

Dispersion Is Not a Nuisance: Penalized Joint Optimal Estimating Functions for Over-Dispersed Regression

Timothy Mutunga^{1*}, Ali Salim¹, Pius Kihara², Justine Okenye¹

¹Department of Mathematics, Egerton University, Egerton, Kenya

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

Email: *mutungatn@gmail.com, asislam54@yahoo.com, piuskihara2@gmail.com, justinoke69@gmail.com

How to cite this paper: Mutunga, T., Salim, A., Kihara, P. and Okenye, J. (2026) Dispersion Is Not a Nuisance: Penalized Joint Optimal Estimating Functions for Over-Dispersed Regression. *Open Journal of Statistics*, 16, 460-493.
<https://doi.org/10.4236/ojs.2026.165022>

Received: August 13, 2026

Accepted: October 20, 2026

Published: October 23, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Penalized regression is often assessed on how well it recovers the mean model. On the other hand, the dispersion parameter is treated as a nuisance quantity estimated from the residuals that the penalty has already altered. This paper asks what penalization does to dispersion inference itself. Working within a joint Optimal Estimating Function (OEF) system that estimates regression and dispersion parameters simultaneously from orthogonalized moment conditions, we apply Ridge, LASSO, Adaptive LASSO and SCAD penalties to the regression equation only leaving the dispersion equation unpenalized. Consequently, we examine what this does to the intra-cluster correlation of the Beta-Binomial model and the heterogeneity parameter of the Negative-Binomial model. Across 500 Monte Carlo replications per scenario in a factorial design crossing over-dispersion level with signal density, penalization disturbs dispersion estimation substantially more than it disturbs regression coefficient estimation. This mechanism is the mean-dispersion coupling of the joint OEF system. Shrinkage of coefficients alters the residual structure from which the dispersion parameter is estimated and since the dispersion equation is unpenalized, the estimate absorbs the mean model error. The disturbance takes a different form in each model. Under the Beta-Binomial, Ridge over-shrinkage drives a compensation cycle that inflates the dispersion estimate and produces multimodal sampling distributions. Under the Negative Binomial, the exponential mean propagates coefficient error quadratically into the variance function and the L1-type penalties underestimate dispersion severely. A log-scale interval raises Negative Binomial dispersion coverage by up to 55 percentage points while the analogous logit-scale interval does almost nothing for the Beta-Binomial, where $\hat{\rho}$ is bounded and already close to symmetric. We also show that the Pearson moment estimator attains better nominal disper-

sion coverage than the Godambe-optimal OEF in the Beta-Binomial design and examine why an asymptotically optimal construction can undercover at $n = 100$ sample size.

Keywords

Penalized Regression, Over-Dispersion, Dispersion Inference, Mean-Dispersion Coupling, Godambe Information, Sandwich Covariance, Variance-Stabilizing Transformation

1. Introduction

Over-dispersion is usually treated as a nuisance. Binary and count responses frequently show variance in excess of the Binomial and Poisson specifications. A scalar dispersion parameter accommodates the excess variation and inference then proceeds on the regression coefficients. In applied work, the dispersion parameter is often the quantity of interest. The heterogeneity parameter of a Negative Binomial fit distinguishes superspreading from uniform transmission [1]. The intra-cluster correlation of a Beta-Binomial fit determines the sample size required for a cluster-randomized trial. In ecology, the same parameter measures spatial aggregation [2] [3].

The dispersion parameter affects inference even when it is a nuisance, as it scales every standard error in the model. Under the Beta-Binomial, the variance carries the factor $1 + (m-1)\rho$ and therefore an error in $\hat{\rho}$ propagates into every interval for every regression coefficient. The distortion is not visible in the fitted mean. A procedure that recovers the mean structure accurately and the dispersion poorly yields miscalibrated inference.

The standard workflows in existing literature conceal this. Penalized likelihood and quasi-likelihood procedures fit the mean model first and estimate dispersion afterwards from the residuals of that fit [4] [5]. Shrinkage alters those residuals. Notably, the dispersion estimate is computed after the mean model is fixed and so the dependence between the two is not exposed at any stage of the procedure.

Optimal Estimating Functions have no such sequential structure. Godambe [6] and Godambe and Thompson [7] established optimality within a class of unbiased estimating functions constructed from moment conditions without a full distributional specification. Regression and dispersion parameters are estimated jointly from a system of orthogonalized estimating equations that uses third and fourth moment information in addition to the first two [8]. Penalizing the regression equation while leaving the dispersion equation unpenalized therefore isolates the effect of shrinkage on dispersion inference.

This paper quantifies that effect. For a penalized joint estimating-function system, we report the bias of the dispersion estimate, the spread of its sampling distribution, and the coverage of intervals constructed from the Godambe sandwich. The regression coefficients are not the object of study in this paper.

Two models are used because the effect depends on the structure of the response rather than on the penalty. Beta-Binomial logistic regression has a bounded support, a logit link and a dispersion parameter on the unit interval. Negative Binomial regression has unbounded support, a log link, a dispersion parameter on $\mathbb{R}^+$ and variance quadratic in the mean. Four penalties are applied: Ridge which shrinks all coefficients without setting any to zero, LASSO which shrinks and selects, Adaptive LASSO which weights the LASSO penalty by an initial estimate and SCAD which shrinks small coefficients and leaves large ones unpenalized. Applying these four to both models separates the properties of the penalty from the properties of the response structure.

Section 2 specifies the moment structure of the two models. Section 3 sets out the joint estimating-function system, its penalized form and the Godambe sandwich covariance. Section 4 describes hyper-parameter selection. Section 5 reports the simulation results for both models. Section 6 concludes the paper.

2. Moment Specifications

The estimating functions adopted in this paper are constructed from moment conditions rather than from a likelihood. The two models therefore get admitted into the OEF system only through their first four central moments and their parameterization of over-dispersion which this section states. The Beta-Binomial arises as a Binomial mixture with Beta-distributed success probability and the Negative Binomial as a Poisson mixture with Gamma-distributed rate.

Throughout, we assume independent observations $\{Y_i, x_i\}_{i=1}^n$, where Y_i is a scalar response variable, $x_i = (1, x_{i1}, \dots, x_{ip})^T$ is the i^{th} row of full rank design/feature matrix $X \in \mathbb{R}^{n \times (p+1)}$ 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.

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

The marginal mean is:

$$\mu_i = E(Y_i) = mp_i \quad (1)$$

Najera-Zuloaga *et al.* [9] assumed that the probability parameter p_i of the beta-binomial is linked to the covariates through the logistic transformation:

$$p_i = \text{logistic}(x_i^T \beta) = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} \quad (2)$$

By the relationship in Equation (2), we define the linear predictor as:

$$\eta_i = \log\left(\frac{p_i}{1-p_i}\right) = x_i^T \beta \quad (3)$$

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 [4] [5] [9].

The variance function can be written as:

$$\text{Var}(Y_i) = mp_i(1-p_i) \left[1 + (m-1) \frac{\phi}{1+\phi} \right] \quad (4)$$

Substituting $p_i = \frac{\mu_i}{m}$ in (4) we have:

$$V(\mu_i; \rho) = \frac{\mu_i}{m} (m - \mu_i) [1 + (m-1)\rho] \quad (5)$$

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:

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

The fourth central moment is defined as:

$$\begin{aligned} \kappa_{4i} = & \frac{\rho^2}{(1+\rho)(1+2\rho)} (\text{Var}(\mu_i; \rho))^{-1} \left[(1-\rho) \left(\frac{1-2\rho(1-3m)+6m^2\rho^2}{\rho^2} \right) \right. \\ & \left. + \frac{3\mu_i(m-\mu_i)}{(m\rho)^2} \left\{ (m-2)(1-\rho)^2 - m\rho(1-\rho)(6-m) - 6m^2\rho^2 \right\} \right] \quad (7) \end{aligned}$$

2.2. Negative Binomial Regression

The Negative Binomial is constructed in the same manner. Counts are Poisson conditional on a Gamma-distributed rate. Marginalising over the rate yields a distribution whose moments are stated below in the heterogeneity parameterisation of Lindén and Mäntyniemi [2].

The moments of the Negative Binomial distribution are as follows.

The mean function is specified using a log-link:

$$\log(\mu_i) = x_i^T \beta \quad (8)$$

so that

$$E(Y_i) = \mu_i = \exp(x_i^T \beta) \quad (9)$$

The linear predictor is therefore expressed as:

$$\eta_i = x_i^T \beta \quad (10)$$

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 [10].

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

$$\text{Var}(\mu_i; \alpha) = \mu_i + \alpha \mu_i^2 \quad (11)$$

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 [11] [12]. This Poisson-Gamma mixture construction and its variance structure are standard in the literature [10]-[12].

The higher-order moments for the Negative-binomial (μ_i, α) 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:

$$\kappa_{3i} = \frac{2\alpha\mu_i + 1}{\sqrt{\text{Var}(\mu_i; \alpha)}} \quad (12)$$

The fourth central moment is defined as:

$$\kappa_{4i} = \frac{1 + 6\alpha\mu_i(1 + \alpha\mu_i)}{\text{Var}(\mu_i; \alpha)} \quad (13)$$

The two models share this construction but differ in the features that matter for penalized estimation. Both attach a single scalar to the excess variation, both nest a one-parameter model at the boundary of the dispersion space, and both supply the four moments required in Section 3. They differ in support, in link, and in how the variance responds to a change in the mean. **Table 1** sets these features side by side. The rows on which the models differ determine the results of Section 4.

Table 1. Structural contrast between the two over-dispersed models.

Feature	Beta-Binomial	Negative Binomial
Response support	Bounded, 0 to m	Unbounded, $\mathbb{R}^+$
Link	logit	log
Mean	$\mu_i = mp_i$	$\mu_i = \exp(x_i^T \beta)$
Variance	$\frac{\mu_i}{m}(m - \mu_i)[1 + (m - 1)\rho]$	$\mu_i + \alpha \mu_i^2$
Variance inflation	Multiplicative, free of the mean	Quadratic in the mean

Continued

Dispersion parameter	$\rho \in (0,1)$, intra-cluster correlation	$\alpha > 0$, heterogeneity
Nested model	Binomial at $\rho = 0$	Poisson at $\alpha = 0$
Effect of shrinking β	Additive on the linear predictor	Multiplicative on the mean

The third and fourth central moments are derived in this section because the Estimating Functions (EFs 3.0) of Section 3 require them. The first two moments fix the mean and the variance and the orthogonalization step draws on skewness and kurtosis to recover information those two leave behind.

3. Estimating Functions (EFs)

This section states the EFs system used in the simulations of Section 4. The construction applies the orthogonalization of Godambe and Thompson [7] to the moment conditions of Section 2. Resultantly, two features are relevant. The regression and dispersion estimating equations are asymptotically uncorrelated by construction and so the dispersion equation is not a residual calculation appended to the mean fit. The penalty is applied to the regression equation alone which leaves the dispersion equation unbiased and Godambe-optimal. Any disturbance to dispersion inference under penalization therefore arises through the coupling between the two equations rather than through the penalty acting on the dispersion parameter.

3.1. Optimal Estimating Functions (OEFs)

Let $y_i, i = 1, \dots, n$ be independent observations from unspecified distribution with mean equation $E(y_i) = \mu_i$.

Where, $\mu_i = \mu_i(\beta)$, and variance equation be defined as:

$$\text{Var}(y_i) = v(\mu_i; \phi) \quad (14)$$

where ϕ is the dispersion parameter.

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

$$\left. \begin{aligned} \pi_{1i} &= y_i - \mu_i \\ \pi_{2i} &= (y_i - \mu_i)^2 - v(\mu_i; \phi) \end{aligned} \right\} \quad (15)$$

It is easy to show both are unbiased, $E(\pi_{ji}) = 0$, for $j = 1, 2$.

