American Journal of Theoretical and Applied Statistics
2026, Vol. 15, No. 5, pp. 276–302
https://doi.org/10.11648/j.ajtas.20261505.17

SciencePG
Science Publishing Group

Research Article

Feature Selection in Over-dispersed Binary and Count Data Models Using Penalized Optimal Estimating Functions

Timothy Mutunga[1, *], Ali Salim[1], Pius Kihara[2], Justine Okenye[1]

[1]Department of Mathematics, Egerton University, Egerton, Kenya

[2]Department Financial and Actuarial Mathematics, The Technical University of Kenya, Nairobi, Kenya

Abstract

Generalized linear models (GLMs) remain a core class of supervised machine learning models for binary and count responses, with feature selection commonly carried out through penalized likelihood or quasi-likelihood methods. This paper develops a feature-selection framework based on penalized Optimal Estimating Functions (OEFs), which attain Godambe optimality within a class of unbiased estimating functions and incorporate higher-order moment information without requiring full likelihood specification. Ridge, LASSO, Adaptive LASSO and SCAD penalties are introduced at the regression estimating-equation level, while dispersion is estimated jointly through an unpenalized OEF. Hyperparameters are selected using cross-validated estimating-function loss with prespecified edge and stability rules. Monte Carlo simulations with 500 replications examine Beta-Binomial and Negative-Binomial regression under moderate and high overdispersion and sparse and moderately dense signals. The results do not show uniform superiority of penalization. The unpenalized OEF generally provides the strongest coefficient coverage and RMSE benchmark, whereas the penalized OEFs provide sparse feature selection with model-dependent trade-offs between sensitivity, specificity and interval calibration. Simple post-selection refitting reduces shrinkage bias in some settings but does not restore nominal coverage because selection uncertainty remains. Penalized OEF is therefore presented as a competitive alternative when sparse selection and joint mean-dispersion estimation are both required, rather than as a uniformly better estimator.

Keywords

Feature Selection, Penalized Estimating Functions, Over-dispersion, Beta-Binomial, Negative Binomial, Godambe Information, Post-selection Inference, Stability Selection

1. Introduction

Feature selection in statistical modelling and supervised machine learning aims at identifying a parsimonious subset of relevant features that adequately explain the relationship between covariates and a response variable. Variable selection serves both inferential and predictive objectives. Penalization techniques such as ridge regression [1], the LASSO [2], Adaptive LASSO [3] and SCAD [4] have become standard tools for mitigating over-fitting and controlling model complexity through variable selection in regression and classification

*Correspondence: Timothy Mutunga (mutungatn@gmail.com)

Received: 12 August 2026; **Accepted:** 5 September 2026; **Published:** 24 September 2026

tasks. These methods are typically embedded within likelihood or quasi-likelihood frameworks for generalized linear models (GLMs) [5].

Classical GLMs such as Logistic and Poisson regression impose restrictive mean-variance relationships that are frequently violated in observational data [5, 6]. Particularly, binary and count data frequently violate the variance assumptions of the Binomial and Poisson models because of unobserved heterogeneity, clustering or model misspecification, which is what motivated quasi-likelihood corrections [7]. While quasi-likelihood methods adjust variance estimates, they rely only on first and second moment conditions and typically treat dispersion as a nuisance parameter. As a result, inference on both regression and dispersion parameters can be unreliable under misspecification.

The reliance on limited moment information and distributional specification can lead to substantial inferential failures. In particular, regression coefficients may exhibit systematic under-coverage while dispersion parameters may be poorly estimated or entirely unidentified in finite samples [7-9]. These shortcomings become more pronounced under penalization. Shrinkage changes the residual structure used to estimate variability which in turn affects dispersion estimates and standard error computations. Standard quasi-likelihood methods and naïve penalized procedures do not explicitly account for this interaction leading to additional inferential distortions in finite samples [4, 10]. These limitations motivate inferential frameworks that remain robust under distributional misspecification.

The theory of Optimal Estimating Functions (OEFs) offers an alternative. Originating in the work of Godambe [11] and extended by Godambe and Thompson [8], OEFs achieve optimality within a class of unbiased estimating functions without requiring full distributional specification. By incorporating higher-order moment information, OEFs allow efficient and improved robustness to misspecification and facilitate joint estimation of regression and dispersion parameters [9]. This joint treatment is particularly important in over-dispersed models where dispersion influences both point estimation and uncertainty quantification.

Little work has applied OEFs to feature selection. Existing work on penalized estimating functions has largely focused on semiparametric regression settings [10] with comparatively little attention to over-dispersed GLMs and the role of dispersion inference. Moreover, the hyper-parameter selection problem which is central to penalized estimation has not been systematically addressed within the estimating functions framework.

Penalization is introduced directly at the estimating-equation level rather than through a likelihood, enabling simultaneous regression and variance estimation in addition to variable selection while preserving the estimating-function structure. Sparsity is induced in the regression coefficients using Ridge, LASSO, Adaptive LASSO and SCAD penalties [1-4]. The penalty is applied exclusively to the regression estimating equation while the dispersion parameter is estimated from an unpenalized OEF, ensuring unbiasedness and preserving Godambe optimality in the dispersion component. This formulation follows the general theory of penalized estimating functions and extends it to settings with explicit dispersion modelling and higher-order moment conditions [10].

This paper develops a framework for feature selection based on penalized OEFs. The framework covers the construction of the penalized estimating equations, hyper-parameter tuning, and stability selection. The remainder of the paper is organized as follows. Section 2 presents the Beta-Binomial and Negative Binomial model frameworks. Section 3 develops the optimal estimating functions and their penalized extensions for feature selection. Section 4 describes hyper-parameter selection and stability procedures. Section 5 reports on the Monte Carlo simulation studies for both models together with a discussion of the findings. Section 6 concludes the research work.

2. Model Framework

This section presents the mathematical framework for the over-dispersed generalized linear regression models considered in this study: Beta-Binomial Logistic Regression (BB) and Negative Binomial - Poisson Regression (NB2). Unlike likelihood-based methods which derive score functions from full distributional specifications, the Optimal Estimating Functions (OEF) framework relies on unbiased Estimating Functions constructed explicitly on successive central moments of the data generating process. Accordingly, we detail the two primary regression models' distributional properties, parameter specifications and moment structures.

Throughout, we assume independent observations $\{Y_i, x_i\}_{i=1}^{n}$, where, Y_i is a scalar response variable, $x_i = \left(1, x_{i1}, ..., x_{ip}\right)^T$ is the i^{th} row of full rank design/feature matrix $\mathrm{X} \in \mathbb{R}^{n\times(p+1)}$ and $\beta \in \mathbb{R}^{p+1}$ is a vector of regression coefficients. In the binary data case, the number of trials m is constant.

2.1. Beta-Binomial Logistic Regression

The Beta-Binomial model accommodates overdispersion in binary data by allowing the heterogeneity in the success probability.

Let $Y_1, Y_2, Y_3, ..., Y_n$ represent the number of successes in m binomial trials, $n \le m.$ In the Beta-Binomial framework, the probability of success π_i is not constant over the set of m trials. Conditionally, we have:

$$Y / \pi_i \sim Bin\ (m, \pi_i) \tag{1}$$

The probability π_i is assumed to follow a Beta probability distribution with parameters p_i/ϕ and $(1-p_i)/\phi$, where $0<p_i<1$ and $\phi>0$. The over-dispersion parameter ϕ controls the degree of heterogeneity in success probability across trials [12].

Under these assumptions, the marginal distribution of the observed successes is obtained by integrating over the distribution of π_i, $g(\pi_i)$.

$$f(y_i)=\int_0^1 P(Y=y_i/\pi_i)\,g(\pi_i)\,d\pi_i=\binom{m}{y_i}\frac{\Gamma\left(\frac{1}{\phi}\right)}{\Gamma\left(\frac{p_i}{\phi}\right)\Gamma\left(\frac{1-p_i}{\phi}\right)}\frac{\Gamma\left(\frac{p_i}{\phi}+y_i\right)\Gamma\left(\frac{1-p_i}{\phi}+m-y_i\right)}{\Gamma\left(\frac{1}{\phi}+m\right)} \tag{2}$$

which yields the Beta-binomial distribution in (2).

The moments of the Beta-binomial distribution are characterized by the following expressions:

The marginal mean is:

$$\mu_i=E(Y_i)=mp_i \tag{3}$$

Najera-Zuloaga et al. [12] assumed that the probability parameter p_i of the beta-binomial is linked to the covariates through the logistic transformation:

$$p_i=\log istic\left(x_i^T\beta\right)=\frac{\exp\left(x_i^T\beta\right)}{1+\exp\left(x_i^T\beta\right)} \tag{4}$$

By the relationship in equation (4), we define the linear predictor as:

$$\eta_i=\log\left(\frac{p_i}{1-p_i}\right)=x_i^T\beta \tag{5}$$

This explicit separation of the mean, link and linear predictor ensures compatibility with the general GLM framework while allowing over-dispersion through the Beta mixing distribution [5, 7, 12].

The variance function can be written as:

$$\mathrm{Var}(Y_i)=mp_i(1-p_i)\left[1+(m-1)\frac{\phi}{1+\phi}\right] \tag{6}$$

Substituting $p_i=\frac{\mu_i}{m}$ in (6) we have:

$$V(\mu_i;\rho)=\frac{\mu_i}{m}(m-\mu_i)\left[1+(m-1)\rho\right] \tag{7}$$

Where $\rho=\frac{\phi}{1+\phi}$ is the intra-class correlation parameter and an alternative parameterization of the over-dispersion parameter. The variance exceeds the Binomial variance $\frac{\mu_i}{m}(m-\mu_i)$ by the multiplicative inflation factor $1+(m-1)\rho$ that explicitly models over-dispersion. When $\rho=0$, the classical binomial model is recovered.

For the higher order moments, we will make the same substitutions $p_i=\frac{\mu_i}{m}$ and $\rho=\frac{\phi}{1+\phi}$. The third central moment is defined as:

$$\kappa_{3i}=\frac{1+(2m-1)\rho}{(1+\rho)}\left(\frac{m-2\mu_i}{m}\right)\left[Var(\mu_i;\rho)\right]^{-\frac{1}{2}} \tag{8}$$

The fourth central moment is defined as:

$$\kappa_{4i}=\frac{\rho^2}{(1+\rho)(1+2\rho)}\left(Var(\mu_i;\rho)\right)^{-1}\left[(1-\rho)\left(\frac{1-2\rho(1-3m)+6m^2\rho^2}{\rho^2}\right)\right.$$

$$\left.+\frac{3\mu_i(m-\mu_i)}{(m\rho)^2}\left\{(m-2)(1-\rho)^2-m\rho(1-\rho)(6-m)-6m^2\rho^2\right\}\right] \tag{9}$$

2.2. Negative Binomial Regression

The Negative Binomial model addresses overdispersion in count data by introducing heterogeneity in the Poisson rate parameter through a Gamma mixing distribution. This hierarchical structure naturally accommodates the excess variability commonly observed in count data [13].

The model is constructed hierarchically. First, conditionally on the rate parameter λ_i, the observed counts y_i, $i = 1,2,3,...,n$ follow a Poisson distribution:

$$y_i / \lambda_i \sim Poisson(\lambda_i) \tag{10}$$

Second, the rate parameter λ_i is assumed to follow a Gamma distribution:

$$\lambda_i \sim Gamma\left(r, \frac{p_i}{1-p_i}\right) \tag{11}$$

where r is the shape parameter and $\frac{p_i}{1-p_i}$ is the rate parameter of the Gamma distribution.

Marginalizing over λ_i yields the Negative-binomial (r, p_i) distribution with probabilty mass function specified as:

$$P(Y = y_i) = \frac{\Gamma(y_i + r)}{y_i!\Gamma(r)} p_i^{\,r} (1-p_i)^{y_i} \quad for\ y_i = 0,1,2,... \tag{12}$$

with mean and variance given as $E(Y_i) = \frac{r(1-p_i)}{p_i}$ and $\mathrm{Var}(Y_i) = \frac{r(1-p_i)}{p_i^2}$ respectively.

Lindén and Mäntyniemi [14] adopted a parameterization scheme where $\mu_i = \frac{r(1-p_i)}{p_i} \Rightarrow p_i = \frac{r}{\mu_i + r}$ and $r = \frac{1}{\alpha}$ in (12) to yield:

$$\frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{y_i!\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \quad for\ y_i = 0,1,2,... \tag{13}$$

The Negative-binomial (μ_i, α) distribution in (13) directly models overdispersion through the heterogeneity parameter $\alpha = \frac{1}{r}$ for $\alpha \geq 0$.

The moments of the Negative-binomial distribution in (13) are characterized by the following expressions.

The mean function is specified using a log-link:

$$\log(\mu_i) = x_i^T \beta \tag{14}$$

so that

$$E(Y_i) = \mu_i = \exp\left(x_i^T \beta\right) \tag{15}$$

The linear predictor is therefore expressed as:

$$\eta_i = x_i^T \beta \tag{16}$$

This log-link ensures that the predicted counts remain positive for all covariate values and provides a natural interpretation of the regression coefficients as multiplicative effects on the mean count [15].

The variance function $\mathrm{Var}(Y_i) = \frac{r(1-p_i)}{p_i^2}$ can be written with proper substitutions for $p_i = \frac{r}{\mu_i + r}$ and $r = \frac{1}{\alpha}$ as:

$$\mathrm{Var}(\mu_i; \alpha) = \mu_i + \alpha\mu_i^2 \tag{17}$$

The variance exceeds the mean by the quadratic term $\alpha\mu_i^2$ with α quantifying the degree of overdispersion. When $\alpha = 0$ the model reduces to the Poisson distribution [13, 16].

This Poisson-Gamma mixture construction and its variance structure are standard in the literature [13, 15, 16].

The higher-order moments for the Negative-binomial distribution are obtained through appropriate substitutions for $p_i = \frac{r}{\mu_i + r}$ and $r = \frac{1}{\alpha}$.

