CHAPTER 1

Estimation of the Loss of an Estimate

Wing Hung Wong

Department of Statistics
Stanford University
Stanford, CA 94305-4065

1. Introduction

Suppose $y_1, \ldots, y_n$ are independent random variables. The density $p_{\theta_i}(\cdot)$ of y_i is supposed to be known up to a parameter θ_i. Let $\hat{\boldsymbol{\theta}} = \big(\hat{\theta}_1(\boldsymbol{y}), \ldots, \hat{\theta}_n(\boldsymbol{y})\big)$ be an estimate of $\boldsymbol{\theta}$ constructed from the sample $\boldsymbol{y} = (y_1, \ldots, y_n)$. Note that the estimate of θ_i can depend on y_j, $j \neq i$. Let $\ell_i(\cdot,\cdot)$ be a loss function so that $\ell_i(\theta_i, \hat{\theta}_i)$ represents the loss of using $\hat{\theta}_i$ as the estimate of θ_i. The purpose of the present paper is to introduce some methods for the estimation of the average loss $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum\limits_{1}^{n} \ell_i(\theta_i; \hat{\theta}_i)$.

Let $\boldsymbol{y}_{(-i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ be the sample with y_i deleted, and write

$$g_i(y_i) = \hat{\theta}_i(y_i; \boldsymbol{y}_{(-i)}),$$

then $$\ell_i(\theta_i, \hat{\theta}_i) = \ell_i\big(\theta_i; g_i(y_i)\big).$$

Conditional on $\boldsymbol{y}_{(-i)}$, $g_i(\cdot)$ is a known function. This leads us to a one-dimensional estimation problem: Given the known functions $\ell_i(\cdot,\cdot)$ and

$g_i(\cdot)$, and an observation y_i from $p_{\theta_i}(\cdot)$, find an estimate $\hat{\ell}_i(y_i)$ of the quantity $\ell_i\big(\theta_i, g_i(y_i)\big)$. One obvious possibility is to estimate $\ell_i\big(\theta_i, g_i(y_i)\big)$ by a Bayes estimate, i.e. setting $\hat{\ell}_i(y_i)$ to be

$$\frac{\int \ell(\theta_i, g_i(y_i) p_{\theta_i}(y_i)\pi(\theta_i)d\theta_i}{\int p_{\theta_i}(y_i)\pi(\theta_i)d\theta_i}$$

where $\pi(\theta_i)$ is a prior for θ_i. This would be a reasonable procedure if n is small and each y_i is strongly informative on θ_i.

In this paper, however, we are mainly interested in the situation when n is large and each y_i by itself is not strongly informative on θ_i. In this case, it is desirable to require $\hat{\ell}_i(y_i)$ to be an unbiased estimator of $\ell_i\big(\theta_i, g_i(y_i)\big)$.

To cite a scenario where unbiasedness is clearly the appropriate requirement, suppose $\{\hat{\theta}_i, i = 1, \ldots, n\}$ are weakly dependent random variables in the sense that each $\hat{\theta}_i$ is approximately independent of most but a small fraction of the other $\hat{\theta}_j$'s. Conditional on $\boldsymbol{y}_{(-i)}$, let $t_i = t\big(y_i; g_i(\cdot)\big)$ be an estimator of $\ell_i\big(\theta_i, g_i(y_i)\big)$, and

$$r_i = t_i - \ell_i\big(\theta_i, g_i(y_i)\big),$$

then $\{r_i, i = 1, \ldots, n\}$ are also weakly dependent variables. Hence $\mathrm{Var}\left(\frac{1}{n}\sum_1^n r_i\right) \to 0$, and

$$\begin{aligned} &E\left(\frac{1}{n}\sum_1^n t_i - \frac{1}{n}\sum_1^n \ell_i(\theta_i, \hat{\theta}_i)\right)^2 \\ &= \mathrm{Var}\left(\frac{1}{n}\sum_1^n r_i\right) + \left(\frac{1}{n}\sum_1^n E(r_i)\right)^2 \end{aligned}$$

will be determined by the average biases of the t_i's. In this case, unbiasedness is clearly desirable if it can be achieved. Otherwise, we should try to keep the bias of $\hat{\ell}_i$ small over a reasonable range of values of θ_i. It is, however, important to note that the variance of $\frac{1}{n}\sum_1^n r_i$ can be small under much more general conditions than weak dependency of the $\hat{\theta}_i$'s. This will be discussed in section 6.

Thus, for each i, we need to find unbiased or approximately unbiased estimator of $\ell_i\big(\theta_i, g_i(y_i)\big)$. If there is more than one unbiased estimator, we choose the one giving, in some sense, the smallest value for $E_{\theta_i}\big(\hat{\ell}_i(y_i) - \ell_i(\theta_i, g_i(y_i)\big)^2$.

In many applications, we are interested in the loss of $\hat{\theta}_i$ only to compare it to the loss of another estimator $\tilde{\theta}_i$. It is clear that in this case it suffices to compare $\ell_i(\theta_i, \hat{\theta}_i)+k(\theta_i)$ to $\ell(\theta_i, \tilde{\theta}_i)+k(\theta_i)$ where $k(\cdot)$ is a constant function of θ_i. We call such a function $\ell_i(\theta_i, \hat{\theta}_i) + k(\theta_i)$ a comparative loss function. For comparison among estimators, it is enough to find unbiased estimates of their comparative losses (corresponding to a common $k(\cdot)$). We then have the freedom of choosing $k(\cdot)$ to make it easy to construct unbiased estimates.

As a final remark, we note that if the family of densities $\{p_{\theta_i}(\cdot)\}$ is a complete family as θ_i vary in its range, then an unbiased estimator of the loss (or a comparative loss), must be unique if it exists.

