A Discussion of Tse and Davidson (2022) "A Note on Universal Inference"

Asher Spector [*] Emmanuel Candès [*†§] Lihua Lei [‡§]

May 5, 2023

1 Introduction

We congratulate the authors on the interesting article, which carefully decomposes the various sources of the power deficit experienced by universal inference (UI). We paid special attention to the case where there are many nuisance parameters, since we had hoped that the e-values from UI would be of use in popular high-dimensional models where it can be challenging to obtain valid p-values; for instance, it is now well established that classical asymptotic theory breaks down in high dimensional logistic regression [Sur and Candès, 2019]. Unfortunately, the excellent article of Tse and Davidson shows that UI has very low power in problems with high-dimensional nuisance parameters.

This short note explores ways to improve UI and make it more practical. In particular, Wasserman et al. [2020] built upon ideas in Grünwald et al. [2020] to introduce the reverse information projection (RIPR) split LRT, another e-value based on UI which is designed to increase power when testing composite null hypotheses. Unfortunately, the RIPR split LRT is often very challenging to compute. In this discussion, we suggest a simple modification called the quasi-RIPR split LRT. Although the quasi-RIPR split LRT is not always an e-value, it is in the multivariate Gaussian case (with a known covariance matrix) and is asymptotically equivalent to an oracle e-value for well-behaved parametric models. Most importantly, it is easy to compute and can dramatically improve the power of UI in settings with high-dimensional nuisance parameters.

2 Improved universal inference with nuisance parameters

2.1 The quasi reverse information projection

Recall that to test $H_0 : \theta \in \Theta_0$ using data $\{Y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_\theta$, the split LRT rejects iff

$$T_{\text{UI}} = \frac{\prod_{i=1}^{n_0} p_{\hat{\theta}_1}(Y_i)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^{n_0} p_\theta(Y_i)} > \frac{1}{\alpha}, \tag{1}$$

where $\{Y_1, \ldots, Y_{n_0}\}$ is the zeroth fold of the data and $\hat{\theta}_1$ is any estimate of θ computed from the first fold, $\{Y_{n_0+1}, \ldots, Y_n\}$. Recall that T_{UI} is a valid e-value in the sense that T_{UI} has expectation at most 1 under any null distribution, and thus Eq. (1) is a valid test by Markov's inequality. As in Tse and Davison [2022], we are interested in tests where we partition θ into (ψ, λ) where $\psi \in \Psi \subset \mathbb{R}^k$ denotes the parameter of interest and $\lambda \in \Lambda \subset \mathbb{R}^d$ denotes the nuisance parameter. Then, testing $H_0 : \psi = \psi_0$ corresponds to testing $H_0 : \theta \in \Theta_0$ where $\Theta_0 = \{(\psi_0, \lambda) : \lambda \in \Lambda\}$ is d-dimensional (one can construct a valid confidence interval for ψ by inverting this test).

[*]Department of Statistics, Stanford University
[†]Department of Mathematics, Stanford University
[‡]Graduate School of Business, Stanford University
[§]Author order determined alphabetically.

This test can lose power for at least three reasons: (a) it uses Markov's inequality to obtain a conservative threshold, (b) it requires splitting the data, and (c) the supremum in the denominator can artificially reduce the value of T_{UI}. In high dimensions, we expect reason (c) to drive the most dramatic loss of power, and indeed, Tse and Davison [2022] show that UI is exponentially conservative in the dimension of Θ_0 in the Gaussian case. Similarly, in logistic regression, under mild assumptions, the denominator will asymptotically equal one whenever $n_0 \leq 2d$, yielding a provably powerless test [Candès and Sur, 2020].

To avoid this, the RIPR split LRT replaces the denominator in Eq. (1) by a single worst-case density which is fixed in advance before seeing the zeroth split of the data. As notation, let $\mathrm{KL}(q,p) = \mathbb{E}_{Y\sim q}[\log q(Y)/p(Y)]$ denote the Kullback-Leibler divergence between two densities q and p. Also, given a null model $\mathcal{P}_{\Theta_0} \triangleq \{p_\theta : \theta \in \Theta_0\}$ and a prior W_0 over Θ_0, let $p_{W_0}(y) \triangleq \mathbb{E}_{\theta\sim W_0}[p_\theta(y)]$ be the mixture density which marginalizes over $\theta \sim W_0$.