The third central moment is defined as:

$$\kappa_{3i} = \frac{2 - p_i}{\sqrt{r(1-p_i)}} = \frac{2\alpha\mu_i + 1}{\sqrt{Var(\mu_i; \alpha)}} \tag{18}$$

The fourth central moment can be derived as:

$$\kappa_{4i} = \frac{p_i^{\,2} - 6p_i + 6}{r(1-p_i)} = \frac{1 + 6\alpha\mu_i(1+\alpha\mu_i)}{Var(\mu_i; \alpha)} \tag{19}$$

The inclusion of third and fourth order moments in the two regression set-ups is essential for the orthogonalization step in the construction of Optimal Estimating Functions (OEFs) in the next section. While the first and second order moment conditions determine mean and variance structure, higher-order moments capture skewness and kurtosis which contain additional information relevant for efficient joint inference under over-dispersion. The orthogonalization process ensures that

the regression and dispersion estimating functions are asymptotically uncorrelated and the higher-order moments contribute information only through components not explained by lower-order moments. Following this, the resulting estimating functions maximize the Godambe information within the class of unbiased estimating functions [8, 9, 11, 35].

This hierarchical construction forms the foundation for the penalized Optimal Estimating Functions developed in the subsequent section.

3. Estimating Functions (EFs)

In this section, we develop the estimating functions (EFs) machinery on which the proposed feature-selection framework is anchored. Firstly, we construct optimal estimating functions (OEFs) for the over-dispersed binary and count data models discussed in Section 2 by orthogonalizing a set of unbiased moment conditions to yield estimators of the regression and dispersion parameters that attain the Godambe information within the class of unbiased estimating functions. We will then introduce penalization directly on the EFs structure with the penalty applied to the regression estimating equation alone. The dispersion estimating equation is left unpenalized since dispersion is a nuisance parameter to be estimated and not a selection candidate. Since the OEFs are based on moment conditions rather than a full likelihood framework, their construction applies uniformly to the Beta-Binomial and Negative-Binomial regression models from Section 2 and this procedure provides the basis for the penalized estimators and hyper-parameter tuning rules examined in the remainder of the paper.

3.1. Optimal Estimating Functions (OEFs)

Let $y_i, i = 1....n$ be independent observations from unspecified distribution with mean equation:

$$E(y_i) = \mu_i \tag{20}$$

where, $\mu_i = \mu_i(\beta)$ and variance equation:

$$\text{Var}(y_i) = \nu(\mu_i;\phi) \tag{21}$$

where ϕ is the dispersion parameter.

A candidate set of unbiased Estimating Functions (EFs) is:

$$\left.\begin{aligned} \pi_{1i} &= y_i - \mu_i \\ \pi_{2i} &= (y_i - \mu_i)^2 - \nu(\mu_i;\phi) \end{aligned}\right\} \tag{22}$$

Utilizing Theorem 3.2 of Godambe and Thompson [8], elementary orthogonal EFs in (23) for $i = 1,2,...,n$ are constructed using the Gram-Schmidt orthogonalization procedure based on the set of unbiased EFs in (22). This orthogonalization removes redundancy between moment conditions and minimizes asymptotic variance.

$$\left.\begin{aligned} h_{1i} &= \pi_{1i} \\ h_{2i} &= \pi_{2i}^{*} \\ &= (y_i - \mu_i)^2 - \nu(\mu_i;\phi) - (y_i - \mu_i)\nu(\mu_i;\phi)^{\frac{1}{2}}\kappa_{3i} \end{aligned}\right\} \tag{23}$$

where, π_{2i}^{*} is the orthogonalized π_{2i} and κ_{3i} is the skewness coefficient defined in (24).

$$\kappa_{3i} = E\left[\frac{(y_i - \mu_i)}{\nu(\mu_i;\phi)^{\frac{1}{2}}}\right]^3 \tag{24}$$

Using the elementary EFs in (23), Theorem 3.2 of Godambe and Thompson [8] is applied to construct a linear combination of elementary EFs in (25)

$$\left.\begin{aligned} g_\beta &= \sum_{i=1}^{n} a_1^1 h_{1i} + \sum_{i=1}^{n} a_2^1 h_{2i} \\ g_\phi &= \sum_{i=1}^{n} a_1^2 h_{1i} + \sum_{i=1}^{n} a_2^2 h_{2i} \end{aligned}\right\} \tag{25}$$

with, $a_1^1 = \dfrac{E\left[\frac{\partial h_{1i}}{\partial \beta}\right]}{E\left[(h_{1i})^2\right]} = \dfrac{-\frac{\partial \mu_i}{\partial \beta}}{\nu(\mu_i;\phi)}$,

$$a_2^1 = \frac{E\left[\frac{\partial h_{2i}}{\partial \beta}\right]}{E\left[(h_{2i})^2\right]} = \frac{\nu(\mu_i;\phi)^{\frac{1}{2}}\kappa_{3i}\frac{\partial \mu_i}{\partial \beta} - \frac{\partial}{\partial \beta}\nu(\mu_i;\phi)}{\nu(\mu_i;\phi)^2\left\{\kappa_{4i} + 2 - \kappa_{3i}^2\right\}},$$

$$a_1^2 = \frac{E\left[\frac{\partial h_{1i}}{\partial \phi}\right]}{E\left[(h_{1i})^2\right]} = 0 \quad \text{and}$$

$$a_2^2 = \frac{E\left[\frac{\partial h_{2i}}{\partial \phi}\right]}{E\left[(h_{2i})^2\right]} = \frac{-\frac{\partial}{\partial \phi}\nu(\mu_i;\phi)}{\nu(\mu_i;\phi)^2\left\{\kappa_{4i} + 2 - \kappa_{3i}^2\right\}}$$

where, $\kappa_{4i} = E\left[\dfrac{(y_i - \mu_i)^4}{\nu(\mu_i;\phi)^2} - 3\right]$ is the fourth order moment.

Substituting the quantities a_1^1, a_2^1, a_1^2 and a_2^2 into (25) results in the jointly OEFs for β and ϕ in (26) and (27) respectively.

$$g_{\beta}^{*}=-\sum_{i=1}^{n}\frac{\frac{\partial\mu_i}{\partial\beta}}{v(\mu_i;\phi)}(y_i-\mu_i)+\sum_{i=1}^{n}\frac{v(\mu_i;\phi)^{\frac{1}{2}}\kappa_3\frac{\partial\mu_i}{\partial\beta}-\frac{\partial}{\partial\beta}v(\mu_i;\phi)}{v(\mu_i;\phi)^2\{\kappa_{4i}+2-\kappa_{3i}^2\}}\left((y_i-\mu_i)^2-v(\mu_i;\phi)-(y_i-\mu_i)v(\mu_i;\phi)^{\frac{1}{2}}\kappa_{3i}\right) \quad (26)$$

$$g_{\phi}^{*}=-\sum_{i=1}^{n}\frac{\frac{\partial}{\partial\phi}v(\mu_i;\phi)}{v(\mu_i;\phi)^2\{\kappa_{4i}+2-\kappa_{3i}^2\}}\left((y_i-\mu_i)^2-v(\mu_i;\phi)-(y_i-\mu_i)v(\mu_i;\phi)^{\frac{1}{2}}\kappa_{3i}\right) \quad (27)$$

Among all unbiased EFs constructed from the moment conditions of the over-dispersed model, the OEFs attain the maximum Godambe information [8, 11, 35] equivalently minimizing the asymptotic variance of the joint regression-dispersion estimator of $\theta=(\beta,\phi)$ through the orthogonalization process. The Godambe information is constructed as follows.

Let $g(\theta)$ be an unbiased estimating function satisfying $E\{g(\theta)\}=0$, where $\theta=(\beta,\phi)$. The sensitivity matrix is defined as:

$$S(\theta)=E\left[\frac{\partial g(\theta)}{\partial\theta^{\mathrm{T}}}\right] \quad (28)$$

The variability matrix is defined as:

$$V(\theta)=Var\{g(\theta)\}=E\left[g(\theta)g(\theta)^{\mathrm{T}}\right] \quad (29)$$

The Godambe information matrix is then the sandwich quantity:

$$G(\theta)=S(\theta)^{\mathrm{T}}V(\theta)^{-1}S(\theta) \quad (30)$$

The asymptotic covariance of the estimator $\hat{\theta}$ obtained by solving $g(\theta)=0$ is the Godambe (sandwich) covariance:

$$Cov(\hat{\theta})\approx G(\theta)^{-1}=S(\theta)^{-1}V(\theta)\left(S(\theta)^{-1}\right)^{\mathrm{T}} \quad (31)$$

Heyde [17] shows that because $G(\theta)$ is the inverse of this covariance, maximizing it in the Loewner (positive semi-definite) ordering is equivalent to minimizing the asymptotic covariance. By Theorem 3.2 of Godambe and Thompson [8], the orthogonal construction of (25) yields the member that attains this maximum g^{*} so that the joint OEFs in (26) and (27) deliver the smallest attainable asymptotic covariance for $\theta=(\beta,\phi)$ within the class of unbiased estimating functions generated by the moment conditions [8, 9, 11].

3.2. Penalized Optimal Estimating Function

Penalization techniques have become central to modern feature selection in statistical and machine learning due to their ability to perform simultaneous estimation and variable selection [18, 19].

In this study, Johnson et al. [10] general penalized estimating functions framework for the parameter vector $\beta=(\beta_1,...,\beta_p)^{\mathrm{T}}$ based on a random sample of size n is adopted to construct penalized OEFs.

The penalized version of (26) is given as:

$$g_{pen}^{*}(\beta)=g_{\beta}^{*}-nq_{\lambda}(\|\beta\|)\mathrm{sgn}(\beta) \quad (32)$$

where, $q_{\lambda}=p_{\lambda}'$ is the derivative of a penalty function $p_{\lambda}(\beta)$.

Four penalty functions are considered.

Let $\beta=(\beta_1,...,\beta_p)^{\mathrm{T}}$ denote the coefficient vector in regression setting and let $\lambda>0$ be a tuning (regularization) parameter.

The Ridge penalty (ℓ_2 – penalty) by Hoerl & Kennard [1] imposes quadratic shrinkage on all coefficients.

$$p_{\lambda}^{ridge}(\beta)=\lambda\sum_{j=1}^{p}\beta_j^2 \quad (33)$$

The Lasso penalty (ℓ_1 – penalty) by Tibshirani [2] induces sparsity through an absolute-value penalty in (34).

$$p_{\lambda}^{lasso}(\beta)=\lambda\sum_{j=1}^{p}|\beta_j| \quad (34)$$

The ℓ_1 – penalty performs simultaneous feature shrinkage and selection.

The Adaptive Lasso [3] introduces coefficient-specific weights and can achieve oracle properties under regularity conditions.

$$p_{\lambda}^{adap.lasso}(\beta)=\lambda\sum_{j=1}^{p}\omega_j|\beta_j|,\qquad \omega_j=\frac{1}{|\beta_j|^{\gamma}} \tag{35}$$

where, β_j is an initial consistent estimator such as the unpenalized OEF utilized in this study and $\gamma>0$ controls the degree of adaptivity.

For fixed adaptive weights, the Adaptive LASSO criterion remains convex. Its coefficient-specific weighting applies stronger shrinkage to weak signals and less shrinkage to large coefficients.

The Smoothly Clipped Absolute Deviation (SCAD) penalty [4] reduces bias for large coefficients while retaining sparsity.

The SCAD penalty is obtained by integration:

$$P_{\lambda}^{scad}(\beta)=\begin{cases}\lambda|\beta_j|, & |\beta_j|\le\lambda\\ -\dfrac{|\beta_j|^2-2a\lambda|\beta_j|+\lambda^2}{2(a-1)}, & |\beta_j|\in(\lambda,\,a\lambda)\\ \dfrac{(a+1)\lambda^2}{2}, & |\beta_j|\ge a\lambda\end{cases} \tag{36}$$

The penalty is defined primarily by its first derivative $p'(\beta)$:

$$p'(\beta)=\lambda\left\{I(\beta<\lambda)+\frac{(a\lambda-\beta)_+}{(a-1)\lambda}I(\beta\ge\lambda)\right\} \tag{37}$$

where, $a>2$ is the shape parameter (commonly $a=3.7$) Fan and Li [4] and $(x)_+=\max(x,0)$.

4. Hyper-parameter Selection and Stability

Selection of the tuning parameter is critical for estimation, sparsity and numerical stability. We use a prespecified hierarchy throughout all scenarios. First, $5-$fold cross-validation defines a common $1-$standard error ($1-SE$) admissible band using held-out joint OEF loss. Second, a penalty-specific point is chosen within that band: the right edge for L1_plain and Adaptive LASSO, an interior or left-edge value for Ridge to limit mean-dispersion compensation, and a stability-guarded value for SCAD because its nonconvex objective may admit several local solutions. These rules are held fixed within each response family and are not selected retrospectively from the reported performance measures.

4.1. K-fold Cross-validated Estimating Function Loss

Since penalized Optimal Estimating Functions (OEFs) do not originate from a likelihood setting, hyper-parameter selection cannot rely on deviance-based criteria or classical information criteria such as Akaike Information Criteria (AIC) or Bayesian Information Criteria (BIC) which are primarily likelihood-driven [5, 20].

In this paper, hyper-parameter tuning is aligned with the EFs framework itself and λ is chosen by measuring how well estimates from the training data satisfy the joint OEFs in (26) and (27) for regression and dispersion paramters on independent validation data that is held out. Since penalized OEFs do not derive from a likelihood, the cross-validation loss is the squared norm of the joint estimating function on held-out data rather than a deviance. The EFs loss construction is presented in equations (38) and (39).

Let K denote the number of folds. For a given value of $\lambda>0$, let $\left(\beta_{\lambda}^{(-k)},\phi_{\lambda}^{(-k)}\right)$ denote the OEF estimates obtained by solving the training fold penalized regression and unpenalized dispersion equations in (38):

$$\left.\begin{aligned} g_{\beta}^{*}-nq_{\lambda}(\|\beta\|)\operatorname{sgn}(\beta)&=0\\ g_{\phi}^{*}&=0\end{aligned}\right\} \tag{38}$$

The cross-validated OEF loss is defined as:

$$L_{OEF}(\lambda)=\sum_{k=1}^{K}\left\|\begin{pmatrix} g_{\beta,k}^{*}\left(\beta_{\lambda}^{(-k)},\phi_{\lambda}^{(-k)}\right)\\ g_{\phi,k}^{*}\left(\beta_{\lambda}^{(-k)},\phi_{\lambda}^{(-k)}\right)\end{pmatrix}\right\|^2 \tag{39}$$

where, $g_{\beta,k}^{*}$ and $g_{\phi,k}^{*}$ denote the jointly optimal estimating functions for regression and dispersion parameters, respectively, as presented in (26) and (27) and evaluated using only observations in fold k of size n_k. The function $q_{\lambda}(.)$ is the derivative of a penalty function $p_{\lambda}(\beta)$ with respect to $\|\beta\|$, $\operatorname{sgn}(\beta)$ is the component-wise sign function and $\|.\|$ is the Euclidean norm.