2. Kullback-Leibler loss and exponential families

Suppose y_i has density $p_{\theta_i}(\cdot)$ and $g_i(\cdot)$ is a given function of y_i as defined in the introduction. To simplify notations, we will suppress the subscript i in the rest of this section. The Kullback-Leibler pseudo-distance between two densities $p(\cdot)$ and $q(\cdot)$ are defined by

$$K(p, q) = \int p \log \frac{p}{q}\, dy.$$

The Kullback-Leibler loss (KL loss) of the estimator $g(y)$ is then defined by

$$\ell\big(\theta, g(y)\big) = K(p_\theta, p_{g(y)}).$$

Let $p_\theta(z)$ be an exponential family distribution, then

$$\log p_\theta(z) = \phi(\theta)t(z) + \alpha(\theta) + m(z)$$

for some functions $\phi(\cdot)$, $\alpha(\cdot)$ of θ and $t(\cdot)$, $m(\cdot)$ of z, and $\ell\big(\theta, g(y)\big) = \phi(\theta)\mu(\theta)+\alpha(\theta)-\phi\big(g(y)\big)\mu(\theta)-\alpha\big(g(y)\big)$, where $\mu(\theta) = E_\theta t(z)$. Even though the exponential family structure leads to a relatively simple form for the KL loss, in most cases it is still not possible to find exactly unbiased estimate of this loss. We will discuss the construction of approximately unbiased estimates in a later section. However, a corresponding comparative loss function $\ell\big(\theta, g(y)\big) - \phi(\theta)\mu(\theta) - \alpha(\theta) = -\phi\big(g(y)\big)\mu(\theta) - \alpha\big(g(y)\big)$ is particularly simple.

To estimate this comparative loss, we only need to solve the following problem:

- (i) Find a function $h(y)$ so that
$$E_\theta h(y) = \mu(\theta) E_\theta \phi\big(g(y)\big) \qquad \text{for all } \theta. \tag{$*$}$$
- (ii) If there exist more than one solution to $(*)$, choose the one that minimizes
$$\int E_\theta \big(h(y) - \mu(\theta)\phi\big(g(y)\big)\big)^2 \pi(\theta) d\theta$$
where $\pi(\cdot)$ is an appropriate weight function.

If a function $h(\cdot)$ satisfying $(*)$ can be found, then $-h(y) - \alpha\big(g(y)\big)$ will be an unbiased estimator of the comparative KL loss $\ell(\theta, g(y)\big) - \phi(\theta)\mu(\theta) - \alpha(\theta)$. Exact solutions to the above problem can be found in several important exponential family models. We list two examples.

Example 1. (Poisson distribution)

Suppose y has a Poisson distribution with mean θ, then $\log p_\theta(y) = y\log(\theta) - \theta - y!$

Hence $\mu(\theta) = E_\theta(y) = \theta$, $\phi(\theta) = \log(\theta)$, $\alpha(\theta) = -\theta$. To obtain an unbiased estimate of the comparative KL loss, notice that

$$\begin{aligned}
\mu(\theta)E_\theta\phi\big(g(y)\big) &= \theta \sum_{y=0}^{\infty} \log\big(g(y)\big) \cdot e^{-\theta}\theta^y / y! \\
&= \sum_{z=1}^{\infty} z \log\big(g(z-1)\big) \frac{e^{-\theta}\theta^z}{z!} \qquad (z = y+1) \\
&= E_\theta y \log\big(g(y-1)\big).
\end{aligned}$$

It follows that $y\log\big(g(y-1)\big)$ is an unbiased estimate of $\mu(\theta)\phi\big(g(y)\big)$, and $g(y) - y\log\big(g(y-1)\big)$ is an unbiased estimator of the comparative KL loss

$K(p_\theta, p_{g(y)}) - \theta\log(\theta) + \theta$. It is the unique unbiased estimator because the Poisson family is complete.

EXAMPLE 2. (Gamma scale family)

Suppose y has a Gamma distribution with a known shape parameter k and an unknown scale parameter θ, then

$$\log p_\theta(y) = -\frac{y}{\theta} - k\log(\theta) + \log\big(y^{k-1}/\Gamma(k)\big).$$

Hence

$$\mu(\theta) = E_\theta(y) = k\theta, \quad \phi(\theta) = -\frac{1}{\theta}, \quad \alpha(\theta) = -k\log(\theta).$$

To obtain an unbiased estimate of $\mu(\theta)E_\theta\phi\big(g(y)\big)$, notice that

$$\begin{aligned}\mu(\theta)E_\theta\phi\big(g(y)\big) &= -k\theta E_\theta\left(\frac{1}{g(y)}\right)\\ &= -\frac{k}{\Gamma(k)\theta^{k-1}}\int_0^\infty \left(\frac{y^{k-1}}{g(y)}\right)(e^{-\frac{y}{\theta}})\,dy\\ &= -\frac{k}{\Gamma(k)\theta^{k-1}}\left\{\Big[v(y)e^{-\frac{y}{\theta}}\Big]_0^\infty - \int_0^\infty v(y)\left(-\frac{1}{\theta}\,e^{-\frac{y}{\theta}}\right)dy\right\}\end{aligned}$$

where $v(y) = \int_0^y \frac{z^{k-1}}{g(z)}\,dz$. We assume that $g(z)$ does not converge to zero faster than z^k as $z \to 0$, so that $v(y)$ exists and $\Big[v(y)e^{-\frac{y}{\theta}}\Big] \to 0$ as $z \to 0$ or $z \to \infty$. Then it follows that

$$E_\theta\mu(\theta)\phi\big(g(y)\big) = E_\theta\big(-ky^{-(k-1)}v(y)\big).$$

Hence $ky^{-(k-1)}v(y) + k\log\big(g(y)\big)$ is an unbiased estimate of the comparative KL loss

$$K(p_\theta, p_{g(y)}) + 1 + k\log(\theta).$$

It is the unique unbiased estimate because of the completeness of the Gamma scale family.

