Journal of the Chinese
Statistical Association
Vol. 61, (2023) 271–293

邏輯斯迴歸的最佳參數估計子抽樣方法

陳俊廷 [1] **蕭維政** [2] **張明中** [3] **黃世豪** [4,1†]

[1] 國立中央大學統計研究所
[2] 東吳大學財務工程與精算數學系
[3] 中央研究院統計科學研究所
[4] 國立中央大學數學系

摘　要

在現今這個大數據世代中，研究者可以輕易地蒐集到許多的樣本與其解釋變數，但這些樣本反應變數之標記可能需要額外的花費才能取得。這時候如何挑選具有代表性的子樣本予以標記來建立精準模型即為主動式學習的研究議題。本文旨在介紹近期文獻中對邏輯斯迴歸的最佳子抽樣方法：GATE [Comput. Stat. Data Anal. 129 (2019) 119–134] 與 MMSE [J. Am. Stat. Assoc. 113 (2018) 829–844]，並對這些方法加以調整，使其能在主動式學習的框架下達到更精準的參數估計。本文以模擬實驗和實際的普查資料分析佐證，調整後的子抽樣方法在邏輯斯迴歸模型下通常有較佳的參數估計效率。

關鍵詞：主動式學習、邏輯斯迴歸、最適設計、最佳子抽樣、參數估計。

JEL classification: C35, C61, C83, C90.

†通訊作者: 黃世豪
E-mail: shhuang@math.ncu.edu.tw

1. 緒論

二分類問題 (binary classification) 為統計學習中很常見的監督式學習 (supervised learning) 問題。當需要統計證據來推論解釋變數怎麼對二元反應變數造成影響並估計反應機率時，研究者常利用邏輯斯迴歸 (logistic regression) 來建立統計模型。比起過去，在現今大數據世代中，我們可以更容易地蒐集到許多樣本與其可能的解釋變數，但有些時候反應變數的標記卻需要額外的花費才能取得。例如，Kohavi (1996) 使用美國人口普查局在 1994 年蒐集的普查資料建立分類模型，用以推論哪些因素會影響成年人的年收入是否超過 50000 美元。這筆資料的解釋變數都是一些短期內不太會改變或容易推算的量，如年齡、性別、種族、教育程度等等。若想要以這筆資料為基礎，研究哪些因素會影響成年人是否支持某項政策，相當於已經擁有許多完整解釋變數資訊的樣本，但反應變數還需要額外調查。而在 Deng *et al.* (2009) 中描述了一個研究銀行帳戶洗錢偵測的問題。對銀行而言，帳戶的基本資料、交易對象、交易金額、交易時間等等可能的解釋變數很容易取得，但該帳戶是否涉及洗錢則需要額外的花費才能得到正確標記。

對影片、影像、音訊、或文稿類型的資料進行統計分析一直是現代科學研究的重要一環，而這些類型的資料同樣會發生標記取得不易的情況。在 Cesa-Bianchi, Gentile, and Zaniboni (2006) 及 Muslea, Minton, and Knoblock (2006) 的研究中，樣本的解釋變數可透過網頁爬蟲取得，相較之下要對網頁的內容進行主題分類會消耗大量時間跟人力資源。另外，在 Imberg *et al.* (2022) 的人類自然駕駛 (naturalistic driving) 研究中，自然駕駛影片資料透過自動化程式可取得車速、前車距離和車道偏離等解釋變數，但作為反應變數的安全臨界事件 (safety critical event) 是否發生則需要專業人員進行人工判斷。在有限的人力與時間之限制下，專業人員只能挑選部分影片予以標記。

在上面這些例子中，如何挑選具有較多資訊的部分樣本來建立精準模型是很重要的。在機器學習的文獻中，主動式學習 (active learning) 為處理此類子抽樣 (subsampling) 問題的顯學 (MacKay, 1992; Cohn, Ghahramani, and Jordan, 1996)。在主動式學習中，一開始會有少量的標記樣本作為訓練集建立起始分類器。根據分類器提供的資訊，主動式學習演算法會指定一個能增加最多資訊量的未標記樣本交由專家予以標記，接著再重複更新訓練集、更新分類器、以及標記下一個樣本這些步驟，直到達成停止條件。根據不同的分類方法以及對「增加最多資訊量的未標記樣本」的

不同定義，陸續有學者提出各式各樣的主動式學習法。對主動式學習的發展與實際應用有興趣的讀者可以參閱 Settles (2009) 及 Xu *et al.* (2019) 這兩篇詳盡的研究回顧。

在統計文獻中，傳統的最適設計 (optimal design) 問題為在給定模型之下考慮如何配置設計點來得到最佳模型估計的問題 (Pukelsheim, 2006)。若將未標記資料所構成的集合視為設計空間，則最佳子抽樣問題就變成了這個離散型設計空間下的最適設計問題。Deng *et al.* (2009) 將實驗設計中序貫式D-最適設計 (sequential D-optimal design) 的想法整合到主動式學習中挑選未標記樣本的準則中，提出了 ALSD 法 (Active Learning through Sequential Design)。而 Hsu, Chang, and Chen (2019) 以 ALSD 演算法為基礎，進一步地考量變數選取，提出了 GATE 法 (Greedy AcTivE learning) 用以增進主動式學習的預測正確率。由於 ALSD 法及 GATE 法主要關注在分類正確性上，需要將邏輯斯迴歸模型可能是錯誤模型的情況納入考量，因此選取子樣本時只考慮反應機率在分類決策邊界之機率值附近的樣本點，該機率值通常設為 0.5。然而，根據最適設計文獻，如 Ford, Torsney, and Wu (1992)、Kabera, Haines, and Ndlovu (2015)、及 Huang, Huang, and Lin (2020)，對邏輯斯迴歸的參數估計而言，最有效率的樣本點之反應機率大約在 0.2 跟 0.8 附近。由此可知，若目標是要對邏輯斯迴歸模型之參數估計做統計推論，則 ALSD 法或 GATE 法這樣僅考慮在反應機率在單一機率值附近的子抽樣方法或許不盡理想。另一方面，Wang, Zhu, and Ma (2018) 在標記已知的前提下，結合實驗設計中的 A-最適準則 (A-optimality criterion) 提出了 MMSE (Minimum asymptotic Mean Squared Error) 法來選擇對估計邏輯斯迴歸模型參數而言最有效的子樣本。但由於 MMSE 法需要所有資料皆為已標記資料，因此需要加以調整以應用在主動式學習中未標記資料的子抽樣問題上。

在本文的研究架構中，我們假設邏輯斯迴歸模型為真實模型，並將問題聚焦於以參數估計效率為主要考量的未標記資料子抽樣問題。我們將對 GATE 法與 MMSE 法進行調整以應用在此問題上。在第 2 節我們將簡介背景知識並介紹 GATE 與 MMSE 子抽樣方法，以及提出調整方案。第 3 節和第 4 節為模擬實驗與實際資料分析，我們將觀察各個方法在不同情境下的表現。第 5 節則為結論與討論。

2. 文獻回顧與方法介紹

2.1 預備知識

我們考慮以下的邏輯斯迴歸模型：

$$\mathbb{E}(Y_j) = \mu(\boldsymbol{x}_j) = F(\boldsymbol{x}_j^\top \boldsymbol{\beta}), \quad j \in \mathcal{A} = \{1, \ldots, N\} \tag{1}$$

其中二元反應變數 Y_j 服從 Bernulli 分佈，取值為 0 或 1；反應機率 $\mu(\boldsymbol{x}_j) \in (0,1)$；$\boldsymbol{x}_j \in \mathbb{R}^p$ 為解釋變數向量，也稱為樣本點；$\boldsymbol{\beta} \in \mathbb{R}^p$ 為參數向量；$F(c) = 1/\{1 + \exp(-c)\}$ 為標準邏輯斯分佈之累積分佈函數；N 為總樣本數；$\mathcal{A}$ 為總指標集。在邏輯斯迴歸模型(1)中，我們通常使用最大概似估計法 (MLE, Maximum Likelihood Estimation) 來估計參數 $\boldsymbol{\beta}$，即

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \log L(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{j \in \mathcal{A}} \{y_j \log(\mu_j) + (1 - y_j) \log(1 - \mu_j)\}$$

其中 y_j 為 Y_j 之觀測值，$\mu_j = \mu(\boldsymbol{x}_j) = F(\boldsymbol{x}_j^\top \boldsymbol{\beta})$。此時參數 $\boldsymbol{\beta}$ 之訊息矩陣 (information matrix) 為

$$I(\mathcal{A}; \boldsymbol{\beta}) = \sum_{j \in \mathcal{A}} w_j \boldsymbol{x}_j \boldsymbol{x}_j^\top \tag{2}$$