The penalty enters the training-fold system in (38) in the $nq_{\lambda}(\|\beta\|)\operatorname{sgn}(\beta)$ form of Johnson et al. [10]. The intercept and the dispersion equation are left unpenalized. In (39), the stacked vector of unpenalized joint OEF for the regression and dispersion parameters as presented in (26) and (27) is evaluated using only the observations in the held-out fold of size n_k at the training-fold estimate.

The construction thus couples standard $K-$fold cross-validation [18, 21] with the penalized estimating functions of Johnson et al. [10]. Although EFs criteria are cross-validated less often than prediction error (PE), moment-based sample-splitting is well established in Generalized Method of Moments (GMM) model selection and averaging [22, 23].

4.2. Method-specific K-SE Rules and Stability Selection

Hastie et al. [18] argue that empirical cross-validation curves for penalized estimators often exhibit flat or weakly convex minima particularly in the presence of over-dispersion and heterogeneous variance structures. Selecting the value of λ that minimizes the cross-validated loss can therefore lead to unstable models with unnecessary complexity. To address this issue, we select λ using a $k-$standard error ($k-SE$) rule which generalizes the classical $1-SE$ rule of Breiman et al. [24] and Hastie et al. [18]. Instead of taking the strict cross-validation minimum, the SE rule retains the band of λ values whose mean cross-validated loss lies within k standard errors of the minimum, that is, all λ satisfying (40).

$$CV(\lambda) \le CV(\lambda_{\min}) + k \cdot SE(\lambda_{\min}) \tag{40}$$

where, $CV(\lambda)$ is the cross-validated OEF loss of equation (39). In our case, $\lambda_{\min}$ is its minimiser and $SE(\lambda_{\min})$ the standard error of the fold-specific losses at $\lambda_{\min}$.

In this paper we use $k=1$ throughout thereby implementing the standard $1-SE$ rule [25, 26]. The band comprises all values of λ whose loss lies within it, and selecting within this band rather than at the raw minimum addresses the over-selection behaviour of minimum CV tuning while improving stability in sparse variable selection. Within this band the left $k-SE$ value denotes the smallest λ translating to the weakest penalization while the right $k-SE$ value represents the largest λ translating to the sparsest model. The specific point chosen is thus penalty-dependent, being the right $k-SE$ value for L1_plain and Adaptive LASSO and the geometric mean of the left and right $k-SE$ values for Ridge, with SCAD handled separately by a stability-augmented rule presented in Section 5.1.2. This family of rules favours sparser and more stable models while maintaining near optimal predictive performance and has been shown to improve stability and interpretability in penalized regression settings [18].

The cross-validation profiles in Figures 1-8 are tuning diagnostics, not a formal sensitivity analysis. Because the simulation archive did not retain results under alternative values of the standard-error multiplier, stability threshold or guard, numerical sensitivity claims are not made here. The effect of varying these tuning constants should be evaluated in future work before a single default rule is recommended across response families.

4.3. Definition of the Selection Metrics

Two related selection metrics are reported. A coefficient counts as selected when the absolute penalized estimate exceeds the penalty-based threshold. The true positive rate (TPR) is the proportion of active coefficients that meet this rule. For null coefficients, the reported quantity applies an additional 2.5-standard-error screen and is therefore denoted the significance-filtered false positive rate (sFPR). Thus, sFPR measures null coefficients that survive both the penalty threshold and the subsequent Godambe significance screen. It should not be interpreted as the false-positive rate of the selected support alone. The additional screen follows the universal-threshold motivation in [27], but direct support-based FPR should be reported in future comparisons.

5. Simulation Studies

In this section, we assess the penalized OEF framework through Monte Carlo simulation under the two over-dispersed regression models of Section 2. For each model six estimators are compared, and these are the Quasi-likelihood baseline, the Unpenalized OEF and the four penalized OEF variants namely Ridge (L2), LASSO (L1_plain), Adaptive LASSO (L1_adapt) and SCAD. Each model is examined over a 2×2 factorial of dispersion level and signal density. We utilize $s \in \{4,8\}$ active features among $p=20$ candidate features with $R=500$ Monte Carlo replications per scenario. The covariates are generated with a first-order autoregressive (AR (1)) correlation structure of 0.5 which introduces the dependence among covariates that is common in practice.

Hyper-parameter tuning is carried out based on the cross-validated OEF loss in (39) with the method-specific $k-SE$ and stability criteria presented in Section 4.2. We present the simulation results in the next sub-sections and their discussion in section 5.3.

5.1. Beta-Binomial Logistic Regression

This section presents the finite-sample performance of the penalized OEFs for feature selection in the Beta-Binomial (BB) logistic regression model. The BB model accommodates over-dispersion in grouped binary data through an intra-cluster correlation parameter ρ that controls excess variability beyond the binomial case [12]. The variance function is as given in equation (7) and it exceeds the binomial variance by the multiplicative inflation factor $1+(m-1)\rho$.

The data-generating process (DGP) is such that, for each replication, $p=20$ covariates are drawn from a mean-zero multivariate normal distribution with a first-order autoregressive correlation structure (AR (1)) with correlation 0.5. This is a moderate level that introduces collinearity without extreme multicollinearity. Given a coefficient vector β, the success probability for i^{th} observation follows the logistic link in (4) and the expected count is as in (3) with $m=40$ trials per cluster. The response is then drawn as $Y_i \sim$ Beta-Binomial (m,μ_i,ρ) generated through the Beta-Binomial mixture in which the binomial success probability is itself Beta-

distributed with shape parameters $a=\mu_i(1-\rho)/(m\rho)$ and $b=(m-\mu_i)(1-\rho)/(m\rho)$, such that ρ is exactly the intra-cluster correlation. The covariates are standardized within each replication and $R=500$ independent datasets are generated per scenario under a common random seed to ensure that all six estimators are compared on identical data.

5.1.1. Simulation Design

Four scenarios are constructed by crossing two overdispersion levels $\rho=0.20$ (moderate over-dispersion) and $\rho=0.65$ (heavy over-dispersion) with two active-set sizes $s=4$ and $s=8$ while holding sample size $n=100$, number of features $p=20$, cluster size $m=40$, and $R=500$ Monte Carlo replications. The sparsity ratios are therefore $s/p=0.2$ (sparse) and $s/p=0.4$ (moderately dense). The shift from $s/p=0.2$ to $s/p=0.4$ is included to probe how the penalized OEF framework behaves under stress rather than to evaluate a regime for which the methods were optimally designed. Table 1 summarizes the design.

Table 1. *Design of the four Beta-Binomial simulation scenarios.*

Scenario	n	p	s	s/p	$\rho\_$ true	Non-zero β	Design
1	100	20	4	0.2	0.20	$\beta_0=0.40$; $\beta_1=0.60$, $\beta_4=-0.85$, $\beta_7=0.70$, $\beta_{12}=0.55$,	Sparse, moderate OD
2	100	20	4	0.2	0.65	$\beta_0=0.40$; $\beta_1=0.60$, $\beta_4=-0.85$, $\beta_7=0.70$, $\beta_{12}=0.55$,	Sparse, high OD
3	100	20	8	0.4	0.20	$\beta_0=0.40$; $\beta_1=0.50$, $\beta_2=0.60$, $\beta_4=-0.85$, $\beta_6=-0.50$, $\beta_7=0.70$, $\beta_9=0.45$, $\beta_{12}=0.55$, $\beta_{20}=0.50$	Dense, moderate OD
4	100	20	8	0.4	0.65	$\beta_0=0.40$; $\beta_1=0.50$, $\beta_2=0.60$, $\beta_4=-0.85$, $\beta_6=-0.50$, $\beta_7=0.70$, $\beta_9=0.45$, $\beta_{12}=0.55$, $\beta_{20}=0.50$	Dense, high OD

n = sample size; p = candidate predictors; s = active predictors; s/p = sparsity ratio; $\rho\_$ true = intra-cluster correlation

Under this design six estimators are evaluated and these are the Quasibinomial GLM, the unpenalized OEF, Ridge (L2-OEF), L1_plain (LASSO-OEF), Adaptive LASSO and SCAD. The Quasibinomial GLM is the standard moment-based benchmark that estimates β by iteratively reweighted least squares with a quasi-binomial variance and estimates ρ by the Pearson Method-of-Moments (MoM) χ^2 statistic. The unpenalized OEF is the Godambe efficient benchmark that jointly estimates the regression and dispersion parameters through the set of orthogonalized estimating functions in (26) and (27) which incorporate the third and fourth-order moment information. The four penalized estimators apply L1, adaptive L1, L2 and SCAD penalties to the regression equation.

5.1.2. Hyperparameter Selection and Tuning Diagnostics

Hyper-parameter selection for the BB model follows the cross-validated OEF loss of equation (39) with the $k-SE$ band and method-specific selection rules as discussed in Section 4.2. For each candidate λ, the regularization parameter is chosen by $K-$ fold cross-validation ($K=5$) of the held-out OEF norm. The remainder of this sub-section reports the BB-specific behavior of this criterion.

We apply the $k-SE$ selection rule and its standard-error band presented in equation (40). For L1_plain and Adaptive LASSO, λ^* is taken as the right $k-SE$ value which represents the largest λ whose cross-validated loss remains within the $1-SE$ band. This selection favours the sparsest model that the data do not distinguish from the cross-validation minimum.

Under the Ridge framework we utilize a different choice. This is because over-shrinking β worsens the mean fit and makes the dispersion equation compensate by absorbing the unexplained variability into ρ. This systematically inflates the dispersion estimate and the coupling between the mean and dispersion equations is well documented in the over-dispersed models literature where variability left unexplained by the mean model is instead absorbed by the dispersion

parameter [28-30]. This means that selecting the right $k-SE$ value would worsen the over-shrinkage while the left edge under-regularizes and so we set λ^* for Ridge to the geometric mean of the left $k-SE$ and right $k-SE$ values. This is an intermediate point within the band that stabilizes joint mean-dispersion inference while retaining near-optimal predictive adequacy.

SCAD requires a further adjustment since the $1-SE$ rule alone does not suffice. This is because the concave penalty produces characteristically flat CV surfaces in the intermediate λ region and because the penalty is nonconcave the optimization objective may admit several local solutions [4, 31, 32] leaving the loss insensitive to λ across a wide range.

Stability selection, as introduced by Meinshausen and Bühlmann [33] gauges whether a variable belongs in the model by how reliably it is chosen when the data or tuning parameters are perturbed. We therefore tune SCAD with a stability-augmented rule ($\tau = 0.90$; guard $=1.05\times$) such that λ^* is placed where the CV loss first turns sharply upward and reverting to the CV minimum whenever the safeguard is not met. The threshold τ is the smallest selection frequency at which a coefficient is treated as reliably retained and is conventionally fixed within the 0.6-0.9 range [33]. On the other hand, the guard borrows the logic of the $1-SE$ rule by admitting the stable λ only when its loss stays within 5% of the best value so that predictive accuracy is not sacrificed [34].

Figures 1-4 present the CV OEF loss profiles for the four penalized scenarios.

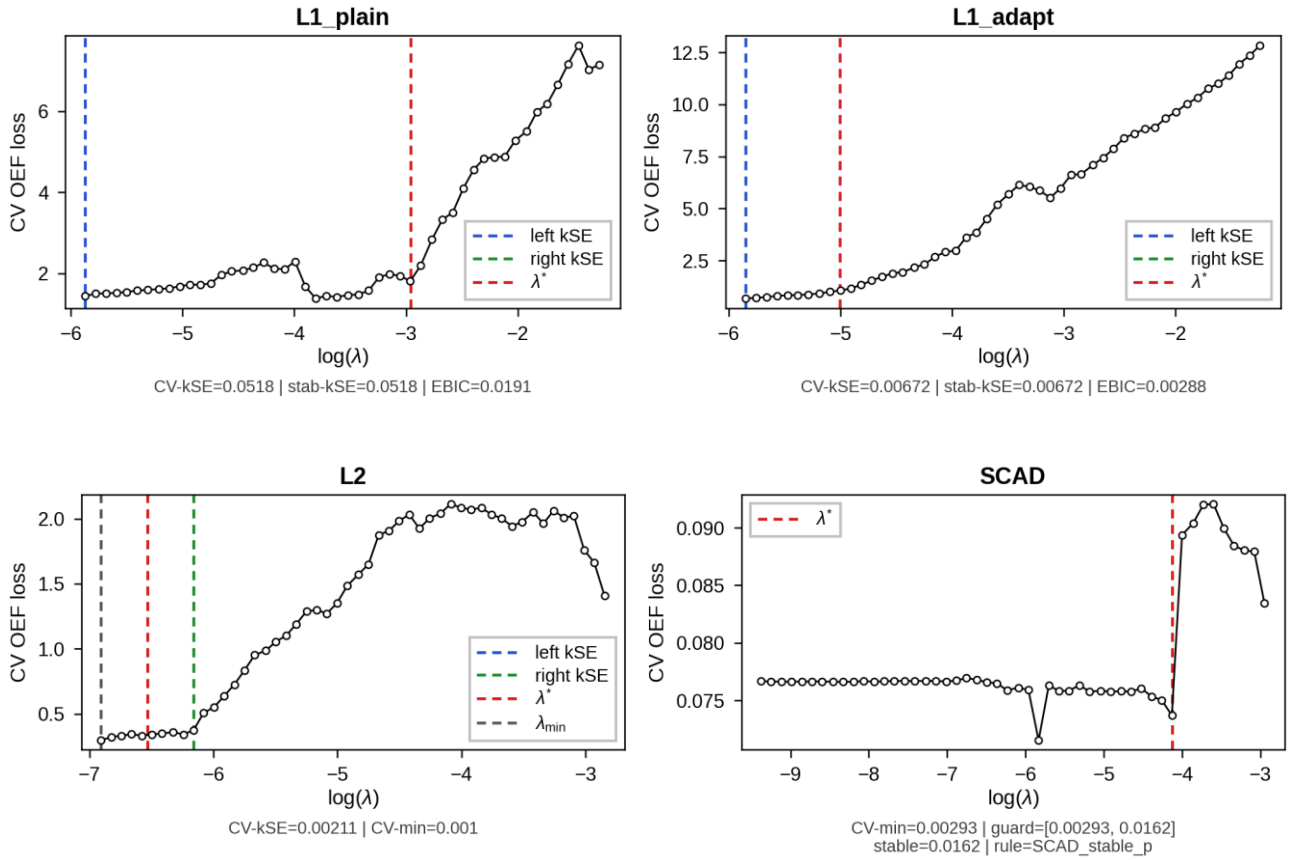


Figure 1. *Cross-validated OEF loss profiles for Scenario 1 ($s = 4$, $\rho = 0.20$).*

