Refined Debiased Lasso for High-Dimensional Generalized Linear Models

Presented By
John Darden & Kaniz Fatema

STAT718-001-FALL-2025
Statistical Learning for Big Data
University of South Carolina

Outline

- Introduction & Motivation
- Methodology: Refined debiased lasso
- Simulation studies
- Application: Twitch Data and Liver RNA Data
- Discussion, Limitations, Conclusion

Introduction & Motivation

- High-dimensional data common in genetics and epidemiology; many covariates ($p > n$).
- When the number of covariates is large and grows with the sample size, standard inference methods for GLMs become unreliable.
- Traditional approaches may suffer from instability, large bias, or non-existent estimates in these settings.
- This motivates the development of a more stable and accurate debiasing method for high-dimensional GLMs.

Limitations of Existing Methods

- Maximum Likelihood Estimation (MLE) can produce unstable or extremely biased estimates when p is moderately large.
- Existing debiased lasso methods attempt to correct bias. **Original debiased lasso** method assumes sparsity in the inverse information matrix, an assumption that commonly fails in GLMs, leading to poor bias correction and inaccurate confidence intervals.

Proposed Refined Debiased Lasso

- Introduces a refined debiased lasso estimator that directly inverts the Hessian matrix, avoiding restrictive sparsity assumptions.
- Designed specifically for “large n, diverging p” settings, improving stability and reducing estimation bias.

Mathematical Notation

Vector and matrix norms

- For a vector $\mathbf{a}$, write $\|\mathbf{a}\|_q$ for its ℓ_q norm ($q \geq 1$).
- For a real matrix $A = (A_{ij})$:

$$\|A\| = \sup_{\|x\|_2=1} \|Ax\|_2 = \left(\lambda_{\max}(A^T A)\right)^{1/2}$$

(spectral norm),

$$\|A\|_1 = \max_j \sum_i |A_{ij}|, \qquad \|A\|_\infty = \max_{i,j} |A_{ij}|.$$

Asymptotic notation

$$a_n = O(b_n) \quad \text{(bounded by constant} \times b_n\text{)}$$
$$a_n = o(b_n) \quad \text{(much smaller than } b_n\text{)},$$

,

$$a_n = o(b_n) \quad \text{if } \frac{a_n}{b_n} \to 0 \text{ as } n \to \infty.$$

$$a_n \asymp b_n \quad \text{when} \quad a_n = O(b_n) \quad \text{and} \quad b_n = O(a_n).$$

Data structure

- Observations: (y_i, x_i) for $i = 1, \ldots, n$, assumed i.i.d. copies of (y, x).
- Covariate vector:

$$x_i = (1, \tilde{x}_i^T)^T \in \mathbb{R}^{p+1},$$

  where the leading 1 is the intercept and $\tilde{x}_i \in \mathbb{R}^p$ are the covariates.

GLM / exponential family (negative log-likelihood)

$$\rho_\xi(y, x) \;=\; \rho(y, x^T\xi) \;=\; -y x^T \xi + b(x^T \xi)$$

where:

- $b(\cdot)$ is a known, twice continuously differentiable function (ensures smoothness),
- $\xi = (\beta_0, \beta^T)^T \in \mathbb{R}^{p+1}$ denotes the parameter vector (intercept β_0 plus coefficients β)

Debiased Lasso: Mathematical Components

Empirical loss and derivatives:

$$P_n g = \frac{1}{n}\sum_{i=1}^{n} g(y_i, x_i),$$

$$P_n \rho_\xi = \frac{1}{n}\sum_{i=1}^{n} \rho_\xi(y_i, x_i), \quad P_n \dot{\rho}_\xi = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \rho_\xi(y_i, x_i)}{\partial \xi},$$

$$\widehat{\Sigma}_\xi = P_n \ddot{\rho}_\xi = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \rho_\xi(y_i, x_i)}{\partial \xi \partial \xi^T}.$$

Lasso estimator:

$$\hat{\xi} = \arg \min_{\xi=(\beta_0, \beta^T)^T} \left\{ P_n \rho_\xi + \lambda \|\beta\|_1 \right\}.$$

Taylor expansion of the score at $\hat{\xi}$:

$$P_n \dot{\rho}_{\xi^0} = P_n \dot{\rho}_{\hat{\xi}} + P_n \ddot{\rho}_{\hat{\xi}} (\xi^0 - \hat{\xi}) + \Delta,$$

Why do we need Taylor expansion and decomposition?