3. Mean square error loss

The mean square error loss (MLE loss) of an estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ is defined by

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{1}^{n} \big(\hat{\theta}_i(\boldsymbol{y}) - \theta_i\big)^2.$$

Write $\hat{\theta}_i(y) = g_i(y_i)$ where $g_i(\cdot)$ is the (random) function of y_i defined in the introduction, then each term in $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is of the form

$$\big(g(y) - \theta\big)^2 = g(y)^2 - 2\theta g(y) + \theta^2,$$

where, for simplicity, the subscript i has been suppressed from g_i and θ_i. Hence, $g(y)^2 - 2\theta g(y)$ is a comparative MSE loss for $g(\cdot)$ and an unbiased estimator of it is of the form $g(y)^2 - 2e(y)$ where $e(\cdot)$ satisfies the equation

$$E_\theta e(y) = \theta E_\theta g(y) \tag{3.1}$$

for all θ. If, furthermore, there is an unbiased estimator $f(y)$ of the term θ^2, then

$$g(y)^2 - 2e(y) + f(y)$$

is an unbiased estimate of the MSE.

Example 3.

Suppose y has a Poisson distribution with mean θ. By the same argument as used in Example 1, it is seen that $e(y) = yg(y-1)$ is an unbiased estimator of $\theta E_\theta g(y)$. Furthermore, it is easy to check that $f(y) = y^2 - y$ is an unbiased estimator of θ^2. Hence

$$g(y)^2 - 2yg(y-1) + y^2 - y$$

is the unique unbiased estimator of $\big(g(y) - \theta\big)^2$.

Example 4.

Suppose y has a Gamma(k, θ) distribution with a known shape parameter k. The arguments used in examples 2 and 3 can be used to calculate an unbiased estimate of $\big(g(y) - \theta\big)^2$. The resulting estimate is

$$g(y)^2 - 2y^{-(k-1)}v(y) + \frac{y^2}{k + k^2}$$

where

$$v(y) = \int_0^y z^{k-1} g(z) dz.$$

4. Location families

Suppose y_i has density $p_i(y_i - \theta_i)$ where, for each i, $p_i(\cdot)$ is a known density on $\mathbb{R}$ with mean zero. Suppressing the subscript i, we write

$$y = \theta + \epsilon$$

where ϵ has density $p(\cdot)$ and satisfies $E(\epsilon) = 0$. It was observed in the last section that to estimate the MSE of an estimator $g(y)$, we need to construct unbiased estimates of $\theta g(y)$ and θ^2. Unbiased estimation of θ^2 is easy: we may use $f(y) = y^2 - \sigma^2$ where $\sigma^2 = \mathrm{Var}(\epsilon) = \int \epsilon^2 p(\epsilon) d\epsilon$. To construct an unbiased estimator of $\theta g(y)$, observe that

$$E\theta g(y) = Eyg(y) - E\epsilon g(\theta + \epsilon).$$

Thus it suffices to find an unbiased estimator $h(y)$ of the term $E\epsilon g(\theta + \epsilon)$, i.e. to find a function $h(\cdot)$ to satisfy the following equation for all θ

$$\int h(\theta + \epsilon)\, p(\epsilon) d\epsilon = \int g(\theta + \epsilon)\, \epsilon p(\epsilon) d\epsilon. \tag{4.1}$$

The solution to this integral equation, if it exists, can be obtained in the following way.

Let $H(\cdot)$, $P(\cdot)$, $G(\cdot)$ be the Fourier transforms of $h(\cdot)$, $p(\cdot)$ and $g(\cdot)$ respectively. For example,

$$P(s) = (\mathcal{F}p)(s) = \int p(\epsilon) e^{-i2\pi\epsilon s} d\epsilon.$$

Let $q(\epsilon) = p(-\epsilon)$, then

$$\int h(\theta + \epsilon)\, p(\epsilon) d\epsilon = h * q(\theta)$$

where $*$ denotes the convolution operation. The Fourier transform of $h*q(\theta)$ is $H(s) \cdot Q(s) = H(s)\bar{P}(s)$ where $\bar{P}(s)$ is the complex conjugate of $P(s)$. Similarly transforming the right hand side of (4.1), and using the fact that the Fourier transform of $\epsilon p(\epsilon)$ is $(i/2\pi)P'(s)$, it is seen that the integral equation (4.1) is equivalent to

$$H(s)\bar{P}(s) = -\frac{i}{2\pi}\,G(s)\bar{P}'(s) \tag{4.2}$$

In other words, a solution $h(\cdot)$ exists for (4.1) iff we can find a L_1 function $H(\cdot)$ which satisfies (4.2) for all s. In particular, if $P(s) \neq 0$ for all s then h is determined uniquely as

$$h = -\frac{1}{2\pi}\,\mathcal{F}^{-1}(iG\bar{P}'/\bar{P}). \tag{4.3}$$

In general, $p(s)$ may vanish for some values of s and (4.3) cannot be used. In this case, we can get approximate solutions of (4.1) by the following device: first approximate $p(\cdot)$ by another density $p_1(\cdot)$ which has a nonvanishing Fourier transform, then compute $h = -\frac{1}{2\pi}\mathcal{F}^{-1}(iG\bar{P}'{}_1/\bar{P}_1)$ and regard it as an approximate solution of (4.1). One possible choice of p_1 is $p_1(\epsilon) = (1-\alpha)p(\epsilon) + \alpha\phi(\epsilon)$ where $\phi(\cdot)$ is a normal density. The above choice of h will then satisfy (4.1) with p replaced by p_1:

$$\int h(\theta+\epsilon)p_1(\epsilon)d\epsilon = \int g(\theta+\epsilon)\,\epsilon p_1(\epsilon)\,d\epsilon. \tag{4.2}$$

$$\left[\int h(\theta+\epsilon)p(\epsilon)d\epsilon - \int g(\theta+\epsilon)\epsilon p(\epsilon)d\epsilon\right] \qquad \text{i.e.}$$

$$= \frac{\alpha}{(1-\alpha)}\left[-\int h(\theta+\epsilon)\phi(\epsilon)d\epsilon + \int g(\theta+\epsilon)\epsilon\phi(\epsilon)d\epsilon\right].$$

