Cross-Fitted Contamination-Aware Generalized Empirical-Bayes Liu Shrinkage for Multinomial Logit Models under Multicollinearity and Outliers

Sadia Saba

s.saba5@campus.unimib.it

University of Milan-Bicocca Faculty of Economics: Universita degli Studi di Milano-Bicocca Scuola di Economia e Statistica https://orcid.org/0009-0006-3166-8953

Muhammad Amir Saeed

University of Milano-Bicocca https://orcid.org/0000-0003-1650-8194

Research Article

Keywords: multinomial logit, Liu estimator, generalized Bayes, empirical Bayes, density power divergence, response misclassification, leverage outliers, multicollinearity, cross-fitting, Godambe information.

Posted Date: August 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10599718/v1

Additional Declarations: The authors declare no competing interests.

Cross-Fitted Contamination-Aware Generalized Empirical-Bayes Liu Shrinkage for Multinomial Logit Models under Multicollinearity and Outliers

Sadia Saba*[1] and Muhammad Amir Saeed†[1]

[1]Department of Economics, Management and Statistics, University of Milano–Bicocca, Milan, Italy

Abstract

Multinomial logit estimation can be unstable when predictors are nearly collinear and can be distorted by response miscoding and high-leverage observations. We develop a cross-fitted contamination-aware generalized empirical-Bayes Liu estimator (CF-CABLS-MNL) that combines weighted density-power-divergence estimation, a diagonal-dominant class-misclassification model, sandwich/Godambe geometry, and direction-specific spectral shrinkage. Out-of-fold robust fits provide both a prior centre and clean-class probabilities for a MAP–EM estimate of the misclassification matrix, thereby avoiding the direct reuse of one full-sample coefficient estimate as both prior centre and data update. The final robust coefficient estimator is rotated into the eigenspace of the estimated Godambe precision, where eigenvalue-adaptive beta priors determine direction-specific Liu factors and are integrated by deterministic quadrature.

The empirical study contains 16 Monte Carlo scenarios, eight estimators, 100 paired replications per scenario, a 30-replication sensitivity study, and five repeated semi-synthetic train–test experiments on the UCI Dry Bean, Vehicle Silhouettes, and Glass Identification data. Ridge regression minimized coefficient error and predictive log loss in all 16 simulation scenarios, emphasizing the strength of direct regularization when the data-generating coefficient vector is dense and the ridge constant is favorable. Relative to maximum likelihood, however, CF-CABLS reduced average SMSE by about 19%, and the version without the misclassification matrix reduced it by about 30%. Full CF-CABLS achieved the best balanced accuracy in six scenarios, especially under joint response and leverage contamination. In the real-data experiments no estimator dominated uniformly: CF-CABLS achieved the lowest log loss for Dry Bean under combined contamination, whereas MLE, DPD, scalar Liu, and ridge were preferable in several clean or small-sample settings. All reported fits completed without recorded optimization failures. Spectral diagnostics showed a strong positive association between robust-information eigenvalues and posterior Liu factors, confirming that poorly identified directions received stronger shrinkage. The evidence supports CF-CABLS as a transparent robustness–shrinkage framework and ablation device, while also showing that explicit misclassification adjustment should be used selectively rather than assumed to improve every dataset.

Keywords: multinomial logit; Liu estimator; generalized Bayes; empirical Bayes; density power divergence; response misclassification; leverage outliers; multicollinearity; cross-fitting; Godambe information.

*Corresponding author: s.saba5@campus.unimib.it

†m.saeed@campus.unimib.it

1 Introduction

Nominal outcomes with more than two categories occur throughout economics, transportation, marketing, medicine, reliability, and pattern recognition. The baseline-category multinomial logit (MNL) model remains one of the most interpretable and computationally accessible tools for such outcomes [3, 20]. Its maximum-likelihood estimator is attractive under a correctly specified model and a well-conditioned design, but these assumptions are often strained in empirical applications. Morphological measurements, chemical compositions, survey indices, and image-derived descriptors frequently contain near-linear relationships. At the same time, recorded classes can be miscoded and a small number of observations can occupy extreme regions of predictor space.

These difficulties interact. Near-collinearity generates small eigenvalues of the information matrix and large sampling variation in selected coefficient directions. A contaminated response can shift the likelihood centre, while a high-leverage design point can amplify that shift. Applying shrinkage to an already distorted estimator may reduce variance around the wrong centre; applying a robust loss without directional regularization may leave the estimator highly variable in weakly identified directions. The statistical problem is therefore not simply robustification plus shrinkage as two independent add-ons. The nuisance mechanisms, uncertainty geometry, and prior centre must be coordinated.

Biased estimation has a long history as a response to multicollinearity. Ridge regression regularizes all directions through a scalar penalty [16], while Liu estimation uses a related biasing transformation [19]. These ideas have been extended to multinomial logistic models through Liu-type, generalized two-parameter, and newer Dawoud–Kibria estimators [1, 2, 4, 11]. Recent penalized MNL work has also emphasized flexible penalty weights and variable-selection structures [12]. Such methods can sharply reduce mean squared error, but their standard constructions begin from the ordinary likelihood and consequently inherit its sensitivity to miscoding and anomalous covariates.

A complementary robust literature replaces or reweights the likelihood. Density-power-divergence (DPD) estimation supplies a continuous trade-off between likelihood efficiency and resistance to unlikely responses [5]. DPD estimators and tests for polytomous logistic regression were developed by Castilla et al. [8]. Miron, Poilane, and Cantoni [21] demonstrated that response-robust methods can remain vulnerable to extreme covariates and proposed B-robust polytomous estimators. Rényi-pseudodistance estimation offers another robust family [7]; weighted-likelihood procedures address ordered and unordered responses [18]; and recent simplex-based methods use truncation, weighting, or label adjustment to avoid unbounded multicategory losses [26]. These approaches primarily target contamination and do not directly solve variance inflation caused by near-zero information eigenvalues.

The present work joins the two strands in a cross-fitted generalized empirical-Bayes construction. The method is called cross-fitted contamination-aware generalized empirical-Bayes Liu shrinkage for multinomial logit models, abbreviated CF-CABLS-MNL. Its design addresses four issues that arise in a naive robust Bayesian combination.

First, a prior centred at a robust estimate should not be updated with a normal approximation based on exactly the same estimate and observations without acknowledging the resulting data reuse. Cross-fitting is used to learn a pilot centre and out-of-fold (OOF) probabilities. Cross-fitting is widely used to isolate nuisance learning from held-out estimating steps [9]; here its role is not causal orthogonalization, but modular separation of the pilot information from the final full-sample robust update.

Second, generalized-posterior curvature does not generally equal frequentist uncertainty under misspecification. General Bayesian updating can coherently connect a loss to a posterior distribution [6], and DPD-based robust Bayes procedures are an important example [14]. We nevertheless use the empirical sandwich covariance and its inverse Godambe precision as the

geometry for shrinkage, rather than treating the inverse loss Hessian as a sampling covariance [15, 25].

Third, unrestricted joint estimation of regression coefficients and a full class-misclassification matrix can be weakly identified. We estimate a row-stochastic, diagonal-dominant matrix from OOF probabilities under an identity-favouring Dirichlet prior, then hold that matrix fixed during the final robust coefficient fit. This provides an interpretable ablation between an identity label model and explicit contamination adjustment.

Fourth, the spectral shrinkage itself is direction-specific. Rather than assigning one Liu parameter to all coefficients, the Godambe precision is diagonalized and each direction receives an eigenvalue-adaptive beta prior for its Liu factor. Stable directions are left comparatively close to the robust centre, whereas directions associated with small robust-information eigenvalues are pulled more strongly toward the cross-fitted pilot centre.

The empirical results are intentionally interpreted without a universal-superiority claim. In the simulation design, ridge is a formidable benchmark and wins all coefficient-error and log-loss comparisons. CF-CABLS nevertheless improves over unregularized robust estimators in coefficient recovery and often improves class-balanced prediction. The real-data study shows that the misclassification component can help in some contaminated, highly collinear settings and harm in small or clean datasets. This mixed evidence is scientifically useful: it identifies which component is responsible for gains and clarifies when a simpler estimator should be preferred.

The contributions are therefore:

(i) an integrated MNL estimator for multicollinearity, response miscoding, and leverage contamination;

(ii) a cross-fitted robust pilot that supplies both OOF probabilities and a prior centre without direct full-sample self-centring;

(iii) a diagonal-dominant MAP–EM class-misclassification stage separated from the final coefficient optimization;

(iv) a sandwich/Godambe spectral representation and direction-specific generalized empirical-Bayes Liu update;

(v) a complete ablation set separating DPD, leverage weighting, misclassification adjustment, cross-fitting, and spectral shrinkage;

(vi) paired Monte Carlo and semi-synthetic real-data experiments with predictive, coefficient, convergence, runtime, and spectral diagnostics; and

(vii) an empirical finding that explicit misclassification modeling is beneficial only in selected regimes, motivating adaptive matrix selection in future extensions.

The remainder of the paper reviews the relevant literature, develops the estimator and its properties, describes the empirical protocol, reports the simulation and real-data findings, and concludes with limitations and future directions.

2 Related Work and Positioning

2.1 Biased and penalized estimation for multinomial logit models

Let the Fisher or observed information matrix have eigendecomposition $Q\Lambda Q^\top$. When some eigenvalues are small, the variance of the MLE becomes large along the corresponding columns of Q. Ridge adds a positive constant to all eigenvalues, while Liu-type transformations introduce a biasing parameter intended to obtain a favorable bias–variance compromise. In multinomial

models, Abonazel and Farghali [2] proposed a Liu-type estimator nesting ridge and Liu structures. Asar and Erisoglu [4] examined rules for selecting a scalar Liu parameter. Farghali et al. [11] proposed generalized two-parameter estimators and algorithms for parameter selection, and Abonazel et al. [1] introduced a generalized Dawoud–Kibria estimator that further targets severe multicollinearity. These developments establish that direct regularization can dominate MLE in coefficient mean squared error.

The main distinction from CF-CABLS is the centre and the geometry. The preceding estimators shrink a likelihood-based estimate using likelihood information. CF-CABLS first replaces the likelihood centre by a weighted robust M-estimator and then uses a sandwich precision to identify uncertain directions. It is therefore closer to a robust spectral post-processing of a generalized-posterior centre than to a conventional penalized likelihood.

A second distinction concerns the objective of Bayesian shrinkage. Shared global–local priors for MNL regression encourage similar variable-selection patterns across category-specific equations [24]. That is a sparsity objective. CF-CABLS does not assume sparsity and instead targets unstable linear combinations of coefficients defined by the Godambe eigensystem. The two ideas could be combined, but they solve different problems.

2.2 Robust polytomous regression

Minimum-DPD estimation downweights observations whose recorded class has very small model probability, with robustness controlled by $\alpha \geq 0$ [5]. Castilla et al. [8] established DPD estimation, testing, and influence-function results for polytomous logistic regression. Castilla [7] developed minimum Rényi-pseudodistance estimators and showed improved behavior under response misclassification. These losses principally limit the contribution of anomalous responses.

Extreme covariates require separate treatment because the score still contains the design vector. Miron, Poilane, and Cantoni [21] showed that leverage contamination can have a larger effect than misclassification and developed robust GLM-type and optimal B-robust polytomous estimators. Iannario and Monti [18] proposed robust M-estimation through weighted likelihood for both ordered and unordered logistic responses. Zhang et al. [26] approached robustness through simplex coding and truncated, weighted, and label-adjusted multicategory learning. CF-CABLS follows the same broad principle that response and design robustness are distinct: DPD controls unlikely recorded categories, while robust-distance weights limit design leverage.

2.3 Misclassification and latent clean labels

Response miscoding may be addressed either by modeling label errors or by constructing losses that are insensitive to them. Robust mislabel logistic regression can avoid explicit estimation of mislabel probabilities [17], while latent-label approaches introduce a transition matrix. The latter is attractive when asymmetric class-to-class confusion is scientifically interpretable, but it is difficult to identify without restrictions. CF-CABLS uses an identity-favouring prior and enforces diagonal dominance. The matrix is estimated from OOF clean-class probabilities, not by unrestricted simultaneous optimization with the regression coefficients.

2.4 Generalized Bayes, sandwich calibration, and cross-fitting

General Bayesian updating replaces the negative log likelihood by a task-relevant loss [6]. The resulting generalized posterior depends on a learning rate and its Hessian need not describe repeated-sampling uncertainty. Under misspecification, sandwich covariance is the natural large-sample uncertainty for an M-estimator [25]. CF-CABLS therefore separates three objects: the robust loss defines the coefficient centre, the sandwich defines the uncertainty geometry, and the empirical-Bayes prior defines the spectral shrinkage.

Cross-fitting is used to reduce overfitting and data-reuse effects in nuisance estimation [9]. In the present method, each held-out observation receives a prediction from a model fitted without that observation. These OOF predictions support estimation of the transition matrix and the fold coefficients support the prior centre. The final estimator still uses the complete training sample, but its prior centre is not the final coefficient vector itself.