With this notation, for any fixed density q, the reverse information projection of q onto the model $\mathcal{P}_{\Theta_0}$, denoted $p^\star_q$, is the density

$$p^\star_q \triangleq p_{W_0^\star} \text{ where } W_0^\star = \arg\min_{W_0} \mathrm{KL}(q, p_{W_0}).^1 \tag{2}$$

Grünwald et al. [2020] show that the minimum can be achieved for most practically relevant statistical models. In other words, $p^\star_q$ is the Bayes-mixture over densities in $\mathcal{P}_{\Theta_0}$ which minimizes the KL divergence with q among all such mixtures. Then, to test $H_0 : \theta \in \Theta_0$, the RIPR split LRT computes the RIPR $p^\star_{\hat\theta_1}$ of $p_{\hat\theta_1}$ onto $\mathcal{P}_{\Theta_0}$ and rejects H_0 if

$$T_{\mathrm{RIPR}} = \frac{\prod_{i=1}^{n_0} p_{\hat\theta_1}(Y_i)}{\prod_{i=1}^{n_0} p^\star_{\hat\theta_1}(Y_i)} > \frac{1}{\alpha}. \tag{3}$$

Theorem 1 of Grünwald et al. [2020] implies that, under mild regularity conditions, the above test is valid; in particular, conditional on $\hat\theta_1$,

$$1 \geq \sup_{\theta\in\Theta_0} \mathbb{E}_{Y_i\sim p_{\hat\theta_1}}\left[\frac{p_\theta(Y_i)}{p^\star_{\hat\theta_1}(Y_i)}\right] = \sup_{\theta\in\Theta_0} \mathbb{E}_{Y_i\sim p_\theta}\left[\frac{p_{\hat\theta_1}(Y_i)}{p^\star_{\hat\theta_1}(Y_i)}\right].$$

This proves that T_{RIPR} is a valid e-value, and thus Eq. (3) is a valid test. Furthermore, by definition, Eq. (3) is uniformly more powerful than the original split LRT in Eq. (1). Indeed, in high dimensions, replacing the supremum over Θ_0 with a single worst-case density ought to give a very large power gain.

However, computing the RIPR can be challenging, since in general it requires solving an infinite-dimensional optimization over W_0 [Grünwald et al., 2020, Ramdas et al., 2022]. This motivates us to introduce the *quasi reverse information projection*, which replaces $p^\star_{\hat\theta_1}$ with $p^{\mathrm{quasi}}_{\hat\theta_1} \triangleq p_{\theta^\star}$ where

$$\theta^\star = \arg\min_{\theta\in\Theta_0} \mathrm{KL}(p_{\hat\theta_1}, p_\theta) = \arg\max_{\theta\in\Theta_0} \mathbb{E}_{Y\sim p_{\hat\theta_1}}[\log p_\theta(Y)]. \tag{4}$$

The key difference between the RIPR and the quasi-RIPR is that the RIPR finds the worst case prior W_0 over Θ_0, whereas the quasi-RIPR finds the worst case value $\theta \in \Theta_0$.

The quasi-RIPR split LRT does not always yield an exact e-value, although in some simple cases it exactly recovers the RIPR (see Section 2.2, Grünwald et al. [2020], Ramdas et al. [2022]). Even when it does not, a forthcoming work by the authors will study the asymptotic validity of quasi-RIPR [Lei, 2023]. Overall, since the RIPR split LRT can be somewhat conservative to begin with (due to the use of Markov's inequality), we expect the quasi-RIPR split LRT to be conservative as well.

Most importantly, computing the quasi-RIPR $p^{\mathrm{quasi}}_{\hat\theta_1}$ is often not much more expensive than computing the MLE $\hat\theta_0$ on the zeroth fold. Indeed, the second objective in Eq. (4) is concave

[1] There are more general definitions of reverse information projections; see Grünwald et al. [2020] Theorem 1, Li [1999] for details.

whenever the log-likelihood $\theta \mapsto \log p_\theta(x)$ is concave, and indeed, Eq. (4) is effectively a re-weighted maximum likelihood problem. For example, consider the case of logistic regression where $Y_i \overset{\text{ind}}{\sim} \text{Bern}((1+\exp(X_i^\top\theta))^{-1})$ for fixed $X_i, \theta \in \mathbb{R}^p$. Although $Y_1, \ldots, Y_n$ are not i.i.d., we can treat them as a single observation from the likelihood $\log p_\theta(\vec{y}) = \sum_{i=1}^n (1-y_i)X_i^T\theta - \log(1+\exp(X_i^T\theta))$. Then, Eq. (4) becomes

$$\theta^\star = \underset{\theta\in\Theta_0}{\arg\max} \sum_{i=1}^{n_0} \mathbb{E}_{\hat{\theta}_1}[1-Y_i]X_i^\top\theta - \log\left(1+\exp(X_i^\top\theta)\right). \tag{5}$$