其中影響係數 $w_j = F'(\boldsymbol{x}_j^\top \boldsymbol{\beta})^2 / [F(\boldsymbol{x}_j^\top \boldsymbol{\beta})\{1 - F(\boldsymbol{x}_j^\top \boldsymbol{\beta})\}] = \mu_j(1 - \mu_j)$ 同時被未知參數 $\boldsymbol{\beta}$ 與樣本點 $\boldsymbol{x}_j$ 所決定。訊息矩陣之反矩陣為 $\widehat{\boldsymbol{\beta}}$ 之漸近變異數-共變異數矩陣 (AVAR 矩陣, Asymptotic VARiance-covariance matrix)，即 $\text{AVAR}(\widehat{\boldsymbol{\beta}}) = I(\mathcal{A}; \boldsymbol{\beta})^{-1}$。

在子抽樣問題中，當給定子樣本數 $n \ll N$ 作為主動式學習的停止條件時，我們希望取出子樣本之參數估計能越精確越好。在這邊我們令 n 為事先給定的數值，由研究預算和對反應變數予以標記之成本所決定。對子樣本之指標集 $\xi = \{j_1, \ldots, j_n\} \subset \mathcal{A}$ 而言，其訊息矩陣為

$$I(\xi; \boldsymbol{\beta}) = \sum_{j \in \xi} w_j \boldsymbol{x}_j \boldsymbol{x}_j^\top \text{。} \tag{3}$$

由於 MLE 的一致性 (consistency)，我們在考量子樣本 ξ 所對應之參數估計的精確程度時，通常會忽略偏誤 (bias) 而只考慮變異 (variance)：其 AVAR 矩陣越小越好，即訊息矩陣越大越好。在此我們引入最適設計理論中用來比較矩陣大小之準則 (如

見 Pukelsheim, 2006)，而最常見的兩個準則為被應用在 GATE 法中的 D-最適準則及被應用在 MMSE 法中的 A-最適準則。其中 D-最適準則為最小化 AVAR 矩陣的行列式值，等價於最大化訊息矩陣之行列式值 $\max_{\xi} |I(\xi;\boldsymbol{\beta})|$，對應到在任意給定信心水準下參數估計之 p 維信賴超橢球之體積達到最小；而 A-最適準則最小化 AVAR 矩陣之跡數 (trace)，即 $\min_{\xi} \mathrm{tr}\{I(\xi;\boldsymbol{\beta})^{-1}\}$，對應到各別參數估計之變異數的總和為最小。除了使用事先決定的子樣本數 n 作為停止條件外，在實務上也可使用訊息矩陣的大小或其增量達到事先給定的閾值作為停止條件。

對邏輯斯迴歸的總樣本訊息矩陣(2)及其子樣本訊息矩陣(3)中，影響係數 w_j 會被未知參數 $\boldsymbol{\beta}$ 所影響。因此在研究開始時我們需要有 $\boldsymbol{\beta}$ 的起始資訊來幫助我們尋找最佳子抽樣方法。常見的起始資訊有兩類：一類是根據過去對 $\boldsymbol{\beta}$ 了解而給出 $\boldsymbol{\beta}$ 的先驗分佈 (prior distribution) 當作起始資訊 (如 Deng *et al.*, 2009)。本文所考慮的情境為另一類：先有一個起始子樣本用來計算 $\boldsymbol{\beta}$ 的起始估計 $\widehat{\boldsymbol{\beta}}_0$ (如 Wang, Zhu, and Ma, 2018; Hsu, Chang, and Chen, 2019)。在文獻中或實務上常見的起始子樣本有以下 4 種：

(i) 最常見的方式是對全部 N 個樣本點用均勻抽樣取出 n_0 個樣本點當作起始子樣本。

(ii) 選用最具有代表性的 n_0 個樣本點當作起始子樣本，如 Mak and Joseph (2018) 提出之 support points 法。

(iii) 若相信各個樣本點的反應機率與樣本點的群聚現象有關，則可先用非監督式學習 (unsupervised learning) 中的分群法將樣本點分群再挑選各群中可能的重要樣本點 (詳見 Yuan *et al.*, 2011)。

(iv) 在一些實際應用中，起始子樣本由該領域專家主觀挑選得來。

有了起始估計之後，我們便可評估每個樣本點對參數估計之重要性並決定後續的 $n_1 = n - n_0$ 個樣本點。

綜上所述，本文考慮的主動式學習情境如下。一開始我們已知所有樣本點之解釋變數 $\{\boldsymbol{x}_j : j \in \mathcal{A}\}$ 並給定一起始子樣本指標集 $\xi_0 = \{j_{01}, \ldots, j_{0n_0}\} \subset \mathcal{A}$。$\xi_0$ 所對應之反應變數 $\{y_j : j \in \xi_0\}$ 為已標記，而其補集 $\xi_0^c = \mathcal{A} \backslash \xi_0$ 所對應之反應變數 $\{y_j : j \in \xi_0^c\}$ 為未標記。接著我們抽出後續的 $n_1 = n - n_0$ 個樣本點，並標記這些樣本點的反應變數。以下我們將聚焦於給定起始資訊之下，後續子抽樣演算法對估計參數效率之影響。此外，由於我們假設研究成本允許標記 n 個樣本點的反應變數，我們不考慮抽取重複的樣本點以避免浪費 n 次標記的機會。

Algorithm 1: GATE 法 (Hsu, Chang, and Chen, 2019)

Input: 已標記之起始樣本 $\{(y_j, \boldsymbol{x}_j) : j \in \xi_0\}$；未標記樣本 $\{\boldsymbol{x}_j : j \in \xi_0^c\}$；後續子樣本數 $n_1 = n - n_0$；決策邊界之反應機率閾值 α；每一步的候選點數 K。

Output: 最終參數估計$\widehat{\boldsymbol{\beta}}$。

1: 令 $i = 0$；根據已標記之起始樣本 $\{(y_j, \boldsymbol{x}_j) : j \in \xi_0\}$ 來計算參數估計 $\widehat{\boldsymbol{\beta}}_0$。

2: 令 $\xi_i^c = \mathcal{A} \backslash \xi_i$；對 $j \in \xi_i^c$，計算剩餘樣本點 $\boldsymbol{x}_j$ 之反應機率估計 $\hat{\mu}_i(\boldsymbol{x}_j) = F(\boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_i)$。

3: 令候選點指標集 $\widetilde{\xi}_i \subset \xi_i^c$ 為包含 $|\hat{\mu}_i(\boldsymbol{x}) - \alpha|$ 之值達到最小的 K 個樣本點之指標。

4: 令下一個要標記的樣本點指標為

$$j_{i+1} = \arg\max_{j \in \widetilde{\xi}_i} \left| I(\xi_i, \widehat{\boldsymbol{\beta}}_i) + \hat{\mu}_i(\boldsymbol{x}_j)\{1 - \hat{\mu}_i(\boldsymbol{x}_j)\}\boldsymbol{x}_j \boldsymbol{x}_j^\top \right| 。 \tag{4}$$

5: 標記 $y_{j_{i+1}}$；令 $\xi_{i+1} = \xi_i \cup \{j_{i+1}\}$。

6: 根據目前已標記資料 $\{(y_j, \boldsymbol{x}_j) : j \in \xi_{i+1}\}$ 來計算參數估計 $\widehat{\boldsymbol{\beta}}_{i+1}$；更新 $i \leftarrow i + 1$。

7: 重複步驟 2 到 6 直到 $i = n_1$；令最終參數估計 $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{n_1}$。

