Statistical Inference for High-dimensional Logistic Regression

Zijian Guo

Department of Statistics, Rutgers University

CMStatistics 2021,London, UK

Joint with Prabrisha Rakshit, Daniel Herman and Jinbo Chen.

Motivation

Electronic health record (EHR)

EHR: Document for medical billing

1. Limited labeled data: 50 to 1000
2. Binary outcome (e.g. disease status.)
3. Many predictors: billing codes, demographics, disease histories, co-morbid conditions, laboratory test results, prescription codes, and concepts extracted from doctors' notes.

EHR phenotyping

High-dimensional inference for binary outcome labeling.

High-dimensional Logistic Regression

For $1 \leq i \leq n$, consider the model for $y_i \in \{0, 1\}$,$X_{i\cdot} \in \mathbb{R}^p$

$$\mathbb{P}(y_i = 1 | X_{i\cdot}) = h(X_{i\cdot}^{\intercal} \beta), \quad h(z) = \exp(z)/[1 + \exp(z)]$$

- $p \gg n$, β is sparse
- Case probability

$$\mathbb{P}(y_i = 1 | X_{i\cdot} = x_{\rm new}) \equiv h\left(x_{\rm new}^{\intercal} \beta\right)$$

EHR phenotyping

$$H_0 : h(x_{\rm new}^{\intercal} \beta) < 1/2.$$

Existing Bias Correction

The penalized log-likelihood estimator $\widehat{\beta}$

$$\widehat{\beta} = \arg\min_{\beta} \ell(\beta) + \lambda \|\beta\|_1,$$

where $\ell(\beta) = \sum_{i=1}^{n} \left[\log\left(1 + \exp\left(X_{i\cdot}^{\intercal}\beta\right)\right) - y_i \cdot \left(X_{i\cdot}^{\intercal}\beta\right)\right].$

Debiasing Inference

- β_j in linear models (Zhang & Zhang '14, Javanmard & Montanari '14)
- β_j in GLM (van de Geer, Bühlmann, Ritov & Dezeure '14)

Bias Correction Intuition

$$\widehat{\beta}_j + \widehat{u}^\intercal \frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}(y_i - h(X_{i\cdot}^\intercal \widehat{\beta})) \tag{1}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}(y_i - h(X_{i\cdot}^\intercal \widehat{\beta})) \approx \widehat{H}(\widehat{\beta})(\beta - \widehat{\beta}) + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i X_{i\cdot},$$

with $\widehat{H}(\beta) = \frac{1}{n}\sum_{i=1}^{n} h(X_{i\cdot}^\intercal \beta)(1 - h(X_{i\cdot}^\intercal \beta))X_{i\cdot}X_{i\cdot}^\intercal$.

Bias Correction Intuition

$$\widehat{\beta}_j + \widehat{u}^{\intercal} \frac{1}{n} \sum_{i=1}^{n} X_{i\cdot}(y_i - h(X_{i\cdot}^{\intercal} \widehat{\beta})) \tag{1}$$

$$\frac{1}{n} \sum_{i=1}^{n} X_{i\cdot}(y_i - h(X_{i\cdot}^{\intercal} \widehat{\beta})) \approx \widehat{H}(\widehat{\beta})(\beta - \widehat{\beta}) + \frac{1}{n} \sum_{i=1}^{n} \epsilon_i X_{i\cdot},$$

with $\widehat{H}(\beta) = \frac{1}{n} \sum_{i=1}^{n} h(X_{i\cdot}^{\intercal} \beta)(1 - h(X_{i\cdot}^{\intercal} \beta)) X_{i\cdot} X_{i\cdot}^{\intercal}$.

$$\begin{aligned} \widehat{u}^{\intercal} \frac{1}{n} \sum_{i=1}^{n} X_{i\cdot}(y_i - h(X_{i\cdot}^{\intercal} \widehat{\beta})) &\approx \widehat{u}^{\intercal} \widehat{H}(\widehat{\beta})(\beta - \widehat{\beta}) \\ &\approx e_j^{\intercal}(\beta - \widehat{\beta}). \end{aligned}$$

Challenge

For $\beta_j = e_j^{\intercal}\beta$,

$$\widehat{H}(\widehat{\beta})\widehat{u} \approx e_j$$

- Sparse $[\mathbb{E}\widehat{H}(\beta)]^{-1}e_j$ (van de Geer et.al., 14)

For $x_{\rm new}^{\intercal}\beta$, we construct $\widehat{u}$ such that

$$\widehat{H}(\widehat{\beta})\widehat{u} \approx x_{\rm new}.$$

Challenge

$[\widehat{H}(\widehat{\beta})]^{-1}x_{\rm new}$ can be **DENSE**!