2.5 Position of CF-CABLS-MNL

CF-CABLS-MNL is best described as a *cross-fitted generalized empirical-Bayes spectral estimator*. It is not fully Bayesian because the transition matrix, pilot centre, sandwich covariance, and hyperparameters are plug-in quantities. It is not merely a Liu estimator because the centre is robust and the direction-specific factors are integrated under adaptive priors. It is not merely a robust M-estimator because its final coefficient vector is spectrally regularized. This positioning is narrower and more defensible than claiming novelty for any individual ingredient.

3 Methodology

3.1 Baseline-category multinomial logit model

For independent observations (Y_i, x_i), let $Y_i \in \{1, \dots, C\}$ and $x_i = (1, z_i^\top)^\top \in \mathbb{R}^q$, where $z_i \in \mathbb{R}^p$ and $q = p + 1$. Category C is the reference. With $K = C - 1$ and coefficient matrix $B = (\beta_1, \dots, \beta_K) \in \mathbb{R}^{q \times K}$,

$$\pi_{ih}(B) = \frac{\exp(x_i^\top \beta_h)}{1 + \sum_{r=1}^{K} \exp(x_i^\top \beta_r)}, \qquad h = 1, \dots, K, \tag{1}$$

$$\pi_{iC}(B) = \frac{1}{1 + \sum_{r=1}^{K} \exp(x_i^\top \beta_r)}. \tag{2}$$

The stacked parameter is

$$\boldsymbol{\beta} = \mathrm{vec}(B) = (\beta_1^\top, \dots, \beta_K^\top)^\top \in \mathbb{R}^P, \qquad P = qK.$$

Numerical evaluation uses a row-wise log-sum-exp stabilization. Continuous predictors are standardized using the training sample only; the intercept is appended afterward and excluded from leverage calculations.

3.2 Class-dependent response misclassification

Let $Y_i^\star$ be the latent clean category and Y_i the recorded category. Define the transition matrix

$$M_{gh} = \Pr(Y_i = h \mid Y_i^\star = g), \qquad M_{gh} \geq 0, \qquad \sum_{h=1}^{C} M_{gh} = 1. \tag{3}$$

If $\boldsymbol{\pi}_i(\boldsymbol{\beta}) = (\pi_{i1}, \dots, \pi_{iC})^\top$, the recorded-category probability vector is

$$\boldsymbol{r}_i(\boldsymbol{\beta}, \mathbf{M}) = \mathbf{M}^\top \boldsymbol{\pi}_i(\boldsymbol{\beta}), \qquad r_{ih} = \sum_{g=1}^{C} \pi_{ig} M_{gh}. \tag{4}$$

The identity matrix corresponds to no misclassification. For each row,

$$M_{g\cdot} \sim \mathrm{Dirichlet}(a_{g1}, \dots, a_{gC}), \qquad a_{gg} = a_D > a_O = a_{gh} \quad (h \neq g), \tag{5}$$

and the fitted matrix is normalized and projected so that $M_{gg} \geq m_0 > 1/2$. The restriction reduces label switching and prevents an estimated clean class from being more likely to transition to a different recorded class than to itself.

3.3 Weighted density-power-divergence objective

Let $y_{ih} = \mathbf{1}(Y_i = h)$. For $\alpha > 0$, the categorical DPD contribution based on recorded probabilities is

$$\rho_{\alpha,i}(\boldsymbol{\beta}; \mathbf{M}) = \sum_{h=1}^{C} r_{ih}^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) \sum_{h=1}^{C} y_{ih} r_{ih}^{\alpha}. \tag{6}$$

As $\alpha \downarrow 0$, the objective converges, up to a parameter-free term, to negative multinomial log likelihood. Given nonnegative leverage weights v_i, the criterion is

$$\mathcal{L}_{\alpha,v}(\boldsymbol{\beta}; \mathbf{M}) = \frac{1}{\sum_i v_i} \sum_{i=1}^{n} v_i \rho_{\alpha,i}(\boldsymbol{\beta}; \mathbf{M}) + \lambda_{\text{num}} \|D\boldsymbol{\beta}\|_2^2, \tag{7}$$

where D excludes intercept coordinates and λ_{num} is a small numerical stabilizer rather than the substantive Liu parameter.

For implementation, write $G_{ih} = \partial \rho_{\alpha,i} / \partial r_{ih}$. For $\alpha > 0$,

$$G_{ih} = (1 + \alpha) r_{ih}^{\alpha} - (1 + \alpha) y_{ih} r_{ih}^{\alpha - 1}. \tag{8}$$

Using $\partial r_{ih} / \partial \eta_{ik} = \pi_{ik}(M_{kh} - r_{ih})$ for $k \leq K$,

$$\frac{\partial \rho_{\alpha,i}}{\partial \eta_{ik}} = \pi_{ik} \sum_{h=1}^{C} G_{ih}(M_{kh} - r_{ih}), \tag{9}$$

and the coefficient gradient follows by premultiplication with x_i. This analytic gradient is used by BFGS, with fallback optimizers retained for difficult cases.

3.4 Robust leverage weights

Let z_i denote the standardized non-intercept predictors. A robust centre and scatter $(\widehat{\mu}_R, \widehat{\Sigma}_R)$ are obtained by minimum covariance determinant when feasible, with robust and classical fallbacks [22]. Define

$$D_i^2 = (z_i - \widehat{\mu}_R)^{\top} \widehat{\Sigma}_R^{-1} (z_i - \widehat{\mu}_R), \tag{10}$$

and

$$v_i = \min\left\{1, \frac{c_{p,\tau}}{\max(D_i, 10^{-12})}\right\}, \qquad c_{p,\tau} = \sqrt{\chi^2_{p,\tau}}. \tag{11}$$

The weights equal one inside the robust distance cutoff and decay inversely outside it. They are recomputed within each cross-fitting training fold and once on the full training sample. No test observation contributes to standardization, scatter estimation, or tuning.

3.5 Cross-fitted robust pilot

Partition the training indices into L stratified folds $\mathcal{I}_1, \ldots, \mathcal{I}_L$. For fold ℓ, fit

$$\widehat{\boldsymbol{\beta}}^{(-\ell)} = \arg\min_{\boldsymbol{\beta}} \mathcal{L}_{\alpha,v}^{(-\ell)}(\boldsymbol{\beta}; \mathbf{I}_C) \tag{12}$$

on the observations outside $\mathcal{I}_\ell$, and predict $\widehat{\boldsymbol{\pi}}_i^{(-\ell)}$ for $i \in \mathcal{I}_\ell$. Collecting these predictions produces the OOF probability matrix $\widehat{\Pi}^{\text{OOF}}$. The pilot centre is the coordinatewise median of fold coefficients,

$$\widetilde{\boldsymbol{\beta}}_{\text{CF},j} = \text{median}_{\ell=1,\ldots,L}\, \widehat{\beta}_j^{(-\ell)}. \tag{13}$$

The median is used rather than the mean to limit the effect of an unstable fold fit.

3.6 MAP–EM estimation of the misclassification matrix

Conditional on OOF probabilities, the posterior probability that observation i has clean class g is

$$T_{ig} = \Pr(Y_i^\star = g \mid Y_i, \widehat{\boldsymbol{\pi}}_i^{\mathrm{OOF}}, \mathbf{M}) = \frac{\widehat{\pi}_{ig}^{\mathrm{OOF}} M_{g,Y_i}}{\sum_{s=1}^{C} \widehat{\pi}_{is}^{\mathrm{OOF}} M_{s,Y_i}}. \tag{14}$$

The MAP M-step is

$$M_{gh}^{\mathrm{new}} = \frac{a_{gh} - 1 + \sum_{i=1}^{n} T_{ig} \mathbf{1}(Y_i = h)}{\sum_{r=1}^{C} \left[a_{gr} - 1 + \sum_{i=1}^{n} T_{ig} \mathbf{1}(Y_i = r)\right]}. \tag{15}$$

Rows are normalized and projected to diagonal dominance after each update. Iteration stops when the maximum absolute matrix change is below a tolerance. The result $\widehat{\mathbf{M}}$ is then fixed.

3.7 Full robust centre

The full-sample robust, contamination-adjusted centre is

$$\widehat{\boldsymbol{\beta}}_R = \arg\min_{\boldsymbol{\beta}} \mathcal{L}_{\alpha,v}(\boldsymbol{\beta}; \widehat{\mathbf{M}}), \tag{16}$$

initialized at $\widetilde{\boldsymbol{\beta}}_{\mathrm{CF}}$. In the no-misclassification ablation, $\widehat{\mathbf{M}}$ is replaced by $\mathbf{I}_C$. In the no-shrinkage ablation, $\widehat{\boldsymbol{\beta}}_R$ itself is used for prediction.

3.8 Sandwich covariance and Godambe precision

Let $\psi_i(\boldsymbol{\beta})$ be the weighted estimating-function contribution associated with (7). Define

$$\mathbf{H} = E\left\{\frac{\partial \psi_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^\top}\right\}, \qquad \mathbf{J} = E\{\psi_i(\boldsymbol{\beta}_0)\psi_i(\boldsymbol{\beta}_0)^\top\}. \tag{17}$$

The asymptotic covariance of the robust centre is

$$\mathbf{V} = \frac{1}{n}\mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1}. \tag{18}$$

The implementation estimates $\mathbf{J}$ by the empirical outer product of per-observation scores and uses a positive-semidefinite Gauss–Newton sensitivity for $\mathbf{H}$. Small eigenvalue floors are applied only to repair numerical non-positive definiteness. The robust precision is

$$\widehat{\mathbf{G}} = (\widehat{\mathbf{V}} + \eta_V \mathbf{I}_P)^{-1} = \mathbf{U}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_P)\mathbf{U}^\top. \tag{19}$$

Set

$$\widehat{\boldsymbol{\theta}}_R = \mathbf{U}^\top \widehat{\boldsymbol{\beta}}_R, \qquad \widetilde{\boldsymbol{\theta}}_{\mathrm{CF}} = \mathbf{U}^\top \widetilde{\boldsymbol{\beta}}_{\mathrm{CF}}. \tag{20}$$

Small λ_j identify weakly estimated directions.

3.9 Direction-specific generalized empirical-Bayes Liu shrinkage

For each direction j,

$$\theta_j \mid \delta_j, \widetilde{\theta}_{\mathrm{CF},j} \sim N\left(\delta_j \widetilde{\theta}_{\mathrm{CF},j}, \kappa_j^{-1}\right), \qquad 0 < \delta_j < 1. \tag{21}$$

Let $\lambda_\star$ be the median positive robust-information eigenvalue and define

$$s_j = \frac{\lambda_j}{\lambda_j + \lambda_\star}, \quad a_j = a_0 + a_1 s_j, \quad b_j = b_0 + b_1(1 - s_j), \quad \delta_j \sim \mathrm{Beta}(a_j, b_j). \tag{22}$$

The prior precision is scaled to the problem by

$$\kappa_j = \kappa_0 \lambda_\star. \tag{23}$$

Under the working approximation

$$\widehat{\theta}_{Rj} \mid \theta_j \dot{\sim} N(\theta_j, \lambda_j^{-1}), \tag{24}$$

the conditional posterior mean is

$$m_j(\delta_j) = \frac{\lambda_j \widehat{\theta}_{Rj} + \kappa_j \delta_j \widetilde{\theta}_{\mathrm{CF},j}}{\lambda_j + \kappa_j}. \tag{25}$$

The marginal density of δ_j is proportional to

$$\delta_j^{a_j - 1}(1 - \delta_j)^{b_j - 1} \phi\Big(\widehat{\theta}_{Rj}; \delta_j \widetilde{\theta}_{\mathrm{CF},j}, \lambda_j^{-1} + \kappa_j^{-1}\Big), \tag{26}$$

where $\phi(\cdot; \mu, \sigma^2)$ denotes a normal density. Gauss–Legendre quadrature yields $\widehat{\delta}_j = E(\delta_j \mid \widehat{\theta}_{Rj})$. The final estimator is

$$\widehat{\theta}_{\text{CF-CABLS},j} = \frac{\lambda_j \widehat{\theta}_{Rj} + \kappa_j \widehat{\delta}_j \widetilde{\theta}_{\mathrm{CF},j}}{\lambda_j + \kappa_j}, \tag{27}$$

$$\widehat{\boldsymbol{\beta}}_{\text{CF-CABLS}} = \mathbf{U} \widehat{\boldsymbol{\theta}}_{\text{CF-CABLS}}. \tag{28}$$

The shrinkage multiplier is not a simple scalar. Both $\lambda_j/(\lambda_j + \kappa_j)$ and $\widehat{\delta}_j$ vary with direction. When λ_j is large, the robust estimate receives high weight and the beta prior tends to allow a larger δ_j. When λ_j is small, the estimator moves more strongly toward a shrunk cross-fitted centre.

3.10 Learning-rate diagnostic

A generalized posterior proportional to $\exp\{-\omega n \mathcal{L}_{\alpha,v}(\boldsymbol{\beta})\}$ has local covariance approximately $(n\omega\mathbf{H})^{-1}$, which cannot generally reproduce the full sandwich. A scalar magnitude diagnostic is

$$\widehat{\omega}_{\mathrm{mag}} = \frac{P}{\mathrm{tr}(\widehat{\mathbf{H}}^{-1}\widehat{\mathbf{J}})}. \tag{29}$$

This value is reported, but the point estimator uses $\widehat{\mathbf{V}}$ directly and is not based on pretending that one learning rate matches every direction.