2.2 GATE 法 (子抽樣的部分)

根據本文之研究情境，我們先將 GATE 法 (詳見 Hsu, Chang, and Chen, 2019, Algorithm 1) 中子抽樣的部分整理如本文 Algorithm 1。我們可以看到其步驟 1 中利用目前的標記資料計算出 $\widehat{\boldsymbol{\beta}}_i$ 之後，在步驟 3 - 4 尋找下一個子樣本點的時候不需要這些候選樣本點的標記資訊，因此可用於本文的研究架構中。GATE 法在步驟 4 中對每個候選點 $\boldsymbol{x}_j$ 評估其對參數估計之有效性的準則即為 D-最適準則，將目前的子樣本加入候選的 $\boldsymbol{x}_j$ 之後的訊息矩陣之行列式值達到最大的點加入子樣本之中，並在步驟 5 時標記其對應之反應變數。而根據 Sylvester's determinant，式(4)可進一步改寫為

$$\begin{aligned} j_{i+1} &= \arg\max_{j \in \widetilde{\xi}_i} \left| I(\xi_i, \widehat{\boldsymbol{\beta}}_i) + \hat{\mu}_i(\boldsymbol{x}_j)\{1 - \hat{\mu}_i(\boldsymbol{x}_j)\}\boldsymbol{x}_j \boldsymbol{x}_j^\top \right| \\ &= \arg\max_{j \in \widetilde{\xi}_i} \left| I(\xi_i, \widehat{\boldsymbol{\beta}}_i) \right| \left[1 + \hat{\mu}_i(\boldsymbol{x}_j)\{1 - \hat{\mu}_i(\boldsymbol{x}_j)\}\boldsymbol{x}_j^\top I(\xi_i, \widehat{\boldsymbol{\beta}}_i)^{-1} \boldsymbol{x}_j \right] \\ &= \arg\max_{j \in \widetilde{\xi}_i} \hat{\mu}_i(\boldsymbol{x}_j)\{1 - \hat{\mu}_i(\boldsymbol{x}_j)\}\boldsymbol{x}_j^\top I(\xi_i, \widehat{\boldsymbol{\beta}}_i)^{-1} \boldsymbol{x}_j 。 \end{aligned} \tag{5}$$

此外，式(5)也是式(4)之對數值 (logarithm) 對新增點 $\boldsymbol{x}_j$ 之方向導數 (詳見附錄 1)。由於計算反矩陣與行列式的計算複雜度皆為 $O(p^{2.373})$，使用式(5)的計算複雜度為

$O(p^{2.373}+Kp)$，和式(4)的計算複雜度 $O(Kp^{2.373})$ 相比會稍微節省一些時間。此外，在步驟 4 和 5 若改為一次標記多個點可以減少計算時間，但會稍微降低參數估計效率。

需特別注意的是步驟 2 - 6 的迭代過程中，候選指標集不是剩下尚未標記的 ξ_i^c，而僅為其部份子集 $\widetilde{\xi}_i$（見步驟 3）。其主要原因為 (i) 時間考量：先藉由步驟 3 將候選點個數由 $N-(n_0+i)$ 降至 K 可以有效減少步驟 4 的計算時間。(ii) 穩健性考量：當真實模型並非邏輯斯模型，但真實的決策邊界 (decision boundary，為樣本空間中反應機率為 α 的點所構成；α 由使用者根據實際問題所決定，通常設為 0.5) 仍為 $\boldsymbol{x}$ 的線性函數時，盡量選取在決策邊界附近的樣本點可以如同隨機近似法 (stochastic approximation，見 Ying and Wu, 1997) 一般，有效降低決策邊界估計的偏誤。

當假設邏輯斯迴歸模型為真，且計算時間不是主要考量時，在步驟 3 中令 $\widetilde{\xi}_i=\xi_i^c$ 或許會是對參數估計來說最有效率的做法。但若總樣本數 N 極大且需要考量計算時間時，在下一節我們希望能利用反應機率估計 $\hat{\mu}_i(\boldsymbol{x})$ 來大致選出對參數估計有較佳效率的 K 個候選點。

2.3 GATE-2 法

在前一節中，我們介紹了 GATE 法為了清楚刻畫出決策邊界，所以在反應機率為閾值 α 附近挑選會使 D-最適準則較佳的點。但在邏輯斯迴歸的最適設計理論中，Ford, Torsney, and Wu (1992) 證明了當只有一個解釋變數時，D-最適設計為在 q^*- 和 $(1-q^*)$-分位數各做一半次數的實驗，其中 $q^*\approx 0.824$。而當有多個非負解釋變數時，Kabera, Haines, and Ndlovu (2015) 及 Huang, Huang, and Lin (2020) 的研究結果顯示 D-最適設計的設計點仍符合反應機率為 q^* 及 $1-q^*$ 的形式，而此時 $q^*\approx 0.8$ 且和解釋變數個數 p 以及截距項參數有關。

當假設邏輯斯迴歸模型為真，子抽樣的目標是對模型參數做統計推論，且希望節省計算時間這些條件下，我們將上述 D-最適設計的理論結果整合進 GATE 法中，提出 GATE-2 法：不使用決策邊界附近樣本點，而改採用反應機率接近 0.2 和 0.8 的樣本點。其演算法與 Algorithm 1 大致相同，主要將步驟 3 調整為下述的 3'，並將步驟 4 中的 $\widetilde{\xi}_i$ 代換為 $\check{\xi}_i$：

3': 令候選點指標集 $\check{\xi}_i\subset\xi_i^c$ 為包含 $|\hat{\mu}_i(\boldsymbol{x})-0.8|$ 之值達到最小的 $K/2$ 個樣本點與 $|\hat{\mu}_i(\boldsymbol{x})-0.2|$ 之值達到最小的 $K/2$ 個樣本點之指標。

Algorithm 2: MMSE 法 (Wang, Zhu, and Ma, 2018)

Input: 已標記之總樣本 $\{(y_j, \boldsymbol{x}_j) : j \in \mathcal{A}\}$；起始子抽樣權重 $\boldsymbol{\pi}_0 = \{\pi_{0j} : j \in \mathcal{A}\}$；起始子樣本數 n_0；後續子樣本數 $n_1 = n - n_0$。

Output: 最終參數估計 $\widetilde{\boldsymbol{\beta}}$。

1: 根據起始子抽樣權重 $\boldsymbol{\pi}_0$ 以重複選取的方式抽出 n_0 個指標 $\xi_0 = \{j_1, \ldots, j_{n_0} \in \mathcal{A}\}$。

2: 根據起始子樣本及其權重資料 $\{(y_j, \boldsymbol{x}_j, \pi_{0j}) : j \in \xi_0\}$，使用 WMLE 計算參數估計 $\widetilde{\boldsymbol{\beta}}_0$。

3: 對 $j \in \mathcal{A}$，計算所有樣本點之後續抽樣權重 $\boldsymbol{\pi}^* = \{\pi_j^* : j \in \mathcal{A}\}$:

$$0 < \pi_j^* \propto |y_j - F(\boldsymbol{x}_j^\top \widetilde{\boldsymbol{\beta}}_0)| \| I(\mathcal{A}; \widetilde{\boldsymbol{\beta}}_0)^{-1} \boldsymbol{x}_j \| \,, \tag{6}$$

其中 $\sum_{j \in \mathcal{A}} \pi_j^* = 1$ 且 $\|\boldsymbol{v}\| = (\boldsymbol{v}^\top \boldsymbol{v})^{1/2}$。

4: 根據後續子抽樣權重 $\boldsymbol{\pi}^*$ 以重複選取的方式抽出 n_1 個指標 $\xi_1 = \{j_1, \ldots, j_{n_1} \in \mathcal{A}\}$。

5: 根據起始及後續子樣本資料及對應的權重 $\{(y_j, \boldsymbol{x}_j, \pi_{0j}) : j \in \xi_0\}$ 及 $\{(y_j, \boldsymbol{x}_j, \pi_j^*) : j \in \xi_1\}$，使用 WMLE 計算參數估計 $\widetilde{\boldsymbol{\beta}}$。

2.4 MMSE 法

Wang, Zhu, and Ma (2018) 提出的 MMSE 法雖然是考慮最佳子抽樣問題，但該文之研究架構與本文有諸多不同之處：其參數估計是使用加權最大概似估計法 (WMLE, Weighted MLE)，其抽樣方式為重複抽樣，而最重要的不同之處在於在該文中全部的反應變數都已標記，使用子抽樣來估計參數主要是考量計算時間。即便如此，該文仍有許多可借鑒之處。我們會先簡介 MMSE 法 (見 Algorithm 2) 並討論其細節，並在下一節對其調整以適用於本文的研究架構。

在 MMSE 法中參數估計 $\widetilde{\boldsymbol{\beta}}$ 由 WMLE 計算得出，其對參數 $\boldsymbol{\beta}$ 仍具有一致性及漸進常態性 (asymptotic normality)，細節見 Wang, Zhu, and Ma (2018)。然而，在一些應用問題中，起始子樣本之抽樣權重可能是未知的，例如起始子樣本由該領域專家主觀決定，此時就無法使用 WMLE 估計參數。此外，重複抽樣會導致可能抽出相同的樣本點，造成浪費一些標記的機會。在 MMSE 關鍵的步驟 3 中，最佳子抽樣機率 $\boldsymbol{\pi}^* = \{\pi_j^* : j \in \mathcal{A}\}$ 來自對 WMLE 下參數估計的 AVAR 矩陣之跡數求最小值：$\boldsymbol{\pi}^* = \underset{\{\pi_j > 0, \sum_{j \in \mathcal{A}} \pi_j = 1\}}{\arg\min} \operatorname{tr}\{\mathrm{AVAR}(\widetilde{\boldsymbol{\beta}})\}$，其中

$$\mathrm{AVAR}(\widetilde{\boldsymbol{\beta}}) \propto I(\mathcal{A}; \boldsymbol{\beta})^{-1} \left[\sum_{j \in \mathcal{A}} \frac{\{y_j - \mu(\boldsymbol{x}_j)\}^2 \boldsymbol{x}_j \boldsymbol{x}_j^\top}{\pi_j} \right] I(\mathcal{A}; \boldsymbol{\beta})^{-1} \,。$$

