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

Naitong Chen

October 17, 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

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

Note that the second term encourages the estimates to be close to the observations, and 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 in Li (1986) to be asymptotically optimal as the numbser of observations approaches infinity. Namely, denote the selected regularization parameter $\lambda^\star$, we have that

$$\frac{\hat{R}\left(\lambda^\star\right)}{\inf_{\lambda \geq 0} R(\lambda)} \xrightarrow{p} 1.$$

SURE has also been widely used in the application of image denoising, where it is most commonly used directly as the objective function under which we find the optimal parameter

setting using a set of training images (noisy and noise-less). Typically these parameters control the threshold used to decide whether the corresponding signal should be removed. While SURE directly applies when the noise is assumed to be normally distributed, there have also been methods developed to handle other distributions of noise (Donoho and Johnstone, 1995; Luisier et al., 2010; Panisetti et al., 2014). Using a slightly different approach, the SURE framework has also been shown in Metzler et al. (2018) to be particularly useful in the setting where training images are not available or only noisy images are available without their noise-less counterparts. This is indeed a very common setting in practice: in medical imaging, microscopy, and astronomy, noise-less ground truth data are rarely available. Here we discuss this work in more detail. Suppose that for an unobserved noise-less iamge $x \in \mathbb{R}^n$, we observe a noisy version of the image y such that

$$y = x + w, \quad w \sim \mathcal{N}\left(0, \sigma^2 I\right).$$

Our goal is to recover the noise-less image x by transforming the noisy observation y through some image denoiser $f_\theta : \mathbb{R}^n \to \mathbb{R}^n$ parameterized by θ. Under the framework of SURE, we can reconstruct the image using the optimal denoiser function obtained by minimizing the unbiased risk estimate

$$\hat{R}(\theta) = -\sigma^2 + \|y - f_\theta(y)\|^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial f_{\theta,i}}{\partial y_i}(y) \approx R(\theta) = \mathbb{E}_w \|x - f_\theta(y)\|^2.$$

The key observations here is that, under the framework of SURE, we no longer require the noise-less ground truth image to obtain a reconstructed image that minimizes the MSE between itself and the true image. Through the lense of bias-variance trade-off, this above formulation also naturally balances between obtaining an approximation close to the observation and overfitting to the noise in the observed image. Therefore, given any noisy image, we can obtain a reconstructed image that minimizes the MSE under normally distributed noise. Furthermore, if we were given a training set of K noisy observations $(y_k)_{k=1}^K$, we can train a denoiser that generalizes to the class of images contaminated with normally distributed noise by minimizing the sum of these individual risk estimates.

While SURE acts as a natural device for image deblurring in the absense of ground truth noise-less training images, there is one major challenge of this approach. Namely, it is difficult to compute the gradient of f_θ with respect to y, often referred to as the divergence. Many of the modern-day image denoisers are neural networks with extremely complicated structure that makes computing its gradient by hand difficult (Dong et al., 2014; Yang et al., 2017; Zhang et al., 2017). However, to optimize the unbiased risk estimate over θ using automatic differentiation as proposed in Metzler et al. (2018), direct computation or approximation of the gradient of f_θ with respect to y is often required. This is because nested automatic differentiation is not typically supported in existing packages. This issue is addressed by resorting to finite-difference type approximation of the divergence term introduced in MC-SURE (Ramani et al., 2008). However, this method still requires the user to specify the spacing parameter ϵ, and the effect of ϵ on the resulting image denoisers remains unexplored.

2 Mini-proposals

2.1 Proposal 1: Using SURE to automate cross validation for principal component regression with L1 regularization

Principal component regression (PCR), a regression method based on principal components (PCs) obtained from principal component analyses (PCA), is commonly used in genomics applications among others. This is because, rather than the effect of individual gene expressions, practitioners are typically more interested in studying the effect of groups of genes, which may describe more complicated processes (Ding et al., 2022).

The general procedure of PCR is typically as follows. Suppose we are given a centred data matrix $X \in \mathbb{R}^{n \times p}$ and corresponding responses $y \in \mathbb{R}^n$ consisting of n observations each with p predictors. Let $X = U\Sigma V^T$ denote the singular value decomposition of the data matrix X, then $W = XV$ denotes the set of all PCs, a set of transformed feature vectors that form an orthonormal basis, ordered by the amount of variance from the data that each PC explains. One can then perform linear regression on the PCs against the response y. Because the PCs are orthogonal to each other, this approach usually leads to better numerical stability than, for example, ordinary least squares.

When the number of predictors is large, a common practice is to perform PCR using only the top k PCs ($k \ll p$), i.e., the PCs that explain the most variance present in the data (Cera et al., 2019; Harel et al., 2019). However, it is important to note that the PCs are fit without knowledge of the response variable, and so the idea that performing PRC using the first k PCs will lead to good a good fit of the data is merely a heuristic. In fact, Jolliffe (1982) provides a number of real examples where performing PCR using the first few PCs that explain the most variance of the data is indeed suboptimal.