3.11 Algorithm and computational flow

Figure 1 summarizes the estimator. The main computational cost is the cross-fitted robust pilot. The ablation study deliberately repeats several stages so that each named estimator is independently reproducible; a production implementation could cache common pilots and robust weights.

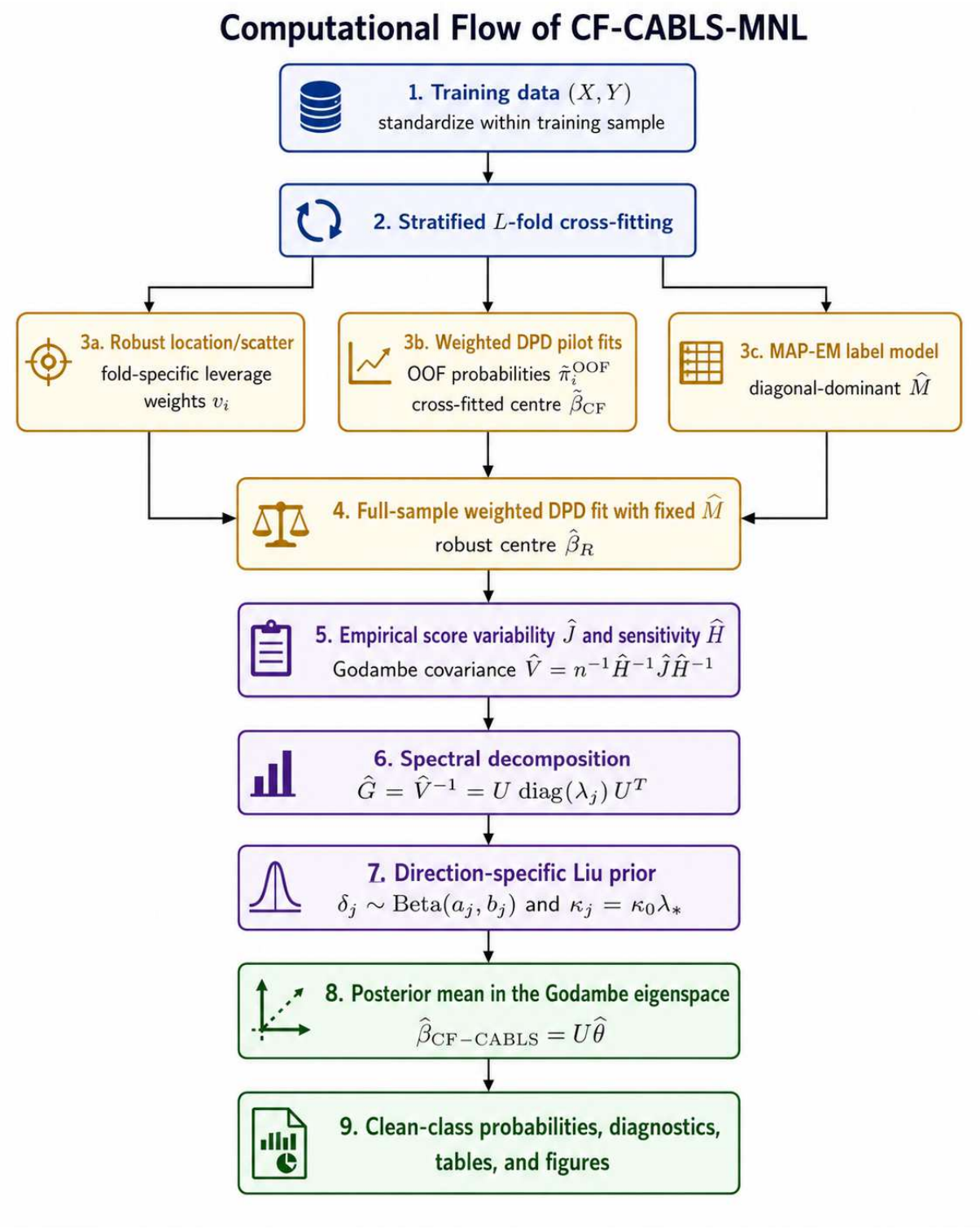


Figure 1: Computational flow of CF-CABLS-MNL. Every fold-specific transformation is learned without the corresponding held-out observations.

Table 1: Estimators and ablation components.

Estimator	DPD	Leverage	$\widehat{\mathbf{M}}$	Liu shrinkage	Role
MLE	No	No	No	No	Likelihood benchmark
Ridge	No	No	No	Scalar penalty	Collinearity benchmark
Scalar Liu	No	No	No	Scalar Liu	Classical biased-estimation benchmark
DPD	Yes	No	No	No	Response robustness
WDPD	Yes	Yes	No	No	Response and leverage robustness
WDPD-M	Yes	Yes	Yes	No	Misclassification-adjusted robust centre
CF-CABLS-noM	Yes	Yes	No	Spectral	Cross-fitted spectral shrinkage
CF-CABLS	Yes	Yes	Yes	Spectral	Complete proposed construction

4 Theoretical Properties

Condition 1 (Regularity) *Observations are independent; the clean MNL parameter is locally identifiable for the selected reference category; standardized predictors have finite second moments; the robust scatter matrix is positive definite after numerical regularization; the diagonal-dominant transition model is locally identifiable;* $\alpha > 0$ *is fixed;* $\mathbf{H}$ *is nonsingular; and all Dirichlet and beta hyperparameters are positive.*

Proposition 1 (Conditional propriety) *Conditional on* $\widehat{\mathbf{M}}$, $\widetilde{\boldsymbol{\beta}}_{\mathrm{CF}}$, $\widehat{\mathbf{V}}$, *and positive* a_j, b_j, κ_j, *the direction-wise generalized empirical-Bayes posterior defined by* (21), (24), *and* (22) *is proper.*

Proof sketch. For each $\delta_j \in (0,1)$, the normal working density and normal prior produce a proper conditional posterior in θ_j. The beta density is integrable because $a_j, b_j > 0$. Tonelli's theorem therefore gives a finite joint integral over (θ_j, δ_j).

Proposition 2 (Asymptotic distribution of the robust centre) *Suppose* $\widehat{\mathbf{M}} \to_p \mathbf{M}_0$ *at a rate sufficient for its first-stage remainder to be* $o_p(n^{-1/2})$. *Under the regularity condition,*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}_0) \xrightarrow{d} N\left(0, \mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1}\right). \tag{30}$$

Proof sketch. The final centre solves a smooth weighted M-estimating equation with a plug-in nuisance matrix. A Taylor expansion around $\boldsymbol{\beta}_0$, a law of large numbers for the derivative, and a central limit theorem for the score yield the sandwich limit. The stated rate condition controls the matrix-estimation remainder. Cross-fitting reduces direct overfitting of the nuisance probabilities, but does not by itself remove the need for a rate argument.

Proposition 3 (Response and leverage control) *For fixed* $\alpha > 0$ *and row-stochastic* $\mathbf{M}$, *the DPD score contribution with respect to the linear predictors is bounded as a function of the recorded category. If* v_i *is defined by* (11), *then* $v_i\|z_i\|$ *is bounded outside a compact set whenever* $\widehat{\Sigma}_R$ *is positive definite.*

Interpretation. DPD limits the effect of observations assigned very small recorded-class probability, while the leverage weight limits growth caused by extreme predictors. Neither protection substitutes for the other. The empirical ablations test the separate and joint value of these components.

Remark 1 (Special cases) *With* $\alpha = 0$, $v_i = 1$, $\widehat{\mathbf{M}} = \mathbf{I}_C$, *and* $\kappa_j \downarrow 0$, *the method reduces to MLE. With* $\kappa_j \downarrow 0$ *and* $\alpha > 0$, *it reduces to weighted DPD. With* $\widehat{\mathbf{M}} \neq \mathbf{I}_C$ *and no spectral update, it is the contamination-adjusted robust centre. If the same coefficient vector is used as both centre and update and every* $\delta_j = d$, *equation* (25) *recovers the familiar Liu-type multiplier after suitable normalization; this is a diagnostic relationship rather than the recommended cross-fitted construction.*

5 Empirical Study

5.1 Empirical protocol and computational configuration

The empirical protocol uses 100 paired Monte Carlo replications for each method–scenario combination, a clean test sample of 2,000 observations per replication, 30 paired replications for the tuning-parameter sensitivity study, and five repeated stratified train–test splits for each real dataset. Within every real-data split, contamination is introduced only into the training sample, while the test sample remains unchanged. All methods are evaluated on common replications or splits, so method comparisons are paired. These settings provide a computationally feasible assessment across 16 simulation scenarios, eight estimators, three real datasets, and multiple contamination mechanisms; small differences between closely performing methods are interpreted cautiously.

Table 2: Empirical configuration reconstructed from the supplied result files.

Component	Setting
Simulation study	16 scenarios, 100 paired replications, 8 estimators, and an independent clean test sample of 2,000 observations per replication.
Sensitivity study	30 paired replications for $\alpha \in \{0.05, 0.10, 0.20, 0.30, 0.50\}$ and $\kappa_0 \in \{0.02, 0.05, 0.10, 0.20, 0.50\}$.
Real-data study	Dry Bean, Vehicle Silhouettes, and Glass Identification; 5 repeated stratified splits; 30% test fraction; contamination applied only to the training portion.
Cross-fitting	Five stratified folds for the robust pilot and out-of-fold probabilities.
Training scenarios	Clean; 10% symmetric response contamination; 10% asymmetric response contamination; 10% leverage contamination; and combined 10% response plus 10% leverage contamination.
Evaluation	SMSE for simulations; clean-test log loss, multiclass Brier score, accuracy, macro-F1, balanced accuracy, optimizer failure rate, and elapsed time.

5.2 Monte Carlo design

Equicorrelated predictors are generated as

$$x_{ij} = \sqrt{1-\rho}\, z_{ij} + \sqrt{\rho}\, z_{i0}, \qquad z_{ij}, z_{i0} \overset{\text{iid}}{\sim} N(0,1), \tag{31}$$

so that $\mathrm{Corr}(x_{ij}, x_{ik}) = \rho$ for $j \neq k$. A deterministic coefficient matrix is scaled to a fixed Frobenius norm and clean responses are sampled from the MNL probabilities. Response contamination is applied after generating clean labels. Leverage contamination shifts selected predictor vectors after their clean labels are drawn, producing discordant high-leverage observations. The 16 scenarios span $n \in \{100, 300\}$, $p \in \{5, 10\}$, $C \in \{3, 5\}$, $\rho \in \{0.50, 0.90, 0.99\}$, and clean, response, leverage, and combined contamination settings.

Table 3: Monte Carlo scenarios represented in the supplied outputs.

ID	n	p	C	ρ	ϵ_y	ϵ_x
S1	100	5	3	0.50	0.00	0.00
S2	100	5	3	0.90	0.00	0.00
S3	100	5	3	0.99	0.00	0.00
S4	100	5	3	0.90	0.10	0.00
S5	100	5	3	0.90	0.00	0.10
S6	100	5	3	0.90	0.10	0.10
S7	100	5	3	0.99	0.10	0.00
S8	100	5	3	0.99	0.00	0.10
S9	100	5	3	0.99	0.10	0.10
S10	300	10	5	0.50	0.00	0.00
S11	300	10	5	0.90	0.00	0.00
S12	300	10	5	0.99	0.00	0.00
S13	300	10	5	0.90	0.10	0.00
S14	300	10	5	0.90	0.00	0.10
S15	300	10	5	0.90	0.10	0.10
S16	300	10	5	0.99	0.10	0.10

The primary coefficient metric is simulated squared error,

$$\text{SMSE} = R^{-1} \sum_{r=1}^{R} \|\widehat{\boldsymbol{\beta}}^{(r)} - \boldsymbol{\beta}_0\|_2^2. \tag{32}$$

Predictive metrics are evaluated on an independently generated clean test set. All estimators use the same generated training and test data within a replication.

5.3 Real-data and semi-synthetic design

The real-data study uses three UCI datasets: Dry Bean [10], Statlog Vehicle Silhouettes [23], and Glass Identification [13]. The original labels are not asserted to be contaminated. Instead, each repeated split preserves an unchanged test set and introduces controlled contamination only into the standardized training predictors or training labels. This design evaluates recovery of clean predictive behavior under known training contamination.

Table 4: Real-data characteristics before train-test splitting.

Dataset	n	p	Classes	Imbalance	Max. $\lvert r \rvert$	Correlation condition no.
Dry Bean	13611	16	7	6.79	0.9999	4.97e+06
Vehicle	846	18	4	1.10	0.9963	2.60e+04
Glass	214	9	6	8.44	0.8104	1.56e+03

Dry Bean provides an extreme collinearity case, with maximum absolute correlation approximately 0.9999 and a predictor-correlation condition number near 5×10^6. Vehicle is also strongly correlated but nearly class-balanced. Glass is small and highly imbalanced, with only nine observations in its smallest class. These structures create substantially different estimation regimes.

