ON STATISTICAL INFERENCE UNDER ASYMMETRIC LOSS FUNCTIONS

Michael Baron

Received:

Abstract

We introduce a wide class of asymmetric loss functions and show how to obtain asymmetric-type optimal decision rules from standard and commonly used Bayesian procedures. Important properties of minimum risk estimators are established. In particular, we discuss their sensitivity to the magnitude of asymmetry of the loss function.

1 Introduction

Consider a statistical decision problem, where underestimating the parameter is, say, cheaper or less harmful than overestimating it by the same amount. Problems of this nature appear in actuarial science, marketing, pharmaceutical industry, quality control, change-point analysis, child's IQ estimation, estimation of water levels for dam constructions, and other situations ([1], [2], [3],

AMS 1991 subject classifications: 62A15, 62C10.

Key words and phrases: loss function, prior distribution, minimum risk estimator, fixed point, range of posterior expectation.

section 4.4, [6], section 3.3, [8], [9], [10], [11]). For example, an underestimated premium leads to some financial loss, whereas an overestimated premium may lead to a loss of a contract.

For the sake of simplicity, it is still common in practical applications to use standard Bayesian decision rules like the posterior mean or median, even in asymmetric situations. However, these procedures do not reflect the difference in losses. It is recognized that some type of correction should be introduced to take account of the asymmetry.

In this paper we introduce a wide class of asymmetric loss functions (1). The corresponding asymmetric-type minimum risk decision rules can then be obtained as fixed points of standard estimators (Theorem 1). A simple alternative is to use linear or quadratic approximations (14) and (15) for situations when over- and underestimation costs are different, but compatible. In this case, a standard Bayes estimator should be corrected by a term proportional to the posterior mean absolute deviation.

Suppose that a continuous or discrete parameter $\theta \in \Theta$ is to be estimated. Here Θ is either a connected subset of $\boldsymbol{R}$ or a possibly infinite collection of points $\{\ldots < \theta_0 < \theta_1 < \ldots\}$. Let Π be a probability measure on Θ, which may be realized as the prior distribution, or the posterior, after a sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ has been observed. As an initial setting, we choose an arbitrary measure Π on Θ and a loss function $L(\delta, \theta)$, where δ is a decision rule. The loss L may be one of the standard symmetric loss functions, however, symmetry is not required for the subsequent discussion.

Introduce an asymmetric loss function of the form

$$W(\delta,\theta) = K_1 L(\delta,\theta) I\{\delta \le \theta\} + K_2 L(\delta,\theta) I\{\delta > \theta\} \tag{1}$$

for some positive K_1 and K_2. It generalizes the linear loss proposed in [3], section 4.4.2, for estimating the child's IQ. Another popular asymmetric loss function is the linex loss ([10], [11]). Unlike the linex, (1) defines a wide family of loss functions, due to an arbitrary choice of $L(\cdot,\cdot)$, where costs of over- and underestimation are compatible.

Without loss of generality we can study the case $K_1 \le K_2$ only, because the other situation obtains by reparameterization $\theta \mapsto -\theta$. Also we can assume that $K_1 = 1$. Then the loss function W has the form

$$W(\delta,\theta) = \begin{cases} L(\delta,\theta) & \text{if } \delta \le \theta, \\ (1+\lambda) L(\delta,\theta) & \text{if } \delta > \theta \end{cases} \tag{2}$$

for some $\lambda \ge 0$, which denotes the additional relative cost of overestimation. Let δ^* be the minimum risk estimator of θ,

$$\delta^* = \delta^*(\lambda) = \arg\min_{\delta} \int W(\delta,\theta) d\Pi(\theta).$$

If overestimation is costly, one will tend to underestimate θ. A way to express this is to consider a family of weighted measures $\{\Pi_t\}$ obtained from Π by increasing probabilities of small values of θ,

$$\Pi_t = \begin{cases} (1+\lambda)\Pi / (1 + \lambda \int_{\theta \le t} d\Pi(\theta)) & \text{if } \theta < t, \\ \Pi / (1 + \lambda \int_{\theta \le t} d\Pi(\theta)) & \text{if } \theta \ge t. \end{cases} \tag{3}$$

For every t let

$$\tilde{\delta}(t) = \arg\min_{\delta} \int L(\delta,\theta) d\Pi_t(\theta).$$

In practice, Π is usually a posterior distribution $\pi(\theta|\boldsymbol{x})$ under some prior $\pi(\theta)$. One defines an altered prior $\pi_t(\theta)$ similarly to (3), by assigning higher probabilities for smaller values of θ. Then Π_t in (3) is the corresponding posterior.