- The lasso estimator $\hat{\xi}$ is **biased** because the ℓ_1 penalty shrinks coefficients.Biased estimates mean we **cannot trust confidence intervals or p-values**.
- To correct this, we need to understand **how** $\hat{\xi}$ differs from the true parameter ξ^0.
- The GLM likelihood is nonlinear, so we apply a **Taylor expansion** to linearize the score function. Multiplying this expansion by a suitable matrix isolates the **bias components**.
- One component (I_j) can be estimated and removed. Others (II_j, III_j) are controlled under assumptions.

Derivation of the Bias Decomposition

$$P_n\dot{\rho}_{\xi^0} = P_n\dot{\rho}_{\hat{\xi}} + P_n\ddot{\rho}_{\hat{\xi}}(\xi^0 - \hat{\xi}) + \Delta.$$

$$P_n\ddot{\rho}_{\hat{\xi}}(\xi^0 - \hat{\xi}) = P_n\dot{\rho}_{\xi^0} - P_n\dot{\rho}_{\hat{\xi}} - \Delta.$$

$$M_j P_n\ddot{\rho}_{\hat{\xi}}(\xi^0 - \hat{\xi}) = M_j P_n\dot{\rho}_{\xi^0} - M_j P_n\dot{\rho}_{\hat{\xi}} - M_j\Delta.$$

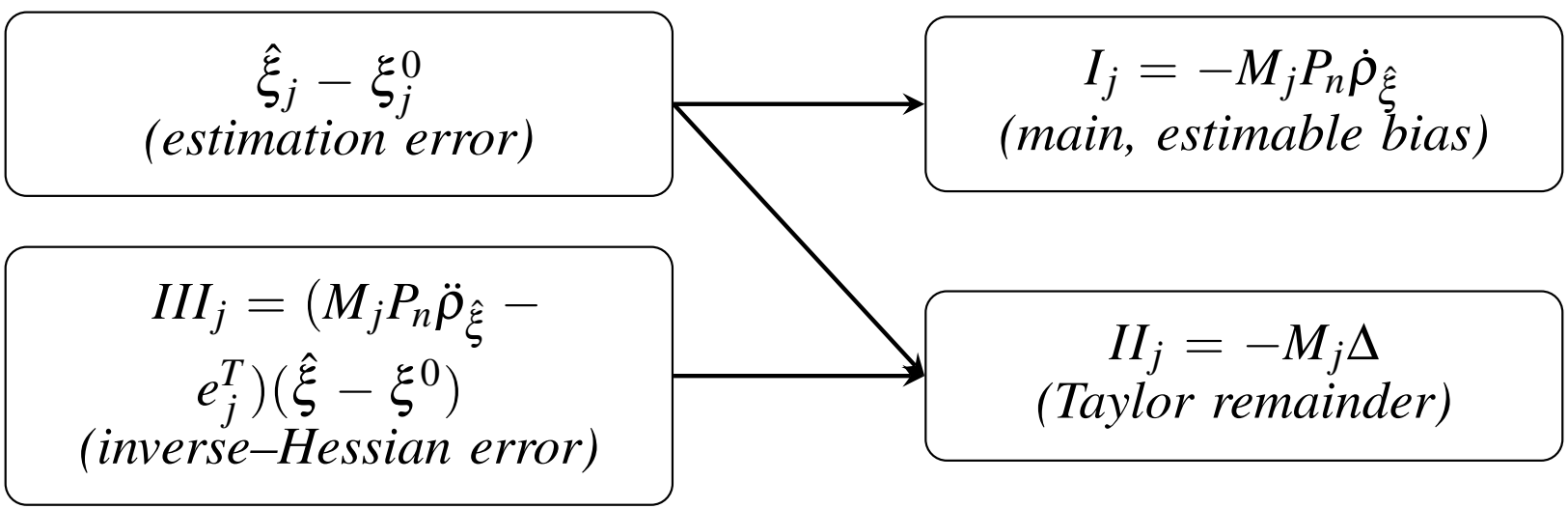
$$e_j^T(\hat{\xi} - \xi^0) + \left(M_j P_n\ddot{\rho}_{\hat{\xi}} - e_j^T\right)(\hat{\xi} - \xi^0) = -M_j P_n\dot{\rho}_{\xi^0} + M_j P_n\dot{\rho}_{\hat{\xi}} + M_j\Delta.$$

$$\boxed{\hat{\xi}_j - \xi_j^0 + (-M_j P_n\dot{\rho}_{\hat{\xi}}) + (-M_j\Delta) + \left(M_j P_n\ddot{\rho}_{\hat{\xi}} - e_j^T\right)(\hat{\xi} - \xi^0) = -M_j P_n\dot{\rho}_{\xi^0}.}$$

$$\boxed{\hat{\xi}_j - \xi_j^0 + I_j + II_j + III_j = -M_j P_n\dot{\rho}_{\xi^0}.}$$