6 Results

6.1 Simulation performance

Figure 2 displays coefficient error on a logarithmic scale. Ridge has the lowest SMSE in every scenario. The advantage becomes especially large when $\rho = 0.99$, where the unregularized MLE, DPD, and WDPD estimators have very large coefficient error. This result confirms that direct regularization is indispensable in the chosen dense-coefficient simulation design.

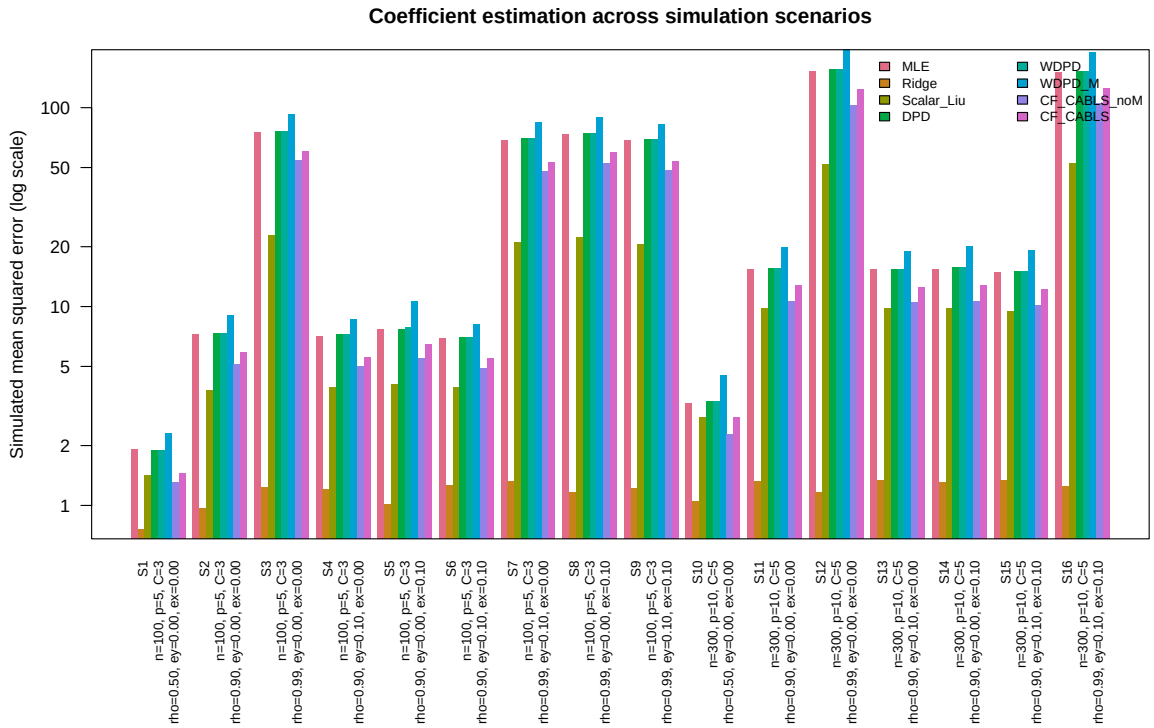


Figure 2: Simulation coefficient error across the 16 scenarios. The vertical axis is logarithmic.

The ranking summary in Table 5 quantifies the pattern. Ridge's average relative SMSE is 0.111, followed by scalar Liu at 0.513, CF-CABLS without the misclassification matrix at 0.698, and full CF-CABLS at 0.810. Thus, relative to MLE, the cross-fitted spectral update reduces average coefficient error even when it does not match the favorable fixed ridge benchmark. The no-matrix version consistently improves on the full version in coefficient error, suggesting that transition-matrix estimation introduces additional variability when the simulated contamination mechanism is not sufficiently identified by the OOF probabilities.

Table 5: Across-scenario simulation rankings and win counts. Lower ranks are better.

Estimator	Mean rel. SMSE	SMSE rank	Log-loss rank	Bal.-acc. rank	SMSE wins	Bal.-acc. wins
Ridge	0.111	1.00	1.00	6.25	16	4
Scalar Liu	0.513	2.12	2.12	4.06	0	1
CF-CABLS-noM	0.698	2.88	2.88	3.75	0	2
CF-CABLS	0.810	4.00	5.16	3.28	0	6
MLE	1.000	5.12	4.81	4.94	0	1
DPD	1.013	6.31	6.06	4.94	0	0
WDPD	1.015	6.56	5.97	4.59	0	0
WDPD-M	1.258	8.00	8.00	4.19	0	2

Note: Relative SMSE is normalized to MLE within each scenario before averaging. Ridge also achieved the smallest mean log loss in all 16 scenarios.

Predictive log loss in Figure 3 follows the SMSE ranking. Ridge has the smallest mean log loss in all 16 scenarios. Differences among MLE, DPD, WDPD, and CF-CABLS are modest on the predictive scale compared with the large coefficient-error differences, illustrating that substantial coefficient instability can coexist with similar predicted probabilities under multicollinearity.

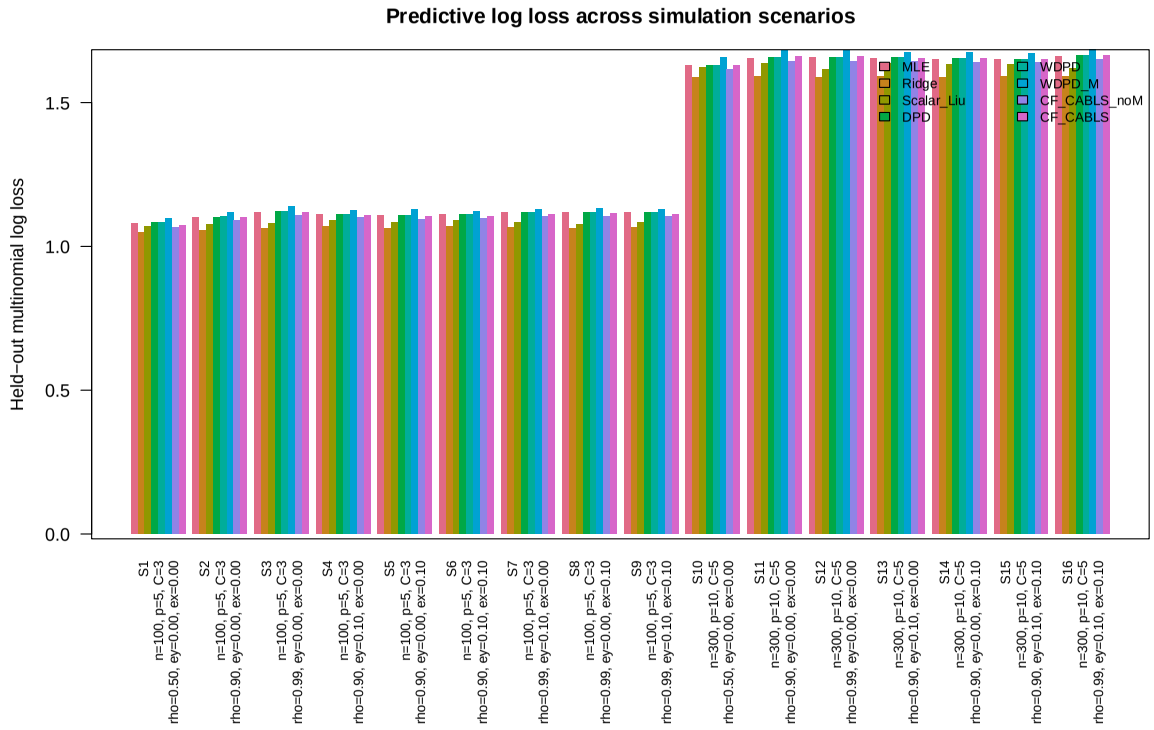


Figure 3: Clean-test multinomial log loss across simulation scenarios.

Balanced accuracy gives a different ranking (Figure 4). The full method is the designated winner in six scenarios, ridge in four, CF-CABLS-noM and WDPD-M in two each, and MLE and scalar Liu in one each. The average balanced-accuracy differences are small, but the ordering indicates that the robust cross-fitted methods can redistribute class-specific errors even when they do not minimize proper scoring rules.

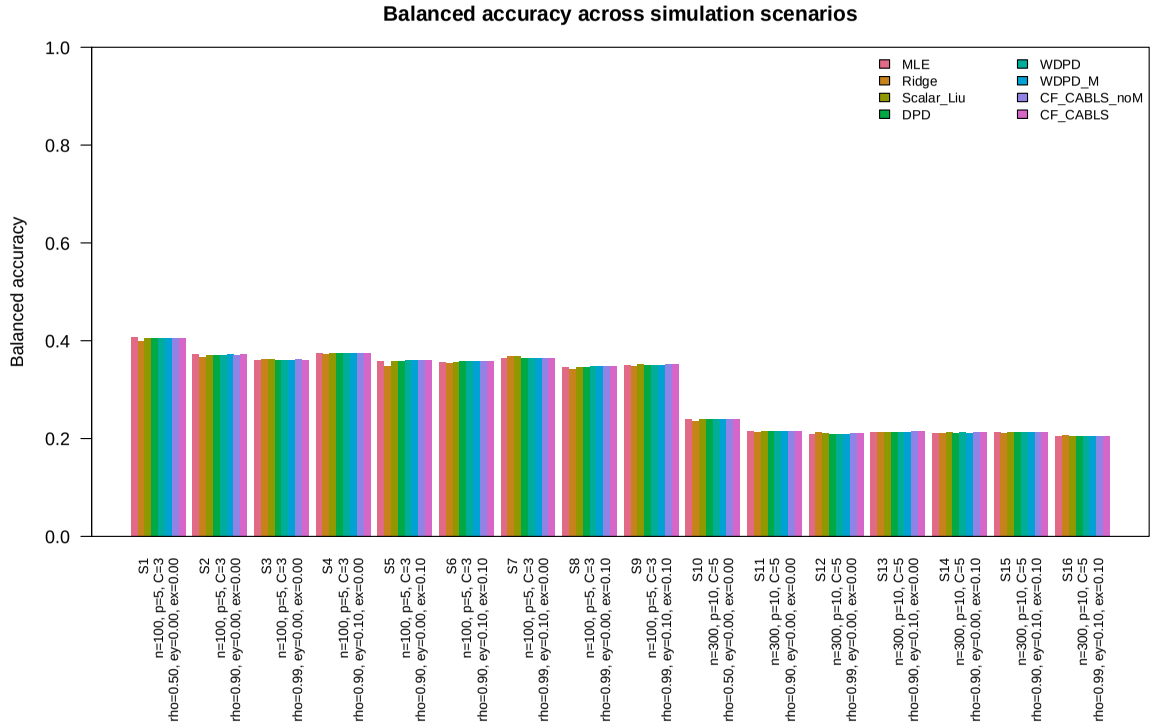


Figure 4: Clean-test balanced accuracy across simulation scenarios.

Selected numerical results appear in Table 6. In severe collinearity, ridge's relative SMSE can be below 0.02, while the cross-fitted spectral estimators retain approximately 67–81% of MLE's error. This is meaningful regularization but not enough to overcome the fixed ridge specification. The result should not be presented as universal superiority of CF-CABLS. Rather, it demonstrates that robust spectral shrinkage improves the robust centre while preserving contamination-oriented components and class-balanced performance.

Table 6: Selected simulation results spanning clean, severely collinear, and jointly contaminated settings.

Scenario	Estimator	SMSE	Rel. SMSE	Log loss	Balanced accuracy
S1	CF-CABLS	1.458	0.766	1.073	0.406
S1	CF-CABLS-noM	1.304	0.685	1.066	0.406
S1	MLE	1.904	1.000	1.081	0.406
S1	Ridge	0.755	0.396	1.048	0.399
S1	Scalar Liu	1.409	0.740	1.068	0.406
S3	CF-CABLS	60.497	0.801	1.117	0.361
S3	CF-CABLS-noM	54.188	0.718	1.107	0.361
S3	MLE	75.503	1.000	1.120	0.361
S3	Ridge	1.234	0.016	1.061	0.363
S3	Scalar Liu	22.856	0.303	1.079	0.362
S6	CF-CABLS	5.464	0.785	1.105	0.359
S6	CF-CABLS-noM	4.916	0.707	1.097	0.358
S6	MLE	6.956	1.000	1.109	0.357
S6	Ridge	1.266	0.182	1.070	0.354
S6	Scalar Liu	3.943	0.567	1.090	0.357
S9	CF-CABLS	53.770	0.786	1.112	0.352
S9	CF-CABLS-noM	48.413	0.708	1.105	0.351
S9	MLE	68.385	1.000	1.116	0.350
S9	Ridge	1.220	0.018	1.067	0.348
S9	Scalar Liu	20.651	0.302	1.081	0.351
S10	CF-CABLS	2.773	0.855	1.630	0.239
S10	CF-CABLS-noM	2.266	0.698	1.615	0.239
S10	MLE	3.244	1.000	1.628	0.239
S10	Ridge	1.047	0.323	1.586	0.236
S10	Scalar Liu	2.770	0.854	1.621	0.239
S15	CF-CABLS	12.182	0.823	1.651	0.214
S15	CF-CABLS-noM	10.196	0.689	1.638	0.214
S15	MLE	14.795	1.000	1.650	0.213
S15	Ridge	1.335	0.090	1.592	0.211
S15	Scalar Liu	9.462	0.640	1.633	0.213
S16	CF-CABLS	125.401	0.834	1.665	0.205
S16	CF-CABLS-noM	104.730	0.697	1.650	0.205
S16	MLE	150.349	1.000	1.662	0.205
S16	Ridge	1.240	0.008	1.591	0.207
S16	Scalar Liu	52.235	0.347	1.619	0.206

Scenario definitions are given in Table 3. Standard errors are available in the supplied full result tables.