Utilizing Theorem 2.2 of Godambe and Thompson [7], elementary orthogonal EFs in (16) for $i = 1, 2, 3, \dots, n$ are constructed using the Gram-Schmidt orthogonalization procedure based on the set of unbiased EFs in (15). This orthogonalization removes redundancy between moment conditions and minimizes asymptotic variance.

$$\left. \begin{aligned} h_{1i} &= \pi_{1i} \\ h_{2i} &= \pi_{2i}^* = (y_i - \mu_i)^2 - v(\mu_i; \phi) - (y_i - \mu_i)v(\mu_i; \phi)^{\frac{1}{2}} \kappa_{3i} \end{aligned} \right\} \quad (16)$$

where, π_{2i}^* is the orthogonalized π_{2i} and κ_{3i} is the skewness coefficient defined in (17).

$$\kappa_{3i} = E \left[\frac{(y_i - \mu_i)^3}{v(\mu_i; \phi)^{\frac{3}{2}}} \right] \quad (17)$$

Using the elementary EFs in (16), Theorem 2.2 of Godambe and Thompson [7] is applied to construct a linear combination of elementary EFs in (18)

$$\left. \begin{aligned} g_\beta &= \sum_{i=1}^n a_1^1 h_{1i} + \sum_{i=1}^n a_2^1 h_{2i} \\ g_\phi &= \sum_{i=1}^n a_1^2 h_{1i} + \sum_{i=1}^n a_2^2 h_{2i} \end{aligned} \right\} \quad (18)$$

The optimal coefficients follow the same projection in every case. For elementary estimating function h_i and parameter θ_j ,

$$a_i^{*j} = \frac{E \left[\frac{\partial h_i}{\partial \theta_j} \right]}{E[h_i^2]}, \quad i = 1, 2, \quad \theta_1 = \beta, \quad \theta_2 = \phi \quad (19)$$

with expectation evaluated at the moments of Section 2. The coefficient linking the dispersion equation to the regression parameter vanishes, $a_2^{*1} = 0$, and this orthogonality is what separates the two equations. Substituting these coefficients into the linear combination in (18) gives the joint OEFs for the regression and dispersion parameters in (20) and (21)

$$g_\beta^* = - \sum_{i=1}^n \frac{\frac{\partial \mu_i}{\partial \beta}}{v(\mu_i; \phi)} (y_i - \mu_i) + \sum_{i=1}^n \frac{v(\mu_i; \phi)^{\frac{1}{2}} \kappa_3 \frac{\partial \mu_i}{\partial \beta} - \frac{\partial}{\partial \beta} v(\mu_i; \phi)}{v(\mu_i; \phi)^2 \{ \kappa_{4i} + 2 - \kappa_{3i}^2 \}} \times \left((y_i - \mu_i)^2 - v(\mu_i; \phi) - (y_i - \mu_i) v(\mu_i; \phi)^{\frac{1}{2}} \kappa_{3i} \right) \quad (20)$$

$$g_\phi^* = - \sum_{i=1}^n \frac{\frac{\partial}{\partial \phi} v(\mu_i; \phi)}{v(\mu_i; \phi)^2 \{ \kappa_{4i} + 2 - \kappa_{3i}^2 \}} \left((y_i - \mu_i)^2 - v(\mu_i; \phi) - (y_i - \mu_i) v(\mu_i; \phi)^{\frac{1}{2}} \kappa_{3i} \right) \quad (21)$$

The regression equation in (20) and the dispersion equation in (21) are coupled through the fitted mean since the dispersion parameter is identified from residuals taken about the fitted mean rather than the true mean. The penalty is applied to (20) alone. This coupling accounts for most of the results in Section 4.

Among all unbiased EFs constructed from the moment conditions of the over-dispersed models, the OEFs attain the maximum Godambe information [6] [7]. This is equivalent to minimizing the asymptotic variance of the joint regression-dispersion estimator $\theta = (\beta, \phi)$.

Let $S(\theta)$ denote the sensitivity matrix which is the expected derivative of the estimating function with respect to the parameter vector.

Let $V(\theta)$ denote the variability matrix which is the variance of the estimating

function.

The asymptotic covariance of the estimator obtained by solving the joint OEF system is the Godambe sandwich covariance in (22).

$$\text{Cov}(\hat{\boldsymbol{\theta}}) \approx G(\boldsymbol{\theta})^{-1} = S(\boldsymbol{\theta})^{-1} V(\boldsymbol{\theta}) (S(\boldsymbol{\theta})^{-1})^T \quad (22)$$

The dispersion block of (22) supplies the standard error of the dispersion estimate.

3.2. Penalized Optimal Estimating Function

In this study, Johnson *et al.* [13] general penalized estimating functions framework for the parameter vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ based on a random sample of size n is adopted to construct penalized OEFs.

The penalized version of (20) is given as:

$$\mathbf{g}_{pen}^*(\boldsymbol{\beta}) = \mathbf{g}_{\beta}^* - nq_{\lambda}(\boldsymbol{\beta}) \text{sgn}(\boldsymbol{\beta}) \quad (23)$$

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

Four penalty functions are considered.

Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ 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 [14] imposes quadratic shrinkage on all coefficients.

$$p_{\lambda}^{Ridge}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^p \beta_j^2 \quad (24)$$

The Lasso penalty (ℓ_1 -penalty) by Tibshirani [15] induces sparsity through an absolute-value penalty in (25).

$$p_{\lambda}^{Lasso}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^p |\beta_j| \quad (25)$$

The ℓ_1 -penalty performs simultaneous feature shrinkage and selection.

The Adaptive Lasso [16] introduces coefficient-specific weights and can achieve oracle properties under regularity conditions.

$$p_{\lambda}^{Adapt-lasso}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^p \omega_j |\beta_j|, \quad \omega_j = \frac{1}{|\tilde{\beta}_j|^{\gamma}} \quad (26)$$

where, $\tilde{\beta}_j$ is an initial consistent estimator such as the unpenalized OEF utilized in this study and $\gamma > 0$ controls the degree of adaptivity. The Adaptive Lasso is a non-convex penalty that provides stronger shrinkage for weak signals and reduced bias for large coefficients.

The Smoothly Clipped Absolute Deviation (SCAD) penalty [17] reduces bias for large coefficients while retaining sparsity.

The SCAD penalty is obtained by integration:

$$P_{\lambda}^{scad}(\boldsymbol{\beta}) = \begin{cases} \lambda |\beta_j|, & |\beta_j| \leq \lambda \\ -\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}, & |\beta_j| \in (\lambda, a\lambda) \\ \frac{(a+1)\lambda^2}{2}, & |\beta_j| \geq a\lambda \end{cases} \quad (27)$$

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

$$p'(\boldsymbol{\beta}) = \lambda \left\{ I(\boldsymbol{\beta} < \lambda) + \frac{(a\lambda - \boldsymbol{\beta})_+}{(a-1)\lambda} I(\boldsymbol{\beta} \geq \lambda) \right\} \quad (28)$$

where, $a > 2$ is the shape parameter (commonly $a = 3.7$) Fan and Li [17] and $(x)_+ = \max(x, 0)$.

3.3. Hyper-Parameter Selection and Stability

Penalized OEFs do not derive from a likelihood and so deviance-based criteria and the classical information criteria are unavailable [5] [18]. The tuning parameter is instead chosen by K -fold cross-validation on the estimating functions themselves. For each candidate value we solve the training folds while penalizing the regression equation and leaving the dispersion equation unpenalized. Further, we score the resulting estimate by the squared norm of the unpenalized joint OEF evaluated on the held-out fold. Moment-based sample splitting of this kind is established in the method-of-moments literature [19] [20].

The cross-validated OEF loss is defined as:

$$L_{OEF}(\lambda) = \sum_{k=1}^K \left\| \begin{pmatrix} \mathbf{g}_{\beta,k}^* \left(\hat{\boldsymbol{\beta}}_{\lambda}^{(-k)}, \hat{\boldsymbol{\phi}}_{\lambda}^{(-k)} \right) \\ \mathbf{g}_{\phi,k}^* \left(\hat{\boldsymbol{\beta}}_{\lambda}^{(-k)}, \hat{\boldsymbol{\phi}}_{\lambda}^{(-k)} \right) \end{pmatrix} \right\|^2 \quad (29)$$

where, K is the number of folds, $\mathbf{g}_{\beta,k}^*$ and $\mathbf{g}_{\phi,k}^*$ denote the jointly optimal estimating functions for regression and dispersion parameters, respectively, as presented in (20) and (21) and evaluated using only observations in fold k of size n_k . The function $q_{\lambda}(\cdot)$ is the derivative of a penalty function $p_{\lambda}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, $\text{sgn}(\boldsymbol{\beta})$ is the component-wise sign function and $\|\cdot\|$ is the Euclidean norm.

The third and fourth order quantities entering (29) are analytic functions of the model parameters. They come from the moment expressions in (6) and (7) for the Beta-Binomial and from (12) and (13) for the Negative Binomial and are evaluated at the estimate obtained from the training folds. No moment of any order is estimated from the held-out observations. The held-out fold contributes an average of as many analytic terms as it holds observations. The size of the fold therefore governs the Monte Carlo noise of that average and not the estimation of a fourth order quantity. The criterion then averages over the K folds and every observation enters exactly once. The effective sample behind the criterion is rather than n divided by K . Small folds therefore leave the criterion intact. However, the variability matrix in (22) is built differently. It substitutes residuals about the fitted

mean into an empirical average and Section 4.1.4 reports how that construction behaves at $n = 100$.

Cross-validation curves for penalized estimators are often flat near the minimum under over-dispersion, so the raw minimiser gives unstable models [21]. The tuning parameter is therefore taken from a k -standard-error (k -SE) band around the minimum:

$$CV(\lambda) \leq CV(\lambda_{\min}) + k \cdot SE(\lambda_{\min}) \quad (30)$$

Throughout this paper we use $k = 1$. This gives the standard 1-standard-error (1-SE) rule popularized in recent literature [22] [23]. The band runs from the smallest λ whose loss lies within one standard error of the minimum to the largest. So, its right edge is the sparsest model not distinguishable from the minimum. The choice within the band is method specific. For LASSO and Adaptive LASSO the right edge is taken. For Ridge the geometric mean of the two edges is used since Ridge shrinks without selecting and over-shrinkage at the right edge inflates the dispersion estimate. The candidate values lie on a logarithmic grid and so this point is the arithmetic midpoint of the band on the scale in which the search is conducted. SCAD is tuned by the same band augmented with the stability criterion of Meinshausen and Bühlmann [24]. The concave penalty admits several local solutions and the loss surface alone does not separate them. In this case, the sparsest λ is retained whose selection frequency stays within 5% of the best value.

4. Simulation Studies

In this section, we assess the penalized OEF framework through Monte Carlo simulation set-up under the two over-dispersed regression models of Section 2. For each model, six estimators are compared. They include 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. Moreover, the covariates/features are generated with a first-order autoregressive (AR(1)) correlation structure of 0.5 presenting dependence among covariates which is a common phenomenon in practice.

Both designs cross an over-dispersion level with a signal density. The dispersion parameters are not comparable across the two models since one is bounded on the unit interval and the other is not.

The two over-dispersion levels in each design are interpreted more readily on the scale of the induced variance than on the scale of the dispersion parameter. **Table 2** gives both. Under the Beta-Binomial with $m = 40$ trials per cluster the variance inflation factor is $1 + (m-1)\rho$. Setting $\rho = 0.20$ inflates the Binomial variance 8.8-fold, a moderate departure from the Binomial and $\rho = 0.65$ inflates it 26.4-

fold, a severe departure from the Binomial. Under the Negative Binomial, the Variance-to-Mean Ratio (VMR) is $1 + \alpha\mu_i$. Therefore, the inflation grows with the fitted mean rather than remaining constant across observations. Values $\alpha = 0.25$ and $\alpha = 0.75$ give ratios of 1.3 and 2.0 at the baseline mean $\exp(0.25) = 1.28$. This contrast between inflation free of the mean and inflation growing with it is the structural difference to which Section 4.3 returns.

Table 2. Over-dispersion levels expressed as induced variance inflation.

Model	Level	Induced inflation	Behaviour
Beta-Binomial	$\rho = 0.20$	$1 + (m-1)\rho = 8.8$	Constant across observations
Beta-Binomial	$\rho = 0.65$	$1 + (m-1)\rho = 26.4$	Constant across observations
Negative Binomial	$\alpha = 0.25$	$\text{VMR} = 1 + \alpha\mu_i \approx 1.3$	Grows with the fitted mean
Negative Binomial	$\alpha = 0.75$	$\text{VMR} = 1 + \alpha\mu_i \approx 2.0$	Grows with the fitted mean

Hyper-parameter tuning is carried out based on the cross-validated OEF loss in (29) with the method specific k -SE and stability criteria presented in Section 3.3. We present the simulation results in the next sub-sections and their discussion in Section 4.3.