Our Proposed Method

LIVE estimator

Existing

$$\widehat{\beta}_j + \widehat{u}^{\intercal} \frac{1}{n} \sum_{i=1}^{n} X_{i\cdot}(y_i - h(X_{i\cdot}^{\intercal}\widehat{\beta}))$$

LInearization and Variance Enhancement

$$\widehat{x_{\mathrm{new}}^{\intercal}\beta} = x_{\mathrm{new}}^{\intercal}\widehat{\beta} + \widehat{u}^{\intercal} \frac{1}{n} \sum_{i=1}^{n} \underbrace{[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1 - h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}}_{\text{weight for } i-th \text{ observation}} X_{i\cdot}(y_i - h(X_{i\cdot}^{\intercal}\widehat{\beta})).$$

Linearization: Logistic to Linear

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}X_{i\cdot}(y_i-h(X_{i\cdot}^{\intercal}\widehat{\beta}))\\
\approx&\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}X_{i\cdot}^{\intercal}(\beta-\widehat{\beta})+\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}\epsilon_i X_{i\cdot}
\end{aligned}
$$

Linearization: Logistic to Linear

$$\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}X_{i\cdot}(y_i-h(X_{i\cdot}^{\intercal}\widehat{\beta}))$$

$$\approx\frac{1}{n}\sum_{i=1}^{n}X_{i\cdot}X_{i\cdot}^{\intercal}(\beta-\widehat{\beta})+\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}\epsilon_i X_{i\cdot}$$

$\widehat{x_{\rm new}^{\intercal}\beta}-x_{\rm new}^{\intercal}\beta$ is decomposed as

$$(\widehat{\Sigma}\widehat{u}-x_{\rm new})^{\intercal}(\beta-\widehat{\beta})+\widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}\epsilon_i X_{i\cdot}.$$

Variance Enhancement: Uniform for x_{new}

$$(\widehat{\Sigma}\widehat{u} - x_{\text{new}})^{\intercal}(\beta - \widehat{\beta}) + \widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1 - h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}\epsilon_i X_{i\cdot}.$$

$$\widehat{\Sigma}\widehat{u} \approx x_{\text{new}}$$

Variance Enhancement: Uniform for $x_{\rm new}$

$$(\widehat{\Sigma}\widehat{u}-x_{\rm new})^{\intercal}(\beta-\widehat{\beta})+\widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}\epsilon_i X_{i\cdot}.$$

$$\widehat{\Sigma}\widehat{u}\approx x_{\rm new}$$

Variance enhancement projection direction.

$$\widehat{u}=\arg\min_{u\in\mathbb{R}^p} u^{\intercal}\widehat{\Sigma}u$$

$$\text{subject to}\quad \|\widehat{\Sigma}u-x_{\rm new}\|_\infty\le\|x_{\rm new}\|_2\lambda_n$$

$$|x_{\rm new}^{\intercal}\widehat{\Sigma}u-\|x_{\rm new}\|_2^2|\le\|x_{\rm new}\|_2^2\lambda_n$$

where $\lambda_n \asymp (\log p/n)^{1/2}$.

What if no additional constraint?

$$\widehat{u} = \arg\min_{u \in \mathbb{R}^p} u^{\intercal} \widehat{\Sigma} u$$

$$\text{subject to} \quad \|\widehat{\Sigma} u - x_{\text{new}}\|_{\infty} \leq \|x_{\text{new}}\|_2 \lambda_n$$

What if no additional constraint?

$$\widehat{u} = \arg\min_{u \in \mathbb{R}^p} u^\intercal \widehat{\Sigma} u$$

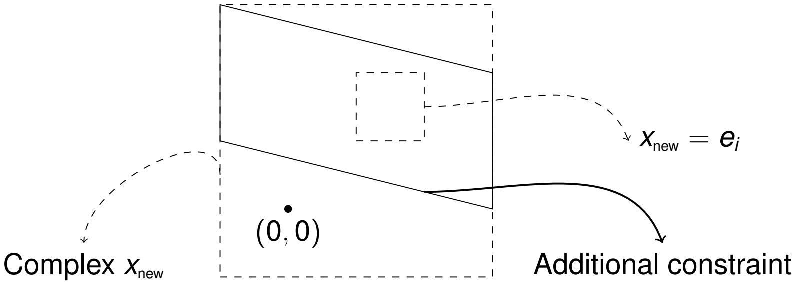
$$\text{subject to} \quad \|\widehat{\Sigma} u - x_{\text{new}}\|_\infty \leq \|x_{\text{new}}\|_2 \lambda_n$$