No simulation optimization failures were recorded for any estimator in any scenario. This is an important implementation result because the proposed estimator requires several fold-specific and full-sample optimizations. It also confirms that the reported performance differences are not artifacts of selectively discarded failures.

6.2 Real-data predictive performance

The real-data results are more heterogeneous than the simulations. Table 7 identifies the best estimator for each dataset, contamination scenario, and metric. MLE and DPD are strongest for Dry Bean under clean or response-only conditions, while CF-CABLS attains the lowest log loss under combined response and leverage contamination. WDPD-M attains the lowest Dry Bean Brier score in the combined and asymmetric scenarios. This is the clearest empirical regime in which the estimated transition matrix is useful.

For Vehicle, scalar Liu is strongest on clean and leverage-only log loss, whereas DPD dominates most contaminated scenarios and balanced-accuracy comparisons. Vehicle combines strong collinearity with balanced classes, making direct robust estimation effective without the additional variability of the transition matrix.

For Glass, ridge or scalar Liu is usually best for proper scoring rules, while DPD often achieves the best macro-F1 and balanced accuracy. The full CF-CABLS and WDPD-M fits can produce very large log loss in this small, imbalanced dataset. This behavior is consistent with weak identification of a 6×6 transition matrix from a small training sample and is an important caution against unconditional use of explicit misclassification adjustment.

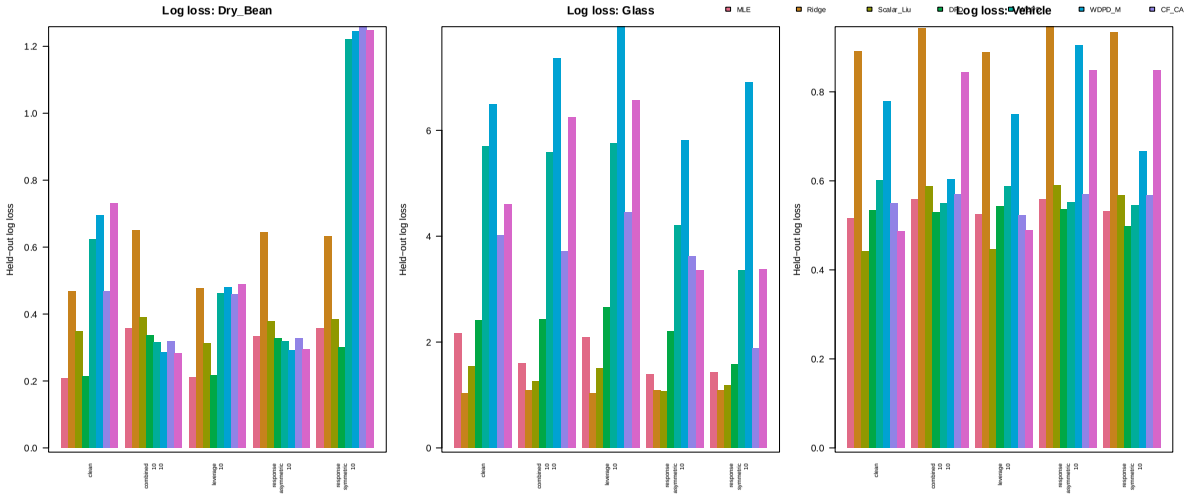
Table 7: Best-performing estimator by real dataset, contamination scenario, and metric.

Dataset	Training scenario	Log loss	Brier	Balanced accuracy
Dry Bean	Clean	MLE	MLE	MLE
Dry Bean	Symmetric 10%	DPD	DPD	DPD
Dry Bean	Asymmetric 10%	WDPD-M	WDPD-M	DPD
Dry Bean	Leverage 10%	MLE	MLE	MLE
Dry Bean	Combined 10%+10%	CF-CABLS	WDPD-M	WDPD-M
Glass	Clean	Ridge	Scalar Liu	DPD
Glass	Symmetric 10%	Ridge	Scalar Liu	DPD
Glass	Asymmetric 10%	Scalar Liu	Scalar Liu	DPD
Glass	Leverage 10%	Ridge	Scalar Liu	DPD
Glass	Combined 10%+10%	Ridge	Scalar Liu	Scalar Liu
Vehicle	Clean	Scalar Liu	Scalar Liu	MLE
Vehicle	Symmetric 10%	DPD	DPD	DPD
Vehicle	Asymmetric 10%	DPD	DPD	DPD
Vehicle	Leverage 10%	Scalar Liu	Scalar Liu	DPD
Vehicle	Combined 10%+10%	DPD	DPD	WDPD-M

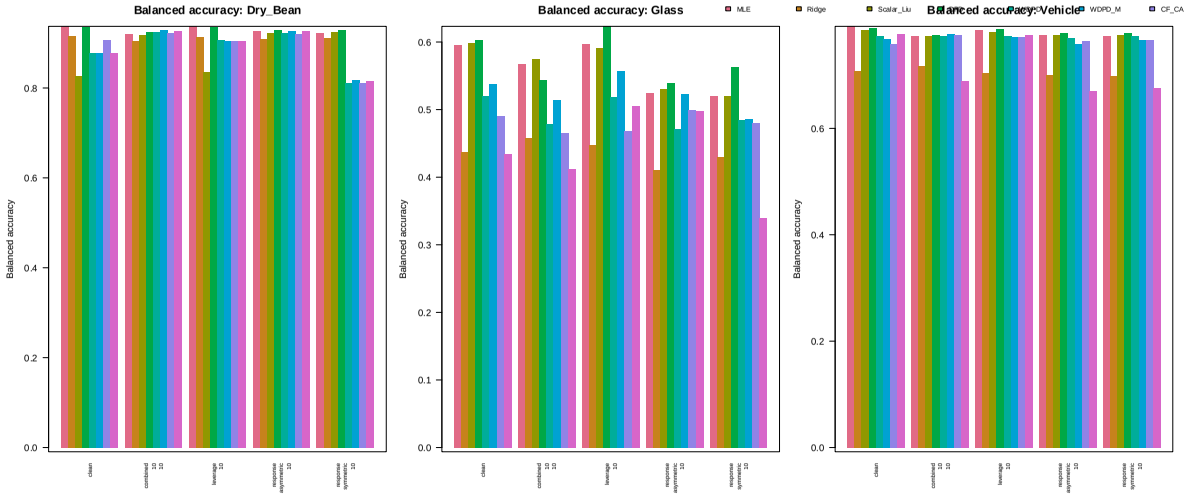
Figure 5 shows the complete log-loss comparison. The panel-specific vertical scales make the dataset heterogeneity clear. Dry Bean supports low log loss for likelihood and DPD estimators; Vehicle produces moderate values; Glass contains several high-loss robust-matrix fits. Figure 5 shows that DPD remains competitive for class-balanced prediction, especially on Vehicle and Glass.

Table 8 reports the clean and combined-contamination settings. On Dry Bean, combined contamination changes the preferred estimator from MLE to CF-CABLS for log loss and to WDPD-M for Brier score. Full CF-CABLS log loss decreases from 0.731 in the clean training condition to 0.283 under combined contamination. This counterintuitive improvement reflects the interaction among the contamination mechanism, the matrix estimate, and the fixed training split design; it should not be interpreted as contamination being intrinsically beneficial. Rather, it indicates that the robust components can regularize an otherwise unstable clean fit in this highly collinear dataset.

On Vehicle and Glass, the simpler estimators remain more reliable. This reinforces the paper's central empirical conclusion: the complete method is a targeted construction for joint difficulties, not a default replacement for ridge, DPD, or MLE in every multiclass problem.



(a) Multinomial log loss.



(b) Balanced accuracy.

Figure 5: Clean-test performance in the repeated semi-synthetic real-data study.

Table 8: Clean-test performance for clean and jointly contaminated training sets. Values are means with Monte Carlo standard errors in parentheses.

Dataset	Scenario	Estimator	Log loss	Brier	Balanced accuracy
Dry Bean	Clean	CF-CABLS	0.731 (0.311)	0.143 (0.021)	0.877 (0.036)
Dry Bean	Clean	CF-CABLS-noM	0.467 (0.238)	0.126 (0.014)	0.906 (0.028)
Dry Bean	Clean	DPD	0.215 (0.014)	0.110 (0.004)	0.936 (0.001)
Dry Bean	Clean	MLE	0.209 (0.013)	0.108 (0.005)	0.937 (0.002)
Dry Bean	Clean	Ridge	0.469 (0.003)	0.222 (0.002)	0.917 (0.002)
Dry Bean	Clean	Scalar Liu	0.347 (0.030)	0.159 (0.008)	0.827 (0.014)
Dry Bean	Clean	WDPD-M	0.694 (0.290)	0.143 (0.021)	0.877 (0.036)
Dry Bean	Combined 10%+10%	CF-CABLS	0.283 (0.012)	0.129 (0.005)	0.927 (0.002)
Dry Bean	Combined 10%+10%	CF-CABLS-noM	0.318 (0.011)	0.144 (0.005)	0.921 (0.002)
Dry Bean	Combined 10%+10%	DPD	0.336 (0.007)	0.151 (0.003)	0.925 (0.002)
Dry Bean	Combined 10%+10%	MLE	0.356 (0.007)	0.162 (0.003)	0.921 (0.002)
Dry Bean	Combined 10%+10%	Ridge	0.650 (0.002)	0.311 (0.001)	0.904 (0.004)
Dry Bean	Combined 10%+10%	Scalar Liu	0.390 (0.004)	0.177 (0.003)	0.917 (0.002)
Dry Bean	Combined 10%+10%	WDPD-M	0.285 (0.011)	0.127 (0.005)	0.928 (0.003)
Vehicle	Clean	CF-CABLS	0.486 (0.014)	0.293 (0.009)	0.777 (0.009)
Vehicle	Clean	CF-CABLS-noM	0.550 (0.014)	0.322 (0.009)	0.758 (0.010)
Vehicle	Clean	DPD	0.533 (0.038)	0.282 (0.011)	0.788 (0.010)
Vehicle	Clean	MLE	0.515 (0.033)	0.277 (0.011)	0.792 (0.009)
Vehicle	Clean	Ridge	0.890 (0.010)	0.483 (0.006)	0.707 (0.009)
Vehicle	Clean	Scalar Liu	0.441 (0.018)	0.272 (0.010)	0.785 (0.009)
Vehicle	Clean	WDPD-M	0.777 (0.034)	0.320 (0.008)	0.769 (0.005)
Vehicle	Combined 10%+10%	CF-CABLS	0.843 (0.036)	0.446 (0.023)	0.689 (0.020)
Vehicle	Combined 10%+10%	CF-CABLS-noM	0.569 (0.008)	0.312 (0.008)	0.775 (0.006)
Vehicle	Combined 10%+10%	DPD	0.529 (0.009)	0.301 (0.007)	0.775 (0.006)
Vehicle	Combined 10%+10%	MLE	0.558 (0.012)	0.312 (0.008)	0.773 (0.005)
Vehicle	Combined 10%+10%	Ridge	0.942 (0.012)	0.504 (0.007)	0.717 (0.011)
Vehicle	Combined 10%+10%	Scalar Liu	0.588 (0.014)	0.319 (0.008)	0.773 (0.007)
Vehicle	Combined 10%+10%	WDPD-M	0.603 (0.032)	0.309 (0.013)	0.777 (0.012)
Glass	Clean	CF-CABLS	4.605 (0.385)	0.867 (0.031)	0.434 (0.050)
Glass	Clean	CF-CABLS-noM	4.020 (0.214)	0.766 (0.020)	0.490 (0.056)
Glass	Clean	DPD	2.413 (0.535)	0.580 (0.047)	0.603 (0.042)
Glass	Clean	MLE	2.166 (0.555)	0.561 (0.051)	0.596 (0.057)
Glass	Clean	Ridge	1.030 (0.027)	0.548 (0.013)	0.437 (0.019)
Glass	Clean	Scalar Liu	1.547 (0.362)	0.536 (0.045)	0.599 (0.058)
Glass	Clean	WDPD-M	6.499 (0.454)	0.795 (0.052)	0.538 (0.040)
Glass	Combined 10%+10%	CF-CABLS	6.251 (0.927)	0.878 (0.060)	0.412 (0.050)
Glass	Combined 10%+10%	CF-CABLS-noM	3.715 (0.458)	0.801 (0.041)	0.465 (0.045)
Glass	Combined 10%+10%	DPD	2.431 (0.674)	0.603 (0.073)	0.544 (0.070)
Glass	Combined 10%+10%	MLE	1.594 (0.364)	0.570 (0.059)	0.567 (0.069)
Glass	Combined 10%+10%	Ridge	1.085 (0.031)	0.575 (0.015)	0.457 (0.033)
Glass	Combined 10%+10%	Scalar Liu	1.259 (0.227)	0.546 (0.046)	0.575 (0.059)
Glass	Combined 10%+10%	WDPD-M	7.350 (0.618)	0.837 (0.053)	0.514 (0.042)

6.3 Ablation evidence

Figure 6 compares the relevant robust and spectral components. In simulation, the no-matrix spectral estimator is generally preferable to full CF-CABLS for log loss and coefficient error. In Dry Bean's combined contamination setting, however, adding the matrix improves the full estimator and the no-shrinkage WDPD-M centre. The matrix therefore acts as a high-variance, occasionally high-value component.