In the next section, we show that δ^* is the local and global minimum, and the fixed point of $\tilde{\delta}(t)$. The rest of the paper gives simple methods of obtaining $\delta^*(\lambda)$ in practice and establishes robust properties of δ^* with respect to the choice of λ. Popular examples are discussed.

2 The fixed point of $\tilde{\delta}(t)$

We assume that $L(\delta, \theta)$ is a strictly convex nonnegative function of δ for any θ, with $L(\theta, \theta) = 0$.

Consider the risks $r(\delta)$ and $r_t(\delta)$ associated with (W, Π) and (L, Π_t) respectively. One has $\tilde{\delta}(t) = \arg\min r_t(\delta)$ for any t, and $\delta^* = \arg\min r(\delta)$. Then $\tilde{\delta}(t)$ also minimizes

$$\psi_t(\delta) = C(\lambda, t) r_t(\delta) = \int L(\delta, \theta) d\Pi(\theta) + \lambda \int_{\theta < t} L(\delta, \theta) d\Pi(\theta),$$

and δ^* minimizes $\psi(\delta) = \psi_\delta(\delta)$. Under our assumptions $\psi_t(\delta)$ and $\psi(\delta)$ are strictly convex functions of δ. The following lemma establishes an important property of $\tilde{\delta}(t)$.

Lemma 1 *The decision rule $\tilde{\delta}(t)$ is a non-increasing function on $(-\infty, T)$ and a non-decreasing function on $(T, +\infty)$, where $T = \inf\{s : \tilde{\delta}(s) \leq s\} = \sup\{s : \tilde{\delta}(s) > s\}$.*

Proof: For any $s \le t$ one has

$$\psi_t(\delta) = \psi_s(\delta) + \lambda \int_{s \le \theta < t} L(\delta, \theta) d\Pi(\theta). \tag{4}$$

Hence, by convexity, the minimum of $\psi_t(\delta)$ is attained between the points of minima of $\psi_s(\delta)$ and $\int_{s \le \theta < t} L(\delta, \theta) d\Pi(\theta)$. In other words, since the latter attains its minimum at some point between s and t,

$$\min\{s; \tilde{\delta}(s)\} \le \tilde{\delta}(t) \le \max\{t; \tilde{\delta}(s)\}. \tag{5}$$

If $\tilde{\delta}(s) \le s$, then (5) results in $\tilde{\delta}(s) \le \tilde{\delta}(t) \le t$, from which $\{s : \tilde{\delta}(s) \le s\}$ is a right half-line and $\tilde{\delta}(t)$ is non-decreasing on it. Let $T = \inf\{s : \tilde{\delta}(s) \le s\}$. Then $\tilde{\delta}(t) > t$ for all $t < T$, and for any $s < t$ (5) implies $\tilde{\delta}(t) \le \tilde{\delta}(s)$. Thus $\tilde{\delta}(t)$ is non-increasing on $(-\infty; T)$. □

We now handle discrete and continuous cases separately. First, let Π be a discrete distribution on $\Theta = \{\theta_k\}$, where θ_k are enumerated in increasing order. Then it follows from (4) that $\psi_t(\delta) = \psi_{\theta_k}(\delta)$ for $t \in (\theta_{k-1}; \theta_k]$, and therefore $\tilde{\delta}(t) = \tilde{\delta}(\theta_k)$. In particular, $\psi(\delta) = \psi_\delta(\delta) = \psi_{\theta_k}(\delta)$, when $\theta_{k-1} < \delta \le \theta_k$.

We show that $\psi(\delta)$ attains a local minimum at $\delta = T$. Then, by convexity of ψ, $\delta^* = T$.

Let $\theta_{k-1} < T \le \theta_k$ for some k. Then

$$T = \sup\{s : \tilde{\delta}(s) > s\} = \sup\{s : \tilde{\delta}(\theta_k) > s\} = \tilde{\delta}(\theta_k), \tag{6}$$

and hence $\psi_{\theta_k}(\delta)$ has a local minimum at T. If $T < \theta_k$, then $\psi(\delta)$ also attains a local minimum at $\delta = T$, because both functions coincide in some neighborhood of T. Otherwise, if $T = \theta_k$ for some k, then

$$T = \inf\{s : \tilde{\delta}(s) \le s\} = \inf\{s : \tilde{\delta}(\theta_{k+1}) \le s\} = \tilde{\delta}_{\theta(k+1)}.$$

Thus, from (6), $\tilde{\delta}(\theta_k) = \tilde{\delta}(\theta_{k+1})$, and T is a point of minimum for both functions ψ_{θ_k} and $\psi_{\theta_{k+1}}$. Since $\psi \equiv \psi_{\theta_k}$ to the left of T and $\psi \equiv \psi_{\theta_{k+1}}$ to the right of it, and it is continuous, it follows that T is a point of minimum for $\psi(\delta)$.

