Lambda(Q(1.5X_1 + 1.5X_2))$, where $Q(u) = (1.5 - 0.1u)\exp\{-(0.25u + 0.1u^2 - 0.04u^3)\}$.

Pref.3 $b(X_1, X_2) = 20$ and $c(X_1, X_2) = 0.75$;

Pref.4 $b(X_1, X_2) = 20 + 40 \cdot \mathbb{1}_{[|X_1+X_2|<1.5]}$ and $c(X_1, X_2) = 0.75$.

Preference

- For the UMPR with any aforementioned penalty, we specify the hierarchy $\{\mathcal{F}_k\}_{k=1}^{\infty}$ of classes as $\mathcal{F}_k = \mathcal{P}_k$ for $k \in \{1, 2\}$ and $\mathcal{F}_k = \mathcal{P}_3$ for all $k \geq 3$.
- For the AIC and BIC, we specify the hierarchy $\{\mathcal{F}_k\}_{k=1}^{\infty}$ of classes as $\mathcal{F}_k = \Lambda(\mathcal{P}_k)$ for $k \in \{1, 2\}$ and $\mathcal{F}_k = \Lambda(\mathcal{P}_3)$ for all $k \geq 3$, where $\Lambda(\mathcal{P}_k) \equiv \{x \mapsto \Lambda(f(x)) : f \in \mathcal{P}_k\}$ for each $k \in \mathbb{N}$.
- We also compute the tenfold cross-validatory LASSO (Tibshirani (1996)) with optimization taken over the class $\Lambda(\mathcal{P}_3)$ and ℓ_1-norm SVM (Fung and Mangasarian (2004)) with optimization taken over the class $\mathcal{P}_3$.

Least Absolute Shrinkage and Selection Operator

Cubic Lasso-logit (i.e., cubic ML-logit with an ℓ_1 penalty)

$$\max_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \left\{ \left(\frac{1+Y_i}{2} \right) \log p(X_i; \boldsymbol{\theta}) + \left(\frac{1-Y_i}{2} \right) \log[1 - p(X_i; \boldsymbol{\theta})] \right\} - \lambda \|\boldsymbol{\theta}\|_1$$

- DGP 1: $p(x; \boldsymbol{\theta}) \equiv \Lambda\left(\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3\right)$,

  $\|\boldsymbol{\theta}\|_1 = \sum_{i=0}^{3} |\theta_i|$

- DGP 2: $p(x; \boldsymbol{\theta}) \equiv \Lambda\Big(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2$

  $+\theta_6 x_1^3 + \theta_7 x_2^3 + \theta_8 x_1^2 x_2 + \theta_9 x_1 x_2^2\Big)$,

  $\|\boldsymbol{\theta}\|_1 = \sum_{i=0}^{9} |\theta_i|$

Support Vector Machine

Lasso-logit (i.e., logistic loss with an ℓ_1 penalty)

$$\max_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{1+Y_i}{2}\right)\log p(X_i;\boldsymbol{\theta}) + \left(\frac{1-Y_i}{2}\right)\log[1-p(X_i;\boldsymbol{\theta})]\right\} - \lambda\|\boldsymbol{\theta}\|_1$$

$$= -\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n}\underbrace{\log\left[1+\exp\left(-Y_i f(X_i;\boldsymbol{\theta})\right)\right]}_{\text{logistic loss}} + \lambda\|\boldsymbol{\theta}\|_1$$

where $p(x;\boldsymbol{\theta}) = \Lambda(f(x;\boldsymbol{\theta}))$ and $f(x;\boldsymbol{\theta})$ is a polynomial in x with coefficient $\boldsymbol{\theta}$.

ℓ_1-norm SVM (i.e., Hinge loss with an ℓ_1 penalty)

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n}\underbrace{\max\{0, 1-Y_i f(X_i;\boldsymbol{\theta})\}}_{\text{Hinge loss}} + \lambda\|\boldsymbol{\theta}\|_1$$

$$\Rightarrow \text{SVM prediction rule } \hat{f}_{\mathsf{SVM}}(x) \equiv \Lambda(f(x;\hat{\boldsymbol{\theta}}_{\mathsf{SVM}}))$$

Note that $\hat{y} \equiv \mathsf{sign}(f(x;\hat{\boldsymbol{\theta}}_{\mathsf{SVM}})) = \mathsf{sign}(\hat{f}_{\mathsf{SVM}}(x) - 1/2)$.

