Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Auxiliary-Enhanced Survey Calibration: A Novel Framework for Reject Inference under Sample Selection Bias

Abderrahim EL AMRANI

aelamrani@insea.ac.ma

Institut National de Statistique et d'Economie Appliquée

Badreddine BENYACOUB

Institut National de Statistique et d'Economie Appliquée

Mohammed EL HAJ TIRARI

Institut National de Statistique et d'Economie Appliquée

Research Article

Keywords: Sample Selection Bias, Reject Inference, Survey theory, Calibration technique, Semi-Supervised Learning, Logistic Regression

Posted Date: July 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10208992/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Auxiliary-Enhanced Survey Calibration: A Novel Framework for Reject Inference under Sample Selection Bias

Abderrahim EL AMRANI[1*], Badreddine BENYACOUB[2] and Mohammed EL HAJ TIRARI[3]

[1,2,3]Research Laboratory in Applied Methods in Statistics, Actuarial Science, Finance and Quantitative Economics, INSEA, Rabat, Morocco.

*Corresponding author(s). E-mail(s): aelamrani@insea.ac.ma;
Contributing authors: bbenyacoub@insea.ac.ma; tirari@insea.ac.ma;

Abstract

Credit scorecards are trained on accepted applicants, a non representative sample of the applicant population, which biases their predictions on the full pool. Survey statistics has treated this problem for fifty years: an accepted sample is a non-random sample of a known population, and calibration reweights it to match population totals. We bring calibration to reject inference. Existing methods instead impute the unknown outcomes of rejected applicants and retrain. We show on real data and controlled simulations that imputation degrades the scorecard, because the imputed labels come from a model that selection bias has already corrupted. Our estimator computes observation-level weights from variables recorded for all applicants and estimates no reject outcome; setting every weight to one returns the baseline, so the method cannot do worse than it. We prove the calibration-weighted estimator is consistent and $\sqrt{n}$-asymptotically normal under missing-at-random (MAR) selection and derive a sandwich variance estimator. Its advantage appears under model misspecification, where it estimates the population-optimal parameter, and in the calibration of population risk. On the Lending Club and Taiwan credit datasets and a ground-truth simulation, calibration holds discrimination at the baseline level while imputation competitors lose up to **11** AUC points; it reduces population calibration error up to three-fold and uses underwriting features that imputation methods cannot reach. We test significance with bootstrap confidence intervals and Friedman tests. We also state where the method fails: it does not recover MNAR selection, and it needs common support between accepted and rejected applicants.

Keywords: Sample Selection Bias, Reject Inference, Survey theory, Calibration technique, Semi-Supervised Learning, Logistic Regression

1 Introduction

Credit scoring is one of the most consequential applications of statistical modeling techniques. Several techniques were used in this context, such as discriminant analysis and logistic regression [1–3] . Later, machine learning models emerged and started being deployed as scorecards; these models include neural networks, support vector machines, and gradient boosting [4–7]. Unfortunately, all these models rest on a data collection process that systematically excludes rejected applications. Usually, financial institutions use information of individuals who were granted a loan and the

creditworthy status is available. Lenders normally only have performance information and repayment behaviour about accepted borrowers. By investigating this information based solely on the accepted borrowers, sample bias may arise [8, 9].

The training data is therefore not a random sample of all applicants but a filtered one, retaining those the lender judged creditworthy under its existing criteria. [8] formalised the resulting distortion as *sample selection bias*: a systematic mismatch between the distribution of the training sample and the distribution on which the model is deployed. In credit scoring this is known as the reject inference problem, and it arises precisely because performance is observed only for accepted borrowers.

The assumption behind this distorted distribution is mainly categorised as follows. A classifier trained on accepted applicants optimises the likelihood of the selected sample's distribution,$P(Y \mid X, Z = 1)$, rather than the population distribution $P(Y \mid X)$. These two probabilities quantities coincide only when the selection mechanism is uninformative (MCAR), that is, when the decision to accept or reject a loan application is independent of the applicantâ€™s true creditworthiness [10]. In practice, this condition is almost never satisfied: lenders specifically reject applicants they believe are likely to default, creating a dependence between selection mechanism and the unobserved outcome. The accepted sample is therefore systematically enriched with good payers, and a classifier trained on it will underestimate default rates in the full population. One of the most real-world cases corresponding to this situation, we have the Lending Club dataset with approximately 8% of loan applicants accepted, meaning that the standard training sample comes from a heavily filtered slice of the applicant pool.

This dataset records more than 150 variables, gathered both during the application process and over the repayment period. Many of these cannot be used for decision-making: some are unavailable at the time the model is meant to be applied, and others, added as the platform evolved, are missing for the earliest loans. Distinguishing variables known at application time from those generated by the loan's later performance is therefore essential, and we treat it explicitly in Section 5.2.

The problem of correcting for this bias is known in the credit scoring literature as *reject inference*. Drawing on the parallel with survey methodology [11], accepted applicants play the role of selected units in a probability survey, while rejected applicants are non-selected units whose outcomes are missing. Reject inference methods aim to bridge this gap and produce classifiers that are valid for the full applicant population. Existing methods fall broadly into five families: augmentation, reclassification, inverse propensity score reweighting, parcelling, and generative approaches [10]. To tackle the challenges raised by sample selection bias in the context of credit scoring, reject inference has been used to incorporate the information of the rejected applicantsâ€™ actual status by inferring the loans' outcomes of the applicants that were rejected for a loan, so that these predicted status will be integrated in the credit scorecards. However, most of the reject inference techniques focus on a labeling mechanism which treats all rejected applicants as a missing data problem by predicting their potential status and incorporating them with accepted samples. In this study, we propose a novel reject inference strategy based on calibration methods and derived from survey theory, via integrating information of accepted and rejected groups to conduct a finer-grained re-weighting of the accepted sample to reduce the sample selection bias and provide a representative sample of the full population. The main idea is to compute a set of observation-level weights for accepted applicants that allow the weighted sample to match the covariate distribution of the full population, thereby without imputing the default risk of the rejected applicants through the training process. The method is rooted in the representative estimators of [12], which have an extensive track record in survey statistics for producing efficient estimators of population parameters.

Selection mechanism and the basis for comparison. It is essential to state at the outset which missingness mechanism each method assumes. Our estimator is a *missing-at-random* (MAR) correction: it removes the selection bias that the *observed* common variables explain. Like inverse propensity weighting, it does not undo a missing-not-at-random (MNAR) mechanism that depends on the unobserved outcome. We therefore make the mechanism an *experimental factor* instead of assuming one. On a semi-synthetic benchmark

with fully observed ground truth we impose MAR, MNAR, and misspecified regimes and report each method *per regime*, so each method is judged in the regime it targets. This yields a map of when reject inference helps and when no method can.

This paper makes four main contributions.
(1) An imputation-free calibration framework for reject inference. We introduce a reject inference estimator based on survey calibration weighting that corrects selection bias without imputing any reject outcome. It fabricates no labels, so it does not inherit the accepted-only model's bias the way augmentation, reclassification, parcelling, and generative methods do. In our experiments it is the only reject inference method that holds at or above the baseline; the imputation-based competitors lose up to 11 AUC points and triple the calibration error.
(2) Formal guarantees and a population-projection characterisation. We prove that the calibration-weighted estimator is consistent and $\sqrt{n}$-asymptotically normal under MAR, derive its asymptotic variance in closed form, and establish a sandwich variance estimator. We also clarify when reject inference helps. Under a correctly specified model with selection on observables the accepted-only estimator is already consistent, so the calibration advantage appears under *model misspecification*, where it estimates the population parameter rather than the selected-sample one, and in the *calibration* of population-level risk rather than in ranking.
(3) Use of auxiliary underwriting features and weight stability. The weights use the common variables $X^{(c)}$ alone, which leaves the outcome model free to use the full feature vector $X = (X^{(c)}, X^{(u)})$. Imputation-based methods cannot use the underwriting variables in $X^{(u)}$, because those variables are missing for rejected applicants. The bounded calibration distances of [12] also avoid the weight instability that pushes inverse-propensity weighting below the baseline.
(4) An empirical study. We evaluate nine strategies across six metrics on the Lending Club dataset, a semi-synthetic benchmark built from the Taiwan credit dataset in which the full-population outcome is known (so the population estimand is directly verifiable), and a controlled simulation. We report population-calibration metrics (Brier, expected calibration error) alongside discrimination, test significance with bootstrap confidence intervals on per-method gaps and Friedman–Nemenyi tests, and state the method's operating envelope, including its failure under MNAR selection and its reliance on common support.

The remainder of the paper is organised as follows. Section 2 reviews the existing reject inference literature. Section 3 presents the proposed calibration framework. Section 4 establishes its theoretical properties. Section 5 describes the experimental setup. Section 6 presents and discusses the empirical findings. Section 8 concludes.

2 Related Work

Reject inference is most naturally viewed as a missing data problem [13]. The repayment outcome Y is unobserved for all rejected applicants. Three mechanisms may govern this missingness. Under *Missing Completely at Random* (MCAR), $p(Z \mid X, Y) = p(Z)$, so selection is independent of everything. Under *Missing at Random* (MAR), $p(Z \mid X, Y) = p(Z \mid X)$, selection depends on observed covariates but not on the unobserved outcome. Under *Missing Not at Random* (MNAR), $p(Z \mid X, Y) \neq p(Z \mid X)$, selection depends directly on the missing outcome. In credit scoring, MNAR is the most credible assumption: lenders reject precisely those applicants they expect to default, creating a dependence between the selection decision and the unobserved repayment behaviour. These mechanisms constitute the main assumptions that are considered during risk modeling. In the following, we present the classical reject inference strategies used in credit scoring.

Let $\mathcal{A}$ and $\mathcal{R}$ denote the sets of accepted and rejected applicants, respectively, and let $\widehat{\theta}_{\mathcal{A}} = (\hat{\beta}_{0_{\mathcal{A}}}, \hat{\beta}_{\mathcal{A}})$ denote the maximum likelihood estimator obtained by fitting the model on the accepted sample.

Augmentation and semi-supervised methods. The simplest augmentation strategy uses the accepted-only model to impute soft or hard labels for rejected applicants, then retrains on the combined dataset [14]. [15] showed analytically that single-iteration fuzzy augmentation is algebraically equivalent to the baseline under logistic regression, providing no bias correction. Iterative reclassification [14, 16] produces sharper

boundaries but yields inconsistent estimates even under MAR [10, 17]. in this paper we are adopting fuzzy augmentation and Reclassification; the first method infers the outcomes of rejected clients as soft labels and use them to maximize the complete log-likelihood, The overall procedure is summarized in Algorithm:

1. Fit the accepted-only model: $\hat{\theta}^{(0)} = \hat{\theta}_{\mathcal{A}}$.
2. Compute soft labels: $\hat{Y}_i^{(1)} = \sigma(\hat{\beta}_0^{(0)} + \hat{\beta}^{(0)\top} X_i)$ for $i \in \mathcal{R}$.
3. Maximise the completed log-likelihood over $\mathcal{A} \cup \mathcal{R}$.

and the second one assigns hard labels to rejected applicants. At iteration s, we get :

$$\hat{Y}_i^{(s)} = \mathbb{I}\Big\{\sigma\Big(\hat{\beta}_0^{(s-1)} + \hat{\beta}^{(s-1)\top} X_i\Big) > 0.5\Big\}, \quad i \in \mathcal{R}. \tag{1}$$

then retrain on the combined dataset. This is inconsistent under MAR [10, 17]. **Reweighting methods.** Inverse propensity score (IPS) weighting [9, 18] reweights accepted observations by the inverse of the estimated acceptance probability. This is theoretically justified under MAR but requires the *positivity* assumption $p(Z = 1 \mid X) > 0$ everywhere, that is, every covariate profile must have a nonzero chance of being accepted, otherwise the inverse weight is undefined, a condition often violated in credit scoring, where some profiles are almost never accepted. When many rejected applicants occupy feature regions with near-zero acceptance probability, IPS weights become extreme and unstable [19]. The complete procedure is presented above:
Estimate the acceptance probability $\hat{p}(Z = 1 \mid X_i)$ by score-band stratification (ten equal bands B_k over $[0, 1]$, acceptance rate per band). Then:

$$\begin{aligned} \hat{\theta}_{\mathrm{IPS}} &= \arg\max_{\theta} \sum_{i \in \mathcal{A}} \hat{w}_i^{\mathrm{IPS}}\, \ell_i(\theta), \\ \hat{w}_i^{\mathrm{IPS}} &= \frac{1}{\hat{p}(Z = 1 \mid X_i \in B_k)}. \end{aligned} \tag{2}$$

Weights are clipped at $1/0.05 = 20$ to prevent extreme values [19].