Formula (6) implies that δ^* solves the equation

$$\tilde{\delta}(t) = t. \tag{7}$$

Moreover, δ^* is the unique root of (7). Indeed, suppose that $\tilde{\delta}(t) = t$ for some $t \neq T$. By the definition of T, the only possibility is $t > T$. Let $\theta_m < t \leq \theta_{m+1}$, and take any $s \in (\max\{T, \theta_m\}\,;\, t)$. Then $\tilde{\delta}(s) = \tilde{\delta}(t) = t > s$, which contradicts that $s > T$.

The fact that δ^* is the unique fixed point of $\tilde{\delta}(t)$ suggests a computational method of numerical evaluation of δ^*.

We also note that from (5) with $s = T$ one has $T \leq \tilde{\delta}(t)$ for any $t > T$. The same inequality obtains for $t < T$, because $\tilde{\delta}(T - \epsilon) = T$ for some $\epsilon > 0$, and by Lemma 1 $\tilde{\delta}_s \geq \tilde{\delta}(T - \epsilon)$ for any $s < T - \epsilon$. Hence, $\delta^* = \min_t \tilde{\delta}(t)$.

Now turn to the case when Π is a continuous distribution on Θ, having a density $\pi(\theta)$. Then, from (4), one has $\psi_s \to \psi_t$ pointwise as $s \to t$, and the following statement holds.

Lemma 2 *If* Π *is absolutely continuous with respect to Lebesgue measure, then* $\tilde{\delta}(t)$ *is a continuous function of* t.

The proof of Lemma 2 is given in the Appendix. By continuity, it follows again from Lemma 1 that $\tilde{\delta}(T) = T = \min_t \tilde{\delta}(t)$. It remains to show that $\delta^* = T$.

Since $\psi_t(\delta)$ increases in t pointwise, for any $t > T$

$$\psi(t) = \psi_t(t) > \psi_t(\tilde{\delta}(t)) \geq \psi_T(\tilde{\delta}(t)) \geq \psi_T(T) = \psi(T).$$

Hence, $\delta^* > T$ is impossible. Suppose that $\delta^* < T$. Choose $\delta^* < t < T$ and a sequence $t_j \downarrow t$, for which $\pi(t_j)$ is bounded by some M. Then, since $\delta^* = \arg\min \psi$, and ψ is convex, one has $\psi(t_j) > \psi(t)$, and

$$\psi(t_j) - \psi_t(t_j) > \psi_t(t) - \psi_t(t_j). \tag{8}$$

Consider both sides of (8). From (4),

$$\begin{aligned} \psi(t_j) - \psi_t(t_j) &= \lambda \int_t^{t_j} L(\nu, t_j)\pi(\nu)d\nu \\ &\leq \lambda M(\boldsymbol{x}) L(t, t_j)(t_j - t) = o(t_j - t), \end{aligned}$$

as $j \to \infty$, because $L(t, t_j) \to 0$. However, by convexity of ψ_t,

$$\psi_t(t) - \psi_t(t_j) \geq m(t_j - t), \text{ where } m = \frac{\psi_t(t) - \psi_t(\tilde{\delta}(t))}{\tilde{\delta}(t) - t} > 0.$$

Now we have a contradiction to (8). Thus we have proved the following theorem.

Theorem 1 *The minimum risk estimator of θ under the loss function W and the prior π is the unique root of the equation $\tilde{\delta}(t) = t$. Also, it equals $\min_t \tilde{\delta}(t)$.*

As an illustration, we consider the problem of estimating the binomial parameter θ, when one variable X from the binomial $(5, \theta)$ distribution is observed, and θ has a uniform prior distribution. For $\lambda \in \{1, 6\}$ and all possible values of X, the graphs of $\tilde{\delta}(t)$ are depicted on Figure 1. Their shapes are in accordance with Lemma 1. The dotted line represents the graph $\tilde{\delta}(t) = t$. By Theorem 1, $\delta^*(\lambda)$ coincides with the intersection points for different values of X. Obviously, higher values of λ imply a higher penalty for overestimation, which leads to lower values of $\tilde{\delta}(t)$ and δ^*.

