5
Prediction

This chapter is about prediction. Prediction is the task most often required of the statistician by the client. In order to understand how there can be good and bad predictions, we start with a brief outline of the more general subject of Decision Theory (Sec. 5.1), and the very powerful Bayes Rule theorem (Sec. 5.2). Sec. 5.3 outlines the special case of prediction, and the reasoning that leads, in mamy cases, to the optimal prediction rule being the hypothetical expectation. There are some tricky implications for Frequentist inference, notably the use of the Maximum Likelihood estimator as a plug-in for the parameter (Sec. 5.4). Finally, Sec. 5.5 discusses a vrey modern prediction problem, in order to explore when it is that computer scientists will make better predictions than statisticians.

5.1 A little Decision Theory

In a decision problem, the client would like to choose an action from among a range of options, on the basis of observations she already has, or expects to acquire.[1] Since this choice ought to be optimal according to some criterion, we formalise this problem as follows.

[1] Or maybe there are no observations: this is an easy special case.

1. A set of random quantities, $\boldsymbol{X} \in \mathcal{X}$, and a set of observables, $\boldsymbol{Y} \in \mathcal{Y}$.
2. A set of possible *actions*, $\mathcal{A}$.
3. A *decision rule* $\delta : \mathcal{Y} \to \mathcal{A}$. The intepretation of '$\delta(\boldsymbol{y})$' is 'the action chosen by decision rule δ based on the outcome $\boldsymbol{Y} \rightsquigarrow \boldsymbol{y}$'.

In other words, the client gets to observe $\boldsymbol{Y}$ before choosing action a, and this choice is represented in the form of a rule δ.

4. A *loss function* $L : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$. The interpretation of '$L(a, \boldsymbol{x})$' is 'the expected loss incurred by the client who chooses action a when $\boldsymbol{X} \rightsquigarrow \boldsymbol{x}$'.

The loss function can depend on $\boldsymbol{y}$ as well as $\boldsymbol{X}$, but I have left this out, for simplicity.

The loss function exists to define what it means for an action or a decision rule to be optimal. An action or decision rule is optimal

exactly when it minimises the expected loss. I assume that every reasonable definition of 'optimal' can be represented in these terms.[2] But is it a task of mind-blowing complexity to specify such a loss function in practice, because loss is such a complex concept.

For example, a client managing the risk of a natural hazard, like a volcano, must include in this one function the loss of life, grief and trauma for the survivors, the financial costs of destruction and rebuilding, and so on. A client designing an early warning system for a volcano must also include reputational loss of the system (e.g. if it produces lots of false alarms and people start to ignore it).

However, decisions have to be made; and, indeed, decisions of this level of complexity are made all the time, decisions which affect us all. My strong preference, for decisions which affect me, is for a decision which is defensible according to a stated loss function which I can inspect and reflect on. This acknowledges the point made at the start of Chapter 1, that people are extremely poor at reasoning under uncertainty. Frankly, I would have little confidence in an an oracular pronouncement such as "We have thought very deeply about this issue and have decided that the best action is a". I would much rather hear a pronouncement along the lines of "According to the admittedly simplistic loss function L, the optimal choice of action is a, with a' a close second. But, taking account of other hard-to-quantify aspects of the problem, we have decided that the best action is a'."

A standard hard problem in the public sector is to balance loss of life against financial cost; for example when deciding whether to make road improvements at an accident blackspot. The common approach is to value each life in money terms, and there are standard values for this purpose. Any such value will undoubtedly be highly contentious. A telling point which is often overlooked is that the optimal action may well be robust to the particular choice of value. This is something that can be checked and demonstrated in a formal framework, but not in an oracular pronouncement. And if it is not true, then one option is to reopen the discussion about what the financial value of a life should be, in these calculations. Of course another option is to reject utterly that these choices should be made according to a financial value of a life. This does not undermine the existence of a loss function, but it makes it much harder to construct one. Smith (2010, ch. ???) discusses 'multi-attribute' loss functions.