Decomposition of the Bias Terms

$$\hat{\xi}_j - \xi_j^0 + I_j + II_j + III_j = -M_j P_n \dot{\rho}_{\xi^0}.$$



Refined debiased lasso: choose $M = \widehat{\Sigma}_{\hat{\xi}}^{-1}$ so $III_j \approx 0$.

Deriving the Refined Debiased Estimator

$$P_n \ddot{\rho}_{\hat{\xi}} (\xi^0 - \hat{\xi}) = P_n \dot{\rho}_{\xi^0} - P_n \dot{\rho}_{\hat{\xi}} - \Delta.$$

$\widehat{\Theta}_{\hat{\xi}} = \widehat{\Sigma}_{\hat{\xi}}^{-1}$:

$$\xi^0 - \hat{\xi} = \widehat{\Theta}_{\hat{\xi}} \left(P_n \dot{\rho}_{\xi^0} - P_n \dot{\rho}_{\hat{\xi}} - \Delta \right),$$

using $\widehat{\Theta}_{\hat{\xi}} \widehat{\Sigma}_{\hat{\xi}} = I$.

$$\hat{\xi} - \widehat{\Theta}_{\hat{\xi}} P_n \dot{\rho}_{\hat{\xi}} - \xi^0 = -\widehat{\Theta}_{\hat{\xi}} P_n \dot{\rho}_{\xi^0} + \widehat{\Theta}_{\hat{\xi}} \Delta.$$

$$\boxed{\hat{b} = \hat{\xi} - \widehat{\Theta}_{\hat{\xi}} P_n \dot{\rho}_{\hat{\xi}}}$$

$$\hat{b} - \xi^0 = -\widehat{\Theta}_{\hat{\xi}} P_n \dot{\rho}_{\xi^0} + \widehat{\Theta}_{\hat{\xi}} \Delta.$$

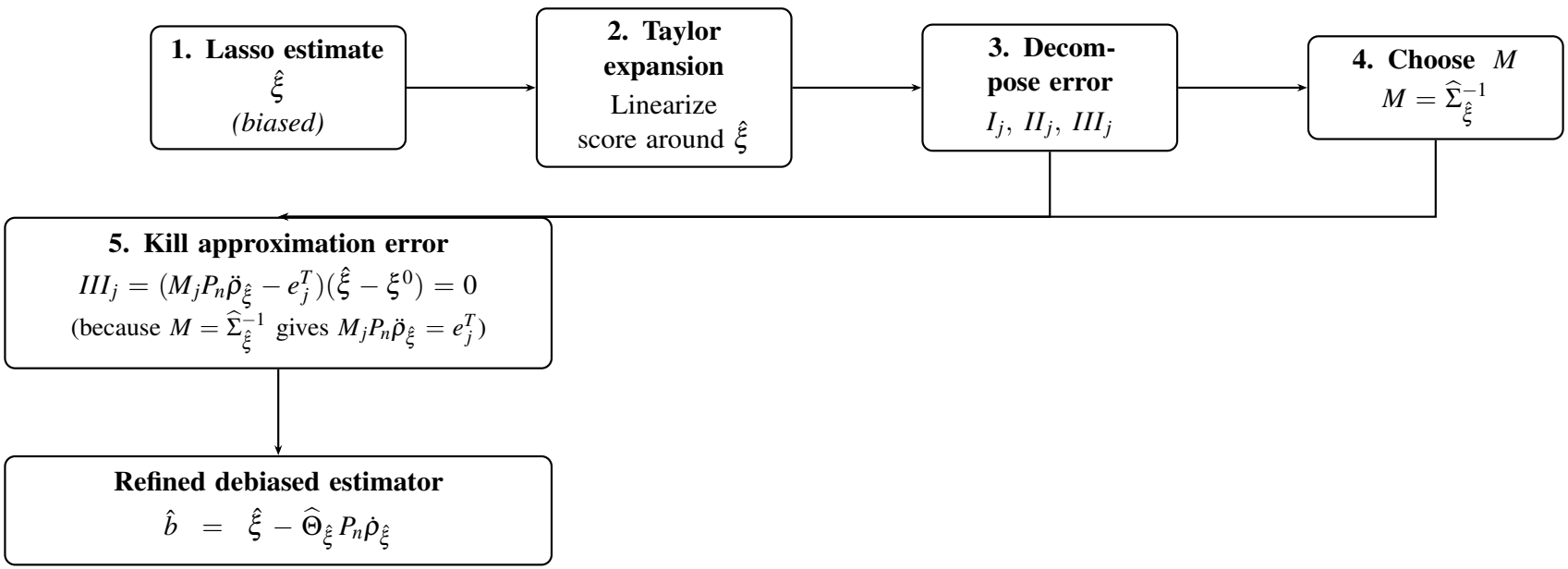
- $-\widehat{\Theta}_{\hat{\xi}} P_n \dot{\rho}_{\xi^0}$ is *mean-zero sampling noise*.
- $\widehat{\Theta}_{\hat{\xi}} \Delta$ is the Taylor remainder; GLM smoothness and $\hat{\xi} \to \xi^0$ imply $\Delta = o_p(n^{-1/2})$.