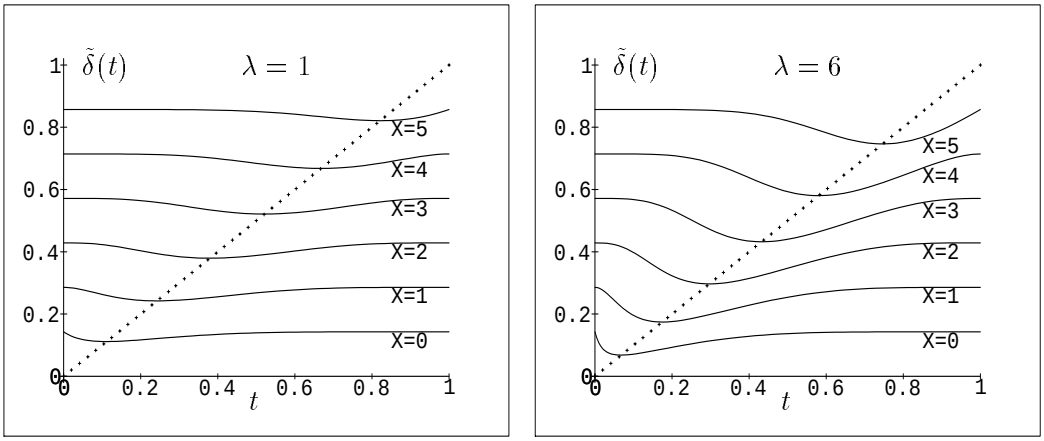

Figure 1. *Behavior of* $\tilde{\delta}(t)$ *and* δ^* *in the binomial case.*

2.1 The range of decision rules

Methods of Bayesian global posterior robustness suggest giving an interval of decision rules corresponding to a family of priors Γ, rather than specifying one optimal procedure for one given prior ([4], [5]). One considers an interval

$$\mathcal{D} = \left[\inf_{\Pi \in \Gamma} \delta^{\pi}; \sup_{\Pi \in \Gamma} \delta^{\pi} \right]$$

of the minimum risk decision rules corresponding to all priors of Γ. In asymmetric problems with a fixed λ, one considers the family of prior distributions $\Gamma_\lambda = \{\Pi_t\}$ for all values of t, as defined in (3). A direct application of Lemma 1 and Theorem 1 gives the form of $\mathcal{D}$,

$$\mathcal{D} = [\delta^*(\lambda), \delta^*(0)]\,, \tag{9}$$

whereas Lemma 4 below bounds the length of $\mathcal{D}$ for the case of the squared-error loss.

Further, one can consider an extended family of priors $\Gamma = \{\Pi_{t,\lambda}\}$ for all $t \in \bar{\Theta}$ and $\lambda \in [\lambda_1, \lambda_2]$, $0 < \lambda_1 < \lambda_2$. Then the range of the posterior expectation is

$$\mathcal{D} = [\delta^*(\lambda_2), \delta^*(0)]\,, \tag{10}$$

independent of λ_1.

If a sample from $f(\boldsymbol{x}|\theta)$ is available, and δ^Π are Bayes rules under the prior distributions $\Pi \in \Gamma$, their ranges are still given by (9) and (10).

3 Absolute error loss

In order to have a unique minimizer, we required that $L(\delta, \theta)$ be a strictly convex function of δ. However, if $L(\delta, \theta) = |\delta - \theta|$, all the minimum risk estimators with respect to the loss function W are still the solutions of (7). Indeed, if L is the absolute error loss, then $\tilde{\delta}(t)$ is any median of Π_t, and (7) is equivalent to a system of two inequalities,

$$\int_{-\infty}^{t} d\Pi_t(\theta) = \int_{-\infty}^{t} \frac{(1+\lambda)d\Pi(\theta)}{(1+\lambda)\int_{-\infty}^{t} d\Pi(\theta) + \int_{t}^{\infty} d\Pi(\theta)} \geq \frac{1}{2}$$

and a similar inequality for $\int_t^\infty d\Pi_t(\theta)$. Any $1/(\lambda+2)$-quantile of Π solves this system. According to [3], section 4.4, the Bayes rule δ^* has the same form.

4 Squared error loss and weighted losses

When $L(\delta, \theta) = (\delta - \theta)^2$, it is easy to check that

$$\tilde{\delta}(t) = \boldsymbol{E}^{\Pi_t}(\theta) = \frac{\mu + \lambda \boldsymbol{E}^\Pi(\theta I_{\theta<t})}{1 + \lambda \boldsymbol{P}^\Pi\{\theta < t\}}. \tag{11}$$