For a dense x_{new}: there is no bias correction,

$$\|x_{\text{new}}\|_\infty \leq \|x_{\text{new}}\|_2 \lambda_n \Rightarrow \widehat{u} = 0!$$

Curse of dimensionality: too much flexibility of searching the direction.

Additional Constraint and Feasible Set



- Large dashed: dense x_{new} without additional constraint.
- Solid parallelogram: dense x_{new} with additional constraint.

$$\left| x_{\text{new}}^{\intercal} \widehat{\Sigma} u - \|x_{\text{new}}\|_2^2 \right| \leq \|x_{\text{new}}\|_2^2 \lambda$$

LIVE Estimator

$$\widehat{x_{\text{new}}^{\intercal}\beta} = x_{\text{new}}^{\intercal}\widehat{\beta} + \widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}\underbrace{[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}}_{\text{weight for } i-th \text{ observation}} X_{i\cdot}(y_i - h(X_{i\cdot}^{\intercal}\widehat{\beta})).$$

with the projection direction $\widehat{u}$ defined as

$$\widehat{u} = \arg\min_{u \in \mathbb{R}^p} u^{\intercal}\widehat{\Sigma}u$$

$$\text{subject to} \quad \|\widehat{\Sigma}u - x_{\text{new}}\|_{\infty} \leq \|x_{\text{new}}\|_2 \lambda_n$$

$$|x_{\text{new}}^{\intercal}\widehat{\Sigma}u - \|x_{\text{new}}\|_2^2| \leq \|x_{\text{new}}\|_2^2 \lambda_n$$

where $\lambda_n \asymp (\log p/n)^{1/2}$.

Statistical Inference

We construct the CI for $\mathbb{P}(y_i = 1 | X_{i\cdot} = x_{\text{new}})$ as,

$$\mathrm{CI}_\alpha(x_{\text{new}}) = \left[h\left(\widehat{x_{\text{new}}^\intercal\beta} - z_{\alpha/2}\widehat{\mathrm{V}}^{1/2}\right), h\left(\widehat{x_{\text{new}}^\intercal\beta} + z_{\alpha/2}\widehat{\mathrm{V}}^{1/2}\right)\right],$$

with

$$\widehat{\mathrm{V}} = \widehat{u}^\intercal \left[\frac{1}{n^2}\sum_{i=1}^{n}[h(X_{i\cdot}^\intercal\widehat{\beta})(1 - h(X_{i\cdot}^\intercal\widehat{\beta}))]^{-1} X_{i\cdot}X_{i\cdot}^\intercal\right]\widehat{u}.$$

EHR phenotyping

$$\phi_\alpha(x_{\text{new}}) = \mathbf{1}\left(\widehat{x_{\text{new}}^\intercal\beta} - z_\alpha\widehat{\mathrm{V}}^{1/2} \geq 0\right).$$

Theory and Optimality

Theoretical Justification

Theorem 1.

Under regularity conditions, if

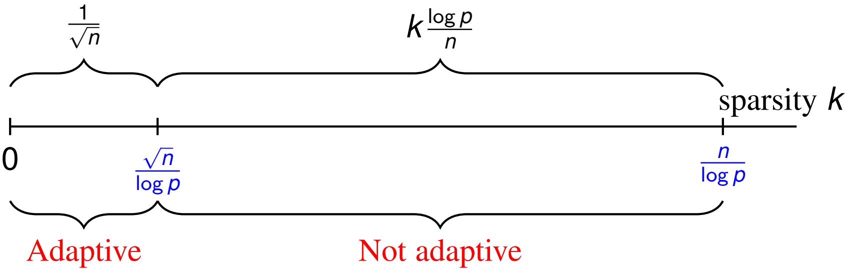
$$k \ll \sqrt{n}/[\log p (\log n)^{1/2}],$$

then

$$\mathbb{P}\left[\mathrm{V}^{-1/2}\left(\widehat{x_{\mathrm{new}}^{\intercal}\beta} - x_{\mathrm{new}}^{\intercal}\beta\right) \geq z_{\alpha}\right] \to \alpha.$$

- No sparsity on Σ^{-1} and x_{new}.
- Approximate $\widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\widehat{\beta})(1-h(X_{i\cdot}^{\intercal}\widehat{\beta}))]^{-1}X_{i\cdot}\epsilon_i$ by $\widehat{u}^{\intercal}\frac{1}{n}\sum_{i=1}^{n}[h(X_{i\cdot}^{\intercal}\beta)(1-h(X_{i\cdot}^{\intercal}\beta))]^{-1}X_{i\cdot}\epsilon_i$
- Contraction Principle.