Asymptotic normality:

$$\sqrt{n}(\hat{b} - \xi^0) \xrightarrow{d} N\left(0,\ \widehat{\Theta}_{\hat{\xi}} V \widehat{\Theta}_{\hat{\xi}}^T\right),$$

where $V = \mathrm{Var}(\dot{\rho}_{\xi^0})$.

Flowchart: How the *Refined* Debiased Lasso Works



Assumptions for the Refined Debiased Lasso

Assumption 1 (Bounded & sub-Gaussian covariates).

$$\|X\|_\infty \le K \quad \text{a.s.}, \qquad X_i \text{ is sub-Gaussian.}$$

Assumption 2 (Well-conditioned information matrix).

$$0 < c_{\min} \le \lambda_{\min}(\Sigma_{\xi 0}) \le \lambda_{\max}(\Sigma_{\xi 0}) \le c_{\max} < \infty.$$

Assumption 3 (Smooth GLM loss). First and second derivatives exist:

$$\dot{\rho}(y,a) = \frac{\partial}{\partial a}\rho(y,a), \qquad \ddot{\rho}(y,a) = \frac{\partial^2}{\partial a^2}\rho(y,a).$$

Assumption 4 (Bounded linear predictor).

$$\|X\xi^0\|_\infty \le C \quad \text{a.s.}$$

Assumption 5 (Well-conditioned covariance).

$$0 < k_{\min} \le \lambda_{\min}\big(E(X^T X/n)\big) \le \lambda_{\max}\big(E(X^T X/n)\big) \le k_{\max} < \infty.$$

Dimensionality requirements. With s_0 = number of nonzero coefficients,

$$\boxed{p^2/n \to 0, \qquad s_0 \log(p)\sqrt{p/n} \to 0.}$$

Simulation Study

- Compare MLE, Original Debiased Lasso, Refined Debiased Lasso in terms of coefficient estimation bias, coverage probability, empirical standard error, and model based standard error.
- Take n = 1000. Use p = 40 and p = 100. Recall p is the dimension of the covariates.
- Covariates independently generated from