A realistic loss function will often depend on many quantities, many more than are represented in the random quantities $\boldsymbol{X}$. Here $\boldsymbol{X}$ might represent some of the most important and easy-to-quantify determinants of loss. The presence of missing quantities implies that the loss function is itself an expectation. For suppose there is a more complete loss function $L[a, (\boldsymbol{x}, \boldsymbol{z})]$, which includes values for

[2] This is bold but—I believe—justifiable. There is a theory of rational choice which asserts exactly this, under some conditions; see, e.g., DeGroot (1970) or Smith (2010). The conditions are strong, for the individual, but I think reasonable for an agent making choices on behalf of others.

the additional tricky quantities $\boldsymbol{Z}$. Then

$$\begin{aligned}
\mathrm{E}\{L[a,(\boldsymbol{X},\boldsymbol{Z})]\} &= \sum_{\boldsymbol{x}} \mathrm{E}\{L[a,(\boldsymbol{X},\boldsymbol{Z})] \mid \boldsymbol{X} \doteq \boldsymbol{x}\} \cdot \mathrm{p}(\boldsymbol{x}) && \text{by the LIE, Thm 3.5} \\
&= \sum_{\boldsymbol{x}} \mathrm{E}\{L[a,(\boldsymbol{x},\boldsymbol{Z})] \mid \boldsymbol{X} \doteq \boldsymbol{x}\} \cdot \mathrm{p}(\boldsymbol{x}) && \text{by Thm 3.7} \\
&\equiv \sum_{\boldsymbol{x}} L(a,\boldsymbol{x}) \cdot \mathrm{p}(\boldsymbol{x}) \\
&= \mathrm{E}\{L(a,\boldsymbol{X})\} && \text{by the FTP, Thm 1.6}
\end{aligned}$$

provided that $L(a,\boldsymbol{x}) := \mathrm{E}\{L[a,(\boldsymbol{x},\boldsymbol{Z})] \mid \boldsymbol{X} \doteq \boldsymbol{x}\}$. Therefore the loss function L is best treated as the expected loss supposing that $\boldsymbol{X} \doteq \boldsymbol{x}$. This is also very challenging to specify.

Therefore we must acknowledge at the very start that difficulties in specifying the loss function imply that the statistician is not going to produce the client's optimal action or decision rule. Rather, he operates more in the role of 'critical friend', helping the client to rule out bad choices, and to think more clearly about her preferences.

5.2 *The Bayes Rule theorem*

From now on I will treat X and Y as scalars, simply to reduce the amount of ink on the page.

A decision rule δ is like a *playbook*. Before knowing the value of the observables, the client is able to say how she would act for each possible outcome in $\mathcal{Y}$. Thus, a decision rule is about being prepared. If the client is responsible for real-time risk management, then she can respond rapidly to the observations as they come in.

Of course it is rarely that simple in practice, as the actual observations will tend not to be precisely the ones that were anticipated, and not a superset of them either. In this situation the decision rule is more about guidance: the client might find a y in the playbook that is sufficiently like the actual observations that $\delta(y)$ is a reasonable candidate for a good action. And presumably the process of computing the decision rule, involving specifying an action set and a loss function and thinking about uncertainty, also equips the client to make better decisions under pressure.[3]

[3] "Plans are worthless, but planning is everything.", Dwight D. Eisenhower, 1957.

In other situations, where a rapid response is not required, decision rules are important in experimental design, but less so in choosing between actions. If the observations are already known then there is little reason to compute the decision rule for any value of Y other than y^{obs}. In this section I consider decision rules in general, and then the choice of optimal action in situtations where Y is known.

As anticipated in the previous section, the optimal rule is one which minimises expected loss. This rule is termed the *Bayes rule*. But we must be careful about the precise definition, because of the interaction between the choice of action, a, and the client's beliefs about X.

For example, suppose that

$$\mathcal{A} = \{\text{wear sandals}, \text{wear shoes}\} \quad \text{and} \quad \mathcal{X} = \{\text{dry}, \text{rainy}\}.$$

I do not believe that my choice of footwear influences the weather, although it might sometimes seem like that, and hence my beliefs about X are invariant to a. On the other hand, suppose that

$$\mathcal{A} = \{\text{don't cloud seed}, \text{cloud seed}\} \quad \text{and} \quad \mathcal{X} = \{\text{dry}, \text{rainy}\}.$$

In this case, the purpose of the action is to influence the weather, and so there is a *prima facie* case that the client's beliefs about X depend on a.[4]

[4] See Levin *et al.* (2010) for an assessment of the long-running Israeli cloud seeding experiment.