Here and later $I_A = 1$ if A holds; $= 0$ otherwise, and $\mu = \boldsymbol{E}^{\Pi}(\theta)$. Note that $\boldsymbol{E}^{\Pi}(\theta I_{\theta<t}) = \boldsymbol{E}^{\Pi}\{\theta \,|\, \theta < t\}\boldsymbol{P}^{\Pi}\{\theta < t\}$. Hence, $\tilde{\delta}(t)$ is a convex linear combination of $\mu = \tilde{\delta}(-\infty)$ and $\boldsymbol{E}^{\Pi}(\theta \,|\, \theta < t)$. Then, according to (11) and Theorem 1, δ^* solves the equation

$$\boldsymbol{E}^{\Pi}\{\theta - t) + \lambda \boldsymbol{E}^{\Pi}(\theta - t) I_{\theta<t} = 0, \tag{12}$$

which is equivalent to

$$\boldsymbol{E}^{\Pi}(\theta - t)^{+} = (1 + \lambda)\boldsymbol{E}^{\Pi}(\theta - t)^{-},$$

where $a^{+} = \max\{a; 0\}$, and $a^{-} = \max\{-a; 0\}$.

In general, for any loss function of the form

$$L(\delta, \theta) = |\delta - \theta|^{\gamma}, \ \gamma > 1,$$

δ^* solves the equation

$$\boldsymbol{E}^{\Pi}[(\theta - t)^{+}]^{\gamma-1} = (1 + \lambda)\boldsymbol{E}^{\Pi}[(\theta - t)^{-}]^{\gamma-1}.$$

The case $\gamma = 1$ yields the $1/(\lambda+2)$-quantile mentioned in the previous section.

Generalization to the case of weighted squared error loss functions $L(\delta, \theta) = w(\theta)(\delta - \theta)^2$ is straightforward. The weights $w(\theta)$ usually allow larger errors for larger true values of the parameter. It is equivalent to the decision problem with the unweighted loss, if the distribution $\Pi_t(d\theta)$ is replaced by

$$\Pi'_t(d\theta) = \begin{cases} \dfrac{(1 + \lambda)\Pi(\theta)w(\theta)}{\boldsymbol{E}^{\Pi}w(\theta) + \lambda \int_{\theta \le t} w(\theta) d\Pi(\theta)} & \text{if } \theta < t, \\ \\ \dfrac{\Pi(\theta)w(\theta)}{\boldsymbol{E}^{\Pi}w(\theta) + \lambda \int_{\theta \le t} w(\theta) d\Pi(\theta)} & \text{if } \theta \ge t. \end{cases}.$$

Any weighted loss function can be considered similarly, as long as the weight function is Π-integrable.

4.1 Sensitivity with respect to λ, and correction of standard decision rules

When the choice of λ, the relative additional cost of overestimation, is not obvious, it is natural to investigate the robustness of δ^* with respect to different λ. The following results concern the rate of change of $\delta^* = \delta^*(\lambda)$ for $\lambda \to 0$ and $\lambda \to +\infty$.

Lemma 3 *Let $L(\delta, \theta)$ be the squared error loss. Then*

$$\left.\frac{d\delta^*(\lambda)}{d\lambda}\right|_{\lambda=0} = Cov(\theta\,,\, I_{\theta<\mu}) = -\frac{1}{2}\boldsymbol{E}^{\Pi}|\theta - \mu|. \tag{13}$$

Lemma 3 shows that $|\delta^*(\lambda) - \delta^*(0)|$ has the linear order in λ as $\lambda \to 0$. As $\lambda \downarrow 0$, one has

$$\delta^*(\lambda) = \mu - \frac{\lambda}{2}\boldsymbol{E}^{\Pi}|\theta - \mu| + O(\lambda^2). \tag{14}$$

In applications, this determines a simple strategy for a decision maker. Every time when overestimation is slightly more costly than underestimation, *the standard Bayes estimator $\delta^*(0) = \mu$ is to be corrected by a term proportional to the mean absolute deviation of θ*. Suppose, for example, that the loss caused by overestimating the parameter is 10% higher than the loss caused by underestimating it by the same amount. Then the corrected decision rule

$$\delta^*(0.1) = \mu - (0.05)\boldsymbol{E}^{\Pi}|\theta - \mu|$$

should be used instead of the standard posterior mean $\mu = \boldsymbol{E}^{\Pi}(\theta)$.

Differentiating once more, one obtains,

$$\left.\frac{d^2\delta^*(\lambda)}{d\lambda^2}\right|_{\lambda=0} = \boldsymbol{E}^{\Pi}|\theta - \mu|\boldsymbol{P}^{\Pi}\left\{\theta < \mu\right\}.$$

This provides a more accurate correction of the standard decision rule,

$$\delta^*(\lambda) = \mu - \boldsymbol{E}^{\Pi}|\theta - \mu| \left(\frac{\lambda}{2} - \lambda^2 \boldsymbol{P}\{\theta < \mu\} + O(\lambda^3) \right), \text{ as } \lambda \to 0. \qquad (15)$$

The next result bounds all possible deviations of δ^*.