4.1. Beta-Binomial Logistic Regression

This section presents results of the finite-sample dispersion inference of the penalized OEFs in the Beta-Binomial (BB) Logistic Regression model. The BB model accommodates overdispersion in grouped binary data through an intra-cluster correlation parameter ρ that controls excess variability beyond the binomial case [9]. The variance function is as given in Equation (7) which 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 $Y_i \sim \text{Beta-Binomial}(m, \mu_i, \rho)$ 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.

4.1.1. 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. Resultantly, the sparsity ratios are $s/p = 0.2$ (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 3** summarizes the design.

Table 3. Design of the four Beta-Binomial simulation scenarios.

Scenario	n	p	s	s/p	ρ_{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; ρ_{true} = intra-cluster correlation.

Under this design, six estimators are evaluated: 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 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 (20) and (21) 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.

4.1.2. Hyperparameter Selection and Tuning Diagnostics

Hyper-parameter selection for the BB model follows the cross-validated OEF loss of Equation (29) with the k -SE band and method-specific selection rules as discussed in Section 3.3. 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 Equa-

tion (29). 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.

The choice of λ within the band differs for Ridge. Over-shrinking β worsens the mean fit and the dispersion equation compensates by absorbing the unexplained variability into ρ . The dispersion estimate is inflated as a result. The same coupling is reported for over-dispersed models generally where variability that the mean model leaves unexplained is taken up by the dispersion parameter [3] [25] [26]. The right k -SE value deepens the over-shrinkage and the left edge under-regularizes. Therefore, we set λ^* to be the geometric mean of the left and right k -SE values, which is a midpoint rule on the scale of the search. Candidate values of λ are generated on a logarithmic grid as is standard for penalized regression paths [21] [27]. The geometric mean of the two edges is the arithmetic midpoint of the band on that grid and the selected value satisfies $\log \lambda = (\log \lambda L + \log \lambda R)/2$ with λL and λR as the two edges. The arithmetic mean of the same two edges falls close to the right edge in log units and sits near the most heavily shrunk model in the band. The cross-validated loss stays within one standard error of its minimum everywhere in the band by construction [28]. The midpoint therefore reduces dispersion bias at a bounded cost in predictive adequacy and needs no further tuning constant.

Finally, SCAD requires yet another adjustment as the 1-SE rule alone does not suffice for it. 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 [17] [27] [29] leaving the loss insensitive to λ across a wide range.

Stability selection, as introduced by Meinshausen and Bühlmann [24] 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 [24]. 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 [30].

4.1.3. Dispersion Coverage

Consider first the two sparse scenarios $s/p = 0.2$. **Table 4** and **Table 5** report unconditional 95% Wald coverage for the dispersion parameter. For each case, two dispersion Confidence Interval (CI) constructions are compared: a direct Wald interval on the ρ scale and a delta-method interval via $\text{logit}(\rho)$ (θ -Wald). The direct interval is given by:

$$\hat{\rho} \pm z_{1-\alpha/2} \widehat{SE}(\hat{\rho}) \quad (31)$$

where $z_{1-\alpha/2} = 1.96$ for 95% coverage and $\widehat{SE}(\hat{\rho})$ taken as the square root of the corresponding diagonal entry of the Godambe (sandwich) covariance in Equation (22). The θ -Wald construction however maps the interval from the $\text{logit}(\rho)$ scale on which the sampling distribution may be more symmetric, back to the ρ -scale. The θ -Wald interval is obtained by the delta method [31].

Writing $\theta = \text{logit}(\rho)$, the standard error is transformed as:

$$\widehat{SE}(\hat{\theta}) = \frac{\widehat{SE}(\hat{\rho})}{\hat{\rho}(1-\hat{\rho})} \quad (32)$$

and the interval is formed on the θ scale and mapped back through the inverse logit such that:

$$CI_{\rho}^{-Wald} = \text{logit}^{-1} \left(\hat{\theta} \pm z_{1-\alpha/2} \widehat{SE}(\hat{\theta}) \right) \quad (33)$$

which confines the interval for ρ to $(0, 1)$ and improves coverage when the sampling distribution of $\hat{\rho}$ is asymmetric [32].

Dispersion coverage varies sharply across methods. Under direct ρ -Wald, L1_plain and Adaptive LASSO come close to nominal (0.928 and 0.924) whereas the unpenalized OEF (0.556), Ridge (0.546), and SCAD (0.582) all under cover. The θ -Wald transformation leaves the L1-type methods essentially unchanged (0.928 and 0.916) and lifts the others only slightly (unpenalized OEF 0.616, Ridge 0.598, SCAD 0.588). The limited gain is expected: a transformation of this kind improves coverage mainly when the untransformed sampling distribution is skewed and adds little when it is already close to symmetric [33] [34]. Since ρ is confined to $(0, 1)$, the distribution of $\hat{\rho}$ is fairly symmetric to begin with so mapping through the logit results to little correction.

Table 4. Dispersion coverage: Scenario 1 ($s = 4$, $\rho = 0.20$).

Method	Cov. ρ (direct Wald)	Cov. ρ (θ -Wald)
QuasiBinomial	0.948	0.948
Unpenalized OEF	0.556	0.616
Ridge (L2)	0.546	0.598
L1_plain (LASSO)	0.928	0.928
Adaptive LASSO	0.924	0.916
SCAD	0.582	0.588

Next, we consider Scenario 2 ($\rho = 0.65$) which is a harder setting for inference. Ridge's coverage drops to 0.520 implying that roughly half of its dispersion intervals miss the true ρ . The θ -Wald transformation has little effect in this simulation Scenario, 0.732 vs. 0.718 for the unpenalized OEF.

Table 5. Dispersion coverage: Scenario 2 ($s = 4$, $\rho = 0.65$).

Method	Cov. ρ (direct Wald)	Cov. ρ (θ -Wald)
QuasiBinomial	0.942	0.942
Unpenalized OEF	0.732	0.718
Ridge (L2)	0.520	0.512
L1_plain (LASSO)	0.928	0.922
Adaptive LASSO	0.912	0.902
SCAD	0.822	0.814

Increasing the active set to $s = 8$ moves the design into a moderately dense regime ($s/p = 0.4$). **Table 6** and **Table 7** report the results.

Table 6. Dispersion coverage: Scenario 3 ($s = 8$, $\rho = 0.20$).

Method	Cov. ρ (direct Wald)	Cov. ρ (θ -Wald)
QuasiBinomial	0.912	0.912
Unpenalized OEF	0.530	0.574
Ridge (L2)	0.456	0.496
L1_plain (LASSO)	0.904	0.910
Adaptive LASSO	0.920	0.924
SCAD	0.592	0.596

Scenario 4 ($\rho = 0.65$) amplifies these tendencies further. Ridge's ρ coverage drops to 0.328, the lowest in all scenarios. The explanation to this is the shape of $\hat{\rho}$ under Ridge at high over-dispersion. Its distribution is multimodal and much of its mass is centred away from ρ_{true} to cover it.

Table 7. Dispersion coverage: Scenario 3 ($s = 8$, $\rho = 0.65$).

Method	Cov. ρ (direct Wald)	Cov. ρ (θ -Wald)
QuasiBinomial	0.874	0.874
Unpenalized OEF	0.740	0.718
Ridge (L2)	0.328	0.326
L1_plain (LASSO)	0.954	0.952
Adaptive LASSO	0.964	0.956
SCAD	0.820	0.818

4.1.4. The Quasi-Binomial Comparison

Tables 4-7 contain a comparison worth treating on its own. The Quasi-Binomial GLM estimates dispersion by the Pearson Method-of-Moments (MoM) statistic and uses no third or fourth moment information. It attains dispersion coverage of 0.948, 0.942, 0.912 and 0.874 across the four scenarios. The unpenalized OEF is the Godambe-efficient member of the class and the benchmark for the penalized variants. It attains 0.556, 0.732, 0.530 and 0.740. The estimator with the smaller asymptotic variance produces the less accurate intervals in every scenario. The penalized variants do not share this. They attain between 0.924 and 0.964 and exceed the quasi-binomial in the two heavy over-dispersion scenarios. The failure is therefore specific to the unpenalized sandwich interval.

The two columns measure the same thing. The quasi-binomial interval is constructed for the dispersion multiplier $1 + (m-1)\rho$ and mapped to the ρ scale by a monotone linear inversion which preserves coverage exactly. The coverage of ρ obtained in this way equals the coverage of ϕ in every scenario.

Godambe optimality concerns the asymptotic variance of the estimator. It states that no other unbiased estimating function in the class attains a smaller variance in the limit [7] [35]. A Wald interval requires two further conditions, that is, that the estimated standard error is close to the true standard error and that the normal approximation holds at the sample size in use. Optimality addresses neither [36].

The difficulty lies in the dispersion block. The variability matrix entering the sandwich covariance in (22) requires the variance of the dispersion estimating function. That function is constructed from squared deviations and so its variance depends on fourth-order moments which is why the construction in Section 3.1 carries the skewness and kurtosis coefficients explicitly. Here those moments are estimated from residuals taken about a fitted mean with 21 parameters fitted from 100 observations and so a large share of the variation is absorbed by the fit. The estimated variability matrix is too small and the standard error inherits the deficiency. An interval of nominal width 1.96 standard errors is then too narrow to attain its stated coverage.

The Pearson estimator does not incur this cost. Its interval is referred to a chi-square distribution on $n - p$ degrees of freedom rather than constructed from an estimated variance. The degrees of freedom adjustment accounts for the fitted parameters directly and no fourth-order moment is estimated anywhere. The interval takes its shape from a reference distribution rather than estimating that shape from fewer than five observations per fitted parameter. The efficiency the OEF gains asymptotically is returned in the finite-sample estimation of its own variance and at $n = 100$ the second effect is the larger.

The logit-scale interval was introduced to correct skewness in the sampling distribution, and it does not improve coverage. The gain is at most 6 percentage points across the four scenarios and in half the cells of **Tables 4-7** the transformed interval covers less often than the direct one. A transformation moves a misplaced interval and does not widen a narrow one. The deficiency is therefore one of width.

The quasi-likelihood advantage is confined to this model. Under the Negative Binomial the ordering reverses. Quasi-Poisson dispersion coverage falls from 0.832 to 0.596 as over-dispersion and signal density increase, while the unpenalized OEF holds between 0.580 and 0.907 once the log-scale interval is used. The advantage under the Beta-Binomial comes from the bounded support and not from the moment estimator.

All four penalized variants are solved from the same estimating system as the unpenalized OEF. They use the same dispersion equation in (21) and take their standard errors from the same dispersion block of (22). They differ only in the penalty term applied to the regression equation. L1_plain covers between 0.904 and 0.954 across the four scenarios and Adaptive LASSO between 0.912 and 0.964, against 0.530 to 0.740 for the unpenalized fit. Coverage near nominal under the same construction points to the estimated variability matrix as the source of the undercoverage rather than to the estimating system. Ridge covers between 0.328 and 0.546 and SCAD between 0.582 and 0.822. Section 4.1.3 traces the Ridge figures to the multimodal shape of the dispersion estimate under over-shrinkage and Section 4.1.2 traces the SCAD figures to local solutions of the nonconcave objective. Both failures follow from the penalty and neither is a property of the interval construction.

The undercoverage has a known source. Residuals taken about a mean fitted with p parameters have an expected sum of squares reduced by approximately the factor $1 - p/n$ [5]. With 21 parameters at $n = 100$ that reduction is substantial and the effect is larger for a fourth order quantity. The estimated variability is therefore too small and the standard error built from it is too small and so an interval of nominal width is narrower than its stated coverage requires. Narrow intervals explain part of the undercoverage but not all of it. The rest is consistent with the dispersion estimating function departing from normality at $n = 100$ which makes a Wald interval the wrong shape as well as the wrong width. Both problems ease as the sample increases. Mancl and DeRouen [37] address the same problem by inflating residuals to account for the leverage of the fitted mean.