So beliefs about X ought to be able to depend on the choice a. At the same time, though, beliefs about Y cannot depend on a, because Y is observed before action a is chosen. Thus the client's PMF over (X, Y) should depend on a and factorise as

$$\mathrm{p}(x, y; a) = \mathrm{p}(x \mid y; a)\, \mathrm{p}(y). \tag{†}$$

Now consider some decision rule δ. The expected loss from using this rule, prior to knowing X or Y, is

$$\mathrm{E}\{L(\delta, X)\} := \sum_y \sum_x L[\delta(y), x]\, \mathrm{p}\{x \mid y; \delta(y)\}\, \mathrm{p}(y) \tag{‡}$$

by the FTP, and using the factorisation in (†). Now we can define the Bayes Rule.

Definition 5.1 (Bayes Rule)**.**
The rule δ^* is a Bayes Rule exactly when

$$\delta^* := \operatorname*{argmin}_{\delta \in \mathcal{D}} \mathrm{E}\{L(\delta, X)\}$$

where $\mathcal{D}$ is the set of all functions from $\mathcal{Y}$ to $\mathcal{A}$.

Now this looks like a hopelessly intractable problem: maximising a function, possibly very complex, over the space of all possible functions mapping from observations to actions. Which makes the following result almost miraculous.

Theorem 5.1 (Bayes Rule theorem)**.**

$$\delta^*(y) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathrm{E}\{L(a, X) \mid Y \doteq y; a\}. \tag{5.1}$$

The following proof is intriguing. It uses the existence of the PMF $\mathrm{p}(x, y; a)$ to establish an ordering between two expectations. This is the power of the FTP: mere existence of the PMF is enough.

Proof. We take an arbitrary rule δ and show that $\mathrm{E}\{L(\delta, X)\} \geq \mathrm{E}\{L(\delta^*, X)\}$:

$$\begin{aligned}
\mathrm{E}\{L(\delta, X)\} &= \sum_y \sum_x L[\delta(y), x]\, \mathrm{p}\{x \mid y; \delta(y)\}\, \mathrm{p}(y) && \text{from (‡)} \\
&\geq \sum_y \min_{a \in \mathcal{A}} \Big\{ \sum_x L[a, x]\, \mathrm{p}\{x \mid y; a\} \Big\}\, \mathrm{p}(y) && \text{as } \mathrm{p}(y) \geq 0 \\
&= \sum_y \Big\{ \sum_x L[\delta^*(y), x]\, \mathrm{p}\{x \mid y; \delta^*(y)\} \Big\}\, \mathrm{p}(y) && \text{by (5.1)} \\
&= \mathrm{E}\{L(\delta^*, X)\} && \text{(‡) again}
\end{aligned}$$

as required, where the CFTP (Thm 3.2) is used to move between (5.1) and the equivalent representation as a sum over x with respect to $\mathrm{p}(x \mid y; a)$. □

This result is a direct consequence of the definition of hypothetical expectation given in Def. 3.2, and the implication that this has for the factorisation of the PMF of (X, Y). It says, in words: in order to determine your optimal action when $Y \to y$, simply imagine yourself in the hypothetical world in which $Y = y$, and choose your best action in this world. In retrospect such a result is highly reassuring, because any other result would seem, again in retrospect, rather bizarre. It is another confirmatory result for the rightness of Def. 3.2 as the defining property of a hypothetical expectation.

* * *

Now consider the case where the client knows the value of the observable to be y^{obs}, and wants to determine her optimal action. If she has a $\mathrm{p}_{X|Y}(\cdot \mid y^{\mathrm{obs}}; a)$ then she can compute a $\mathrm{E}^*\{L(a, X)\}$ for each a, where, as before, the asterisk indicates '$|Y \doteq y^{\mathrm{obs}}$'. In this way she constructs a complete ordering over the elements of $\mathcal{A}$. So she will be able to see immediately which is the best a, i.e. the Bayes action a^*, which other a's are almost as good, which a's are hopeless, and so on.

