“Plus/minus” confidence intervals and thresholding

P. Zietkiewicz

December 4, 2023

Table of contents

Summary of Chee et al. (2023)

- $(Y, X) \in \mathbb{R}^d \times \mathbb{R}^p$ and $D_N = \{(Y_i, X_i) : i = 1, \dots, N\}$

$$\theta_* = \text{argmin}_{\theta \in \Theta} \mathbb{E}[\ell(\theta, Y, X)]$$
$$\hat{\theta}_N = \text{argmin}_{\theta \in \Theta} \sum_{i=1}^{N} \ell(\theta, Y_i, X_i)$$
$$F_* = \mathbb{E}[\nabla \ell(\theta, Y, X) \nabla \ell(\theta, Y, X)^\top]$$

- SGD: $\theta_n = \theta_{n-1} - \gamma_n \nabla \ell(\theta_{n-1}; Y_i, X_i)$ for $i = 1, \dots, N$ and γ_n is the learning rate typically $\gamma_n = \gamma_1 / n$. Let θ_N be the one-pass estimator of θ_*.
- Advantages of one-pass over multi-pass: (1) Asymptotic covariance matrix is known in closed form (2) Covariance matrix can be bounded by a factor that depends only on the learning rate γ_1.

Summary of Chee et al. (2023)

- Propose the SGD-based CIs for each component $\theta_{*,j}$

$$\theta_{N,j} \pm 2\sqrt{\frac{\gamma_1^*}{N}} \text{ for } j = 1, \ldots, p.$$

- Define $\Sigma_* = \gamma_1^2(2\gamma_1 F_* - I)^{-1} F_*$ where γ_1 is large enough such that $2\gamma_1 F_* - I \succ 0$. And has eigenvalues

$$\text{eigen}(\Sigma_*) = \{\frac{2\gamma_1^2 \lambda_j}{2\gamma_1 \lambda_j - 1} : j = 1, \ldots, p\}$$

where λ_j is the jth eigenvalue of F_*.

Results:

Theorem 3.1. *Let $\theta_{N,j}$, denote the j-th component of θ_N in Eq. (4), for $j = 1, \dots, p$. Suppose that $\gamma_1^* \geq 1/\min_j\{\lambda_j\}$, then $\gamma_1^* I - \Sigma_\star \succ 0$. Define the interval*

$$C_{N,j}(D_N) = \left[\theta_{N,j} - z_{\frac{\alpha}{2}}\sqrt{\frac{\gamma_1^*}{N}},\ \theta_{N,j} + z_{\frac{\alpha}{2}}\sqrt{\frac{\gamma_1^*}{N}}\right], \quad (9)$$

where $z_{\frac{\alpha}{2}} = \Phi^{-1}(1-\alpha/2)$ is the critical value of the standard normal. Then, for every $j = 1, \dots, p$,

$$\liminf_{N\to\infty} P\big(\theta_{\star,j} \in C_{N,j}(D_N)\big) \geq 1 - \alpha. \quad (10)$$

Theorem 3.2. *Let θ_N be the one-pass SGD in Eq. (4), and suppose that $\gamma_1^* \geq 1/\min_j\{\lambda_j\}$. Define the following confidence region:*

$$\widehat{\Theta} = \left\{\theta \in \Theta : (1/\gamma_1^*)\ ||\theta - \theta_N||^2 < \chi_{\alpha,p}\right\}, \quad (11)$$

where $\chi_{\alpha,p} = \sup\{x \in \mathbb{R} : P(\chi_p^2 \geq x) \leq \alpha\}$ is the α-critical value of a chi-squared random variable with p degrees of freedom. Then,

$$\liminf_{N\to\infty} P(\theta_\star \in \widehat{\Theta}) \geq 1 - \alpha. \quad (12)$$

Selecting γ_1^*:

Linear asymptote in $\Sigma_\star$. At a high level, the variance bound in Theorem 3.1 holds in the regime where the covariance matrix of θ_N is linear with respect to γ_1. One idea is therefore to try and estimate when such regime has been reached. The idea is visualized in Figure 3. Recall from Eq. (8) that the eigenvalues of $\Sigma_\star$ asymptote to $\gamma_1/2$, and so the trace of $\Sigma_\star$ should asymptote to $p\gamma_1/2$, as shown in the figure. The idea is then to slowly increase the learning rate γ_1 and at the same time monitor the trace of $N\mathrm{Var}(\theta_N)$. When γ_1 is large enough for Theorem 3.1 we expect that a linear regression of $\mathrm{trace}(N\mathrm{Var}(\theta_N))$ with respect to γ_1 will give a coefficient around $p/2$ with high confidence. Only a crude estimate of the variance trace is needed, which can be done via bootstrap. See Appendix D.1 for more details, and a practical example.

An eigenvalue bound. In some settings, an estimate $\tilde{F}$ of $F_\star$ exists that may be too crude to be used directly for inference, but may be acceptable for estimating a bound on $\lambda_{\min}$. Then, an alternative way of selecting γ_1^* is to numerically find the maximum eigenvalue of $\tilde{F}^{-1}$, which implies the minimum eigenvalue of $F_\star$ To this end, we propose using inverse power iteration (Trefethen and Bau III, 1997), which is a simple iterative algorithm. More details of this algorithm and its implementation are in Appendix D.2.

Thresholding and SGD

- In the context of thresholding we define the pivots

$$\frac{\hat{\beta}_j}{\sqrt{\frac{\gamma_1^*}{N}}}$$

  where we have the usual behaviour for $\hat{\beta}_j$ and the same behaviour from $\sqrt{\frac{\gamma_1^*}{N}} = O(N^{-1/2})$.
- Seems to work.
- Next steps: implementing an iterative version so we can build confidence sets.

References

Chee, J., H. Kim, and P. Toulis (2023, 25–27 Apr). "plus/minus the learning rate": Easy and scalable statistical inference with sgd. In F. Ruiz, J. Dy, and J.-W. van de Meent (Eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, Volume 206 of *Proceedings of Machine Learning Research*, pp. 2285–2309. PMLR.