Downward bias of an empirical sandwich variance in small samples is documented in the estimating equation literature [38] [39]. Wald intervals built from it undercover for the same reason. The corrections proposed there address the regression block of the sandwich under clustering. What appears here is the same effect acting on a dispersion block. That block draws its variability from fourth order quantities and the attenuation compounds with the order of the moment.

At $n = 100$ the Godambe interval for the dispersion parameter in the unpenalized Beta-Binomial fit is best read alongside a moment-based or resampling interval. The L1-type fits do not need that support and hold coverage above 0.900 in every scenario, and Section 4.3 explains why.

4.1.5. Dispersion Estimation Accuracy

Tables 8-11 report bias and RMSE for the dispersion parameter estimate $\hat{\rho}$. van de Geer *et al.* [40] note that reporting both quantities separates systematic distor-

tion from total estimation error.

Table 8. Bias and RMSE: Scenario 1 ($s = 4$, $\rho = 0.20$).

Method	Bias $\hat{\rho}$	RMSE $\hat{\rho}$	Mean $\hat{\rho}$
Unpenalized OEF	-0.028	0.092	0.172
Ridge (L2)	-0.018	0.125	0.182
L1_plain (LASSO)	-0.024	0.085	0.176
Adaptive LASSO	-0.023	0.079	0.177
SCAD	+0.019	0.153	0.219

Table 9. Bias and RMSE: Scenario 2 ($s = 4$, $\rho = 0.65$).

Method	Bias $\hat{\rho}$	RMSE $\hat{\rho}$	Mean $\hat{\rho}$
Unpenalized OEF	-0.067	0.087	0.583
Ridge (L2)	-0.098	0.184	0.552
L1_plain (LASSO)	-0.080	0.099	0.570
Adaptive LASSO	-0.077	0.096	0.573
SCAD	-0.060	0.079	0.590

Table 10. Bias and RMSE: Scenario 3 ($s = 8$, $\rho = 0.20$).

Method	Bias $\hat{\rho}$	RMSE $\hat{\rho}$	Mean $\hat{\rho}$
Unpenalized OEF	-0.024	0.108	0.176
Ridge (L2)	+0.018	0.209	0.218
L1_plain (LASSO)	-0.025	0.089	0.175
Adaptive LASSO	-0.028	0.091	0.172
SCAD	+0.024	0.164	0.224

Table 11. Bias and RMSE: Scenario 4 ($s = 8$, $\rho = 0.65$).

Method	Bias $\hat{\rho}$	RMSE $\hat{\rho}$	Mean $\hat{\rho}$
Unpenalized OEF	-0.074	0.094	0.576
Ridge (L2)	-0.119	0.250	0.531
L1_plain (LASSO)	-0.089	0.106	0.561
Adaptive LASSO	-0.089	0.107	0.561
SCAD	-0.063	0.087	0.587

A central finding in this paper is that penalization disturbs dispersion inference substantially. Under the BB model, the RMSE $\hat{\rho}$ varies by a factor of three across penalties, from 0.079 for the Adaptive LASSO to 0.250 for Ridge. Under the NB, the effect is larger still with bias ranging from -0.052 to -0.563 within Scenario 4 alone.

The mechanism behind this is the mean-dispersion coupling of the BB variance function that was described earlier in Section 4.1.2. Errors in β propagate through the success probability in Equation (4) into the dispersion equation where they are absorbed as shifts in $\hat{\rho}$. Therefore, penalizing β alone alters the residual structure used to estimate variability and the dispersion equation compensates. This is not a numerical anomaly but a structural feature of the joint OEF system, one that standard penalized GLM procedures do not expose because they treat dispersion as ancillary.

At $\rho = 0.20$ all penalized methods show modest $\hat{\rho}$ bias but differ sharply in stability. SCAD carries only a small positive pooled bias of 0.019 in Scenario 1 and 0.024 in Scenario 3 yet the largest RMSE of all methods (0.153 and 0.164) nearly twice that of the unpenalized OEF. This mix of near-zero bias with inflated RMSE points to a heavy-tailed rather than systematically shifted $\hat{\rho}$ distribution.

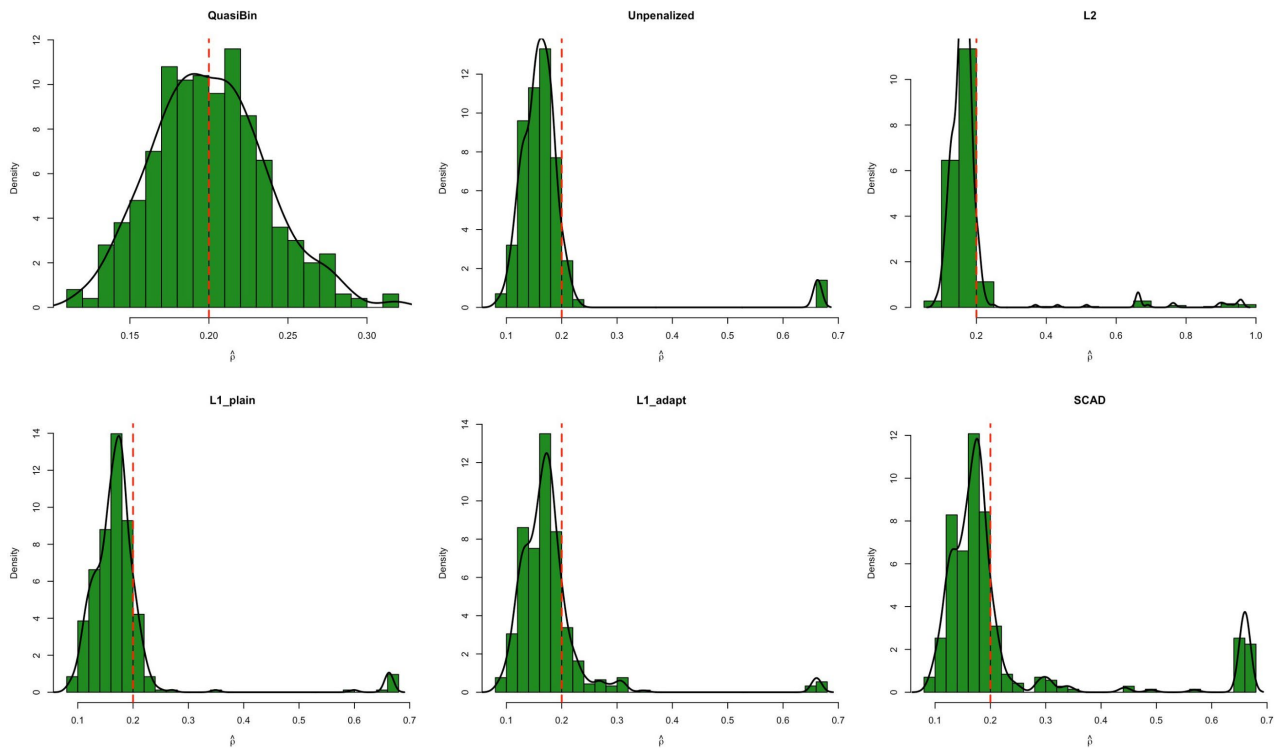
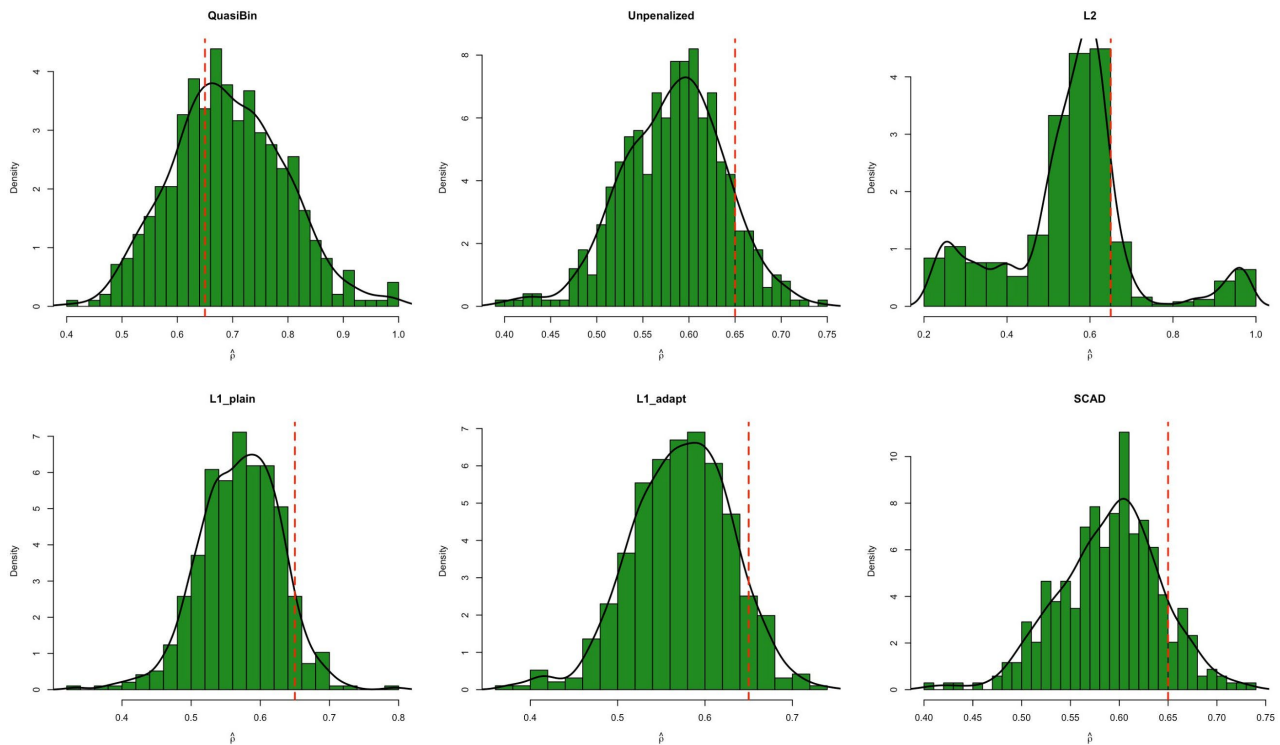
The histograms (**Figure 1** and **Figure 3**) show SCAD's cluster just below ρ_{true} alongside a distinct secondary cluster near 0.65 implying the squared-error summary is driven by this right-tail minority. This is consistent with the multiple local solutions admitted by the nonconvex SCAD penalty [17] [27] [29], under which selection instability feeds into dispersion variability through the coupled system.

L1_plain, by contrast, achieves both a small bias of -0.024 in Scenario 1 and the tightest $\hat{\rho}$ concentration among the penalized methods (RMSE $\hat{\rho} = 0.085$).

Ridge gives the largest RMSE $\hat{\rho}$ at $\rho = 0.20$ of 0.125 in Scenario 1 and 0.209 in Scenario 3. This reflects a multimodal $\hat{\rho}$ distribution. The multimodal structure arises from the over-shrinkage mechanism described in Section 4.1.2. When λ is large, Ridge over shrinks β thus making the mean fit to worsen. This results in the dispersion equation accommodating the residual variability by inflating $\hat{\rho}$ and thus producing the right-tail hump visible in **Figures 1-4**. This is precisely why the geometric mean K -SE rule for Ridge we introduced in Section 4.1.2 is key as it selects the weakest penalization within the acceptable band. Resultantly, the penalty avoids the over shrinkage-dispersion compensation cycle.

At $\rho = 0.65$ all methods underestimate ρ , including SCAD which has a bias of -0.060 in Scenario 2 and -0.063 in Scenario 4. The unpenalized OEF is the most stable method with consistently low RMSE $\hat{\rho}$ (0.087 - 0.094) and moderate negative bias (-0.024 to -0.074). This indicates that much of the dispersion instability under the penalized methods is penalty-induced rather than intrinsic to the OEF system. At this high ρ , L1_plain, Adaptive LASSO, and SCAD give comparably concentrated $\hat{\rho}$ distributions, RMSE $\hat{\rho}$ spans 0.079 - 0.107. This is a clear contrast to the low ρ regime in which SCAD is the most volatile. Under this over-dispersion regime, Ridge remains the most volatile overall recording a RMSE $\hat{\rho}$ of 0.250 in Scenario 4.