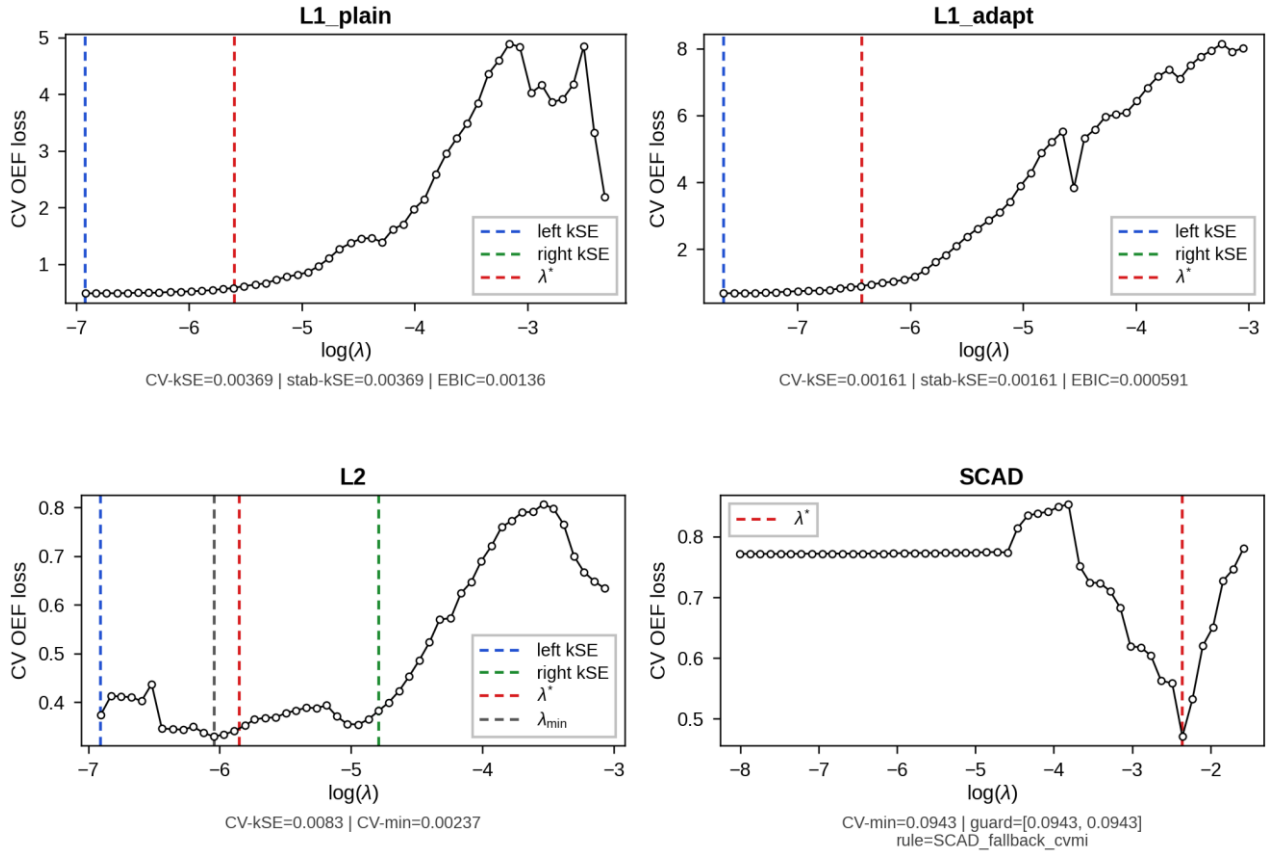


Figure 2. *Cross-validated OEF loss profiles for Scenario 2 ($s = 4$, $\rho = 0.65$).*

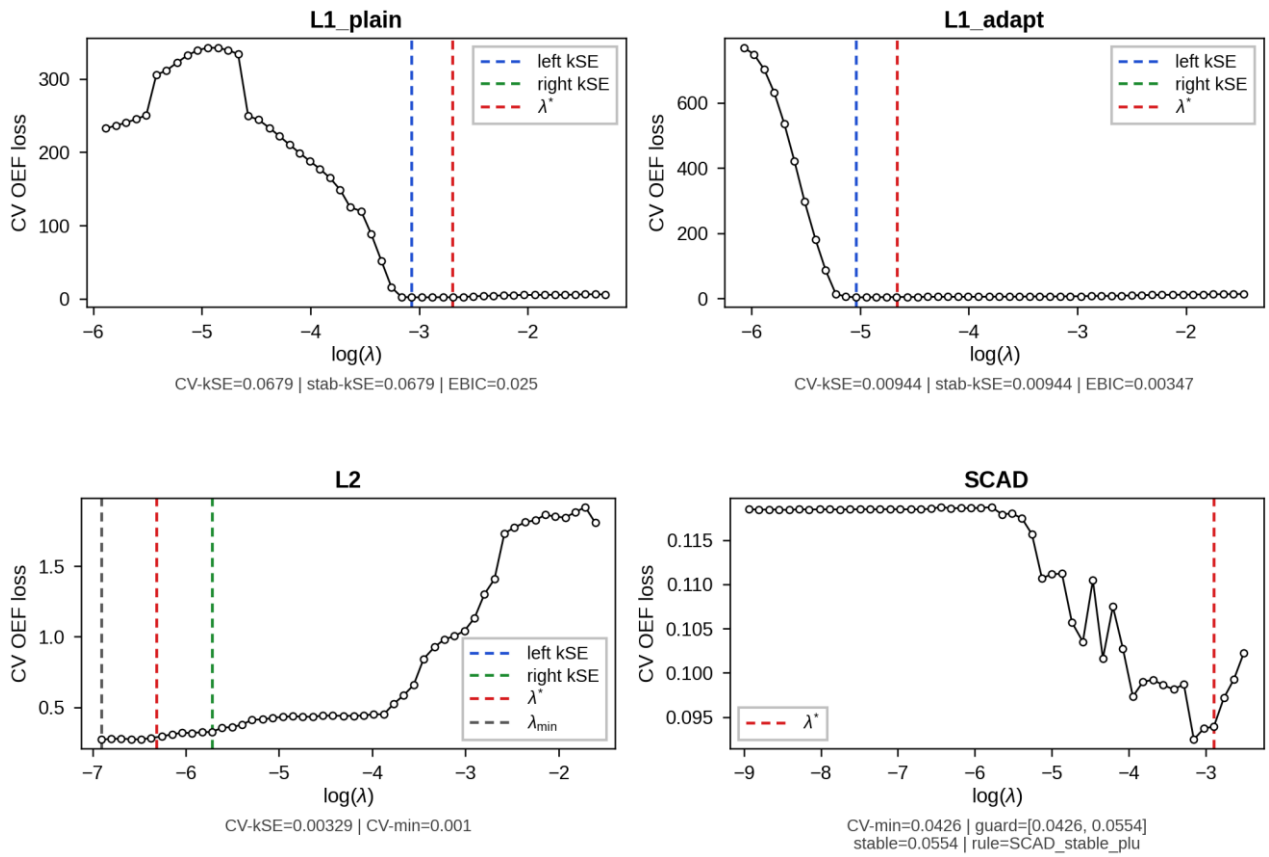


Figure 3. *Cross-validated OEF loss profiles for Scenario 3 ($s = 8$, $\rho = 0.20$).*

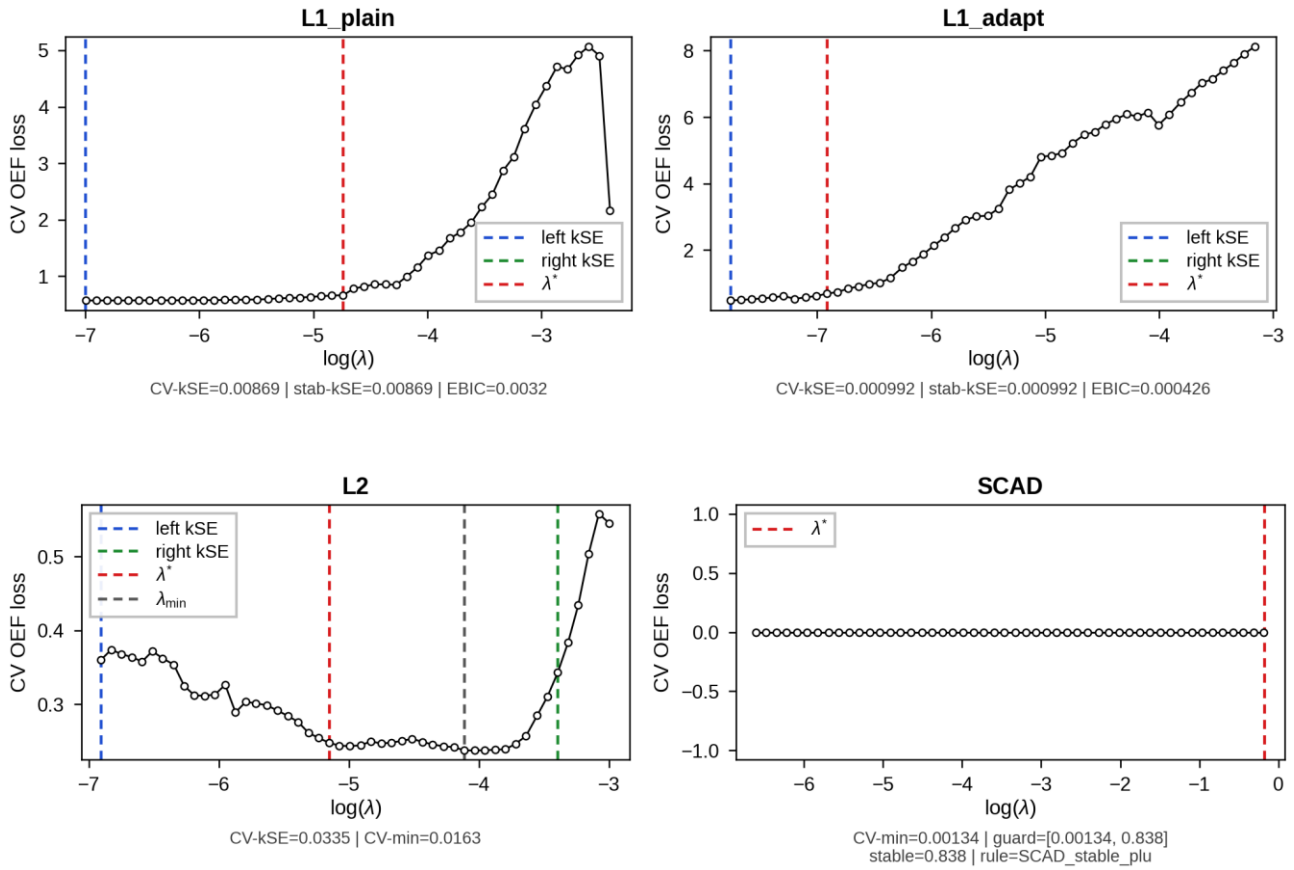


Figure 4. *Cross-validated OEF loss profiles for Scenario 4 ($s = 8$, $\rho = 0.65$).*

Across all scenarios the L1 and Adaptive LASSO profiles are monotone with well-defined minima, and the selected λ^* values increase modestly from $s/p = 0.2$ to $s/p = 0.4$, consistent with the denser active set $(\beta \neq 0)$ requiring slightly heavier penalization. The L2 profile follows the classical ridge shape [1] at $\rho = 0.20$ but develops a broader trough $\rho = 0.65$ reflecting inflated OEF estimating-equation variance. The SCAD panel is flat across the intermediate λ region in all scenarios and it becomes extremely flat at $\rho = 0.65$ and is most pronounced in Scenario 4. This is consistent with the nonconvex penalty's sensitivity to the BB score geometry at high intra-cluster correlation.

5.1.3. Regression Coefficient Coverage and Feature Selection

Consider first the two sparse scenarios $s/p = 0.2$. Tables 2 and 3 report unconditional 95% Wald coverage averaged over the active $(\beta \neq 0)$ and null $(\beta = 0)$ feature sets. Additionally, they report the selection performance of the respective models.

In Scenario 1, the unpenalized OEF yields active-coefficient coverage of 0.899. Although this is the strongest OEF benchmark in the scenario, it remains 0.051 below the nominal 0.95 level and should not be described as near nominal. Ridge coverage is lower at 0.831. Shrinkage estimators trade reduced variance for bias, which can reduce confidence-interval coverage [36, 37].

For the selectors, L1_plain and Adaptive LASSO give active-coefficient coverages of 0.987 and 0.991, respectively. This overcoverage is associated with bootstrap standard-error inflation in the joint OEF system. SCAD coverage is 0.896, below the nominal 0.95 level, and its sFPR is higher at 0.0620.