Thus the bias of this choice of $h(\cdot)$ is of order α. Typically, if α is chosen too small then $h(\cdot)$ will have high variance. In practice, one needs to choose each α_i (in estimating $(g_i-\theta_i)^2$) carefully in order to achieve a good bias/variance trade-off in the estimation of $\frac{1}{n}\sum(\hat{\theta}_i-\theta_i)^2$.

Example 5. (Normal location model).

Let ϵ be $N(0,1)$, then

$$p(\epsilon) = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}\epsilon^2} \qquad \text{and} \qquad P(s) = e^{-2\pi^2 s^2}.$$

Hence $p'(s) = -(2\pi)^2 sP(s)$ and

$$H(s) = i(2\pi)sG(s) = (\mathcal{F}g')(s).$$

It follows that $h(x) = g'(x)$, which leads to the unique unbiased estimate (of the MSE)

$$g(y)^2 - 2\big[yg(y) - g'(y)\big] + (y^2 - 1) = \big(g(y) - y\big)^2 + \big(2g'(y) - 1\big).$$

If the variance of ϵ is σ^2, then the term $(2g' - 1)$ should be multiplied by σ^2. Returning to the whole vector $\hat{\boldsymbol{\theta}} = \big(\hat{\theta}_1(\boldsymbol{y}), \ldots, \hat{\theta}_n(\boldsymbol{y})\big) = \big(g_1(y_1), \ldots, g_n(y_n)\big)$, the above result then leads to

$$\frac{1}{n}\sum_1^n (\hat{\theta}_i - y_i)^2 + \frac{2}{n}\sum_1^n \frac{\partial}{\partial y_i}\,\hat{\theta}_i - \sigma^2$$

as an unbiased estimator of the MSE loss $\frac{1}{n}\sum\limits_1^n(\hat{\theta}_i - \theta_i)^2$. This estimate was first obtained in Stein (1981) as an unbiased estimate of the MSE risk $\frac{1}{n}\sum\limits_1^n E(\hat{\theta}_i - \theta_i)^2$.

Example 6. (Symmetric stable distributions)

Let $p(\cdot)$ be a symmetric stable density with a known scale parameter, then its Fourier transform is given by

$$P(s) = e^{-c|s|^\alpha},$$

where c depends on α and the scale parameter. We assume that the index α is known and $\alpha \in (1, 2]$. It follows that $P'(s) = -\text{sign}(s)\alpha c|s|^{\alpha-1}P(s)$, and, by (4.2),

$$\begin{aligned} H(s) &= \alpha c\big(|s|^{-(2-\alpha)}\big) \cdot \left(\frac{i}{2\pi}\,\text{sign}(s)|s|G(s)\right) \\ &= \alpha c\big(|s|^{-(2-\alpha)}\big) \cdot \left(\frac{i}{2\pi}\,sG(s)\right). \end{aligned}$$

Hence $h(x) = \alpha c(g' * t)(x)$

$$\begin{aligned} t(x) &= \mathcal{F}^{-1}\big(|s|^{-(2-\alpha)}\big)(x) \qquad\qquad \text{where} \\ &= \big[2^{2-\alpha}\pi^{-(\alpha-1)}\Gamma(\alpha - 1)\sin\big((2 - \alpha)\pi/2\big)\big] \cdot |x|^{-(\alpha-1)}. \end{aligned}$$

It is interesting to note that if $\alpha < 2$ then θ^2 cannot be estimated by $f(y) = y^2 - \text{Var}(\epsilon)$ because $\text{Var}(\epsilon)$ is infinite. However, the comparative loss $\big(g(y) - \theta\big)^2 - \theta^2$ can still be estimated by

$$g(y)^2 - 2yg(y) + 2\alpha c \cdot g' * t(y).$$

5. Approximate solutions

In general, exactly unbiased approximation to the loss may not exist. This typically happens when the family $\{p_\theta(\cdot), \theta \in \Theta\}$ is rich but the range $\mathcal{Y}$ of the random variable y is small. In such cases, one has to be satisfied with an estimator which is in some sense close to being unbiased for the loss $\ell_i(\theta, \hat{\theta})$. For example, suppose we want to find a function $e(\cdot)$ to satisfy (3.1). A generally applicable method is as follows: Let $e(\cdot) = \sum_1^\infty c_i f_i(\cdot)$ where $\{f_i(\cdot), i = 1, 2, \dots\}$ is a set of basis functions in a certain space of functions on $\mathcal{Y}$. Let $\gamma(\theta) = \theta E_\theta g(y)$. We can determine the coefficients c_i's so that (3.1) is approximately satisfied in a certain sense. For example

- (a) Restricted unbiasedness: choose a subset $\Theta_0 \subset \Theta$ and requires $e(\cdot)$ to satisfy (3.1) for all $\theta \in \Theta_0$. In particular, if Θ_0 is finite, we may approximate $e(\cdot)$ by taking the first m terms in the expansion and attempt to solve for the coefficients $c_1, \dots, c_m$ in the linear system

$$\sum_1^m c_j \big(E_{\theta_i} f_j(y)\big) = \gamma(\theta_i), \qquad \theta_i \in \Theta_0 = \{\theta_1, \cdots, \theta_m\}.$$