But what if she has only limited beliefs about (X, Y)? In this case, following the template outlined in Chapter 2, she represents her beliefs as best she can (Sec. 2.1 and Sec. 2.3), includes the observations as an additional belief (Sec. 2.4), and then computes a lower and an upper bound for $\mathrm{E}\{L(a, X)\}$ for each a (Sec. 2.2). Now she no longer has a complete ordering over the actions. What she has instead is a (strict) partial order, where $a \prec a'$ exactly when the upper bound of $\mathrm{E}\{L(a', X)\}$ is below the lower bound of $\mathrm{E}\{L(a, X)\}$. In the best situation, there will be an a^* for which $a \prec a^*$ for every $a \neq a^*$. Then she will have a unique optimal action. Otherwise, the best she can do is identify a set $\mathcal{A}^* \subset \mathcal{A}$ such that each member of $\mathcal{A}^*$ is the rightmost element in some sequence; see Figure 5.1. All of the actions not in $\mathcal{A}^*$ are ruled out, but an additional principle is required to select a single action from within $\mathcal{A}^*$.

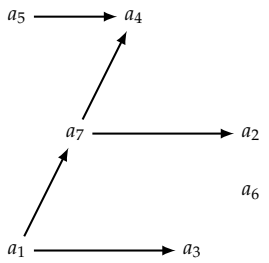

Figure 5.1: A strict partial ordering on $\mathcal{A}$, in which there is a directed path from a to a' exactly when $a \prec a'$. $\mathcal{A}^* = \{a_2, a_3, a_6, a_7\}$.

So limited beliefs allow us to rule out bad actions, but not necessarily to select a single best action. This should not present difficulties, because only the easily-quantified aspects of loss can be represented in the loss function. Even in the case where she has a complete ordering on $\mathcal{A}$, the client ought always to consider additional non-quantifiable aspects of each action, and, as outlined in Sec. 5.1, the purpose of decision theory is to narrow down the set of actions she needs to consider.

5.3 *Prediction problems*

A prediction problem is a special case of a decision problem, in which the objective is to predict the value of the random quantity

X based on the value of Y. This prediction is represented in terms of an action space $\mathcal{A} = \mathbb{R}$, a decision rule $\delta : \mathcal{Y} \to \mathbb{R}$, and a loss function $L : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$. Thus $L(x', x)$ is the expected loss that follows from predicting x' when the actual value is x. Letting the prediction be a value in $\mathbb{R}$ rather than its subset $\mathcal{X}$ is a useful generalisation. If necessary the action space can be restricted to $\mathcal{X}$ by setting $L(x', x) \leftarrow \infty$ when $x' \notin \mathcal{X}$.

There seem to be two main types of prediction problem. In the first, the action is a simple function of the prediction, and hence the loss function on the action translates to a loss function on the prediction. In this type of prediction problem, different clients, or the same client in two different situations, can have different loss functions. Milner and Rougier (2014) give a straightforward example. The amount of a drug that vet gives an animal is a simple function of its weight. With some drugs, such as wormers, an under-dose is more serious than an over-dose. With other drugs, such as anaesthetics, the opposite is true. Therefore the vet's loss function for her prediction of an animal's weight is asymmetric, reflecting the different losses from under- and over-dosing, and depending on the drug.[5] This type of prediction problem falls under the general framework of Sec. 5.1.

[5] In the field, a vet predicts an animal's weight using easier-to-make observations like girth.