Table 2. *Coverage probabilities and feature selection - Scenario 1 ($s = 4$, $\rho = 0.20$).*

Method	Cov. β active	Cov. β null	TPR	sFPR
Quasibinomial	0.768	0.936	-	-

Method	Cov. β active	Cov. β null	TPR	sFPR
Unpenalized OEF	0.899	0.900	-	-
Ridge (L2)	0.831	0.876	-	-
L1_plain (LASSO)	0.987	0.985	0.997	0.0060
Adaptive LASSO	0.991	0.991	0.997	0.0035
SCAD	0.896	0.879	0.992	0.0620

Scenario 2 ($\rho = 0.65$) which is a harder setting for inference. The unpenalized OEF's active coverage drops to 0.879, a little under nominal but still reasonable given the heavier over-dispersion. The selectors show an improved performance: L1_plain and Adaptive LASSO recover almost every true signal with True Positive Rate (TPR) = 0.998 and TPR=0.999 while letting through virtually no noise, that is, sFPR= 0.0001 and sFPR = 0.0000 respectively. Additionally, SCAD, the least conservative of the three, keeps its sFPR at 0.0690.

Table 3. *Coverage probabilities and feature selection -Scenario 2 (* $s = 4$, $\rho = 0.65$ *).*

Method	Cov. β active	Cov. β null	TPR	sFPR
Quasibinomial	0.846	0.932	-	-
Unpenalized OEF	0.879	0.898	-	-
Ridge (L2)	0.856	0.880	-	-
L1_plain (LASSO)	0.999	0.999	0.998	0.0001
Adaptive LASSO	1.000	1.000	0.999	0.0000
SCAD	0.896	0.892	0.998	0.0690

Increasing the active set to $s = 8$ moves the design into a moderately dense regime ($s / p = 0.4$). Tables 4 and 5 report the results. As the model takes on more predictors, sFPRs climb even when tuning across penalties is kept fixed. Bühlmann and van de Geer [19] notes that fitting more true signals within the same sample size leaves less room to separate genuine nulls from weak active effects. In Scenario 3, the numbers show this directly. L1_plain's sFPR nearly triples from 0.0060 in Scenario 1 to 0.0172 and SCAD's rises from 0.0620 to 0.1080. This heightened sensitivity of the concave penalty to signal density is a well documented phenomenon by Zhang [31]. The false positives could have been reduced with more conservative λ choices. Tuning was held fixed across scenarios for comparability.

Table 4. *Coverage probabilities and feature selection - Scenario 3 (* $s = 8$, $\rho = 0.20$ *).*

Method	Cov. β active	Cov. β null	TPR	sFPR
Quasibinomial	0.718	0.933	-	-
Unpenalized OEF	0.889	0.895	-	-
Ridge (L2)	0.844	0.887	-	-
L1_plain (LASSO)	0.970	0.972	0.994	0.0172
Adaptive LASSO	0.996	0.994	0.998	0.0008

Method	Cov. β active	Cov. β null	TPR	sFPR
SCAD	0.848	0.839	0.993	0.1080

Scenario 4 is the most demanding BB setting. Post-refit coverage remains below 0.95 as reported in Section 5.1.5, so the refit cannot be said to restore nominal validity. SCAD records the largest sFPR at 0.1170, whereas L1_plain and Adaptive LASSO record sFPR = 0.0000 and TPR = 0.998 under the significance-filtered definition.

Table 5. *Coverage probabilities and feature selection - Scenario 4 ($s = 8$, $\rho = 0.65$).*

Method	Cov. β active	Cov. β null	TPR	sFPR
Quasibinomial	0.834	0.924	-	-
Unpenalized OEF	0.876	0.890	-	-
Ridge (L2)	0.866	0.889	-	-
L1_plain (LASSO)	1.000	1.000	0.998	0.0000
Adaptive LASSO	1.000	1.000	0.998	0.0000
SCAD	0.849	0.860	0.997	0.1170

5.1.4. Estimation Accuracy

Tables 6-9 report RMSE and signed bias for the regression coefficients. Uhlmann and van de Geer [19] note that reporting both quantities separates systematic distortion from total estimation error.

The unpenalized OEF gives the lowest active RMSE in every scenario, 0.181-0.421. Its active RMSE grows exponentially from $\rho = 0.20$ to $\rho = 0.65$ as the OEF estimating equation variance itself rises under high intra-cluster correlation. Ridge is moderate at $\rho = 0.20$ (0.320-0.359) but climbs sharply at $\rho = 0.65$ (0.700-0.781) and SCAD carries the largest active RMSE at $\rho = 0.65$ (1.139-1.325). This is in line with the convergence instabilities predicted by nonconvex penalty theory [4, 32].

At $\rho = 0.65$ Ridge outperforms the L1-type methods in the sparse case of Scenario 2 whereas L1_plain performs better than Ridge in the dense case of Scenario 4.

Table 6. *Bias and RMSE - Scenario 1 ($s = 4$, $\rho = 0.20$).*

Method	RMSE $\hat{\beta}$ active	Bias $\hat{\beta}$ active	RMSE $\hat{\beta}$ null	Bias $\hat{\beta}$ null
Unpenalized OEF	0.181	+0.016	0.169	0.000
Ridge (L2)	0.359	+0.043	0.251	0.000
L1_plain (LASSO)	0.572	+0.098	0.451	+0.001
Adaptive LASSO	0.605	+0.064	0.480	+0.010
SCAD	0.767	+0.055	0.589	−0.001

Table 7. Bias and RMSE - Scenario 2 ($s = 4$, $\rho = 0.65$).

Method	RMSE $\hat{\beta}$ active	Bias $\hat{\beta}$ active	RMSE $\hat{\beta}$ null	Bias $\hat{\beta}$ null
Unpenalized OEF	0.388	+0.063	0.342	−0.001
Ridge (L2)	0.700	+0.132	0.517	−0.004
L1_plain (LASSO)	0.826	+0.181	0.662	−0.008
Adaptive LASSO	0.898	+0.212	0.776	−0.003
SCAD	1.325	+0.153	0.977	−0.009

Table 8. Bias and RMSE - Scenario 3 ($s = 8$, $\rho = 0.20$).

Method	RMSE $\hat{\beta}$ active	Bias $\hat{\beta}$ active	RMSE $\hat{\beta}$ null	Bias $\hat{\beta}$ null
Unpenalized OEF	0.193	+0.017	0.181	−0.001
Ridge (L2)	0.320	+0.022	0.246	+0.004
L1_plain (LASSO)	0.888	+0.147	0.673	+0.034
Adaptive LASSO	0.479	+0.075	0.401	−0.005
SCAD	0.414	+0.024	0.375	+0.006

Table 9. Bias and RMSE - Scenario 4 (($s = 8$, $\rho = 0.65$).

Method	RMSE $\hat{\beta}$ active	Bias $\hat{\beta}$ active	RMSE $\hat{\beta}$ null	Bias $\hat{\beta}$ null
Unpenalized OEF	0.421	+0.063	0.377	−0.002
Ridge (L2)	0.781	+0.153	0.605	−0.006
L1_plain (LASSO)	0.728	+0.153	0.609	−0.014
Adaptive LASSO	1.013	+0.209	0.796	−0.018
SCAD	1.139	+0.165	0.992	+0.004

5.1.5. Post-selection and Post-refit Inference

Post-selection and post-refit results are reported below across the selectors and simulation scenarios.

Confidence intervals calculated after data-driven selection are not generally valid because selection changes the sampling distribution of the estimator [38, 39]. An unpenalized OEF refit on the selected sub-model can reduce shrinkage bias, but it still conditions on a random selected support and therefore does not, by itself, provide valid post-selection inference [40, 41].

In the OEF framework, the refit uses the unpenalized estimating equations and the sandwich covariance in Equation (31) on the selected sub-model. Table 10 compares naive conditional coverage from the penalized fit with post-refit coverage. The comparison evaluates whether refitting improves calibration; it is not treated as proof that nominal coverage has been restored.

Table 10. *Naive conditional and post-refit coverage - Scenarios 1-4.*

Scenario	Method	Naive active	Naive null	Post-refit active	Post-refit null
1	L1_plain (LASSO)	0.987	0.985	0.834	0.821
	Adaptive LASSO	0.991	0.991	0.874	0.865
	SCAD	0.896	0.879	0.786	0.765
2	L1_plain (LASSO)	0.999	0.999	0.899	0.914
	Adaptive LASSO	1.000	1.000	0.900	0.917
	SCAD	0.896	0.892	0.815	0.817
3	L1_plain (LASSO)	0.970	0.972	0.839	0.838
	Adaptive LASSO	0.996	0.994	0.855	0.855
	SCAD	0.848	0.839	0.761	0.752
4	L1_plain (LASSO)	1.000	1.000	0.889	0.907
	Adaptive LASSO	1.000	1.000	0.896	0.914
	SCAD	0.849	0.860	0.725	0.735

Taking the naive conditional coverage first, for L1_plain and Adaptive LASSO it is near-perfect at $\rho = 0.65$ (0.999-1.000) and moderately above nominal at $\rho = 0.20$ (0.970-0.991) with the excess reflecting bootstrap SE inflation. SCAD's naive active coverage of 0.848 to 0.896 sits near nominal, in line with its oracle property. Active coefficients above the selection threshold ($a\lambda$, $a = 3.7$) receive near zero shrinkage. Therefore, the Godambe CI approximates that of the unpenalized OEF.

Post-refit active coverage remains below 0.95. For L1_plain it ranges from 0.889 to 0.900 in the lower-dispersion scenarios and from 0.834 to 0.874 in the higher-dispersion scenarios; Adaptive LASSO shows a similar pattern. The residual deficit reflects selection uncertainty that a simple unpenalized refit does not remove [38, 39]. Refitting may reduce shrinkage bias, but these results do not support nominal post-selection inference.

For SCAD, post-refit active coverage ranges from 0.725 in Scenario 4 to 0.815 in Scenario 2. The deterioration is strongest under high BB overdispersion. L1_plain and Adaptive LASSO perform better than SCAD on this measure, but their coverage also remains below 0.95; none of the reported refit procedures therefore provides dependable post-selection inference.

5.2. Negative-Binomial Regression

This section reports the finite-sample performance of the penalized OEFs for feature selection in the Negative-Binomial (NB2) regression model in which the variance is quadratic in the mean. The NB2 model parameterizes over-dispersion through a heterogeneity parameter α in the variance function of equation (17) and it nests the Poisson model as the limiting case $\alpha \to 0$ [13, 15, 42]. The conditional mean is specified through the log link in equation (14), which guarantees positive fitted counts, and the exponential mean structure that results has a direct bearing on penalty behaviour. Under the log link a coefficient acts multiplicatively on the mean so that small shrinkage in β produces multiplicative distortion in the fitted μ [15, 42]. This distortion is amplified through the quadratic variance function into the dispersion estimate, as shown by the results in Sections 5.2.3 and 5.2.4.

The data-generating process (DGP) borrows from this specification directly. For each replication, $p = 20$ covariates are drawn from a mean-zero multivariate normal distribution with a first-order autoregressive correlation structure (AR (1) and correlation 0.5. Given a coefficient vector β, the conditional mean for the i^{th} observation follows the log link in equation (14) and the mean count in equation (15). The response is then drawn as $Y_i \sim$ Negative-Binomial (μ_i, α) with the variance as specified in equation (17). Like the Beta-Binomial Logistic regression simulation study in Section 5.1, $R = 500$ independent datasets are generated per simulation scenario under a common random seed. We increase the sample size in this setting to $n = 200$ to accommodate the heavier tails of count responses.

5.2.1. Simulation Design

Four simulation scenarios cross two over-dispersion levels $\alpha = 0.25$ and $\alpha = 0.75$ with two active coefficient set sizes $s = 4$ and $s = 8$ while fixing $n = 200$, $p = 20$, and

$R = 500$ replications are considered. The variance-to-mean ratio is VMR$= 1 + \alpha\mu_i$. Table 11 summarizes the simulation design.

Table 11. *Design of the four Negative-Binomial simulation scenarios.*

Scenario	n	p	s	s/p	α _ true	Non-zero β	Design
1	200	20	4	0.2	0.25	$\beta_0 = 0.25$; $\beta_1 = 0.50$, $\beta_4 = -0.75$, $\beta_8 = 0.65$, $\beta_{15} = 0.45$	Sparse, moderate OD
2	200	20	4	0.2	0.75	$\beta_0 = 0.25$; $\beta_1 = 0.50$, $\beta_4 = -0.75$, $\beta_8 = 0.65$, $\beta_{15} = 0.45$	Sparse, high OD
3	200	20	8	0.4	0.25	$\beta_0 = 0.25$; $\beta_1 = 0.20$, $\beta_4 = -0.75$, $\beta_6 = 0.30$, $\beta_8 = 0.60$, $\beta_{11} = 0.35$, $\beta_{15} = 0.24$, $\beta_{17} = -0.40$, $\beta_{19} = 0.45$	Dense, moderate OD
4	200	20	8	0.4	0.75	$\beta_0 = 0.25$; $\beta_1 = 0.20$, $\beta_4 = -0.75$, $\beta_6 = 0.30$, $\beta_8 = 0.60$, $\beta_{11} = 0.35$, $\beta_{15} = 0.24$, $\beta_{17} = -0.40$, $\beta_{19} = 0.45$	Dense, high OD

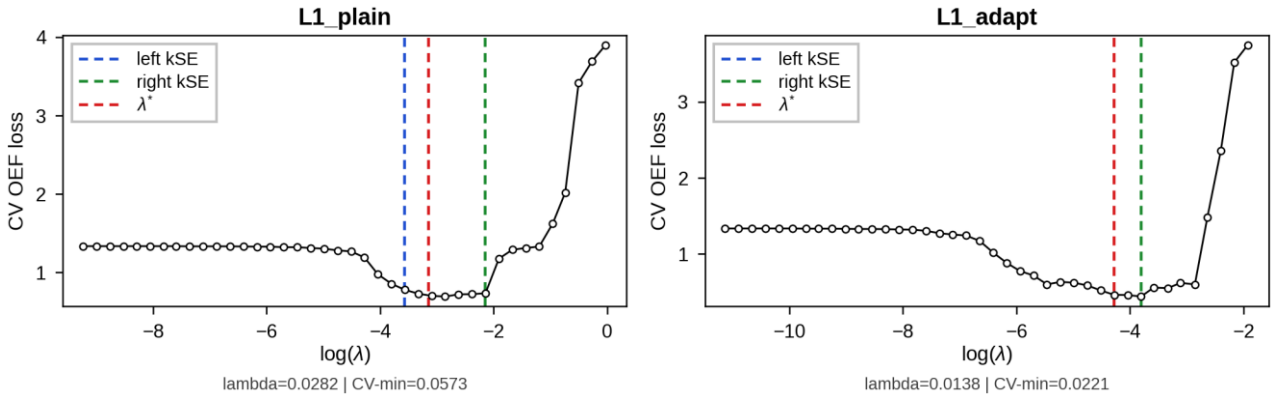
n = sample size; p = candidate predictors; s = active predictors; s/p = sparsity ratio; α _ true = NB2 heterogeneity

The same six estimators are evaluated as in Section 5.1 and these are the Quasi-Poisson GLM, the unpenalized OEF, Ridge (L2-OEF), L1_plain (LASSO-OEF), Adaptive LASSO and SCAD. The QuasiPoisson GLM estimates β by iteratively reweighted least squares with a Quasi-Poisson variance and dispersion by the Pearson Method of Moments (MoM) statistic.