- (b) Least squares solution: Suppose $\gamma(\theta)$ and $\phi_i(\theta) = E_\theta f_i(y)$ $i = 1, 2, \dots$ are all elements of a L_2 space with inner product $\langle \phi, \gamma \rangle = \int \phi(\theta)\gamma(\theta)\, d\mu(\theta)$ where $\mu(\cdot)$ is an appropriate measure on Θ. We may then determine $c_1, \dots, c_m$ by projecting $\gamma(\cdot)$ into the space spanned by $\phi_1(\cdot), \dots, \phi_m(\cdot)$.

Although this approach can be implemented numerically in almost any problem, the degree to which $e(\cdot)$ is "approximately unbiased" must be investigated in each application. We give two examples.

Example 7. (Location family with bounded error)

As discussed in section 4, the key to finding an unbiased approximation to the MSE is the solution of the integral equation (4.1), or equivalently (4.2). Unfortunately, if the error density $p(\cdot)$ has bounded support, then its Fourier transform $P(\cdot)$ may have isolated zeros and (4.2) may not be satisfied by any $H(\cdot)$. For example, if $p(\cdot)$ is the triangular density, i.e. $p(x) = 1 - |x|$, $|x| \leq 1$, then $P(s) = \big[\sin(\pi s)/\pi\big]^2$, and for $s \neq 0$,

$$P'(s)/P(s) = 2\pi\big[\cot(\pi s) - 1/\pi s\big].$$

In this case, the "formal" solution (4.3) will have singularities at $s = \pm 1, \pm 2, \ldots$.

Thus, in many applications the equation (4.1) cannot be satisfied for all θ. A method for constructing approximate solutions has already been given in section 4. We now describe another method which can often be used to construct a function $h(\cdot)$ which satisfies (4.1) for all $\theta \in [-L, L]$ where L is a suitably large constant.

Suppose the support of $p(\cdot)$ is contained in an interval $[-\delta, \delta]$, and $T > L + 2\delta$. Let $\tilde{g}(\cdot)$ be a periodic function with period $2T$ such that $\tilde{g}(y) = g(y)$ for $y \in [-L-\delta, L+\delta]$. It is easy to check that $E_\theta \tilde{g}(y) = E_\theta g(y)$ for all $\theta \in [-L, L]$. Thus it suffices to consider the problem of finding a periodic function $h(\cdot)$ (with period $2T$) to satisfy the equation

$$\int h(\theta + \epsilon)\, p(\epsilon)\, d\epsilon = \int \tilde{g}(\theta + \epsilon)\, \epsilon p(\epsilon)\, d\epsilon$$

for all $\theta \in (-\infty, \infty)$. Since $h(\cdot)$ is periodic, we can expand it in Fourier series:

$$h(x) = \sum_{n=-\infty}^{\infty} H_n e^{i2\pi s_n x}$$

where $s_n = \frac{n}{T}$, $n = 0, \pm 1, \pm 2, \cdots$

and $$H_n = \frac{1}{2T} \int_{-T}^{T} h(x) e^{-i2\pi s_n x} dx.$$

Similarly, $\tilde{g}(x) = \sum_n \tilde{G}_n e^{i2\pi s_n x}$. Putting these into the integral equation and equating coefficients, we have

$$H_n \cdot \bar{P}(s_n) = \tilde{G}_n \cdot \bar{R}(s_n) \tag{5.1}$$

where $P(s_n) = \int_{-\infty}^{\infty} p(\epsilon)e^{-i2\pi s_n \cdot \epsilon} d\epsilon$ and $R(s_n) = \int_{-\infty}^{\infty} \epsilon p(\epsilon)e^{-i2\pi s_n \cdot \epsilon} d\epsilon = \frac{i}{2\pi}P'(s_n)$ are the Fourier transforms of $p(\epsilon)$ and $\epsilon p(\epsilon)$. To solve (5.1) for H_n, we must make sure that $P(s_n) \neq 0 \quad \forall n$. This is usually achievable by choosing T appropriately. For example, if $p(\cdot)$ is the triangular density, then $P(s)$ vanishes only at $s = \pm 1, \pm 2, \ldots$. Thus, we need to make sure that, for all integers $n = \pm 1, \pm 2, \ldots$, $s_n = \frac{n}{T}$ is not a non-zero integer. Theoretically, any positive irrational T will do.

From the point of view of controlling bias, one would like to choose L, (and hence T) as large as possible. However, there is a price to be paid: if T is too large then some of the values of $s_n = \frac{n}{T}$ will be very close to the zeros of $P(s)$ at $s = \pm 1, \pm 2, \ldots$. Consequently, the corresponding values of H_n will be very large, leading to an estimator $h(\cdot)$ with very high conditional variance. The choice of T_i from the considerations of bias/variance trade-off in the estimation of the average MSE $n^{-1}\sum_1^n(\hat{\theta}_i - \theta_i)^2$ is an interesting question which, however, will not be discussed further in this paper.

Example 8. (Binomial distribution)

Suppose each y_i has a Binomial (m_i, θ_i) distribution. From section 3, we know that the construction of unbiased approximation to the MSE depends on the solution of the following problem: Find $e(\cdot)$ such that

$$E_\theta e(y) = \theta E_\theta g(y) \qquad \forall \quad \theta \in [0,1] \tag{5.2}$$

where y is a Binomial (m, θ) variable.