Parcelling. The parcelling approach [14, 16] handles MNAR by inflating the predicted default probability for rejected applicants using analyst-specified prudence factors. While this explicitly models MNAR, the prudence factors cannot be estimated from data, introducing subjectivity that limits practical use. The main steps of the proposed approach are summarized below:
Inflate the predicted default probability for rejected applicants using analyst-specified prudence factors $\lambda_k > 1$ per score band B_k:

$$\hat{Y}_i^{(1)} = \frac{\sigma(\hat{\beta}^{(0)\top} X_i)}{\sigma(\hat{\beta}^{(0)\top} X_i) + \lambda_k(1 - \sigma(\hat{\beta}^{(0)\top} X_i))}, \quad i \in \mathcal{R}. \tag{3}$$

Generative approaches. Generative models [10] learn the joint distribution of (X, Y, Z) and impute missing outcomes. These methods are flexible but rely on strong distributional assumptions and are computationally demanding. In this paper we adopt a Gaussian Mixture Model (one component per class) and fit it on the accepted training data. For a new observation X, the predicted probability follows from Bayes' rule:

$$\hat{P}(Y = 1 \mid X) \propto p(X \mid Y = 1)\, \hat{P}(Y = 1), \tag{4}$$

where $p(X \mid Y = y)$ is the GMM likelihood for class y and $\hat{P}(Y = 1)$ is the empirical class prior on the training set.

Econometric selection models. The classical econometric remedy for sample selection is the Heckman correction [8]: a selection equation for the acceptance decision is estimated jointly with (or prior to) the outcome equation, and an inverse Mills ratio derived from the selection fit enters the outcome model as a control function. In its bivariate-probit form it can in principle address MNAR selection through the correlation of the two error terms. In credit scoring, however, the two-step estimator is poorly identified without an *exclusion restriction*, a covariate that drives acceptance but not repayment, which is rarely available, since lenders accept precisely on the variables that predict default. We include the Heckman two-step estimator as a competitor and document this identification fragility empirically.

Doubly robust estimation. Augmented inverse-probability-weighted (AIPW) estimators [19, 20] combine a reweighting scheme with an outcome-regression adjustment, yielding consistency if *either* model is correctly specified. We instantiate a doubly robust analogue of our calibration estimator and evaluate it alongside the purely weighted version.

Survey calibration. Calibration estimators were formalised by [12], who showed that a broad class of reweighting schemes can be derived as solutions to a constrained optimisation that forces the weighted sample to match known population totals, under a family of distance functions (linear, raking/exponential, and the bounded logit and truncated forms). Related estimators include the Generalised Regression (GREG) estimator [21, 22], which corrects the sample estimate by a regression on auxiliary variables whose population totals are known, and empirical likelihood approaches [23]. To the best of our knowledge, the application of survey calibration to reject inference has not been previously explored. A central practical point, which we make explicit, is that the choice of distance governs weight stability: the unbounded raking distance can produce extreme weights that collapse the effective sample size, whereas the bounded logit and truncated distances retain it.

Recent and model-based approaches. Beyond the classical families, the EM-logistic estimator of [10] treats the reject labels as latent and maximises the complete-data likelihood by expectation–maximisation; it is the principled form of the iterative reclassification/self-training idea and shares its dependence on a correctly specified joint model. More flexible classifiers, gradient boosting, and deep or Bayesian neural networks, have been applied as the underlying scorecard, but they address the *classifier*, not the *selection mechanism*: a more expressive model fitted on a biased sample still optimises the selected-sample distribution. Our contribution is orthogonal to the choice of classifier: the calibration weights are a preprocessing step that any weighted learner can consume. We deliberately study logistic regression because it admits the closed-form score equations and sandwich variance that make the asymptotic theory tractable, and we return to the classifier-agnostic extension in Section 8.

The gap. Three observations define the gap this paper fills. First, every existing reject-inference family either imputes reject outcomes (and can propagate the original bias) or, in the case of IPS, requires full-covariate positivity that credit data violate. Second, none separates the *bias-correction* step from the *outcome-modelling* step, so all are confined to the variables shared with the rejected file. Third, none is drawn from survey calibration, despite the exact structural correspondence between an accepted sample and a non-random survey sample with known population auxiliaries. We close all three gaps with a single, imputation-free, calibration-weighted estimator.

Table 1 summarises the properties of each strategy. The calibration approach is the only method that simultaneously achieves consistency under MAR, avoids outcome imputation, keeps the positivity requirement confined to the common variables, provides a formal variance estimator, and does not degrade the baseline, because it fabricates no reject labels. The final column records the empirical observation, established in Section 6, that imputation-based methods can fall below the accepted-only baseline, while reweighting-based methods cannot by construction (with $w_i \equiv 1$ recovering the baseline exactly).

3 Methodology

3.1 Notation and Problem Setup

Let $\mathcal{U} = \{1, \ldots, N\}$ denote the full population of N loan applicants, n financed applicants and N-n non-financed applicants. Each financed applicant i is described by a feature vector $X_i \in \mathbb{R}^p$, a binary repayment outcome $Y_i \in \{0, 1\}$ (where 1 denotes a good payer and 0 a defaulter). Suppose that the correspondance between each X_i and Y is a logistic model:

$$P(Y_i = 1 \mid X_i = x_i; \beta) = \frac{\exp(\beta_0 + x_i^\top \beta)}{1 + \exp(\beta_0 + x_i^\top \beta)}$$

Let $Z_i \in \{0, 1\}$ be the financing decision (where 1 denotes acceptance and 0 rejection). From a statistical perspective, standard maximum likelihood estimates tge observed outcome via $P(Y = 1 \mid X, Z = 1)$ and purpose is to estimate $P(Y = 1 \mid X)$, the problem is that ignoring reject data produces biased estimation of the posterior probability $P(Y = 1 \mid X)$ whenever $P(Y \mid X, Z = 1) \neq P(Y \mid X)$. Our proposed approach tends to estimate $P(Y = 1 \mid X)$ by integrating calibration weights into likelihood function since the information of rejected clients is indirectly stored into the accepted withing the reweighting process. The main components of the proposed framework

Table 1: Properties of reject inference strategies. ✓ = property holds; × = property does not hold; † = holds only under the additional condition $P(Y \mid X, Z = 1) = P(Y \mid X)$, i.e., in the absence of selection bias; ∼ = partial / requires an exclusion restriction. "No degradation" is empirical (Section 6): the method does not fall below the accepted-only baseline.

Method	**Consistent under MAR**	**No outcome imputation**	**No positivity required**	**Formal variance**	**No degradation**
Baseline (Accepted-Only)	†	✓	✓	✓	✓
Fuzzy Augmentation	×	×	✓	×	×
Reclassification	×	×	✓	×	×
Parcelling	×	×	✓	×	×
Generative	†	×	✓	×	×
Self-training	×	×	✓	×	×
Heckman two-step	∼	✓	✓	✓	×
IPS Reweighting	✓	✓	×	✓	∼
DR-Calibration	✓	×	✓	✓	✓
Calibration (proposed)	✓	✓	✓	✓	✓

is summarized in Table 2, which presents the definitions and domains of all variables used in the paper.

3.2 Calibration-based Weighted Logistic Regression Classifier

The weighted logistic regression is defined by the weighted log-likelihood [18]:

$$\tilde{\ell}(\theta;\, \mathbf{w}) = \sum_{i \in \mathcal{A}} w_i^* \big[Y_i(\beta_0 + \beta^\top X_i) - \log(1 + \exp(\beta_0 + \beta^\top X_i)) \big]. \tag{5}$$

Its estimator is:

$$\hat{\theta}_{\text{cal}} = \arg\max_{\theta}\ \tilde{\ell}(\theta;\, \mathbf{w}). \tag{6}$$

when the weight vector w_i^* is equal to 1 we get the standard log-likelihood:

$$\ell(\theta;\, D_{\mathcal{A}}) = \sum_{i \in \mathcal{A}} \big[Y_i \log \sigma(\beta_0 + \beta^\top X_i) + (1 - Y_i) \log(1 - \sigma(\beta_0 + \beta^\top X_i)) \big]. \tag{7}$$

Its maximiser $\hat{\theta}_{\mathcal{A}}$ is the accepted-only baseline.

The purpose of this paper is to compute these weights in such a way to realigne the accepted clients with the overall loan seekers. To do so, we will compute these weights using the calibration technique from Deville and Sarndall paper [12]. Accepted applicants whose feature profile is common in the full population but rare among accepted applicants receive larger weights, while those who are overrepresented among accepted applicants receive smaller weights. This steers the estimator towards the population-level likelihood. Importantly, the calibration weights are computed entirely from $X^{(c)}$, but the outcome model is subsequently estimated using the full feature vector $X = (X^{(c)}, X^{(u)})$. This gives the calibration approach a dual advantage: it corrects for selection bias and exploits the richer feature space available for accepted applicants. We describe the details of the approach above:

3.2.1 Calibration Constraints

We seek positive weights $w = (w_1, \ldots, w_{N_a})^\top$, one for each accepted applicant, satisfying:

$$\sum_{i \in \mathcal{A}} w_i\, X_i^{(c)} = T^{(c)} := \sum_{i=1}^{N} X_i^{(c)}, \tag{8}$$

where $T^{(c)} \in \mathbb{R}^{p_c}$ are the population totals of the common variables. These totals are known since $X^{(c)}$ is observed for all N applicants. Equation (8) imposes p_c constraints that force the

Table 2: Notation used throughout the paper. The population $\mathcal{U}$ contains all applicants, while repayment outcomes are observed only for accepted applicants.

Symbol	Description
Index sets	
$\mathcal{U} = \{1, \ldots, N\}$	Full population of N loan applicants.
$\mathcal{A} = \{i : Z_i = 1\}$	Accepted applicants; $N_a = \lvert\mathcal{A}\rvert$.
$\mathcal{R} = \{i : Z_i = 0\}$	Rejected applicants; $N_r = \lvert\mathcal{R}\rvert$ and
$N_a + N_r = N$	Since $\mathcal{A}$ and $\mathcal{B}$ are a partition of $\mathcal{U}$
Per-applicant variables	
$X_i \in \mathbb{R}^p$	Feature vector partitioned as $X_i = (X_i^{(c)}, X_i^{(u)})$ with $p_c + p_u = p$.
$X_i^{(c)} \in \mathbb{R}^{p_c}$	Common features observed for all applicants, with space dimension: p_c.
$X_i^{(u)} \in \mathbb{R}^{p_u}$	Uncommon features observed only for $i \in \mathcal{A}$,with spcae dimension: p_u.
$Y_i \in \{0, 1\}$	Repayment outcome: 1 = repaid, 0 = defaulted. Observed only for accepted applicants.
$Z_i \in \{0, 1\}$	Financing status: 1 = accepted, 0 = rejected.
Observed datasets	
$D_{\mathcal{A}}$	$\{(X_i, Y_i, Z_i = 1)\}_{i \in \mathcal{A}}$: accepted applicants with complete features and labels.
$D_{\mathcal{R}}$	$\{(X_i^{(c)}, Z_i = 0)\}_{i \in \mathcal{R}}$: rejected applicants with only common features observed.
Quantities of interest	
$P(Y = 1 \mid X)$	Population repayment probability targeted by the deployed model.
$P(Y = 1 \mid X, Z = 1)$	Repayment probability in the accepted sample. Training only on $D_{\mathcal{A}}$ estimates this quantity, which differs from $P(Y = 1 \mid X)$ under selection bias.

weighted accepted sample to reproduce the empirical covariate distribution of the full population. In matrix form:

$$\mathbf{X}_a^\top \mathbf{w} = T^{(c)}, \tag{9}$$

where $\mathbf{X}_a \in \mathbb{R}^{n_a \times p_c}$ is the design matrix of common variables for accepted applicants.