Figures 1-4 display per-method histograms of $\hat{\rho}$ for the four scenarios.

Distribution of Dispersion Estimates ($\hat{\rho}$) Across Monte Carlo Replications**Figure 1.** Per-method histograms of $\hat{\rho}$: Scenario 1 ($s=4$, $\rho=0.20$).Distribution of Dispersion Estimates ($\hat{\rho}$) Across Monte Carlo Replications**Figure 2.** Per-method histograms of $\hat{\rho}$: Scenario 2 ($s=4$, $\rho=0.65$).

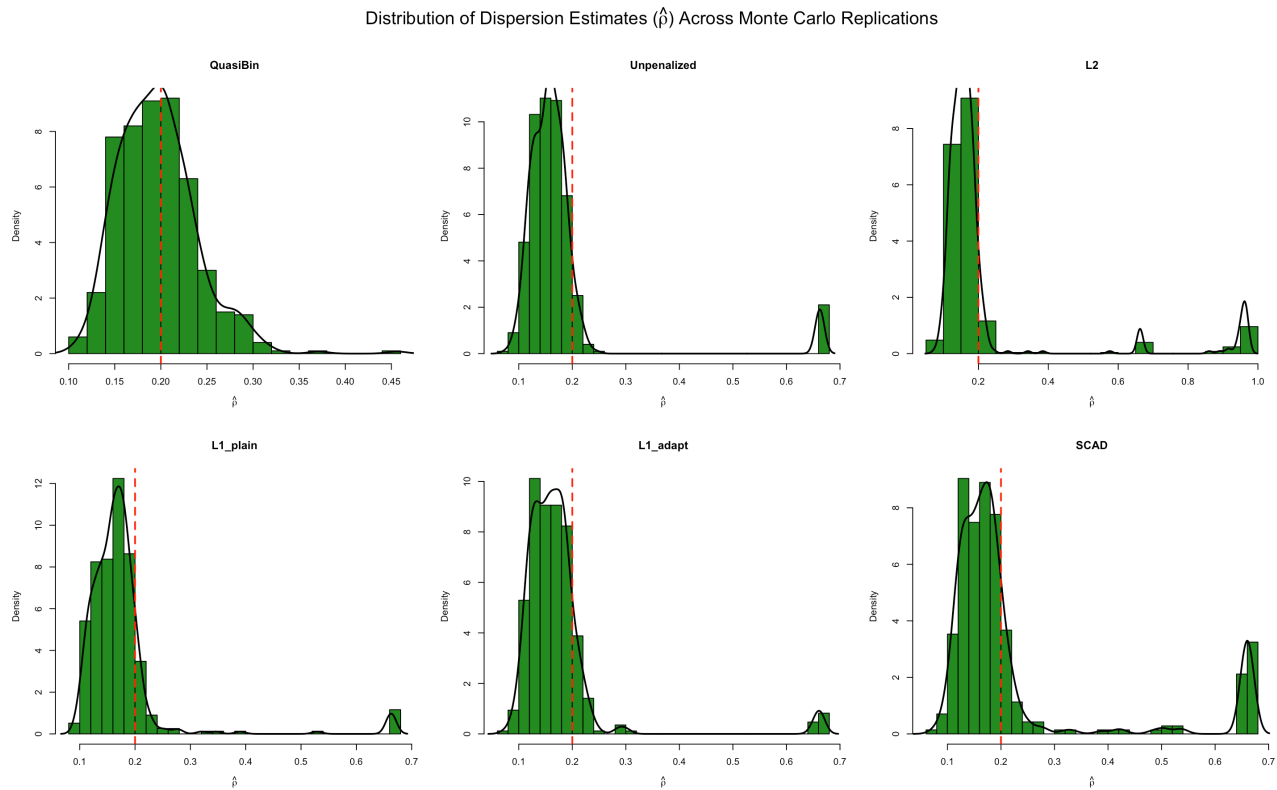


Figure 3. Per-method histograms of $\hat{\rho}$: Scenario 3 ($s = 8$, $\rho = 0.20$).

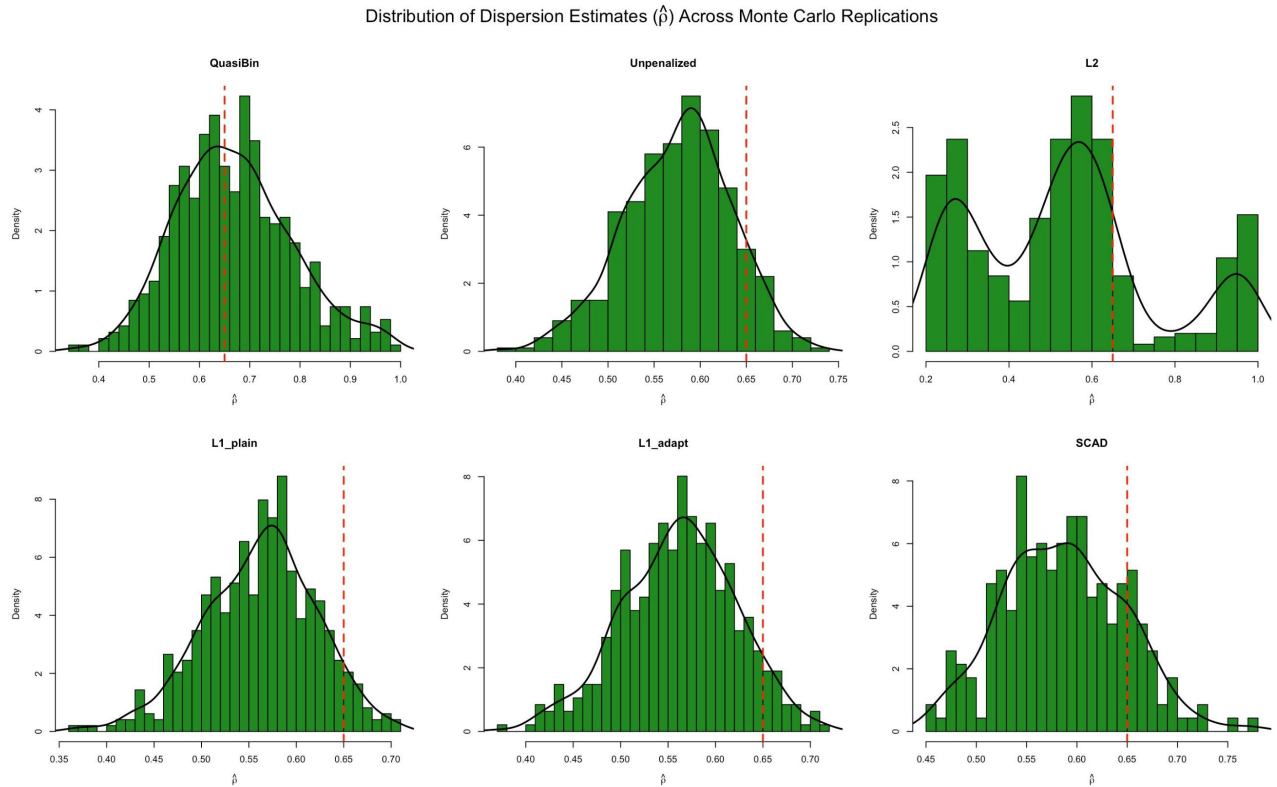


Figure 4. Per-method histograms of $\hat{\rho}$: Scenario 4 ($s = 8$, $\rho = 0.65$).

4.2. Negative-Binomial Regression

This section reports dispersion inference for the penalized OEFs under Negative-Binomial (NB2) regression, where the variance is quadratic in the mean. The NB2 model carries over-dispersion in a heterogeneity parameter α through the variance function in Equation (11) and nests the Poisson model in the limit $\alpha \rightarrow 0$ [10] [11] [41]. The conditional mean uses the log link of Equation (8) and so the fitted counts stay positive. The exponential mean bears directly on penalty behaviour. A coefficient acts multiplicatively on the mean under this link, so shrinkage in β produces multiplicative rather than additive distortion in the fitted mean μ_i [10] [41]. The quadratic variance function then carries that distortion into the dispersion estimate, as Sections 4.2.3 and 4.2.4 show.

The data-generating process follows this specification. For each replication $p = 20$ covariates are drawn from a mean-zero multivariate normal distribution with first-order autoregressive correlation AR(1) and correlation 0.5. Given a coefficient vector β the conditional mean for the i^{th} observation follows the log link of Equation (8) and the mean count of Equation (9). The response is drawn as Negative-Binomial $Y_i \sim \text{NB}(\mu_i, \alpha)$ with the variance of Equation (11). As in the Beta-Binomial study of Section 4.1, $R = 500$ independent datasets are generated per scenario under a common random seed. The sample size is raised to $n = 200$ here to accommodate the heavier tails of count responses.

4.2.1. Simulation Design

Four scenarios cross two over-dispersion levels $\alpha = 0.25$ and $\alpha = 0.75$ with two active set sizes $s = 4$ and $s = 8$ holding $n = 200$, $p = 20$ and $R = 500$ replications fixed. The VMR is $1 + \alpha\mu_i$. Table 12 gives the design.

Table 12. Design of the four negative-binomial simulation scenarios.

Scenario	n	p	s	s/p	α_{true}	Non-Zero β	Design
1	100	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	100	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	100	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	100	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 4.1: the Quasi-Poisson GLM, the unpenalized OEF, Ridge (L2-OEF), L1_plain (LASSO-OEF), Adaptive LASSO and SCAD. The Quasi-Poisson GLM estimates β by iteratively re-

weighted least squares under a Quasi-Poisson variance and estimates dispersion by the Pearson Method of Moments statistic. The unpenalized OEF estimates the regression and dispersion parameters jointly through the orthogonalized EFs in (20) and (21). The penalized estimators apply their penalties to the regression equation only.

4.2.2. Hyperparameter Selection and Tuning Diagnostics

Tuning follows the cross-validated OEF loss in (29) and the method-specific k -SE rules of Section 3.3, the same framework used for the BB model. 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 which avoids the over-shrinkage compensation described in Section 4.1.2 and which operates here with an amplified effect under the exponential mean. SCAD uses the stability-augmented rule of Section 4.1.2 with the NB2 settings $\tau = 0.80$ and $\text{guard} = 1.10\times$. The guard falls back to the CV minimum in all four scenarios.

4.2.3. Dispersion Coverage

Consider first the two sparse scenarios $s/p = 0.20$. **Table 13** and **Table 14** report unconditional 95% Wald coverage for the dispersion parameter. Two interval forms are compared, as in Section 4.1.3. The first is a direct Wald interval on the α -scale. The second is a delta-method interval on the $\log(\alpha)$ scale, denoted the η -Wald, which parallels the θ -Wald construction of Equations (31) to (33) with the log link in place of the logit.

Coverage of α under the direct Wald interval falls below nominal for the unpenalized OEF at 0.526 and for SCAD at 0.434. The η -Wald interval recovers most of this. The unpenalized OEF rises from 0.526 to 0.906 and Adaptive LASSO from 0.966 to 0.974. The transformation works because α is unbounded above, $\alpha \in (0, \infty)$, and the sampling distribution of $\hat{\alpha}$ is severely right skewed under every method. Consequently, the log scale symmetrizes it which is where a transformation of this kind gives most benefit [33] [34]. Section 4.2.4 takes this up.

Table 13. Dispersion coverage: Scenario 1 ($s = 4$, $\alpha = 0.25$).

Method	Cov. α (direct Wald)	Cov. α (η -Wald)
QuasiPoisson	0.832	0.832
Unpenalized OEF	0.526	0.906
Ridge (L2)	0.645	0.657
L1_plain (LASSO)	0.990	0.990
Adaptive LASSO	0.966	0.974
SCAD	0.434	0.660

Scenario 2, $\alpha = 0.75$ repeats the pattern. The η -Wald gain is large. L1_plain rises from 0.352 to 0.882 and Adaptive LASSO from 0.388 to 0.940.

Table 14. Dispersion coverage: Scenario 2 ($s = 4$, $\alpha = 0.75$).