5.2.2. Hyperparameter Selection and Tuning Diagnostics

Tuning of the penalization regularization parameters in the NB2 joint OEF framework follows the cross-validated OEF loss in (39) and the method-specific $k-$SE rules of Section 4.2, the same framework used for the BB model in Section 5.1.2. For L1_plain and Adaptive LASSO, λ^* is the right $k-$SE value within the $1-$SE band. For Ridge it is the left-edge boundary, which avoids the over-shrinkage-dispersion compensation trade-off described in Section 5.1.2 and which operates here with an amplified effect under the exponential mean. For SCAD we apply the same stability-augmented rule described in Section 5.1.2 with the NB2 settings $\tau = 0.80$ and guard = 1.10×. The guard falls back to the CV minimum in all four NB2 scenarios.

Figures 5-8 present the CV OEF loss profiles for the four scenarios.



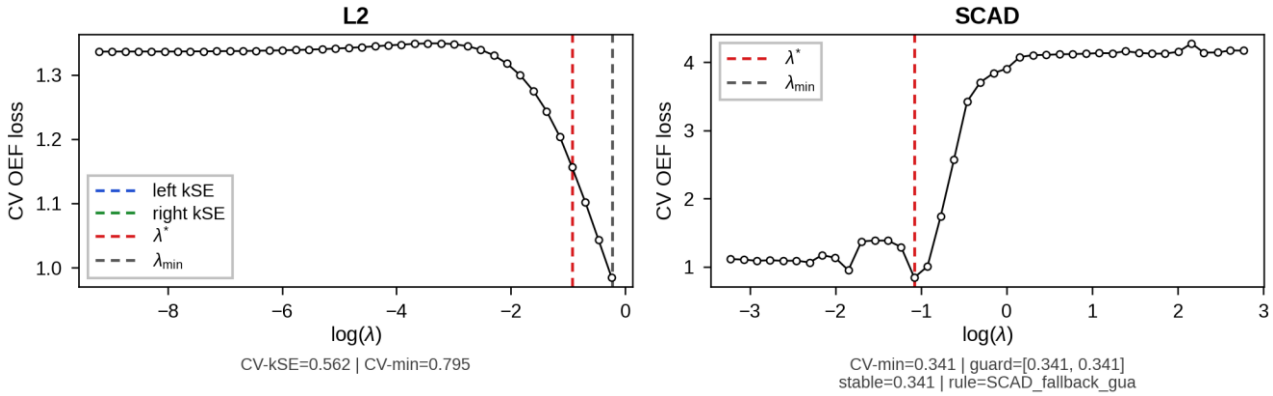


Figure 5. *Cross-validated OEF loss profiles for Scenario 1 ($s = 4$, $\alpha = 0.25$).*

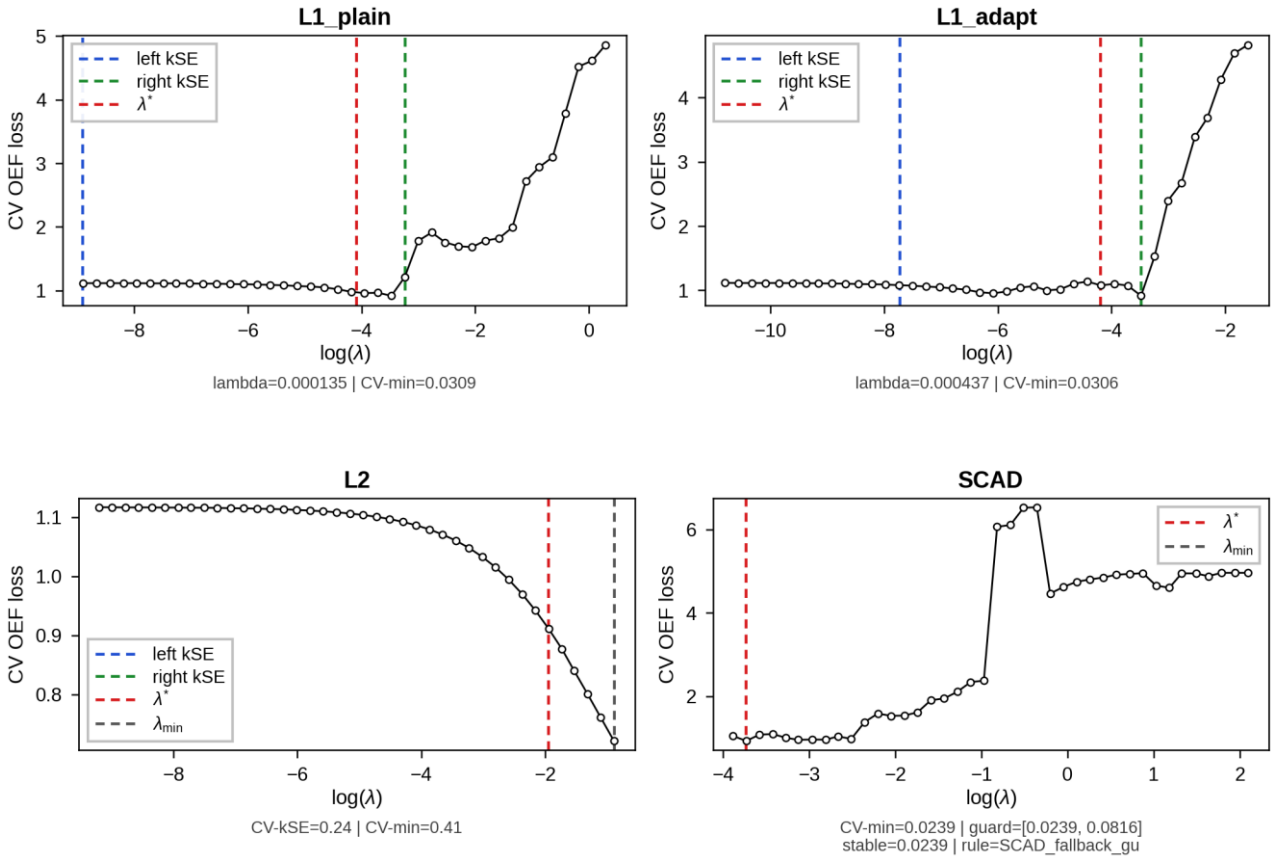
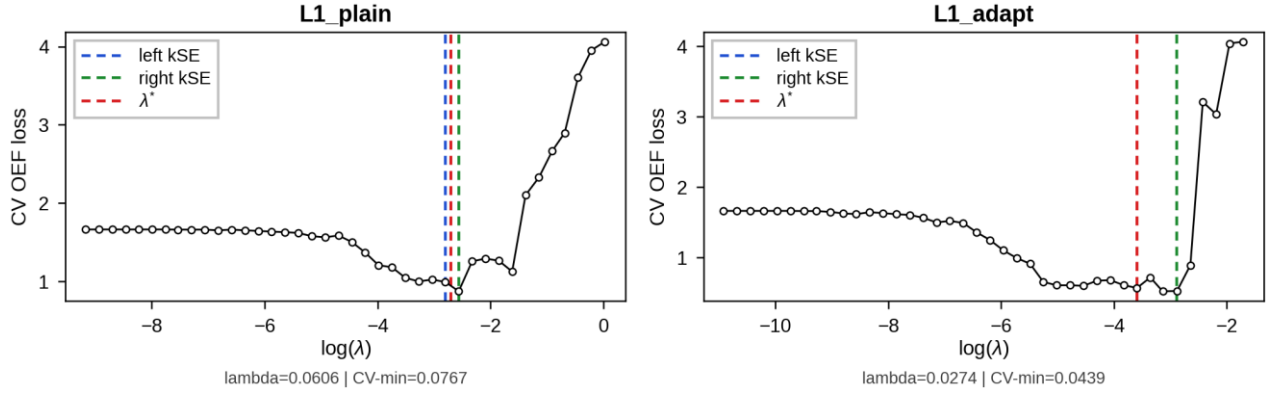


Figure 6. *Cross-validated OEF loss profiles for Scenario 2 ($s = 4$, $\alpha = 0.75$).*



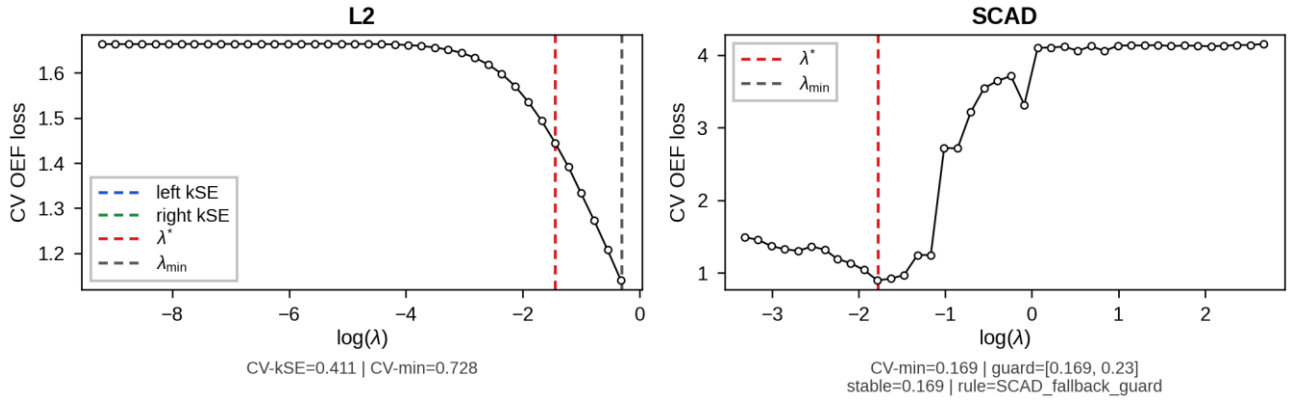


Figure 7. *Cross-validated OEF loss profiles for Scenario 3 ($s = 8$, $\alpha = 0.25$).*

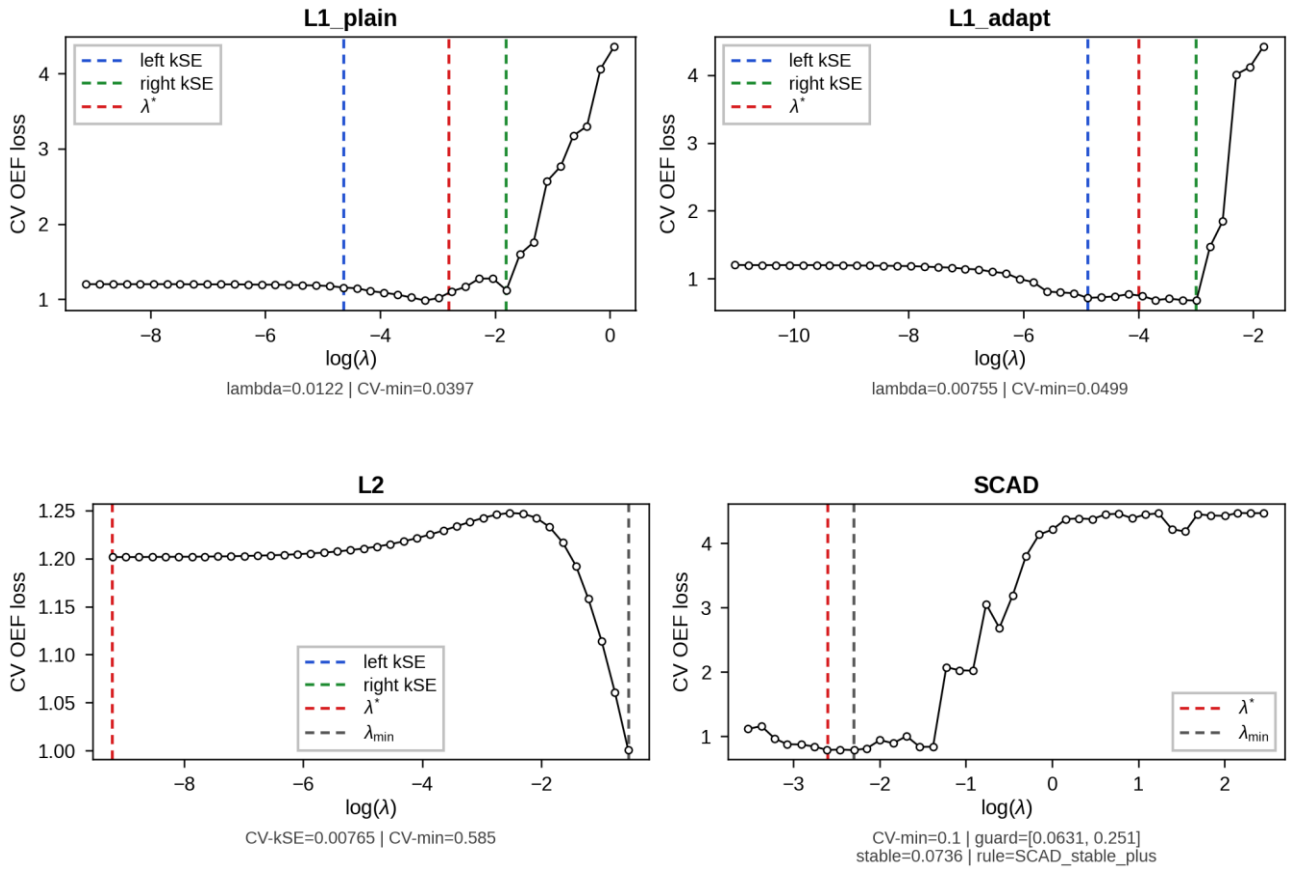


Figure 8. *Cross-validated OEF loss profiles for Scenario 4 ($s = 8$, $\alpha = 0.75$).*

The L1 and Adaptive LASSO profiles are monotone with well-defined minima in every NB2 scenario. The L2 profile keeps the classical ridge U-shape at both dispersion levels with its trough broadening at extreme over-dispersion levels ($\alpha = 0.75$) in Scenarios 2 and 4. The SCAD profiles retain some curvature even at $\alpha = 0.75$. This behaviour reflects the numerical stability of the NB2 score function under high over-dispersion, a property linked in the literature to the favourable geometry of the NB2 score and estimating equations [43]. We interpret the greater stability observed in this case as a result of the smoother estimating function geometry induced by the exponential mean, as described in Section 5.2.4.