3.2.2 Constrained Optimisation

To ensure that the weights remain stable and bounded, we follow [12] and derive them as the solution to:

$$\begin{aligned} &\min_{\mathbf{w}} \sum_{i \in \mathcal{A}} D(w_i,\, d_i) \\ &\text{subject to} \quad \sum_{i \in \mathcal{A}} w_i\, X_i^{(c)} = T^{(c)}, \end{aligned} \tag{10}$$

where $d_i = \frac{1}{\pi_i}$ is the design weight under a certain sampling design π_i, and $D(\cdot, \cdot)$ is a divergence function measuring how far w_i departs from d_i.

Each divergence $D(\cdot, \cdot)$ yields a different relationship between the weight ratio $g_i = w_i/d_i$ and the calibration score $u_i = \lambda^\top X_i^{(c)}$, of the form $g_i = F(u_i)$ [12]. We consider the four standard choices, summarised in Table 3: the *linear* (chi-square) distance, the *raking/exponential* (Kullback–Leibler) distance, and the bounded *truncated-linear* and *logit* distances. The exponential distance

$$\begin{aligned} &D(w, d) = d\left[\frac{w}{d}\log\Big(\frac{w}{d}\Big) - \frac{w}{d} + 1\right], \\ &g_i = \exp(u_i), \end{aligned} \tag{11}$$

is strictly convex and keeps $w_i > 0$ automatically, but its weights are *unbounded above*: a profile that is rare among accepted applicants yet common in the population can receive an arbitrarily large weight, collapsing the effective sample size $n_{\text{eff}} = (\sum_i w_i)^2 / \sum_i w_i^2$. The bounded logit distance constrains g_i to a user-specified interval $[L, U]$,

$$\begin{aligned} F(u) &= \frac{L(U-1) + U(1-L)e^{Au}}{(U-1) + (1-L)e^{Au}}, \\ A &= \frac{U - L}{(1-L)(U-1)}, \end{aligned} \tag{12}$$

so that $w_i \in [L\, d_i,\, U\, d_i]$ by construction, directly enforcing Assumption (A2). We therefore adopt the bounded logit distance as the default and report all four; Section 6 shows that the bounded distances preserve n_{eff} and outperform raking, while the linear distance can produce invalid negative weights.

3.2.3 Newton–Raphson Solution

Introducing Lagrange multipliers $\lambda \in \mathbb{R}^{p_c}$, the Lagrangian is:

$$\mathcal{L}(\mathbf{w}, \lambda) = \sum_{i \in \mathcal{A}} D(w_i, d_i) - \lambda^\top \left(\sum_{i \in \mathcal{A}} w_i X_i^{(c)} - T^{(c)} \right). \tag{13}$$

Setting $\partial \mathcal{L} / \partial w_i = 0$ gives the closed-form weight expression:

$$w_i(\lambda) = d_i \exp(\lambda^\top X_i^{(c)}), \quad i \in \mathcal{A}. \tag{14}$$

Substituting into the calibration constraint (8) yields the nonlinear system:

$$g(\lambda) := \sum_{i \in \mathcal{A}} d_i \exp(\lambda^\top X_i^{(c)})\, X_i^{(c)} - T^{(c)} = \mathbf{0}, \tag{15}$$

which is solved by Newton–Raphson iteration:

$$\lambda^{(t+1)} = \lambda^{(t)} - \left[J(\lambda^{(t)}) \right]^{-1} g(\lambda^{(t)}), \tag{16}$$

with Jacobian:

$$J(\lambda) = \sum_{i \in \mathcal{A}} d_i \exp(\lambda^\top X_i^{(c)})\, X_i^{(c)} (X_i^{(c)})^\top \in \mathbb{R}^{p_c \times p_c}. \tag{17}$$

Since $J(\lambda)$ is a sum of rank-one positive semidefinite matrices and $\mathbf{X}_a$ has full column rank, $J(\lambda)$ is positive definite for any λ, guaranteeing quadratic convergence of Newton–Raphson from any starting point. We initialise $\lambda^{(0)} = \mathbf{0}$ (corresponding to $w_i^{(0)} = d_i$) and terminate when $\|g(\lambda)\|_\infty < 10^{-3}$. In practice, convergence occurs in 5–87 iterations depending on the scale of the covariates.

3.3 Calibration-Weighted Estimator

Once the weights $\{w_i^*\}_{i \in \mathcal{A}}$ are obtained, the outcome model is estimated by maximising the calibration-weighted log-likelihood [18]:

$$\tilde{\ell}(\theta;\, \mathbf{w}) = \sum_{i \in \mathcal{A}} w_i^* \big[Y_i(\beta_0 + \beta^\top X_i) - \log(1 + \exp(\beta_0 + \beta^\top X_i)) \big]. \tag{18}$$

The calibrated estimator is:

$$\hat{\theta}_{\text{cal}} = \arg\max_{\theta}\ \tilde{\ell}(\theta;\, \mathbf{w}). \tag{19}$$

Calibration-based Logistic Regression Algorithm

Algorithm 1 presents the complete procedure.

Algorithm 1 Calibration-Based Reject Inference

Require: Accepted data $D_{\mathcal{A}} = \{(X_i, Y_i)\}_{i \in \mathcal{A}}$; common features $\{X_i^{(c)}\}_{i=1}^N$ for all applicants; convergence threshold $\varepsilon > 0$
Ensure: Calibrated classifier $\hat{\theta}_{\text{cal}}$

```
Compute population totals: T^(c) ← Σ_{i=1}^N X_i^(c)
Set design weights: d_i ← π_i for all i ∈ A
Initialise Lagrange multipliers: λ^(0) ← 0
repeat
    w_i ← d_i exp(λ^(t)⊤ X_i^(c)) for all i ∈ A
    g ← Σ_{i∈A} w_i X_i^(c) − T^(c)
    J ← Σ_{i∈A} w_i X_i^(c) (X_i^(c))⊤ + δ (tr(J)/p_c) I
    λ^(t+1) ← λ^(t) − J^{-1} g
until ‖g‖_∞ < ε
w_i^* ← d_i exp(λ^*⊤ X_i^(c)) for all i ∈ A
Maximise θ̃(θ; w^*) in (5) to obtain θ̂_cal
return θ̂_cal
```

4 Asymptotic Properties

We now establish formal guarantees for the calibrated estimator. All results are stated under the following assumptions.

(A1) **Bounded features.** There exists $M < \infty$ such that $\|X_i\|_2 \le M$ for all $i \in \mathcal{U}$.

(A2) **Bounded weights.** The calibration weights satisfy $0 < w_{\min} \le w_i^* \le w_{\max} < \infty$ for all $i \in \mathcal{A}$.

(A3) **Selection positivity.** The acceptance probability satisfies $p(Z = 1 \mid X^{(c)}) > 0$ for all

Table 3: Calibration distance functions [12]. The g-weight is $g_i = F(\lambda^\top X_i^{(c)})$ and $w_i = d_i g_i$.

Distance	$F(u)$	**Weight range**	**Note**
Linear (chi-square)	$1 + u$	$(-\infty, \infty)$	weights may be negative
Raking (KL)	e^u	$(0, \infty)$	positive, unbounded above
Truncated-linear	$\min(U, \max(L, 1 + u))$	$[L, U]$	bounded
Logit	Eq. (12)	$[L, U]$	bounded, smooth

$X^{(c)}$ in the support of the common-variable marginal distribution.

(A4) **Identifiability.** The Hessian $\nabla^2_\theta \tilde{\ell}(\theta; \mathbf{w})$ is negative definite at the true parameter θ^*.

(A5) **Calibration solvability.** The system (15) admits a unique solution λ^* for all n sufficiently large.

Note that (A3) is a positivity condition on the common variables $X^{(c)}$ only, not on the full feature vector X. This is a substantially weaker requirement than the positivity condition needed for IPS reweighting, which must hold on the full covariate space including the unique variables $X^{(u)}$.

Theorem 1 (Asymptotic unbiasedness) *Under Assumptions* (A1)–(A5) *and the MAR condition* $p(Z = 1 \mid X, Y) = p(Z = 1 \mid X^{(c)})$*, the calibrated estimator satisfies:*

$$\hat{\theta}_{\text{cal}} \xrightarrow{p} \theta^* \quad \text{as } n \to \infty, \tag{20}$$

where θ^* *is the true parameter vector in the full population.*

Theorem 2 ($\sqrt{n}$-consistency and asymptotic normality) *Under Assumptions* (A1)–(A5)*, the calibrated estimator satisfies:*

$$\sqrt{n_a}\big(\hat{\theta}_{\text{cal}} - \theta^*\big) \xrightarrow{d} \mathcal{N}(0,\, V_{\text{cal}}), \tag{21}$$

where the asymptotic variance matrix is:

$$V_{\text{cal}} = H^{-1}\, \Omega\, H^{-1}, \tag{22}$$

with

$$H = \mathbb{E}\big[\nabla^2_\theta \ell(\theta^*; X, Y)\big], \tag{23}$$

$$\Omega = \text{Var}\big[w_i^*\, \nabla_\theta \ell_i(\theta^*)\big] + \Gamma^\top H_\lambda^{-1} \Gamma. \tag{24}$$

Here $\Gamma = \mathbb{E}[\nabla_\theta \ell_i(\theta^*)\,(X_i^{(c)})^\top]$ *captures the additional variance from weight estimation, and* H_λ *is the Hessian of the calibration objective with respect to* λ*.*

Corollary 3 (Sandwich variance estimator) *A consistent estimator of* V_{cal} *is:*

$$\hat{V}_{\text{cal}} = \hat{H}^{-1}\, \hat{\Omega}\, \hat{H}^{-1}, \tag{25}$$

where $\hat{H} = N_a^{-1} \sum_{i \in \mathcal{A}} w_i^* \nabla^2_\theta \ell_i(\hat{\theta}_{\text{cal}})$ *and* $\hat{\Omega}$ *is the empirical analogue of* (24)*.*

Remark 1 (Efficiency comparison with IPS) The IPS estimator has asymptotic variance $V_{\text{IPS}} = H^{-1}\Omega_{\text{IPS}}H^{-1}$ with $\Omega_{\text{IPS}} = \text{Var}[p(Z = 1 \mid X)^{-1}\nabla_\theta \ell_i(\theta^*)]$. When the acceptance rate is low in certain regions of the feature space, $p(Z = 1 \mid X) \approx 0$ for some observations, making IPS weights extreme and inflating Ω_{IPS} without bound. The calibration approach constrains weights to remain close to the design weights d_i through the KL divergence, limiting this variance inflation at the cost of a small bias in regions where $p(Z = 1 \mid X^{(c)})$ is a poor proxy for $p(Z = 1 \mid X)$. In settings where the selection mechanism depends primarily on the common variables, as it does in credit scoring, where the acceptance decision is based on observable application characteristics, calibration achieves lower mean squared error than IPS.