Method	Cov. α (direct Wald)	Cov. α (η -Wald)
QuasiPoisson	0.682	0.682
Unpenalized OEF	0.462	0.606
Ridge (L2)	0.656	0.732
L1_plain (LASSO)	0.352	0.882
Adaptive LASSO	0.388	0.940
SCAD	0.316	0.342

Table 15 and **Table 16** give the denser scenarios $s/p = 0.4$. The move from $s/p = 0.2$ to $s/p = 0.4$ brings out a distinct NB2 pattern.

Table 15. Dispersion coverage: Scenario 3 ($s = 8$, $\alpha = 0.25$).

Method	Cov. α (direct Wald)	Cov. α (η -Wald)
QuasiPoisson	0.790	0.790
Unpenalized OEF	0.524	0.907
Ridge (L2)	0.583	0.549
L1_plain (LASSO)	0.996	0.996
Adaptive LASSO	0.965	0.975
SCAD	0.424	0.657

Quasi-Poisson dispersion coverage falls from 0.832 in Scenario 1 to 0.596 in Scenario 4. The Pearson Method of Moments estimator deteriorates under the combined stress of high over-dispersion and dense signals.

Table 16. Dispersion coverage: Scenario 4 ($s = 8$, $\alpha = 0.75$).

Method	Cov. α (direct Wald)	Cov. α (η -Wald)
QuasiPoisson	0.596	0.596
Unpenalized OEF	0.456	0.580
Ridge (L2)	0.664	0.694
L1_plain (LASSO)	0.416	0.840
Adaptive LASSO	0.348	0.958
SCAD	0.468	0.540

4.2.4. Dispersion Estimation Accuracy

Tables 17-20 report bias and RMSE for the heterogeneity parameter α .

Table 17. Bias and RMSE: Scenario 1 ($s = 4$, $\alpha = 0.25$).

Method	Bias $\hat{\alpha}$	RMSE $\hat{\alpha}$	Mean $\hat{\alpha}$
Unpenalized OEF	-0.107	0.127	0.143
Ridge (L2)	+0.226	0.624	0.476
L1_plain (LASSO)	-0.149	0.178	0.101
Adaptive LASSO	-0.164	0.196	0.086
SCAD	-0.063	0.357	0.187

Table 18. Bias and RMSE: Scenario 2 ($s = 4$, $\alpha = 0.75$).

Method	Bias $\hat{\alpha}$	RMSE $\hat{\alpha}$	Mean $\hat{\alpha}$
Unpenalized OEF	-0.228	0.254	0.522
Ridge (L2)	-0.104	0.237	0.646
L1_plain (LASSO)	-0.425	0.460	0.325
Adaptive LASSO	-0.589	0.619	0.161
SCAD	-0.509	0.526	0.241

Table 19. Bias and RMSE: Scenario 3 ($s = 8$, $\alpha = 0.25$).

Method	Bias $\hat{\alpha}$	RMSE $\hat{\alpha}$	Mean $\hat{\alpha}$
Unpenalized OEF	-0.109	0.128	0.141
Ridge (L2)	+0.250	0.467	0.500
L1_plain (LASSO)	-0.139	0.169	0.111
Adaptive LASSO	-0.155	0.195	0.095
SCAD	-0.100	0.257	0.150

Table 20. Bias and RMSE: Scenario 3 ($s = 8$, $\alpha = 0.75$).

Method	Bias $\hat{\alpha}$	RMSE $\hat{\alpha}$	Mean $\hat{\alpha}$
Unpenalized OEF	-0.230	0.275	0.520
Ridge (L2)	-0.052	0.320	0.699
L1_plain (LASSO)	-0.340	0.382	0.410
Adaptive LASSO	-0.563	0.616	0.187
SCAD	-0.388	0.421	0.362

As in the BB model, penalization disturbs α estimation far more than β esti-

mation and the exponential mean structure amplifies the effect. The variance in Equation (11) grows with the square of the mean in Equation (9) and so errors in β propagate quadratically into the variance function and small β distortions produce large shifts in $\hat{\alpha}$. At $\alpha = 0.25$ every method except Ridge underestimates α and at $\alpha = 0.75$ every method including Ridge underestimates α . The downward bias is most severe for Adaptive LASSO at -0.164 in Scenario 1 and -0.589 in Scenario 2, which matches its sharply concentrated $\hat{\alpha}$ distributions with mean $\hat{\alpha} = 0.086$ against $\alpha_{\text{true}} = 0.25$. Ridge runs the other way at $\alpha = 0.25$ with a positive bias of 0.226 in Scenario 1 and 0.250 in Scenario 3 and the largest RMSE $\hat{\alpha}$ of any method at 0.624 and 0.467 . At $\alpha = 0.75$ its bias turns slightly negative and it becomes the most accurate in central tendency even though its RMSE $\hat{\alpha}$ stays high. The unpenalized OEF gives the most balanced accuracy of RMSE $\hat{\alpha}$ which places the instability of the L1-type penalties in the penalty rather than in the estimating system.

These behaviours follow from the geometry of the log link. The right skewness of $\hat{\alpha}$ under every method arises because the variance function couples α multiplicatively with $\mu^2 = \exp(2\mathbf{x}^T\beta)$. The exponential growth of μ^2 creates heavy right tails and asymmetric estimating equations for α , which is why direct Wald coverage is poor. This is a documented feature of interval estimation for the negative-binomial dispersion parameter [1] [42], and it is why the η -Wald transformation on $\eta = \log(\alpha)$ re-symmetrizes the dispersion geometry so effectively [43]. The same link accounts for the amplified Ridge bias. A coefficient acts multiplicatively on the mean under the log link [10] [41], so linear shrinkage in β becomes multiplicative shrinkage in μ , most severely where α is high.

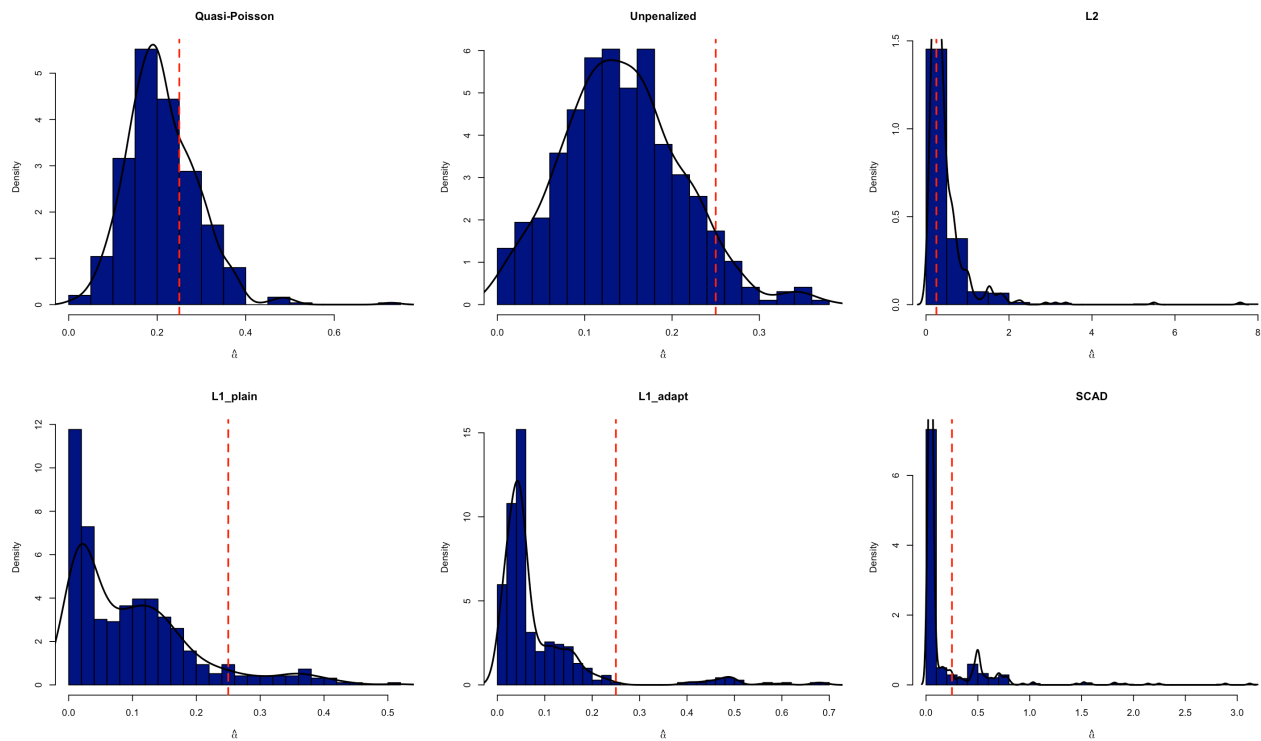
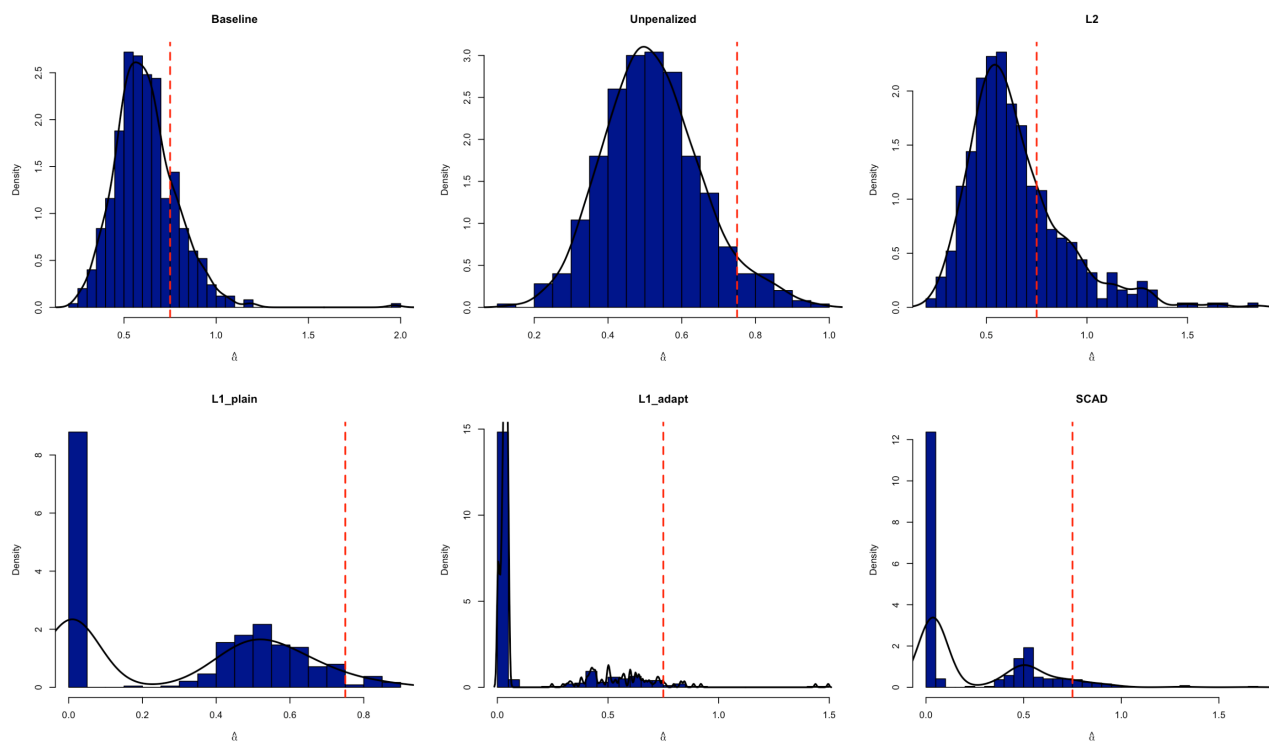
The multiplicative coupling through the exponential link is the structural reason dispersion inference is so penalty-sensitive under NB2.