接著利用 Cauchy-Schwarz 不等式得出最佳子抽樣機率，將未知參數 $\boldsymbol{\beta}$ 代換為起始估計 $\widetilde{\boldsymbol{\beta}}_0$ 後即得到(6)式。由於該式中需要所有標記的資訊，而無法直接應用於本文的實驗架構之中。不過在 MMSE 法的概念中，權重 $\boldsymbol{\pi}^*$ 可被視為樣本點對參數估計的重要性，且一次抽完所有的後續樣本能大幅節省計算時間。因此在下一節中我們提出 MMSE 法的改編版以適應主動式學習的資料型態。

除了 MMSE 法之外，Wang, Zhu, and Ma (2018) 提出了另一 mVc 法，其主要差異在於將步驟 3 中的子抽樣機率(6)式替換為

$$\pi_j^{\mathrm{mVc}} \propto |y_j - F(\boldsymbol{x}_j^\top \widetilde{\boldsymbol{\beta}}_0)| \|\boldsymbol{x}_j\| 。$$

此法較之 MMSE 有較快的計算速度但其參數估計效率較差。在以下的討論中我們將不考慮 mVc 法。

2.5 MEMSE 法與 SMEMSE 法

使用 MMSE 法時，子抽樣機率 π_j^* 越高的樣本點越容易被抽中，因此我們可將 π_j^* 視為對應之樣本點的重要程度。當反應變數未知時，我們對 $E[\mathrm{tr}\{\mathrm{AVAR}(\widetilde{\boldsymbol{\beta}})\}]$ 求最小值，得到該最小值發生時抽樣權重應為 (詳見附錄 2)

$$\pi_j^E \propto \sqrt{\mu(\boldsymbol{x}_j)\{1-\mu(\boldsymbol{x}_j)\}} \left\| I(\mathcal{A};\boldsymbol{\beta})^{-1}\boldsymbol{x}_j \right\| , \tag{7}$$

且與反應變數標記 y_j 無關。

Algorithm 3: MEMSE 法

Input: 總樣本 $\{\boldsymbol{x}_j : j \in \mathcal{A}\}$；起始子樣本指標集 $\xi_0 \subset \mathcal{A}$；起始子樣本之標記資料 $\{y_j : j \in \xi_0\}$；後續子樣本數 $n_1 = n - n_0$。

Output: 後續樣本點指標 $\{j_1, \ldots, j_{n_1}\} \subset \mathcal{A}\backslash\xi_0$；最終參數估計 $\widehat{\boldsymbol{\beta}}$。

1: 根據目前已標記資料 $\{(y_j, \boldsymbol{x}_j) : j \in \xi_0\}$ 來計算參數估計 $\widehat{\boldsymbol{\beta}}_0$。

2: 令 $\xi_0^c = \mathcal{A}\backslash\xi_0$；對 $j \in \xi_0^c$，計算剩餘樣本點 $\boldsymbol{x}_j$ 之重要性估計

$$\hat{\pi}_j^E \propto \sqrt{\hat{\mu}_0(\boldsymbol{x}_j)\{1-\hat{\mu}_0(\boldsymbol{x}_j)\}} \left\| I(\mathcal{A};\widehat{\boldsymbol{\beta}}_0)^{-1}\boldsymbol{x}_j \right\| ,$$

其中 $\hat{\mu}_0(\boldsymbol{x}_j) = F(\boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_0)$。

3: 從 ξ_0^c 中選取 $\hat{\pi}_j^E$ 最大的 n_1 個樣本點之指標 $\xi_1 = \{j_1, \ldots, j_{n_1}\}$ 並予以標記。

4: 根據所有已標記資料 $\{(y_j, \boldsymbol{x}_j) : j \in \xi_0 \cup \xi_1\}$ 來計算參數估計 $\widehat{\boldsymbol{\beta}}$。

在本文的主動式學習架構下，起始子樣本為給定且其子抽樣權重可能未知，因此我們將使用一般的無加權 MLE 來估計參數。然而，我們仍將式(7)之估計值視為每個樣本點的重要性。若仍想保有一次抽出後續子樣本的便利性，可使用本文所提出的 MEMSE (Minimum Expected MSE) 法，詳見 Algorithm 3。由於在 MEMSE 中的後續子抽樣權重需根據起始參數估計 $\widehat{\boldsymbol{\beta}}_0$ 來計算，若要避免起始參數估計誤差太大所造成後續子抽樣對參數估計的效率不佳，也可考慮如同 GATE-2 法逐次加入一個樣本點，並每次僅考慮反應機率估計值最接近 0.2 和 0.8 的 K 個設計點來節省計算時間 (如同 2.3 節之步驟 3' 之考量)。我們將上述演算法稱之為 SMEMSE (Stepwise MEMSE) 並詳述於 Algorithm 4 。值得注意的是因為我們使用一般的無加權 MLE 而非 Wang, Zhu, and Ma (2018) 使用的 WMLE，在後續子抽樣時我們直接挑選權重 $\hat{\pi}_j^E$ 較大的樣本點，如 Algorithm 3 步驟 3 及 Algorithm 4 步驟 5，而不使用 $\hat{\pi}_j^E$ 做為權重進行加權抽樣 (如 Algorithm 2 步驟 4)。根據我們進行模擬實驗時的經驗，前者會有較高的參數估計效率。

Algorithm 4: SMEMSE 法

Input: 總樣本 $\{\boldsymbol{x}_j : j \in \mathcal{A}\}$；起始子樣本指標集 $\xi_0 \subset \mathcal{A}$；起始子樣本之標記資料 $\{y_j : j \in \xi_0\}$；後續子樣本數 $n_1 = n - n_0$；每一步的候選點數 K。

Output: 後續樣本點指標 $\{j_1, \ldots, j_{n_1}\} \subset \mathcal{A} \backslash \xi_0$；最終參數估計 $\widehat{\boldsymbol{\beta}}$。

1: 令 $i = 0$；根據目前已標記資料 $\{(y_j, \boldsymbol{x}_j) : j \in \xi_i\}$ 來計算參數估計 $\widehat{\boldsymbol{\beta}}_i$。

2: 令 $\xi_i^c = \mathcal{A} \backslash \xi_i$；對 $j \in \xi_i^c$，計算剩餘樣本點 $\boldsymbol{x}_j$ 之反應機率估計 $\hat{\mu}_i(\boldsymbol{x}_j) = F(\boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_i)$。

3: 令候選點指標集 $\check{\xi}_i \subset \xi_i^c$ 為包含 $|\hat{\mu}_i(\boldsymbol{x}) - 0.8|$ 之值達到最小的 $K/2$ 個樣本點與 $|\hat{\mu}_i(\boldsymbol{x}) - 0.2|$ 之值達到最小的 $K/2$ 個樣本點之指標。

4: 計算候選指標 $j \in \check{\xi}_i$ 之重要性估計

$$\hat{\pi}_j^E \propto \sqrt{\hat{\mu}_i(\boldsymbol{x}_j)\{1 - \hat{\mu}_i(\boldsymbol{x}_j)\}} \left\| I(\mathcal{A}; \widehat{\boldsymbol{\beta}}_i)^{-1} \boldsymbol{x}_j \right\| 。$$

5: 令 $j_{i+1} = \arg\max_{j \in \check{\xi}_i} \hat{\pi}_j^E$。

6: 標記 $y_{j_{i+1}}$；令 $\xi_{i+1} = \xi_i \cup \{j_{i+1}\}$。

7: 根據目前已標記資料 $\{(y_j, \boldsymbol{x}_j) : j \in \xi_{i+1}\}$ 來計算參數估計 $\widehat{\boldsymbol{\beta}}_{i+1}$；更新 $i \leftarrow i + 1$。

8: 重複步驟 2 到 7 直到 $i = n_1$；令最終參數估計 $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{n_1}$。