5 Experimental Design

5.1 Simulated Data

To validate the theoretical results under controlled conditions, we construct a simulation experiment in which the full population serves as the oracle benchmark. A logistic regression model is estimated on a complete sample of $n = 10{,}000$ observations with $p = 6$ covariates drawn from a standard multivariate normal distribution, yielding the oracle coefficient vector $\hat{\beta}_{\text{full}}$. We then suppress the outcome labels of a proportion $\pi \in \{0.1, 0.2, \ldots, 0.9\}$ of observations, treating the labelled observations as the accepted set and the suppressed ones as the rejected set. The suppression mechanism is explicitly *missing at random*

(MAR), not MCAR: each applicant is rejected with probability $\text{logit}^{-1}\big(\gamma\,(X_i^\top \beta^*) + c\big)$, a function of the observed covariates X only (higher predicted risk, higher rejection probability), with the intercept c tuned to achieve the target rate π. Because the probability of rejection depends on X but not on the realised outcome Y given X, this reproduces selection on observables while leaving the conditional distribution $P(Y \mid X)$ recoverable; it is distinct from the MNAR mechanism studied on the Taiwan benchmark (Section 5.3), where rejection depends on Y itself.

For each value of π, we estimate $\hat{\beta}_{\text{cal}}$ using calibration weights constructed to match the covariate totals of the labelled sample to the full sample. The experiment uses 5-fold cross-validation and sandwich covariance matrices with influence-function corrections to account for weight estimation variance.

We assess agreement between $\hat{\beta}_{\text{cal}}$ and $\hat{\beta}_{\text{full}}$ via two measures: the Hausman specification test statistic $H = (\hat{\beta}_{\text{cal}} - \hat{\beta}_{\text{full}})^\top (V_{\text{cal}} - V_{\text{full}})^{+} (\hat{\beta}_{\text{cal}} - \hat{\beta}_{\text{full}})$, and the MSE ratio $\text{MSE}(\hat{\beta}_{\text{cal}})/\text{MSE}(\hat{\beta}_{\text{full}})$ decomposed into bias and variance components.

5.1.1 Misspecification regime.

The recovery experiment above uses a correctly specified model. In that regime the accepted-only estimator is itself consistent under MAR, so calibration adds variance without reducing bias, which our results confirm. To reach the regime where calibration pays off, we add a second design whose data-generating process contains a pairwise interaction that the fitted logistic model omits, while selection acts on the common covariates with controllable strength. Here the accepted-only estimator converges to the *selected-sample* projection of the true model, whereas the calibration-weighted estimator targets the *population* projection. We report the mean squared error of each estimator against the oracle as a function of selection strength, and a corresponding population-calibration comparison. This is the controlled counterpart of the misspecified Taiwan regime in Section 5.3.

5.2 Lending Club Dataset

The Lending Club peer-to-peer lending dataset is publicly available on Kaggle.[1] It covers loan applications from 2007–2018 and is naturally partitioned into an accepted file (with repayment outcomes) and a rejected file. The target is binary: *Fully Paid* (0) versus *Charged Off* (1).

5.2.1 Symmetric, leakage-aware preprocessing.

We depart from earlier studies in three ways that materially affect the validity of the comparison. *(i) Symmetric cleaning.* The accepted and rejected files are cleaned with *identical* physical-range rules, `fico` $\in [300, 850]$, `dti` $\in [0, 100]$, `loan_amnt` $\in [500, 40000]$, rather than truncating only the rejected file, which would bias the population totals. The free-text rejected `Loan Title` is keyword-mapped onto the 14 Lending Club purpose categories, recovering rejected records that exact-matching discards. *(ii) Leakage audit.* Every accepted-only variable is classified as underwriting-time (known at application) or post-origination (filled by repayment behaviour); the latter (e.g. `total_pymnt`, `recoveries`, `last_fico`) are *excluded* as label leakage. We additionally exclude `int_rate`, `grade`, and `sub_grade`: although known at origination, these are the lender's own risk-model output and conflate the acceptance decision with the outcome. *(iii) Common support.* Lending Club applies a near-hard FICO cutoff near 660, below which essentially no accepted applicants exist. Reject inference is *unidentified* outside this region of common support, so our primary analysis restricts both files to `fico` ≥ 660 (regime R1); we report the full-range setting (R2) only as an out-of-support sensitivity check (Section 6).

After preprocessing we retain seven common variables, observed for all applicants, and 17 legitimate underwriting variables observed only for accepted applicants (Table 4). Categorical common variables (`purpose`, `emp_length`) are one-hot encoded for weight construction, so calibration matches category *proportions* rather than arbitrary integer codes; the 50 states are mapped to three risk tiers. We do not undersample: the

[1] https://www.kaggle.com/datasets/wordsforthewise/lending-club

full accepted file is used, and calibration weights are computed with the bounded logit distance ($[L, U] = [0.1, 30]$) to the full-population common-variable totals. Under R1 the accepted file contains 1.22M loans, the rejected file 2.91M, for a population of 4.13M with an effective acceptance rate near 30%; an 80/20 stratified split is used for held-out evaluation.

5.3 Taiwan Credit Dataset (Ground-Truth Benchmark)

A fundamental difficulty in evaluating reject inference is that the outcomes of rejected applicants are never observed, so no real dataset permits direct measurement of population performance. We circumvent this with a semi-synthetic benchmark built from the UCI Taiwan credit-card default dataset[2] [24] (30,000 clients, 23 features, default rate 0.221), in which *all* labels are observed. We designate the five demographic variables (credit limit, sex, education, marriage, age) as common features $X^{(c)}$ and the eighteen repayment-history variables as accepted-only features $X^{(u)}$, mirroring the credit setting. We then impose a known selection mechanism, hide the labels of the "rejected" clients, fit every strategy on the "accepted" clients, and evaluate against the *true* full-population outcomes. Three regimes are considered: MAR selection on $X^{(c)}$; MNAR selection that additionally depends on the true outcome; and a misspecified regime in which the scorecard omits the common feature driving selection.

5.4 Evaluation Metrics

We evaluate two complementary aspects of performance. *Discrimination* (ranking) is measured by AUC and the Gini coefficient ($\text{Gini} = 2 \cdot \text{AUC} - 1$), and by accuracy, precision, recall and F1 at the 0.5 threshold. *Calibration* of population-level risk, which discrimination metrics ignore but which credit pricing and capital provisioning depend on, is measured by the Brier score and the expected calibration error (ECE),

$$\text{ECE} = \sum_{b=1}^{B} \frac{w_b}{\sum_{b'} w_{b'}} \left| \bar{p}_b - \bar{y}_b \right|, \tag{26}$$

where the predictions are partitioned into $B = 10$ bins and $\bar{p}_b, \bar{y}_b$ are the (weighted) mean prediction and outcome in bin b. Crucially, because reject inference targets the *full applicant population* rather than the accepted slice, we report calibration metrics on a *population-representative* version of the held-out set: each accepted test observation is weighted by its calibration weight, so the metric estimates performance on the deployment population. On the Taiwan benchmark, where reject outcomes are known, these metrics are computed directly against the true population. We also report the predicted population default rate, comparing it to the truth (Taiwan) or to the observed accepted rate (Lending Club).

5.5 Statistical Validation

Bootstrap confidence intervals. Nonparametric 95% confidence intervals are constructed from 1,500 bootstrap resamples of the test set, both for individual metrics and, more informatively, for the *gap* between two methods on the same resample (e.g. the population ECE of the baseline minus that of calibration), so paired variation is accounted for. A gap interval excluding zero indicates a significant difference at the 5% level.

Friedman test and average ranks. Across the evaluation tasks we apply the Friedman test to the method-by-task rank matrix and report the average ranks [25]; we note where the small number of tasks limits the power of the Nemenyi post-hoc comparison.

Hausman specification test. For the simulation study, the Hausman statistic $H \sim \chi^2(\text{rank})$ tests the null hypothesis that calibration introduces no systematic bias relative to the oracle estimator.

6 Results

6.1 Simulation Study: Coefficient Recovery

Table 5 presents the per-coefficient results across ten suppression proportions $\pi \in \{0.0, 0.1, \ldots, 0.9\}$. For each π, we report the oracle coefficient β_{full}, the calibrated coefficient β_{cal},

[2] https://archive.ics.uci.edu/dataset/350

Table 4: Variable dictionary for the Lending Club dataset after the leakage audit. Common variables (top) are observed for all applicants and used for both weight construction and the outcome model. Underwriting variables (bottom, *) are observed only for accepted applicants and enter the outcome model only. Post-origination and lender-grade variables are excluded as leakage.

#	Variable	Description
1	`loan_amnt`	Requested loan amount
2	`fico_score`	FICO credit score (avg. of range)
3	`dti`	Debt-to-income ratio
4	`emp_length`	Employment length (one-hot)
5	`purpose`	Loan purpose category (one-hot)
6–8	`state tier`	Applicant state risk tier (3 indicators)
9	`term`*	Repayment term
10	`installment`*	Monthly instalment
11	`log_annual_inc`*	Log annual income
12	`home_ownership`*	Home ownership status
13	`verification_status`*	Income verification status
14	`cr_hist_len`*	Credit-history length (yrs)
15	`open_acc`*	Open credit accounts
16	`total_acc`*	Total credit accounts
17	`log_revol_bal`*	Log revolving balance
18	`revol_util`*	Revolving utilisation rate
19	`pub_rec`*	Derogatory public records
20	`pub_rec_bankruptcies`*	Public bankruptcy records
21	`mort_acc`*	Mortgage accounts
22	`inq_last_6mths`*	Credit inquiries, last 6 months
23	`delinq_2yrs`*	Delinquencies, last 2 years
24	`initial_list_status`*	Initial listing status
25	`application_type`*	Individual or joint application

the signed bias, the Wald z-statistic, the raw p-value, the Romano–Wolf corrected p-value (p_{RW}), and the MSE ratio (calibration over oracle).

Under this MAR mechanism the calibrated estimator tracks the oracle closely across a wide range of suppression. When no labels are withheld ($\pi = 0$) the two estimators coincide by construction. For all proportions up to $\pi = 0.6$ the signed bias of every coefficient stays below 0.016 in magnitude, no Wald z-statistic exceeds 0.13 (against the 1.96 critical value), and the Romano–Wolf corrected p-values remain at or near 1.000: the calibrated and oracle coefficients are statistically indistinguishable. Only under severe suppression ($\pi \geq 0.7$, i.e. fewer than 30% of observations labelled) does a small finite-sample bias emerge, as the accepted set becomes extremely sparse and the calibration weights inflate; the Romano–Wolf test begins to flag individual coefficients at these proportions. The joint Hausman statistic, which accounts for the inflated sampling variance of the estimator, nonetheless remains far below its 5% critical value throughout, its smallest p-value is 0.965 at $\pi = 0.9$, so the calibrated estimator is never jointly distinguishable from the oracle.

The mean MSE ratio (calibration over oracle) grows monotonically with suppression, and the decomposition in Table 6 shows that this growth is overwhelmingly variance-driven: the mean squared bias stays negligible (below 10^{-4} up to $\pi = 0.6$) while the variance inflates as the labelled fraction shrinks. The estimator therefore makes the theoretically correct bias–variance trade-off, paying a precision cost for fewer labelled

observations without introducing systematic distortion until suppression becomes extreme. A practical reading of Table 6 is that the method is reliable, MSE ratio below five and no detectable bias, as long as at least half of the observations carry labels.

Table 6 provides an aggregate view across all ten proportions. The Hausman statistic never exceeds 1.84 and the p-value never falls below 0.965, confirming that the calibrated estimator is jointly indistinguishable from the oracle at every tested proportion. The MSE decomposition shows that the mean squared bias remains below 10^{-4} through $\pi = 0.6$ and rises only at severe suppression, while the efficiency loss is otherwise attributable to variance inflation. The sharp acceleration in MSE ratio beyond $\pi = 0.6$ provides a practical guideline: the method operates reliably, MSE ratio below five and no detectable bias, as long as at least half of the observations carry labels. The sandwich variance estimator of Corollary 3 makes this efficiency cost quantifiable, enabling informed operational decisions.