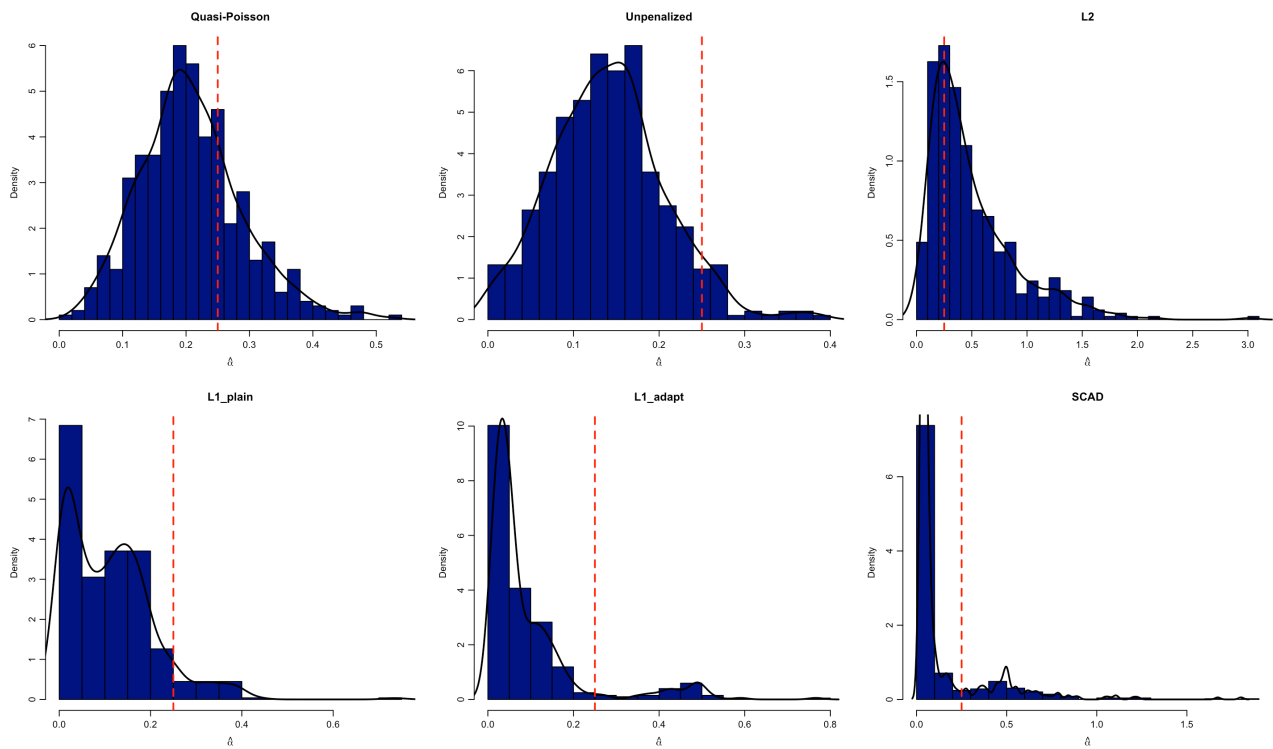
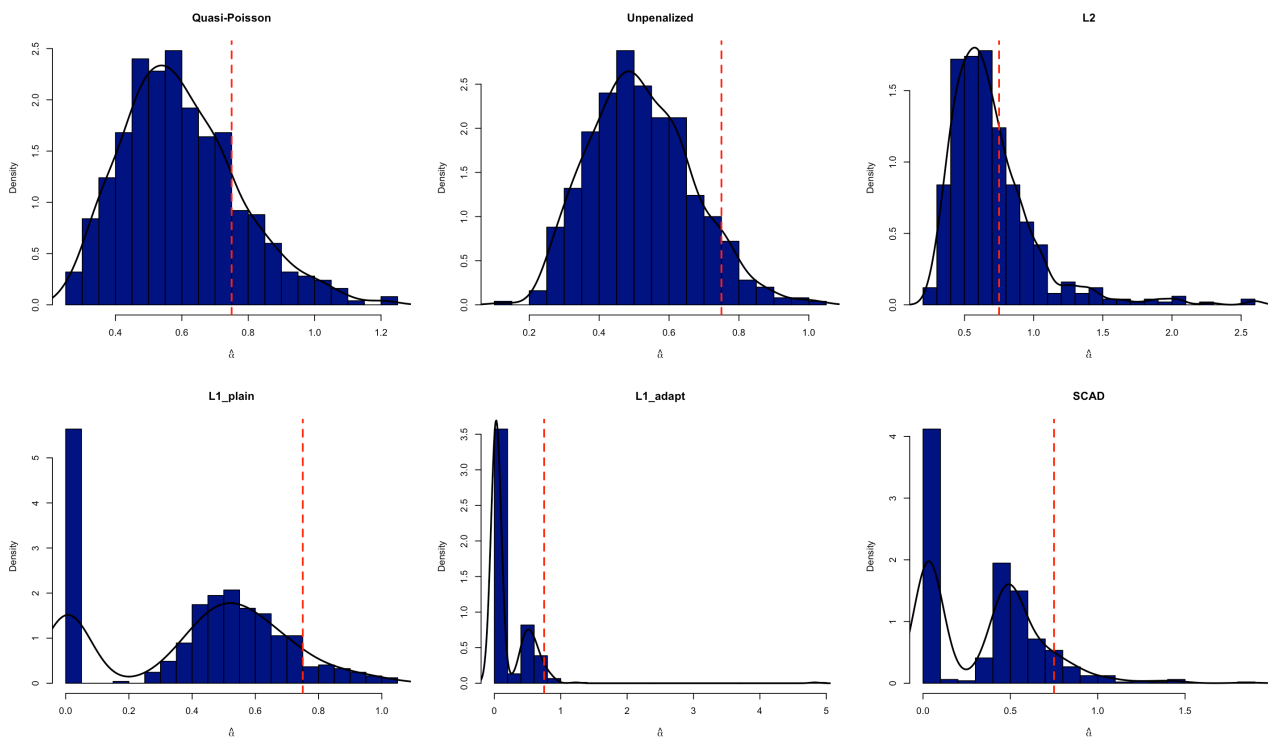
Figures 5-8 display per-method histograms of $\hat{\alpha}$ for the four scenarios.

The η -Wald interval is the practical device for NB2 dispersion inference. Across the four scenarios the η -Wald procedure improves coverage well above direct Wald with the largest gains among the penalized methods and smaller gains for the unpenalized OEF. L1_plain η -Wald coverage exceeds 0.84 everywhere, ranging from 0.840 to 0.998 against direct Wald as low as 0.352 . Adaptive LASSO improves from 0.348 to 0.958 . SCAD is the exception. Its direct Wald coverage is already low at 0.316 to 0.468 and the η -Wald remains modest at 0.342 to 0.660 . The cause is SCAD's severe underestimation of α . A transformation re-symmetrizes an interval but cannot recenter it on a value the point estimate has missed and so with $\hat{\alpha}$ far below α_{true} , the true value stays outside the interval and coverage barely moves.

4.3. Discussion

The two simulation studies examined penalized OEF dispersion inference under Beta-Binomial logistic regression in Section 4.1 and Negative-Binomial regression in Section 4.2. They agree on the behaviour of the framework and disagree on almost every ranking within it. The disagreements trace to the structural differences between the models, that is, the bounded response and logit link of the BB model against the unbounded counts and log link of NB2.

Distribution of Dispersion Estimates ($\hat{\alpha}$) Across Monte Carlo Replications**Figure 5.** Per-method histograms of $\hat{\alpha}$: Scenario 1 ($s=4$, $\alpha=0.25$).Distribution of Dispersion Estimates ($\hat{\alpha}$) Across Monte Carlo Replications**Figure 6.** Per-method histograms of $\hat{\alpha}$: Scenario 2 ($s=4$, $\alpha=0.75$).

Distribution of Dispersion Estimates ($\hat{\sigma}$) Across Monte Carlo Replications**Figure 7.** Per-method histograms of $\hat{\sigma}$: Scenario 3 ($s = 8$, $\alpha = 0.25$).Distribution of Dispersion Estimates ($\hat{\sigma}$) Across Monte Carlo Replications**Figure 8.** Per-method histograms of $\hat{\sigma}$: Scenario 4 ($s = 8$, $\alpha = 0.75$).

Penalization affects dispersion inference more strongly than coefficient inference and this holds in both models. The mean-dispersion coupling of the joint OEF system in Equations (20) and (21) is the reason. Errors in β propagate through the mean function into the dispersion equation which carries no penalty and so shifts the dispersion estimate to restore the moment match. Penalization adds a second source of variation to the dispersion estimate. The tuning parameter is selected from the data in every replication so the dispersion estimate is a random function of it and its variance separates into the variance at a fixed tuning value plus the variance contributed by selecting that value. The first term is the one the unpenalized fit also carries. The second has no counterpart there and it is what the choice of point within the standard-error band controls. NB2 amplifies the coupling because errors propagate multiplicatively through Equation (9) and quadratically through Equation (11). The unpenalized OEF gives the most consistent dispersion RMSE across scenarios with RMSE $\hat{\rho}$ from 0.087 to 0.108 under BB model and RMSE $\hat{\alpha}$ from 0.127 to 0.275 under NB2 model. The instability under the penalized methods is therefore penalty-induced and not intrinsic to the estimating system. Under BB the most volatile estimator is Ridge with RMSE $\hat{\rho}$ up to 0.250, which follows the over-shrinkage compensation cycle of Section 4.1.2. Under NB2, the largest bias comes from Adaptive LASSO at high over-dispersion. The bias reaches -0.589 in Scenario 2 and -0.563 in Scenario 4, where $\alpha_{\text{true}} = 0.75$ and the mean falls to 0.161 and 0.187 respectively. This follows from its weight dependent shrinkage as reported in Section 4.2.4. At the lower over-dispersion level, the largest bias is Ridge in the opposite direction, at $+0.226$ and $+0.250$ in the two scenarios.

Whether a variance-stabilizing transformation helps is likewise model-dependent and for the same structural reason. Under NB2, the η -Wald interval on the log scale raises α coverage by as much as 61 percentage points while the analogous θ -Wald interval on the logit scale gains at most 6 points for ρ coverage under BB model. The dispersion geometry explains the difference. The NB2 heterogeneity parameter $\alpha \in (0, \infty)$ is severely right-skewed under the exponential mean and the log scale symmetrizes it. The BB intra-cluster correlation $\rho \in (0, 1)$ is confined to the unit interval and is already close to symmetric and gains little from the logit scale. Sections 4.1.3 and 4.2.3 give the cell-level results.

SCAD's reliability turns on the smoothness of the estimating-function geometry which again separates the two models. Under the bounded BB support SCAD is the most variable method for $\hat{\rho}$ with dispersion RMSE $\hat{\rho}$ spanning 0.153 to 0.164 at $\rho = 0.20$ despite small pooled bias, and the instability worsens at $\rho = 0.65$. Under the smoother NB2 geometry, it is numerically stable across all four scenarios with reduced sensitivity only at the dense setting $s/p = 0.4$. Both patterns follow from the multiple local solutions admitted by the nonconvex penalty [17] [27]. The bounded BB response range worsens that multiplicity and the monotone NB2 score function suppresses it.

5. Conclusions

This study quantified what penalization does to dispersion inference in over-dispersed generalized linear models using a joint Optimal Estimating Function system in which the regression equation is penalized and the dispersion equation is not. Across 500 replications per scenario in two structurally distinct models, the disturbance to dispersion estimation exceeded the disturbance to coefficient estimation by a wide margin. Coefficient RMSE varied by a factor of two to four across penalties. The bias and RMSE of the dispersion estimate varied by a factor of up to ten.

The attributable mechanism is the mean-dispersion coupling of the joint system. Penalization displaces the fitted mean and the squared displacement enters the dispersion equation as though it were random variation in the response. The dispersion equation carries no penalty and so the dispersion estimate absorbs the mean model error. This is a structural property of joint estimation rather than a numerical artefact. Standard penalized GLM procedures do not expose it because they estimate dispersion after the mean model is fixed.

The form the disturbance takes is model-specific. Under the Beta-Binomial, Ridge over-shrinkage drives a compensation cycle that inflates the dispersion estimate and produces multimodal sampling distributions, giving Ridge the largest dispersion RMSE of any method at 0.250 in Scenario 4. SCAD produces near-zero pooled bias with the heaviest tails, a consequence of multiple local solutions. Under the Negative Binomial, the exponential mean propagates coefficient error quadratically into the variance function and the L1-type penalties underestimate the heterogeneity parameter severely, the Adaptive LASSO reaching a bias of -0.589 in Scenario 2. The unpenalized OEF is the most stable dispersion estimator in both models, which places the instability in the penalty and not in the estimating system.