5.2.3. Regression Coefficient Coverage and Feature Selection

Consider first the two sparse scenarios $s / p = 0.2$. Tables 12 and 13 report unconditional 95% Wald coverage and feature selection performance.

Ridge active coverage by contrast collapses to 0.284 and the cause is the exponential mean. Under the log link linear shrinkage in β becomes multiplicative shrinkage in μ , so a

small reduction in β is exponentially amplified into μ distortion. This inflates the variance estimates and degrades the CI calibration, and it is the effect that the bounded BB response of Section 5.1 avoids.

For the selectors, L1_plain and Adaptive LASSO active-coefficient coverages are 0.889 and 0.892, respectively, while SCAD coverage is 0.887. All three values are below the nominal 0.95 level. SCAD has the lowest sFPR at 0.001 in this scenario.

Table 12. *Coverage probabilities and feature selection - Scenario 1 ($s = 4$, $\alpha = 0.25$).*

Method	Cov. β active	Cov. β null	TPR	sFPR
Unpenalized OEF	0.918	0.905	-	-
Ridge (L2)	0.284	0.819	-	-
L1_plain (LASSO)	0.889	0.913	0.982	0.0280
Adaptive LASSO	0.892	0.829	0.982	0.0510
SCAD	0.887	0.996	0.895	0.0010

Scenario 2 confirms the contrast. Ridge active coverage improves to 0.556 but remains far below nominal. SCAD active coverage is 0.959, slightly above 0.95, with sFPR = 0.0055. The result is scenario-specific and does not establish uniform coverage superiority.

Table 13. *Coverage probabilities and feature selection - Scenario 2 ($s = 4$, $\alpha = 0.75$).*

Method	Cov. β active	Cov. β null	TPR	sFPR
Unpenalized OEF	0.911	0.912	-	-
Ridge (L2)	0.556	0.885	-	-
L1_plain (LASSO)	0.909	0.933	0.999	0.0214
Adaptive LASSO	0.898	0.892	0.996	0.0333
SCAD	0.959	0.987	0.982	0.0055

Tables 14 and 15 present the results to the denser scenarios ($s/p = 0.4$). The move from $s/p = 0.2$ to $s/p = 0.4$ brings out a distinct NB2 selection pattern. L1_plain active coverage falls from 0.889 in Scenario 1 to 0.847 in Scenario 3 and Adaptive LASSO from 0.892 to 0.843. This finding implies that the NB2 OEF system is sensitive to signal density. SCAD active coverage stays high at 0.917 in Scenario 3 and 0.959 in Scenario 4 but its TPR slips to 0.842 and 0.910 under the denser signal regime. This result indicates the concave penalty's difficulty in recovering all active predictors when models are dense. However, its sFPR stays stable at 0.0090 or below.

Table 14. *Coverage probabilities and feature selection - Scenario 3 ($s = 8$, $\alpha = 0.25$).*

Method	Cov. β active	Cov. β null	TPR	sFPR
Unpenalized OEF	0.908	0.912	-	-
Ridge (L2)	0.293	0.772	-	-
L1_plain (LASSO)	0.847	0.926	0.965	0.0230

Method	Cov. β active	Cov. β null	TPR	sFPR
Adaptive LASSO	0.843	0.864	0.934	0.0460
SCAD	0.917	0.992	0.842	0.0050

At extreme over-dispersion setting $\alpha = 0.75$ (Scenario 4), the Adaptive LASSO TPR declines further to 0.921. At large α the NB2 estimation variance grows. This follows from equation (17) and this in turn weakens the signal reaching the selection step.

Table 15. *Coverage probabilities and feature selection - Scenario 4 ($s = 8$, $\alpha = 0.75$).*

Method	Cov. β active	Cov. β null	TPR	sFPR
Unpenalized OEF	0.910	0.906	-	-
Ridge (L2)	0.576	0.852	-	-
L1_plain (LASSO)	0.872	0.943	0.974	0.0190
Adaptive LASSO	0.845	0.895	0.921	0.0360
SCAD	0.959	0.977	0.910	0.0090

5.2.4. Estimation Accuracy

Tables 16-19 report RMSE and signed bias for the regression coefficients.

On coefficient accuracy, the unpenalized OEF again attains the lowest active RMSE across all NB2 scenarios ranging from 0.095 to 0.127. The increase in active RMSE from low to high over-dispersion regimes is mild, from 0.095 in Scenario 1 to 0.127 in Scenario 2. We attribute this to the log link which places estimation on the log mean scale where the NB2 variance function is well behaved. Ridge carries the largest active RMSE throughout ranging from 0.180 to 0.289. For the selectors, SCAD's active RMSE is moderate ranging from 0.141 to 0.193. The L1-type penalties give the smallest null coefficient RMSE at 0.061-0.096 for the L1_plain and 0.067-0.098 for the Adaptive LASSO. This follows from their aggressive shrinkage of null coefficients. However, the unpenalized OEF applies no shrinkage and so it retains the largest null RMSE share ranging from 0.093 to 0.125.

Table 16. *Bias and RMSE - Scenario 1 ($s = 4$, $\alpha = 0.25$).*

Method	RMSE $\hat{\beta}$ active	Bias $\hat{\beta}$ active	RMSE $\hat{\beta}$ null	Bias $\hat{\beta}$ null
Unpenalized OEF	0.095	+0.000	0.097	+0.001
Ridge (L2)	0.289	−0.082	0.070	+0.008
L1_plain (LASSO)	0.095	−0.019	0.065	+0.003
Adaptive LASSO	0.096	−0.013	0.072	+0.002
SCAD	0.193	−0.039	0.069	+0.004

Table 17. Bias and RMSE - Scenario 2 ($s = 4$, $\alpha = 0.75$).

Method	RMSE $\hat{\beta}$ active	Bias $\hat{\beta}$ active	RMSE $\hat{\beta}$ null	Bias $\hat{\beta}$ null
Unpenalized OEF	0.127	−0.001	0.124	−0.001
Ridge (L2)	0.208	−0.043	0.102	+0.004
L1_plain (LASSO)	0.127	−0.018	0.096	+0.001
Adaptive LASSO	0.127	−0.012	0.095	+0.000
SCAD	0.150	−0.019	0.099	+0.001

Table 18. Bias and RMSE - Scenario 3 ($s = 8$, $\alpha = 0.25$).

Method	RMSE $\hat{\beta}$ active	Bias $\hat{\beta}$ active	RMSE $\hat{\beta}$ null	Bias $\hat{\beta}$ null
Unpenalized OEF	0.095	−0.002	0.093	+0.001
Ridge (L2)	0.233	−0.052	0.069	+0.004
L1_plain (LASSO)	0.099	−0.023	0.061	+0.008
Adaptive LASSO	0.098	−0.016	0.067	+0.003
SCAD	0.141	−0.024	0.066	+0.006

Table 19. Bias and RMSE - Scenario 4 ($s = 8$, $\alpha = 0.75$).

Method	RMSE $\hat{\beta}$ active	Bias $\hat{\beta}$ active	RMSE $\hat{\beta}$ null	Bias $\hat{\beta}$ null
Unpenalized OEF	0.125	−0.002	0.125	+0.001
Ridge (L2)	0.180	−0.031	0.101	+0.010
L1_plain (LASSO)	0.127	−0.022	0.094	+0.007
Adaptive LASSO	0.129	−0.017	0.098	+0.004
SCAD	0.142	−0.015	0.109	+0.004

5.2.5. Post-selection and Post-refit Inference

This sub-section examines the select-then-refit strategy under the NB2 model. The same post-selection and post-refit machinery of Section 5.1.5 applies in this case with the select then refit step using the unpenalized OEF and Godambe CIs on the selected sub-model.

Table 20 compares naive conditional coverage with post-refit coverage across the selectors and scenarios.

Table 20. Naive conditional and post-refit coverage (Scenarios 1-4).

Scenario	Method	Naive active	Naive null	Post-refit active	Post-refit null
1	L1_plain (LASSO)	0.890	0.782	0.878	0.729
	Adaptive LASSO	0.915	0.382	0.860	0.562
	SCAD	0.948	0.975	0.851	0.666

Scenario	Method	Naive active	Naive null	Post-refit active	Post-refit null
2	L1_plain (LASSO)	0.909	0.898	0.877	0.830
	Adaptive LASSO	0.901	0.717	0.869	0.727
	SCAD	0.973	0.962	0.878	0.755
3	L1_plain (LASSO)	0.866	0.775	0.876	0.724
	Adaptive LASSO	0.892	0.479	0.878	0.588
	SCAD	0.971	0.955	0.848	0.636
4	L1_plain (LASSO)	0.895	0.905	0.886	0.819
	Adaptive LASSO	0.900	0.731	0.878	0.712
	SCAD	0.978	0.951	0.890	0.790

Taking naive conditional coverage first, SCAD is consistently above nominal for the active coefficients ranging from 0.948 to 0.978. This follows from the oracle property of the concave penalty [4]. Moreover, its naive null coverage is similarly high ranging from 0.951 to 0.975. This is attributable to its near-perfect specificity (sFPR) which leaves few falsely selected null features to evaluate.

L1_plain naive active coverage ranges from 0.866 to 0.909, which is below the nominal 0.95 level, while its null coverage ranges from 0.775 to 0.905. Adaptive LASSO shows marked asymmetry: active coverage ranges from 0.892 to 0.915, but null coverage falls to 0.382 and 0.479 in Scenarios 1 and 3. This pattern is consistent with documented undercoverage after adaptive LASSO selection [44, 49] and may reflect the interaction between adaptive weights and NB2 score curvature.

Post-refit active coverage also remains below nominal: 0.876-0.886 for L1_plain, 0.860-0.878 for Adaptive LASSO and 0.848-0.890 for SCAD. The same limitation appears in the BB results and reflects selection-stage uncertainty that an unpenalized refit cannot remove [38, 39]. Adaptive LASSO null coverage remains low at 0.562-0.727, while SCAD null coverage ranges from 0.636 to 0.790. The NB2 dispersion parameter is itself difficult to estimate under strong overdispersion [47], and interval calibration is sensitive to finite-sample dispersion estimation [48]. Sampling variation in the sandwich covariance estimator provides a further source of undercoverage [50]. The refit should therefore be interpreted as a bias-reduction step, not as a generally valid post-selection procedure.

5.3. Discussion

The two simulation studies evaluated penalized OEFs for feature selection under Beta-Binomial (BB) logistic regression in Section 5.1 and Negative-Binomial (NB2) regression in Section 5.2. Both studies place the unpenalized OEF at the same point as the Godambe efficient benchmark and both show the L1-type penalties recovering the active set almost completely in the sparse regime. What separates them is the ordering of the penalties. Ridge and SCAD trade places between the two models and the reversal is systematic rather than incidental. It follows from the bounded response range and logit link of the BB model set against the unbounded count support and exponential log link of NB2.

In both models the unpenalized OEF is the inferential benchmark. Its active-coefficient coverage ranges from 0.876 to 0.899 for BB and from 0.908 to 0.918 for NB2. These values are consistently below 0.95, especially for BB, so the simulations support relative stability rather than nominal calibration. Finite-sample variation in sandwich covariance estimation can contribute to this undercoverage [50]. The penalized estimators serve a different purpose: they trade some estimation and inferential performance for sparse selection. Ridge, L1_plain, Adaptive LASSO and SCAD therefore should be compared on the joint balance of RMSE, coverage, TPR, sFPR and model sparsity rather than on a claim of uniform superiority.

Selection behaves alike in the sparse regime and degrades in a model-dependent way as the signal becomes denser. L1_plain and Adaptive LASSO achieve near-perfect sensitivity in the sparse regime with TPR up to 1.000 in both models. The Adaptive LASSO provides the best sFPR here and this is consistent with the oracle property and with the sparser support that adaptive weighting affords [3, 19]. Moving to the moderately dense regime $s/p = 0.2$ selection in both models degrades but the mechanism differs between them. Under the BB model the sFPR rises most steeply for SCAD from 0.0620 to 0.1170 while under NB2 it is the sensitivity that drops, with L1_plain TPR ranging 0.965-0.974 and SCAD TPR ranging 0.842-0.910. A higher $s/p = 0.4$ inflates the estimation variance which in turn weakens the signal that reaches the selection step. In both the binary and count cases the decline in selection performance comes from null and weak signals becoming harder to separate rather than from a failure of the estimation procedure [19].

The SCAD estimator's reliability is governed by the smoothness of the score geometry, which again separates the two models. Under the bounded BB support SCAD becomes unstable at high over-dispersion with sFPRs rising to 0.1170 at $\rho = 0.65$ and post-refit active coverage declining to 0.725-0.815. Under the smoother NB2 geometry SCAD is by contrast the numerically stable estimator across all scenarios, its only limitation being reduced sensitivity at the dense setting ρ where TPR ranges 0.842-0.910. This is a loss of power under dense signals rather than the convergence instability seen in the BB model. SCAD's instability under the BB setting and its stability under NB2 are both consistent with the multiple local solutions admitted by the nonconvex penalty [4, 32], a multiplicity worsened by the bounded BB response range but suppressed by the monotone NB2 score function.