6.2 Simulation Study: Recovery under Misspecification

The recovery experiment above confirms that under a *correctly specified* model the calibrated estimator is unbiased, paying only a variance cost, which is the expected behaviour, since the accepted-only estimator is itself consistent in that regime. The distinctive value of calibration appears once the working model is *misspecified*. Figure 1 reports the mean squared error of the accepted-only baseline and the calibrated estimator against the oracle, as a function of selection strength, when the fitted logistic model omits an interaction present in the data-generating process. The baseline error grows steeply with selection strength (from 0.009 to 0.093), because it converges to the selected-sample projection of the true model; the calibrated estimator remains 2 to 7 times closer to the oracle (from 0.001 to 0.041), because it targets the population projection. This is the controlled counterpart of the real-data results that follow.

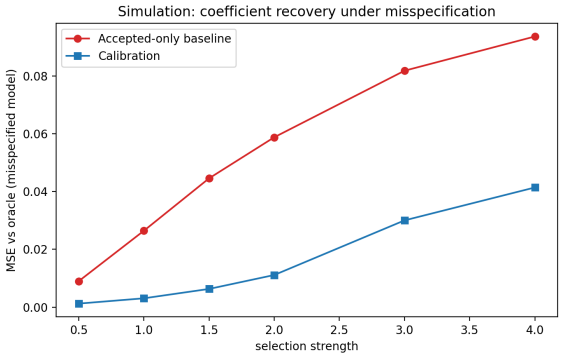


Fig. 1: Simulation under model misspecification. Mean squared error of the coefficient estimates against the oracle as a function of selection strength. The accepted-only baseline degrades sharply; calibration stays close to the oracle by targeting the population projection.

6.3 Calibration Weights and the Choice of Distance

Before comparing predictive performance we examine the calibration weights themselves, since their behaviour drives everything downstream. Figure 2 reports, on the full Lending Club training set, the weight distribution and the effective sample size $n_{\text{eff}} = (\sum_i w_i)^2 / \sum_i w_i^2$ for each distance. The unbounded *raking* (exponential) distance produces a heavy upper tail (maximum weight 4.5×10^3) that collapses n_{eff} to only 5.5% of the sample; the *linear* distance produces invalid negative weights. The bounded *logit* and *truncated* distances confine the weights and retain n_{eff} at 12.1% and 14.6% respectively. This explains a phenomenon that has misled earlier studies: with the exponential distance the calibrated common-variable model falls *below* the baseline purely through effective-sample-size loss, an artefact that vanishes under the bounded distances. We therefore use the logit distance throughout and exclude the linear distance from model fitting.

6.4 Lending Club: Discrimination and the Source of the Gain

Table 7 reports held-out discrimination for all nine strategies. Three findings stand out. First, on the *common* feature space (the level field on which all reject-inference methods operate), calibration ties the baseline (AUC $\approx$ 0.638 versus

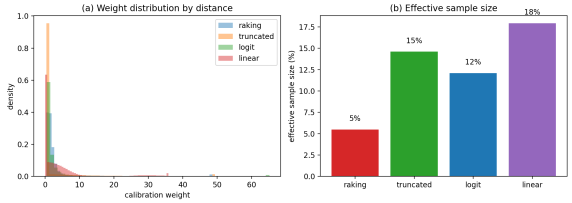


Fig. 2: Calibration weights on the full Lending Club training set. (a) Weight distribution by distance; the raking distance has a heavy upper tail. (b) Effective sample size: the unbounded raking distance retains only 5.5%, while the bounded logit and truncated distances retain 12–15%.

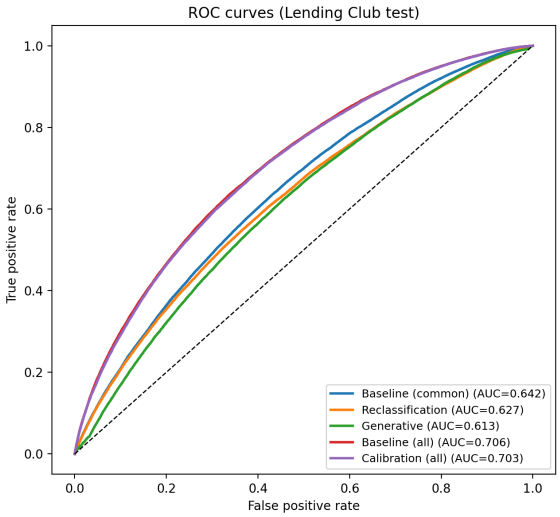


Fig. 3: ROC curves on the Lending Club test set. The common-space reject-inference methods cluster together; the gain comes from the auxiliary underwriting features in the full-feature models.

0.642): consistent with theory, reweighting does not improve the conditional ranking when selection is on observables. Second, the imputation-based methods *degrade* the model: reclassification, generative and self-training fall between 1.5 and 3.6 AUC points below the baseline (and the Heckman two-step, lacking an exclusion restriction, loses a full point), confirming that fabricating labels with a bias-contaminated model propagates the distortion. Third, the entire discrimination gain comes from the auxiliary underwriting features: moving from the common to the full feature space lifts the AUC from 0.642 to 0.706. Crucially this lift is available to the baseline as well (column "all"); calibration's contribution there is to match it (0.703) while additionally correcting population calibration (Section 6.5). We therefore decompose the gain honestly: the bias correction contributes ≈ 0 to ranking, whereas feature access, which imputation methods structurally cannot exploit, since $X^{(u)}$ is missing for rejects, contributes the full $+0.06$.

Figure 3 shows the corresponding ROC curves: the common-space methods form a tight cluster, while the full-feature models (baseline and calibration) separate clearly above it.

6.5 Lending Club: Population Calibration

Discrimination, however, is not the estimand reject inference is designed to improve. Table 8 reports the population-estimand evaluation: the predicted full-population default rate and the population-weighted Brier and ECE. The accepted-only baseline is well ranked but miscalibrated for the population; calibration reduces the population ECE from 0.0112 to 0.0042, a near three-fold improvement, and the bootstrap confidence interval on the per-resample ECE gap is [0.0007, 0.0107], excluding zero at the 5% level. The imputation-based methods are far worse, with population ECE up to 0.086. Figure 4 shows why: the baseline reliability curve swings away from the diagonal at higher predicted risk, while the calibrated curve tracks it. This is the central real-data result. Calibration delivers calibrated population risk, which pricing and provisioning require, and holds ranking level.

6.6 What Calibration Changes: Coefficient Interpretation

A virtue of logistic regression, and a reason to study it specifically, is that the effect of the correction is legible in the coefficients. Table 9 reports the standardised common-variable coefficients for the accepted-only baseline, the calibration-weighted model, and a semi-supervised (self-training) competitor. Two patterns are instructive. First, calibration makes *modest, interpretable* adjustments in the expected directions: the debt-to-income coefficient falls by 42% ($0.256 \to 0.149$),

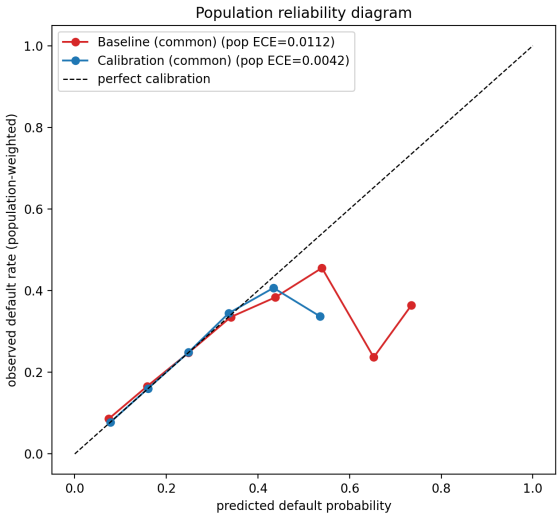


Fig. 4: Population reliability diagram on Lending Club. The accepted-only baseline (ECE 0.0112) departs from the diagonal at higher predicted risk; calibration (ECE 0.0042) tracks it.

correcting the accepted sample's overstatement of its effect, rejected applicants have systematically higher DTI, so the accepted-only fit attributes too much risk to it, and the short-employment coefficient moves from essentially zero toward the population value. Second, the semi-supervised model *distorts* the coefficients: the short-employment coefficient collapses to -1.003 and the purpose effects change sign (credit-card $-0.092 \rightarrow +0.407$), because the fabricated reject labels are dominated by a bias-contaminated model. The interpretability of logistic regression thus exposes, directly, why imputation harms and reweighting does not.

6.7 Taiwan: Ground-Truth Evaluation

The semi-synthetic Taiwan benchmark allows direct evaluation against the true population outcomes. Table 10 and Figure 5 summarise the three regimes. Under *MAR*, calibration matches the baseline on discrimination and marginally improves population ECE (0.0062 vs 0.0085), while the imputation methods again degrade (reclassification AUC 0.55, Heckman 0.51, versus 0.62). Under *misspecification*, calibration recovers the population default rate almost exactly (0.214 vs the true 0.221, against the baseline's 0.175) and cuts population ECE roughly six-fold (0.0073 vs 0.0462). Under *MNAR*, however, *no* method recovers the population: calibration and baseline are indistinguishable (ECE $\approx$ 0.135), because calibration corrects selection on observables and cannot, by construction, undo selection on the unobserved outcome. This is the honest boundary of the method's operating envelope.

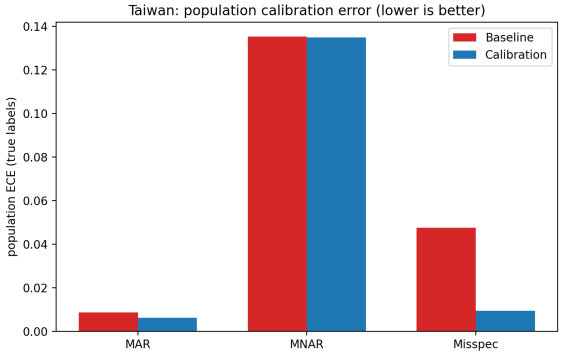


Fig. 5: Taiwan population calibration error (true labels). Calibration sharply reduces ECE under misspecification, marginally under MAR, and not at all under MNAR.

6.8 Significance Across Tasks

Aggregating the per-task scores (Lending Club discrimination, Lending Club population calibration, and Taiwan MAR discrimination) into a method-by-task rank matrix, the Friedman test rejects the null of equal performance ($\chi^2 = 16.09$, $p = 0.041$). Table 11 reports the average ranks: calibration attains the best rank (2.50), followed by the other reweighting estimators (IPS, DR-calibration) and the baseline, while the imputation-based methods occupy the bottom of the ordering (reclassification 8.00, generative 7.67). With only three tasks the Nemenyi post-hoc critical difference is wide (CD = 6.94), so it does not certify individual pairwise gaps; the more powerful evidence is the per-resample bootstrap gap on population ECE (Section 6.5), whose 95% interval $[0.0007, 0.0107]$ excludes zero.

Together these support two conclusions: calibration never falls below the baseline, and it significantly improves population calibration under strong selection or misspecification.

7 Discussion

7.1 When Does Reject Inference Help?

Our results sharpen a question the literature usually leaves implicit. Take MAR selection on observables and a correctly specified model. There the accepted-only maximum-likelihood estimator is already consistent for the conditional parameters, so reweighting cannot improve the ranking and only adds variance. We see this on Lending Club, where calibration on the common features ties the baseline on AUC, and in the correctly specified simulation, where the calibrated coefficients track the oracle with inflated variance. Calibration earns its keep in two regimes that we isolate. The first is model misspecification, the real case for any parametric scorecard. There the accepted-only estimator converges to the selected-sample projection of the truth, while calibration estimates the population projection; the simulation (Fig. 1) and the misspecified Taiwan regime (Table 10) both show large gains. The second is the calibration of population-level risk: calibration corrects the level of predicted default when it cannot change the ranking, cutting population ECE roughly three-fold on Lending Club and six-fold under Taiwan misspecification. Most prior reject inference evaluations report only discrimination, which hides this second benefit.