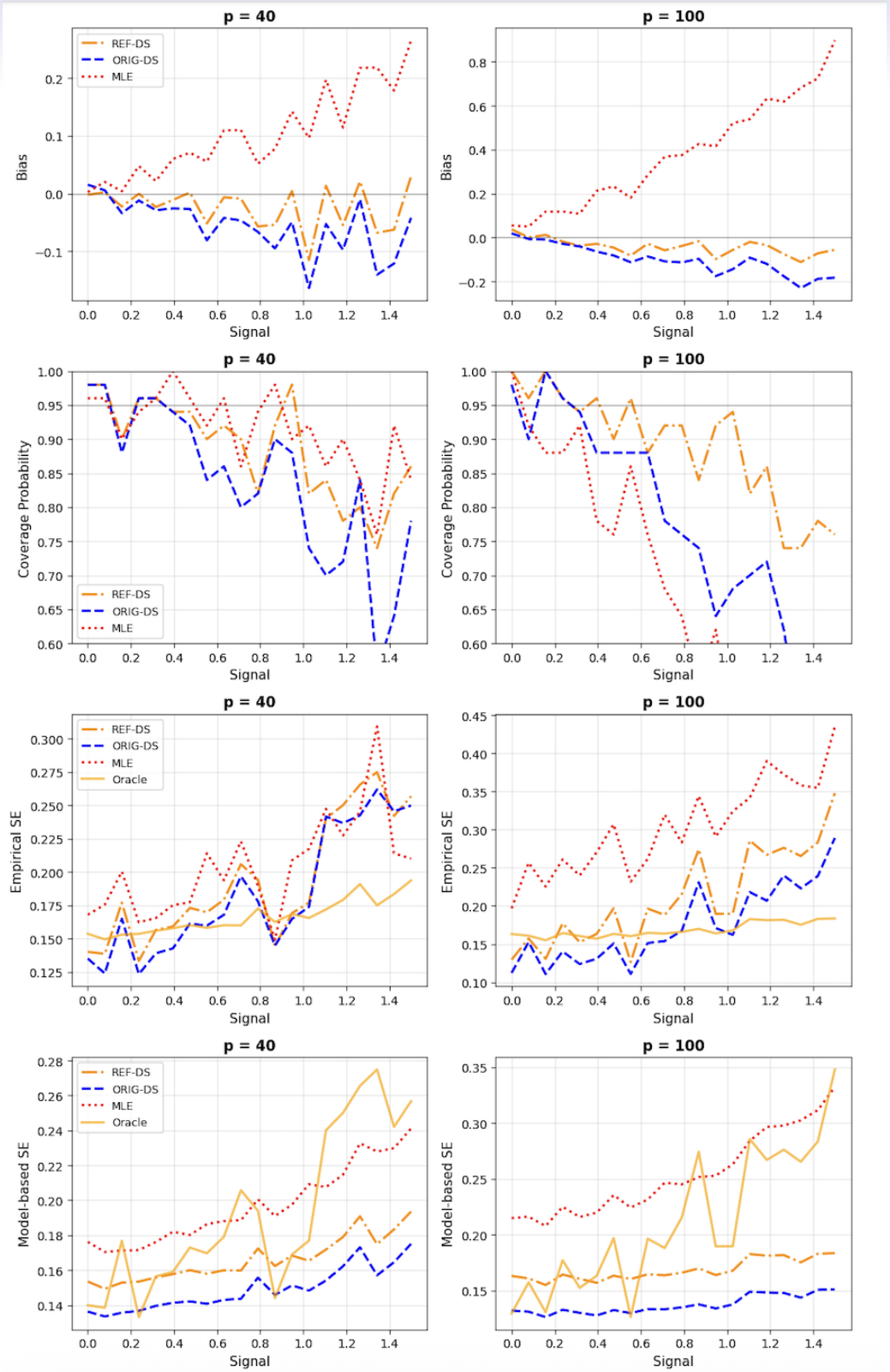
$$N_p\big(0, \Sigma_x\big)$$

- Responses generated from

$$Bernoulli(\mu_i)$$

where

$$\mu_i = exp(x_i^T \xi^0)/1 + exp(x_i^T \xi^0)$$


p = 40
p = 100
REF-DS
ORIG-DS
MLE
Oracle
Bias
Coverage Probability
Empirical SE
Model-based SE
Signal

Real Data Application

- Human Liver RNA data set with 36,000 genes, approximately 1,000 samples.
- Feature selection to remove highly correlated genes before implementation of refined debiased lasso.
- Twitch streaming data set with 116 covariates and approximately 30 observations.
- The application to both data sets shows the versatility of this method with varying numbers of covariates in practice, as well as sample size.

Results: Twitch Data

Predictor	Estimate	SE	p-value	CI Low	CI High
”predictor”				”CI$_{l}ow$”	”CI$_{h}igh$”
”(Intercept)”	2.156	0.062	1.570	2.035	2.278
”n$_{v}ideos_{y}t$”	–0.006	0.098	0.955	–0.198	0.187
”w$_{hw}atched$”	0.011	0.092	0.902	–0.169	0.191
”cuncurrent$_{r}ecord_{t}w$”	0.053	0.136	0.696	–0.213	0.319
”mst$_{c}at_{a}vg_{wc}hannels$”	–0.015	0.085	0.863	–0.181	0.151
”avg$_{f}oll_{g}ained_{p}er_{s}tream_{t}w$”	0.071	0.130	0.587	–0.184	0.326
”female$_{f}lag$”	–0.005	0.082	0.955	–0.165	0.156
”X$_{1}$2.21$_{f}oll_{b}alance$”	–0.046	0.162	0.777	–0.364	0.272
”X$_{0}$1.21$_{f}oll_{b}alance$”	–0.056	0.148	0.707	–0.345	0.234
”X$_{1}$2.21$_{v}iews$”	0.068	0.159	0.671	–0.245	0.380
”X$_{0}$5.20$_{v}iews$”	0.046	0.093	0.625	–0.137	0.228