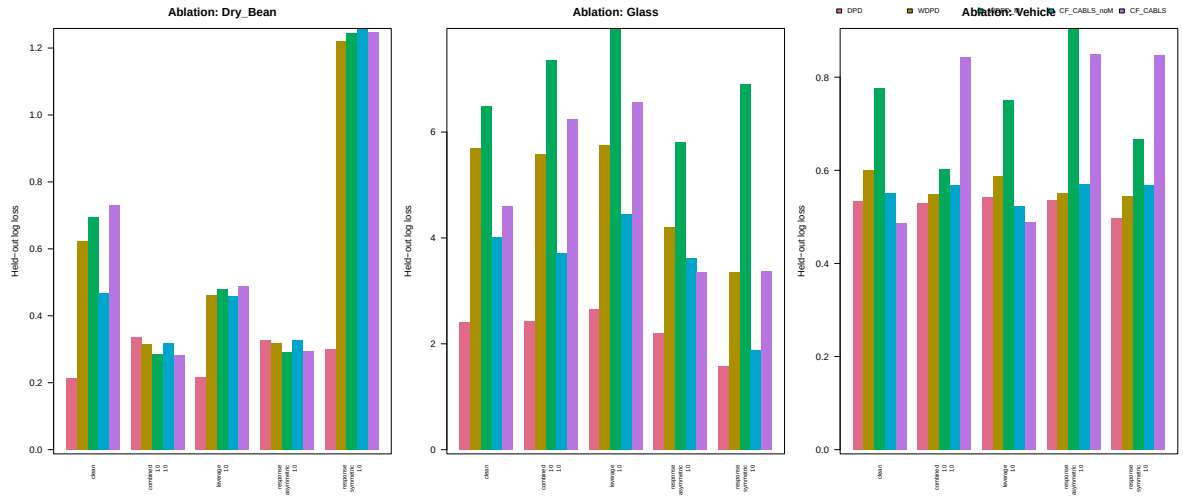


Figure 6: Ablation comparison for predictive log loss. The panels isolate the effects of leverage weighting, matrix adjustment, and spectral shrinkage.

The representative transition matrices in Figure 7 illustrate the diagonal-dominant constraint. The off-diagonal probabilities remain nonzero and adapt between clean and combined-contamination fits, but the matrix is not guaranteed to recover the externally imposed contamination mechanism in small samples. A natural next step is to select between $\widehat{\mathbf{M}}$ and $\mathbf{I}_C$ by training-only cross-fitted risk.

6.4 Spectral diagnostics

The defining mechanism of CF-CABLS is visible in Figure 8. Across datasets and scenarios, posterior mean Liu factors increase with $\log_{10}(\lambda_j)$. The correlations summarized in Table 9 range from 0.817 to 0.964 for the full estimator. Thus, the prior adapts in the intended direction: the smallest robust-information eigenvalues receive the smallest $\widehat{\delta}_j$ and hence the strongest movement toward the cross-fitted centre.

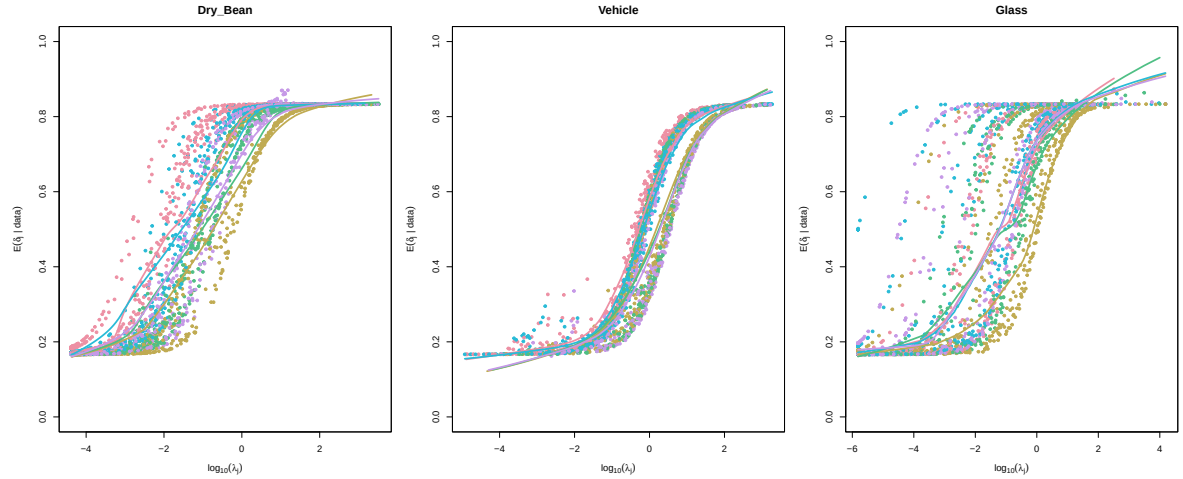
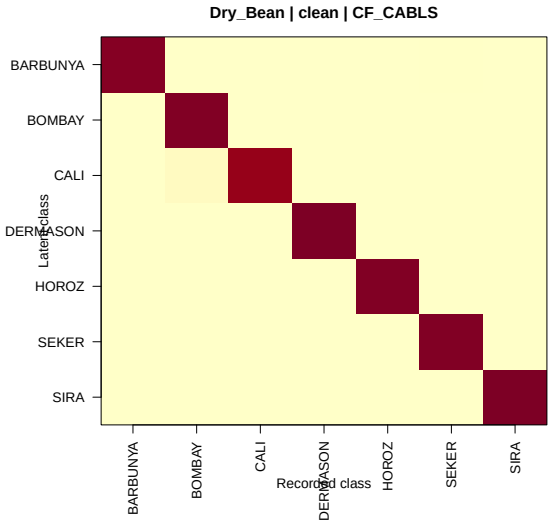
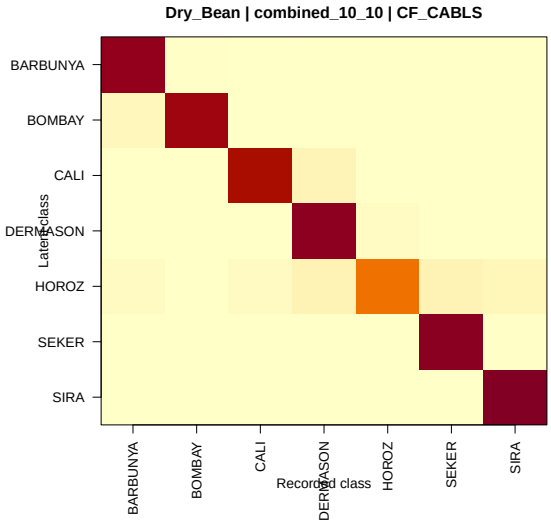


Figure 8: Posterior mean direction-specific Liu factor versus robust-information eigenvalue. Curves and points pool repeated splits and contamination scenarios.

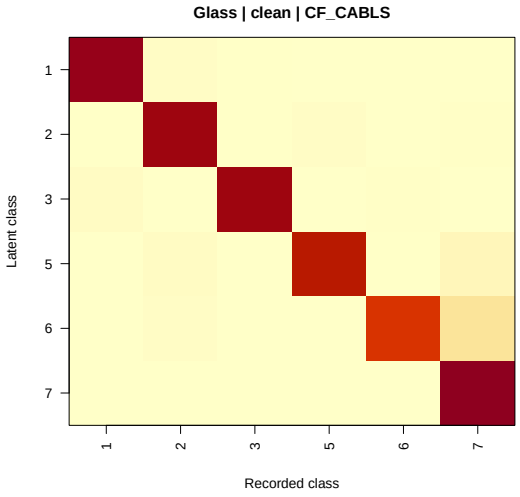
Condition numbers remain extremely large, reaching approximately 10^{10} in some Glass fits. This justifies the use of spectral regularization and positive-definite repair, but it also explains



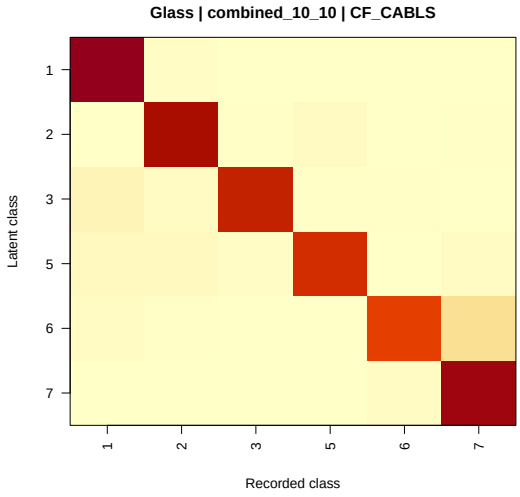
(a) Dry Bean, clean training data.



(b) Dry Bean, combined contamination.



(c) Glass, clean training data.



(d) Glass, combined contamination.

Figure 7: Representative estimated class-transition matrices. The small Glass sample illustrates why explicit matrix estimation can be unstable despite diagonal dominance.

why small changes in the estimated covariance or transition matrix can materially alter the final coefficients.

Table 9: Summary of spectral adaptation across real-data scenarios.

Dataset	Estimator	Mean $\widehat{\delta}$	Correlation range	Median condition no.	Median $\widehat{\omega}_{\mathrm{mag}}$
Dry Bean	CF-CABLS	0.502	[0.949, 0.964]	2.82e+07	1.750
Dry Bean	CF-CABLS-noM	0.504	[0.955, 0.961]	4.20e+07	1.990
Glass	CF-CABLS	0.503	[0.817, 0.936]	2.35e+08	1.893
Glass	CF-CABLS-noM	0.490	[0.885, 0.915]	1.11e+08	2.530
Vehicle	CF-CABLS	0.497	[0.922, 0.937]	1.34e+07	1.838
Vehicle	CF-CABLS-noM	0.514	[0.924, 0.953]	5.43e+05	1.628

The leverage-weight diagnostic in Figure 9 shows substantial downweighting in Dry Bean and more moderate downweighting in Vehicle and Glass. The strong Dry Bean effect is plausible given its extreme correlation structure, but it may also indicate that the theoretical chi-square cutoff is aggressive after robust standardization. Empirical calibration of the cutoff is therefore an important future refinement.

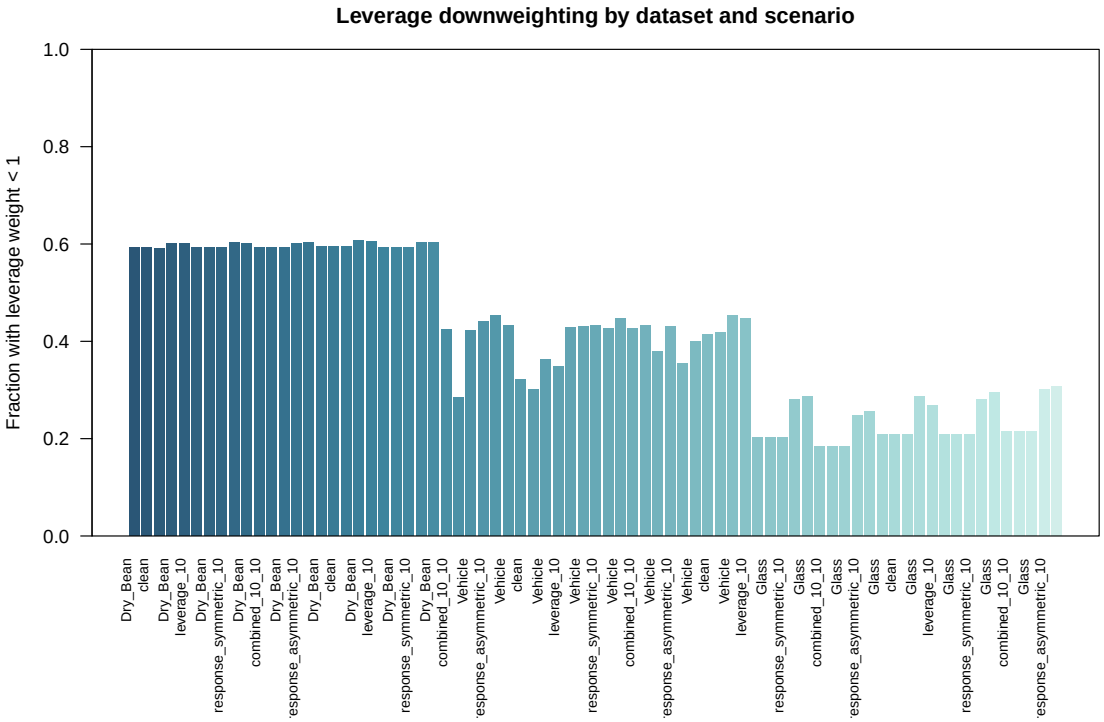


Figure 9: Fraction of training observations assigned leverage weight below one by dataset and contamination scenario.