We will see that for a large class of functions $g(\cdot)$, there are exact solutions to (5.2). Furthermore, even when (5.2) is not exactly solvable, we can often construct $e(\cdot)$'s which satisfy (5.2) to a high degree of accuracy.

Let us represent the functions $e(y)$ and $g(y)$ by the vectors $v_e = (e_0, \ldots, e_m)$ and $v_g = (g_0, \ldots, g_m)$. Choosing $e_0 = 0$ and dividing both sides of (5.2) by θ, we have

$$\sum_{i=0}^{m-1} e_{i+1}\binom{m}{i+1}\theta^i(1-\theta)^{m-1-i} = \sum_{i=0}^{m} g_i\binom{m}{i}\theta^i(1-\theta)^{m-i} \tag{5.3}$$

The first term in the right hand side can be expanded in the following manner:

$$\begin{aligned} g_0(1-\theta)^m &= g_0\big[(1-\theta)^{m-1} - \theta(1-\theta)^{m-1}\big] \\ &= g_0\big[(1-\theta)^{m-1} - \theta(1-\theta)^{m-2} + \theta^2(1-\theta)^{m-2}\big] = \cdots \\ &= g_0\big[(1-\theta)^{m-1} - \theta(1-\theta)^{m-2} + \cdots + (-1)^{m-1}\theta^{m-1} + (-1)^m\theta^m\big]. \end{aligned}$$

Expanding the other terms similarly, we obtain after some calculation that

$$\sum_{i=0}^{m} g_i\binom{m}{i}\theta^i(1-\theta)^{m-i} = \left[\sum_{i=0}^{m-1} \langle c_i, v_g\rangle\theta^i(1-\theta)^{m-1-i}\right] + \langle c_m, v_g\rangle\theta^m \tag{5.4}$$

where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathbb{R}^{m+1}$ and the vectors $c_i, i = 0, \ldots, m$ are defined by the relations

$$\langle c_i, v_g\rangle = \sum_{j=0}^{i} (-1)^j\binom{m}{i-j} g_{i-j}.$$

It can be checked that $\langle c_m, v\rangle = 0$ if v is a vector representing any of the monomials in y of degree $\le m-1$. If $g(y)$ is a polynomial in y of degree at most $m-1$, then $\langle c_m, v_g\rangle = 0$ and it follows that (5.2) is satisfied if we set $e(0) = 0$ and

$$e(i+1) = \frac{\langle c_i, v_g\rangle}{\binom{m}{i+1}} \qquad i = 0, \ldots, m-1. \tag{5.5}$$

Hence, we have the result that, if $g(\cdot)$ is a degree $m-1$ polynomial, then an unbiased estimate of $E\big(g(y)-\theta\big)^2$ is $g(y)^2 - 2e(y) + (y^2-y)/(m^2-m)$.

In general, for an arbitrary $g(\cdot)$, we have $\langle c_m, g\rangle = \|u_g\|^2$ where u_g is the component of v_g perpendicular to the vectors representing polynomials of degree $\le m-1$. If $g(\cdot)$ is any reasonable estimator of θ, this component u_g should be very small. In this general case, the estimator (5.5) can be improved in the following way. Let $\nu(\cdot)$ be an appropriate measure on $[0,1]$ such that θ^m and $\theta^i(1-\theta)^{(m-1)-i}$, $i = 0, \ldots, m-1$, are all square integrable w.r.t. $\nu(\cdot)$. Let $\theta^m = s(\theta) + r(\theta)$ where $s(\theta) = \sum_{i=0}^{m-1} \alpha_i \theta^i (1-\theta)^{(m-1)-i}$ is the projection of θ^m onto the space spanned by $\theta^i(1-\theta)^{m-1-i}$, $i = 0, \ldots, m-1$. Then the expression (5.5) should be modified to

$$e(i+1) = \big(\langle c_i, v_g\rangle + \langle c_m, v_g\rangle \alpha_i\big) \Big/ \binom{m}{i+1}. \tag{5.6}$$

In this case, $e(y)$ is not exactly unbiased for $\theta E_\theta g(y)$, the bias is $\|u_g\|^2 \theta r(\theta)$. The L_2 norm (w.r.t. ν) of this bias is often very small. For example, with $m = 3$, $\nu(\cdot) =$ Lebesque measure, exact calculation shows that the L_2 norm of $r(\theta)$ is 0.0189. If m is larger, the norm of $r(\cdot)$ would be much smaller.

6. Convergence of the loss estimate

In the preceeding sections we have provided constructions of unbiased (or nearly unbiased) estimator of $\frac{1}{n}\sum_1^n \ell_i(\theta_i, \hat{\theta}_i)$. The estimator is of the form $\frac{1}{n}\sum_1^n t_i(\boldsymbol{y})$ where $t_i(\boldsymbol{y})$ is determined by the form of $\hat{\theta}_i(y_1, \ldots, y_n)$ as an univariate function of y_i. The error of this estimator of the loss is $\frac{1}{n}\sum_1^n (t_i - \ell_i)$ where, by construction, each term $(t_i - \ell_i)$ has zero expectation. We now argue that, under quite general conditions, the error $\frac{1}{n}\sum_1^n (t_i - \ell_i)$ is expected to converge to zero.