3. 模擬實驗

在本節的模擬實驗中，我們生成資料 $(Y_j, \boldsymbol{X}_j)_{j=1}^N$ 的模型為

$$Y \mid \boldsymbol{X} = \boldsymbol{x} \sim Ber\left[\{1 + \exp(-\boldsymbol{x}^\top \boldsymbol{\beta})\}^{-1}\right] ,$$

其中總樣本數 $N = 10^5$，解釋變數 $\boldsymbol{X} = (1, X_{[1]}, \ldots, X_{[9]})^\top \in \mathbb{R}^{10}$，其聯合分佈在各個模擬實驗中有不同的設定，$\boldsymbol{\beta} = (0, 0.5, 0.5, \ldots, 0.5)^\top$。起始子樣本 ξ_0 為從 N 筆資料中隨機抽出 $n_0 = 200$ 筆資料，並令後續子樣本數 $n_1 = n - n_0 = 1800$。模擬實驗的總次數設為 1000。在每次的模擬實驗中，我們將考慮以下的抽樣方法並計算其參數估計 $\widehat{\boldsymbol{\beta}}$：

[A] UNIFORM: 以均勻抽樣從剩餘樣本中不重複地抽出後續子樣本。

[B] GATE: 使用 GATE 法進行後續子抽樣，如 Algorithm 1，其中 $\alpha = 0.5$ 且 $K = 500$。

[C] GATE-2: 使用 GATE-2 法進行後續子抽樣，如 Algorithm 1，其中 $K = 500$，但將其步驟 3 改為 2.3 節的步驟 3'。

[D] GATE-0: 使用 GATE 法進行後續子抽樣，如 Algorithm 1，但將步驟 3 的候選指標集令為剩餘指標集，即 $\widetilde{\xi}_i = \xi_i^c$。和 GATE 及 GATE-2 相比，此方法使用更大的候選指標集並同樣使用 D-最適準則挑樣本點。因此預期會耗用較長的計算時間，但參數估計應更為準確。

[E] MEMSE: 使用 MEMSE 法進行後續子抽樣，如 Algorithm 3。

[F] SMEMSE: 使用 SMEMSE 法進行後續子抽樣，如 Algorithm 4，其中 $K = 500$。

[G] SMEMSE-0: 使用 SMEMSE 法進行後續子抽樣，如 Algorithm 4，但步驟 3 的候選指標集令為剩餘指標集，即 $\widetilde{\xi}_i = \xi_i^c$。和 SMEMSE 相比，此方法用更大的候選指標集並同樣使用 A-最適準則挑樣本點。因此預期會耗用較長計算時間，但參數估計應更為準確。

[T] TOTAL: 使用全部樣本估計參數。此結果將作為比較基準。

我們將使用以下兩個指標來評估參數估計的相對效率。對方法 $m = \mathrm{A}, \mathrm{B}, \ldots, \mathrm{G}$，其 A-效率 (Aeff) 與 D-效率 (Deff) 之估計為：

$$\widehat{\mathrm{Aeff}}_m = \frac{\widehat{\mathrm{AMSE}}_{\mathrm{T}}}{\widehat{\mathrm{AMSE}}_m} \times \frac{N}{n} \quad \text{以及} \quad \widehat{\mathrm{Deff}}_m = \frac{\widehat{\mathrm{DMSE}}_{\mathrm{T}}}{\widehat{\mathrm{DMSE}}_m} \times \frac{N}{n} , \tag{8}$$

其中對 $m = \mathrm{A}, \mathrm{B}, \ldots, \mathrm{G}, \mathrm{T}$,

$$\widehat{\mathrm{AMSE}}_m = \frac{1}{10}\mathrm{tr}\left\{\widehat{\mathrm{MSE}}_m\right\} \quad , \quad \widehat{\mathrm{DMSE}}_m = \left|\widehat{\mathrm{MSE}}_m\right|^{1/10} \ ,$$

且 $\widehat{\mathrm{MSE}}_m \in \mathbb{R}^{10\times 10}$ 為均方誤差矩陣之估計，

$$\widehat{\mathrm{MSE}}_m = \frac{1}{1000}\sum_{j=1}^{1000}\left(\widehat{\boldsymbol{\beta}}_{m,j} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}_{m,j} - \boldsymbol{\beta}\right)^{\top} \ , \tag{9}$$

j 表示為第 j 次模擬實驗結果。對方法 $m = \mathrm{A}, \mathrm{B}, \ldots, \mathrm{G}$ 而言，這兩種效率的評估方式都是效率值越大表示參數估計越精準。注意到這邊 $\widehat{\mathrm{MSE}}_m$ 可視為方法 m 之訊息矩陣之反矩陣的估計。因此式(8)可視為平均相對效率

$$\mathrm{Aeff}_m = \frac{E\left(\frac{1}{10}\mathrm{tr}[\{I(\mathcal{A};\boldsymbol{\beta})/N\}^{-1}]\right)}{E\left(\frac{1}{10}\mathrm{tr}[\{I(\xi_m;\boldsymbol{\beta})/n\}^{-1}]\right)} \quad \text{以及} \quad \mathrm{Deff}_m = \frac{E\left|I(\mathcal{A};\boldsymbol{\beta})/N\right|^{-1/10}}{E\left|I(\xi_m;\boldsymbol{\beta})/n\right|^{-1/10}}$$

之估計，其中 $\mathcal{A}$ 為隨機產生之總樣本，ξ_m 為方法 m 所抽出之子樣本。效率接近 1 表示該方法所抽出的 n 個樣本，平均而言每個樣本提供的資訊量與使用全部 N 個樣本時每個樣本的平均資訊量相近。雖然隨著子樣本數的增加，式(3)之訊息矩陣會含有越來越多的資訊，參數估計也會越來越精確；但若選到資訊量較少的樣本點，除以子樣本數的平均相對效率可能不見得是遞增的。

3.1 模擬實驗 1

在模擬實驗 1 中，我們使用獨立且同分佈之標準常態分佈來生成解釋變數 $X_{[1]}, \ldots, X_{[9]}$。這些方法的一些質性特徵與模擬實驗結果見表 1。很明顯地，只需要 1 次後續子抽樣的方法 UNIFORM 和 MEMSE 的計算速度最快；在其他每次抽出 1 個樣本點的方法中，若能限制每次抽樣的候選點個數，如 GATE、GATE-2 和 SMEMSE，相較於每次搜尋所有未標記樣本的 GATE-0 和 SMEMSE-0，也能大幅降低計算時間。

接著我們比較各個子抽樣方法估計參數的效率。顯然地，均勻抽樣 UNIFORM 的 A- 及 D-效率皆接近 1。若為了兼顧穩健性而使用 GATE，在設定 $\alpha = 0.5$ 時參數估計之 D-效率跟均勻抽樣差不多，但 A-效率僅約均勻抽樣的 1/5。相較之下，若採用 GATE-2 或是 SMEMSE，參數估計的 A-效率及 D-效率相比 UNIFORM 均增加了 50% 以上。而需要耗用大量時間的 GATE-0 和 SMEMSE-0 有最高的 D-效率，約為

表 1: 各子抽樣方法之質性特徵與參數估計效率比較。

方法	最佳性準則	後續子抽樣次數	每次抽樣搜尋範圍	平均計算時間 *	$\widehat{\text{Aeff}}$	$\widehat{\text{Deff}}$
UNIFORM	無	1	全部未標記樣本	< 0.1 秒	0.963	0.963
GATE	D	1800	反應機率最靠近 0.5 的 500 個未標記樣本	173.4 秒	0.186	1.003
GATE-2	D	1800	反應機率最靠近 0.2 和 0.8 的各 250 個未標記樣本	185.6 秒	1.513	1.612
GATE-0	D	1800	全部未標記樣本	928.9 秒	1.247	2.111
MEMSE	A	1	全部未標記樣本	0.5 秒	0.803	1.322
SMEMSE	A	1800	反應機率最靠近 0.2 和 0.8 的各 250 個未標記樣本	181.9 秒	1.557	1.648
SMEMSE-0	A	1800	全部未標記樣本	1039.9 秒	1.739	2.072

* 計算環境：R (4.2.2)；Intel Xeon Gold 5218 CPU @ 2.30 GHz ×2；384GB RAM；每次 50 個模擬實驗同時平行運算。

均勻抽樣的兩倍，其中 SMEMSE-0 的 A-效率 (約 1.8) 又較 GATE-0 的 A-效率 (約 1.3) 來得更高。當使用 MEMSE 法一次抽完所有子樣本時，參數估計的效率相當依賴起始子樣本，平均而言會有比均勻抽樣稍高的 D-效率 (約 1.3) 以及稍低的 A-效率 (約 0.8)。