The other type of prediction problem is the 'generic' prediction. In this case the client is a specialist tasked with making a prediction about X, which will then be used by others in their own decision problems. So the client might be, e.g., a meteorologist, a vulcanologist, or a hydrologist. In these areas there is a wide range of stakeholders, each with their own action set and loss function. Ideally, each stakeholder would collaborate with a specialist and a statistican to choose an appropriate action or decision rule using a full set of beliefs about X. But this is often impractical, and costly. A quick-and-dirty alternative is for each stakeholder to approximate their expected loss using

$$\mathrm{E}\{L(a, X) \mid Y \doteq y\} \approx L[a, \tilde{X}(y)] \qquad (\dagger)$$

where $\tilde{X}$ is a 'generic' prediction of X based on the value of Y. This is a lot cheaper, and since cost is part of the loss, there will be many applications where (†) is the right choice, because the fees of the specialist and the statistician will outweight the benefit to the stakeholder from a better choice of action.

As an intriguing aside, note that if the loss function L is linear in some function $g(x)$ for each a then

$$\tilde{X}(y) \leftarrow g^{-1}\,\mathrm{E}\{g(X) \mid Y \doteq y\} \qquad (\ddagger)$$

is the optimal forecast, because in this case (†) is exact. But this is a strong restriction that would only be generally useful if all stakeholders shared the same g.[6]

[6] See the end of this section for where this might be reasonable.

From now on, we treat the client as the specialist, interested in making a good 'generic' prediction about X, helpful to a wide range of stakeholders.

Then there are two widely accepted features of generic prediction problems. First, that the loss function is convex. A loss function is convex exactly when

$$L(x', x) = h(x' - x)$$

for some non-negative convex function h with $h(0) = 0$. In other words, the minimum loss is incurred for a correct prediction, and the loss rises at least proportionately with the error in the prediction. Convex loss functions represent the situation where a near-miss is tolerable, but a large miss is intolerable. Second, that the function h is approximately symmetric in X. After all, as the vetinerary example shows, if losses are asymmetric for one stakeholder, then it is quite possible that they are asymmetric in the other direction for another stakeholder. But it is also possible to represent the sitation where the same asymmetry exists for all stakeholders, as discussed below.

Where both of these common features hold, there is a 'generic' loss function for prediction, the *quadratic loss function*

$$L(x', x) := c(x' - x)^2 \quad c > 0.$$

This follows from the Taylor series expansion

$$\begin{aligned} L(x', x) &= h(x' - x) \\ &= h(0) + (x' - x)\, h'(0) + \tfrac{1}{2}(x' - x)^2\, h''(0) + \text{small terms} \\ &\approx c(x' - x)^2 \end{aligned}$$

where $c := \frac{1}{2}h''(0)$, since $h(0) = h'(0) = 0$, $h''(0) > 0$, and $h'''(0) \approx 0$ for a symmetric function. The quadratic loss function is therefore often taken as the default loss function for a prediction problem, accepting that convexity is a reasonable property, and that the client's loss function has no obvious asymmetry. So the following result is very useful in practice.

Theorem 5.2. *If the loss function is quadratic, and beliefs are not affected by the prediction, then the Bayes Rule is*

$$\delta^*(y) = \mathrm{E}\{X \mid Y \doteq y\}.$$

Proof. Here is a proof that does not involve differentiation. The Bayes Rule theorem and the additional condition about beliefs assert that

$$\delta^*(y) = \operatorname*{argmin}_{x' \in \mathcal{X}} \mathrm{E}\{L(x', X) \mid Y \doteq y\}. \tag{5.2}$$

So let $\psi(y) := \mathrm{E}(X \mid Y \doteq y)$. Then

$$\begin{aligned} L(x', X) &\propto (x' - X)^2 \\ &= (x' - \psi(y) + \psi(y) - X)^2 \\ &= (x' - \psi(y))^2 + 2(x' - \psi(y))(\psi(y) - X) + (\psi(y) - X)^2. \end{aligned}$$

Take expectations conditional on $Y \doteq y$ to get

$$\mathrm{E}\{L(x', X) \mid Y \doteq y\} \propto (x' - \psi(y))^2 + \mathrm{E}\{(\psi(y) - X)^2 \mid Y \doteq y\}.$$

Only the first term contains x', and this term is minimised over x' by setting $x' = \psi(y)$, as was to be shown.[7] □

[7] There is a straightforward extension to the case where $\boldsymbol{X}$ is a vector.