Whether a variance-stabilizing transformation is needed follows the same structure. The log-scale interval raises Negative Binomial dispersion coverage by as much as 61 percentage points, from 0.348 to 0.958 for the Adaptive LASSO in Scenario 4, because the heterogeneity parameter is severely right-skewed under the exponential mean. The logit-scale interval gains at most 6 points under the Beta-Binomial because the intra-cluster correlation is confined to the unit interval and is already close to symmetric. A direct Wald interval for the dispersion parameter should not be reported for count data. For proportion data, the transformation yields little.

The Pearson moment estimator attains coverage closer to nominal than the Godambe-optimal OEF in every Beta-Binomial scenario. This is the clearest practical caution to come out of the study. Asymptotic optimality of the estimating function does not extend to finite-sample calibration of a Wald interval built from an estimated sandwich, and at $n = 100$ the dispersion block of that sandwich is too small. Dispersion intervals in this setting should be built from a moment-based or resampling procedure rather than from the Godambe covariance alone until this

behaviour is better characterized.

Several limitations qualify these conclusions. The simulations were confined to a moderate dimensional setting with $p = 20$, AR(1) covariate correlation and two over-dispersion levels per model. Behaviour under stronger feature dependence or model misspecification remains to be characterized and sensitivity to the tuning rules was not examined. The dispersion instability induced by penalization, in particular the over-shrinkage compensation under Ridge and the local-solution multiplicity under SCAD, invites penalties or tuning rules that protect the dispersion equation directly. Extending the framework to zero-inflated and other over-dispersed models is a natural next step. Within these limits, penalized Optimal Estimating Functions give a coherent route to dispersion inference in over-dispersed regression provided the penalty is chosen with the response geometry of the model in mind.

Three of these directions are worth stating more precisely. Every simulation here used $n = 100$. Tracing dispersion coverage across $n = 100, 200, 500, 1000$ and 2000 would show where the sandwich interval reaches its nominal level. The design is moderate dimensional and how the L1 penalties behave when p exceeds n is the subject of companion work. The tuning criterion in (29) scores each candidate λ by the norm of the held-out estimating function alone and so the criterion measures the predictive fit and ignores what that λ does to the dispersion estimate. Adding a second term, in the gap between the held-out Pearson statistic and its expectation under the fitted model, would make the criterion respond to dispersion bias as well as to predictive fit. A weight would control how much the dispersion term contributes relative to the predictive one. Two questions follow. How the selected λ behaves as n grows and how far the selection moves when that weight is varied. Settling them is a natural direction for future work on tuning criteria in joint estimation.

Author Contributions

Conceptualization, T.M., A.S., P.K. and J.O.; methodology, T.M., A.S., P.K. and J.O.; software, T.M.; validation, A.S., P.K. and J.O.; formal analysis, T.M.; investigation, T.M.; writing—original draft preparation, T.M.; writing—review and editing, T.M., A.S. and P.K.; supervision, A.S. and P.K. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Lloyd-Smith, J.O. (2007) Maximum Likelihood Estimation of the Negative Binomial Dispersion Parameter for Highly Overdispersed Data, with Applications to Infectious Diseases. *PLOS ONE*, **2**, e180. <https://doi.org/10.1371/journal.pone.0000180>
- [2] Lindén, A. and Mäntyniemi, S. (2011) Using the Negative Binomial Distribution to

- Model Overdispersion in Ecological Count Data. *Ecology*, **92**, 1414-1421.
<https://doi.org/10.1890/10-1831.1>
- [3] Harrison, X.A. (2014) Using Observation-Level Random Effects to Model Overdispersion in Count Data in Ecology and Evolution. *Peer Journal*, **2**, e616.
<https://doi.org/10.7717/peerj.616>
 - [4] Wedderburn, R.W.M. (1974) Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, **61**, 439-447.
<https://doi.org/10.1093/biomet/61.3.439>
 - [5] McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. 2nd Edition, Chapman & Hall.
 - [6] Godambe, V.P. (1985) The Foundations of Finite Sample Estimation in Stochastic Processes. *Biometrika*, **72**, 419-428. <https://doi.org/10.1093/biomet/72.2.419>
 - [7] Godambe, V.P. and Thompson, M.E. (1989) An Extension of Quasi-Likelihood Estimation. *Journal of Statistical Planning and Inference*, **22**, 137-152.
[https://doi.org/10.1016/0378-3758\(89\)90106-7](https://doi.org/10.1016/0378-3758(89)90106-7)
 - [8] Desmond, A.F. (1997) Optimal Estimating Functions, Quasi-Likelihood and Statistical Modelling. *Journal of Statistical Planning and Inference*, **60**, 77-104.
[https://doi.org/10.1016/s0378-3758\(96\)00123-1](https://doi.org/10.1016/s0378-3758(96)00123-1)
 - [9] Najera-Zuloaga, J., Lee, D.J. and Arostegui, I. (2017) Comparison of Beta-Binomial Regression Model Approaches to Analyze Health-Related Quality of Life Data. *Statistical Methods in Medical Research*, **27**, 2989-3009.
<https://doi.org/10.1177/0962280217690413>
 - [10] Cameron, A.C. and Trivedi, P.K. (2013). Regression Analysis of Count Data. 2nd Edition, Cambridge University Press. <https://doi.org/10.1017/cbo9781139013567>
 - [11] Lehman, R.R. and Archer, K.J. (2019) Penalized Negative Binomial Models for Modeling an Overdispersed Count Outcome with a High-Dimensional Predictor Space: Application Predicting Micronuclei Frequency. *PLOS ONE*, **14**, e0209923.
<https://doi.org/10.1371/journal.pone.0209923>
 - [12] Suryadi, F., Jonathan, S., Jonatan, K. and Ohhyver, M. (2023) Handling Overdispersion in Poisson Regression Using Negative Binomial Regression for Poverty Case in West Java. *Procedia Computer Science*, **216**, 517-523.
<https://doi.org/10.1016/j.procs.2022.12.164>
 - [13] Johnson, B.A., Lin, D.Y. and Zeng, D. (2008) Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models. *Journal of the American Statistical Association*, **103**, 672-680. <https://doi.org/10.1198/016214508000000184>
 - [14] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Non-orthogonal Problems. *Technometrics*, **12**, 55-67.
<https://doi.org/10.1080/00401706.1970.10488634>
 - [15] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58**, 267-288.
<https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
 - [16] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. <https://doi.org/10.1198/016214506000000735>
 - [17] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360.
<https://doi.org/10.1198/016214501753382273>
 - [18] Claeskens, G. and Hjort, N.L. (2008) Model Selection and Model Averaging. Cambridge University Press.

- [19] Komiyama, J. and Maehara, T. (2018) Cross Validation Based Model Selection via Generalized Method of Moments. <https://doi.org/10.48550/arXiv.1807.06993>
- [20] Wang, F., Mukherjee, S., Richardson, S. and Hill, S.M. (2020) High-Dimensional Regression in Practice: An Empirical Study of Finite-Sample Prediction, Variable Selection and Ranking. *Statistics and Computing*, **30**, 697-719. <https://doi.org/10.1007/s11222-019-09914-9>
- [21] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition, Springer.
- [22] Chen, Y. and Yang, Y. (2021) The One Standard Error Rule for Model Selection: Does It Work? *Stats*, **4**, 868-892. <https://doi.org/10.3390/stats4040051>
- [23] Yates, L.A., Aandahl, Z., Richards, S.A. and Brook, B.W. (2022) Cross Validation for Model Selection: A Review with Examples from Ecology. *Ecological Monographs*, **93**, e1557. <https://doi.org/10.1002/ecm.1557>
- [24] Meinshausen, N. and Bühlmann, P. (2010) Stability Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **72**, 417-473. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>
- [25] Sutradhar, B.C. and Rao, R.P. (1996) On Joint Estimation of Regression and Overdispersion Parameters in Generalized Linear Models for Longitudinal Data. *Journal of Multivariate Analysis*, **56**, 90-119. <https://doi.org/10.1006/jmva.1996.0006>
- [26] Paul, S.R. and Islam, A.S. (1998) Joint Estimation of the Mean and Dispersion Parameters in the Analysis of Proportions: A Comparison of Efficiency and Bias. *Canadian Journal of Statistics*, **26**, 83-94. <https://doi.org/10.2307/3315675>
- [27] Breheny, P. and Huang, J. (2011) Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection. *The Annals of Applied Statistics*, **5**, 232-253. <https://doi.org/10.1214/10-aos388>
- [28] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) Classification and Regression Trees. Wadsworth.
- [29] Zhang, C.H. (2010) Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics*, **38**, 894-942. <https://doi.org/10.1214/09-aos729>
- [30] Ternès, N., Rotolo, F. and Michiels, S. (2016) Empirical Extensions of the Lasso Penalty to Reduce the False Discovery Rate in High-Dimensional Cox Regression Models. *Statistics in Medicine*, **35**, 2561-2573. <https://doi.org/10.1002/sim.6927>
- [31] Fletcher, D., Dillingham, P.W. and Zeng, J. (2022) Model-Averaged Confidence Intervals for Factorial Experiments and Generalized Linear Models. *Environmental and Ecological Statistics*, **29**, 655-678.
- [32] Prataviera, F., Cordeiro, G.M., Ortega, E.M.M. and Suzuki, A.K. (2021) A New Regression Model for Rates and Proportions Data with Applications. *Journal of Applied Statistics*, **49**, 4137-4161. <https://doi.org/10.1080/02664763.2021.1973385>
- [33] DiCiccio, T.J., Monti, A.C. and Young, G.A. (2006) Variance Stabilization for a Scalar Parameter. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **68**, 281-303. <https://doi.org/10.1111/j.1467-9868.2006.00544.x>
- [34] DelosReyes, J., et al. (2023) Bootstrap Correlation Confidence Interval Estimation: The Positive Impact of a Symmetric Distribution. *The Journal of Experimental Education*, **92**, 559-579. <https://doi.org/10.1080/00220973.2023.2196659>
- [35] Godambe, V.P. (1960) An Optimum Property of Regular Maximum Likelihood Estimation. *The Annals of Mathematical Statistics*, **31**, 1208-1211. <https://doi.org/10.1214/aoms/1177705693>
- [36] Heyde, C.C. (1997) Quasi-Likelihood and Its Application: A General Approach to

Optimal Parameter Estimation. Springer.

- [37] Mancl, L.A. and DeRouen, T.A. (2001) A Covariance Estimator for GEE with Improved Small-Sample Properties. *Biometrics*, **57**, 126-134.
<https://doi.org/10.1111/j.0006-341x.2001.00126.x>
- [38] Kauermann, G. and Carroll, R.J. (2001) A Note on the Efficiency of Sandwich Covariance Matrix Estimation. *Journal of the American Statistical Association*, **96**, 1387-1396. <https://doi.org/10.1198/016214501753382309>
- [39] Fay, M.P. and Graubard, B.I. (2001) Small-Sample Adjustments for Wald-Type Tests Using Sandwich Estimators. *Biometrics*, **57**, 1198-1206.
<https://doi.org/10.1111/j.0006-341x.2001.01198.x>
- [40] van de Geer, S., Bühlmann, P. and Zhou, S. (2011) The Adaptive and the Thresholded Lasso for Potentially Misspecified Models (and a Lower Bound for the Lasso). *Electronic Journal of Statistics*, **5**, 688-749. <https://doi.org/10.1214/11-ejs624>
- [41] Hilbe, J.M. (2011) Negative Binomial Regression. 2nd Edition, Cambridge University Press. <https://doi.org/10.1017/cbo9780511973420>
- [42] Saha, K.K. and Paul, S.R. (2011) Interval Estimation of the Negative Binomial Dispersion Parameter. *Journal of Statistical Computation and Simulation*, **81**, 1993-2002.
<https://doi.org/10.1080/00949655.2010.512998>
- [43] Šimkovic, M. and Träuble, B. (2019) Robustness of Statistical Methods When Measure Is Affected by Ceiling and/or Floor Effect. *PLOS ONE*, **14**, e0220889.
<https://doi.org/10.1371/journal.pone.0220889>