Discussion: Optimality of CI for β_j



Cai, T. Tony, Zijian Guo, and Rong Ma. "Statistical inference for high-dimensional generalized linear models with binary outcomes." JASA, to appear.

Numerical Results

Simulation Studies

1. $p = 501$
2. $n \in \{200, 400, 600\}$
3. $\beta_1 = 0$, $\beta_j = (j-1)/20$ for $2 \leq j \leq 11$ and $\beta_j = 0$ for $12 \leq j \leq p$

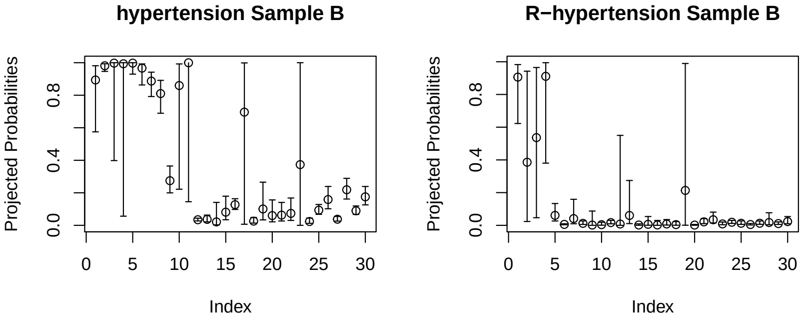
Simulation Studies

		LiVE				Post Selection				`hdi`				`WLDP`			
$\lVert x_{\text{new}} \rVert_2$	n	Cov	ERR	Len	t	Cov	ERR	Len	t	Cov	ERR	Len	t	Cov	ERR	Len	t
16.1	200	0.98	0.05	0.88	5	0.68	0.54	0.42	1	0.97	0.06	0.93	370	1.00	0.00	1.00	34
	400	0.97	0.10	0.81	14	0.71	0.57	0.38	2	0.96	0.10	0.87	751	1.00	0.00	1.00	56
	600	0.95	0.13	0.74	23	0.70	0.68	0.32	6	0.94	0.10	0.83	3212	1.00	0.00	1.00	118
1.90	200	0.96	0.62	0.34	5	0.80	0.77	0.31	1	0.92	0.86	0.31	371	1.00	0.36	0.58	34
	400	0.94	0.92	0.23	14	0.83	0.93	0.24	2	0.92	0.96	0.23	751	1.00	0.45	0.53	54
	600	0.95	0.95	0.19	22	0.82	0.95	0.20	5	0.95	0.97	0.19	3211	1.00	0.47	0.50	118

Real Data Applications

1. Data: extracted from the Penn Medicine clinical data repository, including demographics, laboratory results, medication prescriptions, vital signs, and encounter meta information.
2. 348 patients, 198 predictors in the final analyses
3. Goal: predicting hypertension, hypertension resistant to standard treatment (“R-hypertension").
4. Outcome prevalence: 39.4% and 8.1%

Real Data Results



We randomly sampled 30 patients as the test sample,

- Left, indexes 1 to 11 correspond to hypertension.
- Right, indices 1 to 4 correspond to R-hypertension.

Conclusion and Discussion

1. Non-linear outcome model: Reweighting
2. Uniform inference for $x_{\rm new}$: Additional constraint
3. Optimality of CI construction

Future research

1. Outcome surrogates
2. Model misspecification

Reference and Acknowledgement

Guo, Z., Rakshit, P., Herman, D. S., and Chen, J. (2021). Inference for the case probability in high-dimensional logistic regression. JMLR, 22(254), 1-54.

Cai, T., Guo, Z., and Ma, R. (2021+). "Statistical inference for high-dimensional generalized linear models with binary outcomes." JASA, to appear.

Acknowledgement to NIH and NSF for fundings.

Thank you!