LaTeX Template for STAT 548 Qualifying Paper Report

Naitong Chen

October 13, 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: Efficient cross validation via data subsampling

In the large-data regime where the number of observations is much greater than the number of predictors ($n \gg p$), solving the OLS problem and/or ridge regression can be computationally expensive. Especially during hyperparameter turning in ridge regression, where we'd solve the same problem with a subset of the data k times if we were using a k-fold cross validation procedure.

If we were able to select a sparse, weighted subsample of the data that still contains information about the full dataset, we can greatly reduce the computational cost of this procedure. However, since both the β and $L2$ penalty term λ are unknown prior to solving the regression problem, we need to ensure our selected subsample gives us a good approximation of the full dataset across a wide range of β and λ values, or at least for β and λ values that we will likely get.

This notion of a good approximation over the high density region of some distribution aligns well with the Bayesian point of view. We can therefore leverage the idea of Bayesian coresets to construct our sparse, weighted subsample.

Insert description of Bayesian coresets.

This approach makes intuitive sense as we can associate the ridge regression estimate to the MAP estimate of the coefficients with a Gaussian prior (whose variance is speficied by λ). Note that the Gaussian prior still fixes a λ value. To make our constructed subsample a good approximation over various λ values, we can introduce a prior on the variance of the ridge regression coefficients.

Depending on the choice of the prior on λ, we could potentially have closed form solutions of the posterior distribution over the ridge regression coefficients. This then could enable us to efficiently build a sparse, weighted subsample that well approximates the full dataset over a wide range of β and λ values.

Insert weighted linear regression (with L2 regularization) closed-form solutions.

With the above closed-form solutions to the weighted ridge regression problem, we can now perform cross validation either through the regular approach or the SURE approach discussed in the following section of the report, with much lower computational costs.

2.2 Proposal 2: Using SURE to automate cross validation for principal component regression with L1 regularization

Given the example of using SURE to tune the L2 penalty term through autodiff, it is natural to try and extend this idea to LASSO regression. However, there are a few problems to consider: 1) LASSO regression only has closed form solutions under specific cases (univariate or orthogonal data matrix), which is required for the use of SURE; 2) even when there is a closed-form solution, the solution function is not smooth and hence not differentiable at the non-smooth part.

To address the first problem, we can consider principal compoent regression (PCR) where we perform regression on the principal components resulted from a principal component analysis (PCA). The principal components form a orthogonal basis and so we can obtain a closed-form solution for LASSO regression if we were to treat the principal components as our data.

Insert principal component formulation.

By treating the above as our data, we can write the LASSO regression solution as follows.

Insert LASSO regression solution with orthogonal data matrix.

Now that we have obtained a closed-form solution to the principal component transformed LASSO problem, we can write the SURE as

Insert SURE with PCR + LASSO.

The last term can be replaced with the number of non-zero coefficients (Tibshirani and Wasserman (2015)), and we can use a subgradient method to solve for the optimal λ.

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.