Note that Thm 5.2 is explicitly a property of the quadratic loss function (and those convex loss functions that a quadratic loss function approximates). It is *not* the case that we can always interpret a hypothetical expectation as a prediction, because one cannot make a useful prediction without thinking first about losses, and losses definitely do not have to be quadratic. But we can turn this result around, as a device for assessing hypothetical expectations. I could envisage my hypothetical expectation as the prediction I would make were I to incur a quadratic penalty. Personally, I do not find this helpful, except in a formal sense. Goldstein and Wooff (2007) provide a detailed investigation of this viewpoint, which origined with Bruno de Finetti (see, e.g., de Finetti, 1974).

The condition in Thm 5.2 that the client's prediction does not affect her beliefs about X seems reasonable in many situations, but not all. If the client is the Chief Economist in a Central Bank, then her public prediction of next year's interest rates will affect the market for bonds, and thus affect next year's interest rate. In fact, her prediction of interest rates is an instrument for changing interest rates, and thus affects her beliefs, and other people's too.[8]

[8] The precise effect is very subtle, because of the game-theoretic elements in the way that a Central Bank manages its superior information about financial markets, in the light of its public and private targets.

The quadratic loss function approximates a more general loss function that is convex and symmetric on an interval scale. There are some predictions where another scale seems more appropriate. For example, for population, a loss function on a ratio scale might be better, giving rise to

$$L(x', x) = h(\log(x') - \log(x))$$

for a convex symmetric h. Following the same argument as above, the Bayes Rule prediction in this case is

$$\tilde{X}(y) = \exp \mathrm{E}\{\log(X) \mid Y \doteq y\}$$

with a quadratic approximation. Applying Jensen's inequality (Thm 2.1), this prediction is never larger than that on the interval scale; it is highly likely to be smaller.

5.4 *Estimators and admissibility*

All of this talk of 'Bayes Rules' undoubtedly makes the Frequentist from Sec. 4.4 very nervous. And yet, as discussed in that section, he is also in desparate need of a prediction for the parameter θ to plug into his inference; naturally, he would like to use an optimal prediction if he can. A prediction for the parameter is termed an *estimator*.

The Frequentist has specified a family of PMF's for (X, Y) which is indexed by the parameter θ, written

$$\mathrm{p}(x, y; \theta) \quad \text{for some } \theta \in \Omega.$$

The Frequentist can accept the notion of a loss function, since without it there is no optimality, but not the notion that θ should have a probability distribution. So he is not able to compute an expectation such as $\mathrm{E}\{L[\delta(Y), X]\}$, although he can compute an expectation such as $\mathrm{E}\{L[\delta(Y), X]; \theta\}$. This turns out to be enough to provide a framework to rule out some obviously suboptimal estimators. But it also opens a can of worms.

The *risk function* is defined as

$$R(\delta, \theta) := \mathrm{E}\{L[\delta(Y), X]; \theta\}.$$

If we imagine all of the possible decision rules in one long column, and all the elements of Ω in one long row, then the risk function can be represented as a possibly huge matrix R, with $R_{ij} := R(\delta_i, \theta^{(j)})$. Among the rows we are likely to find rules that are dominated by other rules, in the sense that

$$R_{ij} \geq R_{i'j} \text{ for all } j, \text{ and } R_{ij} > R_{i'j} \text{ for at least one } j.$$

In other words no matter what the value of θ, rule i has a risk which is no lower than that of rule i', and which is higher for at least one value of θ. Rule i is said to be dominated by rule i', or to be *inadmissible*. If a rule is not inadmissible, it is *admissible*. It would appear to be an embarassing error to use a rule which was inadmissible. Thus it appears as though all statisticians require require necessary and sufficient conditions for a rule to be admissible.