We compute the relative generalized expected utility of a prediction rule $f_n^\dagger$

$$\mathsf{RGEU}(f_n^\dagger) \equiv \frac{\mathbb{E}[S(f_n^\dagger)]}{S^*}$$

where $S^* \equiv \sup_f S(f) = S(p^*)$.

The relative expected utility can be approximated via simulation:

$$\mathsf{RGEU}(f_n^\dagger) = \mathbb{E}\left[\frac{S(f_n^\dagger)}{S(p^*)}\right] \simeq \frac{1}{\mathcal{S}} \sum_{j=1}^{\mathcal{S}} \frac{S_{\ell,j}(f_n^\dagger|\mathscr{D}_{n,j})}{S_{\ell,j}(p^*)},$$

- $S_{\ell,j}(f_n^\dagger|\mathscr{D}_{n,j})$ is the j-th (out-of-sample) empirical utility with size ℓ of $f_n^\dagger$, constructed by the j-th in-sample $\mathscr{D}_{n,j}$ with size n,
- $S_{\ell,j}(p^*)$ is the j-th (out-of-sample) empirical utility with size ℓ of p^*, and
- $\mathcal{S}$ is the number of simulation replications.

We set $n \in \{500, 1000\}$, $m = 10$, $\ell = 5000$, and $\mathcal{S} = 500$. Details

Table 1: Relative Generalized Expected Utility of UMPR, AIC, BIC, LASSO and SVM

DGP1 $p^*(x) = \Lambda(-0.5x + 0.2x^3)$

$n = 500$

Preference	$b(x) = 20$ and $c(x) = 0.5$				$b(x) = 20$ and $c(x) = 0.5 + 0.025x$			
UMPR	MD	SMD	RC	BC	MD	SMD	RC	BC
	65.36	66.68	66.86	65.74	55.00	58.87	58.58	57.65
Information	AIC	BIC			AIC	BIC		
Criterion	93.93	89.95			94.70	88.81		
ℓ_1-Penalty	LASSO	SVM			LASSO	SVM		
	60.60	87.77			65.62	83.91		

$n = 1000$

Preference	$b(x) = 20$ and $c(x) = 0.5$				$b(x) = 20$ and $c(x) = 0.5 + 0.025x$			
UMPR	MD	SMD	RC	BC	MD	SMD	RC	BC
	69.32	72.51	72.23	71.75	63.30	67.12	67.01	65.81
Information	AIC	BIC			AIC	BIC		
Criterion	97.21	97.13			97.48	97.29		
ℓ_1-Penalty	LASSO	SVM			LASSO	SVM		
	68.82	93.26			78.92	91.14		

DGP2 $p^*(x_1,x_2) = \Lambda(Q(1.5x_1 + 1.5x_2))$ where $Q(u) = \frac{1.5-0.1u}{\exp\{0.25u+0.1u^2-0.04u^3\}}$

$n = 500$

Preference	$b(x_1,x_2) = 20$ and $c(x_1,x_2) = 0.75$				$b(x_1,x_2) = 20 + 40 \cdot \mathbb{1}_{[\lvert x_1+x_2\rvert<1.5]}$ and $c(x_1,x_2) = 0.75$			
UMPR	MD	SMD	RC	BC	MD	SMD	RC	BC
	68.55	69.52	69.47	69.11	50.41	53.87	53.32	52.90
Information Criterion	AIC	BIC			AIC	BIC		
	60.07	60.27			33.20	30.90		
ℓ_1-Penalty	LASSO	SVM			LASSO	SVM		
	59.75	26.86			32.93	5.92		

$n = 1000$

Preference	$b(x_1,x_2) = 20$ and $c(x_1,x_2) = 0.75$				$b(x_1,x_2) = 20 + 40 \cdot \mathbb{1}_{[\lvert x_1+x_2\rvert<1.5]}$ and $c(x_1,x_2) = 0.75$			
UMPR	MD	SMD	RC	BC	MD	SMD	RC	BC
	71.09 $\uparrow$	71.91 $\uparrow$	71.97 $\uparrow$	71.89 $\uparrow$	57.13 $\uparrow$	59.61 $\uparrow$	60.08 $\uparrow$	58.96 $\uparrow$
Information Criterion	AIC	BIC			AIC	BIC		
	59.72 $\downarrow$	59.06 $\downarrow$			31.49 $\downarrow$	28.16 $\downarrow$		
ℓ_1-Penalty	LASSO	SVM			LASSO	SVM		
	59.68 $\downarrow$	25.93 $\downarrow$			29.08 $\downarrow$	5.10 $\downarrow$		

Pretest

- For $k \in \{2, 3\}$, consider $\begin{cases} H_0^{(k)} : S^*_{(k-1)} = S^*_k \\ H_1^{(k)} : S^*_{(k-1)} < S^*_k \end{cases}$
- Test statistic is developed by Elliott and Lieli (2013).