Lemma 4 *One has*

$$\frac{|\delta^*(\lambda) - \delta^*(0)|}{\sigma^{\Pi}(\theta)} \le \frac{\lambda}{2\sqrt{1+\lambda}}, \qquad (16)$$

where $\sigma^{\Pi}(\theta)$ *denotes the standard deviation of* θ *under the distribution* Π.

Simple calculation shows that if Θ consists of only two points, then $\rho(t) = -1$ everywhere between them. In this case, equality holds in (16).

According to Lemma 4, when Θ is unbounded, $\delta^*(\lambda)$ deviates from $\delta^*(0)$ with the rate $O(\sqrt{\lambda})$, as $\lambda \to +\infty$.

4.2 Examples

Bayesian estimation of the normal mean

We consider Bayesian estimation of a normal mean. Let $X_1, \dots, X_n$ be a sample from the normal distribution with $\boldsymbol{E}\, X = \theta$, $Var(X) = \sigma^2$. Suppose that σ^2 is known, and θ, the parameter of interest, has a prior distribution $\pi(\theta)$, which is normal(μ, τ^2). The corresponding posterior $\Pi = \pi(\theta|\boldsymbol{x})$ is normal(μ_x, τ_x^2), where

$$\mu_x = \frac{n\bar{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} \text{ and } \tau_x^2 = \frac{1}{n/\sigma^2 + 1/\tau^2}$$

(eg. [3], [7]). Then $z = (\theta - \mu_x)/\tau_x$ follows the standard normal distribution under $\pi(\theta|\boldsymbol{x})$. Let $\phi(u)$ and $\Phi(u)$ denote standard normal pdf and cdf

respectively. Note that $\boldsymbol{E}\, z I_{z<u} = -\phi(u)$. From (11),

$$\tilde{\delta}(t) \;=\; \frac{\mu_x + \lambda \boldsymbol{E}(\mu_x + \tau_x z) I_{\mu_x + \tau_x z < t}}{1 + \lambda \boldsymbol{P}\left\{\mu_x + \tau_x z < t\right\}} = \mu_x - \frac{\lambda \tau_x \phi(u)}{1 + \lambda \Phi(u)},$$

where $u = (t - \mu_x)/\tau_x$. Then the fixed-point equation (7) is equivalent to

$$\phi(u) + u\Phi(u) + \frac{u}{\lambda} = 0. \tag{17}$$

Hence, the minimum asymmetric risk decision rule is given by

$$\delta^*(\lambda) = \mu_x + \tau_x u(\lambda), \tag{18}$$

where $u(\lambda)$ is a solution of (17).

According to (18), $\frac{d\delta^*}{d\lambda} = \tau_x u'(\lambda)$, which is independent of $X_1, \ldots, X_n$. In particular,

$$\left.\frac{d\delta^*}{d\lambda}\right|_{\lambda=0} \;=\; -\tau_x \phi(0) \;=\; -\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}},$$

and the same formula can be obtained by Lemma 3.

By (15), $\delta^*(\lambda) \approx \mu_x - \tau_x \lambda(1-\lambda)/\sqrt{2\pi}$ for small λ, and the correction term is the same regardless of the observed sample.

Asymmetric least squares

In the no-intercept linear regression model

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

let $\hat{\beta}$ be selected to minimize

$$\begin{aligned} \psi(b) \;&=\; \sum (y_i - x_i b)^2 + \lambda \sum (y_i - x_i b)^2 I_{y_i < x_i b} \\ &=\; \sum x_i^2 (r_i - b)^2 + \lambda \sum x_i^2 (r_i - b)^2 I_{r_i < b}, \end{aligned} \tag{19}$$

with $r_i = y_i/x_i$. Additional weights assigned to squared residuals reflect the higher cost of overestimating β and overpredicting y.