Results: Liver RNA Data

Variable	REF-DS Est	REF-DS SE	REF-DS 95% CI	ORIG-DS Est	ORIG-DS SE	ORIG-DS 95% CI	MLE Est	MLE SE	MLE 95% CI
Intercept	10.39	0.32	(9.76, 11.02)	13.45	0.91	(11.66, 15.24)	406.53	391.46	(-360.72, 1173.79)
ORM1	5.01	0.24	(4.54, 5.48)	17.65	3.36	(11.06, 24.24)	54.81	201.81	(-340.74, 450.36)
FGG	4.27	0.34	(3.61, 4.93)	14.36	3.14	(8.20, 20.52)	32.90	260.86	(-478.38, 544.18)
FTL	4.20	0.07	(4.06, 4.34)	4.23	0.26	(3.73, 4.74)	-3.88	75.43	(-151.73, 143.96)
GAPDH	-3.71	0.08	(-3.88, -3.54)	-3.76	0.27	(-4.28, -3.23)	-5.17	82.62	(-167.10, 156.77)
FGB	3.55	0.34	(2.88, 4.22)	14.20	3.17	(7.99, 20.41)	-126.90	228.91	(-575.57, 321.77)
SERPINA1	3.47	0.58	(2.35, 4.60)	18.45	2.78	(13.01, 23.90)	211.61	377.95	(-529.18, 952.40)
HP	2.96	0.34	(2.29, 3.64)	15.15	2.12	(10.99, 19.31)	80.26	259.01	(-427.40, 587.92)
MT-CYB	2.76	0.33	(2.12, 3.40)	6.58	1.64	(3.36, 9.79)	4.12	136.04	(-262.52, 270.75)
FGL1	2.40	0.20	(2.01, 2.79)	2.49	0.34	(1.83, 3.15)	-43.48	115.50	(-269.87, 182.90)
HBG2	-2.23	0.09	(-2.40, -2.06)	-2.41	0.58	(-3.56, -1.27)	32.12	114.14	(-191.60, 255.84)
AHSG	2.19	0.54	(1.14, 3.25)	2.21	0.25	(1.71, 2.71)	216.16	360.17	(-489.77, 922.08)
APOH	2.06	0.30	(1.47, 2.66)	2.10	0.24	(1.63, 2.57)	59.89	235.45	(-401.59, 521.38)
VTN	1.96	0.31	(1.36, 2.57)	2.00	0.24	(1.54, 2.47)	166.30	288.47	(-399.11, 731.70)
PLG	-1.93	0.70	(-3.30, -0.55)	-2.00	0.24	(-2.47, -1.53)	-100.07	437.18	(-956.95, 756.80)
CP	1.83	0.26	(1.32, 2.34)	1.88	0.27	(1.35, 2.40)	104.21	161.41	(-212.15, 420.58)

Table: Comparison of REF-DS, ORIG-DS, and MLE estimates with standard errors and 95% confidence intervals.

Discussion

- The refined debiased lasso achieves **substantial bias reduction** compared to both the original debiased lasso and MLE in high-dimensional GLMs.
- Directly inverting the empirical Hessian removes the major approximation error term, leading to **more accurate and stable inference**.
- Simulations show the refined method maintains **near-nominal confidence interval coverage** across a broad range of p/n ratios.
- Overall, the approach provides a **practical and theoretically justified** solution for inference in "large n, diverging p" GLMs.

Limitations & Future Work

- **Invertibility of the Hessian.** The refined method requires $\widehat{\Sigma}_{\hat{\xi}}$ to be invertible, which may fail when p becomes too large or covariates are highly collinear.
- **Taylor remainder.** Although $II_j = -M_j\Delta$ becomes small asymptotically, the remainder may be non-negligible in small samples, especially with highly nonlinear GLMs.
- **Computational cost.** Inverting the $(p+1)\times(p+1)$ Hessian matrix increases computational burden compared to the sparse nodewise lasso approach.
- **Applicability for $p > n$.** The method is not recommended in ultra–high-dimensional settings ($p \gg n$), where $\widehat{\Sigma}_{\hat{\xi}}$ may be nearly singular.