A general-to-specific approach:

$$\hat{k}(\mathsf{G} \to \mathsf{S}) = \begin{cases} 1, & \text{if neither } H_0^{(3)} \text{ nor } H_0^{(2)} \text{ is rejected,} \\ \max\left\{k \in \{2,3\} : H_0^{(k)} \text{ is rejected against } H_1^{(k)}\right\}, & \text{otherwise.} \end{cases}$$

A specific-to-general approach:

$$\hat{k}(\mathsf{S} \to \mathsf{G}) = \begin{cases} 3, & \text{if both } H_0^{(3)} \text{ and } H_0^{(2)} \text{ are rejected,} \\ \min\left\{k \in \{2,3\} : H_0^{(k)} \text{ is not rejected against } H_1^{(k)}\right\} - 1, & \text{otherwise.} \end{cases}$$

Cross-Validation

- We randomly partition the data $\mathscr{D}_n$ into T roughly equal-sized sets. Let $\tau : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, T\}$ be the indexing function such that the observation (Y_i, X_i) is in the validation set $\tau(i)$.
- For each $k \in \{1, 2, 3\}$ and $t \in \{1, 2, \ldots, T\}$, we calculate the MU estimator based on $\mathscr{D}_n^{(-t)}$ by

$$\hat{f}_k^{(-t)} \in \arg\max_{f \in \mathcal{F}_k} \sum_{i:\tau(i) \neq t} s(Y_i, X_i, f).$$

- The cross-validated value of k is defined as

$$\hat{k}_n = \arg\max_{k \in \{1,2,3\}} \sum_{t=1}^{T} \sum_{i:\tau(i)=t} s(Y_i, X_i, \hat{f}_k^{(-t)}).$$

- The cross-validated MU estimator is the MU estimator selected from $\mathcal{F}_{\hat{k}_n}$ based on $\mathscr{D}_n$; specifically,

$$\hat{f}_{\hat{k}_n}^{\mathsf{CV}} \in \arg\max_{f \in \mathcal{F}_{\hat{k}_n}} S_n(f).$$

Table 2: Relative Generalized Expected Utility of UMPR, Pretest, and Cross-Validation

DGP1 $p^*(x) = \Lambda(-0.5x + 0.2x^3)$

$n = 500$

Preference	$b(x) = 20$ and $c(x) = 0.5$				$b(x) = 20$ and $c(x) = 0.5 + 0.025x$			
UMPR	MD	SMD	RC	BC	MD	SMD	RC	BC
	65.36	66.68	66.86	65.74	55.00	58.87	58.58	57.65
Pretest	S→G	G→S			S→G	G→S		
	59.27	62.69			45.63	48.69		
Cross-Validation	61.30				50.42			

$n = 1000$

Preference	$b(x) = 20$ and $c(x) = 0.5$				$b(x) = 20$ and $c(x) = 0.5 + 0.025x$			
UMPR	MD	SMD	RC	BC	MD	SMD	RC	BC
	69.32	72.51	72.23	71.75	63.30	67.12	67.01	65.81
Pretest	S→G	G→S			S→G	G→S		
	62.60	65.20			50.14	53.52		
Cross-Validation	64.81				55.19			

DGP2 $\quad p^*(x_1,x_2)=\Lambda(Q(1.5x_1+1.5x_2))$ where $Q(u)=\frac{1.5-0.1u}{\exp\{0.25u+0.1u^2-0.04u^3\}}$

$n=500$

Preference	$b(x_1,x_2)=20$ and $c(x_1,x_2)=0.75$				$b(x_1,x_2)=20+40\cdot\mathbb{1}_{[\lvert x_1+x_2\rvert<1.5]}$ and $c(x_1,x_2)=0.75$			
UMPR	MD	SMD	RC	BC	MD	SMD	RC	BC
	68.55	69.52	69.47	69.11	50.41	53.87	53.32	52.90
Pretest	S→G	G→S			S→G	G→S		
	68.72	68.34			50.62	49.91		
Cross-Validation	67.30				48.26			

$n=1000$

Preference	$b(x_1,x_2)=20$ and $c(x_1,x_2)=0.75$				$b(x_1,x_2)=20+40\cdot\mathbb{1}_{[\lvert x_1+x_2\rvert<1.5]}$ and $c(x_1,x_2)=0.75$			
UMPR	MD	SMD	RC	BC	MD	SMD	RC	BC
	71.09	71.91	71.97	71.87	57.13	59.61	60.08	58.96
Pretest	S→G	G→S			S→G	G→S		
	70.90	71.20			56.64	56.48		
Cross-Validation	69.93				54.51			

Conclusion

We propose a method of model selection in the framework of maximum utility estimation.