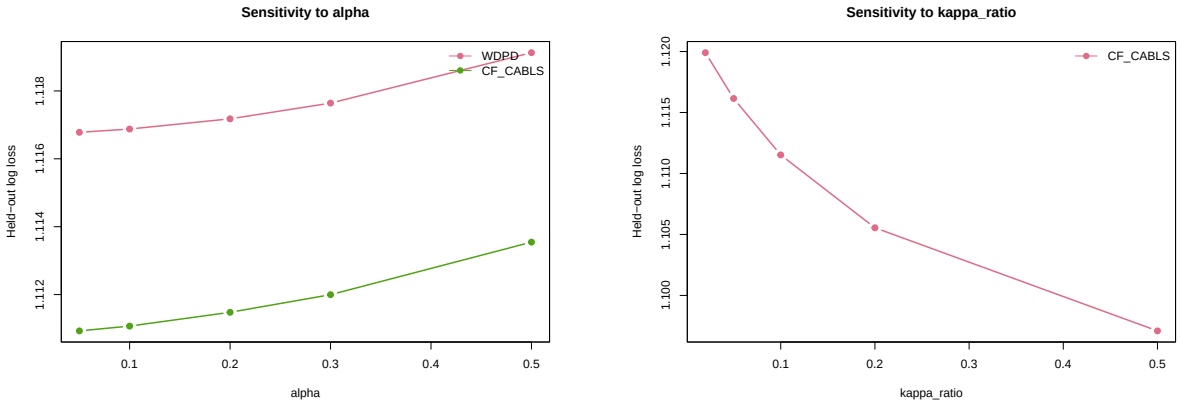
6.5 Sensitivity analysis

The sensitivity results are shown in Table 10 and Figure 10. For both WDPD and CF-CABLS, predictive log loss rises gradually as α increases from 0.05 to 0.50 in the selected combined-contamination scenario. Small α values retain more efficiency while still providing robustness. The CF-CABLS coefficient error is comparatively flat over the lower portion of the grid and worsens at the largest values.

The κ_0 pattern is stronger: increasing κ_0 from 0.02 to 0.50 monotonically reduces log loss and lowers mean SMSE from the upper part of the tested range to 5.46 at 0.50. The default $\kappa_0 = 0.10$ is therefore conservative for this scenario. A broader grid or training-only selection rule should be considered in future applications.

Table 10: Sensitivity to the DPD tuning parameter and spectral precision ratio.

Parameter	Estimator/value	SMSE	SE	Log loss	Balanced accuracy
α	CF-CABLS, 0.05	9.297	0.604	1.111	0.346
α	WDPD, 0.05	11.514	0.779	1.117	0.345
α	CF-CABLS, 0.10	9.339	0.614	1.111	0.346
α	WDPD, 0.10	11.554	0.795	1.117	0.345
α	CF-CABLS, 0.20	9.466	0.637	1.111	0.347
α	WDPD, 0.20	11.674	0.829	1.117	0.345
α	CF-CABLS, 0.30	9.623	0.665	1.112	0.346
α	WDPD, 0.30	11.836	0.871	1.118	0.346
α	CF-CABLS, 0.50	10.026	0.721	1.114	0.347
α	WDPD, 0.50	12.309	0.977	1.119	0.346
κ_0	CF-CABLS, 0.02	12.254	0.948	1.120	0.346
κ_0	CF-CABLS, 0.05	10.974	0.795	1.116	0.346
κ_0	CF-CABLS, 0.10	9.473	0.636	1.112	0.347
κ_0	CF-CABLS, 0.20	7.679	0.477	1.106	0.347
κ_0	CF-CABLS, 0.50	5.462	0.321	1.097	0.347



(a) Sensitivity to α.

(b) Sensitivity to κ_0.

Figure 10: Held-out log-loss sensitivity in the selected simulation setting.

6.6 Convergence and computation

All reported real-data failure rates are zero. Computation time, however, varies substantially. Ridge is nearly instantaneous, while the cross-fitted and matrix-adjusted estimators require repeated fold fits and full-sample optimizations. Some Vehicle cells exhibit unusually long MLE, DPD, or CF-CABLS-noM times because the fallback optimizer sequence is triggered. These costs are part of the practical trade-off for robust cross-fitting.

Table 11: Real-data computation and convergence summarized across 15 dataset-scenario cells.

Estimator	Mean time (s)	Median time (s)	Maximum time (s)	Failure rate
Ridge	0.31	0.06	1.04	0.000
WDPD	8.43	3.24	33.88	0.000
Scalar Liu	11.96	2.70	32.63	0.000
DPD	19.76	2.71	145.53	0.000
MLE	25.63	2.61	204.65	0.000
CF-CABLS	44.39	20.91	155.12	0.000
WDPD-M	51.47	21.88	171.66	0.000
CF-CABLS-noM	70.33	19.70	466.71	0.000

Caching fold-specific weights, reusing a common robust pilot across ablations, and applying the spectral update to one fitted centre for every κ_0 value would reduce runtime without changing the estimator. All publication figures were generated successfully, and the zero-failure record supports numerical reproducibility of the reported analyses.

7 Discussion

The empirical study yields four main lessons.

First, direct regularization is exceptionally effective in the chosen simulation design. Ridge wins every SMSE, log-loss, and Brier comparison. This is not a defect in the study; it is an important benchmark result. The coefficient vector is dense, the ridge constant is favorable, and the independent test distribution matches the clean data-generating mechanism. Under these conditions, a scalar penalty can provide a nearly ideal bias–variance trade-off.

Second, spectral shrinkage improves the robust centre. Both CF-CABLS variants reduce average SMSE relative to MLE, DPD, and WDPD, and the no-matrix variant is consistently the stronger coefficient estimator. Thus, cross-fitted Godambe shrinkage is useful even though it does not dominate ridge. The spectral diagnostic confirms that the mechanism behaves as designed rather than applying an arbitrary global contraction.

Third, explicit misclassification adjustment is regime-dependent. It helps Dry Bean under combined or asymmetric contamination, but often harms Glass and does not improve the simulation's proper scoring rules. A transition matrix has $C(C-1)$ free off-diagonal components before restrictions; with six classes and a small training sample, it can add considerable estimation noise. Diagonal dominance and an informative prior prevent pathological label switching but do not guarantee improved prediction.

Fourth, the appropriate notion of success depends on the metric. Ridge is best for simulation coefficient recovery and proper scoring, while CF-CABLS is more competitive for balanced accuracy. DPD dominates several real-data balanced-accuracy comparisons. Researchers should therefore prespecify whether their priority is coefficient stability, probabilistic calibration, overall accuracy, or minority-class recovery.

These results suggest a decision-oriented use of the framework. In clean or small samples, ridge, scalar Liu, or DPD should remain primary benchmarks. When strong multicollinearity coexists with plausible response and leverage contamination, CF-CABLS offers a principled integrated option. The identity and estimated transition models should ideally be compared by training-only cross-fitted risk rather than always forcing $\widehat{\mathbf{M}}$ into the final fit.

8 Limitations and Future Research

Several limitations must be stated explicitly.

1. **Replication configuration.** The study uses 100 paired simulation replications, 30 paired sensitivity replications, and five repeated real-data splits. This design is adequate for the broad comparisons reported here and balances Monte Carlo precision against the substantial cost of nested robust optimization. Nevertheless, very small differences between closely performing methods should be interpreted cautiously, and future work could extend selected unstable scenarios or real-data splits.

2. **Conditional uncertainty.** The sandwich covariance conditions on the estimated transition matrix and cross-fitted nuisance quantities. It therefore understates uncertainty if $\widehat{\mathbf{M}}$ is noisy. A nonparametric bootstrap that repeats scaling, cross-fitting, EM, and robust fitting is preferable for interval estimation.

3. **Transition-matrix identification.** The matrix is learned from model-based OOF probabilities under diagonal dominance and an identity-favouring prior. Severe model misspecification, rare classes, or small samples can bias or destabilize this stage.

4. **Fixed tuning and comparison fairness.** The empirical run uses fixed primary tuning values and separate sensitivity grids. A stricter comparison would tune ridge, scalar Liu, α, leverage cutoff, and κ_0 with the same training-only validation budget.

5. **Leverage calibration.** The robust-distance cutoff downweights a substantial fraction of Dry Bean observations. Data-adaptive calibration or bounded-influence weights with a target effective sample size could avoid excessive efficiency loss.

6. **Comparator scope.** The study includes eight transparent likelihood, robust, and shrinkage ablations but does not include the recent Rényi, B-robust, weighted-likelihood, or simplex-based estimators as software comparators. Adding one independently implemented robust polytomous competitor would strengthen external positioning.

7. **Reference category and dimensionality.** Baseline-category coefficients depend on the chosen reference class even though exact fitted probabilities are invariant. The method is designed for moderate P and is not a replacement for sparse or structured priors in genuinely high-dimensional settings.

8. **Semi-synthetic real data.** Contamination is imposed on real predictor structures while the clean test labels are retained. This provides a known stress test but does not demonstrate that the original datasets contain real miscoding or leverage anomalies.

Future work should develop an adaptive identity-versus-transition decision, jointly calibrated tuning, bootstrap uncertainty, a symmetric or simplex-coded formulation, and computational caching. The sensitivity results also motivate a wider κ_0 grid and formal selection based on OOF proper scoring rules. Finally, the method could be extended to grouped predictors, sparse high-dimensional MNL models, complex survey designs, and partially verified labels.

9 Conclusion

CF-CABLS-MNL combines robust polytomous estimation and Liu-type spectral stabilization in a cross-fitted generalized empirical-Bayes framework. OOF weighted-DPD fits provide a pilot centre and probabilities for a restricted class-transition model. A full robust centre is evaluated with sandwich/Godambe uncertainty, rotated into robust-information eigen-directions, and shrunk through adaptive beta-distributed Liu factors.

The empirical evidence is informative precisely because it is not uniformly favorable. Ridge is the strongest coefficient and log-loss estimator in the controlled simulations. CF-CABLS nonetheless reduces coefficient error relative to unregularized estimators and is competitive for balanced accuracy. In real predictor structures, the complete method is useful in selected jointly contaminated, highly collinear settings, but simpler DPD, MLE, ridge, or scalar Liu estimators are preferable elsewhere. The transition matrix is therefore a selective component rather than a guaranteed improvement.

The main contribution is a coherent and reproducible decomposition of three simultaneous problems–multicollinearity, response contamination, and leverage–together with diagnostics that reveal when each component helps. The framework provides a foundation for adaptive matrix selection and fair joint tuning in a future definitive implementation.

Acknowledgements

The authors acknowledge the University of Milano–Bicocca for providing PhD scholarship support. The authors also thank the maintainers and contributors of the UCI Machine Learning Repository.

Statements and Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The Dry Bean, Statlog Vehicle Silhouettes, and Glass Identification datasets used in this study are publicly available from the UCI Machine Learning Repository [10, 13, 23]. No restricted or personally identifiable data were used.

Code availability

The analysis code and derived modelling materials are available from the corresponding author upon reasonable request, subject to the terms of the original dataset licences.

References

[1] Mohamed R. Abonazel, Issam Dawoud, Mohammed Naji Al-Ghamdi, and Rasha A. Farghali. "Developing the Generalized Dawoud–Kibria Estimator for the Multinomial Logistic Model: Simulation Study and Application". In: *Scientific African* 29 (2025), e02803. DOI: 10.1016/j.sciaf.2025.e02803.

[2] Mohamed R. Abonazel and Rasha A. Farghali. "Liu-Type Multinomial Logistic Estimator". In: *Sankhya B* 81.2 (2019), pp. 203–225. DOI: 10.1007/s13571-018-0171-4.

[3] Alan Agresti. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: Wiley, 2013.

[4] Yasin Asar and Murat Erisoglu. "Liu Estimator in the Multinomial Logistic Regression Model". In: *arXiv preprint arXiv:2111.03398* (2021). DOI: 10.48550/arXiv.2111.03398.

[5] Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, and M. C. Jones. "Robust and Efficient Estimation by Minimising a Density Power Divergence". In: *Biometrika* 85.3 (1998), pp. 549–559. DOI: 10.1093/biomet/85.3.549.

[6] Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. "A General Framework for Updating Belief Distributions". In: *Journal of the Royal Statistical Society: Series B* 78.5 (2016), pp. 1103–1130. DOI: 10.1111/rssb.12158.

[7] Elena Castilla. "A New Robust Approach for the Polytomous Logistic Regression Model Based on Rényi's Pseudodistances". In: *Biometrics* 80.4 (2024), ujae125. DOI: 10.1093/biomtc/ujae125.