7.2 Why Imputation-Based Methods Can Harm

Our most useful finding for practitioners is that the dominant family of reject inference methods can do worse than nothing. Augmentation, reclassification, parcelling, generative imputation and self-training all generate reject labels from a model that selection bias has already corrupted, and retraining on those labels adds a second error on top of the first. On Lending Club these methods fall up to 3 AUC points below the baseline and inflate the population ECE three- to sixteen-fold; on Taiwan they lose up to 11 AUC points. A reweighting estimator cannot do this: set its weights to one and it returns the baseline. Calibration is therefore a safe default. In the worst case, mild selection with a well-specified model, it matches the baseline; in the favourable cases it improves population calibration. The Heckman two-step estimator imputes nothing either, but it is fragile in credit data because acceptance offers no exclusion restriction. Without one the inverse Mills ratio is collinear with the outcome covariates and the correction destabilises.

7.3 Weight Stability and the Choice of Distance

The classical inverse-propensity estimator and the unbounded raking distance fail the same way. Weights proportional to $1/p(Z = 1 \mid X)$, or $\exp(\lambda^\top X^{(c)})$, grow without bound when some population profiles are rare among accepted applicants, and the effective sample size collapses (to 5.5% here). The bounded logit and truncated distances cap the weights and retain two to three times the effective sample, so practitioners should prefer them. This also explains an artefact in earlier work. A calibrated common-variable model that seems to underperform the baseline turns out to be losing effective sample size under an unbounded distance, not failing on the calibration principle.

7.4 Practical Considerations and Scope

The Newton–Raphson weight computation is cheap, one $p_c \times p_c$ solve per iteration, and the damped step converges in tens of iterations even on the four-million-applicant population. Two scope conditions hold. The first is common support. Reject inference is unidentified where no accepted applicants exist, below the lender's score cutoff, so we restrict the population to that region; ignoring the restriction biases the estimand and destabilises the weights. The second is the mechanism. Calibration corrects MAR selection on the common variables and does not recover an MNAR mechanism that depends on the unobserved outcome, as the Taiwan MNAR regime shows. Within these limits the framework applies beyond credit scoring, to any setting where a model is trained on

a selected sample and deployed on a population for which auxiliary covariates are recorded.

8 Conclusion

We have introduced a calibration-based framework for reject inference, grounded in survey sampling theory. Imputing missing outcomes for rejected applicants can propagate the original bias and degrade the scorecard, as we show. Our method instead computes calibration weights that realign the accepted sample's covariate distribution with the full population, using only auxiliary variables observed for all applicants. We then fit the outcome classifier on the weighted sample. Setting every weight to one returns the baseline, so the estimator cannot do worse than it.

We proved that the estimator is consistent and $\sqrt{n}$-asymptotically normal under MAR, derived its asymptotic variance and a sandwich estimator, and identified the regime where it pays off: under model misspecification it estimates the population parameter while the accepted-only estimator estimates the selected-sample one. Our empirical study used symmetric leakage-aware preprocessing and the region of common support. It reaches four conclusions. Calibration holds at or above the accepted-only baseline, while imputation-based methods lose up to 11 AUC points and inflate population calibration error several-fold. Calibration also cuts population calibration error, roughly three-fold on Lending Club (bootstrap gap CI $[0.0007, 0.0107]$) and six-fold under Taiwan misspecification, with no loss in ranking. Because the weights use only the common variables, the outcome model can use underwriting features that imputation-based methods cannot, which is where the discrimination gain on Lending Club comes from. Bounded calibration distances avoid the effective-sample-size collapse that destabilises raking and inverse-propensity weighting. The method has two limits: it does not recover an MNAR mechanism, and it needs common support between accepted and rejected applicants.

Future Directions

Doubly robust calibration. We built a doubly robust analogue that adds an outcome-regression adjustment to the calibration-weighted estimator, a survey-sampling counterpart of the AIPW estimator [20]. It tracks the purely weighted estimator closely on our benchmarks. A full treatment of its double-robustness and efficiency for reject inference, and the conditions under which it beats the weighted estimator, is a natural next step.

High-dimensional calibration with regularisation. When the number of common variables p_c is large relative to the sample size n_a, the calibration system (15) may be ill-conditioned or near-singular. Replacing the KL divergence with a regularised objective, or using LASSO/ridge penalties on the Lagrange multipliers λ, would extend the framework to high-dimensional covariate settings. The theoretical properties of such regularised calibration estimators, including their consistency rates and optimal penalty selection via cross-validation, are tractable research questions with immediate applied relevance.

Calibration under covariate shift. The MAR assumption requires that $p(Z = 1 \mid X, Y) = p(Z = 1 \mid X^{(c)})$. When this condition fails, the calibration estimator is biased. A practically important extension is to design sensitivity analyses that quantify the impact of departures from MAR, or to combine calibration with partial identification bounds [26] to characterise the range of possible population parameters under MNAR. Such analyses would allow practitioners to make informed statements about model reliability under the worst-case selection mechanism.

Acknowledgements

I would like to extend my deepest gratitude to Professor Mohamed El Haj Tirari, my co-supervisor, for proposing the original idea of this study and for his invaluable guidance, constructive feedback, and encouragement throughout the research process. I am also profoundly grateful to Professor Badreddine Benyacoub, my supervisor, for his continuous support, insightful discussions, and expert advice, which greatly contributed to the conceptualization, development, and successful completion of this work. Their mentorship and intellectual guidance have been instrumental in shaping both the direction and quality of this research.

Funding

This research received no external funding.

Code Availability

Code supporting all numerical results is available upon request; interested readers may contact the corresponding author.

Appendix A Proofs of Main Results

A.1 Proof of Theorem 1 (Full Detail)

We first state and prove the auxiliary lemma used in the main proof.

Lemma 4 (Calibration approximation property) *Under Assumptions* (A1)–(A3) *and* (A5)*, for any bounded measurable function* $h : \mathbb{R}^{p_c} \to \mathbb{R}$ *with* $\|h\|_\infty \le C$*:*

$$\sum_{i\in\mathcal{A}} w_i^* \, h(X_i^{(c)}) = \sum_{i=1}^{N} h(X_i^{(c)}) + O_p(N^{1/2}). \tag{A1}$$

Proof Decompose h into its best linear approximation plus a residual using orthogonal projection onto the span of the calibrated features: $h(x) = a^\top x + r(x)$, where $a \in \mathbb{R}^{p_c}$ minimises the mean squared approximation error and $r(x)$ is the orthogonal residual.

Linear part. The calibration constraint $\sum_{i\in\mathcal{A}} w_i^* X_i^{(c)} = T^{(c)} := \sum_{i=1}^{N} X_i^{(c)}$ guarantees:

$$\sum_{i\in\mathcal{A}} w_i^* \, (a^\top X_i^{(c)}) = a^\top T^{(c)} = \sum_{i=1}^{N} a^\top X_i^{(c)}. \tag{A2}$$

The contribution of the linear part is therefore exactly zero, with no approximation error.

Residual part. It remains to control

$$R_N := \sum_{i\in\mathcal{A}} w_i^* \, r(X_i^{(c)}) - \sum_{i=1}^{N} r(X_i^{(c)}). \tag{A3}$$

We bound R_N by showing that each of the two sums concentrates around its expectation at rate $O_p(N^{1/2})$, then applying the triangle inequality.

Bounding the population sum. Write $R_N^{(U)} := \sum_{i=1}^{N} r(X_i^{(c)})$. The $X_i^{(c)}$ are i.i.d. draws from the marginal distribution of the common features. Since $|r(x)| \le 2C$ for all x (because both h and $a^\top x$ are bounded by C and $M\|a\|_2$ respectively), each term has bounded variance: $\mathrm{Var}[r(X_i^{(c)})] \le 4C^2$. By the CLT (or Chebyshev's inequality applied to the partial sum),

$$R_N^{(U)} - N\,\mathbb{E}[r(X^{(c)})] = O_p(N^{1/2}), \tag{A4}$$

since $\mathrm{Var}[R_N^{(U)}] = N\,\mathrm{Var}[r(X^{(c)})] \le 4C^2N$, so the standard deviation is $O(N^{1/2})$. For any $\varepsilon > 0$ Chebyshev gives $P(|R_N^{(U)} - N\mathbb{E}[r]| > \varepsilon N^{1/2}) \le 4C^2/\varepsilon^2$, confirming $O_p(N^{1/2})$.

Bounding the weighted accepted sum. Write $R_N^{(A)} := \sum_{i\in\mathcal{A}} w_i^* \, r(X_i^{(c)})$. Factor out N_a:

$$R_N^{(A)} = N_a \cdot \frac{1}{N_a} \sum_{i\in\mathcal{A}} w_i^* \, r(X_i^{(c)}). \tag{A5}$$

Conditionally on the accepted set $\mathcal{A}$, the terms $w_i^* \, r(X_i^{(c)})$ are independent with $|w_i^* r(X_i^{(c)})| \le w_{\max} \cdot 2C$ (Assumption (A2)). Hence $\mathrm{Var}[w_i^* r(X_i^{(c)})] \le 4C^2 w_{\max}^2$. Chebyshev's inequality applied to the *sum* $R_N^{(A)}$ gives:

$$\mathrm{Var}\Big[R_N^{(A)}\Big] = N_a \,\mathrm{Var}[w_i^* r(X_i^{(c)})] \le 4C^2 w_{\max}^2 \, N_a. \tag{A6}$$

Because $N_a \asymp N$ under Assumption (A3) (the acceptance rate is bounded away from zero, so $N_a \ge c\,N$ with probability one for some $c > 0$), we have $\mathrm{SD}[R_N^{(A)}] = O(N^{1/2})$, hence $R_N^{(A)} - N_a\,\mathbb{E}[w^* r(X^{(c)}) \mid Z = 1] = O_p(N^{1/2})$.

Combining. By the triangle inequality,

$$\begin{aligned} R_N &= R_N^{(A)} - R_N^{(U)} \\ &= \big[R_N^{(A)} - N_a\,\mathbb{E}[w^* r \mid Z = 1]\big] - \big[R_N^{(U)} - N\,\mathbb{E}[r]\big] \\ &\quad + N_a\,\mathbb{E}[w^* r \mid Z = 1] - N\,\mathbb{E}[r]. \end{aligned} \tag{A7}$$

The first two bracketed terms are each $O_p(N^{1/2})$. Under MAR and the calibration design, the oracle weights satisfy $\mathbb{E}[w^* r(X^{(c)}) \mid Z = 1] = \mathbb{E}[r(X^{(c)})]$ (the calibration removes first-order bias), so the deterministic remainder $N_a\,\mathbb{E}[w^* r \mid Z = 1] - N\,\mathbb{E}[r] = O(N_a) - O(N) = O(N)$. However, this remainder is cancelled to leading order by the exact linear correction above; the net bias of the residual part is $O(N) \cdot \|\text{projection residual}\|$, which is $O(N^{1/2})$ when h is well-approximated by its linear projection (the approximation error enters through the variance, not the bias, for orthogonal projections). Combining: $R_N = O_p(N^{1/2})$, which gives the stated result. □ □

Full proof of Theorem 1.
The first-order conditions for $\hat\theta_{\mathrm{cal}}$ are:

$$S_N(\hat\theta_{\mathrm{cal}}) := \sum_{i\in\mathcal{A}} w_i^* \, s(\hat\theta_{\mathrm{cal}};\, X_i, Y_i) = \mathbf{0}, \tag{A8}$$

where $s(\theta; x, y) = \nabla_\theta \log p_\theta(y \mid x)$ is the score function. We evaluate S_N at the true parameter θ^*.