- The maximum utility estimation proposed by Elliott and Lieli (2013) can be viewed as cost-sensitive binary classification.
- Applying the structural risk minimization in machine learning, we construct a utility-maximizing prediction rule (UMPR) to alleviate the in-sample overfitting of MU estimation.
- Under regularity conditions, the expected utility of the UMPR converges to the maximal expected utility if the approximation error goes to zero.
- Simulation results show that the UMPR, in comparison to some common estimators (AIC, BIC, LASSO, ℓ_1-norm SVM) may have larger relative expected utility if the conditional probability of the binary outcome is misspecified.

References I

ANDERSON, R. (2007): *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press.

BARTLETT, P. L., S. BOUCHERON, AND G. LUGOSI (2002): "Model Selection and Error Estimation," *Machine Learning*, 48, 85–113.

BOUCHERON, S., G. LUGOSI, AND P. MASSART (2013): *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press.

DEVROYE, L., L. GYÖRFI, AND G. LUGOSI (1996): *A Probabilistic Theory of Pattern Recognition*, Springer.

ELLIOTT, G. AND R. P. LIELI (2013): "Predicting Binary Outcomes," *Journal of Econometrics*, 174, 15–26.

ELLIOTT, G. AND A. TIMMERMANN (2016): "Forecasting in Economics and Finance," *Annual Review of Economics*, 8, 81–110.

FRISCH, R. (1933): "Editor's Note," *Econometrica*, 1, 1–4.

FROMONT, M. (2007): "Model Selection by Bootstrap Penalization for Classification," *Machine Learning*, 66, 165–207.

FUNG, G. M. AND O. MANGASARIAN (2004): "A Feature Selection Newton Method for Support Vector Machine Classification," *Computational Optimization and Applications*, 28, 185–202.

GRANGER, C. W. AND M. J. MACHINA (2006): "Forecasting and Decision Theory," Elsevier, vol. 1 of *Handbook of Economic Forecasting*, chap. 2, 81–98.

References II

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2021): *An Introduction to Statistical Learning: with Applications in R*, 2nd ed.

Koltchinskii, V. (2001): "Rademacher Penalties and Structural Risk Minimization," *IEEE Transactions on Information Theory*, 47, 1902–1914.

Lieli, R. P. and H. White (2010): "The Construction of Empirical Credit Scoring Rules Based on Maximization Principles," *Journal of Econometrics*, 157, 110–119.

Manski, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228.

——— (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313–333.

McDiarmid, C. (1989): *On the Method of Bounded Differences*, Cambridge University Press, 148–188, London Mathematical Society Lecture Note Series.

Richter, M. K. and K.-C. Wong (1999): "Non-Computability of Competitive Equilibrium," *Economic Theory*, 14, 1–27.

Su, J.-H. (2021): "Model Selection in Utility-Maximizing Binary Prediction," *Journal of Econometrics*, 223, 96–124.

Talagrand, M. (1996): "A New Look at Independence," *Annals of Probability*, 24, 1–34.

References III

Thomas, L., J. Crook, and D. Edelman (2017): *Credit Scoring and Its Applications*, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2nd ed.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Vapnik, V. (1982): *Estimation of Dependences Based on Empirical Data*, Springer.

Varian, H. R. (2014): "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28, 3–28.

板谷敏彦 (2022): 金融的世界史*:* 泡沫經濟、戰爭與股市, 左岸文化, 陳家豪譯.

Frisch (1933, Econometrica):

> *But there are several aspects of the quantitative approach to economics, and no single one of these aspects, taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics.... Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. And it is this unification that constitutes econometrics.*

computer science: Turing machine (1936), von Neumann architecture (1945)

Varian (2014, JEP):

> *In fact, my standard advice to graduate students these days is go to the computer science department and take a class in machine learning.... There have been very fruitful collaborations between computer scientists and statisticians in the last decade or so, and I expect collaborations between computer scientists and econometricians will also be productive in the future.*

Job opening: full-time/part-time research assistants!

Please contact me:
jhsu@econ.sinica.edu.tw