[8] Elena Castilla, Abhik Ghosh, Nirian Martin, and Leandro Pardo. "New Robust Statistical Procedures for the Polytomous Logistic Regression Models". In: *Biometrics* 74.4 (2018), pp. 1282–1291. DOI: 10.1111/biom.12890.

[9] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. "Double/Debiased Machine Learning for Treatment and Structural Parameters". In: *The Econometrics Journal* 21.1 (2018), pp. C1–C68. DOI: 10.1111/ectj.12097.

[10] *Dry Bean*. 2020. DOI: 10.24432/C50S4B.

[11] Rasha A. Farghali, Muhammad Qasim, B. M. Golam Kibria, and Mohamed R. Abonazel. "Generalized Two-Parameter Estimators in the Multinomial Logit Regression Model: Methods, Simulation and Application". In: *Communications in Statistics—Simulation and Computation* 52.7 (2023), pp. 3327–3342. DOI: 10.1080/03610918.2021.1934023.

[12] Cornelia Fuetterer, Malte Nalenz, Thomas Augustin, Ruth M. Pfeiffer, et al. "A Powerful Penalized Multinomial Logistic Regression Approach". In: *Computational Statistics* 40 (2025), pp. 4565–4587. DOI: 10.1007/s00180-025-01635-0.

[13] B. German. *Glass Identification*. 1987. DOI: 10.24432/C5WW2P.

[14] Abhik Ghosh and Ayanendranath Basu. "Robust Bayes Estimation Using the Density Power Divergence". In: *Annals of the Institute of Statistical Mathematics* 68 (2016), pp. 413–437. DOI: 10.1007/s10463-014-0499-0.

[15] V. P. Godambe. "An Optimum Property of Regular Maximum Likelihood Estimation". In: *The Annals of Mathematical Statistics* 31.4 (1960), pp. 1208–1211.

[16] Arthur E. Hoerl and Robert W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12.1 (1970), pp. 55–67. DOI: 10.1080/00401706.1970.10488634.

[17] Hung Hung, Zhi-Yu Jou, and Su-Yun Huang. "Robust Mislabel Logistic Regression without Modeling Mislabel Probabilities". In: *Biometrics* 74.1 (2018), pp. 145–154.

[18] Maria Iannario and Anna Clara Monti. "Robust Logistic Regression for Ordered and Unordered Responses". In: *Econometrics and Statistics* 38 (2026), pp. 97–121. DOI: 10.1016/j.ecosta.2023.05.004.

[19] Kejian Liu. "A New Class of Biased Estimate in Linear Regression". In: *Communications in Statistics—Theory and Methods* 22.2 (1993), pp. 393–402.

[20] Daniel McFadden. "Conditional Logit Analysis of Qualitative Choice Behavior". In: *Frontiers in Econometrics*. Ed. by Paul Zarembka. New York: Academic Press, 1974, pp. 105–142.

[21] Julien Miron, Benjamin Poilane, and Eva Cantoni. "Robust Polytomous Logistic Regression". In: *Computational Statistics & Data Analysis* 176 (2022), p. 107564. DOI: 10.1016/j.csda.2022.107564.

[22] Peter J. Rousseeuw and Katrien Van Driessen. "A Fast Algorithm for the Minimum Covariance Determinant Estimator". In: *Technometrics* 41.3 (1999), pp. 212–223. DOI: 10.1080/00401706.1999.10485670.

[23] *Statlog (Vehicle Silhouettes)*. 1986. DOI: 10.24432/C5HG6N.

[24] Md Nazir Uddin and Jeremy T. Gaskins. "Shared Bayesian Variable Shrinkage in Multinomial Logistic Regression". In: *Computational Statistics & Data Analysis* 177 (2023), p. 107568. DOI: 10.1016/j.csda.2022.107568.

[25] Halbert White. "Maximum Likelihood Estimation of Misspecified Models". In: *Econometrica* 50.1 (1982), pp. 1–25.

[26] Shunqin Zhang, Sanguo Zhang, Hang Li, and Sheng Fu. "Robust Simplex-Based Multinomial Logistic Regression". In: *Communications in Mathematics and Statistics* (2026), pp. 1–37. DOI: 10.1007/s40304-025-00459-0.

A Supplementary Output Inventory

The supplementary materials include the complete set of dataset–scenario transition-matrix heatmaps and the full generated tables in both CSV and LaTeX formats. The compact tables in the main text were derived directly from those outputs to improve readability.

B Additional Figures

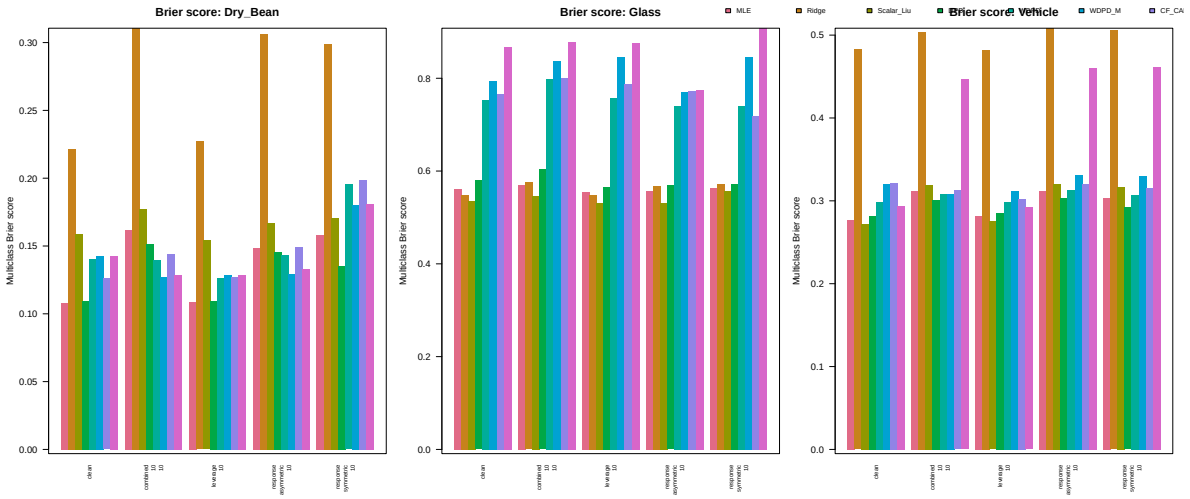


Figure 11: Multiclass Brier score in the real-data study.

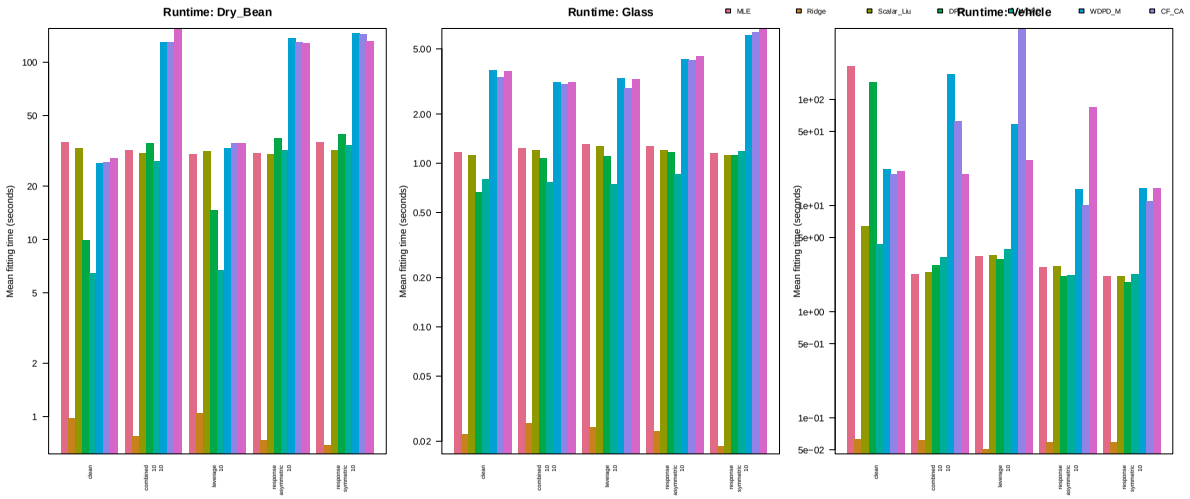


Figure 12: Elapsed computation time by dataset, scenario, and estimator.

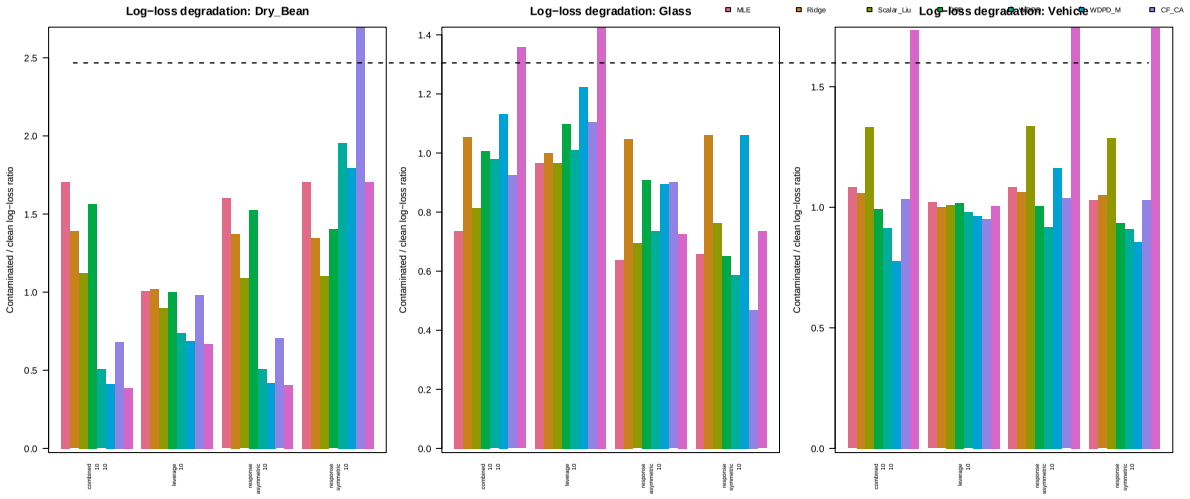


Figure 13: Ratio of contaminated-training log loss to clean-training log loss. Values below one indicate improvement relative to the corresponding clean fit.

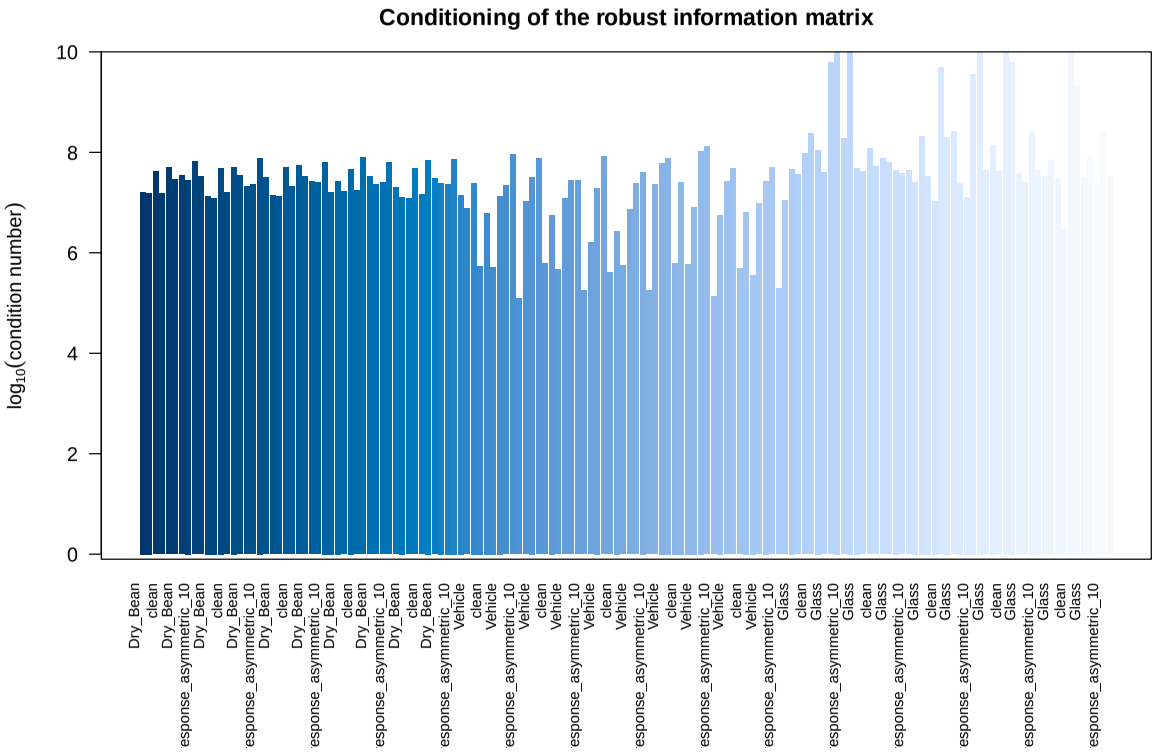


Figure 14: Condition numbers of the robust information matrices.