The objective in Eq. (5) is simply the logistic regression log-likelihood for the zeroth split of the data, except we replace Y_i with its expectation under $\hat{\theta}_1$. Since $\hat{\theta}_1$ is not an optimization variable, solving Eq. (5) is no harder than computing the logistic regression MLE. And a generic strategy to solve Eq. (4) with off-the-shelf MLE solvers is to sample $Y_1^\star, \ldots, Y_K^\star \overset{\text{i.i.d.}}{\sim} P_{\hat{\theta}_1}$ and set $\theta^\star = \arg\max_{\theta\in\Theta_0} \sum_{i=1}^K \log p_\theta(Y_i^\star)$, which recovers the correct solution as $K \to \infty$.

The quasi-RIPR split LRT may be a good default choice over the split LRT in problems with many nuisance parameters, although more work needs to be done to verify its validity. That said, we will see in Section 3 that the quasi-RIPR split LRT does not resolve the other sources of power loss, namely the need to split the data and the use of Markov's inequality.

2.2 Statistical property for multivariate Gaussian problems

Following Tse and Davison [2022], we consider mean estimation for multivariate Gaussian distributions with known covariance matrices. Let

$$Y_i = \begin{bmatrix} Y_{i,\psi} \\ Y_{i,\lambda} \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{bmatrix} \psi \\ \lambda \end{bmatrix}, \Sigma\right), \quad \text{where } \Sigma = \begin{bmatrix} \Sigma_{\psi\psi} & \Sigma_{\psi\lambda} \\ \Sigma_{\lambda\psi} & \Sigma_{\lambda\lambda} \end{bmatrix}.$$

For simplicity, we assume that $\Sigma \succ 0$ and (ψ, λ) is unrestricted (i.e., $\Psi = \mathbb{R}^k, \Lambda = \mathbb{R}^d$). Denote by $\bar{Y}_s = (\bar{Y}_{s,\psi}, \bar{Y}_{s,\lambda})$ the average of Y_i on fold s ($s = 0, 1$). Further, assume that $\hat{\theta}_1$ is given by the maximum likelihood estimator on fold 1, i.e., $\hat{\theta}_1 = (\bar{Y}_{1,\psi}, \bar{Y}_{1,\lambda})$.

Now, for fixed $\psi \in \Psi$, we derive the quasi-RIPR $p^{\text{quasi}}_{\psi,\hat{\theta}_1}$. It is well-known that

$$\text{KL}\Big(\mathcal{N}((\psi,\lambda),\Sigma), \mathcal{N}((\bar{Y}_{1,\psi}, \bar{Y}_{1,\lambda}), \Sigma)\Big) = \frac{1}{2}\|(\psi - \bar{Y}_{1,\psi}, \lambda - \bar{Y}_{1,\lambda})\|^2_{\Sigma^{-1}},$$

where $\|a\|_B^2 = a^\top Ba$. Let $M = \Sigma^{-1}$ and partition it into $\begin{bmatrix} M_{\psi\psi} & M_{\psi\lambda} \\ M_{\lambda\psi} & M_{\lambda\lambda} \end{bmatrix}$. It is easy to show that $\Sigma^{-1}_{\psi\psi} = M_{\psi\psi} - M_{\psi\lambda}M^{-1}_{\lambda\lambda}M_{\lambda\psi}$ and, for any $a \in \mathbb{R}^k, b \in \mathbb{R}^d$,

$$\|(a,b)\|^2_{\Sigma^{-1}} = \|a\|^2_{\Sigma^{-1}_{\psi\psi}} + \|b + M^{-1}_{\lambda\lambda}M_{\lambda\psi}a\|^2_{M_{\lambda\lambda}}. \tag{6}$$

Using this property, we find $p^{\text{quasi}}_{\psi,\hat{\theta}_1}$ is given by the density of $N((\psi, \hat{\lambda}(\psi)), \Sigma)$ where

$$\hat{\lambda}(\psi) = \bar{Y}_{1,\lambda} - M^{-1}_{\lambda\lambda}M_{\lambda\psi}(\psi - \bar{Y}_{1,\psi}). \tag{7}$$