A sufficient condition is easy to derive. A rule δ is admissible if it is a Bayes Rule for some distribution π_θ with support on the whole of θ. For suppose that δ^* were a Bayes rule for π_θ, but that it was inadmissible. In this case there would be a δ' for which

$$R(\delta^*, \theta) \geq R(\delta', \theta) \quad \text{for all } \theta \in \Omega,$$

with a strict inequality for at least one θ. But since for the Bayesian

$$\mathrm{E}\{L[\delta(Y), X]\} = \sum_t R(\delta, t)\, \pi_\theta(t)$$

by the LIE (Thm 3.5), δ^* inadmissible would imply that

$$\mathrm{E}\{L[\delta^*(Y), X]\} > \mathrm{E}\{L[\delta'(Y), X]\},$$

and hence δ^* could not be a Bayes Rule for π_θ, contradicting the original claim. So in fact, Bayesians are largely unconcerned about admissibility: they are automatically admissible.

Much more grippingly, however, there is a converse to this result, first proved by Abraham Wald in 1950 and the subject of much further refinement since; see, e.g., Berger (1985, ch. 8).[9] This converse states that every admissible rule is either a Bayes Rule, or

[9] This converse is not hard to prove in the case where Ω is finite: see Schervish (1995, ch. 3) or Cox and Hinkley (1974, ch. 11).

the improper limit of a sequence of Bayes Rules (in the case where the parameter space is unbounded). But this does not guarantee that the improper limit is actually admissible—some are, but some are not.

This this result is very troubling for the *maximum likelihood estimator (MLE)*,

$$\hat{\theta}(y) := \operatorname*{argmax}_{t \in \Omega} \mathrm{p}(y; t)$$

which appeared in Sec. 4.4 and Sec. 4.7. The MLE is typically not a Bayes Rule for quadratic loss according to some proper prior distribution π_θ. So it fails the sufficient condition. Happily, it often satisfies the necessary condition, being the improper limit of a sequence of Bayes Rules, e.g. from a sequence of π_θ's tending to a flat prior on an unbounded Ω. So there is hope! Unhappily, this is not enough. *Stein's paradox* was a bombshell dropped in the early 1950s, which showed that, in almost the simplest possible case, the MLE was inadmissible for quadratic loss; see Efron and Morris (1977).

This puts the Frequentist in a difficult position. As discussed in Sec. 4.7, the MLE has exactly the invariance property that is required to make sense of an inference made over a family of distributions; not to mention that it is often fairly easy to compute in closed form, or numerically. But it is, typically, inadmissible under quadratic loss. Moreover, the Bayesian's decision to provide a prior distribution π_θ appears to have been vindicated, although this does not make the actual specification of his prior distribution any easier.

This just goes to show that statistical inference is very complicated, involving trade-offs and compromises as the statistician attempts to stitch a defensible line between what is theoretically attractive and what is feasible. The main issue, it seems to me, is that this difficult task should not be undertaken in ignorance. If a statistician chooses to use the MLE as a plug-in estimator for his inference, he must be able to defend his decision. This is the first question I would ask, as an auditor: Did you check that the MLE in your problem is admissible, and, if it is not, why did you use it anyway? I'd be interested in the answer, but, as an auditor, I would also want to see whether that the statistician was knowledgeable and thoughtful. I would be concerned about a statistician who had unknowingly used an inadmissible estimator—what other alarming things might he also have done?

5.5 *The Netflix Prize*

This section is about a modern high-profile prediction problem. It was the prototype Big Data prediction competition, and kicked off a new industry, represented, for example, by the website `http://www.kaggle.com`. In the following, I will simplify slightly. Details are available at `https://en.wikipedia.org/wiki/Netflix_Prize`.

Imagine a huge matrix $\boldsymbol{X}$, where X_{um} is the rating that user u would give to movie m.[10] A small number of the elements of $\boldsymbol{X}$ have been filled-in with actual ratings, on a scale of 1 to 5 stars. Netflix wants to predict the ratings that user u would give to the movies she has not rated, in order to make a helpful recommendation. According to Netflix, a good prediction has a small quadratic loss.

[10] In deference to the N. American origin of this competition, I write 'movie' instead of 'film'.

Netflix organised a competition to find a rating prediction rule that did at least 10% better than its in-house rule. They supplied part of their database for 'training', and kept the rest back for final evaluation, and also for updating a leaderboard during the competition. The data they supplied were in the form of a large number of triples $\boldsymbol{y}^{\text{obs}} := \{(u_i, m_i, r_i)\}_{i=1}^{n}$, where r_i is the rating that user u_i gave to movie m_i.