3.2 模擬實驗 2

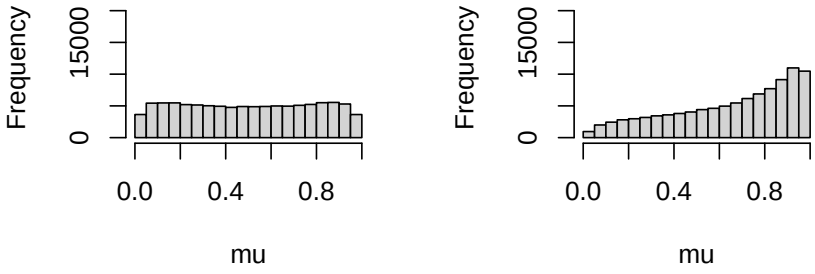
在模擬實驗 2 中，我們將比較幾個計算速度較快的子抽樣方法在：UNIFORM，GATE，GATE-2，MEMSE 及 SMEMSE 在不同的解釋變數分佈與不同的後續子樣本數下的參數估計效率表現。模擬實驗的模型與參數設定大致相同於模擬實驗 1，但後續子樣本數考慮 $n_1 \in \{800, 1000, \ldots, 1800\}$ 這 6 個情況，而解釋變數的生成方式改為考慮下列 4 種：

情境 1 ：$\widetilde{\boldsymbol{X}} = (X_{[1]}, \ldots, X_{[9]})^\top \sim N_9(0, \boldsymbol{\Sigma})$，其中 $\boldsymbol{\Sigma} = 0.3\boldsymbol{I}_9 + (0.1)\mathbf{1}_9\mathbf{1}_9^\top$，$\boldsymbol{I} \in \mathbb{R}^{9\times 9}$ 為單位矩陣，$\mathbf{1}_9 \in \mathbb{R}^9$ 為全為 1 之向量。

情境 2 ：$\widetilde{\boldsymbol{X}} \sim N_9((0.2)\mathbf{1}_9, \boldsymbol{\Sigma})$。

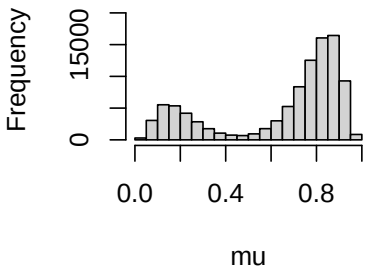
情境 3 ：$\widetilde{\boldsymbol{X}} \sim N_9(\frac{1}{3}(2B-1)\mathbf{1}_9, (0.15)\boldsymbol{\Sigma})$，其中 $B \sim Ber(0.75)$ 。

情境 4 ：$X_{[1]}, \ldots, X_{[9]} \overset{iid}{\sim} \Gamma(0.15, 1)$。

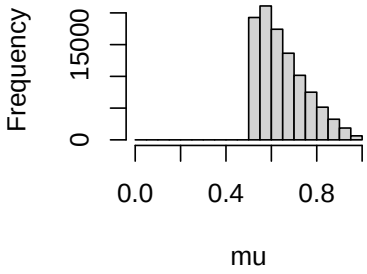


(a) 情境 1

(b) 情境 2



(c) 情境 3

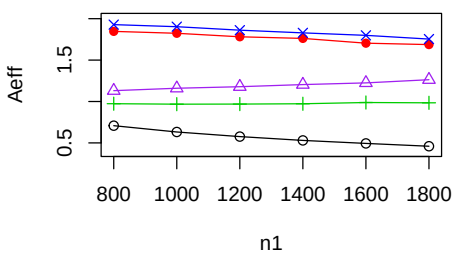


(d) 情境 4

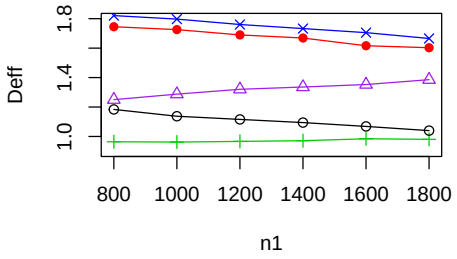
圖 1: 模擬實驗 2 中各情境生成資料反應機率 $\{\mu_j\}_{j=1}^N$ 之直方圖。

各情境生成資料反應機率 $\{\mu_j\}_{j=1}^N$ 之直方圖見圖 1。情境 1 與模擬實驗 1 之 $\{\mu_j\}_{j=1}^N$ 之分佈類似，但情境 1 與模擬實驗 1 不同之處在於各個解釋變數之間具有相關性。在情境 2–4 中我們考慮 3 種不同的不均衡資料生成方式。在固定 Y 之邊際分佈同為 $P(Y = 1) = 0.65$ 的情況下，我們將比較在情境 2–4 中不同分佈之解釋變數對各個子抽樣法表現之影響，其中在情境 2 中反應機率 μ 之密度函數大致隨著 μ 值增加，情境 3 中 μ 之分佈為雙峰，而在情境 4 中 μ 皆大於 0.5。

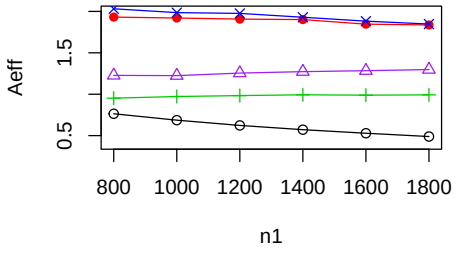
由圖 2 以及表 1 中可看出，在模擬實驗 1 及模擬實驗 2 的這些情境中 SMEMSE 跟 GATE-2 都有最佳的參數估計效率，而在情境 4 這個反應機率皆大於 0.5 的極端情況中，SMEMSE 又明顯較 GATE-2 的效率更好。整體而言，若已知真實模型為邏輯斯迴歸且參數估計是主要考量時，使用 SMEMSE 法會有最佳的參數估計。在極端的情境 4 之中，GATE 和 MEMSE 不管從 A-效率或 D-效率來看都不如均勻抽樣。值得一提的是在情境 3，反應機率與樣本的群聚現象有關 (如見 Yuan *et al.*, 2011)，此時這 4 種子抽樣方法的表現沒有太大的差異，這時候使用 GATE 可以額外兼顧分類正確性的穩健性。



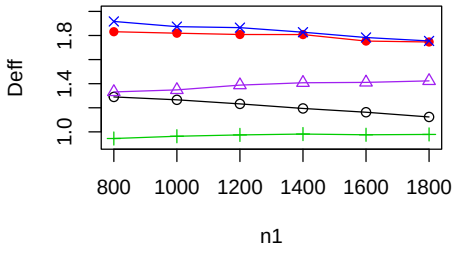
(a) 情境 1 下之 A-效率



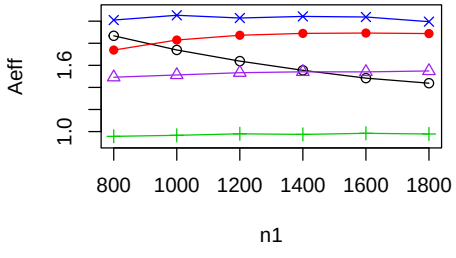
(b) 情境 1 下之 D-效率



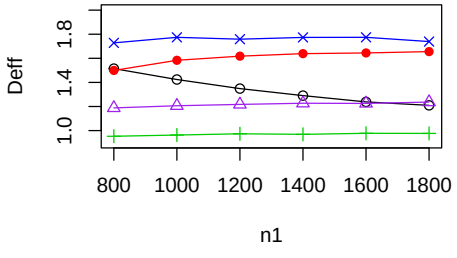
(c) 情境 2 下之 A-效率



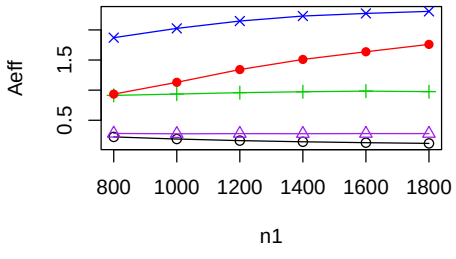
(d) 情境 2 下之 D-效率



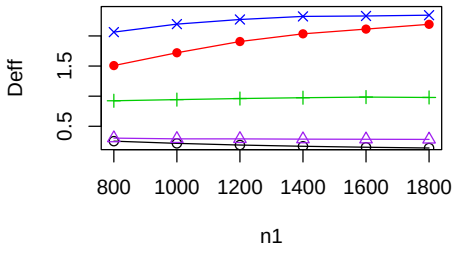
(e) 情境 3 下之 A-效率



(f) 情境 3 下之 D-效率



(g) 情境 4 下之 A-效率



(h) 情境 4 下之 D-效率

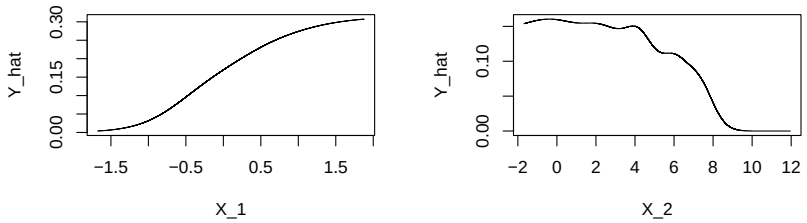
圖 2: 模擬實驗 2 中各方法在情境 1 – 4 及後續子樣本數為 800 – 1800 下之 A-效率與 D-效率。綠色為 UNIFORM，黑色為 GATE，紅色為 GATE-2，紫色為 MEMSE，藍色為 SMEMSE。