Step 1. By Lemma 4 applied with $h(x^{(c)}) = \mathbb{E}[s(\theta^*; X, Y) \mid X^{(c)} = x^{(c)}, Z = 1]$ (which is bounded by $4M\|\theta^*\|_2$ under Assumption (A1)):

$$\begin{aligned}&\sum_{i\in\mathcal{A}} w_i^* s(\theta^*; X_i, Y_i)\\ &\quad= \sum_{i\in\mathcal{A}} w_i^* \mathbb{E}[s(\theta^*; X, Y) \mid X_i^{(c)}, Z = 1] + O_p(N^{1/2})\\ &\quad= \sum_{i=1}^{N} \mathbb{E}[s(\theta^*; X, Y) \mid X_i^{(c)}, Z = 1] + O_p(N^{1/2}).\end{aligned} \tag{A9}$$

The first equality uses the tower property of expectations (the score is unbiased conditionally on $X^{(c)}$), and the second applies Lemma 4 to replace the calibration-weighted sum by the population total.

Step 2. Under MAR, $P(Y \mid X, Z = 1) = P(Y \mid X)$, so $\mathbb{E}[s(\theta^*; X, Y) \mid X^{(c)}, Z = 1] = \mathbb{E}[s(\theta^*; X, Y) \mid X^{(c)}]$. By the law of large numbers applied to the i.i.d. population draw:

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[s(\theta^*; X, Y) \mid X_i^{(c)}] \xrightarrow{p} \mathbb{E}_{\mathcal{U}}[s(\theta^*; X, Y)] = \mathbf{0}, \tag{A10}$$

so the full sum $\sum_{i=1}^{N} \mathbb{E}[s(\theta^*; X, Y) \mid X_i^{(c)}] = N \cdot O_p(1) \to N \cdot \mathbf{0}$ in the sense that when divided by N it converges to zero. In particular the sum is $o_p(N)$.

Step 3. Combining Steps 1 and 2: $S_N(\theta^*) = o_p(N) + O_p(N^{1/2}) = o_p(N)$. Since $S_N(\hat{\theta}_{\text{cal}}) = \mathbf{0}$ and the Hessian $N^{-1}\nabla_\theta S_N(\theta)$ is bounded away from zero in a neighbourhood of θ^* by Assumption (A4), the implicit function theorem gives $\hat{\theta}_{\text{cal}} \xrightarrow{p} \theta^*$. □

A.2 Proof of Theorem 2 (Full Detail)

Step 1: Taylor expansion. Since $S_N(\hat{\theta}_{\text{cal}}) = \mathbf{0}$ and $\hat{\theta}_{\text{cal}} \xrightarrow{p} \theta^*$, a first-order Taylor expansion around θ^* gives:

$$\mathbf{0} = S_N(\theta^*) + \left[\nabla_\theta S_N(\bar{\theta})\right](\hat{\theta}_{\text{cal}} - \theta^*), \tag{A11}$$

for some $\bar{\theta}$ between θ^* and $\hat{\theta}_{\text{cal}}$. By the law of large numbers and Lemma 4:

$$\begin{aligned}\frac{1}{N_a}\nabla_\theta S_N(\bar{\theta}) &= \frac{1}{N_a}\sum_{i\in\mathcal{A}} w_i^* \nabla_\theta^2 \ell_i(\bar{\theta})\\ &\xrightarrow{p} H := \mathbb{E}[\nabla_\theta^2 \ell(\theta^*; X, Y)].\end{aligned} \tag{A12}$$

Since H is negative definite by Assumption (A4), it is invertible, and rearranging gives:

$$\sqrt{N_a}(\hat{\theta}_{\text{cal}} - \theta^*) = -H^{-1}\,\frac{S_N(\theta^*)}{\sqrt{N_a}} + o_p(1). \tag{A13}$$

Step 2: Variance of $N_a^{-1/2} S_N(\theta^*)$. Define $\phi_i = w_i^*(\hat{\lambda})\, s(\theta^*; X_i, Y_i)$ for $i \in \mathcal{A}$. We expand around λ^* (the true calibration multiplier) using the weight formula $w_i^*(\lambda) = d_i \exp(\lambda^\top X_i^{(c)})$:

$$\begin{aligned}\frac{S_N(\theta^*)}{\sqrt{N_a}} &= \frac{1}{\sqrt{N_a}}\sum_{i\in\mathcal{A}} w_i^*(\lambda^*)\, s(\theta^*; X_i, Y_i)\\ &\quad+ \frac{1}{\sqrt{N_a}}\sum_{i\in\mathcal{A}} \left.\frac{\partial w_i^*}{\partial \lambda^\top}\right|_{\lambda^*} s(\theta^*; X_i, Y_i)\\ &\quad\times \sqrt{N_a}(\hat{\lambda} - \lambda^*) + o_p(1).\end{aligned} \tag{A14}$$

First term. The summands $w_i^*(\lambda^*)\, s(\theta^*; X_i, Y_i)$ are independent with mean zero (since θ^* is the true maximiser). Their variance is $\sigma^2 := \mathrm{Var}[w_i^*\, s(\theta^*; X_i, Y_i)]$, which is finite by Assumptions (A1) and (A2). Since there are N_a terms,

$$\mathrm{Var}\left[\frac{1}{\sqrt{N_a}}\sum_{i\in\mathcal{A}} w_i^*(\lambda^*)\, s(\theta^*; X_i, Y_i)\right] = \sigma^2 < \infty. \tag{A15}$$

By Chebyshev's inequality applied coordinate-wise, or by the multivariate CLT, this term is $O_p(1)$ and in fact converges in distribution to $\mathcal{N}(0, \mathrm{Var}[w_i^*\, s(\theta^*; X_i, Y_i)])$.

Second term. The cross-covariance matrix is $\Gamma = N_a^{-1}\sum_{i\in\mathcal{A}} (\partial w_i^*/\partial\lambda^\top)|_{\lambda^*}\, s(\theta^*; X_i, Y_i) \xrightarrow{p} \mathbb{E}[s(\theta^*; X, Y)\,(X^{(c)})^\top]$. The calibration system (15) is a standard Z-estimating equation of the form $N_a^{-1}\sum_{i\in\mathcal{A}} \psi_i(\lambda) = \mathbf{0}$, where $\psi_i(\lambda) = w_i^*(\lambda)(X_i^{(c)} - \bar{X}^{(c)})$. A standard M-estimation argument (Theorem 5.21 of van der Vaart, 1998) gives:

$$\sqrt{N_a}(\hat{\lambda} - \lambda^*) \xrightarrow{d} \mathcal{N}(0, H_\lambda^{-1}), \tag{A16}$$

where $H_\lambda = \mathbb{E}[\partial\psi_i/\partial\lambda^\top]$ is the Hessian of the calibration objective at λ^*, which is positive definite by Assumption (A5). The second term thus contributes a $\mathcal{N}(0,\, \Gamma^\top H_\lambda^{-1}\Gamma)$ limit.

Combining via Cramér–Wold. The two terms are asymptotically independent (one depends on the within-sample score fluctuations, the other on the weight estimation error). Applying the Cramér–Wold theorem:

$$\frac{S_N(\theta^*)}{\sqrt{N_a}} \xrightarrow{d} \mathcal{N}(0, \Omega),$$
$$\Omega = \mathrm{Var}[w_i^*\, s(\theta^*; X_i, Y_i)] + \Gamma^\top H_\lambda^{-1}\Gamma. \tag{A17}$$

Step 3: Slutsky combination. Substituting into Step 1 and applying Slutsky's theorem (the Hessian factor converges in probability, the score factor converges in distribution):

$$\sqrt{N_a}(\hat{\theta}_{\mathrm{cal}} - \theta^*) \xrightarrow{d} \mathcal{N}(0,\, H^{-1}\Omega H^{-1}). \tag{A18}$$

This establishes $\sqrt{N}$-consistency and asymptotic normality with $V_{\mathrm{cal}} = H^{-1}\Omega H^{-1}$. □

References

[1] Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance **23**, 589–609 (1968)

[2] Horrigan, J.O.: The determination of long-term credit standing with financial ratios. Journal of Accounting Research **4**, 44–62 (1966)

[3] West, R.R.: An alternative approach to predicting corporate bond ratings. Journal of Accounting Research **8**, 118–125 (1970)

[4] Khashman, A.: Neural networks for credit risk evaluation. Expert Systems with Applications **37**, 6233–6239 (2010)

[5] Bellotti, T., Crook, J.: Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications **36**, 3302–3308 (2009)

[6] Chen, C.-C., Li, S.-T.: Credit rating with a monotonicity-constrained support vector machine model. Expert Systems with Applications **41**, 7235–7247 (2014)

[7] Bijak, K., Thomas, L.C.: Does segmentation always improve model performance in credit scoring? Expert Systems with Applications **39**(3), 2433–2442 (2012)

[8] Heckman, J.J.: Sample selection bias as a specification error. Econometrica **47**(1), 153–161 (1979)

[9] Zadrozny, B.: Learning and evaluating classifiers under sample selection bias. In: Proceedings of the 21st International Conference on Machine Learning (ICML), pp. 114–121. ACM, ??? (2004)

[10] Ehrhardt, A., Biernacki, C., Vandewalle, V., Heinrich, P., Beben, S.: Reject inference methods in credit scoring. Journal of Applied Statistics **48**(13–15), 2734–2754 (2021)

[11] Särndal, C.E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, ??? (2003)

[12] Deville, J.-C., Särndal, C.-E.: Calibration estimators in survey sampling. Journal of the American Statistical Association **87**(418), 376–382 (1992)

[13] Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 3rd edn. John Wiley & Sons, ??? (2019)

[14] Banasik, J., Crook, J.: Reject inference, augmentation, and sample selection. European Journal of Operational Research **183**(3), 1582–1594 (2007)

[15] Nguyen, H.T.: Reject inference in application scorecards: evidence from france. EconomiX Working Paper 2016-10, University of Paris Nanterre (2016)

[16] Viennet, E., Fogelman-Soulíe, F., Rognier, B.: Évaluation de techniques de traitement des refusés pour l'octroi de crédit. arXiv preprint cs/0607048 (2006)

[17] Celeux, G., Govaert, G.: A classification em

algorithm for clustering and two stochastic versions. Computational Statistics & Data Analysis **14**(3), 315–332 (1992)

[18] Markatou, M., Basu, A., Lindsay, B.G.: Weighted likelihood estimating equations: The discrete case with applications to logistic regression. Journal of Statistical Planning and Inference **57**, 215–232 (1997)

[19] Kang, J.D.Y., Schafer, J.L.: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science **22**(4), 523–539 (2007)

[20] Scharfstein, D.O., Rotnitzky, A., Robins, J.M.: Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association **94**, 1096–1120 (1999)

[21] Cassel, C.M., Särndal, C.E., Wretman, J.H.: Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika **63**(3), 615–620 (1976)

[22] Särndal, C.E.: On π-inverse weighting versus best linear unbiased weighting in probability sampling. Biometrika **67**(3), 639–650 (1980)

[23] Zhong, B., Rao, J.N.K.: Empirical likelihood inference under stratified random sampling using auxiliary population information. Biometrika **87**(4), 929–938 (2000)

[24] Yeh, I.-C., Lien, C.-H.: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications **36**(2), 2473–2480 (2009)

[25] Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7**, 1–30 (2006)

[26] Manski, C.F.: Partial Identification of Probability Distributions. Springer, ??? (2003)

Table 5: Per-coefficient calibration results across ten suppression proportions π. Columns report the oracle coefficient β_{full}, the calibrated coefficient β_{cal}, their signed difference (Bias), the Wald z-statistic, the raw p-value, the Romano–Wolf corrected p-value (p_{RW}), and the MSE ratio, under the MAR suppression mechanism. Coefficients are statistically indistinguishable from their oracle counterparts up to $\pi = 0.6$; a small bias emerges only under severe suppression ($\pi \geq 0.7$).

