Qualifying Paper Report for Unsupervised Learning with Stein's Unbiased Risk Estimator

Naitong Chen

October 16, 2022

1 Summary

Parameter estimation lies in the heart of statistical inference, and the customary maximum likelihood estimator (MLE) may not be optimal in terms of the mean-squared error (MSE). Consider the setting where for some $n \in \mathbb{N}$, we have an observation $x \in \mathbb{R}^n$ that is a realization of $X \sim \mathcal{N}(\mu, I)$. To estimate $\mu \in \mathbb{R}^d$, maximum likelihood estimation would yield $\hat{\mu}(x) = x$. In Stein (1956), a perhaps surprising result shows that when $n \geq 3$, there exists some other estimator $\tilde{\mu}$ such that

$$\mathbb{E}\|\tilde{\mu}(X) - \mu\|^2 < \mathbb{E}\|\hat{\mu}(X) - \mu\|^2.$$

In fact, there are many other cases where the maximum likelihood estimator is not optimal under the MSE, a widely used metric for evaluating the quality of an estimator thanks to its mathematical tractability (Berger, 1975; DeGroot, 2005). In a follow-up work by Charles Stein, he developed what is known as Stein's unbiased risk estimate (SURE), which provides an unbiased estimate of the MSE of an arbitrary estimator for the mean of a normally distributed random variable of the form $\mathcal{N}(\mu, \sigma^2 I)$. In what follows, we present a version of this result as outlined in Tibshirani and Wasserman (2015).

Lemma 1.1. *Let $X \sim \mathcal{N}(\mu, \sigma^2 I)$, where $\mu \in \mathbb{R}^n$ and $\sigma > 0$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function, and let $f(\cdot, x_{-i})$ refer to f as a function of its i^{th} component x_i with all other components x_{-i} held fixed. Suppose 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. If we further assume $\mathbb{E}\|f(X)\|_2 < \infty$, then*

$$\frac{1}{\sigma^2}\mathbb{E}\left[(X - \mu)f(X)\right] = \mathbb{E}\left[\nabla f(X)\right].$$

By decomposing f by its coordinate functions $f = (f_1, \ldots, f_n)$, we have that for each $i = 1, \ldots, n$,

$$\frac{1}{\sigma^2}\mathbb{E}\left[(X - \mu)f_i(X)\right] = \mathbb{E}\left[\nabla f_i(X)\right].$$

Then summing over all n components yields

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

Now suppose $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}^n$ is an arbitrary estimator that satisfies the assumptions laid out in Lemma 1.1, it can be shown that

$$R = \mathbb{E}\|\mu - \hat{\mu}(X)\|^2 = -n\sigma^2 + \mathbb{E}\|X - \hat{\mu}(X)\|^2 + 2\sum_{i=1}^{n}\mathrm{Cov}\left(X_i, \hat{\mu}_i(X)\right),$$

which finally leads to

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

as an unbiased estimator for the MSE of $\hat{\mu}$.

It is worth noting that the SURE can be employed on a very general class of estimators and that it removes the explicit dependence on the unknown μ. These desirable features have enabled SURE to fuel the development of many estimators that are more superior in MSE than the MLE for parameter estimation problems under the normal distribution and beyond. For instance, under the SURE framework, the James-Stein estimator (James and Stein, 1992) can be shown to be a strictly better estimator in terms of MSE for normally distributed vectors with unit covariance. Lemma 1.1 has also been extended to the exponential family, where subsequent estimators outperforming the MLE in terms of MSE have been developed for parameter estimation problems when the underlying distribution is Gamma, Poisson, ect. (Hudson, 1978; Peng, 1975; Tsui, 1978).

The SURE has also been found in a wide range of applications beyond merely parameter estimation. As an example, it can be used to perform model selection for ridge regression. In a typical linear regression setting, we are given a set of n observations such that

$$y_i = x_i^T \beta + \epsilon_i,$$

where $\beta \in \mathbb{R}^p$ for some $p \in \mathbb{N}$ and for all $i = 1, \ldots, n$, $x_i \in \mathbb{R}^p$, $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$. We can equivalently write that

$$Y = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^T \sim \mathcal{N}\left(X\beta, \sigma^2 I\right), \quad \text{where } X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T.$$

Given a regularization parameter $\lambda \geq 0$, we can set

$$\hat{\mu}_\lambda(Y) = \left(X^T X + \lambda I\right)^{-1} X^T Y,$$

the ridge estimator for β, and subsequently the unbiased risk estimate for $\hat{\mu}_\lambda$ takes the form

$$\hat{R} = -n\sigma^2 + \|Y - \hat{\mu}_\lambda(Y)\|^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \hat{\mu}_{\lambda,i}}{\partial y_i}(Y).$$

Note that the second term encourages the estimates to be close to the observations, the last term encourages the estimator to not change much under perturbations of the observations, thus creating a bias-variance trade-off. As a result, selecting $\lambda \geq 0$ by minimizing the SURE can be seen as a model selection procedure that is similar in spirit to cross-validation. This λ selection procedure was first proposed in Mallows (1973), and the corresponding risk estimate was later shown to be asymptotically optimal as the numbser of observations approaches infinity (Li, 1986).

2 Mini-proposals

2.1 Proposal 1: 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.

2.2 Proposal 2: 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.

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

James Berger. Minimax estimation of location vectors for a wide class of densities. *The Annals of Statistics*, pages 1318–1328, 1975.

Morris H DeGroot. *Optimal statistical decisions*. John Wiley & Sons, 2005.

H Malcolm Hudson. A natural identity for exponential families with applications in multiparameter estimation. *The Annals of Statistics*, 6(3):473–484, 1978.

William James and Charles Stein. Estimation with quadratic loss. In *Breakthroughs in statistics*, pages 443–460. Springer, 1992.

Ker-Chau Li. Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, pages 1101–1112, 1986.

C. L. Mallows. Some comments on cp. *Technometrics*, 15(4):661–675, 1973. ISSN 00401706. URL http://www.jstor.org/stable/1267380.

James Chao-Ming Peng. *Simultaneous estimation of the parameters of independent Poisson distributions*. Stanford University, 1975.

Charles Stein. Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability: Contributions to the Theory of Statistics*, volume 1, page 197. University of California Press, 1956.

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

Kam-Wah Tsui. *Simultaneous estimation of the parameters of the distributions of independent Poisson random variables*. PhD thesis, University of British Columbia, 1978.