4. 實例應用

在本節中我們使用 UCI Machine Learning 資料庫中的 Adult 資料集 (Becker and Kohavi, 1996) 作為例子來比較各個子抽樣法在邏輯斯迴歸模型下參數估計的表現。原始資料集共有 48842 位居民，原本的反應變數為該居民的年收入是否超過 50000 美元。在下面的例子中我們主要關注的對象是青壯年 (年齡 $X_1 \leq 50$ 歲)，正常工作時數 (每週工作時數 $X_4 \leq 50$ 小時)，且主要收入來源主要來自薪水 (資產利得為 0 且資產損失為 0) 的居民，樣本數為 30845。除了年齡跟每週工作時數之外，其他的連續型解釋變數包括最終權重 (X_2，為美國普查局人口部門根據社會經濟特徵給出的指標) 及教育程度 X_3(年)。我們想要研究的虛擬情境如下：若想了解這些解釋變數對某個新的二元反應變數，例如是否曾經逃漏稅，之勝率比 (odds ratio) 的影響，我們可以使用邏輯斯迴歸模型的參數估計做統計推論，此時我們希望參數估計越精準越好。在短期內這 4 個連續型的解釋變數都不會改變或是容易推算，但我們需要付出額外的人力物力來標記新的反應變數。

在這個實例應用中我們使用原本的反應變數做為新反應變數之標記值，並由於各個解釋變數尺度差異過大，因此我們將這些變數標準化，使其平均為 0 且變異數為 1。首先我們利用核平滑迴歸 (kernel-smoothing regression) 觀察只考量一個解釋變數時，各個解釋變數和反應變數的關係。由圖 3 可看出每個解釋變數各自對反應變數所建立的核平滑迴歸函數大致上都是單調 (monotone) 函數或近似單調函數，所以只包含所有變數之主效應的邏輯斯迴歸模型大致上算是一個合適的模型。全樣本邏輯斯迴歸的參數估計 $\widehat{\boldsymbol{\beta}}_{\mathrm{T}}$ 如表 2。我們可以看到這些解釋變數對反應機率都有相當顯著影響。

接下來我們模擬需要額外花費才能取得樣本標記的情境。我們將比較 UNIFORM、GATE、GATE-2、MEMSE、以及 SMEMSE 的參數估計效率。假設研究預算足以讓我們標記最多 2000 個樣本，我們使用均勻抽樣取出 $n_0 = 200$ 個樣本做為起始設計，並考慮後續子樣本數為 $n_1 = 800, 1000, \ldots, 1800$ 的情況。1000 次的虛擬子抽樣實驗結果整理如圖 4。其中由於真實參數 $\boldsymbol{\beta}$ 未知，我們將參數估計效率的計算式(8)略作更動，其中做為比較基準的 $\widehat{\mathrm{MSE}}_{\mathrm{T}}$ 矩陣改為 $\widehat{\mathrm{AVAR}}(\widehat{\boldsymbol{\beta}}_{\mathrm{T}}) = I(\mathcal{A}, \widehat{\boldsymbol{\beta}}_{\mathrm{T}})^{-1}$，且用式(9)計算每個方法的 MSE 矩陣時，改用 $\widehat{\boldsymbol{\beta}}_{\mathrm{T}}$ 取代真實參數 $\boldsymbol{\beta}$。從 D-效率來看，GATE、GATE-2、以及 SMEMSE 法都有極佳的參數估計效率，而 MEMSE 的表現和 UNIFORM 差不多。另一方面，從 A-效率的角度來看，最簡單的 UNIFORM 有最佳的參數估計效率，接下來是 GATE-2 和 SMEMSE，最後是 MEMSE 和 GATE。



(a) X_1 versus $\widehat{Y}_{X_1}$ (b) X_2 versus $\widehat{Y}_{X_2}$

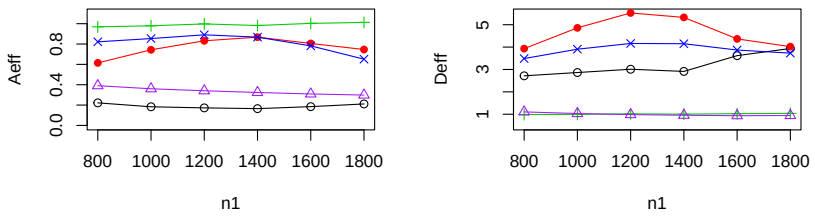
(c) X_3 versus $\widehat{Y}_{X_3}$ (d) X_4 versus $\widehat{Y}_{X_4}$

圖 3: 實例應用中各解釋變數對反應變數之核平滑迴歸函數。

表 2: 實例應用之全樣本邏輯斯迴歸模型參數估計表。

	Estimate	SE	Z value	p-value
Intercept	-2.3539	0.0254	-92.76	< 0.0001
X_1	0.8714	0.0208	41.84	< 0.0001
X_2	0.0658	0.0177	3.71	0.0002
X_3	0.7939	0.0195	40.69	< 0.0001
X_4	0.6538	0.0273	23.94	< 0.0001

註 Residual deviance 20849 on 30840 degrees of freedom (p-value = 1).



(a) 實例應用之 A-效率 (b) 實例應用之 D-效率

圖 4: 實例應用中各方法在後續子樣本數為 800 – 1800 下之 A-效率與 D-效率。綠色為 UNIFORM，黑色為 GATE，紅色為 GATE-2，紫色為 MEMSE，藍色為 SMEMSE。

表 3: 實例應用中當 $n_1 = 1800$，各子抽樣方法之各參數估計 $\widehat{\text{MSE}}$ (括號中為標準差)。

	Intercept	X_1	X_2	X_3	X_4
UNIFORM	0.0090	0.0050	0.0043	0.0061	0.0140
	(0.0141)	(0.0071)	(0.0061)	(0.0087)	(0.0197)
GATE	0.0438	0.0779	0.0032	0.0108	0.0484
	(0.0611)	(0.0308)	(0.0009)	(0.0130)	(0.0371)
GATE-2	0.0079	0.0058	0.0008	0.0011	0.0364
	(0.0130)	(0.0030)	(0.0007)	(0.0017)	(0.0240)
MEMSE	0.0492	0.0337	0.0010	0.0081	0.0383
	(0.0611)	(0.0344)	(0.0014)	(0.0113)	(0.0242)
SMEMSE	0.0054	0.0088	0.0021	0.0019	0.0414
	(0.0050)	(0.0055)	(0.0014)	(0.0014)	(0.0209)

我們進一步研究為何在此實例應用中 UNIFORM 會有比 GATE-2 和 SMEMSE 更高的 A-效率。這些方法在 $n_1 = 1800$ 時 $\widehat{\text{MSE}}$ 矩陣之對角元素值（及其標準差）見表 3。我們可以看到無論是 GATE-2 或是 SMEMSE，幾乎所有的參數估計都比 UNIFORM 更精準，僅在估計 X_4 之參數時表現較差。但由於 X_4 參數之 MSE 明顯較其他參數之 MSE 更大，導致平均而言 UNIFORM 有著更高的 A-效率。

5. 結論與討論

本文主要簡介 Hsu, Chang, and Chen (2019) 及 Wang, Zhu, and Ma (2018) 所提出的 GATE 法和 MMSE 法中子抽樣的部分，並對這兩個子抽樣演算法進行微調，使其在主動式學習的架構下對邏輯斯模型參數估計有更高的估計效率。當使用反應機率估計值 $\hat{\mu}(\boldsymbol{x}_j)$ 初步篩選有效樣本點時，我們結合實驗設計文獻的結果，建議在反應機率 0.2 和 0.8 附近尋找後續樣本點，並得到不錯的結果。在模擬實驗的各個情境中我們發現，微調後的 GATE-2 法及 SMEMSE 法均有較佳的參數估計效率，適合用在已知為邏輯斯模型的主動式學習問題上。在第 4 節中我們發現 GATE-2 及 SMEMSE 應用在此實例時有極高的 D-效率但 A-效率不甚理想。這可能是真實模型與邏輯斯迴歸模型的差異所造成，也可能是在此例資料之聯合分佈下無法同時兼顧 D-效率與 A-效