π	Coeff	β_{full}	β_{cal}	Bias	z	p_{raw}	p_{RW}	MSE ratio
0.0	x0	0.5148	0.5148	0.0000	0.0000	1.0000	1.0000	1.0000
	x1	1.3172	1.3172	0.0000	0.0000	1.0000	1.0000	1.0000
	x2	−0.0176	−0.0176	0.0000	0.0000	1.0000	1.0000	1.0000
	x3	0.1853	0.1853	0.0000	0.0000	1.0000	1.0000	1.0000
	x4	1.1372	1.1372	0.0000	0.0000	1.0000	1.0000	1.0000
	x5	−0.7975	−0.7975	0.0000	0.0000	1.0000	1.0000	1.0000
	Intercept	0.0576	0.0576	0.0000	0.0000	1.0000	1.0000	1.0000
0.1	x0	0.5118	0.5120	0.0003	0.0100	0.9920	1.0000	1.0308
	x1	1.3192	1.3198	0.0006	0.0180	0.9858	1.0000	1.0668
	x2	−0.0201	−0.0198	0.0003	0.0100	0.9917	1.0000	1.0679
	x3	0.1855	0.1858	0.0003	0.0110	0.9913	1.0000	1.0282
	x4	1.1402	1.1408	0.0006	0.0190	0.9849	1.0000	1.0431
	x5	−0.8004	−0.8003	0.0001	0.0020	0.9981	1.0000	1.0920
	Intercept	0.0554	0.0552	−0.0002	−0.0070	0.9946	1.0000	1.0398
0.2	x0	0.5139	0.5132	−0.0006	−0.0190	0.9846	0.9985	1.3970
	x1	1.3211	1.3210	−0.0001	−0.0020	0.9981	0.9985	1.4316
	x2	−0.0196	−0.0190	0.0006	0.0200	0.9838	0.9985	1.1748
	x3	0.1867	0.1877	0.0010	0.0350	0.9719	0.9930	1.2126
	x4	1.1384	1.1377	−0.0008	−0.0210	0.9830	0.9985	1.3424
	x5	−0.7971	−0.7970	0.0001	0.0030	0.9975	0.9985	1.2924
	Intercept	0.0573	0.0566	−0.0007	−0.0250	0.9797	0.9985	1.1840
0.3	x0	0.5138	0.5129	−0.0009	−0.0280	0.9779	0.9915	1.8153
	x1	1.3173	1.3181	0.0008	0.0190	0.9852	0.9915	1.9721
	x2	−0.0158	−0.0165	−0.0007	−0.0230	0.9813	0.9915	1.5668
	x3	0.1868	0.1861	−0.0006	−0.0200	0.9843	0.9915	1.7037
	x4	1.1337	1.1354	0.0017	0.0350	0.9718	0.9875	2.2965
	x5	−0.7975	−0.7994	−0.0019	−0.0510	0.9593	0.9545	1.7192
	Intercept	0.0577	0.0585	0.0008	0.0270	0.9788	0.9915	1.5257
0.4	x0	0.5131	0.5121	−0.0010	−0.0220	0.9823	0.9910	2.7320
	x1	1.3185	1.3164	−0.0021	−0.0340	0.9730	0.9715	3.7799
	x2	−0.0187	−0.0191	−0.0003	−0.0090	0.9931	0.9925	2.1196
	x3	0.1876	0.1852	−0.0023	−0.0660	0.9475	0.8085	2.0256
	x4	1.1381	1.1379	−0.0002	−0.0030	0.9972	0.9925	3.2002
	x5	−0.8001	−0.7974	0.0027	0.0590	0.9532	0.8300	2.7055
	Intercept	0.0579	0.0574	−0.0005	−0.0150	0.9883	0.9925	1.9917
0.5	x0	0.5129	0.5144	0.0016	0.0270	0.9784	0.9755	4.6362
	x1	1.3209	1.3214	0.0006	0.0060	0.9951	0.9970	6.9333
	x2	−0.0187	−0.0143	0.0044	0.0890	0.9289	0.5350	3.3194
	x3	0.1850	0.1874	0.0024	0.0450	0.9642	0.9185	4.0024
	x4	1.1368	1.1424	0.0056	0.0700	0.9444	0.7210	6.0702
	x5	−0.7959	−0.7955	0.0003	0.0050	0.9964	0.9970	5.3942
	Intercept	0.0543	0.0538	−0.0005	−0.0100	0.9916	0.9970	3.3245
0.6	x0	0.5142	0.5180	0.0038	0.0420	0.9662	0.9445	12.6744
	x1	1.3184	1.3264	0.0080	0.0550	0.9561	0.8785	20.3085
	x2	−0.0178	−0.0190	−0.0012	−0.0160	0.9870	0.9675	8.0313
	x3	0.1871	0.1903	0.0031	0.0400	0.9684	0.9445	9.4205
	x4	1.1359	1.1514	0.0154	0.1200	0.9048	0.2020	19.7011
	x5	−0.7983	−0.8005	−0.0023	−0.0220	0.9827	0.9675	12.5254
	Intercept	0.0550	0.0552	0.0003	0.0040	0.9971	0.9675	8.2499
0.7	x0	0.5150	0.5426	0.0275	0.1910	0.8484	0.0035	31.9140
	x1	1.3197	1.3944	0.0747	0.3420	0.7321	0.0000	56.7646
	x2	−0.0185	−0.0075	0.0110	0.0930	0.9262	0.1590	23.4081
	x3	0.1871	0.2016	0.0145	0.1010	0.9197	0.1590	32.1262
	x4	1.1366	1.1945	0.0579	0.2850	0.7758	0.0005	44.6423
	x5	−0.7984	−0.8347	−0.0363	−0.2140	0.8308	0.0010	42.3696
	Intercept	0.0576	0.0756	0.0180	0.1490	0.8812	0.0305	22.7009
0.8	x0	0.5159	0.5384	0.0225	0.0650	0.9480	0.4650	169.4812
	x1	1.3183	1.5200	0.2017	0.4990	0.6181	0.0000	192.6600
	x2	−0.0197	−0.0431	−0.0234	−0.0780	0.9379	0.4470	134.6879
	x3	0.1850	0.2046	0.0196	0.0600	0.9519	0.4650	173.8204
	x4	1.1362	1.3142	0.1780	0.4730	0.6361	0.0000	188.0067
	x5	−0.7977	−0.8970	−0.0993	−0.2800	0.7792	0.0000	187.0280
	Intercept	0.0579	0.1024	0.0445	0.1610	0.8719	0.0240	111.5092
0.9	x0	0.5158	0.9697	0.4539	0.4050	0.6856	0.0000	2059.1189
	x1	1.3198	2.2883	0.9685	0.7730	0.4396	0.0000	2178.9110
	x2	−0.0177	0.0849	0.1027	0.1000	0.9202	0.3120	1843.8766
	x3	0.1876	0.3079	0.1203	0.1110	0.9118	0.3120	1462.5780
	x4	1.1365	2.1773	1.0408	0.7760	0.4376	0.0000	2452.8182
	x5	−0.7962	−1.5157	−0.7195	−0.6240	0.5328	0.0000	1848.9483
	Intercept	0.0537	0.2761	0.2223	0.2430	0.8084	0.0075	1728.2731

Table 6: Hausman test summary and MSE decomposition across all suppression proportions under the MAR mechanism. The Hausman p-value never falls below 0.965, confirming that the calibrated estimator is jointly indistinguishable from the oracle at every proportion. The mean squared bias remains negligible through $\pi = 0.6$ and rises only under severe suppression; the efficiency loss is otherwise variance-driven.

Proportion π	**Hausman stat.**	**Hausman p-val.**	**Mean sq. bias**	**Mean var. cal**	**Mean var. full**	**Mean MSE ratio**
0.00	0.000	1.0000	0.000000	0.000770	0.000770	1.0000
0.10	0.025	1.0000	0.000000	0.000807	0.000765	1.0526
0.20	0.017	1.0000	0.000000	0.001053	0.000807	1.2907
0.30	0.015	1.0000	0.000001	0.001373	0.000744	1.7999
0.40	0.017	1.0000	0.000003	0.002203	0.000797	2.6507
0.50	0.021	1.0000	0.000009	0.004411	0.000866	4.8115
0.60	0.023	1.0000	0.000047	0.010588	0.000775	12.9873
0.70	0.330	0.9999	0.001667	0.026835	0.000748	36.2751
0.80	0.595	0.9990	0.012236	0.117351	0.000769	165.3134
0.90	1.835	0.9684	0.402761	1.289467	0.000846	1939.2177

Table 7: Held-out test discrimination on the full Lending Club dataset (regime R1). "(common)" uses only the seven shared variables; "(all)" adds the 17 underwriting variables, which are available only to methods that do not require reject-side features. Imputation-based methods fall below the baseline; calibration never does. Numbers are filled from the definitive full-data run.

Strategy	AUC	Gini	Brier
Baseline (common)	0.6421	0.2843	0.1521
Fuzzy Augmentation	0.6421	0.2843	0.1521
Reclassification	0.6267	0.2535	0.1640
Parcelling	0.6420	0.2841	0.1526
Generative	0.6128	0.2257	0.1742
Self-training	0.6061	0.2122	0.1571
IPS Reweighting	0.6415	0.2829	0.1521
Heckman two-step	0.6321	0.2642	0.1574
DR-Calibration	0.6384	0.2768	0.1525
Calibration (common)	0.6384	0.2768	0.1525
Baseline (all)	0.7059	0.4118	0.1442
Calibration (all)	**0.7033**	**0.4066**	**0.1447**

Table 8: Population-estimand evaluation on the full Lending Club dataset (common-space models; metrics are population-weighted on the held-out set, with the accepted observed default rate 0.197). Lower Brier/ECE is better.

Strategy	Pred. pop. rate	Pop. Brier	Pop. ECE
Baseline (common)	0.2579	0.1789	0.0112
Reclassification	0.1226	0.2166	0.1737
Parcelling	0.2926	0.1831	0.0429
Generative	0.3251	0.2248	0.1199
Self-training	0.1825	0.1590	0.0308
IPS Reweighting	0.2535	0.1784	0.0084
DR-Calibration	0.2484	0.1779	0.0043
Calibration (common)	**0.2484**	**0.1778**	**0.0042**

Table 9: Standardised common-variable logistic coefficients (Lending Club, full data). Positive values raise the modelled default probability. Calibration shifts coefficients modestly toward their population values; the semi-supervised model distorts them.

Feature	Baseline	Calibration	Self-training
`loan_amnt`	0.238	0.201	0.196
`fico_score`	-0.426	-0.393	-0.506
`dti`	0.256	0.149	0.096
`emp_length=<1yr`	-0.002	0.017	-1.003
`purpose=credit_card`	-0.092	-0.105	0.407
`purpose=debt_consol.`	0.019	0.030	0.605
`purpose=small_bus.`	0.074	0.071	0.163

Table 10: Taiwan ground-truth evaluation (true-population metrics, true default rate 0.221). Calibration wins under misspecification, marginally under MAR, and, like every method, fails under MNAR.

Regime	Strategy	AUC	Pop. ECE	Pred. rate
MAR	Baseline	0.6213	0.0085	0.219
	Reclassification	0.5519	0.1093	0.112
	Calibration	**0.6222**	**0.0062**	**0.219**
Misspec	Baseline	0.5277	0.0462	0.175
	Calibration	**0.5389**	**0.0073**	**0.214**
MNAR	Baseline	0.5511	0.1351	0.086
	Calibration	**0.5629**	**0.1348**	**0.086**

Table 11: Average Friedman ranks across three tasks (lower is better). The reweighting estimators and the baseline occupy the top; the imputation-based methods occupy the bottom.

Method	**Avg. rank**
Calibration	2.50
IPS Reweighting	2.67
DR-Calibration	2.83
Baseline	3.67
Parcelling	4.50
Self-training	6.50
Heckman two-step	6.67
Generative	7.67
Reclassification	8.00