Suppose that ratios r_i are enumerated in their increasing order, $r_1 \le \ldots \le r_n$. For every $k = 1, \ldots, n$, consider

$$\psi_k(b) = \sum_{i=1}^{n} x_i^2 (r_i - b)^2 + \lambda \sum_{i=1}^{k} x_i^2 (r_i - b)^2,$$

This function is minimized at

$$\hat{\beta}(k) = \frac{\sum_1^n x_i^2 r_i + \lambda \sum_1^k x_i^2 r_i}{\sum_1^n x_i^2 + \lambda \sum_1^k x_i^2} = \frac{\int r d\Pi_\lambda(r)}{\int d\Pi_\lambda(r)} = \boldsymbol{E}^{\Pi_\lambda}(r),$$

where the distribution Π_λ puts masses

$$\begin{cases} Cx_i^2 & \text{at } r_i, i > k, \\ C(1+\lambda)x_i^2 & \text{at } r_i, i \le k. \end{cases}$$

Thus $\hat{\beta}(k)$ has the Bayes-type form of $\tilde{\delta}(k)$, and according to Theorem 1, the asymmetric least squares solution $\hat{\beta} = \delta^*$ can be found as the fixed point of $\hat{\beta}(k)$. Clearly, the case $\lambda = 0$ gives the OLS solution.

Change-point estimation

In a classical change-point problem, a sample of independent variables $\boldsymbol{x} = (X_1, \ldots, X_n)$ is such that the distribution of X_j depends on whether $j < \theta$ or $j \ge \theta$. One needs to estimate the change-point parameter θ. Since a change-point has to be detected "as soon as possible", suppose that overestimation leads to a higher loss than underestimation.

Further, assume a (continuous) uniform prior distribution of θ over the interval $\Theta = [0; n]$. If the pre-change and the post-change densities or probability mass functions f and g are mutually absolutely continuous, the standard

Bayes rule under the squared-error loss has the form

$$\mu = \delta^*(0) = \boldsymbol{E}(\theta|\boldsymbol{x}) = \frac{\sum_{k=1}^n k\zeta(k)}{\sum_{k=1}^n \zeta(k)}, \tag{20}$$

where $\zeta(k) = f(X_1) \cdot \ldots \cdot f(X_k)/g(X_1)/\ldots/g(X_k)$ ([8]). Correction for an asymmetric loss is straightforward. According to (15), one obtains the optimal decision rule under the loss function (2),

$$\delta^*(\lambda) = \mu - \left(\frac{\sum_1^n \zeta(k)|k-\mu| + \sum_1^{[\mu]+1} \zeta(k) + (\{\mu\}^2 - 1)\,\zeta([\mu]+1)}{\sum_1^n \zeta(k)} - \frac{1}{2} \right) \times \left(\frac{\lambda}{2} - \lambda^2 \frac{\sum_1^{[\mu]} \zeta(k) + \{\mu\}\zeta([\mu]+1)}{\sum_1^n \zeta(k)} \right) + O(\lambda^3),$$

as $\lambda \to 0$, where μ is given by (20), $[\mu]$ and $\{\mu\}$ denote its integer and fractional parts, respectively (derivation is omitted).

5 Appendix

Proof of Lemma 2. Suppose that there exist such $\Delta > 0$ and a monotone sequence $t_j \to \tau$ that $|\tilde{\delta}(t_j) - \tilde{\delta}(\tau)| > \Delta$ for all j. Let

$$\epsilon = \min\{\psi_\tau(\tilde{\delta}(\tau) - \Delta), \psi_\tau(\tilde{\delta}(\tau) + \Delta)\} - \psi_\tau(\tilde{\delta}(\tau)) > 0.$$

If $t_j \uparrow \tau$, then $\psi_{t_j}(\cdot) \uparrow \psi_\tau(\cdot)$, and $\psi_{t_j}(\tilde{\delta}(\tau) \pm \Delta) > \psi_\tau(\tilde{\delta}(\tau) \pm \Delta) - \epsilon/2$ for sufficiently large j. Then, since $\psi_{t_j}(\tilde{\delta}_\tau) \le \psi_\tau(\tilde{\delta}_\tau) < \psi_\tau(\tilde{\delta}(\tau) \pm \Delta) - \epsilon/2$, it follows from convexity of ψ_{t_j} that it has a minimum on $(\tilde{\delta}(\tau) - \Delta, \tilde{\delta}(\tau) + \Delta)$, which contradicts to $|\tilde{\delta}(t_j) - \tilde{\delta}(\tau)| > \Delta$.

If $t_j \downarrow \tau$, then $\psi_{t_j}(\tilde{\delta}(\tau)) \ge \psi_{t_j}(\tilde{\delta}(t_j)) \ge \psi_\tau(\tilde{\delta}(t_j)) > \psi_\tau(\tilde{\delta}(\tau)) + \epsilon$, and we have a contradiction: $\psi_{t_j}(\tilde{\delta}(\tau)) \not\to \psi_\tau(\tilde{\delta}(\tau))$.

Hence, $\tilde{\delta}(t)$ is continuous.