The select-then-refit strategy reduces shrinkage bias for some retained predictors, but it does not deliver valid post-selection inference in these simulations. Active coverage remains below 0.95 in both model families, and null coverage is particularly weak for Adaptive LASSO and SCAD under NB2. The refit corrects the fitted coefficients conditional on the chosen support. It does not account for the randomness of that support [38, 39].

This strategy is however less complete for null predictors under the NB2 model where Adaptive LASSO null coverage remains low spanning 0.562-0.727 and SCAD null coverage fails to reach nominal in any of the simulation scenarios. Our simulation results therefore indicate that post-refit Godambe intervals do not fully compensate for the sub-model misspecification arising from incorrect null inclusion. Refitting restores coverage for the active predictors but not for the null predictors, which is consistent with the post-selection-inference literature where naive Wald intervals from a selected model are generally invalid and where refitting or conditioning on the selection event is required to restore coverage [38, 39, 45, 46].

6. Conclusion

This study developed and evaluated a penalized OEF framework for feature selection in over-dispersed Beta-Binomial and Negative-Binomial regression. The framework combines joint mean-dispersion estimating equations with Ridge, LASSO, Adaptive LASSO and SCAD penalties on the regression component and tunes them using cross-validated estimating-function loss. The simulations show complementary rather than uniformly superior performance: the unpenalized OEF generally provides the strongest coefficient coverage and RMSE benchmark, while penalization supplies sparse selection with penalty- and response-specific trade-offs.

Penalty performance is not intrinsic to the penalty but is determined by the response-penalty geometry of the model to which the penalty is applied. Bounded against unbounded response and the logit against the log link together reverse the relative performance of the Ridge and SCAD penalties for both coefficient coverage and selection.

In practice the best configuration depends on the data setting and on whether sensitivity or specificity is the priority. For binary and proportion data under the Beta-Binomial model, Adaptive LASSO gives the most dependable feature selection overall by combining the highest sensitivity with TPR up to 0.999 and the tightest control of falsely significant nulls at sFPR no greater than 0.0035. L1_plain comes a close second on sensitivity but admits more falsely significant nulls as the signal grows denser with sFPR up to 0.0172, while SCAD though highly sensitive admits the steepest sFPR from 0.0620 to 0.1170. For count data under the Negative-Binomial model the choice is a sensitivity-specificity trade-off. L1_plain gives the most balanced selection with the highest mean sensitivity at TPR 0.980 and a controlled sFPR. Where false positives are especially costly SCAD is preferable since it gives the best specificity of any method at sFPR no greater than 0.0090, at the cost of reduced sensitivity under dense signals where TPR falls to 0.842. Adaptive LASSO is less suited to this setting since its significance-filtered false positive rate climbs to 0.051 without a matching gain in sensitivity. In both models the unpenalized OEF remains the coverage benchmark against which the penalized variants are judged.

For applied use, penalized OEF may be considered when sparse selection and joint mean-dispersion estimation are both required. Any select-then-refit analysis should be reported as conditional and exploratory unless a procedure that accounts explicitly for selection uncertainty is used. The present results do not justify describing the reported Godambe intervals as dependable post-selection inference.

Several limitations qualify these conclusions. The simulations cover a moderate-dimensional design with 20 candidate predictors, AR (1) correlation and two overdispersion levels per response family. They do not examine stronger dependence, model misspecification or high-dimensional feature spaces. No formal sensitivity analysis was available for the standard-error multiplier, stability threshold or SCAD guard, although the tuning rules were fixed before comparison within each response family. The quasi-binomial and quasi-Poisson models provide unpenalized moment-based benchmarks; matched penalized GLM and GEE comparisons were not included. GEE is not directly matched to the independent aggregate responses studied here, while penalized GLM comparisons in high-dimensional settings are reserved for separate work. Finally, simple refitting does not account for selection uncertainty and does not restore nominal coverage. Within these limits, penalized OEF provides a competitive route to sparse selection with explicit joint mean-dispersion estimation, not a uniformly better alternative to unpenalized OEF or conventional models.

Abbreviations

GLM Generalized Linear Model

OEF	Optimal Estimating Function
EF	Estimating Function
BB	Beta-Binomial
NB2	Negative Binomial
LASSO	Least Absolute Shrinkage and Selection Operator
SCAD	Smoothly Clipped Absolute Deviation
CV	Cross-Validation
SE	Standard Error
TPR	True Positive Rate
sFPR	Significance-filtered False Positive Rate
RMSE	Root Mean Squared Error
MoM	Method of Moments
CI	Confidence Interval
DGP	Data-Generating Process
AR (1)	First-order Autoregressive

Author Contributions

Timothy Mutunga: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing

Ali Salim: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing

Pius Kihara: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing

Justine Okenye: Conceptualization, Methodology, Validation

Data Availability Statement

The data supporting the outcome of this research work has been reported in this manuscript. The simulation code used to generate the datasets is available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Hoerl, A. E., Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67 1970. https://doi.org/10.1080/00401706.1970.10488634

[2] Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288 1996. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[3] Zou, H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476), 1418-1429 2006. https://doi.org/10.1198/016214506000000735

[4] Fan, J., Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456), 1348-1360 2001. https://doi.org/10.1198/016214501753382273

[5] McCullagh, P., Nelder, J. A. Generalized linear models (2nd ed.). Chapman & Hall 1989.

[6] Dean, C. B., Lundy, E. R. Overdispersion. In Wiley StatsRef: Statistics reference online. John Wiley & Sons 2016. https://doi.org/10.1002/9781118445112.stat06788.pub2

[7] Wedderburn, R. W. M. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika, 61(3), 439-447 1974. https://doi.org/10.1093/biomet/61.3.439

[8] Godambe, V. P., Thompson, M. E. An extension of quasi-likelihood estimation. Journal of Statistical Planning and Inference, 22(2), 137-152 1989. https://doi.org/10.1016/0378-3758(89)90106-7

[9] Desmond, A. F. Optimal estimating functions, quasi-likelihood and statistical modelling. Journal of Statistical Planning and Inference, 60(1), 77-104 1997. https://doi.org/10.1016/S0378-3758(96)00120-0

[10] Johnson, B. A., Lin, D. Y., Zeng, D. Penalized estimating functions and variable selection in semiparametric regression models. Journal of the American Statistical Association, 103(482), 672-680 2008. https://doi.org/10.1198/016214508000000184

[11] Godambe, V. P. The foundations of finite sample estimation in stochastic processes. Biometrika, 72(2), 419-428 1985. https://doi.org/10.1093/biomet/72.2.419

[12] Najera-Zuloaga, J., Lee, D.-J., Arostegui, I. Comparison of beta-binomial regression model approaches to analyze health-related quality of life data. Statistical Methods in Medical Research, 27(10), 2989-3009 2018. https://doi.org/10.1177/0962280217690413

[13] Lehmann, B. D., Archer, K. J. Penalized negative binomial models for modeling an overdispersed count outcome with a high-dimensional predictor space: Application predicting micronuclei frequency. PLOS ONE, 14(1), Article e0209923 2019. https://doi.org/10.1371/journal.pone.0209923

[14] Lindén, A., Mäntyniemi, S. Using the negative binomial distribution to model overdispersion in ecological count data. Ecology, 92(7), 1414-1421 2011. https://doi.org/10.1890/10-1831.1

[15] Cameron, A. C., Trivedi, P. K. Regression analysis of count data (2nd ed.). Cambridge University Press 2013. https://doi.org/10.1017/CBO9781139013567

[16] Suryadi, F., Jonathan, S., Jonatan, K., Ohyver, M. Handling overdispersion in Poisson regression using negative binomial regression for poverty case in West Java. Procedia Computer Science, 216, 517-523 2023. https://doi.org/10.1016/j.procs.2022.12.164

[17] Heyde, C. C. Quasi-likelihood and its application: A general approach to optimal parameter estimation. New York, NY: Springer 1997. https://doi.org/10.1007/b98823

[18] Hastie, T., Tibshirani, R., Friedman, J. The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer 2009. https://doi.org/10.1007/978-0-387-84858-7

[19] Bühlmann, P., van de Geer, S. Statistics for high-dimensional data: Methods, theory and applications. Springer 2011. https://doi.org/10.1007/978-3-642-20192-9

[20] Claeskens, G., Hjort, N. L. Model selection and model averaging. Cambridge University Press 2008.

[21] Stone, M. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological), 36(2), 111-147 1974. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

[22] Komiyama, J., Maehara, T. Cross validation based model selection via generalized method of moments. arXiv 2018. https://doi.org/10.48550/arXiv.1807.06993

[23] Wang, F., Mukherjee, S., Richardson, S., Hill, S. M. High-dimensional regression in practice: An empirical study of finite-sample prediction, variable selection and ranking. Statistics and Computing, 30(3), 697-719 2020. https://doi.org/10.1007/s11222-019-09914-9

[24] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. Classification and regression trees. Wadsworth 1984.

[25] Chen, Y., Yang, Y. The one standard error rule for model selection: Does it work?. Stats, 4(4), 868-892 2021. https://doi.org/10.3390/stats4040051

[26] Yates, L. A., Aandahl, Z., Richards, S. A., Brook, B. W. Cross validation for model selection: A review with examples from ecology. Ecological Monographs, 92(1), Article e1557 2022. https://doi.org/10.1002/ecm.1557

[27] Donoho, D. L., Johnstone, I. M. Ideal spatial adaptation by wavelet shrinkage. Biometrika, 81(3), 425-455 1994. https://doi.org/10.1093/biomet/81.3.425

[28] Sutradhar, B. C., Das, K. On joint estimation of regression and overdispersion parameters in generalized linear models for longitudinal data. Journal of Multivariate Analysis, 58(1), 90-108 1996. https://doi.org/10.1006/jmva.1996.0041

[29] Paul, S. R., Islam, A. S. Joint estimation of the mean and dispersion parameters in the analysis of proportions. Canadian Journal of Statistics, 26(1), 83-94 1998. https://doi.org/10.2307/3315675

[30] Harrison, X. A. Using observation-level random effects to model overdispersion in count data in ecology and evolution. PeerJ, 2, e616 2014. https://doi.org/10.7717/peerj.616

[31] Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics, 38(2), 894-942 2010. https://doi.org/10.1214/09-AOS729

[32] Breheny, P., Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. The Annals of Applied Statistics, 5(1), 232-253 2011. https://doi.org/10.1214/10-AOAS388

[33] Meinshausen, N., Bühlmann, P. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4), 417-473 2010. https://doi.org/10.1111/j.1467-9868.2010.00740.x

[34] Ternès, N., Rotolo, F., Michiels, S. Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models. Statistics in Medicine, 35(15), 2561-2573 2016. https://doi.org/10.1002/sim.6927

[35] Godambe, V. P. Estimating functions. Oxford University Press 1991.

[36] Riley, R. D., Snell, K. I. E., Martin, G. P., Whittle, R., Archer, L., Sperrin, M., Collins, G. S. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. Journal of Clinical Epidemiology, 132, 88-96 2020. https://doi.org/10.1016/j.jclinepi.2020.12.005

[37] van de Wiel, M. A., Leday, G. G. R., Heymans, M. W., van Zwet, E. W., Zwinderman, A. H., Hoogland, J. Alternatives to default shrinkage methods can improve prediction accuracy, calibration, and coverage: A methods comparison study. Statistical Methods in Medical Research, 34(7), 1342-1355 2025. https://doi.org/10.1177/09622802251338440

[38] Leeb, H., Pötscher, B. M. Can one estimate the conditional distribution of post-model-selection estimators?. The Annals of Statistics, 34(5), 2554-2591 2006. https://doi.org/10.1214/009053606000000821

[39] Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L. Valid post-selection inference. The Annals of Statistics, 41(2), 802-837 2013. https://doi.org/10.1214/12-AOS1077

[40] Tibshirani, R. J., Taylor, J., Lockhart, R., Tibshirani, R. Exact post-selection inference for sequential regression procedures. Journal of the American Statistical Association, 111(514), 600-620 2016. https://doi.org/10.1080/01621459.2015.1108848

[41] Efron, B., Hastie, T. Computer age statistical inference: Algorithms, evidence, and data science. Cambridge University Press 2016. https://doi.org/10.1017/CBO9781316576533

[42] Hilbe, J. M. Negative binomial regression (2nd ed.). Cambridge University Press 2011. https://doi.org/10.1017/CBO9780511973420

[43] Kenne Pagui, E. C., Salvan, A., Sartori, N. Improved estimation in negative binomial regression. Statistics in Medicine, 41(13), 2403-2416 2022. https://doi.org/10.1002/sim.9361

[44] Kammer, M., Dunkler, D., Michiels, S., Heinze, G. Evaluating methods for Lasso selective inference in biomedical research: A comparative simulation study. BMC Medical Research Methodology, 20, 108 2020. https://doi.org/10.1186/s12874-020-00998-w

[45] Taylor, J., Tibshirani, R. Post-selection inference for $\ell 1$-penalized likelihood models. The Canadian Journal of Statistics, 46(1), 41-61 2018. https://doi.org/10.1002/cjs.11313

[46] Zhang, D., Khalili, A., Asgharian, M. Post-model-selection inference in linear regression models: An integrated review. Statistics Surveys, 16, 86-136 2022. https://doi.org/10.1214/22-SS135

[47] Lloyd-Smith, J. O. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PLoS ONE, 2(2), e180 2007. https://doi.org/10.1371/journal.pone.0000180

[48] Saha, K. K., Paul, S. R. Interval estimation of the negative binomial dispersion parameter. Journal of Statistical Computation and Simulation, 81(9), 1109-1122 2011. https://doi.org/10.1080/00949651003724555

[49] van de Geer, S., Bühlmann, P., Zhou, S. The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). Electronic Journal of Statistics, 5, 688-749 2011. https://doi.org/10.1214/11-EJS624

[50] Kauermann, G., Carroll, R. J. A note on the efficiency of sandwich covariance matrix estimation. Journal of the American Statistical Association, 96(456), 1387-1396 2001. https://doi.org/10.1198/016214501753382309