Then, if $\mathcal{C}^{\text{qRIPR}}_{\psi,1-\alpha}$ denotes the $1-\alpha$ confidence interval based on the quasi-RIPR split LRT, $\psi \in \mathcal{C}^{\text{qRIPR}}_{\psi,1-\alpha}$ iff

$$-\|(\bar{Y}_{1,\psi} - \bar{Y}_{0,\psi}, \bar{Y}_{1,\lambda} - \bar{Y}_{0,\lambda})\|^2_{\Sigma^{-1}} + \|(\psi - \bar{Y}_{0,\psi}, \hat{\lambda}(\psi) - \bar{Y}_{0,\lambda})\|^2_{\Sigma^{-1}} \le -\frac{2}{n_0}\log\alpha.$$

By (6), the LHS can be written as

$$\begin{aligned} &- \|\bar{Y}_{1,\psi} - \bar{Y}_{0,\psi}\|^2_{\Sigma^{-1}_{\psi\psi}} - \|\bar{Y}_{1,\lambda} - \bar{Y}_{0,\lambda} + M^{-1}_{\lambda\lambda}M_{\lambda\psi}(\bar{Y}_{1,\psi} - \bar{Y}_{0,\psi})\|^2_{M_{\lambda\lambda}} \\ &+ \|\psi - \bar{Y}_{0,\psi}\|^2_{\Sigma^{-1}_{\psi\psi}} + \|\hat{\lambda}(\psi) - \bar{Y}_{0,\lambda} + M^{-1}_{\lambda\lambda}M_{\lambda\psi}(\psi - \bar{Y}_{0,\psi})\|^2_{M_{\lambda\lambda}}. \end{aligned} \tag{8}$$

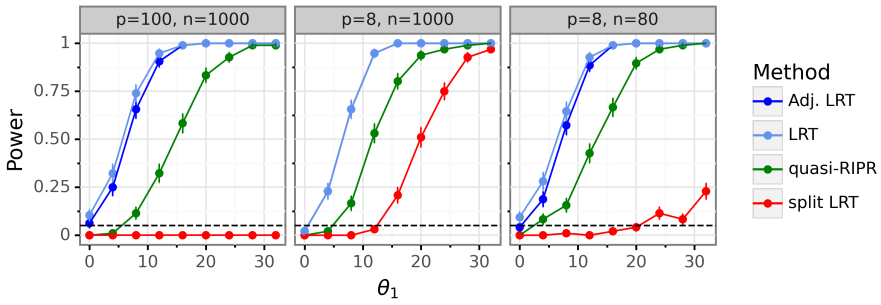

Figure 1: This figure shows the power and Type I error of four methods testing $H_0 : \theta_1 = 0$ in logistic regression. The dashed black line is the nominal level $\alpha = 0.05$; note that when $\theta_1 = 0$, the LRT has an inflated Type I error of 9% in the first and third panel, respectively. In the simulation, we sample $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, and for $j \neq 1$, we set $\theta_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \eta^2)$ for η such that $\text{Var}(X_i^\top \theta) = 15$. The figure shows that the quasi-RIPR split LRT is far more powerful than the split LRT, but still much less powerful than the LRT and adjusted LRT.

By (7),

$$\hat{\lambda}(\psi) - \bar{Y}_{0,\lambda} + M_{\lambda\lambda}^{-1} M_{\lambda\psi}(\psi - \bar{Y}_{0,\psi}) = \bar{Y}_{1,\lambda} - \bar{Y}_{0,\lambda} + M_{\lambda\lambda}^{-1} M_{\lambda\psi}(\bar{Y}_{1,\psi} - \bar{Y}_{0,\psi}).$$

Thus, the second and the fourth terms of (8) cancel, implying that

$$\mathcal{C}^{\text{qRIPR}}_{\psi,1-\alpha} = \left\{ \psi : -\|\bar{Y}_{1,\psi} - \bar{Y}_{0,\psi}\|^2_{\Sigma^{-1}_{\psi\psi}} + \|\psi - \bar{Y}_{0,\psi}\|^2_{\Sigma^{-1}_{\psi\psi}} \leq -\frac{2}{n_0} \log \alpha \right\}.$$