Proof of Lemma 3. According to (12),

$$\int(\theta-\delta^*)d\Pi(\theta)+\lambda\int_{\theta<\delta^*}(\theta-\delta^*)d\Pi(\theta)=0.$$

Differentiating implicitly, one obtains

$$\frac{d\delta^*(\lambda)}{d\lambda}=\frac{-\boldsymbol{E}^{\Pi}(\theta-\delta^*)^-}{1+\lambda\boldsymbol{P}^{\Pi}\{\theta<\delta^*\}}.$$

Notice that $\lambda=0$ corresponds to the symmetric square loss $W\equiv L$, hence $\delta^*(0)=\mu$. Also,

$$-\boldsymbol{E}^{\Pi}(\theta-\mu)^-=\boldsymbol{E}^{\Pi}\left(\theta I_{\theta<\mu}\right)-\mu\,\boldsymbol{E}^{\Pi}I_{\theta<\mu}=Cov(\theta\,,\,I_{\theta<\mu}),\qquad(21)$$

and the first equality in (13) follows.

Next, observe that

$$\boldsymbol{E}(\theta-\mu)I_{\theta>\mu}+\boldsymbol{E}(\theta-\mu)I_{\theta<\mu}=\boldsymbol{E}(\theta-\mu)=0,\qquad(22)$$

and

$$\boldsymbol{E}(\theta-\mu)I_{\theta>\mu}-\boldsymbol{E}(\theta-\mu)I_{\theta<\mu}=\boldsymbol{E}|\theta-\mu|.\qquad(23)$$

Subtracting (23) from (22), one completes the proof.

Proof of Lemma 4. For any t let $\rho(t)$ denote a correlation coefficient between θ and $I_{\theta<t}$ under Π. Then, similarly to (21), one obtains from (11)

$$\begin{aligned}\tilde{\delta}(t)-\delta^*(0) &= \frac{\lambda\,Cov(\theta,\,I_{\theta<t})}{1+\lambda\boldsymbol{P}^{\Pi}\{\theta<t\}}\\ &= \frac{\lambda\rho(t)\sigma^{\Pi}(\theta)\sqrt{\boldsymbol{P}^{\Pi}\{\theta<t\}\,(1-\boldsymbol{P}^{\Pi}\{\theta<t\})}}{1+\lambda\boldsymbol{P}^{\Pi}\{\theta<t\}}.\end{aligned}$$

Clearly, $\tilde{\delta}(t) \le \delta^*(0)$. Then, by Theorem 1, $\delta^*(\lambda)$ maximizes $|\tilde{\delta}(t) - \delta^*(0)|$ over all t, and

$$|\delta^*(\lambda) - \delta^*(0)| \le \lambda\sigma^{\Pi}(\theta) \max_{0\le p\le 1} \frac{\sqrt{p(1-p)}}{1+\lambda p} = \frac{\lambda\sigma^{\Pi}(\theta)}{2\sqrt{1+\lambda}}.$$

Acknowledgment. The author is grateful to Professors A. L. Rukhin, R. J. Serfling, and K. Ostaszewski for their valuable comments to earlier versions of this paper.

References

[1] J. Aitchison and I. R. Dunsmore. *Statistical Prediction Analysis.* Cambridge University Press, London, 1975.

[2] M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application.* PTR Prentice-Hall, Inc., 1993.

[3] J. O. Berger. *Statistical Decision Theory.* Springer–Verlag, New York, 1985.

[4] J. O. Berger. Robust Bayesian analysis: sensitivity to the prior. *J. Stat. Plann. Inf.*, 25:303–328, 1990.

[5] L. DeRobertis and J. A. Hartigan. Bayesian inference using intervals of measures. *Ann. Statist.*, 9:235–244, 1981.

[6] R. V. Hogg and S. A. Klugman. *Loss distributions.* Wiley, New York, 1984.

[7] E. L. Lehmann. *Theory of Point Estimation.* Wiley, New York, 1983.

[8] A. L. Rukhin. Change-point estimation under asymmetric loss. *Statistics&Decisions*, 15:141–163, 1995.

[9] J. Shao and S.-C. Chow. Constructing release targets for drug products: a bayesian decision theory approach. *Appl. Statist.*, 40:381–390, 1991.

[10] H. R. Varian. A Bayesian approach to real estate assessment. In *S. E. Fienberg, A. Zellner, eds., Studies in Bayesian Econometrics and Statistics*, pages 195–208, Amsterdam: North-Holland, 1975.

[11] A. Zellner. Bayesian estimation and prediction using asymmetric loss functions. *J. Amer. Stat. Assoc.*, 81:446–451, 1986.

Michael Baron
Programs in Mathematical Sciences
University of Texas at Dallas
Richardson, TX 75083-0688