As a result, it is desirable to do feature selection among the PCs so that we can balance computational cost and the quality of our PCR model. From the previous section, we know that the SURE can be used as a model selection tool for selecting the regularization parameter for ridge regression. It is then natural to look into extending this framework to LASSO regression on the PCs, which can help us pick the PCs to include in the regression while taking into account the bias-variance trade-off.

Given the transformed data matrix W, a response vector y consisting of n observations, and a regularization parameter $\lambda \geq 0$, LASSO regression finds the regression coefficients of the form

$$\hat{\beta}_{L,\lambda}(y) = \arg\min_{\beta} \frac{1}{n}\|W\beta - y\|_2^2 + \lambda\|\beta\|_1.$$

This L1 penalty term encourages sparse solutions where the regression coefficient for some the transformed predictors are set to 0. To use the SURE framework to select a subset of the PCs, we require that 1) the resulting estimator $\hat{\beta}_{L,\lambda}$ to be available in closed-form, and 2) we can compute the derivative of this estimator with respect to y.

In the context of PCR, we know that W is an orthonormal matrix. We know that when the data matrix is orthonormal, there is indeed a closed-form solution for the LASSO estimator

$$\hat{\beta}_{L,\lambda} = \begin{bmatrix} \text{sgn}(\hat{\beta}_1)(|\hat{\beta}_1| - \lambda)_+ & \cdots & \text{sgn}(\hat{\beta}_p)(|\hat{\beta}_p| - \lambda)_+ \end{bmatrix}^T,$$

where $\hat{\beta}$ denotes the solution of the ordinary least square problem (Gauraha, 2018). Note that $\hat{\beta}_{L,\lambda}$ is implicitly a function of the response y. Furthermore, Tibshirani and Wasserman (2015) provides a derivation that shows the divergence term in the SURE expression equals the number of predictors whose corresponding regression coefficient is nonzero (i.e. $\|\hat{\beta}_{L,\lambda}\|_0$). As a result, we can write the SURE of the LASSO estimator on a set of PCs as

$$\hat{R}(\lambda) = -n\sigma^2 + \|y - \hat{\beta}_{L,\lambda}(y)\|^2 + 2\sigma^2\|\hat{\beta}_{L,\lambda}\|_0.$$

It now remains to find the optimal λ and subsequently the set of PCs with nonzero regression coefficients. Since each dimension of the LASSO estimator, denoted $\hat{\beta}_{L,\lambda,i}$ is non-smooth at $\hat{\beta}_{L,\lambda,i} = \lambda$, we can use subgradient methods in place of regular gradient descent to obtain the optimal λ (Shor, 2012). Instead of selecting the top k PCs to perform PCR, where k is chosen somewhat arbitrarily, an SURE-inspired variable selection procedure proposed here may help achieve a better quality PCR fit with a similar level of reduction in computational cost.

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

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

We can now write

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

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.

Isabella Cera, Laura Whitton, Gary Donohoe, Derek W Morris, Georg Dechant, and Galina Apostolova. Genes encoding satb2-interacting proteins in adult cerebral cortex contribute to human cognitive ability. *PLoS Genetics*, 15(2):e1007890, 2019.

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

Lei Ding, Gabriel E Zentner, and Daniel J McDonald. Sufficient principal component regression for pattern discovery in transcriptomic data. *Bioinformatics advances*, 2(1): vbac033, 2022.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

David L Donoho and Iain M Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, 90(432):1200–1224, 1995.

Niharika Gauraha. Introduction to the lasso. *Resonance*, 23(4):439–464, 2018.

Tom Harel, Naama Peshes-Yaloz, Eran Bacharach, and Irit Gat-Viks. Predicting phenotypic diversity from molecular and genetic data. *Genetics*, 213(1):297–311, 2019.

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.

Ian T Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303, 1982.

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.

Florian Luisier, Thierry Blu, and Michael Unser. Image denoising in mixed poisson–gaussian noise. *IEEE Transactions on image processing*, 20(3):696–708, 2010.

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

Christopher A Metzler, Ali Mousavi, Reinhard Heckel, and Richard G Baraniuk. Unsupervised learning with stein's unbiased risk estimator. *arXiv preprint arXiv:1805.10531*, 2018.

Bala Kishore Panisetti, Thierry Blu, and Chandra Sekhar Seelamantula. An unbiased risk estimator for multiplicative noise—application to 1-d signal denoising. In *2014 19th International Conference on Digital Signal Processing*, pages 497–502. IEEE, 2014.

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

Sathish Ramani, Thierry Blu, and Michael Unser. Monte-carlo sure: A black-box optimization of regularization parameters for general denoising algorithms. *IEEE Transactions on image processing*, 17(9):1540–1554, 2008.

Naum Zuselevich Shor. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012.

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.

Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6721–6729, 2017.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017.