This is identical to the split LRT for ψ based on the subvectors $(Y_{i,\psi})$, in which case the nuisance parameter λ is completely removed. Therefore, for any λ, the e-value corresponding to $\mathcal{C}^{\text{qRIPR}}_{\psi,1-\alpha}$ has mean equal to 1, i.e.,

$$\mathbb{E}_{Y_i \sim N((\psi_0,\lambda),\Sigma)} \left[\prod_{i=1}^{n_0} \frac{p_{\hat{\theta}_1}(Y_i)}{p^\star_{\psi_0,\hat{\theta}_1}(Y_i)} \right] = 1.$$

By contrast, Tse and Davison [2022] prove that the e-value given by the original split LRT is bounded by $(2 + n_0/n_1)^{-d/2}$, which is exponentially more conservative than the RIPR split LRT.

3 Numerical results for logistic regression

We now compare the split LRT against the quasi-RIPR split LRT in the setting of logistic regression, so $Y_i \mid X_i \overset{\text{ind}}{\sim} \text{Bern}((1 + \exp(\theta^\top X_i)^{-1})$ for $\theta, X_i \in \mathbb{R}^p$ fixed. Logistic regression is an important test case because in high dimensions, we know of few methods to compute valid p-values testing (e.g.) $H_0 : \theta_1 = 0$ without making distributional assumptions on X. Indeed, Sur and Candès [2019] showed that classical tests like the LRT or the Wald test are asymptotically invalid in high dimensions unless $\frac{p}{n} \to 0$. Although Candès et al. [2018], Sur and Candès [2019] developed valid methods to test H_0 in high dimensions, these methods require that the distribution of X is either known or has sub-Gaussian tails, which may not always be the case. Furthermore, it is not known how to compute the exact RIPR in this setting, motivating our use of the quasi-RIPR.

Figures 1 and 2 compares the power and Type I error of the classical LRT, the split LRT, the quasi-RIPR split LRT, and the adjusted LRT from Sur and Candès [2019]. As expected, even when $n = 10p$, the classical LRT has a substantially inflated Type I error (please see the figure caption), indicating this is a challenging high-dimensional problem. However, in these settings, the split LRT is almost completely powerless due to the dimensionality of the problem, and even when $n = 125p$, the split LRT is much less powerful than all other methods. In contrast, the quasi-RIPR split LRT is much more powerful, although it is still conservative and substantially less powerful than the adjusted LRT from Sur and Candès [2019], likely due to the use of Markov's inequality and the

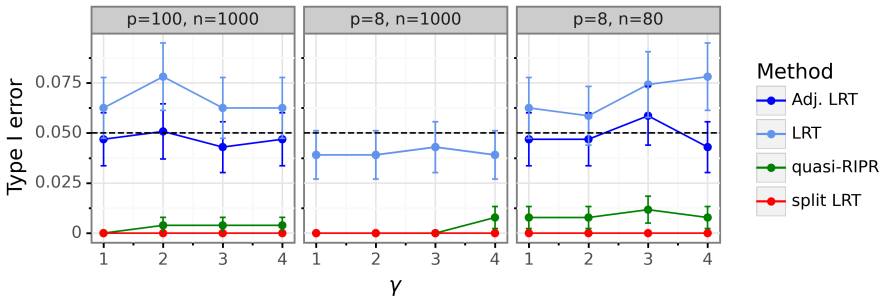

Figure 2: The same setting as Figure 1 except $\beta_1 = 0$ and we vary γ. It shows that the quasi-RIPR split LRT is conservative.

need to split the data. Overall, the quasi-RIPR split LRT gives some hope that UI may be useful in problems with high-dimensional nuisance parameters. Nonetheless, more work remains to be done to make it sufficiently powerful to compete with established alternatives when they exist and to rigorously establish when it controls Type I error.

4 Acknowledgements

We are grateful to Peter Grünwald, who alerted us to an error in a previous version.

References

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577, 2018.

Emmanuel J. Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. 48(1):27–42, 2020. doi: 10.1214/18-AOS1789.

Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–54. IEEE, 2020.

Lihua Lei. Handling nuisance parameters via the quasi-reverse information projection, 2023.

Qiang Jonathan Li. *Estimation of mixture models*. Yale University, 1999.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference, 2022.

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. 116(29):14516–14525, 2019.

Timmy Tse and Anthony C Davison. A note on universal inference. *Stat*, page e501, 2022.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.