LaTeX Template for STAT 548 Qualifying Paper Report

Naitong Chen

September 22, 2022

1 Summary

Tibshirani and Wasserman (2015)

Stein's Lemma:

- (univariate) Let $Z \sim \mathcal{N}(0,1)$. Let $f : \mathbb{R} \to \mathbb{R}$ be absolutely continuous, with derivative f' (and assume that $\mathbb{E}\left[|f'(Z)|\right] < \infty$). Then $\mathbb{E}\left[Zf(Z)\right] = \mathbb{E}\left[f'(Z)\right]$.
- (extesion) Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then $\frac{1}{\sigma^2}\mathbb{E}\left[(x-\mu)f(x)\right] = \mathbb{E}\left[f'(X)\right]$.
- (multivariate) Let $X \sim \mathcal{N}(\mu, \sigma^2 I)$, where $\mu \in \mathbb{R}^n$ and $\sigma^2 I \in \mathbb{R}^{n\times n}$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function such that, for each $i = 1, \cdots, n$ and almost every $x_{-i} \in \mathbb{R}^{n-1}$, $f(\cdot, x_{-i}) : \mathbb{R} \to \mathbb{R}$ is absolutely continuous (and assume $\|f(X)\|_2 < \infty$). Then $\frac{1}{\sigma^2}\mathbb{E}\left[(X-\mu)f(X)\right] = \mathbb{E}\left[\nabla f(X)\right]$.
- (extension) Let $f = (f_1, \cdots, f_n)$, then

$$\begin{aligned}
&\frac{1}{\sigma^2}\mathbb{E}\left[(X-\mu)f_i(X)\right] = \mathbb{E}\left[\nabla f_i(X)\right] \\
\Longrightarrow\ &\frac{1}{\sigma^2}\sum_{i=1}^n \mathrm{Cov}(X_i, f_i(X)) = \frac{1}{\sigma^2}\sum_{i=1}^n \mathbb{E}\left[(X_i-\mu_i)f_i(X)\right] = \mathbb{E}\left[\sum_{i=1}^n \frac{\partial f_i}{\partial X_i}(X)\right].
\end{aligned}$$

Stein's Unbiased Risk Estimate (SURE):

Given samples $y \sim \mathcal{N}\left(\mu, \sigma^2 I\right)$, and a function $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}^n$, $\hat{\mu}$ is a fitting procedure that, from y, provides an estimate $\hat{\mu}(y)$ of the underlying (unknown) mean μ. Then

$$\begin{aligned}
R &= \mathbb{E}_y\|\mu - \hat{\mu}(y)\|^2 \\
&= -n\sigma^2 + \mathbb{E}\|y-\hat{\mu}\|_2^2 + 2\sigma^2 \mathrm{df}(\hat{\mu}) \\
&= -n\sigma^2 + \mathbb{E}\|y-\hat{\mu}\|_2^2 + 2\sum_{i=1}^n \mathrm{Cov}\left(y_i, \hat{\mu}_i\right),
\end{aligned}$$

where $\mathrm{df}(\hat{\mu}) = \frac{1}{\sigma^2}\sum_{i=1}^n \mathrm{Cov}(y_i, \hat{\mu}_i)$. And

$$\hat{R} = -n\sigma^2 + \|y - \hat{\mu}(y)\|_2^2 + 2\sigma^2\sum_{i=1}^n \frac{\partial \hat{\mu}_i}{\partial y_i}(y)$$

is an unbiased estimate for R.

Extending SURE to regularized estimators:

Now suppose $\hat{\mu}_\lambda$ depends on $\lambda \in \Lambda$, which controls the degree of regularization to our estimator (typically $\Lambda = \mathbb{R}_{>0}$), and assume σ is known, we can find the optimal λ, denoted $\hat{\lambda}$ by

$$\hat{\lambda} = \underset{\sigma\in\Sigma}{\arg\min}\, \|y - \hat{\mu}_\lambda(y)\|_2^2 + 2\sigma^2\sum_{i=1}^n \frac{\partial \hat{\mu}_{\lambda,i}}{\partial y_i}(y).$$

2 Mini-proposals

2.1 Proposal 1: MY PROPOSAL TITLE

2.2 Proposal 2: MY OTHER PROPOSAL TITLE

3 Project report

SURE with Ridge Regression:

Let $y \sim \mathcal{N}\left(X^T\beta, \sigma^2\right)$, where $y \in \mathbb{R}$ and $X \in \mathbb{R}^{p+1}$, X constant. Then with $\boldsymbol{X} = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$, we have $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\beta, \sigma^2 I\right)$, where $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{X} \in \mathbb{R}^{n \times (p+1)}$.

We know that $\hat{\beta}_{\text{ridge}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_{p+1}\right)^{-1} \boldsymbol{X}^T\boldsymbol{y}$, then

$$
\begin{aligned}
\hat{\mu}_\lambda(\boldsymbol{y}) &= \boldsymbol{X}\hat{\beta}_{\text{ridge}} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_{p+1}\right)^{-1} \boldsymbol{X}^T\boldsymbol{y} \\
\hat{\mu}_{\lambda,i}(\boldsymbol{y}) &= X_i^T\hat{\beta}_{\text{ridge}} = X_i^T\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_{p+1}\right)^{-1} \boldsymbol{X}^T\boldsymbol{y}
\end{aligned}
$$

Then

$$
\begin{aligned}
\frac{\hat{\mu}_{\lambda,i}(\boldsymbol{y})}{\partial y_i} &= \frac{\partial}{\partial y_i}\left(X_i^T\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_{p+1}\right)^{-1} \boldsymbol{X}^T\boldsymbol{y}\right) \\
&= \frac{\partial}{\partial y_i}F_i\boldsymbol{y} \qquad \left(F_i := X_i^T\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_{p+1}\right)^{-1} \boldsymbol{X}^T \in \mathbb{R}^n\right) \\
&= F_{i,i} \\
&= \left(X_i^T\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_{p+1}\right)^{-1} \boldsymbol{X}^T\right)_i.
\end{aligned}
$$

We can now write

$$
\begin{aligned}
\hat{R} &= -n\sigma^2 + \|\boldsymbol{y} - \hat{\mu}_\lambda(\boldsymbol{y})\|_2^2 + 2\sigma^2\sum_{i=1}^n\left(X_i^T\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_{p+1}\right)^{-1} \boldsymbol{X}^T\right)_i \\
&= -n\sigma^2 + \|\boldsymbol{y} - \hat{\mu}_\lambda(\boldsymbol{y})\|_2^2 + 2\sigma^2 \operatorname{tr}\left(\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_{p+1}\right)^{-1} \boldsymbol{X}^T\right) \\
&= -n\sigma^2 + \|\boldsymbol{y} - \hat{\mu}_\lambda(\boldsymbol{y})\|_2^2 + 2\sigma^2 \operatorname{tr}\left(\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_{p+1}\right)^{-1}\right) \\
&= -n\sigma^2 + \|\boldsymbol{y} - \hat{\mu}_\lambda(\boldsymbol{y})\|_2^2 + 2\sigma^2 \operatorname{tr}\left(H\left(H + \lambda I_{p+1}\right)^{-1}\right),
\end{aligned}
$$

where the last line is by defining $H := \boldsymbol{X}^T\boldsymbol{X}$. We can optimize λ over $\hat{R}$ using autodiff (log-transform λ so that it is nonnegative).

References

Ryan Tibshirani and L Wasserman. Stein's unbiased risk estimate. *Course notes from "Statistical Machine Learning*, pages 1–12, 2015.