率。若要同時兼顧多種準則，可以考慮將 GATE-2 的步驟 4 改為限制型或混合型的準則函數，如 Cook and Wong (1994) 或 Atkinson (2008)。

若邏輯斯模型是真實模型或是跟真實模型差距不大，當參數能被精準估計，分類正確性雖然會與最佳結果相去不遠，但通常不會達到最佳。若目標是要達到最佳的分類正確率，應採用 c-最適準則 (見 Pukelsheim, 2006) 來挑選子樣本，而非以精準估計參數為目標的 D- 或 A-最適準則。一般來說，「分類正確性」與「精準估計參數」在實驗設計中是兩個很接近但不相同的目的。在本文「精準估計參數」背後隱含著邏輯斯迴歸必須是真實模型或近似模型的假設，否則精準估計參數這件事並無意義；另一方面，即便使用之模型與真實模型不同，分類正確性仍是一個可以被量測且有意義的量。若主要考量為分類正確性，且相信決策邊界接近解釋變數之線性函數時，使用 ALSD 法 (Deng *et al.*, 2009) 或 GATE 法 (Hsu, Chang, and Chen, 2019) 可減少錯誤模型導致之偏誤，從而得到正確率較高且較穩健之結果。

此類主動式學習的子抽樣問題仍有一些待解決的實務問題值得更進一步的研究。首先，GATE 法在決策邊界附近進行子取樣以保有隨機近似法對錯誤模型的穩健性，並以 D-最適準則挑點提升其參數估計的品質。但即便如此其參數估計效率仍時常與均勻抽樣差不多。另一方面，若不考慮穩健性，SMEMSE 或 GATE-2 都能精準地估計參數。當模型具有不確定性時，實驗設計的目標可改為最小化整體預測值之均方差，以在固定樣本下取得偏誤與變異之間的平衡。其次，實務中邏輯斯迴歸模型常包含類別型的解釋變數。當各類別的資料筆數極度不均衡時，本研究中所討論的子抽樣方法很可能會錯失某些少見類別。所以當資料包含類別型解釋變數時，需要根據各類別的樣本數來調整子抽樣法，才能精準估計各類別之虛擬變數 (dummy variables) 之參數。最後，在本研究的架構下，我們假設子樣本反應變數之標記均正確無誤。但在實務中，偽陰 (false negative) 及偽陽 (false positive) 有時很難避免。近年來對這樣含有部分錯誤標記資料的議題也逐漸引起討論，如 Hung, Jou, and Huang (2018)，Cannings, Fan, and Samworth (2020)，以及 Lee and Barber (2022)。當樣本中可能含有少部分的錯誤標記時，實驗者該如何進行子取樣，使得在提高參數估計有效性的同時降低錯標帶來的偏誤，將會是個重要的議題。

致謝

本文作者感謝主編、副主編以及一位匿名審稿委員的指正與寶貴建議，讓本文更加清楚與完整。

參考文獻

[1] Atkinson, A. C. (2008). DT-optimum designs for model discrimination and parameter estimation. *Journal of Statistical Planning and Inference*, 138(1), pages 56-64.

[2] Becker, B., and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. URL: https://doi.org/10.24432/C5XW20

[3] Cannings, T. I., Fan, Y., and Samworth, R. J. (2020). Classification with imperfect training labels. *Biometrika*, 107(2), pages 311-330.

[4] Cesa-Bianchi, N., Gentile, C., and Zaniboni, L. (2006). Worst-case analysis of selective sampling for linear classification. *Journal of Machine Learning Research*, 7, pages 1205-1230.

[5] Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1), pages 129-145.

[6] Cook, R. D., and Wong, W. K. (1994). On the equivalence of constrained and compound optimal designs. *Journal of the American Statistical Association*, 89(426), pages 687-692.

[7] Deng, X., Joseph, V. R., Sudjianto, A., and Wu, C. F. J. (2009). Active learning through sequential design, with applications to detection of money laundering. *Journal of the American Statistical Association*, 104(487), pages 969-981.

[8] Ford, I., Torsney, B., and Wu, C. F. J. (1992). The use of a canonical form in the construction of locally optimal designs for nonlinear problems. *Journal of the Royal Statistical Society: Series B*, 54(2), pages 569-583.

[9] Hsu, H.-L., Chang, Y.-C. I., and Chen, R.-B. (2019). Greedy active learning algorithm for logistic regression models. *Computational Statistics and Data Analysis*, 129, pages 119–134.

[10] Huang, S.-H., Huang, M.-N. L., and Lin, C.-W. (2020). Optimal designs for binary response models with multiple nonnegative variables. *Journal of Statistical Planning and Inference*, 206, pages 75-83.

[11] Hung, H., Jou, Z.-Y., and Huang, S.-Y. (2018). Robust mislabel logistic regression without modeling mislabel probabilities. *Biometrics*, 74(1), pages 145-154.

[12] Imberg, H., Lisovskaja, V., Selpi, S., and Nerman, O. (2022). Optimization of two-phase sampling designs with application to naturalistic driving studies. *IEEE Transactions on Intelligent Transportation Systems*, 23(4), pages 3575-3588.

[13] Kabera, G. M., Haines, L. M., and Ndlovu, P. (2015). The analytic construction of D-optimal designs for the two-variable binary logistic regression model without interaction. *Statistics*, 49(5), pages 1169-1186.

[14] Kohavi, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 96, pages 202-207.

[15] Lee, Y., and Barber, R. F. (2022). Binary classification with corrupted labels. *Electronic Journal of Statistics*, 16(1), pages 1367-1392.

[16] MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, 4(4), pages 590-604.

[17] Mak, S., and Joseph, V. R. (2018). Support points. *Annals of Statistics*, 46(6A), pages 2562-2592.

[18] Muslea, I., Minton, S., and Knoblock, C. A. (2006). Active learning with multiple views. *Journal of Artificial Intelligence Research*, 27(1), pages 203-233.

[19] Pukelsheim, F. (2006). *Optimal Design of Experiments*. SIAM, United States.

[20] Settles, B. (2009). Active learning literature survey. *Computer Sciences Technical Report 1648*, University of Wisconsin–Madison.

[21] Wang, H., Zhu, R., and Ma, P. (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522), pages 829-844.

[22] Xu, X., Liang, T., Zhu, J., Zheng, D., and Sun, T. (2019). Review of classical dimensionality reduction and sample selection methods for large-scale data processing. *Neurocomputing*, 328, pages 5-15.

[23] Ying, Z., and Wu, C. F. J. (1997). An asymptotic theory of sequential designs based on maximum likelihood recursions. *Statistica Sinica*, 7, pages 75-91.

[24] Yuan, W., Han, Y., Guan, D., Lee, S., and Lee, Y. K. (2011). Initial training data selection for active learning. *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*. Seoul Korea.

[Received December 2022; accepted September 2023.]

Optimal Subsampling Algorithms for Parameter Estimation in Logistic Regression Model

Chun-Ting Chen[1], Wei-Cheng Hsiao[2], Ming-Chung Chang[3] and Shih-Hao Huang[4,1†]

[1]Graduate Institute of Statistics, National Central University

[2]Department of Finance Engineering and Actuarial Mathematics, Soochow University

[3]Institute of Statistical Science, Academia Sinica

[4]Department of Mathematics, National Central University

ABSTRACT

In the current big data era, the investigators may easily get a lot of samples with explanatory variables, but only limited responses are labelled and to label an unlabelled response is expensive. In order to build a precise model, active learning aims to select informative unlabelled subdata for labels. In this article, we introduce two informative subsampling algorithms for large sample logistic regression in the existing literature, GATE [Comput. Stat. Data Anal. 129 (2019) 119–134] and MMSE [J. Am. Stat. Assoc. 113 (2018) 829–844], and provide variants of them for more precise parameter estimation. Simulation studies and a study on a census data show that the proposed approaches usually have greater efficiency in parameter estimation for logistic regression.

Key words and phrases: Active Learning, Logistic Regression, Optimal Design, Optimal Subsampling, Parameter Estimation.

JEL classification: C35, C61, C83, C90.

[†]Corresponding to: Shih-Hao Huang
E-mail: shhuang@math.ncu.edu.tw