One condition for $\frac{1}{n}\sum_1^n (t_i - \ell_i)$ to converge to zero can be stated loosely as follows: The value of $\frac{1}{n}\sum_1^n (t_i - \ell_i)$ should not depend much on $(y_1, \ldots, y_m)$ if $n \gg m$. Basically, we want events concerning the limiting behavior of $\frac{1}{n}\sum_1^n (t_i - \ell_i)$ to belong to the tail σ-field generated by $y_1, y_2, \ldots$. If this is true, then the zero-one law implies that $\frac{1}{n}\sum_1^n (t_i - \ell_i)$ converges to a constant which must necessarily be zero. Unfortunately, it is not easy to formulate the technical conditions on $\hat{\theta}_i(\cdot,\cdot)$ and $\ell_i(\cdot,\cdot)$ to ensure measurability of $\frac{1}{n}\sum\limits_1^n(t_i-\ell_i)$ with respect to the tail σ-field. Furthermore, the zero-one law does not give any indication on the speed of the convergence. For these reasons we will instead investigate the convergence of the loss estimate by direct variance calculations. We will show that for a very large class of estimators $\hat{\theta}(\boldsymbol{y})$, the variance of $\frac{1}{n}\sum\limits_1^n(t_i-\ell_i)$ is of order n^{-1}. For simplicity, we will only consider the case of mean square error loss.

Recall that we are estimating $\ell_i = g_i(y_i)^2 - 2\theta_i g_i(y_i) + \theta_i^2$ by $t_i = g_i(y_i)^2 - 2e_i(y_i) + f_i(y_i)$ where $e_i(y_i)$ and $f_i(y_i)$ are constructed to be unbiased estimators of $\theta_i g_i(y_i)$ and θ_i^2 respectively. Hence

$$\frac{1}{n}\sum_1^n(t_i-\ell_i) = -\frac{2}{n}\sum_1^n\big[e_i(y_i)-\theta_i g_i(y_i)\big] + \frac{1}{n}\sum_1^n\big[f_i(y_i)-\theta_i^2\big].$$

Since the functions $f_i(\cdot)$ are (non-random) functions of y_i alone, we have

$$\mathrm{Var}\bigg(\frac{1}{n}\sum_1^n(f_i-\theta_i^2)\bigg) \le \frac{c}{n}$$

provided each $\mathrm{Var}(f_i) = \int p_{\theta_i}(y)\big(f_i(y)-\theta_i^2\big)^2 dy \le c$. The analysis of the variance of $\frac{1}{n}\sum(e_i-\theta_i g_i)$ is considerably more complicated. The reason is that both $e_i(\cdot)$ and $g_i(\cdot)$ are random functions, i.e. the values of $g_i(y_i)$ and $e_i(y_i)$ depends not only on y_i but also on $\boldsymbol{y}_{(-i)}$. As a result, all terms in the average are generally dependent on each other. To proceed further, let T_i be the operator which maps the function $g_i(\cdot)$ to the function $e_i(\cdot)$, i.e. T_i is constructed so that for any (non-random) function $g(\cdot)$ of y_i, $(T_i g)(y_i)$ is unbiased for $\theta_i g(y_i)$. T_i is assumed to have the following properties:

- a) (Unbiasedness) For all $g \in \mathbb{G}_i$ where $\mathbb{G}_i$ is a large linear space of (non-random) functions of y_i (which is defined separately in each application) we have

$$E\big[(T_i g)(y_i) - \theta_i g(y_i)\big] = 0 \tag{6.1}$$

- b) (Linearity) For any f, $g \in \mathbb{G}_i$, $a, b \in \mathbb{R}$,
$$T_i(af + bg) = aT_if + bT_ig \tag{6.2}$$

- c) If $g(y) \equiv c$ where c is a constant, then
$$(T_ig)(y) = cu_i(y). \tag{6.3}$$
where $u_i(y)$ is an unbiased estimate for θ.

The forms of T_i in several examples have been obtained in the preceeding sections:

$$(T_ig)(y_i) = y_ig(y_i - 1) \qquad \text{(Poisson example)}$$
$$(T_ig)(y_i) = y_ig(y_i) - g'(y_i) \qquad \text{(Normal)}$$
$$(T_ig)(y_i) = \frac{1}{y_i^{k-1}} \int_0^{y_i} z^{k-1}g(z)dz \qquad \text{(Gamma)}$$
$$(T_ig)(y_i) = y_ig(y_i) + 2\alpha c(t * g')(y_i) \qquad \text{(Stable laws)}$$

T_i can also be applied to a multivariate function $h(y_1, \ldots, y_n)$, in which case $(T_ih)(y_1, \ldots, y_n)$ is obtained by regarding $h(y_1, \ldots, y_n)$ as a univariate function of y_i, with $\boldsymbol{y}_{(-i)}$ fixed, and then applying T_i to this univariate function. For example, it follows from (6.3) that; if h is a function of $\boldsymbol{y}_{(-i)}$, then
$$(T_ih)(y_1, \ldots, y_n) = u_i(y_i)h(\boldsymbol{y}_{(-i)}). \tag{6.4}$$

A function $h(y_1, \ldots, y_n)$ is said to belong to the domain of T_i $\big(h \in \mathcal{D}(T_i)\big)$ if $E(h^2) < \infty$ and
$$E\big[(T_i - \theta_i)h\big]^2 < c_iE(h^2) \tag{6.5}$$
where c_i is a constant which is typically equal to the squared norm of T_i as an operator on the class $\mathbb{G}_i$ of univariate functions of y_i. Also, let R_j be the operator representing expectation over y_j conditional on $\boldsymbol{y}_{(-j)}$, i.e.
$$\begin{aligned}(R_jh)(y_1, \ldots, y_n) &= E(h \mid \boldsymbol{y}_{(-j)}) \\ &= \int h(y_1, \ldots, y_n)p_{\theta_j}(y_i)dy_j.\end{aligned}$$