This is a well-posed prediction problem. A statistician applying the principles of decision theory would represent his beliefs about people and movies as expectations of functions of $\boldsymbol{X}$, or, more comprehensively, as a PMF, $\mathrm{p}_{\boldsymbol{X}}$. Then he would predict the rating given by user u for movie m as $\mathrm{E}\{X_{um} \mid \boldsymbol{Y} \doteq \boldsymbol{y}^{\text{obs}}\}$. In representing his beliefs, he might consult a psychologist, perhaps even talk to people who know people who know big-name Hollywood producers. Unfortunately this statistician, although he might make a useful prediction, will not win the competition; in fact it is unlikely that he will do as well as the Netflix in-house system. It is important to understand why this is.

One reason, which I will set aside, is that this is a Big Data problem, and that many of the calculations that the statistician needs to do will be beyond the scale of his resources. However, this is not the main reason that the statistician will be out-performed.

The main reason is the peculiar nature of $\boldsymbol{X}$. The Netflix random quantities $\boldsymbol{X}$ have the feature that $\boldsymbol{X}$ can be partitioned into sets that are like each other. On this basis Netflix assumes that, for such a partition, a prediction rule that works well on one set will also work well on another. If $\boldsymbol{y}^{\text{obs}}$ is an element of such a partition, and Netflix can find a prediction system that works well on $\boldsymbol{y}^{\text{obs}}$, then it will also work well on all the missing values in $\boldsymbol{X}$.

A process that partitions $\boldsymbol{X}$ into subsets that are like each other might be termed *ignorable*. There are many ways of making a partition of $\boldsymbol{X}$ that are *not* ignorable. For example, partitioning according to whether the user is a man or a woman. I doubt that a prediction rule which works well for women will work as well for men. The users have partitioned $\boldsymbol{X}$ into the set of rated movies and its complement. I doubt that this process is ignorable; in other words, I believe that there are systematic differences in preferences between users who rate movies and those who do not. This implies that a 10% improvement in the measured performance of the rule does not necessarily translate into a 10% improvement in recommendations. But maybe this does not perturb Netflix, who have garnered lots of publicity from their competition.

One process which is definitely ignorable is to partition at random. The training dataset $\boldsymbol{y}^{\text{obs}}$ was a random sample from all of the rated movies, and the evaluation dataset was a random sample from the remainder. Thus competitors could replicate the Netflix scoring procedure by training their aglorithms on a subset of the training dataset—a random subset, naturally—and evaluating it on the remainder. Because the Netflix training dataset was so large, a competitor's training/training subset could still be large, allowing them to develop complex rules. Competitors with huge computing resources could generate a large number of rules using their training/training subset, keep the best ones according to their training/evaluation subset, and iterate the process. Competitors without the computing resources to do this were rapidly eclipsed.

So when will computer scientists make better predictions of unobserved components of $\boldsymbol{X}$ than statisticians? Primarily, when the process selecting the observations is ignorable. It will also help if the number of observations is very large, although this can be somewhat finessed by *cross validation*, in which a series of different randomly-chosen subsets are used for training and evaluation. Finally, the client must put safeguards in place to avoid *overfitting*. Netflix did this by separating the final evaluation dataset from the leaderboard dataset. Thus competitors knew that adjusting their prediction rule purely to climb the leaderboard would not necessarily benefit them in the final assessment.

I would say that most prediction problems fall outside these conditions. For a start, most collections of random quantities lack the simple structure that is a necessary condition for ignorability. This requires replication over different instances of the same basic unit: different people, different schools or hospitals, different regions, and so on. The Netflix random quantities had almost the simplest structure possible: replication over triples of the form (u, m, X_{um}). With this type of structure, ignorability follows from random sampling, but only if there is complete compliance. Random sampling is very popular for surveys, but ignorability does not follow because some people in the sample choose not to respond, and some choose to respond facetiously. The Netflix obervations were not randomly sampled but self-selected. Wherever humans exercise choices, it is prudent to assume that these choices vary systematically with their preferences.