Suppose that $\hat{\theta}_i(y_1, \ldots, y_n)$ has an ANOVA decomposition

$$\hat{\theta}_i(\boldsymbol{y}) = g_i(y_i) = \mu_i + \sum_{j=1}^{n} \alpha_j^i H_j^i(y_j) + \sum_{\{j_1,j_2\}} \beta^i_{\{j_1,j_2\}} H^i_{\{j_1,j_2\}}(y_{j_1}, y_{j_2}) + \cdots \tag{6.6}$$

where H_j^i, $H^i_{j_1,j_2}$ etc. are orthogonal random variables satisfying the conditions

$$R_j H^i_{\{j_1,\ldots,j_k\}} = 0 \qquad \text{if} \qquad j \in \{j_1, \ldots, j_k\}. \tag{6.7}$$

For the construction of such decompositions, see Efron and Stein (1976). We assume that each $\hat{\theta}_i$ has an expansion (6.6) up to m terms where m is independent of n, and that H_j^i, $H^i_{\{j_1,j_2\}}$ etc. are in $\mathcal{D}(T_i)$ and all of them have variance $\leq M$. Finally, the coefficients α_j^i, $\beta^i_{\{j_1,j_2\}}$ etc. are non-negative constants such that

$$\sum_j \alpha_j^i = \sum_{\{j_1,j_2\}} \beta^i_{\{j_1,j_2\}} = \cdots = 1.$$

Theorem. *Suppose T_i, $i = 1, \ldots, n$ satisfy (6.1)–(6.5) with $c_i \leq c_0 < \infty$, and T_i commutes with R_j whenever $j \neq i$. Suppose $\hat{\theta}_i$ $i = 1, \ldots, n$ have ANOVA decompositions satisfying the assumptions of the preceeding paragraph, then there exists a constant $c > 0$ such that*

$$\operatorname{Var}\Big[n^{-1} \sum_1^n \big(e_i(y_i) - \theta_i g_i(y_i)\big)\Big] \leq \frac{c}{n}$$

Proof. ∎

$$\begin{aligned}\sum_1^n (e_i - \theta_i g_i) &= \sum_{i=1}^n (T_i - \theta_i)\hat{\theta}_i \\ &= \Big[\sum_{i=1}^n (T_i - \theta_i)\mu_i\Big] + \Big[\sum_{i=1}^n \sum_j \alpha_j^i (T_i - \theta_i) H_j^i\Big] \\ &\quad + \Big[\sum_{i=1}^n \sum_{\{j_1,j_2\}} \beta^i_{\{j_1,j_2\}} (T_i - \theta_i) H^i_{\{j_1,j_2\}}\Big] + \cdots,\end{aligned}$$

where there are m terms in this expansion. We will only demonstrate the bound for the variance of, say, the third order interaction term. By (6.4), this third order term can be written as $A + B$ where

$$A = \sum_{i=1}^{n} \sum_{i \notin \{j_1,j_2,j_3\}} \gamma^i_{\{j_1,j_2,j_3\}} \, (u_i(y_i) - \theta_i) H^i_{\{j_1,j_2,j_3\}},$$

$$B = \sum_{i=1}^{n} \sum_{\{j_2,j_3\}} \gamma^i_{\{i,j_2,j_3\}} \, (T_i - \theta_i) H^i_{\{i,j_2,j_3\}}.$$

To bound the variance of A, consider

$$E\big[(u_i - \theta_i) H^i_{\{j_1,j_2,j_3\}} (u_k - \theta_k) H^k_{\{\ell_1,\ell_2,\ell_3\}}\big]. \tag{6.8}$$

If $k \notin \{i, j_1, j_2, j_3\}$ we can replace $(u_k - \theta_k)$ by $R_k(u_k - \theta_k) = 0$ in (6.8). Similarly, if $\ell_2 \notin \{i, j_1, j_2, j_3\}$, we can replace $H^k_{\{\ell_1,\ell_2,\ell_3\}}$ by $R_{\ell_2}(H^k_{\{\ell_1,\ell_2,\ell_3\}}) = 0$. Hence (6.8) is zero unless $\{k, \ell_1, \ell_2, \ell_3\} = \{i, j_1, j_2, j_3\}$. Also, by our assumptions on the norm of T_i and the variance of $H^i_{\{j_1,j_2,j_3\}}$, the absolute value of (6.8) is bounded by $c_0 M$. It follows that

$$\begin{aligned}\mathrm{Var}(A) &\le c_0 M \sum_{i=1}^{n} \sum_{i \notin \{j_1,j_2,j_3\}} \gamma^i_{\{j_1,i,j_3\}} \, [\gamma^i_{\{j_1,j_2,j_3\}} + \gamma^{j_1}_{\{i,j_2,j_3\}} + \gamma^{j_2}_{\{j_1,i,j_3\}} + \gamma^{j_3}_{\{j_1,j_2,i\}}] \\ &\le 4 c_0 M \sum_{i=1}^{n} \Big(\sum_{\{j_1,j_2,j_3\}} \gamma^i_{\{j_1,j_2,j_3\}} \Big) \\ &\le 4 c_0 M n.\end{aligned}$$

To bound the variance of B, notice that by applying (6.1) with $g(y_i) = H^i_{\{i,j_2,j_3\}}(y_i, y_{j_2}, y_{j_3})$ where y_{j_2} and y_{j_3} are fixed, we have $R_i(T_i - \theta_i) H^i_{\{i,j_2,j_3\}} = 0$. Also, since $j_2 \neq i$, it follows that

$$R_{j_2}(T_i - \theta_i) H^i_{\{i,j_2,j_3\}} = (T_i - \theta_i) R_{j_2} H^i_{\{i,j_2,j_3\}} = 0.$$

Using these equalities and repeating the same type of arguments used to bound $\mathrm{Var}(A)$, we have

$$\mathrm{Var}(B) \le 3 c_0 M n.$$

This completes the derivation of the bound for the third order interaction term. The same argument can be applied to bound the variance of any other term. ∎

References

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9**, 586–596.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135–1151.